UNIWERSYTET EKONOMICZNY W KRAKOWIE
DZIEDZINA NAUKI: NAUKI SPOŁECZNE
DYSCYPLINA NAUKOWA: NAUKI O ZARZĄDZANIU I JAKOŚCI

mgr Grzegorz Migut

Identyfikacja optymalnej ścieżki budowy modeli *data mining* w obszarze retencji klientów

Praca doktorska

Promotor
dr hab. Mariusz Łapczyński, prof. UEK

Kraków, 2023

Dla ukochanej żony,
z podziękowaniem za jej trud i poświęcenie.

Spis treści

Wstęp

W obecnych czasach ilość danych gromadzonych przez przedsiębiorstwa niezależnie od branży nieustannie rośnie. Rośnie również potrzeba syntezy informacji zawartych w tych danych i w konsekwencji optymalizacji procesów biznesowych. Jednym z kluczowych aspektów działania przedsiębiorstwa są aktywności związane ze sprzedażą i marketingiem. Dużym wyzwaniem w tym obszarze jest zdolność zrozumienia motywacji klientów, umiejętność dopasowania oferty do szybko zmieniających się preferencji nabywców oraz pogłębianie i poszerzanie relacji.

Wyzwanie to może być podejmowane poprzez budowę modeli ekonometrycznych oraz modeli uczenia maszynowego. Modele dobrej jakości przekładają się na trafne decyzje biznesowe, wspierają racjonalne gospodarowanie dostępnymi zasobami oraz umożliwiają opracowanie skutecznych strategii zatrzymania klientów.

Termin „jakość modelu" jest pojęciem pojemnym, który najczęściej rozumiany jest jako zdolność modelu do generalizacji, mierzona za pomocą wybranej miary dobroci dopasowania. Takie rozumienie jakości modelu wymaga w praktyce poszerzenia o wymiary związane z wartością wynikającą z interpretacji reguł odkrytych przez model czy też stabilnością działania modelu w czasie.

Do głównych determinant wpływających na jakość uzyskanych modeli należy zaliczyć: techniki czyszczenia danych oraz selekcji zmiennych, przyjętą metodę analityczną, strategie optymalizacji hiperparametrów oraz dodatkowe strategie uczenia, takie jak segmentacja analizowanego zbioru, hybrydyzacja czy agregacja modeli. Lepsze zrozumienie relacji między wymienionymi determinantami oraz zbadanie wpływu poszczególnych czynników na jakość modelu może przyczynić się do poprawy procesu jego budowy, skrócenia czasu jego powstawania oraz łatwiejszej pielęgnacji.

Merytoryczny obszar powyższych dociekań został ograniczony do modeli lojalności klienta. Wybór tego obszaru podyktowany został niesłabnącym zapotrzebowaniem rynku na budowę tego typu modeli. Utrzymanie klientów, zrozumienie czynników wpływających na ich lojalność i wreszcie umiejętność przewidywania ich

zachowania jest nieodmienne kluczowym aspektem pracy zespołów odpowiedzialnych za marketing oraz wsparcie procesów obsługi klienta.

Praca ma na celu wypełnienie luki badawczej w obszarze wieloaspektowej oceny jakości modeli retencji klientów. W literaturze brakuje również – w ocenie autora – opracowań syntezujących wpływ wielu determinant jakości modeli na końcowy rezultat modelowania. Synteza powyższych wymiarów może przyczynić się do identyfikacji strategii budowy modeli najczęściej prowadzących do uzyskania pożądanych przez badacza rezultatów. Wiedza oparta na wynikach badań może być przyczynkiem do opracowania narzędzi wspierających pracę analityka na przykład w postaci kreatorów w bezpłatnych lub komercyjnych narzędziach analitycznych.

Ze względu na eksploracyjny charakter pracy, hipotezy badawcze nie zostały sformułowane. W to miejsce określono cel główny pracy oraz cele poboczne.

Za cel główny obrano identyfikację determinant wpływających na jakość modeli migracji klientów oraz określenie relacji między nimi. Przyjętymi celami pobocznymi są:

1. Ocena wpływu wybranych technik czyszczenia i transformacji danych na jakość budowanych modeli klasyfikacyjnych.
2. Identyfikacja optymalnej ścieżki budowy modelu ekonometrycznego, na przykładzie regresji logistycznej, budowanego przy wykorzystaniu szeregu technik selekcji zmiennych opartych zarówno na filtrach, jak również metodach wbudowanych (np. LASSO) czy metodach opakowujących (metody krokowe, *Branch&Bound*).
3. Ocena skuteczności modeli drzew klasyfikacyjnych budowanych za pomocą alternatywnych ścieżek podziału.
4. Porównanie skuteczności działania modelu regresji logistycznej z modelami uczenia maszynowego zbudowanymi za pomocą drzew klasyfikacyjnych, perceptronu wielowarstwowego oraz drzew wzmacnianych.
5. Ocena wpływu hybrydyzacji, segmentacji oraz agregacji na jakość budowanych modeli.
6. Wykazanie możliwości budowy modeli o zadowalających własnościach za pomocą metod interpretowalnych przez badacza (biała skrzynka), porównywalnych z zaawansowanymi metodami nieinterpretowalnymi (czarna skrzynka), przy użyciu odpowiednich technik przygotowania danych oraz hybrydyzacji modeli.

Budowa modeli retencji klienta zrealizowana została zgodnie z paradygmatem budowy modeli *data mining* zakładającym wtórne wykorzystanie danych gromadzonych w wyniku realizacji standardowych procesów biznesowych. Implikuje to pracę na zastanych zbiorach danych zgromadzonych w systemach informatycznych przedsiębiorstw. W pracy wykorzystane zostało doświadczenie autora w budowie modeli retencji klienta na rzeczywistych zbiorach danych klientów z branży telekomunikacyjnej, ubezpieczeniowej i usługowej. Ze względu na ograniczenia związane z poufnością danych ich bezpośrednie wykorzystanie nie było możliwe podczas wykonanych symulacji i eksperymentów. Wnioski i wiedza płynące ze zrealizowanych projektów zostały uwzględniane w pracy w sposób pośredni. Podstawą analizy był zbiór danych dostępny w domenie publicznej (repozytoriach on-line) cechujący się wystarczającą złożonością oraz wolumenem.

Procedura badawcza polegała na wykonaniu badań symulacyjnych oceniających wpływ determinant wpływających na jakość modeli migracji klientów oraz określenie relacji między nimi. W oparciu o dostępny zbiór danych zbudowanych zostało szereg modeli zgodnie z metodyką CRISP-DM. Podczas symulacji wzięte zostały pod uwagę następujące czynniki:

- ***T**ransformation* – sposób przygotowania predyktorów, dyskretyzacja, standaryzacja itp.,
- ***I**nteraction* – fakt uzupełnienia zbioru danych o zmienne pochodne (*derived variables*),
- ***V**ariables* – sposób doboru zmiennych do modelu,
- ***H**yperparameters* – metody optymalizacji hiperparametrów,
- ***E**nsembles* – dodatkowe strategie uczenia: segmentacja, hybrydyzacja, agregacja modeli.

Aspekty TIVHE były brane pod uwagę w sposób uwzględniający specyfikę wykorzystywanych metod analitycznych.

Punktem odniesienia oraz podstawą do szczegółowych porównań był model ekonometryczny zbudowany za pomocą regresji logistycznej. Budowa modelu logistycznego była realizowana zgodnie z koncepcją, która zakłada, że zbudowany model jest wypadkową oczekiwań biznesu oraz chęci sprostania wymaganiom statystycznym. Metodyka zakłada dyskretyzację predyktorów oraz przekodowanie istniejących kategorii

do wartości WoE (*Weight of Evidence*). Podejście to zapewnia liniowy wpływ zmiennych na logarytm szansy modelowanego zjawiska. Dodatkowo dzięki temu przekształceniu redukowany jest wpływ wartości odstających na model, a także w naturalny sposób imputowane są braki danych, które tworzą osobną kategorię – wartość WoE. W przypadku modelu regresji logistycznej kryteria identyfikacji modelu zostały poszerzone o: liczbę zmiennych w finalnym modelu, zgodność znaków ocen parametrów regresji z ich biznesową interpretacją oraz analizę poziomu współliniowości predyktorów.

Zasadniczą część analizy poprzedził etap zrozumienia i przygotowania danych. Dostępne zmienne zostały zidentyfikowane pod kątem skali pomiarowej na jakiej są mierzone. Ocenione zostały rozkłady zmiennych ze szczególnym naciskiem na rozkład zmiennej zależnej oraz kwestie jej zbilansowania.

W kolejnym kroku wykonana została analiza braków danych oraz przeprowadzona została ich imputacja. W zależności od charakteru zmiennej, charakteru oraz skali pomiarowej braku dobrana została odpowiednia metoda imputacji. Następnie przeprowadzona została wstępna selekcja zmiennych. Za pomocą filtrów odrzucone zostały te zmienne, które nie wykazywały współzmienności ze zmienną zależną.

W procesie identyfikacji zmiennych pochodnych wykorzystane zostało podejście automatyczne oparte na metodzie losowego lasu. Podejście to polega na zbudowaniu zestawu relatywnie płytkich drzew o głębokości 2 lub 3. W przypadku realizowanej pracy badawczej przyjęto głębokość na poziomie 2. Po zbudowaniu zestawu drzew, każdy liść przekształcany jest na niezależną regułę opisującą podziały od korzenia drzewa do liścia. Fakt losowego doboru predyktorów dla każdego z drzew zaimplementowany w losowym lesie pozwolił na identyfikację potencjalnie interesujących reguł, niemożliwych do identyfikacji za pomocą tradycyjnych algorytmów drzew klasyfikacyjnych i regresyjnych. Uzyskane reguły zostały przefiltrowane pod kątem liczby przypadków (powszechność reguły) oraz na podstawie odsetka nielojalnych klientów spełniających tę regułę i wartości przyrostu (siła reguły). Najbardziej interesujące reguły zostały przekształcone na zmienne binarne (na zasadzie informacji, że dany przypadek spełnia regułę lub jej nie spełnia) i zasiliły zbiór potencjalnych predyktorów.

Dane, które przeszły wstępną selekcję zmiennych zostały poddane transformacjom mogącym potencjalnie ułatwić proces budowy modeli o pożądanych właściwościach. Pierwszym przekształceniem, jakie zastosowano dla pierwotnego zbioru danych była standaryzacja logistyczna zaproponowana przez Pyle'a. Ten rodzaj standaryzacji zastosowano dla predyktorów ilościowych, redukując tym samym skalę występowania

wartości odstających w analizowanym zbiorze. W analizie przyjęto granicę na poziomie 6 odchyleń standardowych. Powyżej tej granicy wartości predyktorów zostały zredukowane do wartości bliskich 0 lub 1 w zależności od kierunku odchylenia.

Drugim rodzajem zastosowanej standaryzacji będzie standaryzacja WoE. Wykonana została dla wszystkich predyktorów; zarówno dla predyktorów jakościowych, jak i dla poddanych uprzedniej dyskretyzacji predyktorów ilościowych. Dyskretyzacja zmiennych ilościowych została przeprowadzona dwoma sposobami: za pomocą podziału na decyle oraz za pomocą podziału algorytmem CART, przy założeniu, że maksymalna głębokość drzewa wynosi 5. Analizie poddane zostały zatem cztery zbiory danych: 1) oczyszczone zmienne surowe, 2) zmienne przekształcone za pomocą standaryzacji Pyle'a, 3) zmienne WoE na podstawie decyli (WoE-Decyle) i 4) zmienne WoE na podstawie algorytmu CART (WoE-CART).

Kontynuacją analizy była segmentacja statystyczno-ekspercka wykonana za pomocą algorytmu CART. Segmentacja umożliwiła podział zbioru danych na dwa podzbiory, dla których wykonano budowę odrębnych modeli migracji klientów. Podejście to pozwoliło zatem na budowę modeli hybrydowych, a w zależności od przyjętej metody modelowania był to na przykład *CART – logit*, bądź *CART – sieć neuronowa*. Poza modelami hybrydowymi do celów porównawczych zostały również zbudowane modele na podstawie całego zbioru danych, bez hybrydyzacji.

Analizowane zbiory danych różniły się zatem ze względu na fakt zastosowania w ich przypadku trzech rodzajów modyfikacji lub ich braku (w nawiasie podano liczbę wariantów): 1) standaryzacja zmiennych (4), 2) dodanie zmiennych pochodnych (2) oraz 3) segmentacja zbioru danych (2).

Daje to 16 wariantów zbiorów danych dla których wykonano badania. W pracy badawczej porównaniu zostały poddane powyższe kombinacje dla wybranych metod analitycznych. Dla każdego z wariantów wykonano optymalizację hiperparametrów. W obrębie danej metody przygotowano dodatkowo agregację najlepszych modeli w celu oceny wpływu tego zabiegu na uzyskaną siłę predykcyjną.

Łączna liczba modeli (większa od 10 000), głównie ze względu na przeprowadzoną optymalizację hipeparametrów o kilka rzędów wielkości przekraczała możliwości ręcznej parametryzacji i strojenia poszczególnych metod, stąd też krytycznym elementem procedury stało się opracowanie skryptów automatyzujących większą część obliczeń. Podczas analizy wykorzystano pakiet TIBCO Statistica, pakiety R oraz biblioteki dostępne w języku Python.

Praca składa się z pięciu rozdziałów. Pierwszy rozdział wprowadza czytelnika w zagadnienie retencji klientów jako obszaru modelowania marketingowego. Omówiono w nim rolę modelowania w badaniach marketingowych, przedstawiono koncepcję marketingu relacji w budowie lojalności klientów a także zaprezentowano definicje i determinanty lojalności nabywców. W dalszej części rozdziału omówione są rodzaje i poziomy lojalności konsumentów oraz wybrane podejścia pomiaru i modelowania lojalności. Rozdział zamyka przedstawienie metodyki budowy modeli *data mining* w obszarze retencji klientów.

Rozdział drugi podejmuje tematykę przygotowania danych na potrzeby budowy modeli retencji klientów. Omówiono w nim zagadnienia związane z określaniem kluczowych parametrów projektu analitycznego, oceną wiarygodności danych oraz identyfikacją i imputacją braków danych. W rozdziale przedstawiono także sposoby łagodzenia problemów związanych z niejednorodnym zbiorem danych oraz zaprezentowano techniki związane z transformacją zmiennych i przygotowaniem zmiennych pochodnych. Ostatnim zagadnieniem opisanym w tym rozdziale jest problem niezbilansowanego rozkładu zmiennej zależnej.

Rozdział trzeci przedstawia zagadnienia związane z budową optymalnego modelu klasyfikacyjnego. Omówione są w nim techniki wyboru zmiennych do budowy modelu, a także wybrane metody modelowania, począwszy od regresji logistycznej poprzez sieci neuronowe a skończywszy na zespołach modeli. Rozdział podejmuje również tematykę optymalizacji hiperparametrów i jej wpływu na wynik modelowania.

W rozdziale czwartym omawiane są kwestie odnoszące się do walidacji i stosowania modeli retencji klientów. Przedstawione zostały w nim miary dobroci dopasowania modeli retencji klientów uzupełnione o obszerną dyskusję ich wrażliwości na zjawisko niezbalansowanego rozkładu zmiennej zależnej. W kolejnych częściach rozdziału zaprezentowane zostały kwestie wyboru strategii walidacji modeli retencji klientów oraz określenia optymalnego punktu decyzyjnego. Rozdział uzupełnia zagadnienie związane z wyborem klientów, którzy wymagają kontaktu w ramach strategii sprzedażowych i retencyjnych.

Rozdział piąty przedstawia wyniki pracy badawczej związanej z identyfikacją optymalnej ścieżki selekcji modelu migracji klientów. W pierwszej części rozdziału zaprezentowano metodykę przeprowadzonych badań oraz cel wykonanych analiz. W kolejnej części przedstawione są kolejne kroki pracy badawczej związane ze zrozumieniem i przygotowaniem danych. Rozdział wieńczy część związana z zasadniczym

eksperymentem analitycznym mającym na celu identyfikację kluczowych czynników wpływających na jakość modeli migracji klientów.

Rozdział 1
Retencja klientów jako obszar modelowania marketingowego

1.1. Rola modelowania w badaniach marketingowych

Rzeczywistość biznesowa otaczająca przedsiębiorstwa działające na rynku jest niezwykle złożona i skomplikowana. Dodatkową trudność w jej zrozumieniu wprowadza jej dynamiczny charakter. Skuteczne działanie wymaga zatem odpowiednich narzędzi, które pozwolą na racjonalne i umiejętne wsparcie procesów decyzyjnych. Naturalnymi instrumentami używanym przez każdego człowieka pozwalającymi radzić sobie ze złożoną rzeczywistością są modele. Modele, którymi człowiek posługuje się w życiu codziennym są wypadkową spuścizny społecznej przekazywanej w formie tradycji, zasad, uprzedzeń, stereotypów czy praw oraz modeli budowanych na bazie własnych doświadczeń. Skuteczność takich modeli zależy od prawdziwości przesłanek, na których są budowane oraz umiejętności ich twórcy do ich dostosowania do zmieniającej się rzeczywistości.

Badania naukowe, w tym także badania marketingowe oraz praktyka menedżerska to obszary, w których modele wykorzystywane są w sposób powszechny. Modele mogą przyjmować różną formę. P. Leeflang i inni [2013, 2016] wyróżniają cztery podstawowe ich reprezentacje. Najbardziej oszczędnymi w formie są modele niejawne oparte na doświadczeniu i intuicji decydenta. Nie są one w żaden sposób formalnie zapisane, a tworzone są na bieżąco w jego umyśle. Kolejny poziom to modele polegające na ustnym formułowaniu zasad skutecznego działania w formie twierdzeń opisujących relacje pomiędzy podjętymi krokami a efektem tych kroków. G. Lilien i inni [1995] również wymieniają tę formę jako jeden ze sposobów tworzenia modeli. Modele oparte na werbalnych stwierdzeniach nie muszą mieć charakteru naukowego i mogą wynikać z indywidualnych doświadczeń osób zajmujących się marketingiem. Granica między

naukowością a praktycznym doświadczeniem nie musi być przy tym jednoznacznie postawiona. Nauka może być bowiem traktowana jako wyszkolony i zorganizowany zdrowy rozsądek [Huxley, 1854][1]. Trzeci poziom reprezentują modele sformalizowane, przedstawione w formie pisemnej. Przykładem mogą być logiczne modele przepływu będące swoistym zapisem modeli werbalnych czy modele objaśniające relacje pomiędzy zjawiskami rynkowymi za pomocą zmiennych w formie zestawu zdań logicznych.

Najbardziej złożonym sposobem reprezentacji są modele określone numerycznie na podstawie przyjętych założeń oraz zgromadzonych danych empirycznych bądź symulacyjnych. Modele budowane zgodnie z tą strategią opisują zjawiska za pomocą równań bądź nierówności matematycznych przy użyciu symboli oznaczających zmienne marketingowe. Strategia ta zaczęła być stosowana od lat 60 XX wieku [Wierenga, 2008]. W tym czasie w organizacjach pojawiła się możliwość wykorzystania komputerów do obliczeń matematycznych a rozwój technologii informatycznych umożliwił gromadzenie większych niż dotychczas ilości danych marketingowych, co istotne w ilości niepozwalającej na efektywną ich tradycyjną analizę. Naturalnym następstwem tych okoliczności stała się potrzeba formułowania modeli matematycznych. Dodatkowym czynnikiem wzmacniającym pojawienie się modeli matematycznych w marketingu stało się przesunięcie metodologii stosowanej w naukach o zarządzaniu w kierunku metodologii stosowanej w naukach przyrodniczych, w których powszechnie stosowano modele matematyczne.

Marketing jest interdyscyplinarną dziedziną nauki[2], stosującą wiele różnorodnych podejść badawczych opartych na różnych niejednokrotnie sprzecznych paradygmatach. W badaniach marketingowych wykorzystywane są paradygmaty nauk społecznych [Mazurek-Łopacińska, Sobocińska, 2013]:

- neopozytywistyczno-funkcjonalistyczno-systemowy,
- interpretatywno symboliczny,
- radykalnego strukturalizmu,
- postmodernistyczny.

Każdy z wymienionych paradygmatów bazuje na odmiennych założeniach. Pierwszy z nich zakłada, że wiedza ma charakter obiektywny. Faworyzuje on podejście analityczne,

[1] Za [Lilien i inni, 1995].
[2] Formalnie rzecz ujmując marketing do subdyscyplina nauk o zarządzaniu.

zakładające możliwość uogólniania i matematycznego modelowania wyników badań marketingowych. Modele budowane zgodnie z tym paradygmatem oparte są na podejściu ilościowym. Według pozostałych paradygmatów wnioski płynące z badań nie muszą mieć charakteru obiektywnego. Dominują w nich metodyka jakościowa wykorzystująca różnorodne strategie i modele. Bez większego wysiłku można dostrzec analogię pomiędzy paradygmatami a modelami (budowanymi niezależnie od przyjętego paradygmatu). Paradygmaty określając ramy metodologiczne, w których porusza się badacz również w naturalny sposób tworzą pewien model, zgodnie z którym opisywana jest rzeczywistość będąca przedmiotem dociekań naukowych.

Mimo, że nie można odmówić wartości żadnemu z wymienionych paradygmatów, dominującym jest paradygmat neopozytytywistyczno-funkcjonalistyczno-systemowy [Mazurek-Łopacińska, Sobocińska, 2013]. Pozwala on na opis relacji pomiędzy podejmowanymi działaniami marketingowymi a uzyskanymi efektami za pomocą metod ilościowych, wykorzystując zmienne oraz zależności przyczynowo-skutkowe. Paradygmat ten szczególnie mocno kładzie nacisk na budowanie modeli określonych numerycznie, budowanych przy użyciu metod ilościowych opartych na statystyce, programowaniu liniowym czy metodach uczenia maszynowego. Modele budowane w tym obszarze są związane z opisem, wyjaśnianiem i predykcją reakcji rynkowych na ofertę marketingową przedsiębiorstw [Sagan, 2016].

Numeryczne modele marketingowe stanowią złożony i niejednorodny zbiór. Kolejna oś podziału może przebiegać na linii celu dla jakiego budowane są modele. Rozróżniane są tutaj modele opisowe (deskryptywne), predykcyjne oraz preskrypcyjne [Leeflang i inni, 2016]. Zadaniem modeli opisowych jest pomoc w zrozumieniu badanych procesów. Modele tego typu mają na przykład tłumaczyć relacje popytu i podaży na rynkach czy też wskazać jakie instrumenty marketingowe mają wpływ na sprzedaż. Mają też na celu uogólnienie analizowanych zjawisk marketingowych. Głównym celem modeli predykcyjnych jest poprawne przewidywanie przyszłych obserwacji. Do ich budowy wykorzystuje się zaawansowane modele *data mining*, duży nacisk kładąc na zdolności generalizacyjne budowanych modeli. Modele normatywne (preskrypcyjne) mają na celu ocenę kierunków działań marketingowych i wybór najlepszego wariantu z punktu widzenia przyjętego kryterium [Sagan, 2016].

Ostatnim sposobem rozróżniania modeli marketingowych jest poziom agregacji. Budowane modele mogą opisywać zależności na poziomie rynku, sklepu czy segmentu (modele makromarketingowe) oraz gospodarstwa domowego czy pojedynczego

konsumenta (modele mikromarketingowe). Modele makromarketingowe mają charakter opisowy. Jedną z głównych zalet analizy na poziomie zagregowanym są niewielkie wymagania dotyczące danych. Modele takie dają jednak mniej szczegółowy obraz wpływu przyjętych instrumentów marketingowych na zachowanie konsumentów [Fok, 2013]. Obszar wykorzystania modeli mikromarketingowych jest bardziej uniwersalny, mogą one służyć zarówno do opisu, jak i predykcji czy preskrypcji.

Na bazie paradygmatów stosowanych w marketingu rozwinęło się wiele szkół i tradycji badawczych, których wyczerpujący przegląd można znaleźć na przykład w pracy A. Sagana [2014]. W ramach istniejących szkół wyróżnić można dwa podstawowe kierunki rozwoju modeli marketingowych [Sagan, 2016]. Pierwszy odnoszący się do pomiaru i analizy zjawisk związanych z konsumentem oraz jego działaniami, skutkujący budową modeli mikromarketingowych służących zarówno do opisu jak również predykcji oraz preskrypcji. Analizowane zmienne dotyczą głównie cech społeczno-demograficznych konsumentów, ich profilu psychograficznego i behawioralnego. Na ich podstawie budowane są modele zachowań lub segmentacji rynku związane między innymi z preferencjami, satysfakcją czy lojalnością klienta. Drugi kierunek związany jest z budową modeli makromarketingowych związanych z funkcjonowaniem całych rynków lub sektorów. Są one najczęściej modelami eksplanacyjnymi oraz teoriopoznawczymi.

Jedną ze szkół marketingu stosujących pierwszy z wymienionych kierunków rozwoju jest szkoła marketingu relacyjnego a w szczególności wywodząca się z niego tradycja badawcza oparta na CRM (*Customer Relationship Managment*). Jest ona zgodna z dominującym od początku wieku podejściem skoncentrowanym na kliencie [Wierenga, B., Van der Lans, 2008] podkreślającym wagę jego potrzeb. Działania marketingowe są podporządkowane budowaniu relacji z klientem, zwiększaniu jego lojalności oraz satysfakcji co docelowo ma przekładać się na zwiększenie jego dochodowości wyrażonej najczęściej za pomocą wartości życiowej klienta (*CLV, Customer Lifetime Value*). Strategie te stosowane są z powodzeniem w stosunku do klientów na rynkach masowych, w oparciu o informacje zawarte w bazach danych eksplorowanych za pomocą zaawansowanych technik uczenia maszynowego. Charakter relacji w tym ujęciu jest zautomatyzowany, zarządzany przez systemy informatyczne, za pomocą których realizowana jest strategia masowej indywidualizacji [Sagan, 2014].

1.2. Koncepcja marketingu relacji w budowie lojalności klientów

Marketing relacji jest jedną z ważniejszych obecnie szkół marketingu. Podwaliny pod jego koncepcję położone zostały w latach 60. oraz 70. ubiegłego stulecia. Jedna z pierwszych definicji tego podejścia wskazywała na konieczność pozyskania, utrzymania i rozwoju relacji z nabywcami [Berry, 1983][3]. Częścią wspólną wielu proponowanych obecnie definicji tego terminu jest akcent położony na budowanie trwałych i długotrwałych więzi z klientami w celu wzrostu wartości dla obu stron [Łapczyński, 2016]. Ważnym aspektem jest tutaj retencja klientów oraz znalezienie optymalnego balansu pomiędzy zdobywaniem nowych klientów a utrzymaniem obecnych. Duży nacisk kładzie się także na optymalizację wartości życiowej klienta, która zakłada budowanie długotrwałych relacji, dbanie o jakość usług oraz wysoki poziom satysfakcji klienta. Podejście to zakłada budowanie relacji opartej na zaufaniu korzystnej zarówno dla dostawcy, jak i odbiorcy – klienta. Marketing relacyjny zakłada, że w proces tworzenia relacji z otoczeniem zaangażowani są nie tylko pracownicy działu marketingu, ale wszyscy pracownicy przedsiębiorstwa.

Zasady marketingu relacyjnego mogą być stosowane zarówno na rynkach B2B jak i B2C, a także na rynku masowym jak i niszowym. Oczywiście widoczne tutaj jednak będą różnice w specyfice tych relacji, jak również w narzędziach stosowanych w celu ich podtrzymania. Obszarem, w którym największe znaczenie w zarządzaniu relacjami z klientem ma wykorzystanie zaawansowanych modeli marketingowych jest niewątpliwie masowy rynek B2C. Masowość klienta i jego relatywnie niewielki wpływ na przychód firmy ograniczają możliwość zazwyczaj kosztownej indywidualnej opieki ze strony dostawcy oraz budowania rzeczywistych relacji. Zaawansowane algorytmy analizy danych umożliwiając masową indywidualizację mogą zapewnić zadowalający dla obydwóch stron substytut relacji. Warunkami umożliwiającym realizację strategii budowania więzi na rynkach masowych jest [Bhattacharya, Bolton, 2000]:

- dostateczne zróżnicowanie oferty dostawcy, umożliwiające dostosowanie jej do indywidualnych potrzeb nabywcy, im większy wybór tym większa szansa na zaangażowanie się konsumenta w relację z dostawcą,

[3] Za [Terlutter i Weinberg 2006].

- wykorzystanie mediów masowych do budowania więzi emocjonalnych z nabywcą na przykład poprzez przekaz reklamowy angażujący klientów w proces zakupowy,
- zapewnienie komunikacji pomiędzy klientem a dostawcą na przykład poprzez system obsługi reklamacji czy infolinii,
- zwiększenie liczby epizodów zakupowych, wpływ na to mogą mieć koszty zmiany dostawcy, oraz postrzegana jakość produktu wpływająca na poziom satysfakcji oraz lojalności.

Chęć zapewnienia indywidualizacji oferty na rynku masowym spowodowała, że marketing relacji zaczął wykorzystywać potencjał technologii informacyjnej. Odpowiedzią na potrzebę zarządzania relacjami z klientami stała się metodyka CRM będąca praktyczną realizacją filozofii marketingu relacyjnego wykorzystującą zaawansowane możliwości technologii informacyjnej. Metodyka ta nastawiona jest na optymalizację praktycznych działań nakierowanych na rozwój i pielęgnację relacji z klientami.

W CRM można wyróżnić trzy podstawowe poziomy (podsystemy): operacyjny, analityczny oraz interakcyjny [Migut, 2004]. CRM operacyjny ma na celu zbieranie danych transakcyjnych i danych o klientach, ich opinii o produktach, sprzedawcach i komunikacji. Zadaniem operacyjnego CRM jest również wsparcie telemarketingu. Do tego celu służą teleinformacyjne systemy automatycznego rozdzielania rozmów przychodzących (*Automatic Call Distribution*) czy też systemy interaktywnej obsługi głosowej (*Interactive Voice Response*). Zadaniem analitycznego CRM jest przetwarzanie i analiza danych w celu segmentacji klientów czy optymalizacji kampanii marketingowych. CRM interakcyjny ma na celu wspieranie bezpośrednich kontaktów z klientem za pomocą różnorodnych kanałów dystrybucji i komunikacji. Uzupełnieniem wyróżnionych podsystemów może być strategiczny CRM (*Strategic CRM*) związany z budowaniem strategii marketingowej oraz tworzeniem wartości dla klienta [Payne, 2005].

Z wymienionych tutaj podsystemów CRM, to poziom analityczny jest najmocniej związany z budową modeli marketingowych oraz podejmowaniem na ich podstawie strategicznych decyzji odnośnie różnego rodzaju akcji i kampanii marketingowych. Analityczny CRM jest często łączony z marketingiem opartym na bazach danych czy też z marketingiem wspieranym danymi [Łapczyński, 2016]. Termin ten jest również ściśle związany z pojęciem CI (*Customer Intelligence*), który można interpretować jako analityczne przetwarzanie danych o kliencie w celach marketingowych [Surma, 2009].

Do najważniejszych obszarów analitycznych wspierających doskonalenie relacji z klientem należy zaliczyć [Migut, 2004]:

- analizy związane z cyklem życia klienta,
- segmentację, mogącą pełnić pomocniczą rolę w stosunku do analiz cyklu życia klienta,
- analizę satysfakcji klientów.

Dwa pierwsze obszary mocniej wiążą się z marketingiem opartym na bazach danych polegającym na wykorzystaniu konsumenckich baz danych do zwiększenia skuteczności działań marketingowych. Trzeci obszar łączony jest z kolei silniej z marketingiem wspieranym danymi, w którym kładzie się nacisk na pomiar efektów działalności marketingowej. Obszar ten nie będzie przedmiotem rozważań w niniejszej pracy[4].

Analizy związane z cyklem życia klienta są kluczowymi zadaniami analitycznego CRM. Cykl życia klienta zakłada, że relacja pomiędzy nabywcą a dostawcą podlega ewolucji przechodząc z czasem przez kolejne fazy, poziomy zaangażowania. Proces nawiązywania relacji rozpoczyna się w momencie odebrania przez klienta przekazu marketingowego i pozytywnej reakcji na odebrany przekaz. Widocznym momentem nawiązania relacji jest dokonanie przez niego pierwszego zakupu. Początkowo siła relacji pomiędzy dostawcą a klientem jest niska, a klient dokonuje stosunkowo niewielu zakupów.

W kolejnej fazie więź staje się silniejsza, czego manifestacją jest większa częstotliwość oraz wielkość zakupów. Z czasem siła relacji słabnie a klient odchodzi do konkurencji. Na tej podstawie można wyróżnić trzy podstawowe fazy cyklu życia klienta [Wuebben, 2008]:

- faza pozyskiwania klienta,
- faza rozwoju klienta,
- faza utrzymania klienta.

Każda z nich stawia przed dostawcami innego rodzaju problemy i zadania. Podczas fazy pozyskiwania klienta podstawowe pytania menedżerów brzmią [Reinartz, Venkatesan, 2008]:

- jakich klientów powinno się pozyskiwać?

[4] Szczegółowe rozróżnienie obydwóch podejść przedstawiono w [Łapczyński, 2016].

– w jaki sposób należy to robić (jaki przekaz może być skuteczny)?

W fazie rozwoju klienta pytania mają postać:

– na rozwoju, których klientów należy się koncentrować?

– jak powinno się przydzielać zasoby, aby rozwój klientów był efektywny?

Analogiczne pytania w fazie utrzymania brzmią:

– których klientów powinno się zatrzymać?

– które relacje z klientami należy wygaszać?

– jak powinno się przydzielać zasoby, aby proces utrzymania klientów był efektywny?

Znajomość odpowiedzi na powyższe pytania pozwalają na realizację fundamentalnego celu jakim jest maksymalizacja wartości życiowej klienta (*Customer Lifetime Value, CLV, LTV*). Na wartość tą składają się zarówno kwestie związane z wielkością zakupów w danymi czasie jak i z czasem trwania relacji. Obydwie wielkości są ze sobą powiązane, a ich wzajemny wpływ na CLV może być wyrażony za pomocą wzoru [Leeflang i inni, 2016]:

$$CLV = \sum_{t=0}^{T} \frac{m_t \times r_t}{(1+i)^t} - AC$$

Gdzie:

AC – koszt pozyskania klienta,

T – horyzont czasu dla oszacowania CLV,

m_t – marża w czasie t (przychód – koszt),

r_t – wskaźnik retencji (prawdopodobieństwo, że klient będzie aktywny w czasie t)

i – stopa dyskontowa

Według M. Wuebbena [2008] na wartość klienta składa się ocena:

– długości relacji definiowanej jako czas trwania relacji,

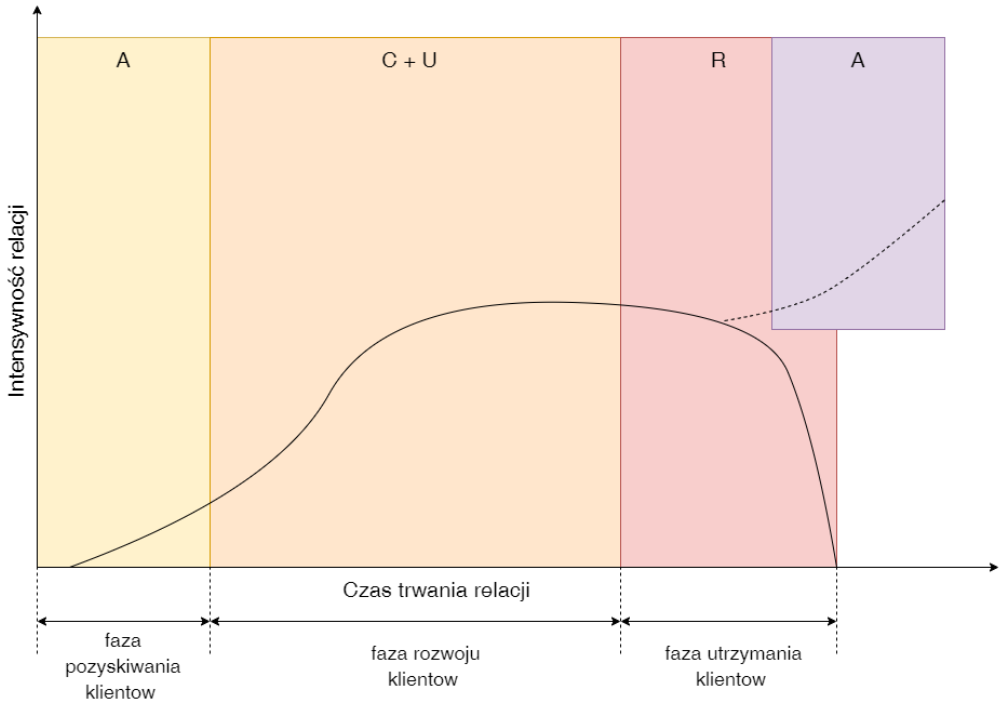
– głębokości relacji znajdującej odzwierciedlenie w częstotliwości korzystania z usług oraz skłonności klientów do wyboru produktów o wyższych marżach,

– szerokości relacji powiązanej z liczbą różnych produktów lub usług zakupionych od firmy w danym czasie.

Koncepcją spójną z przedstawionym powyżej cyklem życia klienta jest model ACURA [Łapczyński, 2016]. Nazwa modelu to skrót angielskich terminów:

- ***A**cquisition* (pozyskanie klienta),
- ***C**ross-sell* (sprzedaż krzyżowa),
- ***U**p-sell* (sprzedaż uzupełniająca),
- ***R**etention* (utrzymanie klientów),
- ***A**dvocacy* (promocja klienta na stanowisko ambasadora marki).

Relację pomiędzy modelem ACURA a cyklem życia klienta przedstawia Rysunek 1.



Rysunek 1 Cykl życia klienta a model ACURA
Źródło: M. Łapczyński [2016].

Faza pozyskania klienta pokrywa się z elementem ***A**cquisition*. Faza rozwoju klienta związana jest z członami ***C**ross-selling* oraz ***U**p-selling*. Faza utrzymania klienta jest natomiast równoważna z elementem ***R**etain*. Relacja może następnie zaniknąć lub rozwinąć się, co jest po części zgodne z ostatnim elementem modelu - ***A**dvocacy*.

Niezależnie od etapu cyklu życia, na którym znajduje się klient, jest on obiektem zainteresowania dostawcy, pragnącego poznać jego potrzeby, dostarczyć mu wartość w postaci produktu bądź usługi, a dzięki temu zapewnić satysfakcję oraz lojalność klienta.

Niezwykle istotne jest też poszerzenie zakresu współpracy oraz wydłużenie czasu jej trwania. Modele predykcyjne budowane za pomocą technik analizy danych mogą znacząco ułatwić pogłębianie relacji, zwłaszcza na rynku masowym, gdzie masowa indywidualizacja nie pozostawia alternatywy dla innych działań.

Na każdym etapie cyklu życia klienta pożądane jest trafne przewidywanie jego potrzeb i zachowań. Niniejsza praca koncentruje się jedynie na jednym z elementów modelu ACURA jakim jest retencja klientów. Należy mieć jednak na uwadze, iż modele, choć optymalizujące różne aspekty relacji z klientem powinny być rozpatrywane i stosowane w holistycznym kontekście wartości życiowej klienta. Jednie wtedy mogą zapewnić rzeczywiste zwiększenie skuteczności realizowanych działań.

Analizy segmentacyjne pozwalają na wyodrębnienie z niejednorodnej zbiorowości klientów jednorodnych grup o podobnych cechach oraz podobnie reagujących na podjęte wobec nich środki marketingowego oddziaływania. Jej wyniki mogą być podstawą dla masowej indywidualizacji – dany sposób oddziaływania przypisywany jest do całego segmentu klientów. Wyniki segmentacji pozwalają ponadto na lepsze zrozumienie cech klientów oraz poprawiają komunikację w zespole handlowców. Same wyniki segmentacji mogą sugerować działania, jakie można podjąć w stosunku do wyróżnionych grup. Wyniki segmentacji mogą być podstawą do dalszych pogłębionych analiz na przykład związanych z cyklem życia klienta. Szczegółowe analizy wykonywane są nie dla całego niejednorodnego zbioru klientów, lecz co jest bardziej skuteczne i poprawne merytorycznie, dla wybranego homogenicznego segmentu[5].

Do podejść często wykorzystywanych do segmentacji rynku można zaliczyć analizę według następujących zmiennych [McDonald, Dunbar, 2013, Kotler i inni, 2020]:

- Oferowanych produktów, usług i kanałów sprzedaży - pozwala to na zrozumienie, które konkretne cechy produktów oddziałują na różnych klientów, a w następstwie na opracowanie propozycji opartych na różnych potrzebach klientów.
- Danych demograficznych, które wspomagają identyfikację charakterystycznych cech klientów należących do danego segmentu pomagając przez to w określeniu optymalnego kanału komunikacji czy sposobu dotarcia do klientów.
- Danych geograficznych – pozwala to na podział rynku względem miejsca zamieszkania klientów, bazując na różnym poziomie szczegółowości: od państwa

[5] Dyskusja związana z analizą niejednorodnego zbioru danych została przedstawiona w rozdziale 2.

czy regionu po pojedyncze dzielnice czy ulice. Dane geograficzne same z siebie nie mogą definiować segmentu ze względu na niewielką moc wyjaśniania postaw zakupowych klientów. Wynika to z dużej niejednorodności klientów zamieszkujących na danym obszarze (problem analogiczny do segmentacji demograficznej). Zmienne te mogą jednak pomóc wskazać najbardziej prawdopodobne lokalizacje klientów z wyróżnionych segmentów ułatwiając tym samym sposób dotarcia do nich.

- Danych związanych z kanałami dystrybucji i komunikacji – wybrane segmenty klientów mogą być powiązane z określonymi formami komunikacji czy sprzedaży, ich znajomość może pozwolić na przyjęcie adekwatnych sposobów komunikowania się. Istotną kwestią jest tutaj zrozumienie motywu wyboru określonego kanału przez klienta.
- Danych psychograficznych – takie zmienne umożliwiają podział kupujących na różne segmenty na podstawie ich stylu życia lub cech osobowości. Identyfikacja wewnętrznych czynników determinujących zachowanie klientów może umożliwić określenie sposobu komunikacji z nimi, zwiększając szansę przyciągnięcia ich uwagi i dotarcie do nich z odpowiednim przekazem.
- Danych behawioralnych, które pozwalają na wyodrębnienie segmentów na podstawie zachowań zakupowych klientów, ich reakcji na produkt lub stosowane wobec nich środki marketingowego przekazu. Dane behawioralne mogą być postrzegane jako najlepszy punkt wyjścia do budowy segmentów rynku [Kotler i inni, 2020], natomiast pozostałe zmienne mogą stanowić ich – cenne poznawczo – uzupełnienie.

1.3. Lojalność nabywców – definicje i determinanty

Zagadnienie lojalności klienta doczekało się w literaturze szeregu definicji oraz prób opisu. Bogate ich zestawienie można znaleźć w licznych pracach polskich autorów [Skowron, Skowron, 2012, Rudawska, 2005, Urban, Siemieniako, 2008]. W definicjach tych można dostrzec dualizm będący z jednej strony efektem wpływu dwóch odrębnych szkół marketingu: marketingu relacji oraz marketingu transakcyjnego, a z drugiej wieloaspektowego wymiaru tego pojęcia.

Lojalność klientów w ujęciu marketingu transakcyjnego koncentruje się przede wszystkim na aspekcie behawioralnym. Lojalny klient cechuje się wysoką powtarzalnością zakupów, nabywaniem innych produktów i usług danego dostawcy. Dodatkowo charakteryzuje się mniejszą skłonnością do zmiany dostawcy nawet w przypadku pewnych braków w jego ofercie lub sposobie obsługi, a także rekomendowaniem dostawcy znajomym. Innym aspektem lojalności w aspekcie behawioralnym jest doradzanie dostawcy korzystnych dla niego zmian. Lojalne zachowania klienta nie zawsze muszą wynikać z jego poczucia więzi z dostawcą. Mogą być swoistym „małżeństwem z rozsądku" związanym z brakiem wyboru, wygodą, przyzwyczajeniem czy lenistwem.

Drugie podejście – wywodzące się z marketingu relacji – podkreśla kluczową rolę przeżyć wewnętrznych klienta, jego nastawienia do dostawcy oraz poziomu przywiązania. Relacyjny punkt widzenia jest mocno widoczny w pracy M. Tesławskiego [2012], który lojalnego klienta utożsamia z lojalnością emocjonalną i przywiązaniem do marki, które skutkują zakupami niewymagającymi dodatkowych zachęt oraz rekomendacjami innym uczestnikom rynku. Aspekt behawioralny w postaci częstych i dużych zakupów nie jest w tym ujęciu istotą lojalności, a jedynie jej niekoniecznym aczkolwiek mile widzianym skutkiem ubocznym.

Inne ujęcia lojalności akceptują dualizm tego zjawiska uwzględniając zarówno jego wymiar behawioralny, jak również aspekt postawy klienta wobec dostawcy. Częste zakupy negatywnie nastawionego klienta, podobnie jak głębokie pozytywne zaangażowanie nie mające wymiaru zakupowego nie mogą być utożsamiane z lojalnością. Definicja łącząca oba aspekty określa lojalność jako postawę w stosunku do konkretnych obiektów związanych z dostawcą, prowadzącą do wyrażania zachowań lojalnościowych [Urban, Siemieniako, 2008].

Zachowania zakupowe noszące zewnętrzne znamiona lojalności mogą zatem, ale nie muszą wynikać z jego wewnętrznego nastawienia, postawy. Postawa klienta może być w dużej mierze scharakteryzowana poprzez poziom jego przywiązania do dostawcy. Powyższe ujęcia lojalności umożliwiają przeprowadzenie kategoryzacji klientów biorącej pod uwagę poziom natężenia obydwu wymiarów. Na tej podstawie można zaproponować podział klientów na cztery różne kategorie (Rysunek 2).

Rysunek 2 Typologia lojalności klientów
Źródło: Opracowanie własne na podstawie [Skowron, Skowron, 2012].

Zaproponowana typologia wpisuje się w opisywaną diadę ”postawa-zachowanie” i opisuje różne kombinacje wartości tych wymiarów. Prawdziwie lojalni klienci to ci wyrażający swoją lojalność zarówno za pomocą zachowania jak i wewnętrznego pozytywnego nastawienia. Na drugim biegunie znajdują się klienci nielojalni o zgodnym, negatywnym natężeniu nastawienia oraz zachowania. Pozostałe dwie grupy to klienci charakteryzujący się pozytywnym nastawieniem jednak nieokazujący go w sposób praktyczny (ukryta lojalność) oraz klienci przejawiający zachowania lojalne niewynikające z pozytywnego nastawienia, a z innych zewnętrznych czynników (bierna lojalność).

Zachowania zakupowe klienta, pomimo że posiadające zewnętrzne znamiona lojalności nie muszą wynikać z jego wewnętrznego nastawienia, ale mogą być determinowane przez inne czynniki. Oczywiście mogą one wynikać z jego świadomej postawy jednak mogą być także wynikiem jego nieuświadomionych, automatycznych procesów psychicznych. Końcowy efekt w postaci zakupu może być też mieszanką czynników automatycznych oraz wynikających ze świadomej postawy.

Wewnętrzna postawa klienta może być zmodyfikowana za pomocą czynników zewnętrznych takich jak działania marketingowe firmy i konkurencji, informacje przekazywane przez środki masowego przekazu oraz informacje uzyskiwane drogą nieformalną. Lojalne zachowania automatyczne mogą być z kolei wzmacniane przez wielokrotne przypominanie klientowi o danym produkcie w sposób wywołujący określone skojarzenie emocjonalne. Innym czynnikiem wzmacniającym automatyczne postawy może być rola norm społecznych i chęć działania zgodnie z nimi. Czynnik społeczny pełni dodatkowo ważną rolę w kształtowaniu się postaw lojalnościowych związanych z chęcią wyróżnienia się od innych czy też przynależności do określonej grupy [Urban, Siemieniako, 2008].

Lojalna postawa oraz jej fizyczna realizacja może wynikać z wielu wzajemnie powiązanych motywów. Ich znajomość pozwala na bardziej świadome kształtowanie oferty firmy oraz na dobór stosowanych przez nią środków marketingowego przekazu. Motywy te można zaliczyć do czterech grup [Urban, Siemieniako, 2008]:

- wynikających z relacji klienta z firmą,
- bazujących na relacjach społecznych,
- opartych na ocenie korzyści,
- powodowanych przymusem wewnętrznym.

Pierwsza grupa wynika z relacji klienta z firmą. Zalicza się do niej motywy związane z poczuciem więzi z personelem firmy lub wartościami, które ona reprezentuje. Grupa ta jest także związana z poczuciem bycia ważnym dla firmy, wpływem na jej ofertę oraz przekonaniem, że relacje będą w przyszłości kształtowały się zgodnie z oczekiwaniem klienta. Do motywów z tej grupy zalicza się też mniej pozytywne aspekty, jak chęć uniknięcia wysiłku poznawczego.

Kolejna grupa motywów ma swoje źródło w kwestiach społecznych, do których zaliczyć można potrzebę wyróżnienia się, bycia zauważonym, docenionym przez innych czy też z potrzebą przynależności do określonej grupy dzielącej podobne wartości czy zainteresowania.

Trzecią grupą motywów są motywy związane z oceną korzyści. W tej grupie mieszczą się motywy związane z możliwością uzyskania większej korzyści używania z produktu w porównaniu z produktami oferowanymi przez znaną klientowi konkurencję. Z drugiej strony mogą się wiązać z chęcią realizacji zakupu optymalnego pod względem cenowym. Innym motywem z tej grupy jest ponawianie zakupów ze względu na spodziewane przyszłe korzyści związane na przykład z nagrodą lub „darmowym" kolejnym produktem.

Ostatnia grupa motywów związana jest z odczuwanym przez klientów swoistym przymusem wewnętrznym. Mogą one wynikać z braku alternatywy w stosunku do obecnego dostawcy, wygody związanej z łatwą dostępnością dostawcy, bądź też wysokimi karami za zerwanie umów terminowych z dostawcą.

1.4. Rodzaje i poziomy lojalności konsumentów

Pojęcie lojalności jest złożone i wieloaspektowe. Z tej złożoności wynika wiele prób jej definiowania oraz szereg prób jej kategoryzacji. Przedstawioną we wcześniejszym podrozdziale podstawową klasyfikację lojalności wynikającą z diady „postawa-zachowanie" warto uzupełnić o odmienne ujęcia. Istotny podział lojalności wynika z obserwacji dynamicznego charakteru tego zjawiska. Lojalność ewoluuje, zmienia się w czasie. Patrząc na nią pod tym kątem można wyróżnić następujące jej poziomy [Studzińska, 2015, Skowron, Gąsior, 2017]:

- Lojalność poznawcza (*cognitive loyalty*) – jest to lojalność na wstępnym etapie. Ze swojej natury jest płytka i narażona na przerwanie. Klient podejmuje decyzje o zakupie na podstawie zaufania do marki oraz na podstawie wybranych atrybutów takich jak cena bądź jakość obsługi klienta. Klient na tym poziomie jest bardzo wrażliwy na ofertę konkurencji.
- Lojalność wynikająca z zauroczenia (*affective loyalty*) – pojawia się w sytuacji, gdy czynniki determinujące zakup z poprzedniego poziomu zostają wzmocnione przez satysfakcję z kontaktów z firmą. Wrażliwość klienta na zmianę upodobań jest niższa, jednak wciąż utrzymuje się na wysokim poziomie.
- Lojalność wynikająca z głębokiego przekonania (*conative loyalty*) – pojawia się w momencie wielokrotnie powtarzanego doświadczenia satysfakcji. Doświadczenie to buduje w kliencie przekonanie o wartości relacji z firmą. Przekonanie to niekoniecznie musi manifestować się wysoką częstotliwością zakupów.
- Lojalność czynna (*action loyalty*) – na tym poziomie wewnętrzna postawa jest manifestowana poprzez zakupy. Pozycja dostawcy staje się ugruntowana, a zakupy wchodzą w nawyk. Pojawia się niewrażliwość na oferty konkurencji.

Dynamiczny charakter lojalności oraz jej poziomy można przedstawić za pomocą piramidy lojalności [Hill, Alexander, 2003].

Rysunek 3 Piramida lojalności
Źródło: Opracowanie własne na podstawie [Hill, Alexander, 2003].

Prawdopodobni nabywcy to wszyscy kupujący dany produkt lub usługę na rynku. Są to osoby, które albo nie wiedzą o istnieniu oferowanego przez firmę produktu, bądź też nie wykazują chęci jego zakupu. Potencjalni klienci widzą potencjalną korzyść z zakupu produktu lub usługi, nie zdecydowali się jednak jeszcze na żadną aktywność w tym kierunku. Okazjonalni klienci czyli osoby, które sporadycznie dokonują zakupu produktu lub usługi, nie wykazują przy tym przywiązania do dostawcy. Stali klienci to osoby kupujące często, wykazujący wysokie przywiązanie do dostawcy, odczuwający pozytywne uczucia do firmy jednak z niewielką skłonnością do ich okazywania. Zwolennicy to osoby o podobnych cechach do stałych klientów, z tą różnicą, że pozytywne uczucia do firmy są przez nie przekuwane na rekomendacje produktu lub usługi innym osobom. Partnerzy to osoby cechujące się najsilniejszą formą relacji pomiędzy dostawcą a klientem. Dla obydwu stron relacja jest postrzegana jako korzystna.

Kolejną koncepcją wykorzystującą dynamiczne ujęcie lojalności jest przedstawienie klientów na pięciu szczeblach drabiny lojalności (Rysunek 4). Szczeble te odzwierciedlają z jednej strony poziomy lojalności, z drugiej zaś odnoszą się do procesu kreowania lojalnego klienta przez firmę.

Rysunek 4 Drabina lojalności
Źródło: Opracowanie własne na podstawie [Skowron, Skowron, 2012].

Pierwsze dwa stopnie odnoszą się do reflektanta oraz nabywcy. Na tych poziomach akcent kładziony jest na pozyskiwanie nowych klientów, zapewnienie im odpowiedniego poziomu obsługi oraz zaoferowanie oferty korzystniejszej w stosunku do konkurencji. Odpowiedni poziom obsługi może sprawić, że klienci przejdą na kolejne, wyższe poziomy, w których akcent kładziony jest na utrzymanie klientów. Na kolejnych etapach firma podejmuje działania mające na celu przywiązanie klienta do firmy, regularne zakupy, bliskie relacje z firmą, a na najwyższym poziomie – rzecznika – doprowadzenie do sytuacji, w której klient rozpowszechnia pozytywne informacje o firmie i jej ofercie.

Kolejny podział lojalności może odnosić się do czynników warunkujących zachowanie klientów. Czynniki te można podzielić na dwie grupy [Skowron, Gąsior, 2017]:

- o podłożu emocjonalnym – wynikające z więzi, jaka została nawiązana przez klienta z firmą,
- racjonalne - wynikające z wiedzy klienta oraz porównania oferty firmy z ofertami konkurencyjnymi.

Ważnym atrybutem lojalności jest podmiot, bądź przedmiot, który może być jej obiektem. W tym ujęciu lojalność można dzielić na [Studzińska, 2015]:

- lojalność wobec produktu bądź marki - klient jest przywiązany do konkretnej marki, wykazując się małą wrażliwością na produkty marek konkurencyjnych;

– lojalność wobec producenta- klient kupuje różne marki danego producenta, na przykład ze względu na jakość czy pochodzenie (krajowe / inne);

– lojalność wobec miejsca zakupu - klient kupuje w danym sklepie ze względu na dostępność lub presję czasu (np. blisko miejsca zamieszkania), poziom dochodu itp.

W. Urban oraz D. Siemieniako [2008] zwrócili uwagę na trzy determinanty lojalności, jakimi są 1) zaufanie, wiążące się z wiarą, że dostawca będzie dbał o interesy klienta, 2) przyzwyczajenie, rozumiane jako nawyk podejmowania działań wynikający z ich powtarzania oraz 3) zaangażowanie powodujące, że klient gotowy jest na podjęcie pewnych kosztów dla podtrzymania relacji z firmą. Dwa pierwsze czynniki są traktowane jako wymiary dychotomiczne. W przypadku zaangażowania rozpatruje się jego trzy poziomy: brak zaangażowania, zaangażowanie negatywne oraz zaangażowanie pozytywne. Kombinacje tych czynników prowadzą do szesnastu teoretycznych wariantów lojalności[6], spośród których autorzy wyróżniają 9 prawdopodobnych typów, których zestawienie przedstawia Tabela 1.

Tabela 1 Typy lojalności w funkcji jej determinant

Typ lojalności	Zaufanie	Przyzwyczajenie	Zaangażowanie
Świadoma	Tak	Nie	Brak
Z rozsądku	Tak	Tak	Brak
Zaangażowana	Tak	Nie	Pozytywne
Partnerska	Tak	Tak	Pozytywne
Z rutyny	Nie	Tak	Brak
Nieakceptowana z przymusu	Nie	Nie	Negatywne
Bezradna z przymusu	Nie	Tak	Negatywne
Wyrozumiała	Tak	Tak	Negatywne
Warunkowa	Tak	Nie	Negatywne

Źródło: Opracowanie własne na podstawie [Urban, Siemieniako, 2008].

Uwzględnienie w rozważaniach wymiaru przywiązania klienta oraz skłonności do kontynuowania relacji może prowadzić do typologii [Furtak, 2003 za Studzińska, 2015] przedstawionej w Tabela 2.

[6] Autorzy rozpatrują również warianty, w których zaangażowanie pozytywne oraz negatywne występują łącznie.

Tabela 2 Profile lojalności ze względu na przywiązanie oraz skłonność do kontynuacji związku

Profil lojalności	Przywiązanie	Skłonność do kontynuowania związku
Adwokaci	Wysokie	Wysoka
Bierni	Niskie	Wysoka
Tymczasowi	Niskie	Niska
Zdrajcy	Wysokie	Niska

Źródło: Opracowanie własne na podstawie [Studzińska, 2015].

Adwokaci, identyfikowani z „prawdziwą" lojalnością reprezentują wysokie przywiązanie do firmy oraz wykazują wysoką skłonność do kontynuowania relacji. Bierni realizują lojalność pozorną, ich skłonność do kontynuowania związku nie jest bowiem powiązana z przywiązaniem, a wynika z innych czynników. Klienci tymczasowi są wrażliwi na pojawienie się konkurencyjnych ofert ze względu na niski poziom przywiązania do firmy. Zdrajcy są skłonni do zmiany dostawcy pomimo przywiązania do firmy, a ich decyzja może wynikać z zewnętrznych przyczyn np. zmiany miejsca zamieszkania.

Ważnym czynnikiem wpływającym na lojalność jest satysfakcja. Warto zauważyć, że satysfakcja, podobnie jak lojalność sama w sobie jest pojęciem złożonym, mieszczącym w swoim obrębie wiele znaczeń i warstw. Satysfakcja może być odczuwana zarówno na poziomie emocjonalnym jak i poznawczym, może być wynikiem konkretnego zdarzenia, transakcji bądź też może być ogólnym odczuciem wynikającym z danej relacji. Może być zdefiniowana jako reakcja lub postawa wynikająca z różnicy pomiędzy oczekiwaniem klienta w stosunku do nabywanego produktu a faktycznym poziomem zaspokojenia tych oczekiwań. Ma więc ona charakter subiektywny, a ponadto kolejne zakupy danego produktu przez klienta, pomimo stałego poziomu zaspokojenia oczekiwań, mogą skutkować niższym jej poziomem ze względu na rosnący poziom jego oczekiwań. S. Skowron i Ł. Skowron [2012] wymieniają dwie kategorie satysfakcji:

- satysfakcję transakcyjną wynikającą z oceny samej transakcji zakupu, mającą charakter tymczasowy,
- satysfakcję skumulowaną budowaną na podstawie procesu użytkowania nabytego produktu.

Autorzy zwracają uwagę, że satysfakcja skumulowana jest czynnikiem, który bezpośrednio wpływa na budowę procesu lojalności wobec dostawcy bądź marki.

Skojarzenie lojalności z satysfakcją pozwalana na utworzenie kolejnej typologii klientów (Rysunek 5).

Rysunek 5 Podział klientów względem lojalności i satysfakcji
Źródło: Opracowanie własne na podstawie [Urban, Siemieniako, 2008].

Klienci cechujący się wysoką satysfakcją oraz lojalnością należą do grupy tzw. lojalistów, którzy rozpowszechniają pozytywne opinie o firmie i rekomendują innym jej produkty lub usługi. Klienci o wysokiej lojalności oraz niskiej satysfakcji nazywani są zakładnikami. Ich zachowanie może wynikać z braku alternatywy na rynku bądź wysokich kosztów rezygnacji z usług dostawcy. Osoby o niskim poziomie lojalności i jednocześnie wysoko usatysfakcjonowane z wartości dostarczanej przez firmę nazywane są „interesownymi". Pojawienie się konkurencyjnej oferty na rynku może najprawdopodobniej powodować zmianę przez nich dostawcy. Klienci nieprzejawiający ani wysokiej lojalności ani wysokiej satysfakcji zaliczani są do dwóch grup. Dezerterzy to klienci skłonni do rezygnacji z usług firmy bez skłonności do okazywania swojego ewentualnego niezadowolenia otoczeniu. Ich postawa odróżnia ich od tzw. szkodników, którzy wraz z odejściem rozpowszechniają negatywny wizerunek firmy.

Badania związków pomiędzy satysfakcją a lojalnością wskazują, że [Skowron, Skowron, 2012]:

- ponadprzeciętny poziom zadowolenia klienta wpływa na podniesienie się poziomu lojalności,
- związek pomiędzy oboma czynnikami jest wyraźniej zarysowany na rynku usług.

Należy także zwrócić uwagę, że zarówno lojalność jak i satysfakcja klienta są zjawiskami dynamicznymi, zmieniającymi się w czasie. Satysfakcja zmienia się wraz ze

zmieniającymi się oczekiwaniami konsumenta i podobnie przywiązanie klienta do dostawcy może z czasem ulegać osłabieniu lub wręcz zmienić swój kierunek. Czynnikami osłabiającymi postawę lojalności na etapie „dojrzałej" relacji pomiędzy klientem a dostawcą mogą być [Urban, Siemieniako, 2008]:

- potrzeba różnorodności wynikająca z nasycenia się klienta danym dobrem i chęci wprowadzenia urozmaiceń,
- skłonność klienta do zwiększania swoich oczekiwań prowadząca do spadku satysfakcji pojmowanej jako różnica pomiędzy jego oczekiwaniami a realizacją tych oczekiwań,
- niewystarczający poziom zainteresowania ze strony firmy oraz działania konkurencji.

Czynniki te mogą być osłabiane lub wzmacniane przez specyfikę produktów lub usług oferowanych przez dostawcę, czas trwania relacji czy też występowanie barier wyjścia. Warty podkreślenia jest indywidualny charakter reakcji każdego klienta na docierające do niego bodźce. Zakłada się jednak, że indywidualne zachowania mają swoje wspólne cechy i na poziomie masowym możliwe jest stosowanie modeli do segmentacji klientów oraz predykcji ich zachowań.

Satysfakcja klienta jest nieodzownym wymiarem wpływającym na ocenę relacji z klientem. Nie jest jednak jedynym czynnikiem budującym czy też wzmacniającym tę relację. Poza satysfakcją, innymi czynnikami branymi pod uwagę przez badaczy zjawiska lojalności są [Urban, Siemieniako, 2008]:

- jakość produktów lub usług,
- integracja wokół marki oparta na zaufaniu i zaangażowaniu klientów,
- wizerunek firmy,
- czynniki społeczne takie jak uznanie czy osobiste traktowanie.

Spośród wymienionych powyżej czynników na szczególną uwagę zasługuje aspekt jakości, czyli zespołu cech decydujących o ocenie danego produktu bądź usługi. S. Skowron, Ł. Skowron [2012] traktują wręcz triadę jakość, satysfakcja oraz lojalność jako bazę do modelowania zjawiska lojalności (i satysfakcji). Autorzy wyróżniają trzy składowe jakości wynikające z perspektywy marketingowej:

- jakość obiektywna, mierzalna, dotyczy tych atrybutów produktów, które da się zweryfikować za pomocą eksperckiej oceny,
- jakość doświadczana przez klienta, mająca charakter subiektywny,
- jakość w relacjach odnosząca się bardziej do procesu zakupowego i obsługi pozakupowej niż samej istoty produktu bądź usługi.

Z jakości nabywanych dóbr wynika wartość konsumencka produktu bądź usługi nieodzowna do uzyskania lojalności klientów. Wartość dla klienta może być rozumiana jako suma korzyści uzyskiwanych przez klientów, a wynikająca z kombinacji [Skowron, Skowron, 2012]:

- korzyści użytkowych wiążących się z zaspokajaniem podstawowych potrzeb klienta,
- korzyści ekonomicznych związanych z relacją pomiędzy ceną produktu lub usługi a jej subiektywnie postrzeganą jakością,
- korzyści indywidualnych, związanych z wygodą zakupu, prestiżem itp.

Pozostałe wymienione czynniki mogą uzupełniać wyróżnioną triadę i są używane przez twórców modeli wyjaśniających zjawisko lojalności.

Lojalność klienta może być także determinowana dostępnością alternatywnych ofert. Analiza tego wymiaru wraz ze wzmiankowaną wcześniej satysfakcją może doprowadzić do następnej klasyfikacji, w której kolejnym poziomom towarzyszy rosnąca liczba ofert alternatywnych [Curasi, Kennedy, 2002 za Studzińska, 2015]:

- więźniowie – klienci kontynuujący współpracę z firmą ze względu na brak alternatyw, relacji towarzyszy niski poziom satysfakcji,
- kupieni lojaliści – klienci współpracujący z firmą ze względu na korzyści finansowe, przy jednoczesnym niskim poziomie satysfakcji,
- oderwani lojaliści – klienci utrzymujący relację z firmą z powodu wysokich kosztów rezygnacji z usług firmy, pomimo niskiego poziomu satysfakcji,
- zadowoleni klienci – klienci, których potrzeby są w wysokim stopniu zaspokojone, odczuwający satysfakcję z relacji z firmą, niewidzący potrzeby przejścia do konkurencji,
- apostołowie – klienci wysoce usatysfakcjonowani z kontaktów z firmą, a dodatkowo dzielący się z innymi osobami swoimi pozytywnymi wrażeniami.

Innym wymiarem mogącym wchodzącym w interakcję z lojalnością jest dochodowość klientów. Rozpatrując łącznie obydwa wymiary można zdefiniować cztery typy klientów różniących się poziomem natężenia wymienionych cech. Szczegóły przedstawia Tabela 3**Błąd! Nieprawidłowy odsyłacz do zakładki: wskazuje na nią samą.**.

Tabela 3 Typologia klientów ze względu na poziom lojalności oraz poziom dochodowości

Typ klienta	Poziom lojalności	Poziom dochodowości
Motyl	Niski	Wysoki
Prawdziwy przyjaciel	Wysoki	Wysoki
Skorupiak	Wysoki	Niski
Obcy	Niski	Niski

Źródło: Opracowanie własne na podstawie [Skowron, Gąsior, 2017].

Motyle to klienci cechujący się wysokim poziomem dochodowości, charakteryzujący się przy tym wysoką skłonnością do zmiany dostawcy. Skorupiaki oraz obcy to klienci charakteryzujący się niskim poziomem dochodowości. Ci pierwsi są przy tym osobami o wysokim poziomie lojalności. Najbardziej wartościową grupą klientów są tzw. prawdziwi przyjaciele. Ich wysoki poziom dochodowości idzie w parze z wysoką lojalnością.

Zestawienie pojęcia lojalności wraz z dochodowością klientów ma szczególne znaczenie dla końcowego wyniku finansowego firmy. Zabieganie przez przedsiębiorstwa o lojalność klientów ma racjonalne uzasadnienie, jeżeli jest związane z korzyściami (wartością dla przedsiębiorstwa), do których zaliczyć można:

- brak kosztów związanych z pozyskaniem klientów,
- niższe koszty obsługi,
- akceptację wyższej ceny za oferowaną wartość,
- rekomendacje firmy innym osobom,
- możliwość podnoszenia lojalności przejawiającej się na przykład przez zwiększenie ilości zakupów.

Powyższe postulaty pomimo, że wydają się rozsądne i oczywiste, to nie zawsze znajdują odzwierciedlenie w rzeczywistości. Można wykazać, że [Reinartz, Kumar, 2002]:

- istnieją lojalni klienci, których obsługa jest droższa od obsługi klientów nowopozyskanych,

- lojalni klienci są w stanie wynegocjować lepsze warunki cenowe w porównaniu do pozostałych klientów,
- jedynie część lojalnych klientów poleca produkt innym osobom.

Z kolei R. East i inni [2016] wskazują jedynie niewielki związek (współczynnik korelacji równy 0,09) pomiędzy długością relacji z klientem a wysokością wydawanych przez niego środków. Większe wydatki lojalnych klientów były obserwowane jedynie w niewielu badanych obszarach. Autorzy wskazują na wyższe koszty obsługi lojalnych klientów, a także na brak lub wręcz negatywny związek pomiędzy długością relacji a poziomem rekomendacji. Lojalni klienci mogą również cechować się większą wrażliwością na zwiększenie poziomu cen.

Spostrzeżenia te prowadzą do konkluzji, że przywiązanie klienta do marki nie może być jedynym wyznacznikiem jego wartości. Działania retencyjne podejmowanie w odniesieniu do klientów powinny być poprzedzone dogłębną analizą innych ich cech, zwłaszcza ich dochodowości. W stosunku do lojalnych, dochodowych klientów powinno się podejmować kroki mające na celu w pierwszej kolejności zrozumienie przyczyn ich utraty, najlepiej na poziomie ich motywów i postaw. Utrata klientów może wynikać z obniżenia postrzeganego przez klienta poziomu wartości oferowanych produktów lub usług, negatywnego rozpatrzenia reklamacji czy też działań konkurencji. Może to też nastąpić bez specjalnej przyczyny. W tej sytuacji rzeczywistym powodem jest najczęściej brak relacji, która mogłaby powstrzymać klienta przed odejściem. Kolejnym krokiem jest dogłębna analiza zachowań klientów wsparta budową modeli marketingowych umożliwiających selektywnie działanie w ramach masowej indywidualizacji. Wyniki analiz mogą być następnie podstawą do działań zmierzających do podniesienia lojalności najbardziej wartościowych klientów oraz wzrostu wartości tych już lojalnych.

1.5. Sposoby pomiaru i modelowania lojalności – wybrane podejścia

Jak wskazano we wcześniejszej części rozdziału, lojalność może być postrzegana jako złożenie wewnętrznej postawy klienta z jej zewnętrzną realizacją. Obydwa aspekty nie muszą być ze sobą zgodne, co prowadzi do wyróżnienia określonych typów lojalności. W każdym ze swoich wymiarów można wskazać składowe lojalności, które mogą być

przedmiotem praktycznej oceny jej poziomu. W odniesieniu do zachowań związanych z lojalnością można wyróżnić następujące aspekty [Urban, Siemieniako, 2008]:

- powtórne zakupy produktów,
- poziom wierności marce kupowanej przez klienta,
- wielkość środków finansowych przeznaczanych na wydatki u dostawcy,
- inne zachowania niezwiązane z zakupami.

Do miar lojalności związanych z postawami czy ocenami klientów zaliczyć można z kolei [Skowron, Skowron, 2012]:

- czas trwania relacji,
- częstotliwość korzystania z oferty,
- zadowolenie klienta,
- wrażliwość na niekorzystne doświadczenia z produktem lub usługą,
- ryzyko rezygnacji na rzecz konkurencji.

Dobór miar lojalności klientów uzależniony jest dodatkowo od perspektywy badawczej. Inne miary dobierane będą w sytuacji badania realizowanego z punktu widzenia oddziaływania na pojedynczych klientów, inne w przypadku oceny samej firmy pod kątem lojalności jej klientów. Z perspektywy niniejszej pracy znacznie istotniejszy wydaje się być drugi z wymienionych aspektów. Istotnym wymiarem pozwalającym ocenić poziom lojalności jest czas trwania relacji, który może być podstawą do formułowania opartych na nich wskaźników. Pierwszym z nich jest prosta stopa zatrzymania klientów określana wzorem [Urban, Siemieniako, 2008]:

$$Stopa\ zastrzymania\ klientów = \frac{liczba\ klientów\ powtarzających\ zakupy\ w\ bieżącym\ okresie}{liczba\ klientów\ z\ poprzedniego\ okresu} \times 100\%$$

Jej modyfikacją jest stopa zatrzymania klientów ważona wielkością zakupów dokonanych przez klientów w okresach uwzględnionych w analizie. Wskaźnikiem komplementarnym do stopy zatrzymania klientów jest stopa utraty klienta, obydwa wskaźniki powinny sumować się do 100%.

Bardziej złożoną miarą opisującą poziom lojalności jest wskaźnik RFM złożony z trzech podstawowych miar związanych z:

- aktualnością relacji z klientem (*recency*) powiązaną z czasem trwania relacji. Wymiar ten wyrażany jest za pomocą czasu jaki upłynął od ostatniego zakupu w stosunku do momentu dokonania pomiaru;
- częstością zakupu produktu przez klienta (*frequency*) wyrażaną na przykład poprzez liczbę zakupów w określonym okresie;
- wielkością wydatków na produkt (*money* lub *monetary*).

Wskaźnik ten zakłada, że najbardziej prawdopodobny jest zakup przez klientów, którzy nabyli produkt niedawno, kupowali go często i wydali na zakupy wysokie kwoty. Niespełnienie tych warunków w poszczególnych wymiarach wskaźnika RFM adekwatnie obniża prawdopodobieństwo zakupu. Wskaźniki te są niewątpliwe cennymi miarami pozwalającymi uzupełnić obraz związany z lojalnością klienta. Oczywiście nie wyczerpują możliwości jego wielowymiarowej oceny. Na podstawie powyższych wymiarów wskaźnik RFM można obliczać w inny sposób, np. jako ważoną sumę poszczególnych wymiarów, które dodatkowo mogą podlegać wcześniejszej dyskretyzacji. W modelach *data mining* wymiary te pełnią rolę potencjalnie cennych predyktorów wpierając proces predykcji zmiennych zależnych odnoszących się do modelu ACURA, zwłaszcza na bardziej dojrzałych etapach relacji z klientem.

Nieco inną propozycję dekompozycji zjawiska lojalności na mierzalne wskaźniki można znaleźć w pracy R. East i inni [2016]. Lojalność może być opisywana za pomocą trzech zmiennych zastępczych odnoszących się do oceny:

- udziału (*share*) produktu danej marki w koszyku produktów danej kategorii u analizowanego klienta[7];
- retencji (*retention*) odnoszącej się do czasu korzystania przez klienta z produktu danej marki;
- rekomendacji (*recommendation*) przez konsumenta produktu danej marki innym klientom.

Miarą opartą na wymiarze udziału może być również wykorzystywany w branży finansowej oraz detalicznej wskaźnik SOW (*share of wallet*) oceniający w jakim stopniu

[7] Pomiar udziału możliwy jest na przykład poprzez wykorzystanie panelu konsumentów i regularne badanie ich koszyków zakupów. Badania tego typu pozwalają wyróżnić różne wzorce lojalności od konsumentów wiernych danej marce, poprzez klientów dzielących swoją lojalność pomiędzy markami, zmieniającymi swoje preferencje aż po konsumentów nie wykazujących oznak lojalności.

dany klient zaspokaja wszystkie swoje potrzeby u danego dostawcy [Skowron, Skowron, 2012].

Wskaźniki stosowane do oceny lojalności mogą bazować na różnych jej aspektach. Pod uwagę brany może być wymiar [Skowron, Gąsior, 2017]: behawioralny, afektywny, włożonego wysiłku lub mieszany.

Wskaźnikiem odnoszącym się do lojalności behawioralnej jest wskaźnik CLR (*Customer Loyalty Ratio*). Określa on, jaka frakcja spośród badanych osób deklaruje warunkową lub bezwarunkową chęć zakupu. Bezwarunkowa chęć zakupu dotyczy sytuacji deklaracji zakupu pomimo porównywalnej oferty konkurencji oraz konieczności przezwyciężenia pewnych niedogodności związanych z podjęciem działania. Deklaracja warunkowa wyklucza możliwość wystąpienia niedogodności. Ze względu na deklaratywny charakter odpowiedzi wskazana jest ostrożna ocena uzyskanej wartości. Wskaźnik ten jest wyrażony wzorem [Kozielski, 2008 za Skowron, Gąsior, 2017]:

$$CLR = \frac{liczba\ klientów\ deklarujacych\ chęć\ zakupu}{liczba\ badanych} \times 100\%$$

Wzmocniona wersja wskaźnika w liczniku bierze pod uwagę jedynie klientów deklarujących bezwarunkową chęć zakupu.

Afektywny aspekt lojalności mierzy wskaźnik NPS (*Net Promoter Score*), który w swoim zamyśle ma oceniać skłonność klientów firmy do udzielenia rekomendacji jej produktów lub usług. Wskaźnik ten bazuje na 11-punktowej skali oceniającej skłonność klienta do polecania firmy swoim znajomym. Wartość 0 oznacza brak skłonności do rekomendacji, wartość 10 pewność rekomendacji. Udzielone odpowiedzi podlegają dyskretyzacji na trzy grupy, promotorów (odpowiedzi 9 lub 10), pasywnie zadowolonych (odpowiedzi 7 lub 8) oraz malkontentów (odpowiedzi poniżej 7). Na tej podstawie wskaźnik można wyrazić za pomocą wzoru:

$$NPS = \frac{liczba\ promotorów - liczba\ malkontentów}{liczba\ badanych} \times 100\%$$

Do zalet tego wskaźnika zaliczyć można uniwersalność, krótki czas badania oraz dużą wartość diagnostyczną. Wadą jest wrażliwość na sytuacje w których świadomi klienci podają zawyżone oceny w celu uniknięcia dodatkowych pytań.

Aspekt włożonego wysiłku jest z kolei mierzony przez wskaźnik CES (*Customer Effort Score*). Jego wartość określa się na podstawie odpowiedzi respondentów na

pięciostopniowej skali oceniającej wysiłek włożony przez klienta w celu rozwiązania problemu, jaki pojawił się w związku z zakupem lub użytkowaniem produktu. Niewielki wysiłek powiązany jest ze skłonnością do powtórnych zakupów, z drugiej strony wysoki wysiłek zwiększa skłonność klientów do rozpowszechniania negatywnych ocen na temat firmy.

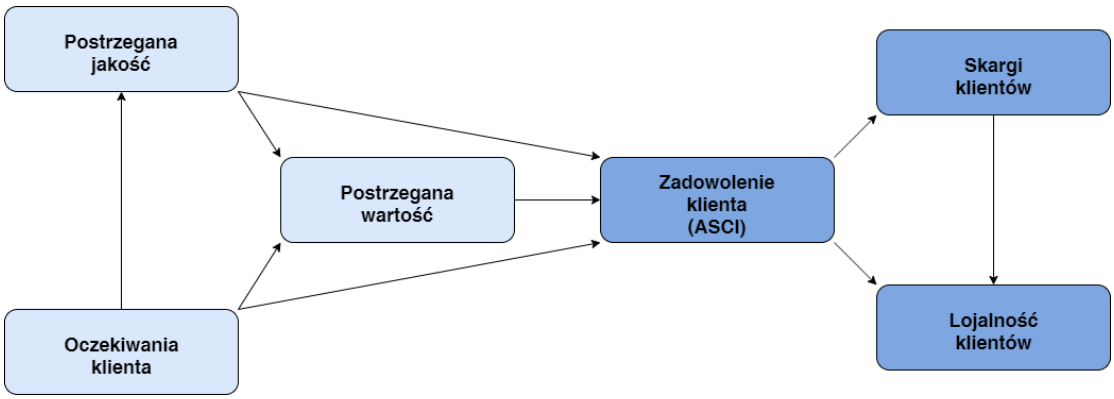
Wskaźnikiem biorącym pod uwagę wiele aspektów związanych z lojalnością jest wskaźnik TRI*M. Jego nazwa pochodzi od trzech słów *Management – Monitoring – Measurement*. Wskaźnik obejmuje następujące elementy [Drapińska, 2013]:

- ogólną ocenę firmy przez klienta,
- skłonność klientów do rekomendacji,
- skłonność do podtrzymywania relacji z przedsiębiorstwem,
- korzyści czerpane ze współpracy na tle potencjalnych korzyści płynących ze współpracy z konkurencją.

Wskaźnik ten przyjmuje wartości od 0 do 100 punktów. Przyjmuje się, że klienci lojalni to osoby uzyskujące powyżej 70 punktów.

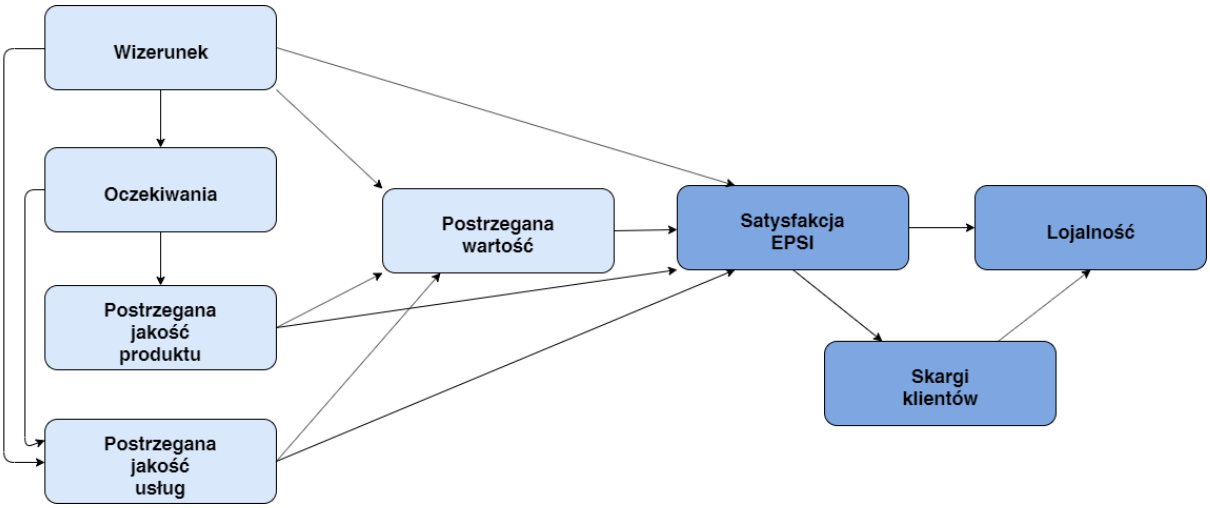
Wymienione powyżej cztery wskaźniki definiowały lojalność nie łącząc jej z satysfakcją klienta z oferty firmy. W drugim podejściu bada się złożone relacje pomiędzy satysfakcją, lojalnością oraz innymi elementami. Jednym z najbardziej popularnych wskaźników [Drapińska, 2013] jest wskaźnik CSI (*Customer Satisfaction Index*). Jest on obliczany jako średnia ważona ocen zadowolenia klienta z poszczególnych atrybutów produktu. Jego związek z lojalnością wynika z założenia, że satysfakcja leży u podstaw lojalności, a zadowolony klient dokona ponownych zakupów.

Bardziej złożone ujęcie satysfakcji i lojalności można odnaleźć we wskaźnikach ACSI (*American Customer Satisfaction Index*) oraz EPSI (*European Performance Satisfaction Index*) powstałych na bazie modelu SCSB (*Swedish Customer Satisfaction Barometer*).



Rysunek 6 Model ASCI (American Customer Satisfaction Index)
Źródło: Opracowanie własne na podstawie [Skowron, Skowron, 2012].

Model ACSI zakłada, że wzrost lojalności klientów wynika ze wzrostu poziomu satysfakcji oraz spadku poziomu skarg.



Rysunek 7 Model EPSI (European Customer Satisfaction Index)
Źródło: Opracowanie własne na podstawie [Drapińska, 2013].

Podobną zależność zaobserwować można w modelu EPSI różniącym się od poprzednika liczbą branych pod uwagę niezależnych konstruktów. Badania przeprowadzone z zastosowaniem tego modelu wykazują dodatnią korelację pomiędzy satysfakcją a lojalnością klientów [Drapińska, 2103].

Spośród wymienionych przez R. Easta [2016] mierzalnych składników lojalności klientów tj. udziału, retencji oraz rekomendacji, w sposób szczególny w niniejszym opracowaniu potraktowana zostanie retencja klientów. Brak lojalności w tym wymiarze wiązał będzie się ze zjawiskiem migracji klientów, którą można rozumieć jako częściową

lub całkowitą rezygnację klienta z produktów bądź usług oferowanych przez dostawcę [Łapczyński, 2016].

Samych przyczyn migracji jest wiele, jednak można je podzielić na klika głównych kategorii [Berry, Linoff, 2004, Mattison, 2005, Łapczyński, 2016]:

- migracja dobrowolna (*voluntary churn*) wynikająca z inicjatywy klienta, która dzielić może się dalej na
    - przypadkową,
    - zaplanowaną,
- migracja wymuszona (*invluntary churn*), w której inicjatywa oraz decyzja o zerwaniu relacji podejmowana jest przez dostawcę,
- migracja oczekiwana, wynikająca z naturalnej zmiany potrzeb klienta związanej na przykład z fazami cyklu życia rodziny.

Przyczyny przypadkowe niezwiązane są z siłą i jakością relacji z dostawcą. Związane są z poważną zmianą w życiu klienta na przykład utratą pracy, koniecznością zmiany miejsca zamieszkania wykluczającego korzystanie z usług dostawcy itp. Migracja zaplanowana jest spowodowana z kolei niesatysfakcjonującym charakterem relacji pomiędzy klientem a dostawcą. Sama rezygnacja z usług może być spowodowana przez jedną przyczynę lub kombinację kilku różnych przyczyn. Do najważniejszych zaliczyć można [Parvatiyar, Sheth, 2000, Rudawska, 2005]:

- niezadowolenie klienta z oferty dostawcy,
- niesatysfakcjonujący poziom obsługi reklamacji,
- brak zaangażowania pracowników firmy,
- znudzenie klienta i wywołana nim chęć zmiany,
- interesująca oferta konkurencji,
- wygaśnięcie umowy stanowiącej barierę wyjścia.

Obok powyższych przyczyn klienci odchodzić mogą również bez konkretnego powodu [Griffin, 1995] co można interpretować jako odejście w wyniku braku relacji, braku zaangażowania w stosunku do dostawcy.

Migracja wymuszona jest z kolei inicjowana przez dostawcę. Relacja z klientem zrywana jest głównie z powodu niewywiązywania się klienta z warunków umowy, na przykład nieregulowania miesięcznych opłat.

Powyższe rozróżnienie przekłada się wprost na wybór docelowej grupy klientów mających być podstawą budowy modelu retencji klientów. Z analizowanego zbioru wyeliminowane muszą być przypadki migracji wymuszonych. Korzystne dla jakości modelu oraz płynących z niego wniosków byłoby też wyeliminowanie przypadków migracji oczekiwanych oraz przypadkowych, tak aby w docelowym zbiorze danych model miał za zadanie rozróżnienie klientów lojalnych od klientów, którzy odeszli w sposób zaplanowany. Nie zawsze jednak takie rozróżnienie jest możliwe z technicznego punktu widzenia, ponieważ przyczyny odejścia klientów nie są znane.

Modele migracji budowane na podstawie historycznych obserwacji będą miały zarówno wymiar predykcyjny, pozwalający na przewidywanie odejścia klientów, jak również wymiar objaśniający, wskazujący na najważniejsze czynniki towarzyszące zjawisku migracji klientów. Złożenie obu wymiarów może być podstawą do konstruowania skutecznej strategii marketingowej, odpowiedniego formułowania kampanii retencyjnych, ale przede wszystkim do podejmowania zindywidualizowanych działań w stosunku do konkretnych klientów narażonych na ryzyko odejścia zwieńczonych poprawą relacji pomiędzy dostawcą a klientem. Analiza retencji klientów może być uznana za kluczowy aspekt w procesie mającym na celu wzrost wartości klientów dla firmy [Christopher i inni, 2008].

Modele analizujące zjawisko migracji klientów mogą być różnie określane w zależności od branży. Najczęściej mianem używa się terminu *churn analysis* wywodzącego się z branży telekomunikacyjnej, a obecnie używanego powszechnie, niezależnie od branży. Analiza migracji może być również określana jako *attrition analysis*, bądź *retention/defection analysis*. Określeń tych można używać zamiennie. Taką konwencję przyjęto również w niniejszym opracowaniu. Modele migracji klientów budowane są najczęściej zgodnie z eksploracyjnym podejściem do analizy danych. Ogólne założenia tego podejścia oraz szczegóły związane ze specyfiką tych modeli zostały przedstawione w kolejnym podrozdziale.

1.6. Metodyki budowy modeli *data mining* na potrzeby retencji klientów

Obecnie w terminologii zarówno naukowej, jak również w praktycznym użyciu stosuje się wiele bliskoznacznych terminów związanych z analizą danych. W obiegu naukowym oraz biznesowym pojawiają się terminy takie jak sztuczna inteligencja (*artificial intelligence*), uczenie maszynowe (*machine learning*), eksploracja danych (*data mining*), eksploracyjna analiza danych (*exploratory data analysis*), *big data*, *data science, database marketing,* głębokie uczenie *(deep leraning)*. Wszystkie te terminy bądź wprost wywodzą się ze wspólnego pnia jakim jest statystyczna analiza danych, bądź statystyczna inspiracja jest jednym z ważniejszych aspektów składowych danego terminu.

Punktem centralnym pomocnym w uchwyceniu różnic i zależności pomiędzy tymi terminami jest pojęcie statystycznej analizy danych. Mówiąc najogólniej, celem statystyki jest zwykle weryfikacja modeli opartych na wiedzy teoretycznej (znany jest ogólny charakter zależności, a badacz stara się oszacować parametry) lub weryfikacja pewnych hipotez badawczych [Migut, 2019]. Innymi słowy, modele statystyczne mają najczęściej za zadanie potwierdzić istniejąca teorię (mają charakter konfirmacyjny). Kolejnym wymiarem znamionującym statystyczne podejście do analizy jest sposób gromadzenia danych. W przypadku statystyki zazwyczaj najpierw planuje się badanie lub eksperyment, a następnie gromadzi dane. Głównym celem gromadzenia danych jest ich późniejsza analiza statystyczna. Również wolumen analizowanych danych liczony jest co najwyżej w setkach (rzadziej kilku tysiącach [8]) przypadków opisanych za pomocą kilku, bądź kilkunastu cech. Podczas doboru próby do analizy kładziony jest nacisk na oddanie struktury populacji bazowej. Statystyczne narzędzia analityczne są często wrażliwe na obserwacje nietypowe, braki danych, niespełnione założenia dotyczące rozkładów zmiennych czy współliniowości zmiennych [Łapczyński, 2016]. Innymi słowy, podejście statystyczne cechuje się eleganckim podejściem do procesu wnioskowania, skutecznym pod warunkiem spełnienia formalnych założeń, które w praktyce trudno jest badaczowi spełnić. Klasyczne statystyczne podejście do procesu wnioskowania przedstawiono na Rysunek 8. Należy zauważyć, że model oparty jest na wiedzy badacza, a wyniki analizy danych mają go jedynie potwierdzić.

[8] Na przykład badania *exit pools* w dniu wyborów.

Rysunek 8 Proces wnioskowania zgodny ze statystyczną analizą danych
Źródło: opracowanie własne na podstawie [Ratner, 2017].

Klasyczne podejście zostało zakwestionowane przez J. Tukeya, który zaproponował gruntowną zmianę strategii wnioskowania. Nowe podejście nazwane eksploracyjną analizą danych lub częściej EDA[9] zostało opisane w przełomowym dziele z 1977 roku [Tukey]. Tukey zaproponował nowe, pozbawione założeń, nieparametryczne podejście do rozwiązywania problemów, w którym analiza opiera się na samych danych a w celu uzyskania wiarygodnych wyników wykorzystuje techniki polegające na iteracyjnym testowaniu i modyfikowaniu rezultatów analizy na podstawie informacji zwrotnych [Ratner, 2017]. Największym przełomem była zmiana schematu wnioskowania, w którym model oparty był na danych a badacz w wyniku analizy uzyskiwał wiedzę na temat jego natury. Podejście do wnioskowania zgodne z EDA przedstawiono na Rysunek 9. Strzałki skierowane w obie strony podkreślają iteracyjny charakter procesu wnioskowania.

Rysunek 9 Proces wnioskowania zgodny z EDA
Źródło: opracowanie własne na podstawie [Ratner, 2017].

W przypadku EDA nie przyjmuje się żadnych założeń dotyczących charakteru modelowanego zjawiska. Model powstaje w oparciu o struktury i prawidłowości zawarte w danych. Dopuszczalny jest element subiektywizmu w postaci osądu analityka, ponieważ celem analizy jest wyciągnięcie praktycznych a nie statystycznie istotnych wniosków. Rodzi to z kolei ryzyko identyfikacji pozornych zależności. W celu ograniczenia ryzyka ich wystąpienia konieczna jest zewnętrzna walidacja (ocena) uzyskanych wyników[10]. Jeśli model przejdzie kontrolę poprawności, zostanie uznany za ostateczny i gotowy do stosowania w praktyce. Jeśli nie, dokonuje się ponownych analiz, bądź wraca do etapu

[9] Skrót od *Exploratory Data Analysis.*
[10] Strategie walidacji modeli zostały zaprezentowane dalszej części pracy.

gromadzenia danych, dopóki nowe struktury nie pozwolą na utworzenie wiarygodnego, (pozytywnie ocenionego) modelu.

Koncepcja EDA była przełomem w podejściu do analizy danych oraz do sposobu wyciągania wniosków na ich podstawie. Niezależnie od koncepcji EDA, w 1959 roku [Samuel] pojawiła się i rozwinęła w kolejnych latach koncepcja uczenia maszynowego (*machine learning*). Pojęcie to oraz metody analityczne z nim związane było rozwijane przez środowisko informatyków (*computer science*). Celem uczenia maszynowego jest odtworzenie procesu generowania danych pozwalając analitykom na uogólnienie z obserwowanych danych na nowe nieobserwowane przypadki [Guidici, 2003]. W szczególności uczenie maszynowe może być definiowane jako zestaw metod pozwalających na automatyczne wykrywanie wzorców zawartych w danych, a następnie na wykorzystaniu odkrytych wzorców, aby przewidzieć przyszłe zdarzenia, bądź realizować inne działania związane z podejmowaniem decyzji w warunkach niepewności [Murphy, 2012]. W początkowej fazie rozwój metod uczenia maszynowego można powiązać z badaniami nad dwoma rodzinami metod: sieciami neuronowymi oraz drzewami decyzyjnymi (nazywanymi przez statystyków – drzewami klasyfikacyjnymi i regresyjnymi). Pierwszym algorytmem zaliczanym do narzędzi uczenia maszynowego był perceptron przedstawiony przez F. Rosenblatta [1958]. Na jego podstawie w połowie lat 80. opracowano sieci neuronowe [Guidici, 2003]. W podobnym czasie [Morgan, Sonquist, 1963] opracowali pierwszy algorytm oparty na drzewach decyzyjnych (AID). Wielu statystyków uważa, że AID oznaczało początek stosowania technik uczenia maszynowego do rozwiązywania problemów statystycznych [Ratner, 2017]. W kolejnych latach opracowano szereg odmian oraz ulepszeń obu ogólnych koncepcji uczenia. Mimo iż obecnie dostępnych jest szereg innych algorytmów opartych na odmiennej filozofii, algorytmy te stanowią do dzisiaj klasyczny rdzeń metod uczenia maszynowego. Powodzenie algorytmu uczącego się zależy w dużej mierze od użytych danych. Uczenie maszynowe jest zatem nieodłącznie związane z analizą danych i statystyką. W koncepcji *machine learning* dochodzi do syntezy podstawowych pojęć informatycznych z pomysłami z obszarów statystyki, prawdopodobieństwa i optymalizacji [Mohri i inni, 2018].

Rozwój technik uczenia maszynowego oraz ugruntowanie się wśród badaczy koncepcji EDA zbiegł się w czasie z kolejną rewolucją związaną z rozwojem technik bazodanowych oraz możliwością przechowywania oraz dostępu do coraz znaczniejszych ilości danych. W kolejnych latach dane te zaczęto wykorzystywać do budowy modeli za pomocą technik uczenia maszynowego posługując się procesem wnioskowania zgodnym z EDA. Głównym

celem budowy modeli było wsparcie marketingu (*database marketing*). Wraz z tym zastosowaniem pojawił się zapomniany już nieco termin KDD (*Knowledge Discovery in Databases*). Obejmował on cały proces wydobywania wiedzy z danych począwszy od identyfikacji problemu aż po sformułowanie reguł biznesowych [Guidici, 2003]. Na tym gruncie pojawił się termin *data mining*[11], który początkowo odnosił się do etapu KDD związanego z analizą danych, później rozszerzył swoje znaczenie do tożsamego z KDD, pozbawiając je tym samym popularności. Termin *data mining* powstał zatem jako synteza kilku dziedzin, który można opisać za pomocą wzoru [Ratner, 2017]:

$$data\ mining = statystyka + duże\ zbiory\ danych + uczenie\ maszynowe$$

Popularna definicja określa *data mining* jako proces wyboru, eksploracji oraz modelowania dużych ilości danych w celu wykrycia nieznanych wcześniej wzorców, bądź reguł, które mają dostarczyć przejrzyste użyteczne wyniki posiadaczowi bazy danych [Guidici, 2003]. Definicja ta podkreśla eksploracyjny charakter analizy zgodny z EDA oraz praktyczny wymiar analiz tego typu.

Największa popularność terminu *data mining* przypada na pierwszą dekadę bieżącego stulecia. W tym czasie termin „duże dane" był kojarzony raczej z gigabajtami, a co najwyżej z dziesiątkami, rzadko z setkami gigabajtów danych. Ilości te były możliwe do przeanalizowania w scentralizowanym środowisku informatycznym. Obserwowany obecnie rozwój wolumenów gromadzonych danych postawił przed badaczami zadanie wykorzystania technik uczenia maszynowego w sytuacji, gdy dane są rozproszone na wielu fizycznych maszynach (dyskach) i ich objętość jest większa o rząd bądź nawet kilka rzędów wielkości. W kontekście obserwowanego przyrostu skali gromadzonych danych pojawił się termin *big data*, który jest definiowany jako zbiory danych których ilość (*volume*), prędkość generowania (*velocity*) bądź różnorodność (*variety*) jest tak duża, że jest trudna do składowania, zarządzania, przetwarzania oraz analizy za pomocą tradycyjnych baz danych oraz narzędzi do przetwarzania danych [Bahga, Madisetti, 2016].

Podstawowe cechy *big data* obejmują [Bahga, Madisetti, 2016]:

– Ilość (*Volume*) - ilość gromadzonych danych jest tak duża, że nie mieści się na jednej maszynie, stąd do ich przechowywania i przetwarzania konieczne jest użycie specjalizowanych narzędzi oraz klastrów maszyn.

[11] Termin ten nie posiada ugruntowanego tłumaczenia na język polski, częściej spotykanym jest termin eksploracja danych.

- Prędkość (*Velocity*) – prędkość generowania danych jest jednym z głównych powodów wykładniczego wzrostu ich ilości. Niektóre wymagania biznesowe [12] implikują konieczność analizy danych w czasie rzeczywistym. Aby sprostać wyzwaniu przetworzenia danych pojawiających się z dużą częstotliwością wymagana jest specjalistyczna infrastruktura oraz narzędzia informatyczne.
- Różnorodność (*Variety*) odnosi się do form danych. Duże dane występują w formie danych uporządkowanych, pozornie uporządkowanych, nieuporządkowanych lub częściowo uporządkowanych [Łapczyński, 2020]. Zalicza się do nich dane tekstowe, obrazy, audio, wideo i dane z czujników. Systemy dużych zbiorów danych muszą być wystarczająco elastyczne, aby obsługiwać tak różnorodne dane.
- Prawdziwość (*Varacity*) odnosi się do jakości danych. Aby możliwe było uzyskanie wartości na podstawie danych, dane muszą zostać wyczyszczone w celu usunięcia szumu. Aplikacje oparte na danych mogą czerpać korzyści z dużych zbiorów danych tylko wtedy, gdy dane są dobrej jakości.
- Wartość (*Value*) odnosi się do przydatności danych w realizacji zamierzonego celu (rozwiązania problemu badawczego). Celem każdego systemu analizy dużych zbiorów danych jest wydobycie wartości z danych. Wartość danych jest również związana z ich prawdziwością. W przypadku niektórych aplikacji wartość zależy również od tego, jak szybko jesteśmy w stanie przetwarzać dane.

Big data jest zatem adaptacją i rozwinięciem pojęcia *data mining* w sytuacji, gdy wolumen danych oraz wymagania biznesowe wymagają przygotowania specjalnej infrastruktury oraz zaadaptowanych do nowych wymagań algorytmów uczenia maszynowego. Można zauważyć, że termin ten bardzo często jest odnoszony do danych rzędu nawet megabajtów oraz gigabajtów tym samym zastępując pojęcie *data mining*.

Pojawienie się olbrzymich wolumenów danych otworzyło możliwości rozwoju nowych technik uczenia maszynowego. Swój kolejny renesans przeżywają sieci neuronowe, które dzięki rozwojowi procesorów graficznych są coraz bardziej powszechnie wykorzystywane do analizy ogromnych zbiorów danych. Popularny obecnie termin głębokiego uczenia (*deep learning*) odnosi się do stosowanych obecnie rozbudowanych architektur sieci neuronowych składających się z większej liczby warstw ukrytych[13].

[12] Na przykład wykrywanie nadużyć czy automatyzacja operacji giełdowych
[13] Więcej informacji na temat sieci neuronowych przedstawiono w rozdziale 3.

Ogólna koncepcja uczenia maszynowego może być realizowana za pomocą kilku strategii budowy modelu na podstawie danych. Pierwszą, najbardziej popularną strategią uczenia jest tak zwane uczenie z nauczycielem bądź uczenie nadzorowane (*supervised learning*). Podejście to stosowane jest w sytuacji, gdy w zgromadzonym zbiorze danych oprócz predyktorów zawarta jest również zmienna zależna. Zmienna ta ukierunkowuje proces uczenia modelu będąc swoistym nauczycielem procesu szacowania jego parametrów. W przypadku, gdy zmienna zależna jest mierzona na skali ilościowej (ilorazowej bądź przedziałowej) za pomocą modelu rozwiązywany jest problem regresyjny. W przypadku, gdy zmienna zależna mierzona jest na skali jakościowej rozwiązywany jest problem klasyfikacyjny. Model klasyfikacyjny budowany dla dwustanowej zmiennej zależnej nazywany bywa modelem skoringowym. (*scoring model*) lub modelem reakcji (*response model*). To tej grupy zalicza się modele migracji klientów (*churn)*. Zmienna zależna przyjmuje w nich dwa stany – „lojalny" oraz „nielojalny". Modele uczone za pomocą uczenia z nauczycielem noszą również miano modeli predykcyjnych.

Przeciwieństwem modeli predykcyjnych są modele opisowe (deskryptywne). Budowane są one w sytuacji, gdy w zbiorze danych nie występuje zmienna zależna – wszystkie zmienne pełnią w analizie rolę predyktorów. Odpowiednią strategią uczenia jest wtedy uczenie bez nauczyciela bądź uczenie nienadzorowane (*unsupervised learning)*. Popularnym zastosowaniem tej strategii uczenia jest analiza skupień polegająca na identyfikacji w zbiorze niejednorodnych obiektów (na przykład klientów) grup elementów podobnych do siebie, różniących się tym samym od pozostałych grup. Innym przykładem jest analiza koszykowa i analiza sekwencji polegające na identyfikacji i ocenie reguł (na przykład zakupowych) występujących w danych.

Sytuacją pośrednią w stosunku do przedstawionych powyżej jest zbiór danych, w którym wartość zmiennej zależnej jest znana jedynie dla ograniczonej liczby przypadków, dla większości jest ona nieznana. Na podstawie przypadków ze znaną wartością zmiennej zależnej budowany jest model, który następnie jest stosowany dla pozostałych przypadków. Wynik stosowania modelu jest podstawą do oszacowania dla nich wartości zmiennej zależnej. Po ich oszacowaniu kolejny model budowany jest na całym zbiorze. Wynik jego działania ponownie służy do oszacowania wartości zmiennej zależnej w grupie, w której była ona nieznana. Budowa i ponowna klasyfikacja są powtarzane iteracyjne aż do osiągnięcia zbieżności. Taką strategię uczenia nazywa się uczeniem częściowo nadzorowanym (*semi-supervised learning*). Ogólnie przedstawiony schemat

uczenia ma szereg odmian, wśród których popularną strategią jest uczenie wspólne (*co-training*) zaproponowane przez A. Blum i T. Mitchel [1998].

Poza trzema wymienionymi powyżej strategiami uczenia należy wspomnieć o uczeniu ze wzmocnieniem *(reinforcement learning*). Podczas procesu uczenia algorytm uzyskuje odpowiedź zwrotną, czyli odpowiedź, którą wskazał jest poprawną, czy też błędną. W sytuacji błędnej odpowiedzi nie uzyskuje wskazówki, jak ją poprawić. Musi zbadać i wypróbować różne możliwości, dopóki nie uda się znaleźć właściwej odpowiedzi. Uczenie się przez wzmocnienie jest czasami nazywane uczeniem się z krytykiem [Marsland, 2014].

Proces wnioskowania zgodny z EDA przedstawiony na Rysunek 9 jest ogólnym zarysem koncepcji budowy modeli za pomocą metod analitycznych. W literaturze przedmiotu można znaleźć szereg propozycji uszczegóławiających i rozwijających tę koncepcję. Wyczerpujące ich zestawienie można znaleźć na przykład w pracy M. Łapczyńskiego [2016]. Zgodnie z konkluzją autora można stwierdzić, że wszystkie przedstawione podejścia są w dużej mierze zgodne. Opisują one realizację procesu wydobywania wiedzy z danych jako zbiór wykonywanych w sposób iteracyjny etapów. Na podstawie syntezy proponowanych strategii można określić następujące etapy analizy:

- *Analiza biznesowa* –jest to najtrudniejszy i najważniejszy etap procesu analizy danych. Błędy popełnione na tym etapie mają najpoważniejsze konsekwencje w kontekście całego realizowanego projektu. Etap wymaga od badacza zrozumienia problemu biznesowego stawianego przez menedżerów. Na tym etapie określa się cel biznesowy modelowania oraz kryteria sukcesu planowanego projektu. W przypadku modeli *churn*, na tym etapie należy określić wstępną definicję rezygnacji klienta. W tej fazie określane są cele, jakie powinny być spełnione dzięki analizie danych w języku szeroko pojętego biznesu, niejako w oderwaniu od zagadnień *data mining*. Na tym etapie oceniana jest sytuacja, możliwe działania oraz dostępność zasobów dla proponowanych działań [Migut, 2019]. Po ustaleniu celów biznesowych przekłada się je na cele *data mining*. Po zakończeniu tego etapu powinno się otrzymać plan projektu ze ściśle zdefiniowanymi celami, kryteriami sukcesu oraz określonym zakresem odpowiedzialności.
- *Zrozumienie i przygotowanie danych* – jest to najbardziej pracochłonnych etap analizy. W zależności od projektu może zajmować od 50% do nawet 80% jego całkowitego czasu. Tak duży procent czasu wynika z faktu, że dane, które są

przedmiotem analizy są wykorzystywane do niej w sposób wtórny. Dbałość o format oraz jakość danych wymaga znacznego nakładu pracy, szeregu przekształceń, agregacji, czy łączenia zbiorów z wielu źródeł. Produktem końcowym tego etapu jest zbiór danych, w którym wiersze reprezentują analizowane obiekty (w zależności od przyjętego podejścia na przykład klientów, rachunki lub konta), natomiast kolumny zawierają opisujące je zmienne. Na tym etapie w zbiorze danych uzupełnia się braki danych, przeprowadza wstępną eliminację zmiennych ze względu na ich jakość oraz stopień powiązania ze zmienną zależną. W tym miejscu powstaje także zestaw zmiennych pochodnych (na podstawie wiedzy biznesowej klientów). Zbiór danych przygotowany na tym etapie powinien umożliwić wykonanie iteracji budowy i oceny modelu. W trakcie projektu zazwyczaj dochodzi do wielu iteracji czyszczenia danych oraz modelowania i oceny[14]. Na tym etapie przeprowadza się również jedną z najważniejszych, w przekroju całego projektu, analiz, jaką jest segmentacja ze względu na ryzyko odejścia[15].

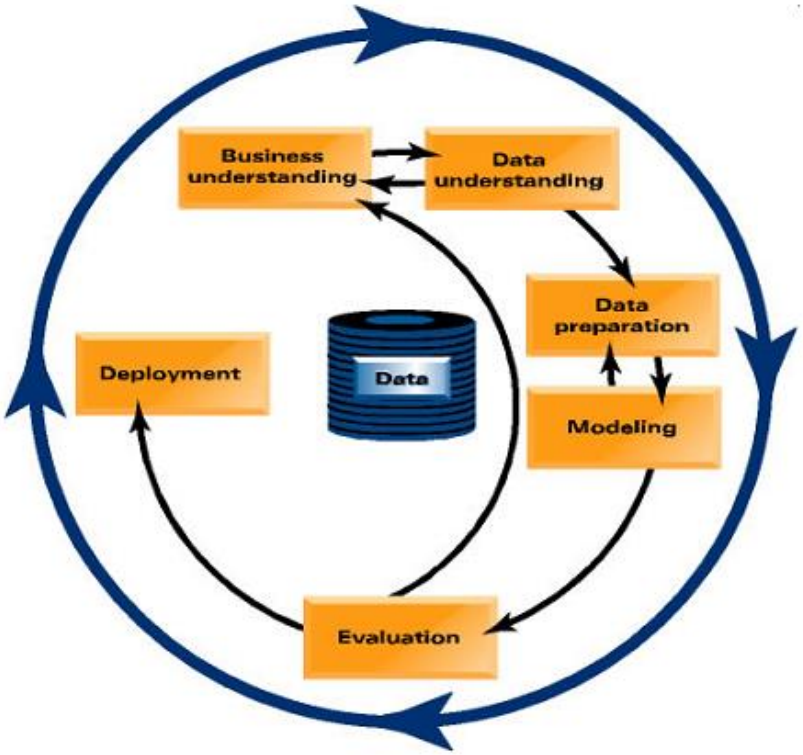
- *Budowa modelu* – w tej fazie wybiera się odpowiednią technikę modelowania, ustala najbardziej odpowiednie wielkości parametrów początkowych analiz, buduje model lub modele oraz wybiera te najlepiej rokujące [Migut, 2015]. Podczas tego etapu dokonuje się ostatecznej selekcji zmiennych, określa optymalny układ hiperparametrów modelu. Prace na tym etapie mają największy potencjał, jeśli chodzi o możliwość ich automatyzacji.
- *Walidacja modelu* – na tym etapie ocenia się model oraz dokonuje przeglądu kroków wykonanych w celu jego stworzenia, aby mieć pewność, że model służy prawidłowo założonym celom biznesowym [CRISP-DM, 2000]. Kolejnym kryterium oceny jest poprawność działania modelu na zewnętrznych zbiorach walidacyjnych. Na tym etapie należy podjąć decyzję, czy można wdrażać uzyskanie modele, czy też należy wprowadzić poprawki do modeli. Oceniając uzyskane wyniki możliwa jest również decyzja by rozpocząć cały proces od nowa [Migut, 2019]. Jeżeli model spełnia kryteria odnośnie jakości, przed jego wdrożeniem określa się punkt, bądź punkty odcięcia (*cut-off point*).

[14] Podczas relatywnie dużego projektu kierowanego przez autora konieczne było wykonanie sześciu iteracji czyszczenia danych oraz modelowania.

[15] Więcej na temat segmentacji przedstawiono rozdziale 2.

- *Wdrożenie modelu* – wiedza odkryta przez model jest bezużyteczna bez należytej popularyzacji. Wynikiem projektu może być zestaw reguł, które po asymilacji przez organizację może wpłynąć na jej procesy biznesowe. Model przewidujący skłonność klientów do odejścia musi zostać wdrożony i zintegrowany z istniejącym otoczeniem informatycznym, aby umożliwić bieżącą ocenę ryzyka odejścia klienta. Na tym etapie przygotowuje się także końcowy raport lub prezentację dla osób zainteresowanych wynikami (np. osób finansujących, decydentów, końcowych użytkowników).
- *Monitorowanie modelu* – jednym z założeń budowy modeli predykcyjnych (w tym modeli retencji klienta) jest przyjęcie, że wzorce obserwowane w analizowanym zbiorze danych będą pojawiać się również w przyszłości. Zmiany zachodzące w otoczeniu biznesowym sprawiają, że założenie to jest niespełnione w zasadzie już w momencie wdrożenia modelu. Kolejne okresy działania modelu zwiększają ryzyko jego błędnego działania. Po określeniu sposobu wdrożenia modelu konieczne jest zatem ustalenie mechanizmów dotyczących monitorowania poprawności jego działania i wszystkich działań związanych z jego aktualizacją i modyfikacją. Prawidłowe zdefiniowanie strategii związanych z monitorowaniem może pozwolić na uniknięcie okresów, w których model *data mining* jest wykorzystywany nieprawidłowo, np. w związku ze zmianą zależności, które ma odtwarzać.

Rysunek 10 przedstawia przebieg projektu analitycznego zgodnie z metodyką CRISP-DM.



Rysunek 10 Kolejne etapy projektu analitycznego wg metodyki CRISP-DM
Źródło: [CRISP-DM, 2000].

Modele migracji klientów budowane zgodnie z powyższą strategią dosyć często są omawiane w literaturze przedmiotu. Opracowania, których dotyczą wiążą się z rynkiem telefonii komórkowej [Phadke i inni, 2013, Idris i inni, 2013, Lalwani i inni, 2022, Ahmad i inni, 2019], usługami finansowymi [Naveen i inni, 2010], handlem detalicznym [Migueis, 2012], rynkiem ubezpieczeniowym czy różnego rodzaju usługami. M. Łapczyński [2016] wymienia ponad 20 takich opracowań z kolei D. García i inni [2017] w zbiorczym artykule na temat tego typu modeli wymieniają ich około 60. W zasobach czasopism naukowych dostępna jest bardzo duża liczba kolejnych opracowań tego typu. Do budowy modeli wykorzystywanych jest szereg metod począwszy od modeli ekonometrycznych po zaawansowane metody uczenia maszynowego. Przegląd wybranych metod modelowania przedstawiony zostanie w rozdziale trzecim.

Rozdział 2

Przygotowanie danych podczas budowy modeli retencji klientów

2.1. Określanie kluczowych parametrów projektu analitycznego

Przygotowanie danych do analizy jest najdłuższym i najbardziej pracochłonnym etapem procesu budowy modelu retencji klientów. W wyniku pracy na tym etapie powstaje zbiór danych, w którym każdy wiersz odpowiada modelowanemu obiektowi (najczęściej klientowi), a kolumny reprezentują cechy analizowanych obiektów. Prace nad zbiorem danych przebiegają w sprzężeniu zwrotnym z określeniem kluczowych parametrów projektu analitycznego. Tej tematyce poświęcony zostanie pierwszy podrozdział. Kolejne podrozdziały są poświęcone technikom mającym na celu poprawę jakości danych. W szczególności, aby zbiór danych był podstawą potencjalnie użytecznego modelu powinien:

- posiadać wiarygodne wartości,
- być kompletny (nie posiadać braków danych),
- być jednorodny, czyli zawierać informacje wyłącznie o badanej zbiorowości.

Na etapie przygotowania danych należy także rozpatrzyć kwestie związane z przygotowaniem zmiennych pochodnych i standaryzacją analizowanych zmiennych. Powyższe zagadnienia zostały omówione w kolejnych częściach niniejszego rozdziału.

Analiza biznesowa jest etapem, na którym określane są kluczowe parametry projektu – definiujące jego zakres, kryteria sukcesu, strategię doboru metod itp. Wyznaczenie ich ma fundamentalny wpływ na kolejne etapy projektu analitycznego.

Pierwszym parametrem, krytycznym z punktu widzenia przyszłej skuteczności modelu oceniającego lojalność klientów, jest definicja zbiorowości, dla której model ma być w przyszłości stosowany. Podczas przygotowywania danych mających być podstawą do budowy modelu, należy pamiętać, że dane te powinny być możliwie najbardziej zbliżone do zbiorowości, którą określono na tym etapie. Brak zgodności definicji zbiorowości docelowej i tej, na podstawie której budowany będzie model w konsekwencji prowadził będzie do niestabilnego działania modelu oraz jego mniejszej skuteczności, zwłaszcza w odniesieniu do grupy klientów, która nie miała swojej dostatecznej reprezentacji w zbiorze uczącym.

Zgodności struktury zbioru uczącego oraz grupy docelowej, dla której stosowany będzie model nie można zrealizować w sposób idealny. W wyniku zmian zachowań klientów oraz naturalnych procesów demograficznych grupy te w miarę upływu czasu będą się od siebie coraz bardziej różnić. W gestii badacza leży minimalizacja ryzyka wystąpienia znaczących różnic poprzez uwzględnianie w zbiorze uczącym relatywnie aktualnych danych[16]. Po wdrożeniu modelu, na etapie jego monitorowania konieczna jest cykliczna ocena zgodności struktury zbioru uczącego oraz ocenianych przez model klientów. Istotna zmiana struktury populacji jest sygnałem do odświeżenia modelu na podstawie bardziej aktualnych danych [Migut, 2015]. Model zbudowany za pomocą metod uczenia maszynowego wymaga, aby analizowane wzorce miały dostateczną reprezentację w zbiorze danych. To wymaganie determinuje eliminację ze zbioru uczącego przypadków nietypowych. Ogólnie rzecz ujmując, w zbiorze uczącym powinny znaleźć się jedynie takie przypadki, które typowo pojawiają się podczas codziennej pracy oraz odpowiadają docelowej grupie klientów, dla których przygotowywany jest model. Ze zbioru uczącego odrzucić należy zatem wszelkie przypadki klientów VIP, kluczowych klientów[17], czy korzystających z usług na wyjątkowych warunkach.

Spośród przypadków, jakie powinny znaleźć się w próbie uczącej należy rozważyć wykluczenie osób mających utrudnioną możliwość odejścia – wieloletni kredyt mieszkaniowy, lokatę z brakiem premii w wypadku jej zerwania, długoletnią umowę, w której klient otrzymał „telefon za złotówkę” itp. Tego typu przypadki należy wykluczyć

[16] Motywacją do sięgania do starszych danych jest najczęściej chęć zgromadzenia jak największej bazy do budowy modelu.

[17] Klienci ci analizowani są za pomocą rzeczywistej inteligencji pracowników firmy a nie sztucznej inteligencji modelu.

z modelowania albo ewentualnie zbudować dla nich osobne modele odejść „związanych" klientów[18] [Migut, 2015].

Jednym z najbardziej kluczowych parametrów projektu jest formalna definicja klienta „lojalnego" oraz „nielojalnego". Definicja „nielojalnego" klienta musi być zgodna z przyjętym celem biznesowym i odnosić się do klientów, których warto zatrzymać. Istotnym czynnikiem jest łatwość jej interpretacji oraz monitorowania klientów zgodnie z przyjętą definicją. W przypadku organizacji, w których klienci nie są obowiązani do cyklicznych opłat definicja odejścia klienta może różnić się od segmentu klientów. Na przykład w segmencie klientów kupujących kilka razy w tygodniu odejściem może być brak aktywności przez dwa tygodnie, natomiast w segmencie mniej aktywnych klientów odejście może być równoznaczne z brakiem aktywności przez miesiąc lub dłużej. Warto również rozważyć kwestię częściowych rezygnacji. O rezygnacji nie koniecznie musi świadczyć całkowita rezygnacja z usług. Odejście może być także rozumiane jako znaczący spadek wartości zakupów czy liczby produktów kupowanych przez klienta. Definicja zmiennej zależnej każdorazowo wymaga uchwycenia szerszego kontekstu biznesowego organizacji i musi być kompromisem pomiędzy chęcią idealnego rozróżnienia „dobrych" i „złych" klientów a jej możliwościami technicznymi i biznesowymi [Migut, 2015].

Zbiór danych, na podstawie którego budowany jest model migracji klientów zawiera zarówno cechy statyczne (np. zmienne demograficzne), jak i zmienne opisujące dynamikę korzystania przez klienta z produktów firmy. Uchwycenie dynamiki wymaga obserwacji zachowania klienta przez więcej niż jeden okres[19]. W zależności od branży może wahać się od 3 do 12 miesięcy.

Profile zachowania klienta w kolejnych miesiącach mogą różnić się w zależności od okresu, w którym rozpoczynana jest obserwacja. W zachowaniu klientów może występować sezonowość mająca wpływ na intensywność z korzystania z usług oraz skłonność do odejścia. Okres próby definiuje zatem zakres miesięcy, w których rozpoczynana jest obserwacja klienta. Przyjmuje się, że optymalny okres to 12 miesięcy ze względu na możliwość uchwycenia w takim zbiorze pełnego zestawu profili klientów, które w zależności od okresu w roku mogą się znacznie od siebie różnić.

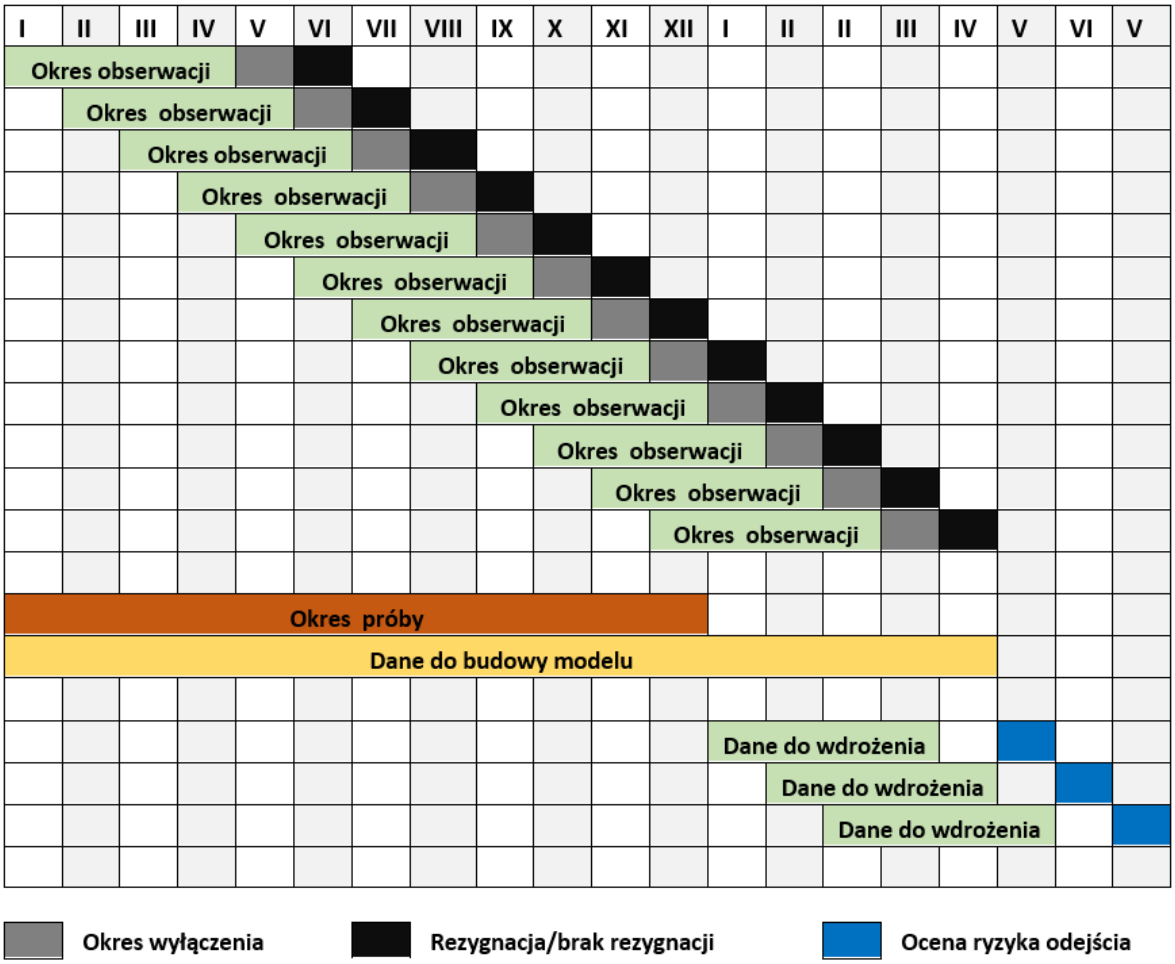
[18] Choć w takim przypadku odejście klienta może być *de facto* w krótkim terminie korzystne dla dostawy ze względu na kary lub brak uzgodnionej premii.

[19] Okresem tym jest najczęściej miesiąc.

Poza wymienionymi powyżej dwoma okresami determinującymi sposób i czas gromadzenia danych, konieczne jest wzięcie pod uwagę jeszcze jednego krytycznego parametru jakim jest okres wyłączenia. Okres ten pełni w modelowaniu rolę swoistego „horyzontu prognozy" i określa czas, po upłynięciu którego prognozowane jest odejście klienta.

Długość okresu wyłączenia jest definiowana przez ekspertów biznesowych. Ma on dawać czas na reakcję ze strony firmy i kontakt z klientem zanim podejmie on realne kroki związane z rezygnacją z usług. Okres wyłączenia wynosi zazwyczaj miesiąc. Krótszy może nie dawać odpowiedniej ilości czasu na reakcję, dłuższy powoduje że modele predykcyjne mają mniejszą skuteczność.

W wymiarze technicznym okres wyłączenia uwzględniany jest na etapie przygotowania danych. Jest to czas wykluczony z obserwacji, który bezpośrednio poprzedza moment odejścia klienta. Na Rysunek 11 przedstawiono schematycznie realizację procesu przygotowania danych, w którym okres próby wyniósł 12 miesięcy, okres obserwacji 4 miesiące, natomiast okres wyłączenia 1 miesiąc. Należy zauważyć, że powyższe parametry implikują konieczność zgromadzenia obserwacji o zachowaniu klientów z okresu 17 miesięcy. Kolejną konsekwencją takiego ustawienia tych parametrów jest niemożność dokonania oceny klienta, który korzysta z usług dostawcy krócej niż 4 miesiące.



Rysunek 11 Schemat przygotowania danych i stosowania modelu lojalności klientów
Źródło: Opracowanie własne na podstawie [Berry, Linoff, 2004, Łapczyński, 2016].

2.2. Wiarygodność danych, identyfikacja i imputacja braków danych

Wiarygodny model retencji klientów wymaga zbioru danych zawierającego prawdziwe, wiarygodne wartości. Pierwszym etapem wstępnej analizy powinno być zatem zrozumienie zgromadzonych danych. Skuteczna ocena wiarygodności analizowanego zbioru wymaga połączenia wiedzy merytorycznej wraz ze statystyczną oceną zgromadzonych danych. Na tym etapie potwierdza się skale pomiarowe, na których mierzone są analizowane predyktory[20]. Następnie bada się ich rozkłady stosując tabele liczności, histogramy oraz dodatkowo wykresy ramkowe w przypadku zmiennych ilościowych. Po zbadaniu rozkładów zmiennych oblicza się odpowiednie dla danego typu i

[20] W analizie przyjętą praktyką jest korzystanie z uproszczonego podziału na zmiennie ilościowe i jakościowe z pominięciem zmiennych porządkowych, które są traktowane jak jakościowe oraz podziału na skalę ilorazową i przedziałową.

charakteru rozkładu statystyki opisowe. Charakterystyki te pozwalają uzyskać syntetyczny opis wartości przeciętnego poziomu i zmienności analizowanych cech.

Po zbadaniu każdej zmiennej oddzielnie należy również zbadać współzależności pomiędzy zmiennymi za pomocą tabel krzyżowych oraz macierzy korelacji. Te stosunkowo proste analizy umożliwiają ogólny wgląd w dane, pozwalają ocenić wiarygodność zebranych danych, mogą przyczyniać się do poprawy ich opisu oraz być punktem wyjścia dla operacji czyszczenia danych, transformacji danych i innych czynności związanych z ich przygotowaniem wymaganym dla dalszych analiz [Migut, 2019]. Wartości, które na tym etapie zostały zidentyfikowanie jako nieprawdziwe zostają zamienione na braki danych, o ile nie ma możliwości uzyskania ich rzeczywistych wartości.

Braki danych są powszechnym zjawiskiem występującym w niemal każdym analizowanym zbiorze danych. Wprowadzenie do modelowania niekompletnego zbioru danych spowoduje, że algorytmy uczenia maszynowego w sposób automatyczny pominą przypadki, w których występuje przynajmniej jeden brak[21]. Okoliczność ta wymusza na badaczu przeprowadzenie wstępnej eksploracji danych oraz imputacji brakujących wartości. Zaniechanie tej czynności może w konsekwencji prowadzić do redukcji zbioru danych, a tym samym do utraty części zawartych w nim wzorców. W przypadku, gdy braki danych są brakami nielosowymi utracone przypadki wpłyną na zmianę obserwowanych zależności skutkując tym samym budową obciążonego, nieskutecznego modelu.

Proces imputacji braków danych powinien być poprzedzony analizą przyczyn ich występowania. Eksploracyjna analiza danych może być tutaj bardzo użytecznym narzędziem pozwalającym ocenić wzorce zawarte w brakach danych. W pierwszej kolejności należy ocenić odsetek brakujących wartości w każdej ze zmiennych. Następnie należy zbadać czy występowanie braków danych jest losowe, czy też można wyróżnić zależność pomiędzy wystąpieniem braku danych a [Migut, 2019]:

- rozkładem zmiennej zależnej,
- rozkładami pozostałych predyktorów,
- brakami danych innych predyktorów.

[21] Wyjątkiem są wybrane algorytmy drzew klasyfikacyjnych i regresyjnych, które obsługują braki danych poprzez wyszukiwanie zmiennych zastępczych, na podstawie których model generuje odpowiedź w przypadku wystąpienia baku danych.

Przydatną techniką ułatwiającą analizę braków danych jest dyskretyzacja wszystkich zmiennych, tak aby nowe dwuwariantowe zmienne informowały o ich występowaniu (kategorie: brakujące dane / kompletne dane). Dla tak przygotowanego zbioru obserwacji bada się wpływ predyktorów na zmienną zależną. Technika ta pozwala określić, które z zaobserwowanych braków mogą wnieść dodatkową informację do budowanego modelu.

Kolejną metodą pozwalającą lepiej zrozumieć strukturę braków danych oraz ich ewentualne współwystępowanie jest analiza skupień (zmiennych) dla danych poddanych dyskretyzacji [Harrell, 2015]. Powala ona zidentyfikować grupy zmiennych, w których braki danych występują dla tych samych przypadków. Eksploracyjna analiza danych oraz wiedza biznesowa badacza pozwala przypisać zaobserwowane braki danych do jednej z trzech grup[22]:

- Braki czysto losowe (MCAR - *Missing Completely At Random*). Ich pojawienie się nie jest związane z żadnym zdarzeniem czy wartością innej zmiennej zrejestrowanej w analizowanym zbiorze. Mogą one wynikać z np. chwilowej awarii systemu czy też innych zdarzeń losowych.
- Braki losowe (*MAR – Missing At Random*). W przypadku tego rodzaju braków przyjmuje się założenie, że nie istnieje zależność pomiędzy rzeczywistą wartością danej cechy a prawdopodobieństwem, że będzie ona brakiem. Dodatkowo zakłada się występowanie powiązania jej wartości z wartościami innych zmiennych w zbiorze danych.
- Braki informacyjne (*Informational Missing*). Występują one w sytuacji, gdy istnieje większe prawdopodobieństwo, że wartości są brakującymi danymi, jeżeli ich prawdziwe wartości są w sposób systematyczny niższe, bądź wyższe od wartości niebrakujących.

Po wstępnej eksploracji należy podjąć decyzję o sposobie ich zastępowania (imputacji). Większość technik uzupełniania braków danych została zaproponowana przez autorów reprezentujących nurt statystyczny, głownie w kontekście przeprowadzania poprawnej procedury testowania hipotez statystycznych [Khun, Johnson, 2013]. Do najczęściej proponowanych metod tego nurtu można zaliczyć:

[22] Szerszy opis rodzajów braków danych oraz ich charakterystyki można znaleźć między innymi w [Allison, 2001, Vittinghoff i inni, 2011, Harrell, 2015].

- Usuwanie przypadków z brakami danych. Jest to dopuszczalne podejście w sytuacji braków czysto losowych, które stanowią niewielki odsetek zbioru danych. Jeżeli braki są czysto losowe, podzbiór kompletnych przypadków będzie losową podpróbą oryginalnego zbioru [Allison, 2001] Jest to przyjęta technika w przypadku, gdy brak danych został zaobserwowany dla zmiennej zależnej [Harrell, 2015].
- Zamiana braków za pomocą średniej lub mediany w przypadku predyktorów ilościowych[23]. Ta popularna metoda imputacji braków danych jest krytykowana ze względu na fakt zaniżania oszacowań wariancji oraz kowariancji i raczej powinno się jej unikać [Allison, 2001].

Osobną strategią uzupełniania braków danych jest wykorzystanie informacji zawartych w pozostałych zmiennych niezależnych do oszacowania ich najbardziej prawdopodobnych wartości. Strategia ta zakłada, że braki danych są typu MAR. W oparciu o nią można rozpatrywać szereg metod, spośród których należy wymienić [Allison, 2001, Vittinghoff i inni, 2011, Harrell, 2015]:

- Metoda „*Hot deck*", polegająca na uzupełnianiu wartości zmiennych (ilościowych i jakościowych) na podstawie zestawu zmiennych jakościowych powiązanych z uzupełnianą zmienną. Brakująca wartość jest uzupełniana losowo spośród przypadków posiadających takie same kategorie powiązanych zmiennych jakościowych.
- Metoda pojedynczej imputacji (*single imputation*) polega na budowie modelu regresji[24], w którym zmienna z brakami danych pełni rolę zmiennej zależnej, natomiast wybrane zmienne pełnią rolę predyktorów. Oszacowaniami braków danych są przewidywania dopasowanych w ten sposób modeli. Oszacowania są następnie korygowane o niewielką losową wartość w celu zachowania zmienności uzupełnianej cechy.
- Powyższa procedura jest wykorzystywana w wielokrotnej imputacji (*multiple imputation*). Powtarza się ją zwykle 5 lub 10 razy [Vittinghoff i inni, 2011] uzyskując adekwatną liczbę zbiorów danych. Dla każdego z nich buduje się

[23] Analogicznie braki danych dla zmiennych jakościowych mogą być zastępowane za pomocą dominanty.

[24] W zależności od skali pomiarowej uzupełnianej zmiennej może to być regresja wieloraka, logistyczna lub wielomianowa logistyczna.

docelowy model predykcyjny po czym uśrednia się jego parametry oraz błędy standardowe. Błędy koryguje się dodatkowo o zmienność uzyskanych parametrów.

– Rozszerzeniem metody *Hot deck* umożliwiającą wykorzystanie zmiennych ilościowych jest metoda predykcyjnego dopasowania średniej (*Predictive Mean Matching*). Podobnie jak w pojedynczej imputacji budowany jest model predykcyjny. Następnie na podstawie modelu obliczane są prognozy zarówno dla brakujących, jak i niebrakujących danych. Dla danego przypadku z brakującą obserwacją odszukiwana jest określona liczba przypadków kompletnych z podobną predykcją. Spośród nich losowana jest rzeczywista wartość, którą uzupełnia się brak danych [Allison, 2001].

Powyższe metody są zalecane w sytuacji, gdy głównym celem analizy jest próba wyjaśnienia zjawiska odchodzenia klientów, a głównym narzędziem modelowania jest model regresji logistycznej lub model pokrewny. Najczęściej jednak modele retencji klientów są budowane w celu skutecznej predykcji przyszłego zachowania klientów. Aspekt statystycznej poprawności oszacowań jest mniej istotny od aspektu skuteczności zbudowanego modelu.

Przy założeniu losowości braków danych (MAR), gdy celem budowy modelu jest skuteczne przewidywanie skłonności do odejścia, najbardziej popularnym podejściem jest budowa modelu predykcyjnego dla każdego predyktora zawierającego braki. Podejście to jest analogiczne do metody pojedynczej imputacji z tą różnicą, że algorytm modelowania jest dobierany z szerokiej gamy metod uczenia maszynowego. W tym celu wykorzystać można między innymi drzewa klasyfikacyjne (regresyjne), drzewa wzmacniane czy sieci neuronowe. Najpopularniejszą metodą imputacji wydaje się metoda *k*-najbliższych sąsiadów. Metoda ta zapewnia, że uzupełniana wartość będzie mieściła się w zakresie wartości obserwowanych w zbiorze treningowym. Jej wadą jest konieczność dostępności całego zbioru uczącego [Khun, Johnsonn, 2013][25]. Metody imputacji braków danych za pomocą technik uczenia maszynowego cechują się wyższą skutecznością od technik używanych w podejściu tradycyjnym, jeśli skuteczność tę mierzyć siłą predykcyjną zbudowanego modelu [Jerez i inni, 2010].

Inną strategią zarządzania brakami danych jest kategoryzacja zmiennych ilościowych, a następnie dodanie do niej nowej klasy „brak danych". W przypadku predyktorów

[25] W momencie stosowania modelu konieczne jest utrzymanie stałej dostępności zbioru uczącego w celu imputacji braków. Proces imputacji jest bardziej czasochłonny od innych metod.

jakościowych sprowadza się ona do dodania osobnej kategorii reprezentującej brak. Metoda ta bywa krytykowana [Harrell, 2015] jako niedopuszczalna z punktu widzenia interpretacji modeli statystycznych, bardzo dobrze sprawdza się podczas budowy modeli predykcyjnych. Dodatkowo w przypadku, gdy braki danych są brakami informacyjnymi podejście to umożliwia wykorzystanie tego faktu i poprawę skuteczności modelu[26].

Skrajnym podejściem do kwestii braków danych jest propozycja zawarta w pracy F. Cholleta [2019], aby braki danych zastępować zerami (lub innymi stałymi wartościami spoza zakresu zmienności). Model[27] w wyniku procesu uczenia rozpozna, że 0 oznacza brak danych i nauczy się go ignorować. Dodatkowo autor zaleca sztuczne wprowadzanie do zbioru braków danych, aby niejako „uodpornić" model, który będzie mniej wrażliwy na takie sytuacje w trakcie jego wykorzystywania w przyszłości.

2.3. Sposoby łagodzenia problemów związanych z niejednorodnym zbiorem danych

Jednorodność zbioru danych jest kolejnym krytycznym aspektem jakości danych, jaki powinien zostać poddany badaniu. Zbiór możemy traktować jako jednorodny, jeżeli w przestrzeni wielowymiarowej tworzy on zwartą figurę [Jajuga 1993]. Niespełnienie warunku jednorodności może przejawiać się przez [Lula 1999]:

- wystąpienie pojedynczych wartości nietypowych,
- wystąpienie grup obserwacji.

Obecność wartości odstających może być symptomem pojawienia się w badanym zbiorze nowej grupy klientów, która dopiero zaczyna się formować [Khun, Johnson, 2013] lub może wynikać z innych bardziej losowych przyczyn, na przykład błędów systemu. Obecność wartości odstających ma zazwyczaj negatywny wpływ na proces modelowania oraz jakość zbudowanego modelu. Przypadki nietypowe mogą wywierać nieadekwatnie duży wpływ na wartości parametrów modelu w stosunku do swojej wartości informacyjnej. Ich obecność może utrudniać modelowanie głównej struktury zależności zawartej w zbiorze danych. Konieczna jest zatem ich identyfikacja oraz odpowiednia

[26] W praktyce autor miał do czynienia ze zmiennymi w których kategoria „brak danych" była najcenniejszą informacją, jaką niosła dana zmienna.

[27] W tym przypadku sieć neuronowa.

transformacja lub eliminacja ze zbioru uczącego. Zaniechanie tych działań może mieć destabilizacyjny wpływ na model[28].

Operacyjne kryterium nietypowości może być oparte na kryterium odległości. Za wartości nietypowe można uznać obserwacje, których średnia odległość od swoich *k* najbliższych sąsiadów jest największa. Inne kryteria opierają się na analizie rozkładu. Za nietypowe uznaje się wtedy punkty leżące w obszarach o niskiej gęstości [Ting i inni, 2018].

Identyfikacja wartości nietypowych może być przeprowadzona za pomocą wielu technik graficznych oraz analitycznych. Ich obszerny przegląd można znaleźć w pracy C.C. Aggarwala [2017] oraz V. Chandola i innych [2009]. Do klasycznych metod identyfikacji wartości nietypowych można zaliczyć LOF[29] (*Local Outlier Factor*), która klasyfikuje dany punkt jako nietypowy, jeśli jego odległość od sąsiadów jest większa niż sąsiadów do samych siebie [Emmott i inni, 2015]. Metoda ABOD (*Angle-Based Outlier Detection*) ocenia wariancję kątów, których wierzchołkiem jest analizowany punkt, natomiast ramiona przechodzą przez wszystkie pary punktów w zbiorze. Za odstające uznaje się przypadki o niewielkiej wariancji tak wyznaczonych kątów. Innymi klasycznymi metodami analizy wartości odstających są [Emmot i inni, 2015]: *One-Class SVM, Support Vector Data Description* czy *LODA* (*Lightweight Online Detector Of Anomalies*).

Spośród, relatywnie nowych algorytmów należy wymienić *Isolation Forest* [Hariri i inni, 2019] działający w trybie bez nauczyciela, polegający na budowie zespołu drzew rekurencyjnie dzielących zbiór danych na dwa podzbiory na podstawie wylosowanego predyktora oraz losowo wybranej wartości. Wartości odstające są identyfikowane jako punkty, dla których izolacji wystarczająca była przeciętnie mniejsza w przekroju wielu modeli głębokość drzewa[30]. Inną metodą identyfikacji wartości odstających jest *iNNE* (*isolation using Nearest Neighbour Ensemble*) [Bandaragoda i inni, 2018]. W przypadku tej metody ze zbioru danych wielokrotnie losowany jest podzbiór przypadków[31], a następnie dla każdego z przypadków w danym podzbiorze wyznaczany jest wskaźnik izolacji w oparciu o odległość od jego najbliższego sąsiada. Po wykonaniu zadanej liczby

[28] W przypadku niektórych metod, takich jak na przykład drzewa klasyfikacyjne i regresyjne, wpływ ten może być co najwyżej neutralny.

[29] Możliwe tłumaczenie tej metody to współczynnik lokalnej nietypowości.

[30] Na tej odstawie oblicza się syntetyczny wskaźnik izolacji (*isolation score*).

[31] Najczęściej wielkości od 2 do 128 przypadków [Ting i inni, 2018].

powtórzeń dla każdego przypadku uśrednia [32] się jego wskaźnik izolacji. Wartości odstające identyfikuje się wśród przypadków o najwyższym wskaźniku izolacji.

Tak zaawansowane narzędzie analityczne nie zawsze musi być skuteczne w odniesieniu do modeli retencji klientów. Często wystarczającym minimum pozwalającym zidentyfikować wartości odstające jest analiza jednowymiarowa przeprowadzona za pomocą wykresów ramka-wąsy. Jako uzupełnienie służyć może identyfikacja wielowymiarowych wartości odstających za pomocą metody *k*-średnich, w której określa się relatywnie dużą liczbę skupień a za wartości nietypowe przyjmuje mało liczne grupy oddalone od pozostałych skupień. Przydatna może być również analiza wykresów głównych składowych całego zbioru [Lula, 1999].

Po zidentyfikowaniu wartości odstających wymagane jest określenie strategii ich neutralizacji. Jeśli zidentyfikowane przypadki odstające reprezentują klientów spoza obszaru zainteresowania modelowanego zjawiska zalecane jest ich usunięcie. Możliwa jest także transformacja danych zmniejszająca stopień ich nietypowości, przykładem takiej transformacji może być logarytmowanie. W przeciwnym przypadku najprostszą i najczęściej wystarczającą transformacją jest zastąpienie wartości odstających wartościami granicznymi (*clamp transformation*). Wartości te mogą zostać określone na podstawie rozstępu kwartylowego [Kelleher, 2015][33].

$$dolna\ granica = pierwszy\ kwartyl - 1{,}5 * rozstęp\ kwartylowy$$

$$górna\ granica = trzeci\ kwartyl + 1{,}5 * rozstęp\ kwartylowy$$

Powyższa reguła jest zgodna z regułą Tukeya stosowaną na wykresach ramka-wąsy. Inną propozycją [Ratner, 2017] jest wykonanie transformacji SRD (*Symmetrizing Ranked Data*), która ranguje, a następnie symetryzuje rozkłady. Według autora, przekształcenie to poza redukcją wartości odstających pozwala zwiększyć siłę dyskryminacyjną analizowanych zmiennych. Kolejną strategią redukcji wartości odstających, szerzej opisaną w dalszej części tego podrozdziału jest dyskretyzacja zmiennych, a następnie standaryzacja wartości należących do wyznaczonych przedziałów za pomocą transformacji WoE.

[32] Przyjmuje się, że wystarczająca liczbą powtórzeń dla każdego przypadku mieści się w przedziale od 20 do 100 [Ting i inni, 2018].

[33] Innym popularnym choć mniej zalecanym podejściem jest określanie granic w odległości dwóch odchyleń standardowych od średniej.

Innym zagadnieniem związanym z tematem jednorodności danych jest kwestia występowania w zbiorze danych różnych segmentów klientów. Jeśli w danych występują wyraźnie odrębne skupienia (grupy), to możliwe jest wydzielenie istniejących grup i niezależna analiza każdej z nich. Dla każdej z grup uzyskuje się niezależny model. Możliwe jest również zignorowanie tego faktu i budowa łącznego modelu dla całego zbioru danych. Zalety i wady tych dwóch strategii omówione zostały przez P. Lulę [1999]. Podział danych na jednorodne grupy i budowa modelu dla każdej z nich prowadzi do mniej skomplikowanych modeli, które łatwiej się uczą i generują mniejsze błędy. Wadą modeli tego typu jest natomiast niewykorzystane informacji wynikających z łącznego rozpatrywania obiektów należących do różnych grup wydzielonych ze względu na wymóg jednorodności. Wada ta nie ujawnia się w sytuacji, gdy budowany jest jeden model na podstawie zbioru danych, ale w takim przypadku może on nie opisywać pewnych zachowań o mniejszym nasileniu. Wymaga ponadto zastosowania rozbudowanych modeli, co pociąga za sobą dodatkowe problemy związane z doborem ich odpowiedniej struktury.

Trzecim sposobem łączącym dwie powyższe strategie jest budowa modeli hybrydowych, które w pierwszej swojej warstwie koncentrują się uchwyceniu głównych zależności oraz identyfikacji jednorodnych podzbiorów, w warstwie drugiej modelują zależności o mniejszym nasileniu, właściwe dla poszczególnych grup klientów. Popularnym przykładem modelu hybrydowego jest model CART-Logit, w którym za pomocą modelu CART dokonuje się wstępnego podziału zbioru danych na kilka podzbiorów reprezentowanych przez liście drzewa [34]. Podzbiory te są podstawą do zbudowania niezależnych modeli regresji logistycznej. Podejście to jest standardem w modelowaniu ryzyka kredytowego, gdzie wymogiem jest budowa modeli interpretowalnych przez człowieka [Siddiqui, 2017]. Są one również z powodzeniem stosowane w obszarze marketingu, a w szczególności utrzymaniu klienta, czego przykładem może być monografia M. Łapczyńskiego [2016]. Podejście to wydaje się być najbardziej rekomendowane.

Przesłanką świadczącą o zasadności modelu hybrydowego jest występowanie w zbiorze danych zmiennych, których siła predykcyjna mierzona za pomocą kryterium informacyjnego (*IV, Information Value*)[35] przekracza 0,5 [Siddiqui, 2017]. Zmienne takie są potencjalnie użyteczne do przeprowadzenia na ich postawie podziałów

[34] Zazwyczaj są waha się od dwóch do czterech. Większa liczba grup należy do rzadkości. Ograniczeniem jest zazwyczaj niewystarczająca liczba przypadków nielojalnych klientów.

[35] Miara ta została przybliżona w dalszej części rozdziału.

segmentacyjnych za pomocą drzewa. Ich wykonanie powinno być wypadkową wiedzy biznesowej badacza oraz przesłanek statystycznych. Najlepszy z matematycznego punktu widzenia predyktor niekoniecznie musi dawać najlepszy podział segmentacyjny z punktu widzenia biznesu. Na etapie przygotowania jednorodnych segmentów przydatne są implementacje drzew decyzyjnych umożliwiające samodzielne określenie sposobu podziału zbioru przez badacza. Przesłanką statystyczną świadczącą o dobrym podziale segmentacyjnym jest zróżnicowana siła predykcyjna zmiennych niezależnych w przekroju utworzonych segmentów.

2.4. Transformacje zmiennych oraz przygotowanie zmiennych pochodnych

Do najbardziej popularnych transformacji predyktorów zalicza się normalizację zmiennych. Przez normalizację zmiennych rozumie się rodzinę przekształceń mających na celu ujednolicenie zakresu zmienności analizowanych predyktorów. Jej celem jest ułatwienie przeprowadzenia procesu modelowania. Ma ona szczególne znaczenie w metodach opartych na analizie odległości jak na przykład metoda k najbliższego sąsiedztwa (KNN). Normalizacja jest również powszechnie stosowana w przypadku sieci neuronowych ułatwiając tam proces optymalizacji wag modelu. Najbardziej popularną metodą normalizacji jest standaryzacja klasyczna [Walesiak, 2011]:

$$z = \frac{x - \overline{x}}{s_x}$$

gdzie $\overline{x}$ oznacza średnią zmiennej podlegającej transformacji, a s_x jej odchylenie standardowe. W wyniku przekształcenia uzyskuje się zmienne, których średnia wynosi 0, a odchylenie standardowe jest równe 1. Wartości przekształconej zmiennej informują o liczbie odchyleń standardowych różniących daną wartość od wartości średniej[36].

Kolejną popularną metodą normalizacji jest unitaryzacja zerowana wyrażona wzorem [Walesiak, 2011]:

$$z = \frac{x - x_{min}}{x_{max} - x_{min}}$$

[36] Na bazie tej transformacji opracowano szereg skal takich jak skala stenowa, staninowa czy tenowa.

Gdzie x_{min} oznacza minimalną wartość zmiennej a x_{max} jej wartość maksymalną. Przekształcenie to zawęża zakres zmienności pierwotnej zmiennej do przedziału [0; 1]. Powyższe formuły należą do najczęściej stosowanych przekształceń, głównie ze względu na swoją prostotę. W literaturze można znaleźć szereg analogicznych przekształceń przedstawionych na przykład w pracach M. Walesiaka [2011] czy L. Błażejczyk-Majki [2018]. Cechują się one różnymi własnościami, między innymi odpornością na brak symetrii rozkładu przekształcanej cechy czy też na występowanie wartości odstających[37].

W pracy D.Pyle [1999] zaproponowano przekształcenie wykazujące odporność na występowanie wartości odstających a także zapewniające, że uzyskane wartości zawsze będą mieściły się w zakresie [0;1], nawet jeśli przyszłe wartości będą wykraczały poza obserwowany dotychczas zakres zmienności. Przekształcenie bazuje na funkcji logistycznej:

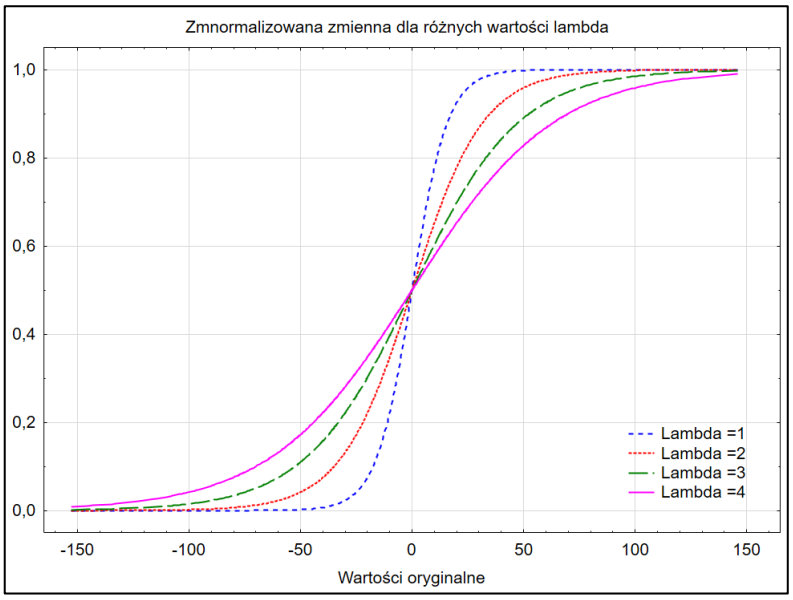
$$v = \frac{e^z}{1 + e^z}$$

gdzie e jest stałą Eulera, natomiast z jest obliczane na podstawie wzoru [Pyle, 1999]:

$$z = \frac{x - \overline{x}}{\lambda * \frac{S_x}{2\pi}}$$

Wartość λ jest hiperparametrem określanym przez użytkownika, definiującym, jaki zakres wartości w sensie odchyleń standardowych od średniej ma zostać uwzględniony w części liniowej przekształcenia. Wartości wychodzące poza ten zakres będą przyjmowały wartości odpowiednio bliskie 0 oraz 1. Rysunek 12 przedstawia wartości znormalizowane dla różnych wartości λ. Warto zauważyć, że im większa wartość λ tym większy zakres wartości danej zmiennej trafia do części liniowej przekształcenia.

[37] Przykładem takiego przekształcenia może być standaryzacja pozycyjna.



Rysunek 12 Znormalizowana zmienna dla różnych wartości lambda
Źródło: Opracowanie własne.

Pomimo pożądanych właściwości przekształcenie to nie zdobyło większej popularności, być może ze względu na swój relatywnie wysoki poziom komplikacji. W ocenie autora zasługuje na zdecydowanie większą popularyzację[38].

Kolejnym popularnym przekształceniem analizowanych predyktorów jest ich dyskretyzacja, czyli podział obserwowanego zakresu wartości na przedziały klasowe. Przedziały mogą zostać wyznaczone „ręcznie" przez badacza na podstawie jego wiedzy biznesowej, jednak częściej tworzone są w sposób automatyczny na podstawie:

- określonej liczby przedziałów o stałej liczności (podział na percentyle),
- określonej liczby przedziałów o stałej szerokości,
- algorytmu drzew decyzyjnych[39].

Warto zauważyć, że dyskretyzacja zmiennych pozwala na redukcję wartości odstających, które trafiają w naturalny sposób do skrajnych przedziałów. Dyskretyzacja upraszcza proces imputacji braków danych, dla których może zostać stworzona osobna kategoria „brak danych". Transformacja ta ułatwia także wykorzystanie w modelach linowych zmiennych o niemonotonicznym wpływie na zjawisko odejścia klientów.

[38] W dalszej części pracy będzie ono nazywane standaryzacją Pyle'a
[39] Przykładowo moduł „Dyskretyzacja zmiennych" w programie Statistica Zestaw Skoringowy proponuje dyskretyzację za pomocą drzew CART, CHAID oraz CHAID na podstawie wyników CART.

Po przeprowadzonej dyskretyzacji wyzwaniem staje się liczbowa reprezentacja uzyskanej zmiennej jakościowej. Liczbowa reprezentacja jest wymagana w wielu algorytmach uczenia maszynowego. Domyślną opcją w literaturze oraz większości implementacji jest przekodowanie zmiennej na zestaw sztucznych zmiennych zero-jedynkowych (*dummy variables*), gdzie liczba zmiennych sztucznych jest równa liczbie kategorii oryginalnej zmiennej[40] lub o jeden mniejsza od liczby tych kategorii. W tej drugiej sytuacji najczęściej stosowanymi odmianami kodowania są:

- kodowanie z sigma ograniczeniami (*sigma restricted*),
- kodowanie z poziomem odniesienia (*effect coding*)[41].

Interesującą alternatywą dla kodowania zero-jedynkowego jest przekształcenie uzyskanych kategorii za pomocą formuły *Weight of Evidence* (*WoE*) [Siddiqui, 2017]:

$$WoE = ln\left(\frac{Distr\ Good}{Distr\ Bad}\right) \times 100$$

gdzie określenia *Distr Good* oraz *Distr Bad* odnoszą się odpowiednio do odsetka osób lojalnych (w odniesieniu do wszystkich osób lojalnych) i odsetka osób nielojalnych (w odniesieniu do wszystkich osób nielojalnych) w analizowanej grupie. Warto zauważyć, że przekształcenie WoE można stosować zarówno dla zmiennych ilościowych po uprzedniej ich dyskretyzacji, jak również dla zmiennych pierwotnie mierzonych na skalach słabszych. W wyniku przekształcenia uzyskuje się wystandaryzowane zmienne przyjmujące wartość 0 dla kategorii, w której rozkład zmiennej zależnej jest zgodny z rozkładem tej zmiennej w całym zbiorze. Kategorie, w których obserwowana jest nadreprezentacja lojalnych klientów przyjmują wartości dodatnie (tym większe im ta nadreprezentacja jest większa) a ujemne w przeciwnym przypadku. Na podstawie wartości WoE możliwe jest graficzne przedstawienie profilu ryzyka co umożliwia lepsze zrozumienie charakteru wpływu analizowanego predyktora na modelowane zjawisko, a także ekspercką korektę pierwotnego podziału na kategorie.

Innym rodzajem eksperckiej ingerencji w przedziały klasowe może być zmiana ich granic. Zasadnym może okazać się określenie wartości 24 miesięcy granicą analizowanego przedziału. Po określeniu liczby oraz granic przedziałów pierwotne zmienne są

[40] Tego typu kodowanie nazywane kodowaniem przeparametryzowanym (*overparametrized*) stosowane jest domyślnie w na przykład w sieciach neuronowych i innych metodach uczenia maszynowego.

[41] To kodowanie jest najczęściej wykorzystywane w modelach regresji logistycznej i innych modelach liniowych.

standaryzowane do wartości WoE i w takim formacie są wprowadzane do algorytmu uczącego. Przekształcenie to zapewnia monotoniczność wpływu na ryzyko tak wystandaryzowanych zmiennych, pozwala także unikać przygotowania dużej liczby zmiennych sztucznych. Dla każdej pierwotnej zmiennej X tworzona jest tylko jedna zmienna pochodna.

Osobną grupą przekształceń danych jest tworzenie zmiennych pochodnych. Zmienne pochodne (*derived variables*) to nowe zmienne przygotowane najczęściej na podstawie dwóch, bądź większej liczby zmiennych pierwotnych. Ich rolą jest pomoc w odkrywaniu ważnych zależności pomiędzy zmiennymi oraz ułatwienie identyfikacji wzorców przez metody uczenia maszynowego. Zmienne pochodne często są kluczowe dla jakości końcowego modelu. W literaturze można spotkać twierdzenie, iż poprawa jakości modeli bardziej zależy od przygotowania zmiennych pochodnych niż od doboru metody modelowania [Mitrović i inni, 2017].

Najważniejszym źródłem informacji o sposobie przygotowania zmiennych pochodnych jest wiedza ekspertów biznesowych rozumiejących proces od strony praktycznej i potrafiących na podstawie swojego doświadczenia sformułować szereg reguł przekładających się na nowe predyktory. Z tego faktu wynika konieczność zapewnienia dobrej komunikacji pomiędzy osobami budującymi modele a praktykami posiadającymi specyficzną wiedzę biznesową.

Popularnymi zmiennymi pochodnymi są wyniki analizy RFM (*Recency, Frequency, Monetary*). W analizie tej bierze się pod uwagę interesujące badacza zdarzenie z uwzględnieniem kryteriów przedstawionych w rozdziale 1. Podejście RFM ma trzy ważne cechy, dzięki którym zyskało popularność w badaniach naukowych oraz biznesie [Mitrović i inni, 2017]:

- opiera się na prostej koncepcji, która pozwala łatwo obliczyć oraz zinterpretować nowe cechy pochodne,
- elastycznie podchodzi do definicji analizowanego zdarzenie dzięki czemu może być stosowana w wielu branżach,
- zmienne RFM cechują się wysoką mocą predykcyjną.

Innym przykładem zmiennych pochodnych zdefiniowanych ekspercko mogą być różnego rodzaju indeksy dynamiki informujące o zmianie intensywności korzystania przez klientów z oferowanych produktów. Zmienne pochodne mogą być również generowane w

sposób automatyczny na podstawie zadanej z góry grupy metod przekształceń. Najczęściej spotykanym przekształceniem jest obliczanie iloczynów par, bądź rzadziej większej liczby analizowanych predyktorów[42]. W wyniku automatycznych przekształceń zdecydowana większość nowych cech nie ma wartości predykcyjnej, dlatego też istotne jest zastosowanie filtrów pozwalających na eliminację z dalszej analizy nieistotnych predyktorów.

Ogólnie należy przyjąć, że im prostsza metoda uczenia maszynowego zostanie użyta podczas modelowania tym potencjalnie większe znaczenie dla końcowego wyniku może mieć etap przygotowania zmiennych pochodnych.

F. Chollet [2019] zawraca uwagę na zmianę paradygmatu odnoszącą się do przygotowania zmiennych pochodnych i związaną z pojawieniem się metod uczenia głębokiego. Według niego algorytmy uczenia głębokiego dysponują na tyle skomplikowaną strukturą (przestrzenią hipotez), aby samodzielnie uczyć się przydatnych cech. Metody te jednak lepiej sprawdzają się w obszarach badań o nieustrukturyzowanych zbiorach danych. Ich wykorzystanie w budowie modeli retencji klientów jako substytut procesu przygotowania zmiennych pochodnych jest nadal przedmiotem badań [Castandedo i inni, 2014, Spanoudes, Thomson, 2017].

2.5. Niezbilansowany rozkład zmiennej zależnej jako problem w prognozowaniu zjawisk o charakterze rzadkim

W większości firm zjawisko migracji klientów należy do zdarzeń relatywnie rzadkich. W skali roku rezygnuje z usług maksymalnie do 25% klientów[43]. Zbiór danych będący podstawą budowy modelu odzwierciedlający rzeczywistą stopę rezygnacji klientów, zawiera zmienną zależną o niezbilansowanym rozkładzie. Jest to okoliczność wpływająca negatywnie na proces identyfikacji wzorców przez metody uczenia maszynowego. Negatywny wpływ braku równowagi w zbiorze danych na proces identyfikacji wzorców wynika w dużej mierze z używanych podczas uczenia kryteriów optymalizacji. Najczęściej oparte są one na mierze ACC (*Accuracy*), bądź innych miarach wrażliwych na

[42] Tego typu przekształcenia nazywane są interakcjami.
[43] Na podstawie [www.statista.com, 2020].

niezbilansowanie rozkładu zmiennej zależnej [44]. Podczas procesu uczenia algorytm optymalizuje ogólny błąd, przez co ma skłonność do uczenia się prawidłowości informującej o generalnie wysokiej lojalności klientów nie koncentrując się na czynnikach różnicujących osoby lojalne od nielojalnych. Skutkuje to faworyzowaniem przypadków należących do bardziej licznej klasy oraz brakiem zdolności dyskryminacyjnych modelu[45].

Występujące w literaturze strategie łagodzenia negatywnego wpływu niezbilansowanego zbioru danych na proces uczenia można podzielić na następujące grupy [Haixiang i inni, 2017, Kubus, 2020, Loyola-Gonzalez i inni, 2016]:

- korygujące oryginalny zbiór danych przez różnego rodzaju strategie próbkowania (*resampling*) bądź generowania nowych, sztucznych przypadków;
- korygujące działanie istniejących algorytmów uczenia maszynowego poprzez zmianę ich hiperparametrów, najczęściej kosztów błędnych klasyfikacji, wag przypadków bądź punktu odcięcia;
- wykorzystywanie technik uczenia maszynowego mniej wrażliwych na problem niezbilansowania klas zmiennej zależnej, bądź tworzenie modyfikacji istniejących algorytmów odpornych na ten problem.

Więcej informacji o technikach związanych z korektą hiperparametrów oraz z metodami uczenia maszynowego zostało przedstawionych w rozdziale 3. W dalszej części przedstawiono strategie korekty oryginalnego zbioru danych.

Strategie bilansowania zmiennej zależnej polegające na próbkowaniu najogólniej można podzielić na dwie kategorie [Fernandez i inni, 2018].

- *Undersampling*, polegający na utworzeniu nowego zbioru uczącego, który zachowuje wszystkie przypadki nielojalnych klientów. Z grupy klientów lojalnych (klasy większościowej) są losowane przypadki w liczbie zbliżonej do liczności klasy mniejszościowej. Niewylosowane przypadki są zatem pomijane w analizie.
- *Oversampling*, polegający na utworzeniu nowego zbioru uczącego, który zachowuje wszystkie przypadki klasy lojalnych klientów (większościowej). Przypadki klientów nielojalnych (klasa mniejszościowa) są losowane ze zwracaniem do momentu, gdy ich liczność będzie zbliżona do liczności przypadków z grupy liczniejszej. Po tej

[44] Miary oceny modeli oraz ich wrażliwość na niezbilansowanie próby opisano w rozdziale 4 oraz w [Migut, 2020].

[45] Model informujący, że wszyscy klienci są lojalni, w sytuacji gdy liczba nielojalnych klientów wynosi 5% ma dokładność na poziomie 95%.

operacji poszczególne przypadki osób nielojalnych mogą być wielokrotnie reprezentowane w nowym zbiorze.

Zaletą techniki *undersampling* jest fakt, iż uzyskany w jej wyniku zbiór danych jest zazwyczaj niewielkich rozmiarów, co znacząco przyspiesza proces uczenia modelu. Podejście to generuje jednak szereg problemów wynikających z faktu iż wylosowane przypadki mogą w niedostatecznym stopniu reprezentować zbiór lojalnych klientów prowadząc tym samym do zmniejszenia zdolności klasyfikatora do generalizacji. *Oversampling* jest z kolei *de facto* inną formą zwiększenia wag przypadków rzadszych, bądź podniesieniem kosztów ich błędnej klasyfikacji. Technika ta może wzmacniać szumy obecne w zbiorze nielojalnych klientów oraz prowadzić do nadmiernego dopasowania do poszczególnych punktów zawartych w zbiorze danych. Możliwe są również działania hybrydowe, łączące cechy powyższych strategii obarczone jednak wadami obydwu podejść [Loyola-Gonzalez i inni, 2016].

Wady przedstawionych powyżej strategii próbkowania spowodowały wzrost zainteresowania technikami bilansowania opartymi na uzupełnianiu oryginalnego zbioru danych poprzez wygenerowane sztuczne przypadki. Przełomową okazała się praca N.V. Chawla i innych, [2002] prezentująca algorytm SMOTE (*Synthetic Minority Over-sampling Technique*). Nowe dane są tworzone za pomocą interpolacji na podstawie przypadków należących do mniej licznej klasy, znajdujących się w ich sąsiedztwie. W ogólnym zarysie algorytm losuje przypadek z rzadszej klasy i wyznacza jego najbliższych sąsiadów[46]. Z każdym z sąsiadów z osobna tworzy parę, na podstawie której interpolowany jest nowy przypadek. Nowy przypadek leży na linii łączącej wylosowany punkt z wybranym sąsiadem. Przed wyznaczeniem punktu losowana jest liczba z przedziału (0,1). Do wyznaczenia nowego punktu dla każdej zmiennej ilościowej wyznaczana jest różnica pomiędzy wylosowanym przypadkiem i jego sąsiadem. Różnica ta jest mnożona przez wylosowaną liczbę, a następne dodawana do oryginalnej wartości wylosowanego przypadku. Dla zmiennych nominalnych klasa nowego przypadku jest losowana.

Podejście to pozwala uniknąć niedoskonałości wymienionych powyżej algorytmów próbkowania. Unika wad związanych z techniką *undersampling* ponieważ zachowuje bardziej liczną grupę klientów w oryginalnej formie nie dopuszczając do utraty informacji. Pozwala też uniknąć ryzyka nadmiernego dopasowania do pojedynczych obserwacji co cechowało metodę *oversampling*. Pomimo zalet, SMOTE nie rozwiązywało wszystkich

[46] Domyślnie pięciu.

problemów związanych z niezbilansowanymi zbiorami. W opracowaniu A. Fernandez i in., [2018] przedstawiono ponad 85 rozszerzeń oraz udoskonaleń tego podejścia opracowanych w ciągu 15 lat, jakie minęły od opracowania oryginalnej wersji algorytmu. Spośród wymienionych siedmiu typów rozszerzeń algorytmu najważniejszymi wydają się być:

- korekta sposobu losowania przypadków,
- korekta sposobu interpolacji.

Przykładem algorytmu korygującego sposób losowania przypadków jest *Safe-Level SMOTE*, który tworzy sztuczne przypadki jedynie w regionach bezpiecznych. Bezpieczeństwo danego przypadku jest definiowane na podstawie analizy *k* jego najbliższych sąsiadów. Przypadki, które wśród swoich *k* najbliższych sąsiadów nie mają żadnego innego przypadku z rzadszej klasy są traktowane jako szum i nie uczestniczą w generowaniu sztucznych danych. Przypadki, które wśród swoich *k* najbliższych sąsiadów miały przypadki z rzadszej klasy w liczbie *k* bądź do niej zbliżonej były traktowane jako bezpieczne i na ich podstawie realizowano interpolację [Loyola-Gonzalez i inni, 2016]. Innym przykładem korekty wyboru przypadków może być użycie wektorów wspierających uzyskanych za pomocą metody SVM [Fernandez i inni, 2018]. Odwrotna filozofia przyświeca metodzie ADASYN (*Adaptive Synthetic Sampling Approach for Imbalanced Learning*) przedstawionej w 2008 roku [He i inni]. Przypadki z rzadszej klasy, w otoczeniu których znajdują się jedynie przypadki z klasy dominującej są uznawane za trudne od nauczenia i ich udział w procesie generowania nowych przypadków jest wzmacniany. Przypadki, w których otoczeniu znajdują się jedynie przypadki z rzadszej klasy nie biorą udziału w generowaniu sztucznych danych. Autorzy algorytmu zwracają uwagę na podobieństwo do wzmacniania (*boosting*) używanego na przykład w metodzie drzew wzmacnianych.

Korekta sposobu generowania nowych przypadków może polegać między innymi na wykorzystaniu więcej niż jednego sąsiada, wykonaniu analizy skupień i wyznaczeniu nowych przypadków za pomocą środków skupień, czy też za pomocą metod jądrowych[47] [Fernandez i inni, 2018]. Oczekiwaną własnością algorytmu generowania nowych przypadków jest umiejętność odfiltrowania artefaktów z danych i równocześnie wzmocnienia grup przypadków na obszarach słabiej reprezentowanych oraz przypadków

[47] Przykładem może być algorytm ROSE [Menardi, Torelli, 2014].

bliskich granicy decyzyjnej. Wydaje się, że na ten moment nie istnieje przyjęty standard realizujący powyższe postulaty.

Rozdział 3 Budowa optymalnego modelu klasyfikacyjnego

3.1. Wybór zmiennych podczas budowy modelu

Opisany powyżej proces czyszczenia danych skutkuje uzyskaniem zbioru danych zawierającego kompletne obserwacje, przyjmującego prawdopodobne wartości i wzbogaconego o zmienne pochodne. Kolejnym krokiem procesu analitycznego jest zadanie polegające na zbudowaniu modelu, który na podstawie dostępnych danych najlepiej spełni postawione cele biznesowe. Proces budowy modelu jest realizacją działań związanych z:

- wyborem optymalnego podzbioru zmiennych,
- wyborem metody bądź metod analitycznych[48],
- określeniem optymalnego zestawu hiperparametrów dla każdej z uwzględnianych metod,
- wyborem optymalnej strategii agregacji zbudowanych modeli.

Poza wymienionymi działaniami związanymi bezpośrednio z procesem budowy modeli, niejako w bezpośrednim sąsiedztwie tego procesu, należy wyróżnić dodatkowe działania mające równie znaczący, aczkolwiek pośredni wpływ na wynik modelowania. Do tej grupy należy zaliczyć:

- sposób doboru próby do budowy modelu,
- wybór kryterium oceny modeli,
- określenie sposobu walidacji modeli[49].

[48] Docelowy model może być modelem hybrydowym, czyli modelem wykorzystującym wyniki dwóch lub więcej metod.

[49] Zagadnienia te zostaną poruszone w dalszej części pracy.

Prawidłowa realizacja wymienionych powyżej działań ma umożliwić zbudowanie modelu o zadowalającej mocy predykcyjnej, opartym na możliwie najmniejszym zestawie zmiennych X, odpornym na niewielkie zmiany w strukturze populacji klientów, a co najważniejsze skutecznie wskazującym osoby zagrożone odejściem do konkurencji.

Istotą procesu wyboru zmiennych jest odnalezienie możliwie najmniejszego ich podzbioru, który użyty do budowy modelu pozwoli na uzyskanie zadawalających wyników. Podzbiór ten powinien składać się jedynie ze zmiennych wpływających na modelowane zjawisko oraz być pozbawiony nadmiarowych cech [Khun, Johnson, 2013]. Poniżej przedstawiono najczęstsze przyczyny przeprowadzania procedury wyboru zmiennych. Ich znaczenie może się zmieniać w odniesieniu do różnych metod analizy danych. Redukcja zestawu zmiennych może wynikać z chęci [Radcliffe, Surry, 2011]:

- uproszczenia modelu – model uwzględniający nadmierną liczbę zmiennych generuje ryzyko dopasowania się do szumów zawartych w danych, co w konsekwencji przekłada się na spadek jego zdolności do generalizacji;
- uniknięcia nadmiernej korelacji pomiędzy predyktorami – redundancja może prowadzić do problemów numerycznych podczas budowy modelu oraz utrudniać interpretację uzyskanych wyników, zmniejsza również stabilność zbudowanego modelu;
- poprawy jakości modelu – usunięcie zmiennych może poprawiać zdolność modelu zarówno do aproksymacji zależności, jak i do generalizacji;
- poprawy stabilności modelu – niestabilny w czasie wpływ zmiennej na modelowane zjawisko może powodować niestabilne działanie modelu w przyszłości;
- ułatwienia interpretacji modelu – budowa modelu zrozumiałego dla odbiorców biznesowych może powodować konieczność wyboru jedynie zmiennych, których wpływ na zmienną zależną ma biznesowe uzasadnienie.

Na podstawie powyższych kryteriów można stwierdzić, iż celem wyboru zmiennych jest uzyskanie takiego podzbioru dostępnych cech, który przy pomocy wybranej metody modelowania umożliwi zbudowanie modelu o możliwie największej zdolności do generalizacji, odpornego na spadek mocy predykcyjnej pojedynczych cech, bazującego na możliwie niewielkiej liczbie zmiennych, najlepiej mających interpretację biznesową, których wpływ jest stabilny w czasie.

Należy zauważyć, że dla d dostępnych zmiennych istnieje 2^d ich możliwych podzbiorów. Znalezienie optymalnego podzbioru musiałoby się wiązać z budową modelu dla każdej kombinacji cech i wyborem najlepszego z nich, czyli realizacją tzw. „metody siłowej" (*brute force*). Dla mniejszych zbiorów danych może to być strategia możliwa do realizacji, natomiast dla większych staje się niepraktyczna ze względu na czas potrzebny do przeprowadzenia obliczeń.

W praktyce, badacz nie poszukuje modelu najlepszego, a jedynie modelu wystarczająco dobrego (a co za tym idzie wystarczająco dobrego zestawu zmiennych), który mógłby zostać zbudowany w akceptowalnym dla niego czasie. Konsekwencją tego faktu jest stosowanie różnego rodzaju strategii heurystycznych gwarantujących identyfikację jedynie rozwiązań suboptymalnych. Strategie te zakładają najczęściej konieczność użycia kombinacji kilku metod selekcji, pozwalających na stosowanie różnorodnych kryteriów ważności zmiennych.

Metody wyboru zmiennych można podzielić na następujące grupy [Kuhn, Johnsonn, 2013, Guyon i inni, 2008]:

- niezależne od metody analitycznej – filtry (*filter methods*):
    - nieukierunkowane (*unsupervised*),
    - ukierunkowane (*supervised*),
- zależne od metody analitycznej:
    - metody opakowujące metody analityczne (*wrapper methods*),
    - metody wbudowane w metody analityczne (*embedded methods*).

Filtry są stosowane we wstępnej fazie procesu wyboru zmiennych ze względu na szybkość działania. Głównym zadaniem filtrów jest preselekcja zmiennych. Na ich podstawie z dostępnego zestawu cech odrzucane są zmienne X o małej mocy dyskryminacyjnej, które innymi słowy mają niewielkie szanse na odegranie znaczącej roli w procesie predykcji [Guyon i inni, 2008]. Podzbiór cech powstały po ich zastosowaniu nadal może zawierać pewną liczbę predyktorów nieistotnych z punktu widzenia celu modelowania, ale będzie on jednak znacząco mniejszy od zbioru pierwotnego. Zredukowana liczba predyktorów umożliwia zastosowanie metod zależnych od metody analitycznej, które są bardziej wymagające pod względem mocy obliczeniowej. Metody te cechują się lepszą dokładnością w wyborze optymalnego zestawu zmiennych. Skutecznie zrealizowana heurystyczna strategia wyboru zmiennych powinna zatem polegać na

wykorzystaniu efektywności filtrów oraz dokładności metod opakowujących i wbudowanych.

3.1.1. Filtry ukierunkowane i nieukierunkowane

Metody nieukierunkowane umożliwiają dokonanie wyboru zmiennych bez oceny ich związku ze zmienną zależną. Opierają się na analizie predyktorów kandydujących do modelu pod kątem ich rozkładu i współzmienności z innymi zmiennymi.

Analiza rozkładu cech pozwala na eliminację predyktorów niewykazujących zmienności (stałe zmienne) oraz zmiennych nominalnych o liczbie klas zbliżonych do liczby przypadków. Osobną grupę zmiennych stanowią zmienne rzadkie (*sparse data*), które przyjmują stałą wartość dla więcej niż 95% przypadków. Zmienne tego typu zazwyczaj są eliminowane przez filtry ukierunkowane. Ich niewielka zmienność w konsekwencji prowadzi do ograniczonej siły predykcyjnej. W sytuacji, gdy zmienne rzadkie stanowią znaczny odsetek predyktorów, bądź wiedza biznesowa badacza wskazuje na ich znaczący, merytoryczny wpływ na skłonność klientów do odejścia, dobre wyniki daje procedura polegająca na agregacji zmiennych rzadkich do jednej bądź kilku zmiennych pochodnych. Szereg metod tworzenia tego typu zmiennych zostało zaproponowanych w pracy D. Pyle [1999].

Utworzone zmienne informują zatem ile razy wystąpiło pewne zdarzenie rzadkie. Dzięki agregacji nowopowstałe predyktory mogą wykazywać większą zmienność oraz większą siłę predykcyjną i tym samym odegrać istotną rolę w procesie budowy modelu. Słabą stroną stosowania tej oraz innych metod agregacji zmiennych rzadkich jest jej pracochłonność zarówno na etapie budowy modelu jak i jego wdrożenia w praktyce.

Analiza współzmienności ma na celu eliminację z procesu modelowania nadmiarowych zmiennych. Pomimo pojawiających się w literaturze głosów [Guyon i inni, 2008], iż użycie dwóch idealnie skorelowanych cech niekoniecznie musi prowadzić do redundancji danych, powszechną praktyką jest wybór reprezentantów opisujących określone źródło zmienności. Wyboru zmiennych można dokonać na podstawie analizy macierzy korelacji[50], eliminując zmienne wysoko skorelowane z innymi predyktorami. Strategią eliminacji zmiennych opartej na tej zasadzie jest metoda eliminacji wektorów przedstawiona w pracy S. Chomątowskiego i A. Sokołowskiego [1978]. Odmienna

[50] W zależności od skali pomiaru predyktorów może być to korelacja liniowa Pearsona, korelacja Spearmana, bądź spółczynnik V Cramera.

procedura eliminacji nadmiarowych zmiennych została przedstawiona w pracy G. Miguta i innych [2014]. Proces identyfikacji i eliminacji zmiennych został oparty na analizie czynnikowej realizowanej za pomocą metody głównych składowych z rotacją czynników. Procedura po identyfikacji liczby głównych składowych [51] i wykonaniu analizy czynnikowej z rotacją macierzy ładunków wymaga analizy uzyskanej macierzy ładunków. W każdej kolumnie macierzy ładunków identyfikowane są zmienne wysoko skorelowane z daną składową[52]. Z grupy tych zmiennych wybiera się od dwóch do kilku reprezentantów. Pozostałe zmienne z danej grupy są odrzucane z analizy.

Poza przedstawionymi heurezami w procesie wstępnego wyboru zmiennych zastosowanie mogą mieć inne filtry nieukierunkowane. Przykładem takich filtrów są *spectral feature selection (SPEC)* oraz *Laplacian Score (LS)* wykorzystywane jako narzędzia wyboru zmiennych w analizie skupień, jednak mogące mieć zastosowanie również w zagadnieniach klasyfikacyjnych [Alelyani i inni, 2013].

Stosowanie filtrów ukierunkowanych polega na obliczeniu dla każdego potencjalnego predyktora wybranej miary bądź miar oceniających siłę jego związku ze zmienną zależną. Następnie na podstawie uzyskanych miar tworzy się ranking predyktorów. Kolejny krok to określenie punktu odcięcia na podstawie zadanej z góry liczby predyktorów, które mogłyby być brane pod uwagę w dalszej analizie, bądź też na podstawie progu[53], powyżej którego dana zmienna mogłaby zostać uznana za ważną.

Miary siły predykcyjnej używane w praktyce budowy modeli lojalności można podzielić na następujące grupy [Jović i inni, 2015]:

- oparte na teorii informacji,
- statystyczne, oparte na korelacji lub odległości pomiędzy rozkładami,
- oparte na odległości lub podobieństwie.

W miarach opartych na teorii informacji stosuje się tzw. kryterium informacyjne (*IV, Information Value*). Jest to miara oparta na entropii. Pozwala ocenić całkowitą siłę dyskryminacyjną analizowanej zmiennej X. Oblicza się ją za pomocą następującego wzoru [Siddiqui, 2017]:

[51] Na przykład za pomocą kryterium Kaisera lub Cattela.
[52] Najczęstszą wartością progową jest $r \geq |0{,}7|$.
[53] Określanego metodą ekspercką.

$$IV = \sum_{i=1}^{n} (Distr\ Good_i - Distr\ Bad_i) \times ln\ (\frac{Distr\ Good_i}{Distr\ Bad_i})$$

gdzie n jest liczbą kategorii analizowanego predyktora, natomiast *Distr Good* oraz *Distr Bad* odnoszą się odsetka osób lojalnych (w odniesieniu do wszystkich osób lojalnych) i nielojalnych (w odniesieniu do wszystkich osób nielojalnych) w danej grupie. Miara ta pozwala obliczyć siłę wpływu zmiennych jakościowych na zmienną zależną. Predyktory mierzone na skalach mocnych, przed jej obliczeniem, są poddawane kategoryzacji, najczęściej na 10 równolicznych grup.

Miara ta jest szeroko wykorzystywana w procesie budowy modeli klasyfikacyjnych w różnych branżach. Na jej podstawie można przyjąć następującą interpretację siły predykcyjnej analizowanej cechy [Siddiqui, 2006]:

- poniżej 0,02[54] – brak mocy predykcyjnej,
- od 0,02 do 0,10 – słaba moc predykcyjna,
- od 0,10 do 0,30 – przeciętna moc predykcyjna,
- powyżej 0,30 – duża moc predykcyjna.

Wartość *IV* powyżej 0,5 świadczy o bardzo dużej sile predykcyjnej zmiennej. Zmienna taka może być źródłem niestabilnego działania modelu zwłaszcza w sytuacji, gdy pozostałe zmienne charakteryzują się znacznie mniejszą siłą predykcyjną [Siddiqui, 2017]. Identyfikacja zmiennej o tak dużej sile predykcyjnej jest z jednej strony okolicznością pozytywną, implikuje bowiem możliwość budowy dobrze dopasowanego modelu, z drugiej strony narzuca konieczność redukcji ryzyka nadmiernego jej wpływu na końcową postać modelu. Preferowanym rozwiązaniem stosowanym w takiej sytuacji jest budowa modelu hybrydowego. Za pomocą drzew klasyfikacyjnych dokonuje się najczęściej jednego podziału zbioru danych względem dominującego predyktora. Uzyskane podzbiory są podstawą do budowy niezależnych modeli z niezależnie przeprowadzonym procesem wyboru zmiennych objaśniających.

Mocną stroną miary *IV* jest jej addytywny charakter. Należy zwrócić uwagę, że jest ona sumą wartości *Weight of Evidence* (*WoE*) ważonych różnicą w odsetku osób lojalnych i nielojalnych w danej kategorii. Ten fakt pozwala ocenić wkład każdej z kategorii predyktora w ogólną wartość siły predykcyjnej. Kolejną zaletą tej miary jest jej

[54] Na podstawie doświadczeń autora należałoby rozważyć przesunięcie tej granicy do poziomu 0,05.

niewrażliwość na problem niezbalansowanych proporcji klas zmiennej zależnej. Słabą stroną jest natomiast jej wrażliwość na liczbę klas, na którą podzielono przed jej obliczeniem predyktor ilościowy. Poprawne obliczenie wartości *IV* wymaga również, aby w każdej kategorii znajdowała się przynajmniej jedna osoba lojalna i jedna osoba nielojalna.

Innym popularnym filtrem opartym na teorii informacji jest tzw. zysk informacji (*information gain*). Miara ta, podobnie jak opisane dalej współczynnik zysku (*gain ratio*) oraz symetryczna niepewność (*symmetrical uncertainty*) obok pełnienia roli filtra dla problemów klasyfikacyjnych [55] jest wykorzystywana w procesie budowy drzew decyzyjnych do poszukiwania optymalnego punktu podziału. Stosowanie jej dla predyktorów ilościowych wymaga ich wcześniejszej dyskretyzacji. Obliczenie miary *information gain* wymaga uprzednio określenia wartości informacji dla zmiennej zależnej Y. W odniesieniu do modeli lojalności można wyrazić ją za pomocą następującego wzoru [Lin, 2018]:

$$I(Z) = (-1) \times \left[\frac{L}{T} \times log_2\left(\frac{L}{T}\right) + \frac{N}{T} \times log_2\left(\frac{N}{T}\right)\right]$$

W równaniu *L* odnosi się do liczby osób lojalnych a *N* do liczby osób nielojalnych. Wartość *T* to liczba wszystkich osób w zbiorze treningowym. Wartość *I(Z)* jest również nazywana entropią zbioru uczącego. W kolejnym kroku oblicza się informację uwzględniającą klasy analizowanego predyktora *Pred*:

$$I_{Pred}(Z) = \sum_{j=1}^{m} \frac{T_j}{T} \times (-1) \times \left[\frac{L_j}{T_j} \times log_2\left(\frac{L_j}{T_j}\right) + \frac{N_j}{T_j} \times log_2\left(\frac{N_j}{T_j}\right)\right]$$

gdzie T_j odnosi się do liczebności *j*-tej klasy analizowanego predyktora, L_j oraz N_j odpowiednio do liczby osób lojalnych i nielojalnych w obrębie *j*-tej klasy predyktora. Na podstawie powyższych wzorów *information gain* oblicza się w następujący sposób:

$$Information\ gain = I(Z) - I_{Pred}(Z)$$

Wadą miary *information gain* jest jej wrażliwość na liczbę klas analizowanego predyktora skutkującą faworyzowaniem zmiennych posiadających dużą liczbę klas. Miara

[55] Miara używana w popularnych programach do analizy danych, na przykład w programie Weka.

gain ratio redukuje to obciążenie przez uwzględnienie w swoim wzorze informacji, jaka jest zawarta w samym predyktorze [Lin, 2018]:

$$I(Pred) = (-1) \times \sum_{j=1}^{m} \frac{T_j}{T} \times \ log_2(\frac{T_j}{T})$$

Gain ratio oblicza się na podstawie poniższego wzoru:

$$Gain\ ratio = \frac{Information\ gain}{I(Pred)}$$

Kolejną miarą opartą na teorii informacji jest *symmetrical uncertainty* wyrażoną wzorem [Weka, 2022].

$$Symmetrical\ uncertainty = \frac{2 \times Information\ gain}{I(Pred) \times I(Z)}$$

Miara ta redukuje obciążenie miary *information gain* związane z liczbą klas analizowanych predyktorów. Przyjmuje wartości w zakresie [0,1]. Innymi filtrami opartymi na teorii informacji są dywergencja Kullbacka-Leiblera (*Mutual Information*), minimalna długość opisu (*Minimum Description Length, MDL*) czy odległość Matrasa (*Matras distance*) przedstawione na przykład w pracy I. Guyon i inni [2008].

Do popularnych filtrów statystycznych zalicza się miary V Cramera oraz separację Fishera. V Cramera jest miarą siły związku dla cech jakościowych (nominalnych). Miara ta przyjmuje wartości w zakresie [0,1], gdzie 0 oznacza brak korelacji a 1 idealną korelację. W przypadku dwuwariantowej zmiennej zależnej może być sformułowana jako [Mynarski, 2003]:

$$V = \sqrt{\frac{\chi^2}{n}}$$

Gdzie χ^2 oznacza wartość statystyki chi-kwadrat, a *n* oznacza liczbę przypadków w analizowanym zbiorze danych.

Separacja Fishera jest miarą opartą na różnicy średniej wartości danej zmiennej w grupie osób lojalnych m_L, oraz w grupie osób nielojalnych m_N, znormalizowanej z wykorzystaniem ich wariancji σ_L^2 i σ_N^2. Przyjmuje wartości z przedziału [0; +∞). Oblicza się ją za pomocą wzoru [Guyon i inni, 2008]:

$$FS = \frac{(m_L - m_B)^2}{\sigma_L^2 + \sigma_N^2}$$

Im większa wartość tego wskaźnika, tym większa siła predykcyjna analizowanej zmiennej. Przypisuje on najwyższy wynik predyktorom, dla których przypadki z różnych klas są dalekie od siebie przy jednoczesnym wymogu, aby przypadki należące do tej samej klasy znajdowały się blisko siebie. Podobną miarą opartą na separacji dwóch średnich jest statystyka *t-Studenta*.

Do filtrów statystycznych zalicza się również statystkę K-S (Kołmogorowa-Smirnowa), wykorzystywaną częściej jako miara siły predykcyjnej zbudowanego modelu, a która została opisana w dalszej części niniejszego rozdziału. Innymi filtrami statystycznymi opisanymi w [Guyon i inni, 2008] są wskaźnik rozróżnienia dwóch rozkładów normalnych (*Bi-normal separation*) czy odległość Jeffreya-Matusita (*Jeffreys-Matusita distance)*.

Kolejną grupą filtrów ukierunkowanych są miary oparte na odległości. Do popularnych sposobów selekcji zmiennych należących do tej grupy zalicza się algorytm *Relief* oraz jego odmianę *ReliefF*. *Relief* jest procedurą identyfikacji zmiennych opartą na metodzie k-najbliższych sąsiadów. Zaletą przedstawionej metody jest fakt, że ocena siły predykcyjnej danej zmiennej uwzględnia również jej interakcje z innymi predyktorami bez konieczności tworzenia w trakcie analizy zmiennych pochodnych. Co ważne algorytmy z rodziny *Relief* zachowują zalety innych filtrów to znaczy relatywną szybkość obliczeń oraz niezależność od docelowego algorytmu budowy modelu [Urbanowicz i inni, 2018]. Podobnie jak inne filtry ukierunkowane metoda ta nie jest w stanie wyeliminować redundancji obecnej w zbiorze danych. Algorytm *Relief* jest wykonywany w następujący sposób:

1. *Ze zbioru danych liczącego N obiektów i Z zmiennych:*
    a. *wylosuj za pomocą losowania prostego zależnego L przypadków,*
    b. *zainicjalizuj wektor wag W[Z], przypisując każdej zmiennej z_j wagę $W[z_j]=0$.*
2. *Dla wszystkich L przypadków:*
    a. *weź wzorcowy przypadek l_i,*
    b. *odszukaj w zbiorze zdanych najbliższy przypadkowi wzorcowemu obiekt h, dla którego klasa zmiennej zależnej jest zgodna z klasą wzorcowego przypadku l_i,*
    c. *odszukaj w zbiorze zdanych najbliższy przypadkowi wzorcowemu obiekt m, dla którego klasa zmiennej zależnej jest niezgodna z klasą wzorcowego przypadku l_i.*
    d. *Dla każdej zmiennej z_i:*

i. oblicz różnicę $diff(z_j, l_i, h)$ *pomiędzy przypadkiem wzorcowym* l_i *oraz obiektem h,*

ii. oblicz różnicę $diff(z_j, l_i, m)$ *pomiędzy przypadkiem wzorcowym* l_i *oraz obiektem m,*

iii. skoryguj wagę zmiennej zgodnie z poniższym wzorem:

1. $W[z_j] = W[z_j] - \frac{diff(z_j, l_i, h)}{L} + \frac{diff(z_j, l_i, m)}{L}$.

3. Zwróć wektor W[Z] zawierający oceny ważności analizowanych zmiennych.

Dla zmiennych jakościowych różnica pomiędzy przypadkami wynosi 0 w sytuacji zgodności oraz 1 w sytuacji niezgodności. Dla zmiennych ilościowych przyjęto ją obliczać jako wartość bezwzględną różnicy wartości znormalizowaną przez podzielenie przez rozstęp[56]. W analogiczny sposób obliczane są odległości pozwalające zidentyfikować przypadki *h* oraz *m*.

Podzielenie każdej różnicy przez wartość L równą liczbie przypadków wzorcowych gwarantuje, że wartości $W[z_j]$ będą znajdowały się w przedziale [-1,1]. W zbiorze danych zawierającym zarówno zmienne ilościowe jak i jakościowe, wpływ zmiennych ilościowych może być niedoszacowany.

W praktyce powyższy algorytm jest rzadko stosowany. W praktycznych implementacjach został zastąpiony jego rozwinięciem *ReliefF* [Urbanowicz, 2018], głównymi różnicami w tym podejściu są:

- poleganie na większej liczbie sąsiadów zamiast jednego, najczęściej na 10 sąsiadach[57],
- obsługa braków danych – w przypadku braku danych wartość różnicy jest zastępowana warunkowym prawdopodobieństwem, że dla danej zmiennej dwa przypadki mają różne wartości,
- wykorzystanie zmiennych zależnych o wielu klasach (mniej istotna cecha z punktu widzenia analizowano problemu lojalności klientów),
- uwzględnienie w analizie wszystkich przypadków jako przypadków wzorcowych.

Niedoszacowanie wpływu zmiennych jakościowych skorygowane może być przez dyskretyzację predyktorów. Należy zwrócić uwagę, że problem oceny siły predykcyjnej predyktora jest w dużej mierze analogiczny do oceny siły predykcyjnej całego modelu.

[56] Zgodnie z formułą na odległość miejską (Manhattan). Inne miary odległości dają podobny wynik, odległość miejska stała się najbardziej popularna ze względu na swoją prostotę [Urbanowicz, 2018].

[57] Dodatkowo wpływ poszczególnych sąsiadów może być odwrotnie proporcjonalny do ich odległości od obiektu wzorcowego czego nie przewiduje algorytm *RelieF*.

Wyniki modelu klasyfikacyjnego można interpretować jako zmienną rangującą, a siłę predykcyjną modelu ocenia się na podstawie powiązania tej zmiennej ze zmienną zależną. Szereg miar siły predykcyjnej jest wykorzystywanych na etapie wyboru zmiennych. Dwiema popularnymi miarami z tej grupy są współczynnik Giniego oraz miara *F-Score*. Własności tych oraz innych miar zostały opisane w podrozdziale dotyczącym walidacji modelu.

Osobną metodą wyboru zmiennych opartą na filtrach jest CFS (*Correlation based Feature Selection*) przedstawiona pierwotnie przez M.A. Halla [1999]. Istotą tej metody jest realizacja postulatu, że dobrze wybrany podzbiór cech powinien charakteryzować się wysokim powiązaniem ze zmienną zależną oraz niskim z pozostałymi wybranymi predyktorami. CSF ocenia jakość całego podzbioru zmiennych a nie pojedynczego predyktora. Jakość ta jest obliczana według wzoru [Hall, 1999],

$$M_S = \frac{k \times \overline{r_{cf}}}{\sqrt{k + k \times (k-1) \times \overline{r_{ff}}}},$$

gdzie k oznacza liczbę zmiennych w proponowanym podzbiorze, $\overline{r_{cf}}$ oznacza średnią miarę powiązania pomiędzy predyktorami a zmienną zależną, $\overline{r_{ff}}$ oznacza średnią miarę powiązania pomiędzy parami predyktorów. Metoda umożliwia wykorzystanie preferowanego przez badacza filtra. Miarami powiązania użytymi w oryginalnej pracy były *symmetrical uncertainty*, *ReliefF* oraz *MDL*. Optymalny podzbiór predyktorów maksymalizuje wartość M_S[58].

Miarą realizującą postulat analogiczny do CFS jest przedstawiona przez E. Gatnara [2008] integralna pojemność informacji. Dla wybranego podzbioru Z zawierającego k predyktorów oblicza się ją następująco:

$$H(Z_k) = \sum_{i=1}^{k} \frac{r^2(Y, X_i)}{\sum_{j=1}^{k} \left| r(X_i, X_j) \right|}$$

Najlepszym podzbiorem zmiennych jest zestaw maksymalizujący powyższe wyrażenie. Podobnie jak w przypadku CFS współczynnik korelacji może zostać zastąpiony inną miarą powiązania odpowiednią dla dwustanowej zmiennej zależnej. Znalezienie optymalnego

[58] Pod warunkiem, że większa wartość wybranej miary powiązania oznacza silniejszy wpływ predyktora na zmienną zależną.

podzbioru zmiennych maksymalizującego wyrażenia M_s oraz $H(Z_k)$ jest możliwe za pomocą algorytmów opisanych w kolejnym punkcie.

3.1.2. Metody opakowujące

Realizacja strategii opakowujących polega na wielokrotnym powtarzaniu procesu budowy i oceny modelu za pomocą wybranej metody analitycznej, na różnych podzbiorach predyktorów. Konieczność budowy modelu dla każdego podzbioru zmiennych powoduje, że podejście to jest wymagające obliczeniowo. Znalezienie (sub)optymalnego zestawu zmiennych wymaga zazwyczaj wykonania wielu iteracji. Dla każdej z nich podzbiór zmiennych do analizy jest przygotowany zgodnie z określoną strategią, której najpopularniejsze odmiany zostały przedstawione poniżej. Wysokie wymagania pod względem obliczeniowym ograniczają stosowanie tych strategii dla bardziej skomplikowanych metod analitycznych.

Wykorzystanie w procedurze wyboru zmiennych konkretnej metody analitycznej jest źródłem przewagi tej metody nad filtrami. Pozwala bowiem dostosować wybór zmiennych do specyfiki konkretnej metody analitycznej, znacząco zawęzić zbiór danych, redukując redundancję. Z drugiej strony, ponieważ zestaw zmiennych wskazanych przez procedurę jest oparty na konkretnej metodzie analitycznej nie może być traktowany jako rozwiązanie ogólne.

Wybór kolejnego podzbioru powinien być przeprowadzany na podstawie oceny działania modelu na zbiorze testowym. Wiarygodna ocena jakości zbudowanego modelu a tym samym potwierdzenie trafności wyboru podzbioru zmiennych wymaga użycia niezależnego zbioru walidacyjnego oraz niezależnie innego algorytmu do budowy modelu potwierdzającego skuteczność wyboru [Jović i inni, 2015]. Metody opakowujące można zaklasyfikować do trzech grup:

- wykładnicze,
- sekwencyjne,
- randomizowane.

Metody wykładnicze zgodnie z ich nazwą cechują się rosnącą wykładniczo złożonością obliczeń wraz ze wzrostem wymiarowości przestrzeni przeszukiwania. Przykładem algorytmu z tej grupy jest wspomniana wcześniej metoda siłowa (*brute force*) sprawdzająca wszystkie kombinacje predyktorów. Innym algorytmem z tej grupy jest

metoda podziału i ograniczeń (*branch and bound*). Metoda ta zakłada, że usunięcie zmiennej ze zbioru danych nie może spowodować poprawy działania modelu (założenie monotoniczności kryterium oceny modelu[59]). Przyjmuje też założenie, że dla ustalonego progu jakości modelu najlepszy zestaw zmiennych to zestaw najmniejszy. Uwzględniając te założenia, z rozwiązania zawierającego wszystkie zmienne eliminowane są kolejno zmienne, aż do momentu osiągnięcia przez model granicznego poziomu jakości. Algorytm następnie bada inne sposoby usuwania zmiennych zgodnie z zasadą przeszukiwania w głąb (*depth-first search*), które nie naruszają progu, i które mogą zawierać mniej cech. W praktyce, aby złagodzić założenie monotoniczności wprowadza się próg tolerancji, który umożliwia przeszukiwanie poniżej granicznego poziomu jakości modelu [60]. Innym przykładem wykładniczego przeszukiwania optymalnego zestawu predyktorów jest przeszukiwanie wiązkowe (*beam search*). W algorytmie tym wprowadza się pojęcie kolejki o określonej przez badacza długości *d*. W pierwszym kroku budowane są modele z jedną zmienną niezależną, które są sortowane względem przyjętego kryterium jakości. Dla *d* najlepszych zmiennych budowane są modele z dwiema zmiennymi niezależnymi uwzględniające wszystkie pozostałe zmienne. Po posortowaniu uzyskanych wyników wybieranych jest *d* najlepszych modeli. Proces jest powtarzany do momentu, gdy nie można dodać żadnej zmiennej poprawiającej model, lub gdy osiągnięto graniczny poziom jakości modelu. Przy zbiorach liczących więcej niż 30 zmiennych nie jest zalecana kolejka dłuższa niż 10 [Doak, 1992].

Metody sekwencyjne są szybkimi i relatywnie prostymi metodami odnajdywania suboptymalnego zestawu predyktorów. Najprostszą odmianą algorytmu sekwencyjnego jest wprowadzanie postępujące (*Sequential Forward Selection – SFS*). Algorytm rozpoczyna działanie od pustego zbioru danych a następnie sekwencyjne dodaje do zestawu zmienną, która w największym stopniu poprawia model składający się z dotychczas wybranych zmiennych. Algorytm ten jest równoważny przeszukiwaniu wiązkowemu o długości kolejki równej 1. Wadą tej strategii poszukiwania jest niemożność usunięcia zmiennych wprowadzonych we wcześniejszych iteracjach, które straciły na znaczeniu po wprowadzeniu do modelu innych predyktorów. Analogicznie działającym algorytmem jest metoda eliminacji wstecznej (*Sequential Backward Selection - SBS*). Metoda ta rozpoczyna działanie od modelu zawierającego wszystkie zmienne. W

[59] Założenie to jest w praktyce nieprawdziwe dla wielu metod, np. dla sieci neuronowych.

[60] Do puli zestawów kandydujących trafiają jednak jedynie zestawy, które nie przekroczyły bazowego poziomu jakości.

kolejnych krokach z modelu eliminowane są zmienne, które w najmniejszym stopniu pogarszają jego działanie. Metoda ta działa najlepiej, gdy optymalny zestaw zmiennych jest relatywnie duży w stosunku do dostępnego zbioru predyktorów. Wadą tego podejścia jest niemożność sprawdzenia użyteczności wcześniej usuniętych zmiennych po usunięciu w kolejnych krokach innych nieważnych predyktorów.

Metodami próbującymi złagodzić wady powyższych algorytmów wynikające z niemożności wykonywania dwustronnych korekt są metody krokowe (*step-wise*), które w przypadku metody krokowej postępującej po etapie wprowadzenia zmiennych testują możliwość usunięcia cech wprowadzonych do modelu we wcześniejszych krokach. W sposób analogiczny działa metoda krokowa wsteczna. Kolejnym uogólnieniem metod *SFS* oraz *SBS* jest metoda plus-L minus-R (*LRS*) w których L oraz *R* reprezentują liczby naturalnie i oznaczają odpowiednio liczbę zmiennych, jaka w danej iteracji powinna zostać wprowadzona (L) oraz usunięta (R) z modelu. Jeżeli L>R algorytm rozpoczyna działanie od pustego zbioru zmiennych a następnie dodaje do modelu L a następnie usuwa R zmiennych. W przeciwnym przypadku poszukiwanie optymalnego zestawu cech rozpoczyna się od zbudowania modelu ze wszystkimi predyktorami. Słabą stroną tej metody jest konieczność eksperymentalnego określenia wartości L oraz R. Algorytmy *Sequential Floating Selection* rozwiązują ten problem w sposób elastyczny określając wielkość kroku L oraz R. W przypadku metody postępującej *SFFS*, proces budowy modelu rozpoczyna się od wprowadzenia pierwszego predyktora a następnie w każdym kroku porównywane są ze sobą opcja wprowadzenia do modelu kolejnego „najlepszego" predyktora oraz usunięcia z modelu „najgorszego" predyktora. Wybierana jest opcja maksymalizująca przyjęte kryterium jakości modelu. Metoda wsteczna SFBS działa w sposób analogiczny rozpoczynając swoje działanie od pełnego modelu i usunięcia z niego jednej zmiennej.

Kolejnym rozszerzeniem przedstawionych metod jest wyszukiwanie oscylacyjne (*Oscillating Search, OS*). Punktem startu tej metody jest wynik uzyskany za pomocą wcześniej opisywanych metod – na przykład SFFS, jednak zakłada się przy tym, że model powinien posiadać k zmiennych. Uzyskany w ten sposób wynik traktowany jest jako dobry punkt wyjścia do dalszych poszukiwań. W kolejnych krokach wyszukiwane są rozwiązania posiadające nieznacznie więcej bądź mniej zmiennych. Zatem sprawdzane wyniki oscylują wokół zadanej wcześniej liczby k zmiennych. Motywacją do stosowania tego podejścia jest przeniesienie punktu ciężkości poszukiwań na rozwiązania bliskie oczekiwanym [Guyon i inni, 2008].

Przedstawione powyżej metody działały w sposób deterministyczny gwarantując powtarzalność wyników w wypadku powtórzenia eksperymentu dla tych samych hiperparametrów. Przedstawione poniżej algorytmy włączają element losowości do procedury przeszukiwania. Główną motywacją ich stosowania jest chęć uniknięcia lokalnych minimów, w jakich mogły zatrzymać się metody sekwencyjne. Do metod selekcji zmiennych należących do tej grupy należy zaliczyć algorytmy ewolucyjne:

- algorytm genetyczny (*genetic algorithm*),
- algorytm mrówkowy (*ant colony optimization*),
- optymalizacja rojem cząstek *(particle swarm optimisation)*,
- przeszukiwanie tabu *(tabu search)*.

Popularną metodą randomizowaną jest metoda symulowanego wyżarzania[61]. Algorytm rozpoczyna swoje działanie od losowo dobranego zestawu zmiennych. W każdej iteracji następuje losowa niewielka korekta zestawu zmiennych. Jeśli losowa zmiana będzie skutkować poprawą jakości modelu stanie się ona bieżącym rozwiązaniem. W przypadku, gdy losowa zmiana pogorszy wynik modelu będzie miała szansę z określonym prawdopodobieństwem zostać bieżącym rozwiązaniem w przeciwnym przypadku zostanie odrzucona. Akceptacja przez algorytm rozwiązań lokalnie pogarszających rozwiązanie ma na celu redukcję ryzyka wpadnięcia procedury w lokalne optimum i zwiększenie tym samym szansy na osiągnięcie globalnego optimum dla poszukiwanego rozwiązania.

3.1.3. Metody wbudowane

Wbudowane metody doboru zmiennych stanowią integralną cześć algorytmu służącego do budowy modeli. Z tego względu metody te są zazwyczaj bardziej efektywne obliczeniowo od metod opakowujących [Urbanowicz i inni, 2018]. Są także mniej podatne na nadmierne dopasowanie [Guyon i inni, 2008]. Strategie selekcji wbudowane w metody analityczne można podzielić na trzy grupy oparte na:

- regularyzacji,
- wprowadzaniu kolejnych zmiennych do modelu,
- eliminacji zmiennych z modelu.

[61] Metoda jest inspirowana procesem wyżarzania (sekwencyjnego podgrzewania i schładzania) stosowanym w metalurgii.

Pierwsza strategia wbudowanej procedury selekcji zmiennych stosowana jest zarówno w odniesieniu do metod liniowych na przykład regresji logistycznej [62] jak i bardziej zaawansowanych metod jak sieci neuronowe. Mechanizm, który umożliwia redukcję zmiennych podczas procesu budowy modelu, polega na dodaniu do błędu dopasowania minimalizowanego podczas estymacji, członu kary zależnego od wartości szacowanych parametrów.

$$Funkcja\ celu = błąd\ dopasowania + kara\ za\ wartości\ parametów\ modelu$$

Uzyskane w ten sposób rozwiązanie jest kompromisem pomiędzy dopasowaniem modelu a jego złożonością. Mechanizm polegający na dodaniu dodatkowego kryterium do optymalizowanego zadania nosi miano regularyzacji. Człon kary może być określony na różne sposoby. W praktyce, w przypadku regresji logistycznej wyróżnia się trzy podstawowe strategie:

- regresję LASSO (*Least Absolute Shrinkage and Selection Operator Regression*), w której kara zależy od sumy wartości bezwzględnych szacowanych parametrów;
- regresję grzbietową (*Ridge regression*), w której kara zależy od sumy kwadratów szacowanych parametrów podzielonej przez dwa,
- regresję *Elastic Net* będącą rozwiązaniem łączącym powyższe strategie, w różnych proporcjach wykorzystującym obydwa sposoby obliczania kary[63].

Regularyzacja zależna od sumy wartości bezwzględnych szacowanych parametrów nazywana jest regularyzacją L1. Podejście zależne od sumy kwadratów szacowanych parametrów nazywane jest regularyzacją L2.

W przypadku, gdy badacz oczekuje wyboru jedynie niewielkiego podzbioru cech spośród dostępnego zestawu metoda LASSO oraz *Elastic Net* przynoszą zazwyczaj lepsze wyniki od regresji grzbietowej ponieważ mają większą skłonność do redukcji wartości parametrów modelu do zera. W sytuacji, gdy w zbiorze danych zaobserwowano dodatkowo silne korelacje między predyktorami metoda *Elastic Net* powinna być preferowanym rozwiązaniem [Geron, 2017].

[62] Metoda opisana w dalszej części rozdziału.

[63] W przypadku wszystkich strategii w gestii badacza jest określenie hiperparametru α z przedziału [0,1] określającego stopień wkładu członu kary w optymalizowaną funkcję celu. W przypadku regresji Elastic Net dodatkowym hiperparametrem jest wartość r określająca proporcję pomiędzy obydwoma sposobami obliczania członu kary.

Regularyzacja jest również wykorzystana w procesie budowy modeli opartych na sieciach neuronowych czy metodzie SVM, nie prowadzi ona wtedy do wyłączenia z modelu zmiennych lecz do zmniejszenia liczby parametrów modelu oraz redukcji ich wartości co w znaczącym stopniu pozwala zredukować ryzyko nadmiernego dopasowania.

Najbardziej reprezentatywnym przykładem wprowadzania postępującego są algorytmy oparte na drzewach decyzyjnych na przykład CART, CHAID C4.5 czy drzewach wzmacnianych (*boosted trees*). Proces wyboru zmiennych polega w tych metodach na obliczaniu rankingu predyktorów przed każdą realizacją rekurencyjnego podziału zbioru obserwacji. Podział jest dokonywany każdorazowo na podstawie zwycięzcy danego rankingu. W wyniku tego procesu model tworzą jedynie zwycięzcy poszczególnych lokalnych rankingów.

Strategia eliminacji wstecznej może być realizowana w przypadku metody SVM czy sieci neuronowych. Jest ona podobna do metod opakowujących (na przykład eliminacji wstecznej) lecz mniej kosztowna obliczeniowo oraz mniej podatna na nadmierne dopasowanie [Guyon i inni, 2008] Polega ona na zbudowaniu modelu na podstawie wszystkich dostępnych zmiennych, gdzie następnie dla każdej ze zmiennych użytych w modelu przeprowadza się analizę wrażliwości polegającą na porównaniu błędu uzyskanego modelu z błędem otrzymanym w sytuacji eliminacji danej zmiennej. W przeciwieństwie do metod opakowujących, eliminację zmiennej przeprowadza się poprzez wprowadzenie do oszacowanego modelu zbioru danych, w którym wyłączana zmienna jest pozbawiana zmienności poprzez zamianę wartości na średnią lub modalną. Zmienna, która najbardziej pogarsza model, bądź najmniej go poprawia jest eliminowana z zestawu predyktorów. Po wyeliminowaniu najgorszego predyktora proces budowy modelu jest ponawiany dla ograniczonego zestawu zmiennych.

Przedstawione powyżej strategie mogą być inspiracją do konstruowania bardziej skomplikowanych hybrydowych strategii łączących wiele podejść podstawowych. Najczęściej spotyka się połączenie metod filtrowania z metodami opakowującymi bądź metod filtrowania z metodami wbudowanymi.

Przykładem hybrydowego podejścia łączącego filtry z metodami wbudowanymi jest algorytm CFSH (*Correlation-based Feature Selection with the Hellwig heuristic*) zaprezentowany przez E. Gatnara [2008]. Celem algorytmu jest zbudowanie zespołu drzew decyzyjnych. Dla każdego ze składowych drzew losowany jest podzbiór predyktorów równy połowie dostępnego zestawu zmiennych. W kolejnym kroku liczba zmiennych jest ograniczana za pomocą metody filtrowania opartej na integralnej pojemności informacji

(z poprawką M.Walesiaka). Wstępnie określony podzbiór zmiennych jest następnie podstawą do wyboru zmiennych wbudowanego w algorytm drzewa klasyfikacyjnego.

3.2. Wybór techniki modelowania

Wybór techniki modelowania jest kolejnym bardzo ważnym etapem cyklu życia modelu. Badacz ma do dyspozycji szereg technik i strategii o różnych własnościach, wykazujących szereg charakterystycznych dla danej metody zalet oraz braków. Spośród kilku obecnych w literaturze podziałów oraz klasyfikacji dostępnych metod wydaje się, że należy wymienić podział na metody działające na zasadzie:

- „białej skrzynki", dostarczającej badaczowi nie tylko informację o ocenie ryzyka odejścia klienta, ale również pozwalającej na ocenę przyczyn wpływających na to ryzyko;
- „czarnej skrzynki", która umożliwia ocenę ryzyka odejścia klienta bez informacji o zależnościach oraz wzorcach zawartych w danych.

Metody działające na zasadzie „białej skrzynki" mają zazwyczaj mniejszą siłę predykcyjną od metod „czarnoskrzynkowych". Są one jednak ciągle bardzo popularne, ponieważ sama siła predykcyjna nie jest jedynym wymiarem oceny jakości zbudowanego modelu. Bardzo ważna z praktycznego punktu widzenia jest przejrzystość oraz możliwość łatwej interpretacji zbudowanego modelu. Nie bez znaczenia jest także łatwość wdrożenia modelu w środowisku informatycznym oraz aspekt psychologiczny, wiążący się z zaufaniem do metod automatycznej detekcji oraz ich wyników. We wszystkich tych dodatkowych wymiarach przewagę mają metody działające na zasadzie „białej skrzynki". Do metod tych zaliczyć można regresję logistyczną oraz drzewa klasyfikacyjne (regresyjne) opisane w dalszej części rozdziału. Metody działające na zasadzie „czarnej skrzynki" są zalecane w sytuacji, gdy głównym celem modelu jest predykcja. Sprawdzają się lepiej w sytuacji występowania w danych słabszych, nieliniowych zależności. Do metod tych zaliczyć można na przykład sieci neuronowe, metodę wektorów nośnych czy metodę najbliższych sąsiadów.

Drugim fundamentalnym podziałem pozwalającym sklasyfikować dostępne metody modelowania jest podział wynikający ze strategii łączenia modeli w zespoły. Punktem wyjścia jest pojedynczy model zbudowany na przykład za pomocą jednej z wymienionych

metod. Pojedyncze modele mogą być traktowane jako budulec do budowy bardziej złożonych zespołów. Na podstawie tego wymiaru możemy zatem wyróżnić:

- modele proste,
- zespoły modeli złożone z tych samych metod,
- zespoły modeli łączące wiele metod – modele hybrydowe.

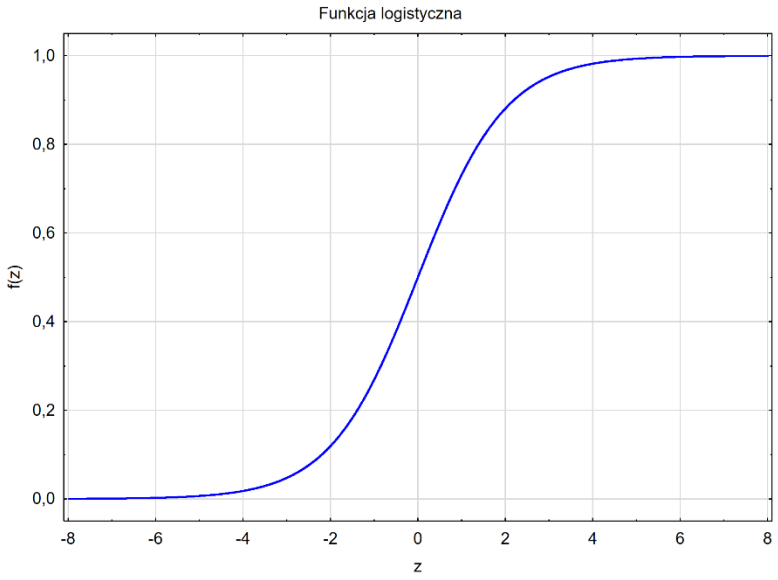
Niektóre strategie łączenia modeli w zespoły zyskały odrębną nazwę i są traktowane jako odrębne metody analityczne. Do tej grupy należy zaliczyć metodę drzew wzmacnianych (*boosted trees*) oraz losowego lasu (*random forest*), które jako swój podstawowy budulec wykorzystują drzewa klasyfikacyjne bądź regresyjne, różnią się natomiast sposobem ich agregacji. Metody te zostaną przybliżone w dalszej części podrozdziału. Zespoły modeli mogą być budowane na podstawie innych metod na przykład sieci neuronowych czy metody wektorów nośnych. Podejście hybrydowe, wykorzystujące w końcowym rozwiązaniu modele zbudowane za pomocą różnych metod będzie przybliżone w osobnym podrozdziale.

3.2.1. Regresja logistyczna

Regresja logistyczna jest bardzo popularną metodą modelowania wykorzystywaną powszechnie zarówno na polu naukowym jak i w biznesie. Jej popularność wynika z wielu czynników. Do tych najbardziej istotnych niewątpliwie można zaliczyć możliwość interpretacji ocen parametrów uzyskanego modelu. Kolejnym czynnikiem jest relatywnie niewielka liczba założeń. Model regresji logistycznej, w przeciwieństwie do regresji wielorakiej nie zakłada normalności oraz homoskedastyczności reszt. W porównaniu do analizy dyskryminacyjnej nie ma tutaj wymagania wielowymiarowego rozkładu normalnego predyktorów. Model logistyczny bazuje na funkcji logistycznej o postaci:

$$f(z) = \frac{exp\ (z)}{1 + exp\ (z)}$$

Przebieg funkcji logistycznej przedstawia Rysunek 13.



Rysunek 13 Wykres funkcji logistycznej
Źródło: opracowanie własne.

Można zauważyć, że wartości funkcji logistycznej przyjmują wartości od 0 do 1, co umożliwia interpretację jej wartości jako prawdopodobieństwo modelowanego zdarzenia. Jej kształt przypomina literę S, dzięki czemu funkcja pozwala modelować zjawiska, które charakteryzują się zmianą natężenia szacowanego prawdopodobieństwa po osiągnięciu pewnej wartości progowej. Początkowo zmiany wartości funkcji są minimalne i oscylują blisko 0, po osiągnięciu wartości progowej gwałtownie wzrastają do 1.

Przed rozpoczęciem budowy modelu regresji logistycznej klasy zmiennej zależnej są przekodowywane odpowiednio do wartości 1 dla modelowanej klasy oraz 0 dla drugiej klasy. W przypadku modeli retencji klientów modelowana klasa najczęściej odnosi się do klientów nielojalnych. Regresja logistyczna może być przedstawiona za pomocą wyrażenia [Vittinghoff i inni, 2012]:

$$P(x_1, x_2, \dots, x_p) = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}$$

gdzie $P(x_1, x_2, \dots, x_p)$ oznacza warunkowe prawdopodobieństwo, że analizowany przypadek należy do modelowanej klasy zmiennej zależnej, $\beta_0, \beta_1, \dots, \beta_p$ są ocenami parametrów regresji, $x_1, x_2, \dots, x_p$ jest wektorem cech opisujących konkretnego klienta.

Oceny parametrów regresji są optymalizowane podczas budowy modelu na podstawie zbioru uczącego za pomocą metody największej wiarygodności. W surowej postaci nie mają one merytorycznej interpretacji. Przekształcenie ich za pomocą wyrażenia exp(β)

pozwala interpretować siłę efektu związaną z jednostkową zmianą danego predyktora. Uzyskaną w wyniku przekształcenia miarą efektu jest iloraz szans (*odds ratio*, *OR*), informujący o zmianie szansy (*odds*) [64] wystąpienia modelowanego zjawiska przy jednostkowej zmianie wartości predyktora.

Przyjętą praktyką budowy regresji logistycznej jest uwzględnianie w modelu jedynie zmiennych, dla których wartości ocen parametrów są istotnie różne od zera. Ocenę istotności przeprowadza się za pomocą testu Walda [65]. Statystyka Walda wyraża się następującym wzorem [Hosmer i inni, 2013]:

$$W = \frac{\beta}{SE(\beta)}$$

gdzie β oznacza ocenę parametru regresji uzyskaną w wyniku estymacji metodą największej wiarygodności, $SE(\beta)$ oznacza błąd standardowy oceny. Wyrażenie W^2 podlega rozkładowi chi-kwadrat z jednym stopniem swobody i jest podstawą do przeprowadzenia statystycznego testu istotności. Innym popularnym testem wykorzystywanym do diagnostyki modelu jest test LR (*Likelihood Ratio*). Statystyka LR opiera się na porównaniu ilorazów maksymalnych wartości funkcji największej wiarygodności otrzymanych podczas estymacji dwóch różnych modeli, modelu bieżącego oraz referencyjnego. Model bieżący zawiera wszystkie zmienne modelu referencyjnego oraz jedną bądź więcej nowych zmiennych niezależnych. Test sprawdza czy model bieżący jest istotnie lepszy od modelu referencyjnego a tym samym czy dodanie do modelu referencyjnego nowych zmiennych w istotny sposób go poprawia. Statystykę LR można przedstawić za pomocą następującego wzoru [Hosmer i inni, 2013]:

$$LR = -2ln\left(\frac{L\ Ref}{L\ Biez}\right) = -2[ln(L\ Ref) - ln\ (L\ Biez)]$$

Statystyka LR podlega rozkładowi chi-kwadrat z liczbą stopni swobody równą różnicy w liczbie parametrów porównywanych modeli.

Istotnym założeniem regresji logistycznej mającym praktyczne przełożenie na jakość modelu jest liniowość predyktorów na skali logarytmu szansy. Założenie to wynika wprost ze wzoru na regresję logistyczną, który po wykonaniu przekształceń można przedstawić w postaci liniowej względem parametrów [Vittinghoff i inni, 2012]:

[64] Szansa definiowana jest jako iloraz prawdopodobieństwa zajścia modelowanego zdarzenia i prawdopodobieństwa zajścia zdarzenia przeciwnego. Szczegóły patrz [Hosmer i inni, 2013].

[65] Alternatywnym, rzadziej stosowanym testem jest test wartości punktowej (*Score test*).

$$Ln[\frac{P(x_1, x_2, \dots, x_p)}{1 - P(x_1, x_2, \dots, x_p)}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Wyrażenie po lewej stronie równania będące logarytmem naturalnym szansy modelowanego zjawiska określa się mianem logitu. Linowość można testować za pomocą testu LR dodając do modelu z danym predyktorem zmienną pochodną będącą przekształceniem predyktora za pomocą funkcji kwadratowej bądź innej nieliniowej [Vittnghoff i inni, 2012]. Bardziej zasadna jest jednak wizualna ocena profilu danego predyktora na wykresie przedstawiającym wartość logarytmu szansy rezygnacji klienta obliczoną dla przedziałów analizowanego predyktora skategoryzowanego za pomocą podziału na percentyle. W przypadku stwierdzenia nieliniowego wpływu predyktora na modelowane zjawisko zaleca się jego dyskretyzację.

Regresja logistyczna jest metodą budowy modelu wykorzystywaną do budowy kart skoringowych (*scorecard*). Karta skoringowa jest popularnym formatem modelu klasyfikacyjnego, który umożliwia intuicyjną interpretację oraz łatwe stosowanie modelu przez osoby niezwiązane z analizą danych. Metodykę budowy kart skoringowych przedstawiono między innymi w [Siddiqui, 2017, Baesens, 2014, Migut i inni, 2013]. Proces budowy modelu opiera się na koncepcji profilu ryzyka, który zakłada możliwość eksperckiej ingerencji badacza w proces budowy modelu. W wyniku analizy ma powstać model z jednej strony poprawny pod względem statystycznym oraz o zadowalającej sile predykcyjnej, z drugiej strony musi być on zrozumiały dla biznesu. Elementami wyróżniającymi metodykę budowy karty skoringowej od innych strategii budowy modeli uczenia maszynowego są:

- segmentacja klientów przy użyciu wiedzy eksperckiej oraz drzew klasyfikacyjnych, w wyniku której modele regresji budowane są na jednorodnych podsegmentach,
- dyskretyzacja zmiennych ilościowych na jednorodne pod względem ryzyka odejścia klasy,
- łączenie kategorii predyktorów jakościowych podobnych pod względem ryzyka,
- skalowanie ocen parametrów regresji, aby każdej klasie predyktora możliwe było przypisanie odpowiedniej liczby punktów.

Problem niejednorodności zbioru danych jest zatem łagodzony przez budowę modelu hybrydowego łączącego drzewo klasyfikacyjne lub regresyjne z modelem regresji.

Dyskretyzacja zmiennych umożliwia redukcję problemów z brakami danych (tworzą osobną kategorię), wartościami odstającymi oraz przybliżonym powyżej problemem nieliniowości. Powtórna kategoryzacja zmiennych jakościowych pozwala na eliminację mało licznych klas redukując ryzyko nadmiernego dopasowania modelu. Model regresji logistycznej budowany jest na zmiennych wystandaryzowanych do wartości WoE, bądź przekodowanych za pomocą kodowania z sigma-ograniczeniami[66]. Po zbudowaniu modelu następuje przeskalowanie jego ocen parametrów umożliwiające wygenerowanie oceny punktowej. Konieczne do tego parametry skalujące; mnożnik (*factor*) oraz przesunięcie (*offset*) oblicza się po eksperckim określeniu parametrów umożliwiających przyszłą interpretację uzyskanej punktacji:

- punktów podwajających szansę (*points to double the odds, pdo*), określających co ile punktów szansa bycia lojalnym klientem będzie się podwajać[67],
- poziomu odniesienia pozwalającego na określenie i interpretację jednego poziomu punktacji, pozostałe oblicza się stosując parametr *pdo*.

Poziom odniesienia jest określany za pomocą szansy (*odds*) dla wskazanej liczby punktów (*score*). N. Siddiqui [2006] sugeruje określenie wartości parametru *pdo* na poziomie 20, co oznaczać będzie podwojenie szansy bycia dobrym klientem co 20 punktów. Poziom odniesienia jest sugerowany jako szansa 50 do 1 dla 600 punktów. Po eksperckim określeniu parametrów *pdo*, *odds* oraz *score*, parametry skalujące oblicza się za pomocą poniższych wzorów [Siddiqui, 2006]:

$$factor = \frac{pdo}{ln\,(2)}$$

$$offset = score - [factor \times ln\,(odds)]$$

Obliczone parametry skali pozwalają na obliczenie punktacji na podstawie ocen parametrów regresji. Wzory różnią się w zależności od sposobu kodowanie predyktorów. Dla kodowania typu WoE, punktację dla jednej kategorii danego predyktora oblicza się za pomocą wzoru [StatSoft, 2013]:

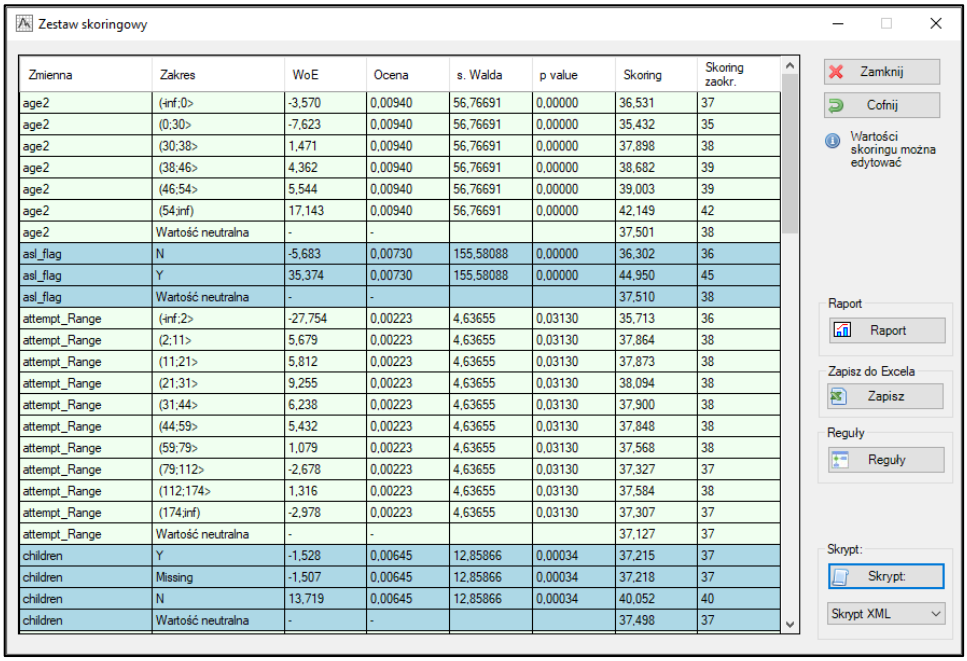
$$score = \left(\beta \times WoE + \frac{\alpha}{m}\right) \times factor + \frac{offset}{m}$$

[66] Standaryzację WoE oraz kodowanie z sigma-ograniczeniami przedstawiono w rozdziale 2.
[67] Wraz z liniowym wzrostem punktacji obserwowany będzie zatem wykładniczy wzrost szansy.

natomiast w przypadku kodowania z sigma-ograniczeniami punktację oblicza się jako [StatSoft, 2013]:

$$score = \left(\beta + \frac{\alpha}{m}\right) \times factor + \frac{offset}{m}$$

gdzie β oznacza ocenę parametru regresji dla danego predyktora, α jest wyrazem wolnym natomiast m oznacza liczbę predyktorów w modelu. Uzyskane wartości punktacji są następnie zaokrąglane do wartości całkowitych. W wyniku przekształceń model regresji logistycznej może być przedstawiony w postaci karty skoringowej. Przykładowy fragment karty skoringowej wygenerowany w programie Statistica Zestaw Skoringowy prezentuje Rysunek 14.



Zmienna	Zakres	WoE	Ocena	s. Walda	p value	Skoring	Skoring zaokr.
age2	(-inf;0>	-3,570	0,00940	56,76691	0,00000	36,531	37
age2	(0;30>	-7,623	0,00940	56,76691	0,00000	35,432	35
age2	(30;38>	1,471	0,00940	56,76691	0,00000	37,898	38
age2	(38;46>	4,362	0,00940	56,76691	0,00000	38,682	39
age2	(46;54>	5,544	0,00940	56,76691	0,00000	39,003	39
age2	(54;inf)	17,143	0,00940	56,76691	0,00000	42,149	42
age2	Wartość neutralna	-	-			37,501	38
asl_flag	N	-5,683	0,00730	155,58088	0,00000	36,302	36
asl_flag	Y	35,374	0,00730	155,58088	0,00000	44,950	45
asl_flag	Wartość neutralna	-	-			37,510	38
attempt_Range	(-inf;2>	-27,754	0,00223	4,63655	0,03130	35,713	36
attempt_Range	(2;11>	5,679	0,00223	4,63655	0,03130	37,864	38
attempt_Range	(11;21>	5,812	0,00223	4,63655	0,03130	37,873	38
attempt_Range	(21;31>	9,255	0,00223	4,63655	0,03130	38,094	38
attempt_Range	(31;44>	6,238	0,00223	4,63655	0,03130	37,900	38
attempt_Range	(44;59>	5,432	0,00223	4,63655	0,03130	37,848	38
attempt_Range	(59;79>	1,079	0,00223	4,63655	0,03130	37,568	38
attempt_Range	(79;112>	-2,678	0,00223	4,63655	0,03130	37,327	37
attempt_Range	(112;174>	1,316	0,00223	4,63655	0,03130	37,584	38
attempt_Range	(174;inf)	-2,978	0,00223	4,63655	0,03130	37,307	37
attempt_Range	Wartość neutralna	-	-			37,127	37
children	Y	-1,528	0,00645	12,85866	0,00034	37,215	37
children	Missing	-1,507	0,00645	12,85866	0,00034	37,218	37
children	N	13,719	0,00645	12,85866	0,00034	40,052	40
children	Wartość neutralna	-	-			37,498	37

Rysunek 14 Karta skoringowa
Źródło: Opracowanie własne.

Każdemu atrybutowi, czyli każdej kategorii zmiennej jakościowej oraz każdemu przedziałowi zmiennej ilościowej przypisana jest liczba punktów. Ocenę skłonności klientów do rezygnacji określa się poprzez zsumowanie punktów odnoszących się do atrybutów danego klienta. Im większa liczba punktów tym mniejsza skłonność klienta do rezygnacji z usługi. Ważną cechą karty skoringowej jest możliwość oceny poziomu lojalności klientów nawet w przypadku braku danych dla wybranych predyktorów

zawartych w modelu. W takiej sytuacji ocena zastępowana jest punktacją neutralną (*neutral score*) obliczaną jako [StatSoft, 2013]:

$$neutral\ score = \sum_{i=1}^{k} score_i \times distr_i$$

gdzie k określa liczbę kategorii danego predyktora, $score_i$ oznacza punktację dla i-tej kategorii, $distr_i$ oznacza frakcję przypadków należących do i-tej kategorii.

Zbudowany model regresji logistycznej niezależnie od jego ostatecznej postaci może być oceniany za pomocą standardowego zestawu miar jakości modelu opisanych w rozdziale 4. Poza ogólnymi miarami siły predykcyjnej regresja logistyczna umożliwia wykonanie pogłębionej diagnostyki modelu za pomocą dodatkowych, specyficznych miar oraz testów.

Do jednego z najbardziej popularnych testów oceniających dobroć dopasowania modelu regresji logistycznej należy zaliczyć test Hosmera-Lemeshowa (HL). Ogólną ideą testu jest sprawdzenie zgodności oszacowanych prawdopodobieństw z wartościami empirycznymi. Korzysta z faktu, że suma oszacowanych prawdopodobieństw przynależności do modelowanej klasy równa jest liczbie przypadków należących do modelowanej klasy. Test sprawdza czy zgodność ta zachodzi również w podgrupach ryzyka (najczęściej w decylach ryzyka). Po posortowaniu przypadków względem prawdopodobieństwa przynależności do modelowanej klasy oceniana jest zgodność oczekiwanych częstości wyników w tych grupach z ich empirycznymi odpowiednikami. Następnie oblicza się statystykę $\hat{C}$ na podstawie wzoru [Hosmer i inni, 2013]:

$$\hat{C} = \sum_{k=1}^{g} \frac{o_k - n_k\overline{\pi}_k}{n_k\overline{\pi}_k(1-\overline{\pi}_k)}$$

gdzie g jest liczbą grup, na który podzielony został zbiór uczący,o_k jest liczbą przypadków należących do modelowanej klasy zawartych w k-tej grupie, n_k jest liczbą przypadków zawartych w k-tej grupie, $\overline{\pi}_k$ jest średnim poziomem prawdopodobieństwa przynależności do modelowanej klasy przypadków należących do k-tej grupy.

Przyjmuje się, że statystyka $\hat{C}$ podlega rozkładowi chi-kwadrat z liczbą stopni swobody g-2[68]. Hipoteza zerowa H_0 zakłada, że oszacowane i obserwowane liczności są zgodne. Tak

[68] W pozycji [Hosmer i inni, 2013] zawarto obszerną dyskusję na temat modyfikacji testu oraz przybliżeń wartości jego stopni swobody.

więc statystycznie istotny wynik (tj. odrzucenie H_0) wskazuje brak dopasowania. Brak podstaw do odrzucenia H_0 wyklucza rażący brak dopasowania. Pomimo swojej popularności test HL ma szereg ograniczeń. Jedną z nich jest zależność od liczby wyszczególnionych grup, a także od rozkładu wartości predyktorów w tych grupach [Vittognhoff i inni, 2013]. Może być bardzo wrażliwy na dość małe rozbieżności dopasowania w dużych próbkach, dlatego statystycznie istotny wynik może w takich przypadkach nie sygnalizować poważnego problemu z dopasowaniem. Podobnie brak znalezienia statystycznie istotnego wyniku niekoniecznie oznacza, że model dobrze pasuje do danych. Ten test jest najbardziej przydatny jako bardzo podstawowy sposób badania problemów z dopasowaniem i nie powinien być traktowany jako ostateczna diagnoza dobrego dopasowania.

Maksymalne wartości funkcji największej wiarygodności uzyskane w wyniku estymacji będące podstawą przedstawianego testu LR są także podstawowym składnikiem używanym do konstrukcji tak zwanych współczynników pseudo R^2, służących do oceny dobroci dopasowania zbudowanego modelu logistycznego. Współczynniki te są próbą odzwierciedlenia współczynnika determinacji wykorzystywanego w modelu regresji wielorakiej, informującego jaka część zmienności zmiennej zależnej jest wyjaśniona za pomocą zbudowanego modelu. Przykładem miary pseudo R^2 jest miara Coxa-Snella, którą można obliczyć za pomocą wzoru [Stanisz, 2016]:

$$R_{CS}^2 = 1 - (\frac{L_0}{L_m})^{2/n}$$

gdzie n oznacza liczbę przypadków uczących, L_0, L_m odnoszą się do wartości funkcji największej wiarygodności odpowiednio dla modelu z wyrazem wolnym oraz modelu analizowanego. Maksymalna wartość współczynnika Coxa-Snella jest zawsze mniejsza od 1, zatem, aby zapewnić temu wskaźnikowi pożądaną własność stosuje się jego modyfikację zaproponowaną przez Nagelkerke'a przyjmującą postać [Stanisz, 2016]:

$$\overline{R}^2 = \frac{R_{CS}^2}{1 - L_0^{2/n}}$$

Współczynnik ten przyjmuje wówczas wartości z przedziału [0, 1]. Poza wymienionymi miarami w literaturze spotkać można szereg analogicznych współczynników jak na przykład R^2 McFaddena, R^2 Cragga-Uhlera, R^2 McKelveya i Zavoina czy R^2 Efrona [Stanisz, 2016].

Kolejną miarą oceniającą dobroć dopasowania modelu regresji logistycznej jest statystyka odchylenia D (*deviance*) porównująca wartość funkcji największej wiarygodności zbudowanego modelu z analogiczną funkcją modelu nasyconego [69] (*saturated*). Statystykę D można przedstawić za pomocą wzoru [Hosmer i inni, 2013]:

$$D = -2\,ln\left(\frac{L_m}{L_s}\right) = -2ln\,(L_m)$$

gdzie L_m odnosi się do wartości funkcji największej wiarygodności ocenianego modelu, L_s odnosi się do wartości funkcji największej wiarygodności modelu nasyconego, która wynosi 1 [Hosmer i inni, 2013]. Ocena dobroci dopasowania na podstawie statystyki D polega na porównaniu jej wartości z liczbą stopni swobody. Jeżeli iloraz jej wartości przez odpowiadającą jej liczbą stopni swobody jest bliski 1, można wnioskować o dobrym dopasowaniu modelu [Stanisz, 2016].

Niewątpliwą zaletą regresji logistycznej jest jej prostota oraz możliwość interpretacji parametrów zbudowanego modelu. Metoda nie posiada hiperparametrów które mogłyby być optymalizowane w trakcie uczenia. Ponieważ uwzględnia jedynie efekty główne, które powinny dodatkowo mieć liniowy wpływ na logarytm szansy zajścia modelowanego zjawiska, bardzo duży wpływ na końcową jakość modelu mają etapy związane z przygotowaniem danych pochodnych. Podczas poszukiwania optymalnego podzbioru zmiennych w modelu najczęściej wykorzystywane są metody opakowujące (*wrapper*) opisane wcześniej w tym rozdziale.

3.2.2. Sieci neuronowe

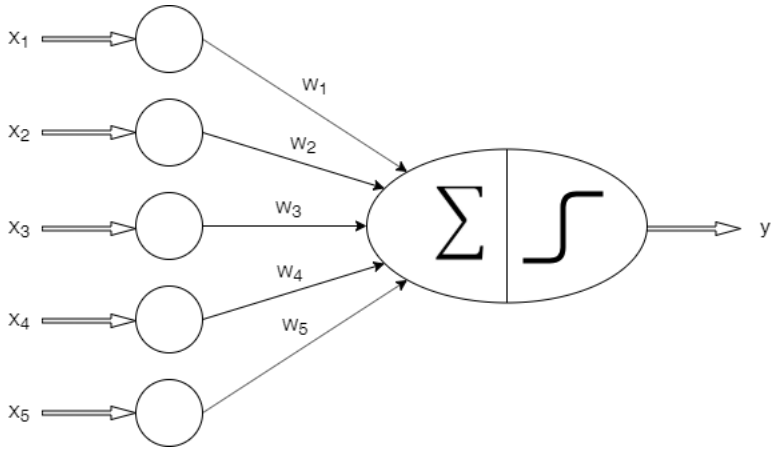
Sieci neuronowe to niejednorodna grupa metod uczenia maszynowego wykorzystująca podczas uczenia zarówno strategie uczenia z nauczycielem jak i strategie uczenia bez nauczyciela. Do sieci neuronowych zaliczyć można perceptron wielowarstwowy (*multilayer perceptron*), sieć o radialnych funkcjach bazowych (RBF), sieć Kohonena (*self organizing maps*), sieć splotową (*convolutional neural network*), sieci rekurencyjne, sieć Hopfielda czy maszyny Boltzmanna. Elementem wspólnym wszystkich metod zaliczanych do sieci neuronowych jest konieczność stworzenia przez badacza, na drodze eksperymentu, struktury powiązanych ze sobą jednostek obliczeniowych zwanych sztucznymi neuronami. Niezależnie od rodzaju sieci wspólnym elementem każdego sztucznego neuronu jest:

[69] Model nasycony to model idealnie dopasowany do danych.

– agregacja sygnałów wprowadzanych do neuronu,

– przekształcenie zagregowanego sygnału i przekazanie go na wyjściu.

Powyższe działania można w sposób schematyczny przedstawić za pomocą Rysunek 15.



Rysunek 15 Schemat sztucznego neuronu
Źródło: opracowanie własne na podstawie [Aggarwal, 2018].

Na wejściu neuronu prezentowany jest wektor wartości X=[x_1 … x_d] z wartościami cech opisujących analizowany obiekt[70]. Surowe wartości najczęściej są standaryzowane, a następnie wprowadzane do neuronu. Przed wprowadzeniem do neuronu są one kojarzone z wektorem wag W=[w_1 … w_2], które pełnią w rolę parametrów optymalizowanych w trakcie nauki sieci. Po wprowadzeniu do neuronu wektora wartości wejściowych oraz wektora wag są one agregowane. Sztuczne neurony używane w poszczególnych rodzajach sieci mogą różnić się od siebie sposobem agregowania sygnału. Popularnym sposobem jest agregacja liniowa [Aggarwal, 2018]:

$$X \times W = \sum_{i=1}^{d} x_i w_i$$

Po zagregowaniu wartości sygnał jest przekształcany za pomocą funkcji aktywacji. Jeśli funkcją tą byłby funkcja logistyczna, wtedy pojedynczy neuron odpowiadałby działaniu regresji logistycznej opisywanej we wcześniejszym podrozdziale.

[70] Na przykład cech opisujących klienta.

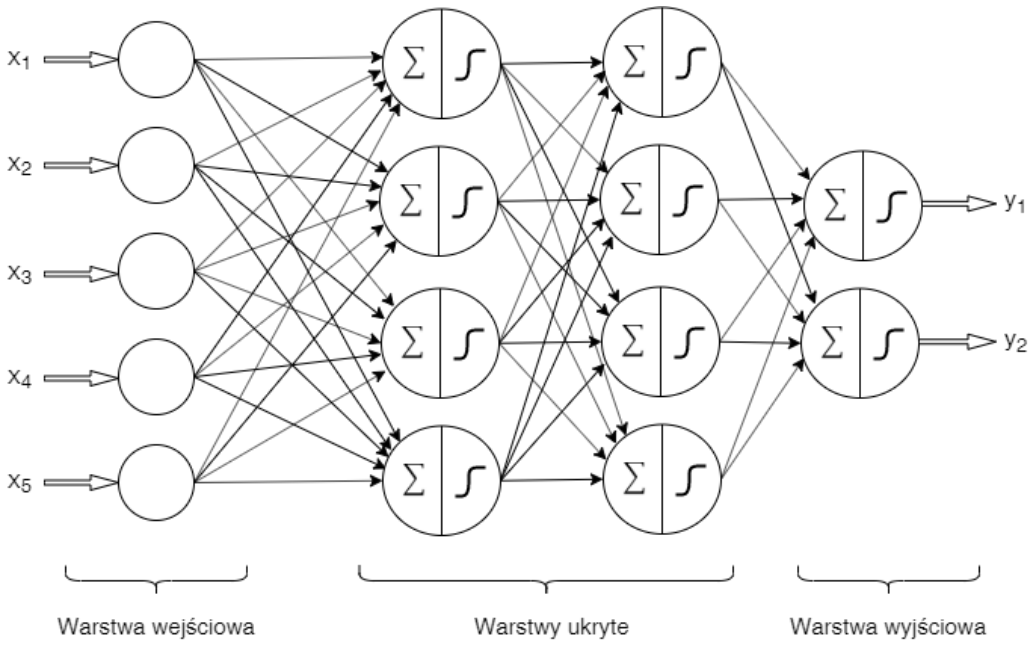
Odpowiednikiem ocen parametrów regresji byłyby wagi neuronu[71]. Oczywiście poza wspomnianym sposobem agregacji oraz funkcją logistyczną w roli funkcji aktywacji dostępnych jest szereg innych metod obliczania sygnału wyjściowego. Nie zmienia to jednak faktu, że pojedynczy neuron jest jednostką wykonującą niezbyt skomplikowane działania matematyczne. Zdolność do dopasowania się przez sieci neuronowe do dowolnego problemu [Aggarwal, 2018] wynika zatem nie ze zdolności pojedynczego neuronu, a z faktu, że na ich podstawie tworzone są bardziej skomplikowane, wielowarstwowe struktury – sieci neuronowe.

W zagadnieniach związanych z budową modeli retencji klienta najczęściej używanym rodzajem sieci neuronowej jest perceptron wielowarstwowy. Konkurencyjnym do niego podejściem może być użycie sieci o radialnych funkcjach bazowych. W najnowszych opracowaniach na przykład [Spanoudes, 2018, Mishra, Reddy, 2017] można znaleźć również próby wykorzystania sieci splotowych, najczęściej wykorzystywanych do klasyfikacji obrazów[72].

Perceptron wielowarstwowy jest siecią jednokierunkową. Nie występują w niej sprzężenia zwrotne. Perceptron składa się z jednej warstwy wejściowej, której rolą jest wprowadzanie wartości wejściowych do kolejnej warstwy sieci. Perceptron posiada jedną, bądź większą liczbę warstw ukrytych oraz jedną warstwę wyjściową, prezentującą wyniki działania sieci. Schemat perceptronu przedstawia Rysunek 16.

[71] Wyraz wolny modelu zostałby odtworzony poprzez dodanie kolejnego wejścia neuronu, do którego wprowadzana byłaby zawsze wartość 1.

[72] Opracowania te przygotowują dane związane z aktywnością każdego klienta w postaci obrazu, w którym poszczególne piksele odpowiadają danemu rodzajowi aktywności klienta w konkretnym okresie, a ich kolor dopowiada jego natężeniu.



Rysunek 16 Schemat sieci neuronowej
Źródło: opracowanie własne na podstawie [Aggarwal, 2018].

Jak już wspomniano, neurony w warstwie wejściowej standaryzują wprowadzane sygnały i przekazują je do kolejnej warstwy. Ich liczba jest zależna od liczby i skali pomiarowej (poziomu pomiaru) predyktorów zawartych w zbiorze danych. Z każdym połączeniem pomiędzy warstwami skojarzona jest (nieprzedstawiona na rysunku) waga. Wagi pełnią rolę parametrów modelu i są optymalizowane w trakcie uczenia sieci. Funkcje aktywacji są stałe w ramach jednaj warstwy, ale mogą różnić się między warstwami. Liczba warstw ukrytych oraz liczba neuronów w tych warstwach zależy od decyzji badacza. Warstwa wyjściowa prezentuje wyniki obliczeń. Sieć posiadająca dwa neurony w warstwie wyjściowej pozwala na budowę modelu klasyfikacji binarnej.

Perceptron wielowarstwowy jest metodą pozwalająca zbudować skuteczny model klasyfikacyjny. Wymaga jednak od badacza pewnego doświadczenia wynikającego z dużej liczby hiperparametrów wpływających na sposób działania sieci, a określenie których leży w jego gestii. Do najważniejszych hiperparametrów należy zaliczyć:

- liczbę warstw ukrytych sieci neuronowej,
- liczbę neuronów w danej warstwie ukrytej,
- rodzaj funkcji aktywacji w poszczególnych warstwach,
- sposób inicjalizacji wag,
- wybór algorytmu uczenia,

- określenie technik regularyzacji,
- liczba epok (cykli) podczas których budowana jest sieć,
- wybór funkcji straty.

Dla wielu problemów klasyfikacyjnych wystarczająca jest budowa sieci neuronowej posiadającej jedną warstwę ukrytą. Za pomocą sieci neuronowej z jedną warstwą ukrytą można skutecznie modelować nawet najbardziej skomplikowaną funkcję, o ile sieć zawierała będzie dostateczną liczbę neuronów. Okazuje się, że tę samą funkcję można aproksymować zwiększając liczbę warstw i redukując jednocześnie liczbę neuronów [Aggarwal, 2018]. W praktyce sieci o większej liczbie warstw wykazują się zatem większą efektywnością swoich parametrów. Mogą modelować skomplikowane zależności za pomocą wykładniczo mniejszej liczby neuronów niż jej płytsze odpowiedniki. W praktyce, liczba warstw ukrytych jest wyznaczana (podobnie jak inne hiperparametry) metodą prób i błędów. Zaleca się rozpoczynanie budowy modelu od jednej, bądź dwóch warstw ukrytych, a następnie stopniowe zwiększanie ich liczby do momentu, w którym zostanie zaobserwowane nadmierne dopasowanie budowanego modelu [Geron, 2017].

Liczba neuronów w sieci neuronowej jest zdeterminowana kilkoma czynnikami. Po pierwsze liczba neuronów w warstwie wejściowej wynika wprost z liczby zmiennych, jakie mają zostać uwzględnione w modelu. Każdemu predyktorowi ilościowemu odpowiada jeden neuron w warstwie wejściowej. Predyktory jakościowe wymagają zmiany swojej reprezentacji. Najczęściej wykonywane jest przekodowanie do zmiennych zero-jedynkowych (*dummy variables*), każdej klasie zmiennej jakościowej odpowiadał będzie jeden neuron w warstwie wejściowej (kodowanie typu *n*). Liczba neuronów w warstwie wyjściowej jest zdeterminowana przez charakter modelowanego zjawiska. W przypadku klasyfikacji binarnej najczęściej będą to dwa neurony przypisane do każdej z modelowanych klas.

Zbyt prosta topologia sieci prowadzić będzie do budowy niedostatecznie dopasowanego modelu. Zbyt rozbudowana struktura będzie z kolei skutkować skłonnością sieci do nadmiernego dopasowania. Znalezienie optymalnej topologii bywa zazwyczaj czasochłonne i wymaga przeprowadzenie wielu prób. Pomocna może być tutaj automatyzacja działań pozwalająca na przeszukaniu wielu możliwych rozwiązań i wybór najlepszego z nich[73].

[73] Zagadnienie optymalizacji hiperparametrów zostało omówione w dalszej części rozdziału.

Liczba neuronów w warstwie (bądź warstwach) ukrytych jest również dobierana metodą prób i błędów. Nie ma uniwersalnej, skutecznej metody określania jej struktury. W przypadku modeli płytkich w doborze optymalnej struktury modeli stosowane mogą być algorytmy optymalizujące. Ze względu na sposób działania można podzielić je na trzy grupy [Lula, 1999]:

- redukujące, dla których punktem wyjścia jest sieć o rozbudowanej strukturze, a ich celem jest zmniejszenie liczby neuronów[74];
- rozbudowujące - w trakcie działania dodawane są warstwy, neurony lub połączenia;
- genetyczne, za pomocą których tworzone i testowane są różne warianty sieci.

Popularną strategią jest również wielokrotny, losowy dobór topologii sieci z zadanego zakresu, budowa na tej podstawie kilkudziesięciu lub kilkuset modeli, a następnie wybór najlepszego z nich. Poprawna jest także strategia ręcznego doboru topologii sieci, metodą prób i błędów. W określeniu wstępnej topologii pomocne mogą być różnego rodzaju nieformalne „reguły kciuka". Jedna z nich proponuje przyjąć liczbę neuronów w warstwie ukrytej jako średnią geometryczną liczby neuronów wejściowych i wyjściowych. Kolejna wiąże wielkość sieci z liczbą przypadków w zbiorze uczącym, wskazując, że liczba wag w sieci powinna być 10 razy większa od liczby przypadków w zbiorze uczącym. Powyższe reguły nie mają jednak formalnego uzasadnienia.

W przypadku modeli głębokich czas potrzebny na wykonanie jednego eksperymentu jest znacząco dłuższy. Akceptowalną strategią jest świadome zaprojektowanie nadmiarowej topologii i zastosowanie techniki wczesnego zatrzymania (*early stopping*) uczenia w momencie, gdy zaobserwowano wzrost błędu szacowanego w oparciu o próbę służącą do monitorowania procesu uczenia. Technika ta może być zaliczona do grupy technik regularyzacji [Aggarwal, 2018].

Kolejnym krytycznym etapem procesu projektowania sieci neuronowej jest wybór funkcji aktywacji. Funkcja aktywacji służy do obliczenia wartości sygnału wyjściowego neuronu. Wspólną cechą wszystkich funkcji aktywacji jest ich różniczkowalność. Wymaganie to wynika ze sposobu uczenia wag sieci, które polega na optymalizacji w oparciu o gradient funkcji błędu[75]. Kolejną cechą funkcji aktywacji, choć w tym

[74] Przykładem takiego podejścia jest dodanie to funkcji celu członu kary, którego wartość jest zmniejszana, gdy zmniejszane są wartości wag.

[75] Szereg algorytmów uczenia oprócz gradientu dodatkowo oblicza, bądź przybliża wartości pochodnych wyższego rzędu.

przypadku zdarzają się wyjątki, jest ich monotoniczność. Wybór funkcji aktywacji znacząco wpływa na jakość zbudowanego modelu oraz tempo jego uczenia. Decyzja o wyborze funkcji aktywacji leży w gestii badacza, który eksperymentalnie określa ich rodzaj zawsze w obrębie danej warstwy. W Tabela 4 przedstawiono popularne funkcje aktywacji stosowane w sieciach typu perceptron wielowarstwowy.

Tabela 4 Wybrane funkcje aktywacji perceptronu wielowarstwowego

Nazwa funkcji	Wzór
Liniowa	$\Phi(v) = v$
Signum	$\Phi(v) = sign(v)$
Logistyczna	$\Phi(v) = \frac{1}{1+e^{-v}}$
Tangens hiperboliczny (tanh)	$\Phi(v) = \frac{e^{2v}-1}{e^{2v}+1}$
Wykładnicza	$\Phi(v) = exp\,(-v)$
Sinus	$\Phi(v) = sin\,(v)$
Funkcja typu ReLU (*Rectified Linear Unit*)	$\Phi(v) = max\,\{v, 0\}$
Funkcja typu Leaky ReLU	$\Phi_\alpha(v) = max\,\{\alpha v, v\}$
Funkcja typu ELU (*Exponential Linear Unit*)	$\Phi_\alpha(v) = \begin{cases} \alpha(exp(v)-1) & if\ v < 0 \\ v & if\ v \geq 0 \end{cases}$
Funkcja typu Hard tanh	$\Phi(v) = max\,\{min[v, 1], -1\}$

Źródło: Opracowanie własne na podstawie [Geron, 2017, Aggarval, 2018, TIBCO Statistica, 2017].

Do momentu spopularyzowania sieci głębokich najczęściej wykorzystywanymi funkcjami aktywacji były funkcje logistyczna oraz tanh. Obecnie straciły one popularność

na rzecz funkcji przedstawionych w drugiej części tabeli (począwszy od funkcji typu ReLU). Jedną z motywacji wzrostu popularności tych funkcji jest krótszy czas potrzebny na uczenie sieci [Aggarwal, 2018]. Innym czynnikiem determinującym wybór funkcji aktywacji jest ryzyko wystąpienia zjawiska zanikania, bądź eksplozji gradientu (*gradient exploding*) obserwowanym wraz ze wzrostem liczby warstw sieci. Funkcja typu ReLU okazała się lepszym wyborem od funkcji sigmoidalnej oraz tanh ze względu na fakt, iż nie osiąga ona poziomu nasycenia dla dodatnich wartości. Jej mankament polega na skłonności do udzielania odpowiedzi na poziomie 0[76]. Zjawisko to może dotknąć znaczącą liczbę neuronów w modelu. A. Geron [2017] wskazuje, że lepszym wyborem jest funkcja typu Leaky ReLU, przy czym wartość α=0,2 wydaje się dawać lepsze rezultaty niż α=0,01. Funkcja typu ELU dała lepsze rezultaty od funkcji typu ReLU oraz jej odmian [Clevert i inni, 2015]. Wadą funkcji typu ELU jest dłuższy czas potrzebny na wykonanie obliczeń. Podczas wyboru funkcji aktywacji można kierować się poniższą zasadą [Geron, 2017]:

$$ELU > Leaky\ ReLu > ReLU > Tanh > Logistyczna$$

Należy jednak pamiętać, że w konkretnym zadaniu analitycznym powyższa reguła może nie być zachowana i do wyłonienia optymalnej funkcji aktywacji konieczne będzie przeprowadzenie wielu dodatkowych eksperymentów.

Po wyborze funkcji aktywacji, kolejnym zadaniem jest określenie sposobu inicjalizacji wag. Przed rozpoczęciem procesu uczenia wagom nadawane są (pseudo) losowe wartości. Najczęstszym sposobem inicjalizacji wag jest wylosowanie ich z rozkładu normalnego o średniej równej 0 i wariancji równej 1, bądź jednostajnego z przedziału [-1,1]. Ten sposób inicjalizacji wag okazał się źródłem trudności w uczeniu sieci o większej liczbie warstw ukrytych. Powodował, że wynikające z gradientu, relatywnie duże korekty wag obserwowane w końcowych warstwach sieci zanikały wraz z przejściem procesu uczenia do warstw wcześniejszych nie wywołując praktycznie żadnych zmian wartości wag. Zjawisko to określone zostało problemem zanikającego gradientu (*vanishing grdient*). Problem zanikającego gradientu został powiązany ze sposobem inicjalizacji wag przez X. Glorota i Y. Bengio [2010], którzy w swoim artykule zaproponowali korektę parametrów rozkładu w sytuacji stosowania logistycznej funkcji aktywacji. Korekta sposobu inicjalizacji wag dla dwóch kolejnych funkcji aktywacji Tanh oraz typu ReLU została zaproponowana przez K. He i innych [2015]. Ich propozycje zawarto w Tabela 5.

[76] Tak zwane martwe neurony.

Tabela 5 Sposób inicjalizacji wag w zależności od funkcji aktywacji

Rodzaj funkcji aktywacji	Rozkład jednostajny [-r, r]	Rozkład normalny
Logistyczna	$r = \sqrt{\frac{6}{n_{in} + n_{out}}}$	$\sigma = \sqrt{\frac{2}{n_{in} + n_{out}}}$
Tangens hiperboliczny	$r = \sqrt[4]{\frac{6}{n_{in} + n_{out}}}$	$\sigma = \sqrt[4]{\frac{2}{n_{in} + n_{out}}}$
ReLU i jej odmiany	$r = \sqrt[\sqrt{2}]{\frac{6}{n_{in} + n_{out}}}$	$\sigma = \sqrt[\sqrt{2}]{\frac{2}{n_{in} + n_{out}}}$

Źródło: Opracowanie własne na podstawie [Geron, 2017, Glorot, Bengio, 2010, He i inni, 2015].

Innym sposobem, który pozwala ograniczyć zjawisko zanikania gradientu jest normalizacja wsadowa (*batch normalization*). Technika ta jest analogiczna do standaryzacji wartości wejściowych przed prowadzeniem ich do sieci. Zasadniczą różnicą jest fakt, że jest on realizowana kolejno warstwa po warstwie po obliczeniu wartości funkcji aktywacji, a przed wprowadzeniem ich do kolejnej warstwy. Po wystandaryzowaniu wartości funkcji aktywacji w obrębie warstwy uzyskane wyniki są skalowane, a następnie przesuwane za pomocą dwóch parametrów, które podlegają korekcie w trakcie procesu uczenia. Parametry te umożliwiają ekspozycję wejść na nieliniowe obszary funkcji aktywacji kolejnej warstwy. Autorzy tej techniki [Ioffe, Szegedy, 2015] zauważyli, że redukuje ona problem zanikających gradientów, dodatkowo redukując o rząd wielkości liczbę cykli uczenia. Technika ta może ograniczać ryzyko przeuczenia modelu, przez co redukuje konieczność wykorzystywania technik regularyzacji opisanych w dalszej części podrozdziału. Wadą tego podejścia jest większa komplikacja modelu oraz dłuższy czas potrzebny na oszacowanie prognozy przez model [Geron, 2017].

Kolejnym hiperparametrem wpływającym na czas oraz skuteczność procesu uczenia jest wybór algorytmu uczenia. Jak dotąd, opracowano wiele algorytmów uczenia sieci, które można podzielić na trzy grupy [Lula, 1999]:

- metody gradientowe,

- metody hesjanowe lub korzystające z przybliżenia hesjanu,
- metody nie korzystające z informacji o gradiencie.

Metody gradientowe korzystają z informacji o wartości gradientu minimalizowanej funkcji błędu. Przegląd tych metod można znaleźć w pracy P. Luli [1999]. Ze względu na relatywnie prosty obliczeniowo mechanizm optymalizacji są powszechnie wykorzystywane w uczeniu głębokich sieci neuronowych. Spośród dostępnych algorytmów gradientowych, rekomendowane jest użycie przyspieszonego gradientu Nesterova (*Nesterov Accelerated Gradient,* NAG) [Geron, 2017]. Metody hesjanowe oprócz gradientu wykorzystują podczas optymalizacji również informacje o drugich pochodnych (lub ich przybliżeniu), których macierz nazywamy hesjanem. Algorytmy hesjanowe są skuteczniejsze w poszukiwaniu minimum funkcji błędu, jednak ze względu na wymagania obliczeniowe stosowane są jak dotąd z powodzeniem jedynie dla płytkich sieci neuronowych. Spośród dostępnych algorytmów godnym polecenia wydaje się być algorytm Marquardta-Levenberga [Lula, 1999]. Do technik niekorzystających z informacji o gradiencie zaliczyć można szereg metod heurystycznych opartych na algorytmach genetycznych bądź algorytmie symulowanego wyżarzania. Metody te mogą być traktowane jako niezależne metody uczenia lub jako wstępny etap zwieńczony nauką za pomocą metod gradientowych.

Techniki regularyzacji to różnego rodzaju strategie pozwalające uniknąć nadmiernego dopasowania modelu do danych. Poza wspomnianą normalizacją wsadową do technik regularyzacji można zaliczyć:

- Wczesne zatrzymanie (*early stopping*) polegającą na zatrzymaniu procesu uczenia w momencie, gdy obserwowany jest wzrost błędu w próbie testowej, bądź brak poprawy jakości modelu w zadanym oknie cykli uczenia. Po zatrzymaniu uczenia wagi sieci są przywracane do poziomu, jaki przejęły dla cyklu o najmniejszym błędzie.
- Regularyzacja wag polegająca na dodaniu do optymalizowanej funkcji błędu członu kary. Kara jest wprost proporcjonalna do wartości bezwzględnych wag. Użycie tej techniki redukuje wartości wag, redukując ryzyko nadmiernego dopasowania. Wykorzystywana może być tu zarówno regularyzacja L1 jak i L2, a także ich połączenie.

– *Technika dropout* przedstawiona w pracy N. Srivastava i innych [2014], polegająca na wyłączaniu losowo wybranych neuronów i przypisanych do nich wag z procesu uczenia w danej epoce. Procent wyłączanych neuronów jest regulowany przez hiperparametr zwykle przyjmowany na poziomie 50% [Geron, 2017]. Po zakończonym procesie budowy modelu, wynik sieci dla nowych przypadków obliczany jest na podstawie wszystkich neuronów. Użycie tej techniki poza redukcją ryzyka nadmiernego dopasowania modelu, dodatkowo zwiększa jego skuteczność.

– Technika *Max Norm Regularization* powoduje, że każdego dla neuronu sprawdzany jest warunek [Geron, 2017]:

$$\|w\|_2 \leq r,$$

gdzie r jest hiperparametrem określanym przez badacza. W przypadku niespełnienia tego warunku, wagi są korygowane zgodnie ze wzorem [Geron, 2017]:

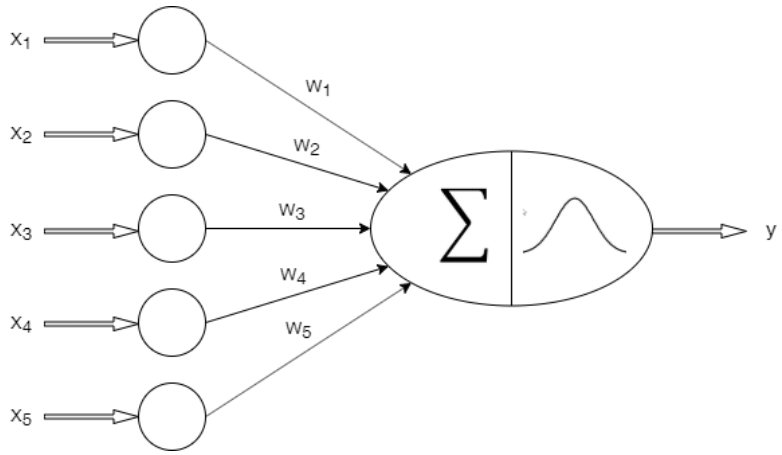
$$w' = w\frac{r}{\|w\|_2}$$

podejście to pozwala również uniknąć problemu zanikających gradientów.

– Rozszerzenie danych (*data augmentation*) polega na wygenerowaniu sztucznych przypadków uczących na podstawie istniejącego zbioru danych. Dane mogą zostać wygenerowane poprzez dodanie do przypadku uczącego losowego szumu, bądź poprzez zastosowanie algorytmu SMOTE lub jego odmian[77]. Działanie to może ograniczać ryzyko nadmiernego dopasowania się modelu do danych.

Nieco zapomnianą architekturą sieci neuronowych są sieci o radialnych funkcjach bazowych (*Radial Basis Function, RBF*). Mimo, że na ich podstawie tej architektury nie są tworzone sieci głębokie, mają znaczny potencjał w rozwiązywaniu problemów klasyfikacyjnych [Aggarwal, 2018]. Sieci te składają się z jednej warstwy ukrytej, która uczona jest w sposób nienadzorowany. Warstwa wyjściowa uczona jest w sposób nadzorowany.

[77] Technika ta najczęściej stosowana jest w klasyfikacji obrazów i polega na stosowaniu różnego rodzaju przekształceń wejściowych obrazów – obrotu, przesuwania, rozciągania, zmiany kontrastu itp.



Rysunek 17 Schemat neuronu radialnego
Źródło: opracowanie własne.

Napływające do neuronu sygnały są agregowane za pomocą wzoru [Tadeusiewicz, 2001]:

$$X \times W = \sum_{i=1}^{d} (x_{i-} w_i)^2$$

W części agregacji obliczany jest zatem kwadrat odległości pomiędzy wektorem wejściowym a wektorem wag. Wagi neuronów w warstwie ukrytej uczone są za pomocą analizy skupień i po zakończonym uczeniu reprezentują środki wyznaczonych skupień tworząc zespół prototypów, do których porównywane są przypadki wejściowe. W części aktywacji znajduje się funkcja rozkładu normalnego. Im bardziej dany przypadek jest bliski wagom danego neuronu, tym silniejsza jest wartość aktywacji. Wyniki wytrenowanej w ten sposób warstwy ukrytej trafiają do warstwy wyjściowej, której neurony działają analogicznie do przedstawionych w opisie perceptronu. Warstwa ta jest uczona w trybie z nauczycielem. Podejście to jest bardzo podobne do metody k-najbliższych sąsiadów, z tą różnicą, że wagi w drugiej warstwie zapewniają dodatkowy poziom dopasowania [Aggarwal, 2018]. Sieci te mają również wiele cech z wspólnych z metodą wektorów nośnych opisaną w dalszej części podrozdziału.

3.2.3. Metoda wektorów nośnych

Metoda wektorów nośnych (*Support Vector Machines*, *SVM*) w swoich ogólnych założeniach została opracowana przez Vladimira Vapnika w połowie lat 60 XX w. W kolejnych latach metoda ta była rozwijana i jest obecnie uważana za jedną z najbardziej

elastycznych i efektywnych metod uczenia maszynowego [Khun, Johnson, 2013]. Wektorami nośnymi określa się przypadki leżące blisko granicy decyzyjnej. Przypadki te są wyznaczane w trakcie procesu estymacji. Po jego zbudowaniu są one częścią zbudowanego modelu umożliwiając przeprowadzenie klasyfikacji nowych obiektów. Granica decyzyjna metody SVM jest definiowana jako linowa kombinacja wektorów nośnych [Flach, 2012]. Pojęciem często używanym w kontekście metody wektorów nośnych jest margines (*margin*). Jest to odległość pomiędzy najbliższymi przypadkami przeciwnych klas leżącymi po przeciwnych stronach granicy decyzyjnej. Im większy margines, tym lepsza generalizacja modelu, co oznacza, że jest on zatem maksymalizowany w trakcie procesu estymacji [Burkov, 2019]. Szukanie granicy decyzyjnej (granicy separującej klasy) o największym marginesie jest tożsame z szukaniem wektorów nośnych. Przypadki identyfikowane jako wektory nośne w sposób jednoznaczny i kompletny określają granicę decyzyjną [Flach, 2012]. Proces klasyfikacji przypadków za pomocą metody SVM dla problemu klasyfikacji dwustanowej dla przypadku u może zostać wyrażony za pomocą następującego wzoru [Khun, Johnson, 2013]:

$$D(u) = \beta_0 + \sum_{i=1}^{n} y_i \alpha_i x_i' u,$$

gdzie: n oznacza liczbę przypadków w zbiorze uczącym, y_i oznacza wartość klasy zmiennej zależnej dla i-tego przypadku uczącego, kodowanej za pomocą wartości 1 dla jednej klasy oraz -1 dla drugiej klasy, $x_i' u$ jest iloczynem skalarnym pomiędzy klasyfikowanym przypadkiem u a i-tym przypadkiem uczącym x, α_i jest parametrem modelu uzyskanym podczas optymalizacji[78], przyjmującym wartości nieujemne [Flach, 2012], β_0 jest odpowiednikiem wyrazu wolnego w modelu liniowym.

Sam proces uczenia modelu jest realizowany za pomocą metody mnożników Lagrange'a i sprowadza się do optymalizacji poniższej funkcji [Flach, 2012]:

$$\alpha_1^*, \dots, \alpha_n^* = \underset{\alpha_1,\dots,\alpha_n}{arg\ max} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i\, y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j + \sum_{i=1}^{n} \alpha_i,$$

[78] Jest to równocześnie wartość mnożnika Lagrange'a. Szczegóły obliczeniowe związane z procesem estymacji można znaleźć na przykład we [Flach, 2012, Baesens, 2014].

z zastrzeżeniem, że $\alpha_i \geq 0$ oraz $\sum_{i=1}^{n} \alpha_i y_i = 0$. Wyrażenie $\boldsymbol{x}_i \cdot \boldsymbol{x}_j$ jest iloczynem skalarnym wartości kolejnych przypadków uczących.

Dla przypadków uczących leżących poza granicą decyzyjną parametr α przyjmuje wartość 0. Innymi słowy, wektorami nośnymi są przypadki, dla których parametr ten jest większy od 0. Dany przypadek klasyfikowany jest do pierwszej klasy, jeśli uzyskany wynik jest większy od 0 i do drugiej, jeśli wynik jest ujemny. Na podstawie uzyskanego wyniku za pomocą osobnej formuły obliczane jest prawdopodobieństwo przynależności do modelowanych klas. Do tego celu wykorzystywana jest funkcja logistyczna pozwalająca zmapować odległość od granicy decyzyjnej na oszacowanie prawdopodobieństwa wystąpienia modelowanej klasy [Flach, 2012]. Przedstawione równanie wyznacza granicę decyzyjną w postaci hiperpłaszczyzny co pozwala na rozwiązywanie problemów, które są linowo separowalne. Nieliniowa granica decyzyjna jest uzyskiwana poprzez zamianę iloczynu skalarnego $x_i' u$ przez funkcję jądrową wektorów x oraz u obliczaną według wzoru [Khun, Johnson, 2013].

$$D(u) = \beta_0 + \sum_{i=1}^{n} y_i \alpha_i K(x_i, u),$$

Najbardziej popularnymi funkcjami są funkcja liniowa, wielomianowa, tangens hiperboliczny oraz radialna funkcja bazowa (RBF, *Radial Basis Function*) [Baesens, 2014]. Ta ostatnia jest obliczana według wzoru:

$$RBF\ K(x_i, u) = exp(\frac{-\|x - u\|^2}{\sigma^2})$$

Zastosowanie funkcji RBF daje najlepsze wyniki modelowania [Baesens, 2014], chociaż jej wadą jest konieczność znalezienia optymalnej wartości hiperparametru σ.

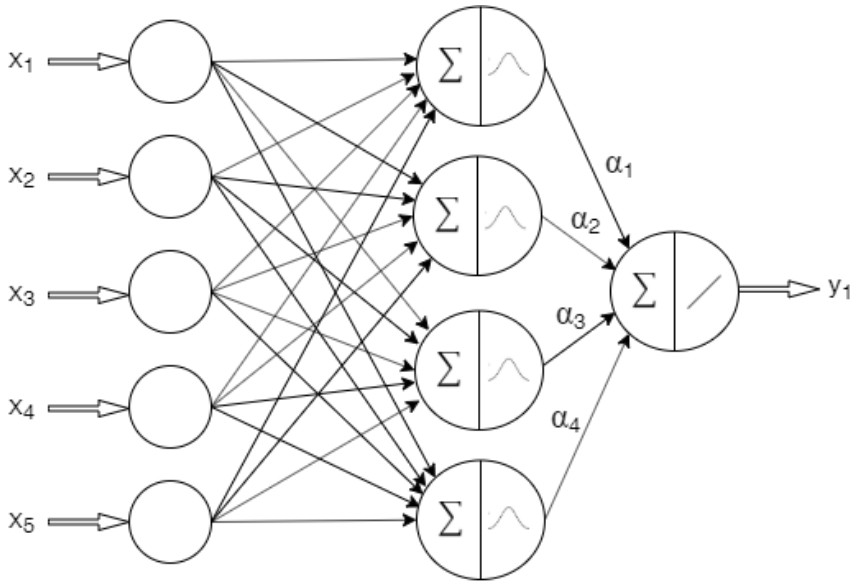
W przypadku rzeczywistych zbiorów danych, idealna separacja jest praktycznie niemożliwa, niezależnie od wybranej funkcji bazowej. Przypadki z obu klas zmiennej Y najczęściej zachodzą na siebie, utrudniając proces identyfikacji granicy decyzyjnej. Rozwiązaniem jest umożliwienie niektórym przypadkom znalezienie się po przeciwnej stronie granicy decyzyjnej. Optymalizowany problem jest kompromisem pomiędzy maksymalizacją marginesu (*maximum margin classifier*) a minimalizacją przypadków znajdujących się po niewłaściwej stronie granicy decyzyjnej. Relacja pomiędzy dwoma optymalizowanymi elementami jest regulowana za pomocą hiperparametru C dobieranego

eksperymentalnie. Dodanie hiperparametru C nie zmienia postaci optymalizowanego wyrażenia, wprowadza natomiast dodatkowe restrykcje na wartości α [Flach, 2012].

$$\alpha_1^*, \dots, \alpha_n^* = \underset{\alpha_{1,\dots,\alpha_n}}{arg\ max} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i\, y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j + \sum_{i=1}^{n}\alpha_i,$$

z zastrzeżeniem, że $0 \leq \alpha_i \leq C$ oraz $\sum_{i=1}^{n} \alpha_i\, y_i$=0.

Niska wartość hiperparametru C zwiększa wagę wielkości marginesu, mniejszą wagę przykładając do przypadków znajdujących się po niewłaściwej stronie granicy decyzyjnej. Wraz z jego wzrostem spada tolerancja metody na błędnie zaklasyfikowane przypadki kosztem szerokości marginesu. Model SVM może zostać przedstawiony symbolicznie w postaci sieci neuronowej [Baesens, 2014].



Rysunek 18 Model SVM w postaci sieci neuronowej
Źródło: opracowanie własne na podstawie [Baesens, 2014].

Sieć SVM posiadała będzie trzy warstwy. W warstwie wejściowej prezentowane są przypadki uczące, a każdy neuron tej warstwy odpowiada jednej zmiennej zawartej w zbiorze uczącym. Neurony w warstwie ukrytej są tożsame z wektorami nośnymi, a ich wagami są wartości konkretnych zmiennych danego wektora nośnego. Wewnątrz neuronów warstwy ukrytej, na podstawie danego przypadku ze zbioru danych (wektora wejściowego), wektora wag oraz przyjętej funkcji aktywacji (funkcji jądrowej) następuje agregacja sygnału do wartości skalarnej oraz jego transformacja. Wagi pomiędzy warstwą ukrytą a warstwą wyjściową są reprezentowane przez parametry α. Neuron w warstwie

wyjściowej jest neuronem liniowym. Największą różnicą pomiędzy modelem SVM a sieciami neuronowymi jest sposób doboru neuronów w warstwie ukrytej. W przypadku sieci neuronowych liczba neuronów jest hiperparametrem ustalanym w sposób eksperymentalny, natomiast w przypadku metody SVM jest uzyskiwana w sposób automatyczny w wyniku procesu budowy modelu.

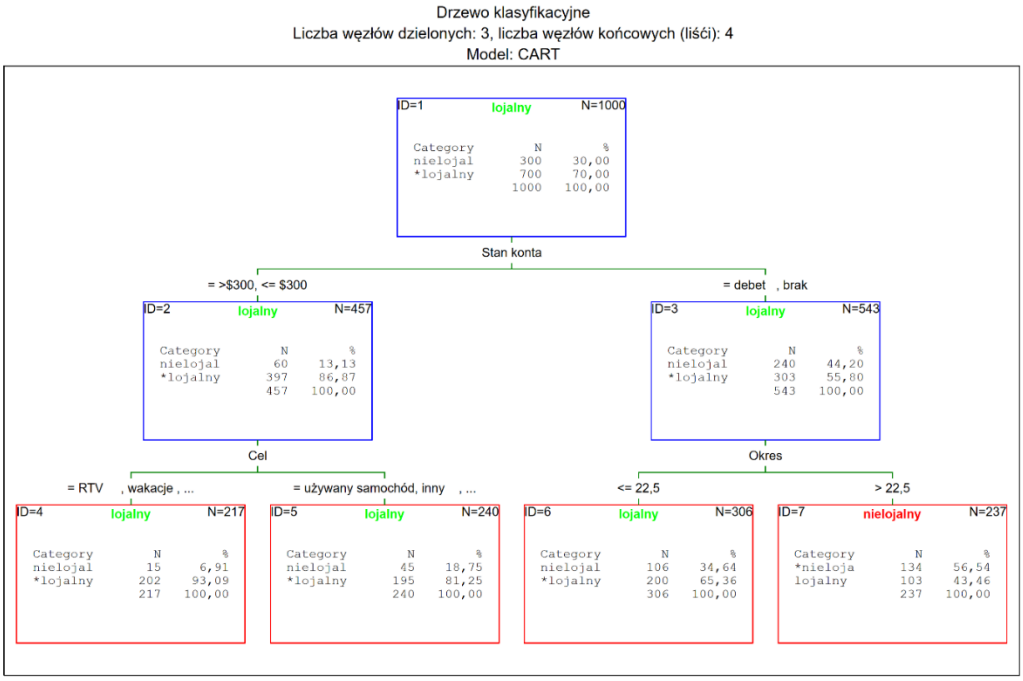
3.2.4. Drzewa klasyfikacyjne i regresyjne

Drzewa klasyfikacyjne (oraz regresyjne)[79] są popularną metodą uczenia maszynowego cechującą się dużą prostotą, elastycznością oraz łatwością interpretacji [Łapczyński, 2010]. Algorytm drzewa działa na zasadzie rekurencyjnego podziału zboru obserwacji. Na podstawie przyjętego kryterium optymalizacyjnego oraz zgromadzonych danych algorytm określa optymalny podział zbioru obserwacji względem każdego z predyktorów z osobna. W kolejnym kroku wybierana jest ta zmienna niezależna, względem której podział w największym stopniu poprawia przyjęte kryterium optymalizacji. Każdy podział tworzy, w zależności od algorytmu, dwa bądź większą liczbę podzbiorów[80], węzłów potomnych (*child nodes*), które w kolejnych krokach mogą podlegać dalszemu podziałowi. Graficzną reprezentacją procesu poddziału jest drzewo rozgałęziające się od góry ku dołowi. Proces budowy drzewa jest zatrzymywany w momencie uzyskania względnie jednorodnego węzła, bądź w przypadku spełnienia określonego przez użytkownika warunku zatrzymania związanego najczęściej z licznością podzbioru, bądź z liczbą podziałów.

Dzielony zbiór nosi nazwę węzła macierzystego (*parent node*). W kolejnych krokach procedury węzeł potomny, który jest dalej dzielony zaczyna pełnić rolę węzła potomnego dla kolejnego etapu. Węzeł który nie podlega dalszemu podziałowi staje się węzłem końcowym, nazywanym liściem (*leaf, terminal node*). Przykładowe drzewo klasyfikacyjne przedstawia Rysunek 19.

[79] W sytuacji, gdy zmienna zależna jest zmienną nominalną bądź porządkową mówimy o drzewach klasyfikacyjnych. Drzewa regresyjne rozwiązują problemy, w których zmienna zależna jest mierzona na skali przedziałowej bądź ilorazowej.

[80] W przypadku algorytmów dzielących zbiór zawsze na dwa podzbioru mówimy o drzewach binarnych, w przeciwnym przypadku o drzewach dowolnych.



Rysunek 19 Przykładowe drzewo klasyfikacyjne
Źródło: Opracowanie własne w programie TIBCO Statistica 13.3.

Celem budowy modelu jest uzyskanie podzbiorów maksymalnie jednorodnych pod względem wartości zmiennej zależnej. Jest to proces wieloetapowy, a w każdym kroku można wykorzystywać inną zmienną niezależną. Na każdym etapie analizuje się wszystkie predyktory i wybiera ten, który zapewnia najlepszy podział węzła, czyli wydziela najbardziej homogeniczne podzbiory.

Powyższy sposób działania odnosi się do wielu algorytmów, które różnią się od siebie szczegółami kryterium optymalizacyjnego. Do najbardziej popularnych należą algorytm CART, CHAID oraz C4.5.

Za pomocą drzewa klasyfikacyjnego można przedstawić relacje pomiędzy zmienną zależną a zbiorem predyktorów. Metoda ta ma wiele zastosowań, z których najważniejsze to [Ma, 2018]:

- Segmentacja, czyli identyfikacja przypadków, które prawdopodobnie należą do określonej grupy klientów;
- Podział na grupy ryzyka związanego z realizacją zdarzenia opisywanego przez analizowaną zmienną zależną, na przykład ryzyka odejścia do konkurencji;
- Ocena prawdopodobieństwa wystąpienia modelowanego zdarzenia. Prawdopodobieństwo to szacowane jest na podstawie frakcji obiektów danej klasy

w węźle końcowym. P. Flach [2012] zaleca dodatkowo, aby wartość tę wygładzić za pomocą poprawki Laplace'a (*Laplace correction*) bądź techniki *m-estimate*;

- Identyfikacja interakcji umożliwiająca zdefiniowanie zależności pomiędzy dwoma bądź większą liczbą predyktorów a zmienną zależną. Poza walorami informacyjnymi, interakcje mogą zostać użyte jako podstawa do tworzenia zmiennych pochodnych, wykorzystywanych w innych metodach analitycznych [Migut i inni, 2013, Srivastava, 2013];
- Selekcja zmiennych umożliwiająca wybór niewielkiej liczby zmiennych niezależnych najsilniej powiązanych ze zmienną zależną. Wybrane w ten sposób zmienne objaśniające mogą być użyte do budowy modeli za pomocą innych metod;
- Transformacje zmiennych pozwalające na dyskretyzację predyktorów ilościowych oraz na łączenie klas zmiennych jakościowych. W takich przypadkach budowane jest drzewo z jednym predyktorem, a podziały utworzone przez drzewo wskazują odpowiednio granice klas w przypadku predyktora ilościowego, bądź grupy w przypadku predykora jakościowego.

Uzyskany węzeł końcowy uznawany jest za czysty (*pure*), jeżeli należą do niego przypadki należące jedynie do jednej klasy zmiennej zależnej. Na podstawie względnej liczności przypadków w węźle należących do wszystkich klas zmiennej zależnej można wyróżnić kilka miar zanieczyszczenia (*impurity*) węzła. Miary te są oparte na [Flach, 2012]: względnej liczności mniej licznej klasy (*minority class*), indeksie Giniego, entropii czy statystyce chi-kwadrat[81].

Miara zanieczyszczenia oparta na względnej liczności mniej licznej klasy jest obliczana za pomocą wzoru [Flach 2012]:

$$impurity = min(p, 1 - p)$$

gdzie p oznacza frakcję modelowanej (zwykle mniej licznej) klasy w analizowanym węźle[82]. Miara ta jest też nazywana wskaźnikiem błędu (*error rate*). Nie jest zalecane stosowanie tej miary zanieczyszczenia, głównie ze względu na wysoką wrażliwość na niezbilansowany rozkład klas zmiennej zależnej [Flach, 2012].

[81] W programie Tibco Statistica można dodatkowo użyć miary G kwadrat będącej modyfikacją statystyki chi-kwadrat obliczaną za pomocą metody największej wiarygodności.

[82] Wzór (podobnie jak kolejne) odnosi się do przypadku, w którym zmienna zależna przyjmuje dwa stany.

Inną miarą zanieczyszczenia węzła jest indeks Giniego [Łapczyński, 2010] powszechnie wykorzystywany w drzewach CART i wyrażony wzorem:

$$Gini\ index = 2 \times p \times (1 - p) = 1 - p^2 + (1 - p)^2$$

Alternatywną miarą opartą na indeksie Giniego jest zaproponowany przez P. Flacha [2012] pierwiastek tej miary $\sqrt{Gini}$. Jego przewagą nad klasycznym indeksem oraz miarami z pozostałych grup jest niewrażliwość na niezbilansowany rozkład zmiennej zależnej w modelowanym zbiorze danych.

Kolejną grupą miar zanieczyszczenia węzła są te oparte na entropii. Wyróżnia się tutaj przyrost informacji (*information gain)* oraz *gain ratio* przedstawione w podrozdziale opisującym filtry służące do wstępnej selekcji zmiennych. Preferowaną miarą zanieczyszczenia jest *gain ratio* ze względu na jej niewrażliwość na liczbę klas predyktorów[83].

Zanieczyszczenie może odnosić się nie tylko do pojedynczego węzła, ale również dla gałęzi lub całego drzewa. Określa się ją wtedy jako średni poziom zanieczyszczenia węzłów końcowych ważony frakcjami przypadków, jakie do nich należą [Ma, 2018]. Zasada te jest wykorzystywana w procesie poszukiwania optymalnego podziału węzła macierzystego na węzły potomne. Za optymalny przyjmuje się podział, po którym zanieczyszczenie wszystkich węzłów potomnych jest najmniejsze.

Alternatywnym podejściem do wyboru optymalnego podziału na podstawie zanieczyszczenia jest wykorzystanie statystyki chi-kwadrat. Węzły uzyskane na podstawie optymalnego podziału charakteryzują się maksymalną w stosunku do podziałów konkurencyjnych wartością tej statystyki. Podejście oparte na statystyce chi-kwadrat pozwala na uzyskanie drzew analogicznych do drzew zbudowanych na podstawie indeksu Giniego [Grabmeier, Lamble, 2007].

Istotnym zagadnieniem związanym z budową drzewa klasyfikacyjnego jest określenie jego optymalnych rozmiarów. Stosuje się tutaj dwa przeciwstawne kryteria. Pierwszym z nich jest oczekiwanie, aby poziom zanieczyszczenia uzyskanego drzewa był jak najmniejszy, a drugim jest oczekiwanie, aby model posiadał zdolność do generalizacji. Wyzwaniem jest zatem zbudowanie drzewa o możliwie niskim poziomie zanieczyszczenia bez konieczności jego zbytniej rozbudowy. Jest to klasyczny przykład konieczności

[83] Miara *information gain* faworyzuje predyktory o większej liczbie klas.

odnalezienia kompromisu pomiędzy obciążeniem (*bias*) a wariancją modelu, który opisany jest między innymi przez T Hastie i inni [2009].

Strategie poszukiwania drzewa optymalnych rozmiarów można podzielić na dwie grupy. Pierwsza związana jest z określeniem kryteriów zatrzymania procesu uczenia (*stopping*), czyli ma na celu niedopuszczenie do nadmiernego rozrostu drzewa. Druga grupa polega na przycinaniu (*pruning*) drzewa nadmiernie rozbudowanego.

Pierwsza strategia zakłada uzyskanie optymalnego rozmiaru poprzez niedopuszczenie do wykonania zbędnego podziału, który spowodowałby nadmierny rozrost drzewa. Popularnymi sposobami na uzyskanie tego celu są określane przez badacza kryteria zatrzymania takie jak:

- liczba węzłów drzewa (wielkość),
- liczba poziomów drzewa (głębokość),
- liczność węzła podlegającego podziałowi.
- liczność potomka węzła potomnego.

Hiperparametry te mogą zostać określone na podstawie wiedzy eksperckiej badacza uzupełnionej o dostępne w literaturze reguły kciuka, które sugerują np., aby liczność węzła końcowego była większa bądź równa 50, bądź też nie mniejsza od 5% ogólnej liczby przypadków.

Inną stosowaną tutaj metodą jest zatrzymanie podziału na podstawie wyniku testu zgodności chi-kwadrat w oparciu formułę [Ma 2018]:

$$\chi^2 = \frac{(n_{1L} - P \times n_1)^2}{P \times n_1} + \frac{(n_{2L} - P \times n_2)^2}{P \times n_2}$$

gdzie symbole są zgodne z oznaczeniami umieszczonymi w Tabela 6, a $P = \frac{n_L}{n}$.

Tabela 6 Schematyczny podział węzła macierzystego na węzły potomne

	Zmienna zależna – klasa 1	Zmienna zależna – klasa 2	Suma w wierszu
Węzeł potomny L	n_{1L}	n_{2L}	n_L
Węzeł potomny R	n_{1R}	n_{2R}	n_R
Węzeł macierzysty	n_1	n_2	n

Źródło: opracowanie własne.

Podejście to jest podejściem zachowawczym. Dodatkowo w przypadku dużej liczby podziałów wzrasta ryzyko popełnienia błędu I rodzaju co implikuje konieczność korekty poziomu istotności za pomocą na przykład poprawki Bonferroniego.

Kolejną popularną strategią zatrzymania procesu budowy drzewa jest podział zbioru danych na próbę uczącą oraz próbę testową[84]. Na podstawie próby uczącej realizowane są podziały drzewa, a następnie w oparciu o przypadki z próby testowej sprawdzany jest błąd modelu[85]. Proces uczenia zatrzymywany jest w momencie wzrostu tego błędu.

Zbyt wczesne zatrzymanie procesu uczenia wiąże się z ryzykiem pominięcia potencjalnie wartościowego podziału, który mógłby być zrealizowany na późniejszym etapie budowy modelu. Z tego powodu bardziej zalecanymi strategiami budowy drzewa o optymalnej głębokości są strategie polegające na przycinaniu drzewa. W podejściach tych powstaje nadmiernie rozbudowane drzewo o relatywnie niskich hiperparametrach zatrzymania (czasem zalecana jest wręcz budowa drzewa z liśćmi zawierającymi pojedyncze przypadki). Drzewo to jest następnie zmniejszane poprzez przycinanie (*pruning*) nadmiarowych węzłów lub gałęzi. Za nadmiarowe uznaje się te węzły, które w niewielkim stopniu redukują zanieczyszczenie drzewa.

Zaczynając od rozbudowanego drzewa kolejne gałęzie są przycinane do poziomu, który w największym stopniu redukuje minimalny koszt złożoności (*mnimal cost-complexity*) wyrażony wzorem [Breiman i inni, 1984]:

$$R_\alpha(T) = R(T) + \alpha|T|$$

[84] W przypadku małych zbiorów danych strategia ta może zostać zastąpiona przez wielokrotną walidację krzyżową (*v-fold cross vallidation*) opisaną w rozdziale 4.

[85] Miary jakości modelu zostały omówione w rozdziale 4.

$R(T)$ jest przyjęta miarą jakości modelu, α jest współczynnikiem kary większym od zera, natomiast $|T|$ oznacza liczbę liści w modelu. $R_\alpha(T)$ jest zatem liniową kombinacją jakości modelu oraz jego złożoności. Po zastosowaniu powyższej formuły końcowy sposób przycięcia drzewa jest zwykle wyznaczany za pomocą reguły 1 błędu standardowego (*1 SE rule*).

Ostateczny model jest najmniejszym drzewem, którego błąd spełnia poniższy warunek:

$$R(T)_{Finalne} \leq R(T)_{Minimalne} + 1SE$$

Jedynym ograniczeniem tego podejścia są większe wymagania obliczeniowe [Ma, 2018].

Algorytm drzew decyzyjnych CART ma dwa interesujące rozwiązania, dzięki którym badacz wpływa na strukturę i jakość rozwiązania modelu: koszty błędnych klasyfikacji oraz prawdopodobieństwo *a priori*. Te właściwości algorytmu CART są wykorzystywane w przypadku zmiennych Y o niezbilansowanym rozkładzie, co jest szczególnie ważne podczas modelowania migracji klientów.

Negatywny wpływ niezbilansowanego rozkładu może oczywiście zostać złagodzony za pomocą zastosowania specjalnych strategii przygotowania zbioru danych przedstawionych w rozdziale 2, zaś koszty błędnych klasyfikacji uwzględnione po zbudowaniu modelu na etapie określania punktu odcięcia, nie mniej jednak dostępność tych opcji umożliwia elastyczne działania już na etapie budowy modelu.

Obydwa hiperparametry są wykorzystywane w procesie wyboru optymalnych podziałów drzewa korygując obliczanie miar zanieczyszczenia węzła. Ich korekta powoduje, że finalnie możliwe jest uzyskanie innej struktury modelu. Fakt ten rodzi oczywiście pytanie o optymalne wartości kosztów i prawdopodobieństw *a priori*. M. Łapczyński [2010] zaleca, aby pierwszy model budować przy równych prawdopodobieństwach *a priori* (oraz kosztach błędnych klasyfikacji), gdyż wtedy każda klasa zmiennej zależnej jest traktowana jednakowo. Ponieważ badaczowi zazwyczaj zależy na poprawnej predykcji nielojalnych klientów, może zwiększyć on prawdopodobieństwo *a priori* tej klasy (ponad poziom 50%) lub zwiększyć koszt klasyfikacji fałszywie negatywnej.

W praktyce zmiana kosztów błędnych klasyfikacji wynika częściej z chęci uwzględnienia w modelu uwarunkowań biznesowych, natomiast zmiana prawdopodobieństwa *a priori* z chęci oddania faktycznego rozkładu zmiennej zależnej. W przypadku modelowania niezbilansowanej zmiennej Y zmiana tych opcji może być stosowana zamiennie [Ma, 2018].

Drzewa klasyfikacyjne do metoda prosta w użyciu i interpretacji, pozwalająca analizować wiele typów predyktorów oraz zbiory z brakami danych. Jest odporna na występowanie wysokiej korelacji pomiędzy predyktorami oraz na zmienne nie wykazujące siły predykcyjnej. Wymienionym zaletom towarzyszą jednak pewne wady. Drzewa są bardzo niestabilne[86] tj. wrażliwe na niewielkie zmiany w zbiorze uczącym, a ponadto charakteryzują się względnie niską mocą predykcyjną [Khun, Johnsonn, 2013].

3.2.5. Zespoły modeli

Jedną z najbardziej skutecznych strategii budowania rozwiązań analitycznych, często przewyższających inne podejścia jest tworzenie zespołów modeli (*model ensembles*). Główną ideą jaka przyświeca budowie modelu zagregowanego jest tworzenie modeli bazowych, które różniłyby się od siebie, pokrywając różne obszary zmienności analizowanych danych. Dzięki ich wspólnemu działaniu oczekuje się redukcji słabości i braków pojedynczego modelu [Baesens, 2014]. Prognozowanie za pomocą modeli zagregowanych jest dokonywane przez uśredniane predykcji pojedynczych modeli dla zmiennej ilościowej Y bądź przez głosowanie dla zmiennej jakościowej Y [Flach, 2012]. Taki zabieg może prowadzić do bardziej stabilnych i niezawodnych oszacowań poprzez redukcję losowej wariancji związanej z pojedynczym modelem. Warunkiem otrzymania satysfakcjonującego rozwiązania jest tu jednak osiągniecie zadawalającej różnorodności modeli składających się na zespół. Wykorzystywane podejścia, będące zazwyczaj kombinacją elementarnych strategii polegających na budowie modeli składowych, to:

- losowanie podzbiorów zbioru uczącego,
- losowanie podzbiorów dostępnych predyktorów,
- wykorzystywanie różnych metod analitycznych,
- różne ustawienia hiperparametrów.

Osiągnięcie tej różnorodności ułatwia stosowanie metod wrażliwych na zmiany zbioru uczącego (na przykład drzewa klasyfikacyjne) bądź hiperparametrów (na przykład sieci neuronowe). Niewątpliwą wadą zespołu modeli jest zwiększone zapotrzebowanie na moc obliczeniową oraz większy poziom komplikacji uzyskanych modeli.

[86] Ta wada okaże się zaletą podczas konstruowania zespołów modeli opisanych w kolejnej części opracowania.

Bagging jest prostą, lecz skuteczną strategią, w której powstaje duża liczba różnorodnych modeli składowych na podstawie różnych podzbiorów zbioru uczącego. Zbiory uczące są tworzone jako próby bootstrapowe, co powoduje, że niektóre przypadki w danym zbiorze pojawiają się wielokrotnie, niektóre zaś są pomijane. Przy założeniu, że próba bootstrapowa ma taką samą liczebność jak zbiór uczący, to każda z nich pomija około jedną trzecią przypadków. Różnice pomiędzy poszczególnymi próbami generują pożądaną przez badacza zmienność w modelach składowych. Metoda *bagging* jest szczególnie użyteczna w połączeniu z drzewami klasyfikacyjnymi, które są wrażliwe na zmiany w obrębie zbioru uczącego.

W połączeniu z drzewami klasyfikacyjnymi lub regresyjnymi stosowana jest dodatkowo druga technika zwiększająca zmienność finalnych modeli polegająca na losowaniu do każdego modelu składowego określonej liczby predyktorów. Połączenie tych dwóch technik spowodowało powstanie metody o nazwie losowy las (*Random Forests*)[87].

Bardzo podobny mechanizm budowy zespołu modeli może być stosowany w przypadku innych metod cechujących się wrażliwością wyniku na zmianę wybranych elementów, a tym samym zapewniających różnorodność modeli składowych. W programie Statisitca 13.3, zaimplementowany jest na przykład mechanizm budowy zespołów modeli neuronowych. Mechanizm różnicowania oparty jest na losowaniu prób bootstrapowych, na podstawie których budowane są kolejne modele. Dodatkowym czynnikiem różnicującym są wagi inicjalizowane niezależnie dla każdego z modeli składowych.

Inną metodą opartą na idei losowania prób bootstrapowych jest rotacyjny las (*Rotation Forests*) zaprezentowany w 2006 roku [Rodriguez i inni]. W metodzie tej drzewa składowe budowane są na podstawie sztucznych zmiennych będących liniowymi kombinacjami pierwotnych predyktorów. Wartości nowych zmiennych uzyskuje się za pomocą metody głównych składowych (PCA). Przygotowanie zbioru do budowy pojedynczego drzewa przebiega w następujący sposób:

- Zbiór danych dzieli się na podzbiory zawierające wszystkie przypadki i (zazwyczaj rozłączny) podzbiór zmiennych objaśniających podział zbioru danych względem predyktorów na zadaną przez użytkownika liczbę grup.
- Część przypadków z każdego podzbioru jest usuwana a na podstawie pozostałych tworzona jest próba boostrapowa licząca 75% ogólnej liczby przypadków [Łapczyński, 2017].

[87] Metoda i nazwa została zaproponowana przez L. Briemana [2001].

– Na tak przygotowanych podzbiorach wykonywana jest niezależnie analiza PCA.
– Uzyskane transformacje ze wszystkich prób są stosowane na oryginalnym zbiorze danych. Obliczone w ten sposób sztuczne zmienne tworzą zestaw predyktorów do budowy modelu.

Ograniczeniem rotacyjnego lasu jest niewątpliwie możliwość stosowania tego algorytmu jedynie dla predyktorów ilościowych. M. Łapczyński [2017] zwraca uwagę na możliwość ominięcia tego ograniczenia poprzez wykorzystanie na przykład korelacji tetrachorycznych w klasycznej analizie PCA bądź analogicznych procedur dostosowanych do obsługi zmiennych jakościowych takich jak wielowymiarowa analiza korespondencji bądź skalowanie optymalne. Studium porównawcze szeregu metod analitycznych [Bagnall i inni, 2018] wykazała, że na tle innych metod (SVM, sieci neuronowe, losowy las, drzewa wzmacniane) rotacyjny las daje ponadprzeciętne wyniki.

Losowanie prób boostrapowych o liczebności pierwotnego zbioru danych może w przypadku większych zbiorów wymagać odpowiednio długiego czasu na realizację obliczeń. Alternatywą może być losowanie prób do budowy modeli o mniejszej liczebności od pierwotnego zboru. M. Łapczyński [2021] opisuje dwie strategie tego typu. Pierwsza z nich nazwana bootstrapem typu *m* z *n* (*m-out-of n bootstrap*, BMN) została przedstawiona w [Bickel i inni, 2012], która zastosowana do budowy modeli skutkować będzie losowaniem mniejszych prób. Wielkość podzbiorów jest hiperparametrem określanym przez użytkownika. Bardziej skomplikowaną strategią jest *bag of little bootstraps* (BLB) [Kleiner i inni, 2014]. W tym przypadku oryginalny zbiór danych dzieli się na określoną liczbę rozłącznych prób. Dla każdej z prób stosuje się tradycyjną strategię znaną na przykład z losowego lasu. W ostatnim kroku łączy się modele budowane na podstawie rozłącznych prób.

Wzmacnianie (*boosting*) jest kolejną techniką budowania zespołów modeli wykorzystującą bardziej skomplikowany niż *bagging* mechanizm budowy modeli składowych. U podstaw tej strategii leży idea stopniowej poprawy słabego klasyfikatora, aby w kolejnych krokach stawał się on coraz bardziej skuteczny. Wzmacnianie w swojej pierwotnej koncepcji polegać miało na odfiltrowaniu obserwacji, z poprawną klasyfikacją których bieżący model „radził sobie” w zadowalającym stopniu i koncentracji na klasyfikowaniu pozostałych „trudnych” obserwacji.

Pierwszym algorytmem, w którym zastosowano koncepcję wzmacniania był *Adaptive Boostig* najczęściej nazywany *AdaBoost*[88]. Najpopularniejszą metodą do budowy modeli bazowych są drzewa klasyfikacyjne (regresyjne) CART składające się jedynie z dwóch liści. Drzewa takie noszą miano pniaków (*stumps*). Zbudowany pniak ze względu na swoją prostotę jest słabym klasyfikatorem, niezdolnym do aproksymacji złożonych zależności.

W pierwszym kroku każdemu przypadkowi ze zbioru uczącego przypisywana jest taka sama waga, których suma wynosi 1. Zatem przed zbudowaniem pierwszego pniaka każdy z przypadków jest równie ważny. Po zbudowaniu modelu obliczany jest jego błąd (*total error*), jako suma wag niepoprawnie zaklasyfikowanych przypadków. Wartość błędu mieści się w granicach [0;1], ponieważ suma wag wynosi 1. Błąd ten jest podstawą do określenia siły głosu (*amount of say*) zbudowanego drzewa na podstawie poniższego wzoru [Baesens, 2014]:

$$amount\ of\ say = \frac{1}{2} \times ln\ (\frac{1 - total\ error}{total\ error})$$

Im błąd jest bliższy zeru, tym większe wartości przyjmuje siła głosu danego drzewa. Błąd powyżej 0,5 skutkuje wartościami ujemnymi[89]. Przed zbudowaniem kolejnego drzewa korekcie podlegają wagi przypadków zbioru uczącego. W pierwszej kolejności zwiększane są wagi przypadków błędnie zaklasyfikowanych zgodnie ze wzorem [Baesens, 2014]:

$$w_i = w_{i-1} \times e^{amount\ of\ say}$$

Im większa wartość siły głosu, tym większa korekta wag niepoprawnie zaklasyfikowanych przypadków. W kolejnym kroku zmniejszane są wagi poprawnie zaklasyfikowanych przypadków zgodne ze wzorem [Baesens, 2014]:

$$w_i = w_{i-1} \times e^{-\ amount\ of\ say}$$

Na koniec wszystkie wagi są skalowane w taki sposób, aby ich suma wynosiła 1. Kolejne drzewo budowane jest na podstawie próby przypadków wylosowanych ze zbioru uczącego w sposób niezależny (ze zwracaniem). Podczas losowania uwzględniane są wagi przypadków. Po zakończonym losowaniu wybranym przypadkom przypisuje się równe

[88] Algorytm przedstawiony przez Y. Freund i innych [1999].

[89] W praktyce do błędu dodawana jest niewielka wartość, w celu uniknięcia dzielenia przez zero dla skrajnych przypadków błędu.

wagi. Po zbudowaniu zespołu modeli ostateczna klasyfikacja jest uzyskiwana przez głosowanie poszczególnych modeli, przy czym wkład każdego z nich jest proporcjonalny do jego siły głosu.

Nowszym algorytmem, który właściwie wyparł AdaBoost jest algorytm wzmacniania gradientowego (*gradient boosting*). Zasadniczą różnicą tej metody w stosunku do *AdaBoost* jest budowa drzew, w których zmienna zależna reprezentuje reszty z modelu uzyskanego w poprzednim kroku. Przed rozpoczęciem procesu modelowania zmienna zależna przekodowywana jest w sposób analogiczny do opisywanego w modelu regresji logistycznej. Przypadkom reprezentującym osoby nielojalne przypisywana jest wartość 1, a pozostałym wartość 0. Model inicjalizowany jest za pomocą stałej wartości. Wszystkim przypadkom przypisywana jest wartość przewidywana równa logarytmowi szansy wystąpienia w zbiorze osób nielojalnych. Po inicjalizacji wykonywane są następujące kroki:

- Za pomocą funkcji logistycznej logarytm szansy jest przekształcany do prawdopodobieństwa występowania w zbiorze przypadków osób nielojalnych.
- Na podstawie uzyskanych prawdopodobieństw obliczane są reszty modelu.
- Budowany jest model drzewa regresyjnego[90] w którym rolę zmiennej zależnej pełnią obliczone wartości reszt.
- Po zbudowaniu modelu predykcję każdego z liści wyraża się w postaci logarytmu szansy za pomocą poniższego wzoru [Starmer, 2019]:

$$wartość\ prognozowana\ liścia = \frac{\sum_{i=1}^{N} r_i}{\sum_{i=1}^{N} [p_i \times (1 - p_i)]}$$

gdzie N oznacza liczbę przypadków w danym liściu, r_i oznacza wartość zmiennej zależnej[91] dla i-tego przypadku w liściu, p_i oznacza prawdopodobieństwo bycia nielojalnym klientem obliczone w poprzednim kroku algorytmu dla i-tego przypadku w liściu.

[90] W odróżnieniu od algorytmu AdaBoost, drzewa składowe nie są zazwyczaj pniakami, ale składają się zazwyczaj z od 8 do 32 liści.

[91] Należy pamiętać, że zmienna zależna jest resztą pomiędzy stanem faktycznym a prawdopodobieństwem określonym we wcześniejszym etapie.

– Obliczone wartości prognozowane (wyrażone w postaci logarytmu szansy) są podstawą do korekty wcześniejszych prognoz (wyrażonych w analogiczny sposób) dla wszystkich przypadków w zbiorze danych za pomocą wzoru [Starmer, 2019]:

$$Pred_i = Pred_{i-1} + \nu \times Pred\ drzewa_i$$

Współczynnik ν nazywany jest współczynnikiem uczenia, bądź współczynnikiem redukcji (*shrinkage*) i jest on zwykle ustalany na poziomie 0,1. Ma to na celu zapewnić, że każde kolejne drzewo jedynie w niewielkim stopniu poprawi predykcję w stosunku do zastanego wyniku.

Powyższe kroki powtarzane są wielokrotnie do momentu zbudowania zadanej liczby drzew. Możliwe jest wzbogacenie algorytmu o mechanizm wczesnego zatrzymania w momencie gdy wartości reszt są bardzo małe tj. mniejsze od ustalonego przez użytkownika progu. Dodatkowo liczba drzew w modelu może zostać ograniczona na podstawie analizy działania modelu na zbiorze walidacyjnym. W przypadku zastosowania tej techniki do oceny zdolności modeli do generalizacji konieczne staje się użycie trzeciego zbioru (testowego). Wynikowy model składa się z prognoz wszystkich modeli składowych. Końcowa predykcja jest przekształcana do postaci prawdopodobieństwa za pomocą funkcji logistycznej w sposób analogiczny do realizowanego podczas budowy drzew składowych.

Popularną implementacją algorytmu wzmacniania drzew jest biblioteka *eXtreme Gradient Boost* (*XGBoost*) przeznaczona do budowy modeli uczenia maszynowego, której użytkownicy odnoszą sukcesy w konkursach analitycznych na najlepszy model predykcyjny. Proces budowy modelu przebiega w analogiczny sposób do wzmacniania gradientowego. Po inicjalizacji modelu za pomocą logarytmu szansy (domyślnie na poziomie 0), a następnie przeliczeniu wartości na prawdopodobieństwa za pomocą funkcji logistycznej obliczane są reszty, które pełnią rolę zmiennej zależnej. W odróżnieniu do wcześniejszego algorytmu, w którym wykorzystywane było klasyczne drzewo regresyjne CART w przypadku *XGBoost* używana jest miara czystości węzła o nazwie punktacja podobieństwa (*similarity score*) obliczana za pomocą wzoru [Starmer, 2020]:

$$similarity\ score = \frac{\left(\sum_{i=1}^{N} r_i\right)^2}{\sum_{i=1}^{N}[p_i \times (1 - p_i)] + \lambda},$$

gdzie N oznacza liczbę przypadków w danym liściu, r_i oznacza wartość zmiennej zależnej dla i-tego przypadku w liściu, p_i oznacza prawdopodobieństwo bycia nielojalnym klientem

obliczone w poprzednim kroku algorytmu dla i-tego przypadku w liściu. Wyrażenie λ jest współczynnikiem regularyzacji, który ma na celu zmniejszenie wrażliwości modelu na wpływ pojedynczych obserwacji. Redukuje ona również wartość *similarity score*. Należy zwrócić uwagę, że wartości sumowane w liczniku mogą mieć różne znaki, redukując się tym samym. Im czystszy węzeł, tym większa będzie w nim przewaga przypadków o tych samych znakach, a tym samym większa wartość wyrażenia. Podziału dokonywany jest dla predyktora oraz punktu podziału maksymalizującego współczynnik korzyści (*gain*), obliczany na podstawie formuły [Starmer, 2020]:

$$gain = similarity\ score_{left} + similarity\ score_{right} + similarity\ score_{parent}$$

Kolejną różnicą w stosunku do klasycznego modelu CART jest sposób zatrzymania procesu uczenia. W przypadku drzew budowanych za pomocą biblioteki *XGBoost* każdy potomek musi charakteryzować się współczynnikiem pokrycia (*cover*), obliczanym jako [Starmer, 2020]:

$$cover = \sum_{i=1}^{N} [p_i \times (1 - p_i)]$$

większym od ustalonej przez użytkownika wartości (domyślnie 1). Po zbudowaniu modelu następuje etap jego przycinania. Na podstawie określonego współczynnika γ przycina się gałęzie, dla których *gain* jest od niego mniejszy. Wartość prognozowaną dla liścia obliczamy jako [Starmer, 2020]:

$$wartość\ prognozowana\ dla\ liścia = \frac{\sum_{i=1}^{N} r_i}{\sum_{i=1}^{N} [p_i \times (1 - p_i)] + \lambda}$$

Wyrażenie to jest analogiczne do używanego w algorytmie wzmacniania gradientowego. Różni się jedynie współczynnikiem regularyzacji, który zmniejsza wartość oszacowanej prognozy. Korekta wyniku jest realizowana w sposób analogiczny do klasycznego algorytmu wzmacniania gradientowego.

Algorytm wzmacniania gradientowego może być dodatkowo rozbudowywany o elementy znane z losowego lasu:

- losowanie zmiennych przed zbudowaniem każdego z drzew składowych,
- losowanie przypadków (ze zwracaniem) przez zbudowaniem każdego z drzew składowych.

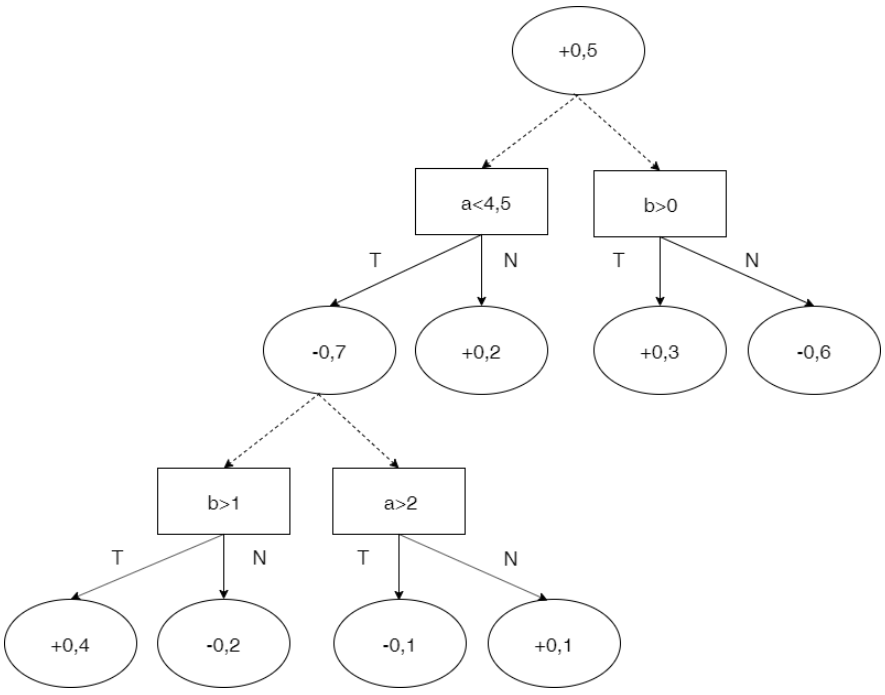
Obydwie techniki mają na celu uniknięcie nadmiernego dopasowania modelu do danych. Jak twierdzą niektórzy autorzy [Chen, Guestrin, 2016] bardziej skuteczną techniką jest losowanie zmiennych, która skraca czas obliczeń.

Kolejną techniką pozwalającą na redukcję ryzyka nadmiernego dopasowania modelu do danych w drzewach wzmacnianych jest DART (*Dropouts meet multiple Additive Regression Trees*) [Rashimi, Gilad-Bachrach, 2015]. Wykorzystuje ona koncepcję znaną z metody *dropout* stosowanej w sieciach neuronowych. Autorzy zauważyli, iż pomimo stosowania współczynnika redukcji, kolejne drzewa dodawane do modelu w końcowych iteracjach wpływają jedynie na poprawę predykcji niewielu przypadków. Z drugiej strony wpływ początkowych drzew na ostateczną postać modelu jest bardzo duży. Obydwa te fakty negatywnie świadczą o zdolności modelu do generalizacji. Zastosowana tutaj technika polega na wyłączaniu działania losowo wybranych drzew podczas kolejnych iteracji budowy modelu. Wkład nowo zbudowanego drzewa utworzonego na podstawie niepełnego zestawu wcześniej zbudowanych drzew jest następnie skalowany, podobnie jak wkład drzew usuniętych z modelu w danej iteracji. Autorzy wykazali, że model zbudowany w ten sposób cechuje się większym zbalansowaniem wkładu poszczególnych drzew składowych oraz wysoką skutecznością zbudowanego w ten sposób modelu.

Relatywnie nową metodą budowy zespołu modeli jest metoda RGF (*Regularized Greedy Forest*) przedstawiona w 2014 roku [Johnson, Zhang]. Podejście to jest podobne do opisywanych powyżej metod wzmacniania. Różni się od standardowego modelu wzmacniania gradientowego tym, iż podczas rozbudowy modelu rozpatrywana jest zarówno możliwość wprowadzenia nowego podziału do bieżącego drzewa, jak również stworzenie nowego drzewa. W związku z tym w algorytmie tym nie zakłada się maksymalnej głębokości drzewa. Kolejną różnicą jest sposób optymalizacji wag modelu. Podejście stosowanie w drzewach wzmacnianych polega na korekcie wag jedynie dla modelu budowanego w bieżącym etapie. Są one dopasowywane, aby możliwie najlepiej przybliżyć rozwiązanie przy założeniu stałości wag dla dotychczas uzyskanych liści (strategia *stagewise*). W algorytmie RGF stosowana jest globalna optymalizacja wszystkich wag modelu. Ze względu na wysokie wymagania obliczeniowe autorzy metody zalecają wykonywanie tego kroku co około 100 cykli rozbudowy modelu. Istotnym elementem algorytmu jest regularyzacja wag pozwalająca na zmniejszenie wrażliwości modelu na przeuczenie.

Duży rozmiar oraz czarno-skrzynkowy charakter drzew wzmacnianych skłoniły badaczy do opracowania tzw. zmiennych drzew decyzyjnych (*Alternating Decision Tree* –

ADTree). W zamierzeniu autorów [Freund, Mason, 1999] miały być uogólnieniem zarówno drzew klasyfikacyjnych jak również uogólnieniem zespołu głosujących pniaków (*decission stumps*) przedstawionym w opisie algorytmu *AdaBoost*. Algorytm *ADTree* choć oparty na ogólnej koncepcji drzew klasyfikacyjnych, ma szereg cech istotnie go od nich różniących. Pierwszą zasadniczą różnicą jest sposób rozbudowy drzewa. W klasycznym podejściu rozbudowie podlegają jedynie węzły końcowe (liście). W przypadku *ADTree* można rozbudowywać również węzły macierzyste. Zatem z jednego węzła macierzystego może zostać utworzonych wiele podziałów charakteryzujących się różnymi warunkami. Drugą różnicą jest sposób tworzenia reguł na podstawie zbudowanego drzewa. W klasycznym drzewie klasyfikacyjnym reguła decyzyjna dla danego przypadku jest zestawem koniunkcji warunków opisujących drogę od korzenia drzewa do liścia. W przypadku metody *ADTree* na decyzję składa się suma reguł wspierających dany przypadek. Przy czym pojedyncza reguła może kończyć się zarówno w liściu, jak i w węźle macierzystym. Przykład drzewa *ADTree* zbudowanego dla dwóch predyktorów *a* oraz *b* przedstawia Rysunek 20.



Rysunek 20 Przykładowy model ADTree
Źródło: Opracowanie własne na podstawie [Freund, Mason, 1999].

Dla przykładu przypadku, w którym zmienna a = 2 oraz zmienna b=2 klasyfikacja będzie sumą następujących reguł:

Tabela 7 Reguły klasyfikacji dla przypadku a=2 oraz b=2

Reguła	Wynik reguły
Jeżeli (Prawda)	+0,5
Jeżeli (Prawda) i Jeżeli (a<4,5)	-0,7
Jeżeli (Prawda) i Jeżeli (b>0)	+0,3
Jeżeli (Prawda) i Jeżeli (a<4,5) i Jeżeli (b>1)	+0,4
Jeżeli (Prawda) i Jeżeli (a<4,5) i Jeżeli nie (a>2)	+0,1

Źródło: opracowanie własne.

Wynikowa klasyfikacja jest określona za pomocą funkcji signum, której argumentem jest suma wyników składowych reguł. Dodawanie kolejnych składowych drzewa realizowane jest w sposób analogiczny do drzew wzmacnianych na przykład za pomocą algorytmu *AdaBoost*. Niewątpliwą zaletą tej metody jest uzyskiwanie mniej skomplikowanych modeli w porównaniu do klasycznych drzew wzmacnianych. Inną zaletą jest fakt, iż każda reguła decyzyjna jest z góry znana i może być rozpatrywana w oderwaniu od pozostałych [Po-Leen Ooi i inni, 2017]. Algorytm *ADtree* podlegał wielu modyfikacjom co skutkowało pojawianiem się opartych na nim metodach *Fisher's ADTree*, *Sparse ADTree* czy *regularized logistic ADTree*, których opis można znaleźć na przykład w pracy z 2017 roku [Po-Leen Ooi i inni]. Algorytm ten był również inspiracją dla twórców opisywanej metody RGF, którzy zwracali uwagę na jego słabą stronę, jaką był brak mechanizmu regularyzacji [Johnson, Zhang, 2014].

Opisane powyżej strategie agregacji modeli zakładały, że zespół tworzony jest przez modele składowe zbudowane za pomocą tej samej metody. Odmienną strategią jest łączenie ze sobą modeli zbudowanych za pomocą różnych metod. Podejście takie nazywamy hybrydowym.

Popularną strategią budowy modelu hybrydowego jest połączenie drzew klasyfikacyjnych CART z regresją logistyczną opisywane na przykład w [Łapczyński, 2016, Siddiqui, 2017]. Drzewa decyzyjne budowane są w celu wyodrębnienia bardziej jednorodnych ze względu na ryzyko odejścia grup klientów. Wyodrębnione podzbiory służą do budowy niezależnych od siebie modeli regresji logistycznej. Podejście to jest jedną z bardziej skutecznych metod budowy modeli w sytuacji, gdy oczekiwana jest możliwość jego interpretacji. Sprawdza się zwłaszcza w sytuacji, gdy wśród predyktorów można wyróżnić grupę silnie powiązanych ze zmienną zależną. Ich moc predykcyjna zostanie wtedy wykorzystana w podziałach modelu CART, dzięki czemu nie zdominują one modeli regresji, w których uwzględnione mogłyby zostać pozostałe predyktory.

Chociaż znane są implementacje automatyzujące proces budowy tego typu modeli[92], ich budowa zazwyczaj wymaga jednak pewnej interakcji z badaczem w celu określania optymalnej postaci drzewa.

3.3. Optymalizacja hiperparametrów

Ważnym czynnikiem wpływającym na końcową jakość budowanego modelu jest odpowiednie określenie hiperparametrów (ustawień) wybranej metody analitycznej. Praktycznie każda metoda analityczna ma hiperparametry, których wartości na końcową postać modelu. Jakość dopasowania modelu może znacząco różnić się w zależności od ustawień przyjętych dla danej metody przez badacza. Pewne ustawienia początkowe skutkować będą lepiej dopasowanymi modelami a inne będą prowadziły do niezadowalających wyników. W związku z tym proces wyboru ustawień również może podlegać i często podlega optymalizacji.

Proces optymalizacji hiperparametrów (*hyperparameter optimization*, HPO) powiązany jest z uwarunkowaniami, które powodują, że stanowi on wyzwanie dla badaczy. Optymalizowane hiperparametry mogą być określane na różnych skalach pomiarowych i dodatkowo nieznany jest wpływ poszczególnych ustawień na jakość modelu oraz ich wzajemne interakcje. Nie jest też zazwyczaj możliwe użycie metod gradientowych w cel ich optymalizacji, dodatkowo optymalizowana funkcja nie jest ani wypukła ani gładka [Feurer, Hutter, 2019]. W efekcie stosowane metody najczęściej generują rozwiązania suboptymalne w ramach dostępnego czasu lub mocy obliczeniowej. Idealna metoda przeszukiwania hiperparametrów powinna [Falkner i inni, 2018]:

- pozwalać na określanie dobrych konfiguracji przy ograniczonym budżecie związanym z mocą obliczeniową i czasem obliczeń,
- poświęcać większą część budżetu bardziej obiecującym konfiguracjom,
- umożliwiać zrównoleglenie obliczeń,
- skalować się nawet dla dużej liczby hiperparametrów,
- cechować się elastycznością, wyrażającą się poprzez możliwość stosowania danej metody przeszukiwania dla różnych metod analitycznych oraz różnych rodzajów

[92] Na przykład w programie Weka zaimplementowana jest metoda *logistic model trees* [Landwehr i inni, 2005] realizująca tę strategię.

hiperparametrów (na przykład binarnych, jakościowych, ciągłych czy warunkowych[93]).

Znane metody przeszukiwania realizują te postulaty w różnym i zazwyczaj ograniczonym stopniu. Najbardziej podstawową i intuicyjną metodą przeszukiwania przestrzeni hiperparametrów jest metoda ręcznej zmiany ustawień algorytmu. Sprawdza się w sytuacji, gdy liczba możliwych ustawień jest relatywnie niewielka a badacz posiada już pewną wiedzę i doświadczenie umożliwiające ukierunkowane przeszukiwanie poziomów dostępnych opcji. Jest to jednak podejście zawodne, czasochłonne oraz bardzo słabo skalujące się w sytuacji wzrostu liczby hiperparametrów. Na przestrzeni lat opracowano szereg metod pozwalających na automatyzację procesu doboru optymalnego zestawu ustawień. Pierwsza grupa metod to tak zwane metody niezależne od modelu (*model free methods*), do których zaliczyć można [Ryan, 2019]:

- systematyczne przeszukiwanie przestrzeni hiperparametrów na podstawie przyjętej siatki możliwych rozwiązań (*grid seach*),
- losowe przeszukiwanie przestrzeni hiperparametrów (*random search*),
- przeszukiwanie heurystyczne,
- przeszukiwanie oparte na algorytmach genetycznych i pokrewnych metodach,
- algorytmy wielu rozdzielczości (*multi-resolution algorithms*).

Poszukiwanie na podstawie przyjętej siatki możliwych rozwiązań (*grid search*) sprawdza się wszystkie możliwe kombinacje wartości hiperparametrów przy założonej granulacji kroku dla każdego z nich. Wadą rozwiązania jest konieczność sprawdzenia dużej liczby możliwych rozwiązań, zaletą niezależność poszczególnych kroków przeszukiwania co umożliwia zrównoleglenie obliczeń. Działa dobrze dla niewielkiej liczby ustawień. J. Bergstra oraz Y. Bengio [2012] wykazali, że ręczne przeszukiwanie przestrzeni hiperparametrów oraz poszukiwanie po siatce są mniej efektywne od przeszukiwania losowego. Nieefektywność przeszukiwania po siatce wynika z faktu, że najczęściej na końcowy efekt ma relatywnie niewielka liczba hiperparametrów[94] a wraz ze wzrostem liczby hiperparametrów wykładniczo rośnie liczba ich kombinacji. Przeszukiwanie po siatce przy danej liczbie powtórzeń pozwala zatem sprawdzić mniejszą

[93] Przykładem hiperparametru warunkowego może być wielkość kroku uczenia dla sieci neuronowych. Niektóre algorytmy uczenia umożliwiają jego określenia inne nie (np. algorytm BFGS). Zatem wybór algorytmu warunkuje ewentualną konieczność określenia kolejnego hiperparametru.

[94] Dodatkowo dla różnych zbiorów danych zestaw znaczących hiperparametrów jest inny.

liczbę ustawień kluczowych hiparparametrów ze względu na systematyczne przeszukiwanie również w przestrzeni nieistotnych ustawień. Podejście losowe testuje przy tej samej liczbie powtórzeń większą liczbę kombinacji znaczących hiperparametrów. Zaletą losowego przeszukiwania przestrzeni hiperparametrów jest niewątpliwie łatwość z jaką można zrównoleglić tę metodę. Jest ona też dobrym punktem wyjścia do stosowania na jej podstawie bardziej zaawansowanych metod optymalizacji. Wadą jest konieczność wykonania dużo większej liczby powtórzeń niż bardziej zaawansowane metody pozwalające na korektę działania na podstawie wcześniej uzyskanych wyników.

Kolejną grupą metod niezależnych od modelu są algorytmy oparte na heurystycznym przeszukiwaniu przestrzeni hiperparametrów. Można do nich zaliczyć algorytmy takie jak algorytm mrówkowy (*Ant Colony Optimization)* oraz optymalizacja rojem cząstek (*Particle Swarm Optimization)*. Inną klasą takich metod są algorytmy genetyczne i im pokrewne, które do populacji istniejących rozwiązań wprowadzają losowe zmiany (mutacje) oraz kombinacje istniejących rozwiązań (krzyżówki) w celu uzyskania zestawu (populacji) lepszych konfiguracji. [Feurer, Hutter, 2019] wskazują na algorytm CMA-ES (*Covariance Matrix Adaption Evolutionary Strategy*) jako jeden z najbardziej konkurencyjnych strategii doboru hiperparametrów. Działanie algorytmu polega na losowaniu nowych konfiguracji z wielowymiarowego rozkładu normalnego, którego parametry są aktualizowane na podstawie sukcesów przypadków z wcześniejszego pokolenia.

Ostatnią grupą metod niezależnych od modelu są algorytmy wielu rozdzielczości. Reprezentantem tej grupy jest sukcesywne przepoławianie (*succesive halving*), które opiera się na następujących założeniach:

- pozwala na wczesne zatrzymanie procesu uczenia i podanie przybliżonej wartości dopasowania modelu,
- pozwala na równoległe testowanie wielu konfiguracji hiperparametrów
- pozwala określić budżet związany z czasem obliczeń oraz zaangażowaniem mocy obliczeniowej,
- po zrealizowaniu określonej frakcji zadanego budżetu zachowuje jedynie połowę najlepszych konfiguracji poświęcając im proporcjonalnie więcej zaplanowanego budżetu,

– powyższa czynność jest powtarzana, aż do zrealizowania całości zaplanowanego budżetu.

Podejścia to pozwala na poświęcenie większej części budżetu na bardziej obiecujące konfiguracje. Jego wadą jest zależność od liczby punktów przepoławiania[95]. Im większa liczba punktów przepoławiania tym większa liczba konfiguracji może zostać sprawdzona, jednak wraz ze wzrostem punktów rośnie ryzyko odrzucenia obiecującej konfiguracji. *HyperBand* [Li i inni, 2017] jest rozwinięciem metody sukcesywnego przepoławiania w swoim zamyśle redukującą wadę konieczności określania liczby podziałów na pół w obrębie budżetu. Równolegle wykonuje ona kilka eksperymentów opartych na metodzie sukcesywnego przepoławiania dla różnych jego ustawień od niewielkiej (zbliżającą to podejście do losowego przeszukiwania) do znacznej liczby punktów przepoławiania.

Strategie niezależne od modelu są elastyczne i skalowalne nie są jednak efektywne. Wymagają wykonania znacznej liczby eksperymentów co pociąga za sobą również konieczność zaangażowania znacznej mocy obliczeniowej w relatywnie długim czasie. Przyjętą praktyką mającą na celu zwiększenie efektywności jest wczesne przerywanie procesu uczenia. Skutkuje to uzyskaniem jedynie przybliżonej oceny jakości danej konfiguracji, jednak w praktyce akceptowalną dla badacza.

Druga grupa metod optymalizacji hiperparametrów to metody oparte na modelu (*model based methods*). Podejście to wymaga wykonania wstępnej liczby eksperymentów metodami niezależnymi od modelu. Wynik tych eksperymentów są podstawą do rozpoczęcia procesu optymalizacji przeszukiwania nowego zestawu hiperparametrów. Wykonane eksperymenty są zatem niejako przypadkami uczącymi, predyktorami są hiperparametry a zmienną zależną jest ocena jakości modelu dla danych ustawień.

Najbardziej znaną metodą należącą do tej grupy jest optymalizacja Bayesowska. Jest stosowana do optymalizacji dowolnej funkcji, dla której nieznana jest jej postać analityczna (nie można zatem stosować optymalizacji gradientowej i im podobnych). W przypadku modelowania hiperparametry traktowane są jako argumenty funkcji, wartością funkcji jest przyjęta miara dobroci dopasowania. Start algorytmu optymalizacji Bayesowskiej wymaga wykonania pewnej liczby modeli na podstawie losowych ustawień, które będą podstawą do wykonania pierwszej iteracji optymalizacji hiperparametrów. Ustawienia dla zbudowanych modeli są traktowane jako przypadku uczące. Na ich podstawie wskazywany jest nowy układ ustawień, dla których spodziewana jest poprawa

[95] Liczba ta jest zatem hiperparametrem metody wyszukiwania hiperparametrów.

działania budowanego modelu. Algorytm ze swojej natury działa w sposób sekwencyjny co w przeciwieństwie do wcześniej przedstawionych metod uniemożliwia zrównoleglenie wykonywania obliczeń. Podejście to ma sens w sytuacji, gdy liczba hiperparametrów jest relatywnie duża i jednocześnie budowa i ocena pojedynczego modelu zajmuje relatywnie dużo czasu.

Obiecującą syntezą obydwóch podejść jest algorytm BOHB (*Bayesian Optimization and HyperBand*) przedstawiony w [Falkner i inni, 2018]. W pierwszym etapie algorytm realizuje strategię *HyperBand*, zapamiętując wszystkie wylosowanie zestawy hiperparametrów oraz powiązane z nimi oceny dopasowania niezależnie od momentu, w którym uczenie dla danej konfiguracji zostało przerwane. Na podstawie uzyskanych konfiguracji realizowana jest optymalizacja Bayesowska w celu określenia nowego układu ustawień. Podejście to pozwoliło według autorów uzyskać znaczące skrócenie czasu przeszukiwania[96] oraz poprawić jakość dopasowania zbudowanych modeli w stosunku do obydwóch podejść składających się na tę hybrydową strategię.

Niewątpliwie zaletą automatycznego przeszukiwania hiperparametrów jest [Feurer, Hutter, 2019]:

- redukcja czasu, jaki badacz musi poświęcić na identyfikację zadowalającego rozwiązania,
- poprawa jakości zbudowanych modeli,
- porównywalność metod analitycznych oraz odtwarzalność uzyskanych wyników niemożliwa do uzyskania podczas ręcznego doboru.

Wybór najbardziej odpowiedniej metody zależy w dużej mierze od czasu jaki należy poświęcić jednej iteracji budowy modelu. W przypadku, gdy budowa jednego modelu trwa krótko z punktu widzenia badacza, skuteczną strategią są metody wolne od modelu. Przy bardziej skomplikowanych układach zasadne może być stosowanie algorytmu CMA-ES. Metody oparte na modelu należy rozważyć w sytuacji, gdy czas budowy jednego modelu jest relatywnie długi i nie jest możliwe wykonanie setek czy tysięcy powtórzeń.

[96] W porównaniu do przeszukiwania losowego uzyskano ponad 55 krotną redukcję czasu.

Rozdział 4 Walidacja i wdrażanie modeli retencji klientów

4.1. Miary dobroci dopasowania modeli retencji klientów

Jednym z najważniejszych etapów cyklu życia modelu predykcyjnego jest etap jego oceny (walidacji). Walidację przeprowadza się bezpośrednio po zakończonym procesie budowy modelu na podstawie próby testowej lub za pomocą bardziej zaawansowanych technik opartych na wielokrotnym próbkowaniu. Ocena na tym etapie życia modelu nosi miano walidacji *ex ante*. Do jej głównych zadań należy:

- porównanie konkurencyjnych modeli zbudowanych różnymi metodami lub przy użyciu innego zestawu hiperparametrów[97] i wybór modelu najlepiej spełniającego kryteria biznesowe (niekoniecznie posiadającego najwyższe wskaźniki dopasowania),
- potwierdzenie, że najlepszy ze zbudowanych modeli jest odpowiedniej jakości i może zostać wdrożony w środowisku informatycznym.

Drugim momentem cyklu życia modelu, w którym niezbędna staje się jego ocena jest okres po wdrożeniu modelu. Może to być na przykład okres kilku miesięcy po jego wdrożeniu. Walidację taką wykonuje się na przychodzącej populacji klientów. Ocena na tym etapie życia modelu nosi miano walidacji *ex post*. Głównym jej celem jest wykazanie, że działający w praktyce model nie stracił dotychczasowej siły predykcyjnej i może być nadal stosowany. Wnioski przeciwne mogą prowadzić do decyzji o przebudowie istniejącego modelu, co zamyka *de facto* cykl życia modelu. Poniżej przedstawiono miary

[97] Hiperparametry to ustawienia algorytmów, wpływające na sposób działania danej metody. Są one ustalane przez badacza, w przeciwieństwie do parametrów, ustalanych przez algorytm w trakcie procesu nauki. Dla sieci neuronowych przykładowymi hiperparametrami będą: liczba warstw sieci, rodzaj funkcji aktywacji, czy liczba neuronów w danej warstwie. Dla drzew klasyfikacyjnych i regresyjnych będą to: maksymalna głębokość drzewa czy też koszty błędnych klasyfikacji.

dobroci dopasowania (miary siły dyskryminacyjnej, *goodness of fit measures, performance measures*). Mogą one być stosowane zarówno w wypadku walidacji *ex ante*, jaki i *ex post*.

Klasyfikacyjne modele lojalności po zakończonym procesie uczenia umożliwiają prognozowanie skłonności klientów do odejścia do konkurencji. Skłonność ta jest wyrażana wartością z przedziału od 0 do 1. W sytuacji gdy:

- rozkład klas zmiennej zależnej w zbiorze uczącym był zgodny z obserwowanym odsetkiem osób nielojalnych (prawdopodobieństwem *a priori*);
- do budowy modelu użyto metody zachowującej prawdopodobieństwo *a priori* modelowanego zdarzenia (np. regresja logistyczna, naiwna metoda Bayesa);

wartość ta może traktowana jako ocena prawdopodobieństwa zajścia modelowanego zdarzenia (prawdopodobieństwo *a posteriori*). Zazwyczaj jednak bezpieczniej jest założyć odstępstwo od tych wymogów i traktować uzyskany wynik jako miarę rangującą analizowanych klientów względem ich skłonności do odejścia[98]. Odpowiedź modelu poddawana jest następnie dyskretyzacji. Dla uzyskanych wyników wprowadzany jest punkt odcięcia (*cut-off point*, punkt graniczny) co pozwala na uzyskanie binarnej odpowiedzi modelu informującej o klasyfikacji danego przypadku do kasy „lojalny" bądź „nielojalny".

Powyższe rozróżnienie pozwala na wprowadzenie klasyfikacji miar siły predykcyjnej na trzy grupy [Ferri i inni, 2009, Berrar, 2019]:

- bazujące na macierzy błędnych klasyfikacji (po wprowadzeniu punktu odcięcia),
  - podstawowe,
  - złożone,
- przyjmujące, że odpowiedzi modelu należy traktować jako zmienną rangującą,
- oparte na probabilistycznej interpretacji odpowiedzi modelu.

W przypadku modeli lojalności najczęściej stosuje się miary należące do dwóch pierwszych grup. Miary oparte na probabilistycznej interpretacji odpowiedzi modelu są stosowane rzadziej.

[98] Uzyskane wartości można poddać procesowi kalibracji, aby nadać im probabilistyczną interpretację. Jedna z metod kalibracji została przedstawiona w dalszej części rozdziału.

4.1.1. Miary obliczone na podstawie macierzy błędnych klasyfikacji

Popularnymi i prostymi miarami pozwalającymi ocenić jakość zbudowanego klasyfikatora są miary oparte na macierzy błędnych klasyfikacji (*confusion matrix, misclassification matrix*). W macierzy tej porównujemy stan faktyczny z prognozą uzyskaną na podstawie modelu. W niniejszym opracowaniu przyjęto konwencję, że wartość „1" symbolizującą wystąpienie zdarzenia (*event*) przypisano klientom nielojalnym, „0" (*non-event*) przypisano klientom lojalnym.

Tabela 8 Macierz błędnych klasyfikacji

	Obserwowane 1 (Nielojalni)	Obserwowane 0 (Lojalni)	Suma
Przewidywane 1 (Nielojalni)	TP	FP	PP
Przewidywane 0 (Lojalni)	FN	TN	PN
Suma	RP	RN	N

Źródło: opracowanie własne.

Poszczególne pola w tabeli oznaczają odpowiednio:

- TP (*True Positives*) – liczba przypadków prawdziwie pozytywnych[99], czyli klientów, którzy zdecydowali się na odejście do konkurencji i jednocześnie zostali poprawnie wskazani przez model jako nielojalni;
- FN (*False Negatives*) – liczba klientów, którzy zdecydowali się na odejście do konkurencji i jednocześnie model błędnie zaliczył ich do grupy osób lojalnych;
- FP (*False Positives*)– liczba lojalnych klientów, którzy w sposób niepoprawny zostali uznani przez model za nielojalnych;
- TN (*True Negatives*) – liczba lojalnych klientów, którzy zostali przez model poprawnie zaklasyfikowani jako lojalni.

Wartości brzegowe można interpretować jako:

- RP (*Real Positives*) – liczba nielojalnych klientów,
- RN (*Real Negatives*) – liczba lojalnych klientów,
- PP (*Predicted Positives*) – osoby wskazane przez model jako nielojalne,

[99] Słowo „pozytywny" użyte w tym kontekście nie musi oznaczać stanu pożądanego przez badacza. Na przykład pozytywny wynik testu na obecność choroby nie wiąże się z pożądanym przez pacjenta wynikiem.

– PN (*Predicted Negatives*) – osoby wskazane przez model za lojalne,

– N – liczba wszystkich klientów.

Tabela 8 jest podstawą do definiowania wielu miar jakości klasyfikatora. W Tabela 9 przedstawiono przegląd miar wykorzystywanych w procesie oceny modelu.

Tabela 9 Przegląd miar jakości modelu dla macierzy błędnych klasyfikacji

Nazwa	Wzór	Interpretacja
Dokładność klasyfikacji (*Accuracy*) – ACC	$\frac{TP+TN}{N}$	Odsetek przypadków poprawnie zaklasyfikowanych przez model
Błąd klasyfikacji (*Error rate*) – ER	$\frac{FN+FP}{N}$	Odsetek przypadków niepoprawnie zaklasyfikowanych przez model
Czułość (Sensitivity) – SENS, Recall, TPR (True Positive Rate)	$\frac{TP}{RP}$	Odsetek przypadków pozytywnych poprawnie zaklasyfikowanych przez model
Specyficzność (Specificity) – SPEC, TNR (True Negative Rate)	$\frac{TN}{RN}$	Odsetek przypadków negatywnych poprawnie zaklasyfikowanych przez model
Zrównoważona dokładność klasyfikacji (Balanced accuracy) – BACC	$\frac{SENS+SPEC}{2}$	Średnia arytmetyczna czułości i specyficzności
Wartość predykcyjna pozytywnego wyniku (Positive predictive value) – PPV, Precision	$\frac{TP}{PP}$	Odsetek przypadków pozytywnych w grupie uznanych przez model za pozytywne,
Wartość predykcyjna negatywnego wyniku (Negative predictive value) – NPV	$\frac{TN}{PN}$	Odsetek przypadków negatywnych w grupie uznanych przez model za negatywne,
Wskaźnik fałszywie pozytywnych (False positive rate) – FPR, Fallout	$\frac{FP}{RN}$	Odsetek przypadków pozytywnych błędnie zaklasyfikowanych przez model (błąd I rodzaju)
Wskaźnik fałszywie negatywnych (False negative rate) – FNR	$\frac{FN}{RP}$	Odsetek przypadków negatywnych błędnie zaklasyfikowanych przez model (błąd II rodzaju)
Dodatni iloraz wiarygodności (+Likelihood ratio) - LR (+)	$\frac{SENS}{1-SPEC} = \frac{TPR}{FPR}$	O Ile wzrasta szansa rzeczywistego wystąpienia stanu pozytywnego, w sytuacji gdy model prognozuje stan pozytywny

Nazwa	Wzór	Interpretacja
Ujemny iloraz wiarygodności (-Likelihood ratio) - LR (-)	$\frac{1-SENS}{SPEC} = \frac{FNR}{TNR}$	O ile maleje szansa wystąpienia stanu pozytywnego, w sytuacji gdy model prognozuje stan negatywny.
Indeks Youden'a J (Informedness)	$SENS + SPEC - 1$	Syntetyczna miara jakości klasyfikacji. Wartość 0 informuje o braku siły predykcyjnej klasyfikatora, 1 o idealnej klasyfikacji.
Miara F (F -score, F-measure)	$2 * \frac{SENS * PPV}{SENS + PPV}$	Syntetyczna miara jakości klasyfikacji. Wartość 0 informuje o braku siły predykcyjnej klasyfikatora, 1 o idealnej klasyfikacji. Średnia harmoniczna wskaźników SENS i PPV[100]
G – mean (Fowlkes–Mallows index)	$\sqrt{SENS * PPV}$	Średnia geometryczna wskaźników SENS i PPV.
Przyrost (Lift)	$\frac{PPV}{\frac{RP}{N}}$ lub $\frac{NPV}{\frac{RN}{N}}$	Informuje o stopniu poprawy jakości klasyfikacji w stosunku do losowego wyboru przypadków. Jest ilorazem odsetka przypadków pozytywnych w grupie uznanych przez model za pozytywne (PPV) przez odsetek pozytywnych (osób nielojalnych) w całym zbiorze danych. Może być również liczony dla klasy negatywnej (drugi wzór).
Współczynnik korelacji Matthewsa (Matthews correlation coefficient) - MCC	$\frac{TP * TN - FP *}{\sqrt{PP * RP * RN}}$	Bierze pod uwagę wszystkie składowe macierzy pomyłek. Wartość 0 informuje o braku siły predykcyjnej, wartość 1 informuje o idealnej klasyfikacji.
Wyrazistość[101] (Markedness) - MK	$PPV + NPV - 1$	Syntetyczna miara jakości klasyfikacji. Wartość 1 informuje o idealnej klasyfikacji. Wrażliwa na niezbalansowany rozkład przypadków pozytywnych i negatywnych [Powers, 2011]

[100] W przedstawionej postaci wzór zakłada wkład wskaźników SENS oraz PPV w równych proporcjach.

[101] Określenie „wyrazistość" to propozycja tłumaczenia wprowadzona przez autora. W polskiej literaturze nie znaleziono odpowiednika.

Nazwa	Wzór	Interpretacja
Indeks Jaccarda	$\frac{TP}{TP + FP + FN}$	Ignoruje przypadki prawdziwie negatywne TN. Wrażliwy na występowanie niezbalansowanego rozkładu przypadków pozytywnych i negatywnych [Tharwat, 2018].

Źródło: opracowanie własne na podstawie [Łapczyński, 2016,Tharwat, 2018, Kuhn, Johnson, 2013, Berrar, 2019].

4.1.2. Miary jakości a prawdopodobieństwo *a posteriori*

Należy zwrócić uwagę, że obliczenie powyższych miar dla modelu klasyfikacyjnego jest możliwe jedynie w sytuacji, gdy badacz określi punkt odcięcia wynikowego prawdopodobieństwa *a posteriori*[102]. W wielu programach analitycznych (Statistica, IBM SPSS, Rapid Miner) punkt ten jest wyznaczany w sposób automatyczny na poziomie 0,5 co niekoniecznie musi być wartością optymalną z punktu widzenia celu modelowania W zależności od zaproponowanego punktu odcięcia wartości poszczególnych miar będą ulegały zmianie.

Dla każdej z powyższych miar można zatem utworzyć wykres ich wartości w zależności od wartości prawdopodobieństwa *a posteriori* analizowanego modelu. Poniżej przedstawiono ilustracje wybranych miar siły predykcyjnej w przekroju wartości prawdopodobieństw *a posteriori* z uwzględnieniem sytuacji, w której zbiór służący do oceny modelu zawiera:

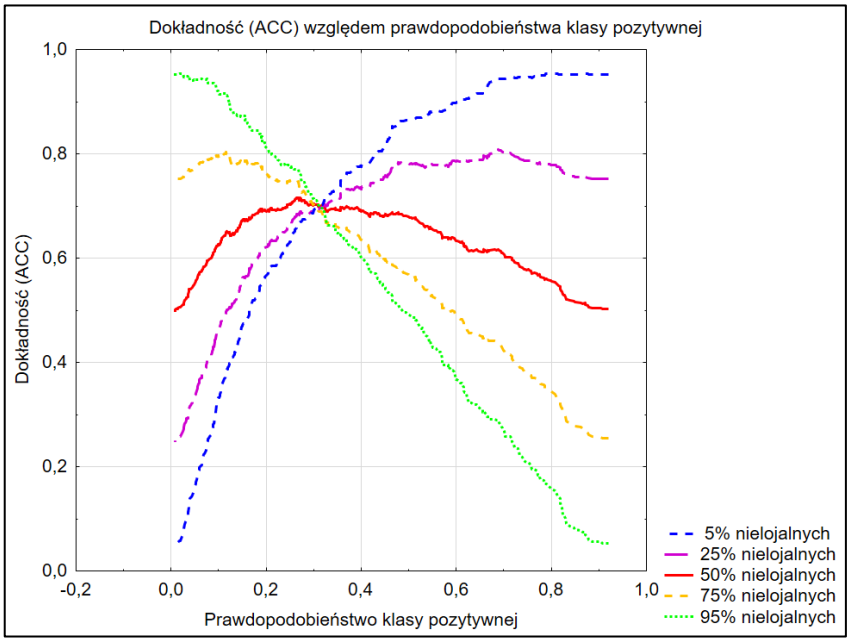
- 5% przypadków osób nielojalnych,
- 25% przypadków osób nielojalnych,
- 50% przypadków osób nielojalnych,
- 75% przypadków osób nielojalnych,
- 95% przypadków osób nielojalnych.

Zbiór, na podstawie którego budowany był model zawierał 30% przypadków osób nielojalnych. Model zbudowano za pomocą regresji logistycznej ze względu na jej niewielką wrażliwość na nierówne proporcje klas zmiennej zależnej. Dodatkowe zbiory przygotowano zwiększając odpowiednio liczebność jednej z analizowanych klas zależnej (*oversampling*).

[102] Znajdowanie optymalnego punktu odcięcia stanowi osobne zagadnienie, które będzie przedmiotem rozważań w dalszej części pracy

Analizę uzupełnia prezentacja wybranych miar siły predykcyjnej, przyjmujących, że odpowiedzi modelu należy traktować jako zmienną rangującą.

Dokładność jest jedną z najbardziej intuicyjnych miar jakości modelu. Przedstawia odsetek poprawnie zaklasyfikowanych przypadków w stosunku do wszystkich przypadków. W przypadku modeli migracji (*churn models*) jest to miara, najczęściej niewiele mówiąca badaczowi ze względu na jej wrażliwość na występowanie nierównych proporcji przypadków klientów lojalnych i nielojalnych w zbiorze danych. Okoliczność ta jest powszechnym zjawiskiem w modelach tego typu. Poniżej przedstawiono wykresy dokładności dla tego samego modelu dla pięciu wersji próby walidacyjnej zgodnie z powyższym opisem.



Rysunek 21 Dokładność (ACC) względem prawdopodobieństwa klasy pozytywnej
Źródło: opracowanie własne.

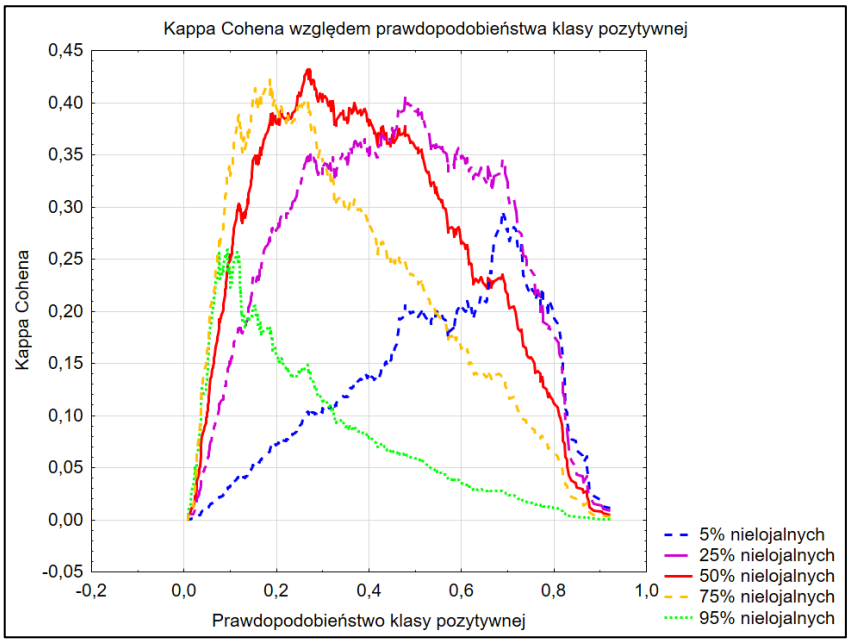
Na podstawie wykresu (Rysunek 21) można zauważyć, że dla punktu odcięcia na poziomie 0,3 reprezentującego udział nielojalnych klientów w zbiorze uczącym, wartość ACC wynosi około 0,68 dla wszystkich zbiorów. Dla zbioru zbilansowanego najwyższy poziom wskaźnika ACC równy 0,71 model osiąga dla punktu odcięcia około 0,26. Jeżeli punkt odcięcia zostałby ustalony na poziomie 0 (prognozujemy, że wszyscy są lojalni) to i tak wartość ACC wyniesie 0,5, mimo że model będzie w praktyce bezużyteczny. W przypadku modeli skrajnie niezbilansowanych wartość ACC osiąga swoje maksimum na poziomie 0,95, przy równoczesnym braku mocy predykcyjnej. Miara ACC może zatem

posłużyć badaczowi do wyboru najlepszego modelu (najlepszy model to taki, który cechuje się najwyższą wartością ACC), nie pozwala natomiast na ocenę czy dany model jest dobrej jakości (wartość 0,95 może oznaczać zarówno model prawie idealny, jak i model bezużyteczny). Należy dodać, że w przypadku zbiorów danych z niezbilansowanymi próbami dokładność jest miernikiem o niewielkim znaczeniu. Badaczowi zależy przede wszystkim na prognozowaniu klasy mniej licznej i to jej trafność predykcji jest zazwyczaj najważniejsza.

Aby zredukować niepożądane własności miary ACC możliwe jest jej porównanie z wartością progową równą odsetkowi przypadków częstszej klasy zmiennej zależnej. Za potencjalnie wartościowe uznaje się jedynie modele o ACC powyżej tego progu. Alternatywą jest zastosowanie miary, która w swojej konstrukcji uwzględnia rozkład klas zmiennej zależnej w próbie uczącej. Przykładem takiej miary jest współczynnik Kappa Cohena oryginalnie zaprojektowany do oceny zgodności pomiędzy dwoma sędziami. Miara ta bierze pod uwagę poziom trafności, jaki mógłby być uzyskany w wyniku przypadku. Miarę tę oblicza się zgodnie z następującą formułą [Kuhn, Johnson, 2013]:

$$Kappa = \frac{O - E}{1 - E}$$

gdzie O jest obserwowaną wartością ACC, natomiast E jest oczekiwaną wartością ACC dla modelu naiwnego obliczoną na podstawie sum brzegowych macierzy pomyłek. Statystyka ta przyjmuje wartości od -1 do 1. Wartość 1 informuje o modelu idealnym (idealnej zgodności przewidywań z rzeczywistością), wartość 0 o modelu losowym. Wartości ujemne świadczą o wyniku gorszym niż model losowy.

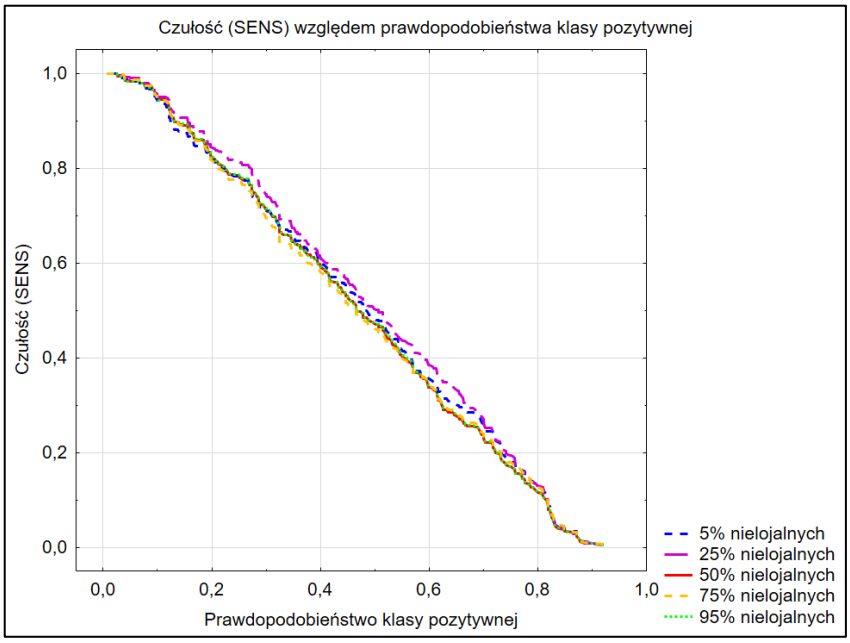


Rysunek 22 Kappa Cohena względem prawdopodobieństwa klasy pozytywnej
Źródło: opracowanie własne.

Na podstawie Rysunek 22 można stwierdzić, że pomimo korekty o wartość oczekiwaną E, miara Kappa Cohena jest wrażliwa na nierówne proporcje klas zmiennych zależnych. Mają one wpływ zarówno na wartość tego współczynnika, jak również na optymalny punkt odcięcia. Należy zauważyć, że w przeciwieństwie do miary ACC, współczynnik Kappa Cohena premiuje zbiory o zbilansowanych proporcjach.

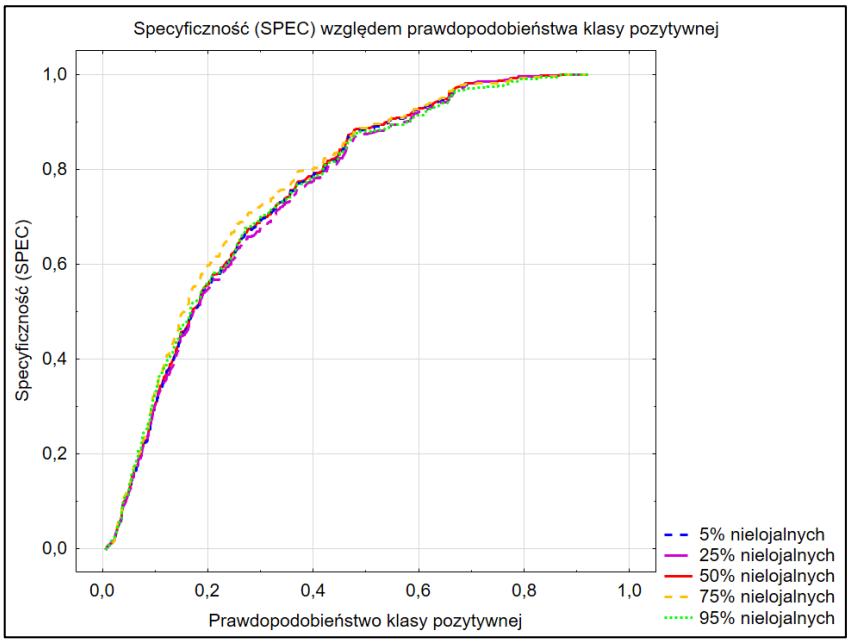
Kolejnymi miarami, którym należy się przyjrzeć nieco bliżej są czułość (SENS) oraz specyficzność (SPEC). Czułość informuje, jaki procent spośród osób nielojalnych został poprawnie wskazany przez model. Przyjmuje wartości od 0 do 1, gdzie 0 oznacza brak zdolności do wskazania osób nielojalnych zaś 1 wiąże się z idealną zdolnością do ich wyróżniania. Zatem model czuły to taki, który skutecznie potrafi wychwycić osoby nielojalne (i ogólnie rzecz ujmując, przypadki z klasy pozytywnej). Model skuteczny to model cechujący się wysoką czułością, natomiast zależność ta nie działa w drugą stronę. Jeżeli model w sposób naiwny wskaże, że wszyscy klienci są nielojalni, to wprawdzie będzie cechował się idealną czułością, ale nie jest to w żaden sposób tożsame z jego skutecznością. Przykład ten pokazuje, że miara ta nie może być stosowana autonomicznie w procesie oceny modeli. Ważną cechą czułości jest fakt, że nie zależy ona od udziału klasy pozytywnej w analizowanym zbiorze danych. Poniżej przedstawiono wykres tej miary (Rysunek 23) w zależności od punktu odcięcia odpowiedzi modelu dla

przedstawionych pięciu zbiorów. Różnice w przebiegach wykresów wynikają jedynie z zaburzenia zbioru spowodowanego próbkowaniem.



Rysunek 23 Czułość (SENS) względem prawdopodobieństwa klasy pozytywnej
Źródło: opracowanie własne.

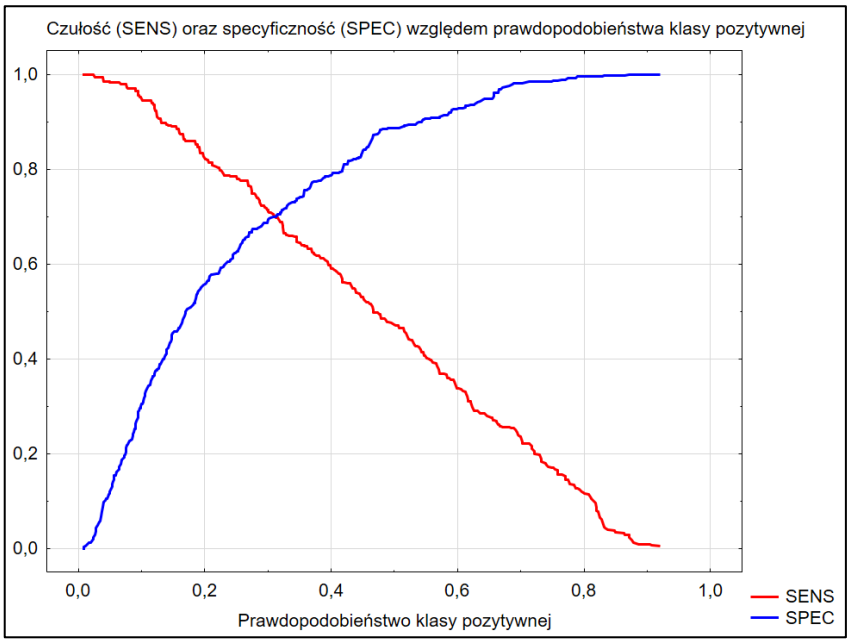
Miarą komplementarną do czułości jest specyficzność. Informuje ona, jaki procent lojalnych klientów został poprawnie wskazany przez model. Specyficzność przyjmuje wartości od 0 do 1, gdzie wartość 0 informuje o braku zdolności modelu do poprawnego klasyfikowania lojalnych klientów, a wartość 1 o idealnych zdolnościach modelu do ich wyróżniania. Model o specyficzności równej 1 nie musi być jednak modelem użytecznym. Jeżeli klasyfikator naiwnie wskaże, że wszyscy klienci są lojalni to osiągnie specyficzność równą 1. Podobnie jak w przypadku czułości, wartość specyficzności nie zależy od wielkości udziału klasy negatywnej w analizowanym zbiorze danych.



Rysunek 24 Specyficzność (SPEC) względem prawdopodobieństwa klasy pozytywnej
Źródło: opracowanie własne.

Ocena dobroci dopasowania modelu wymaga zatem łącznej oceny jego czułości i specyficzności. Kolejny wykres (Rysunek 25) prezentuje obydwie miary dla zbioru zbilansowanego[103]. Można zauważyć, że wzrost jednej miary wiąże się ze spadkiem drugiej.

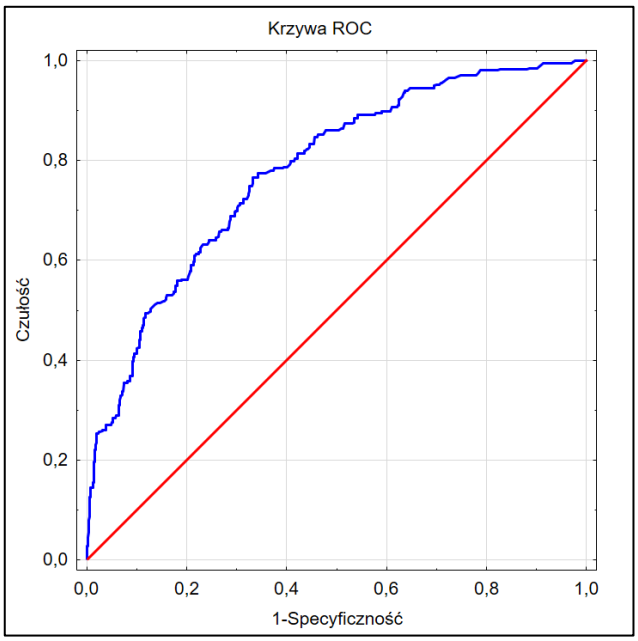
[103] Wykresy dla pozostałych zbiorów wyglądałyby dokładnie tak samo z dokładnością do błędu próbkowania



Rysunek 25 Czułość (SENS) oraz specyficzność (SPEC) względem prawdopodobieństwa klasy pozytywnej

Źródło: opracowanie własne.

Łączna ocena tych miar prowadzi bezpośrednio do konstrukcji krzywej ROC (*Receiver Operating Characteristic*). Krzywa ROC jest tworzona przez przedstawienie na jednym wykresie odsetka prawdziwie dodatnich (czułość) oraz odsetka fałszywie dodatnich (1-specyficzność).



Rysunek 26 Krzywa ROC

Źródło: opracowanie własne.

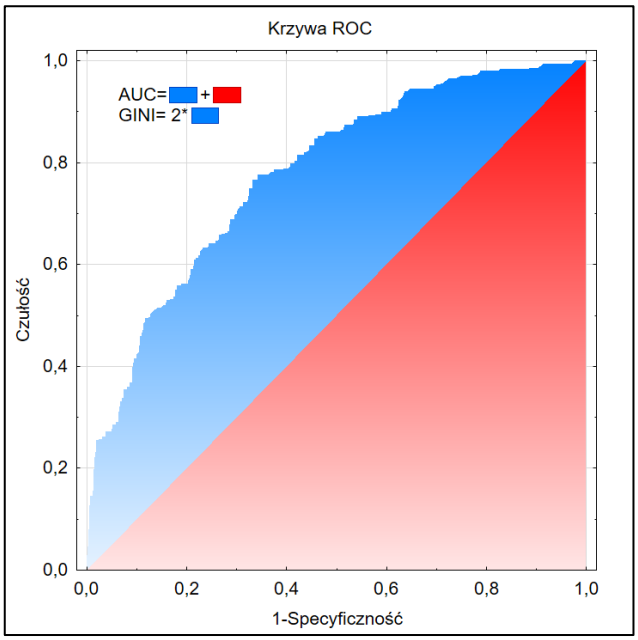
Rysunek 26 przedstawia krzywą ROC prezentującą łącznie obydwie wartości. Najbardziej popularnym wskaźnikiem związanym z analizą ROC jest pole powierzchni pod krzywą - AUC (*Area Under Curve*). Pole to stanowi syntetyczną miarę mocy predykcyjnej modelu. Konstrukcja tej miary zakłada, że odpowiedzi modelu wystarczy traktować jako zmienną rangującą. W przypadku modelu idealnie separującego klientów lojalnych od nielojalnych pole powierzchni przyjmuje wartość 1, w przypadku modelu losowego, pole pod krzywą wynosi 0,5[104]. Wartość AUC nie zależy od proporcji klas modelowanej zmiennej zależnej co pozwala na normatywną ocenę siły predykcyjnej modelu. Miara ta informuje, jaki jest przeciętny poziom przypadków prawdziwie pozytywnych dla wszystkich możliwych wskaźników fałszywie pozytywnych [Krzanowski, Hand, 2009]. AUC można również interpretować jako prawdopodobieństwo, że wybrana losowo osoba z grupy nielojalnych będzie miała według modelu większą skłonność do odejścia od wybranej losowo osoby z grupy lojalnych.

W praktyce budowy modeli klasyfikacyjnych często można spotkać użycie wskaźnika Giniego (GINI) stosowanego w analizie koncentracji. Miara ta jest ściśle powiązana z krzywą ROC oraz wskaźnikiem AUC (Rysunek 27). Można ją obliczyć na podstawie następującego wzoru [Krzanowski, Hand, 2009]:

$$GINI = 2 * AUC - 1$$

Graficznie jest to zatem podwojone pole pod krzywą ROC ograniczone jedynie do części nad linią odniesienia (linią modelu losowego). Wskaźnik Giniego przyjmuje wartości od 0 do 1, gdzie 0 odpowiada modelowi losowemu a 1 modelowi idealnemu. Wartość GINI pokrywa się z wartością statystyki D Somersa. W praktyce oceny modeli obie te miary są używana zamiennie.

[104] Teoretycznie mógłby istnieć klasyfikator, dla którego wartość AUC byłaby poniżej 0,5 musiałby jednak działać gorzej od losowego wyboru.



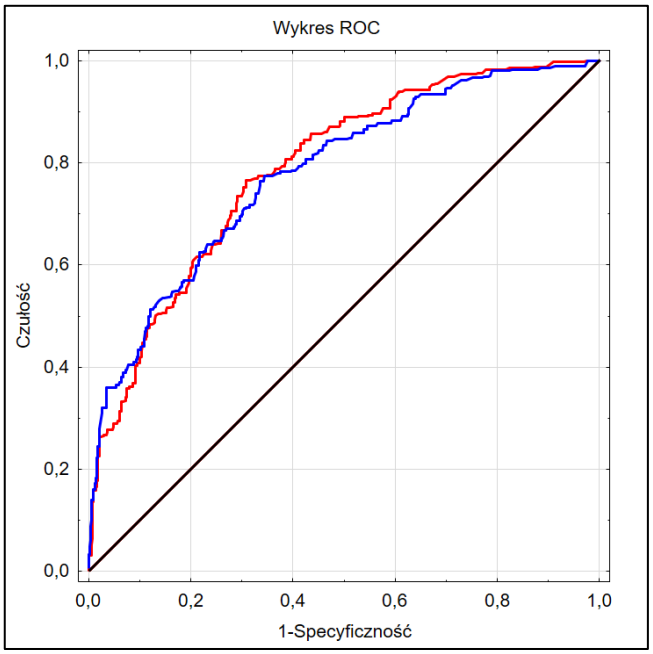
Rysunek 27 Wskaźnik GINI a krzywa ROC
Źródło: opracowanie własne.

Poniżej przedstawiono interpretację siły predykcyjnej modelu w zależności od wartości wskaźnika GINI [Migut i inni, 2013]:

- Poniżej 0,2 – model do odrzucenia,
- 0,2 – 0,4 – słaba siła dyskryminacyjna,
- 0,4 – 0,6 – akceptowalna siła dyskryminacyjna,
- 0,6 – 0,95 – duża siła dyskryminacyjna,
- Powyżej 0,95 – nadmiernie optymistycznie duża siła dyskryminacyjna (zbyt dobrze by było prawdziwie).

Miara AUC/GINI bardzo często służy badaczom do porównania siły predykcyjnej modeli i jest podstawą wyboru najlepszego z nich. Podejście to, choć popularne, nie powinno być stosowane w oderwaniu od analizy kształtu krzywej ROC. Jak napisano powyżej, pole powierzchni AUC przedstawia średni poziom odsetka prawdziwie pozytywnych dla całego zakresu odpowiedzi modelu, zatem nie odnosi się do żadnego konkretnego punktu odcięcia. Wybór „najlepszego" modelu na podstawie miary AUC bez uwzględnienia aspektu biznesowego może prowadzić do nieoptymalnych wyborów. Aby przybliżyć to zagadnienie należy rozważyć dwa modele o identycznym wskaźniku AUC. Na poniższym wykresie (Rysunek 28) widać dwie krzywe odpowiadające rozważanym

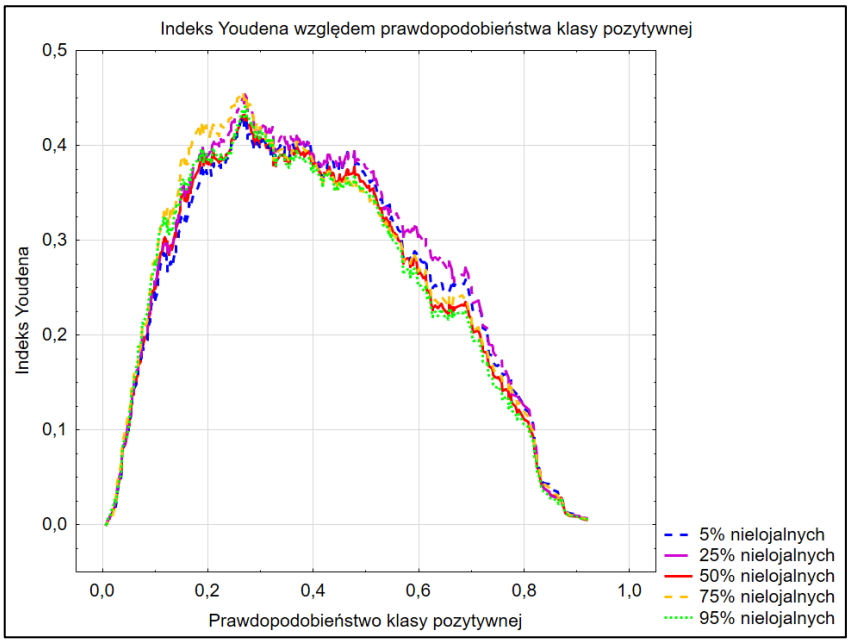
modelom. Ich linie się przecinają, zatem każdy z nich działa lepiej od drugiego jedynie w zakresie, w którym cechuje się większą czułością dla zadanego odsetka fałszywie pozytywnych (1-SPEC). W takiej sytuacji możliwa strategia polega na stworzeniu łącznego klasyfikatora bazującego na lokalnie najlepszych modelach. W praktyce jednak unika się nadmiernej komplikacji procesu decyzyjnego i dokonuje się wyboru jednego modelu, działającego najlepiej w realnym obszarze decyzyjnym[105]. Wybrany model może mieć zatem niższy wskaźnik AUC od modeli odrzuconych.



Rysunek 28 Krzywe ROC – porównanie dwóch modeli
Źródło: opracowanie własne.

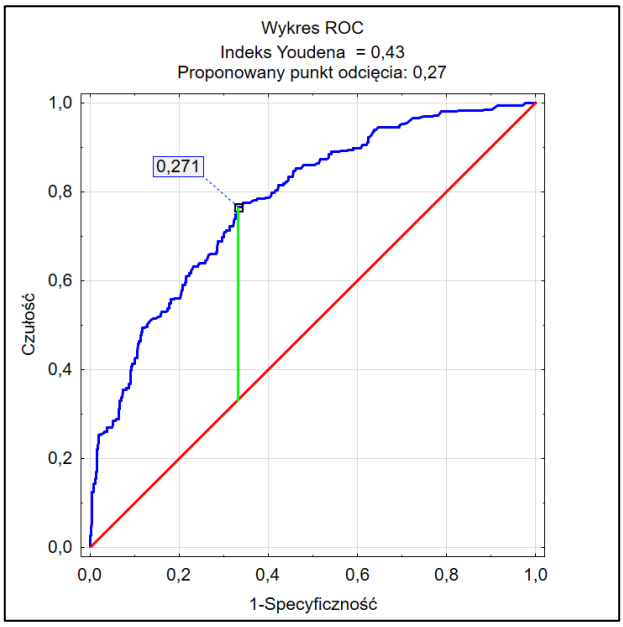
Indeks J Youdena jest syntetyczną miarą siły predykcyjnej modelu. Przyjmuje wartości od 0 (o ile model działa nie gorzej od modelu losowego) do 1. Wartość 1 przyjmuje dla modelu idealnego, dla którego nie występują ani przypadki FP ani FN [Youden, 1950]. Miara ta nie zależy od frakcji klasy pozytywnej w analizowanym zbiorze danych co jest widoczne na Rysunek 29.

[105] Na przykład w wyniku działania modelu za nielojalnych może zostać uznane nie więcej niż 3% bazy danych klientów – większy odsetek wiązałby się z nieakceptowalną skalą interwencji w stosunku do klientów.



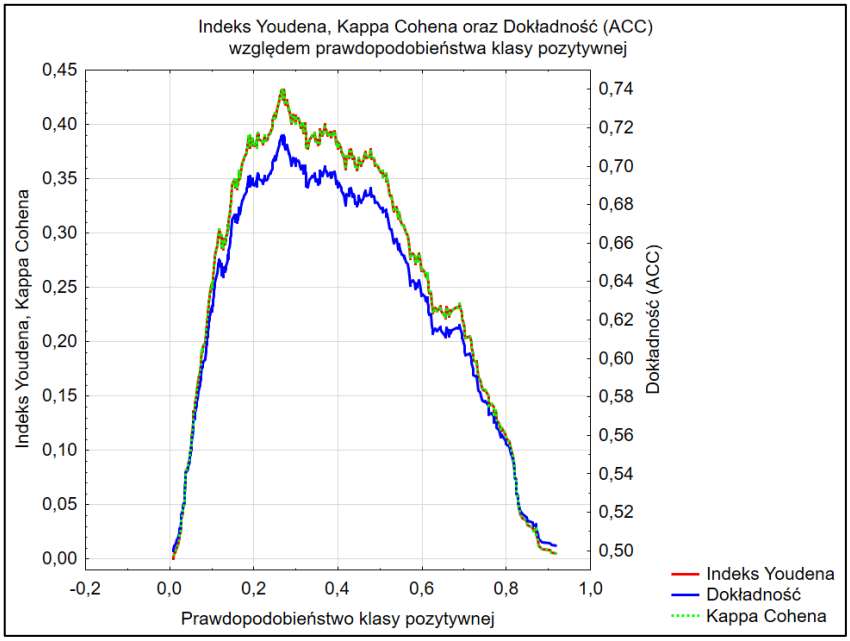
Rysunek 29 Indeks J Youdena względem prawdopodobieństwa klasy pozytywnej
Źródło: opracowanie własne.

Indeks J jest miarą powiązaną z krzywą ROC. Na podstawie krzywej ROC dla danego punktu odcięcia wartość indeksu J Youdena może być wyznaczona jako długość pionowej linii łączącej linę modelu losowego z krzywą ROC. Wartość odpowiedzi modelu, dla którego indeks J Youdena osiąga maksimum może być rozpatrywany jako potencjalny punkt odcięcia modelu.



Rysunek 30 Indeks J Youdena a krzywa ROC
Źródło: opracowanie własne.

Dla zbiorów danych o zbalansowanym rozkładzie klas zmiennej zależnej indeks J Youdena pokrywa się z wartościami Kappy Cohena, jest również mocno skorelowany z wartościami dokładności (ACC).

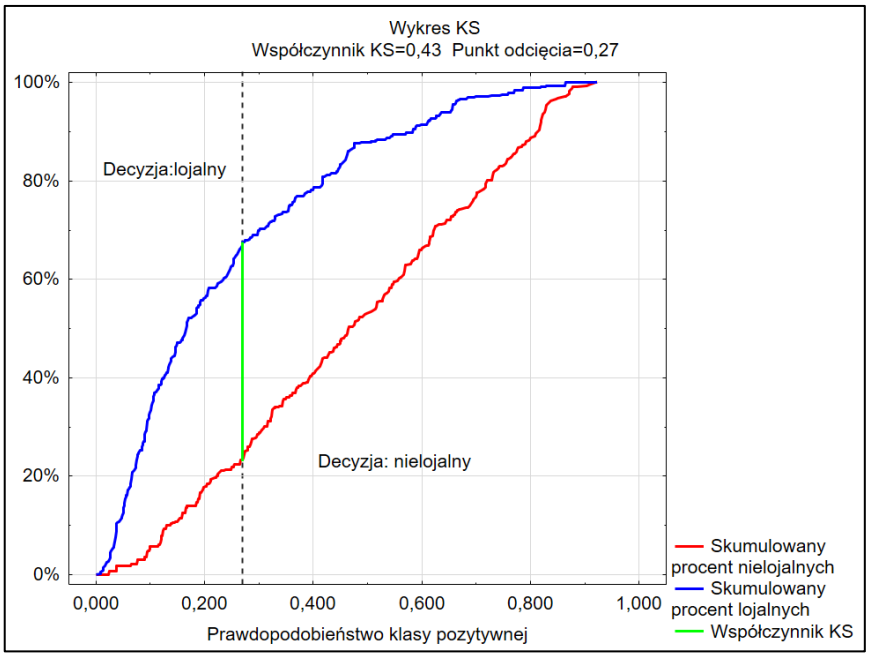


Rysunek 31 Porównanie indeksu J Youdena, Kappy Cohena oraz dokładności dla zbioru zbalansowanego
Źródło: opracowanie własne.

Miarą tożsamą z indeksem J Youdena, wywodzącą się z odmiennej tradycji oceny modeli dodatkowo wnoszącą pewne dodatkowe informacje do całościowego obrazu jakości modelu jest statystyka KS (Kołmogorowa-Smirnova). Statystyka KS określa maksymalną różnicę między skumulowanym rozkładem (dystrybuantą) klientów lojalnych i nielojalnych.

$$KS = \max_j |G(x_j) - B(x_j)|,$$

gdzie: $G(x_j)$ – dystrybuanta rozkładu klientów lojalnych; $B(x_j)$ – dystrybuanta rozkładu klientów nielojalnych. Dystrybuanty konstruujemy dla klientów uprzednio uporządkowanych rosnąco względem prawdopodobieństwa *a posteriori* bycia lojalnym klientem uzyskanego na podstawie zbudowanego modelu.



Rysunek 32 Wykres KS
Źródło: opracowanie własne.

Na podstawie powyższego wykresu (Rysunek 32) można zauważyć, że maksymalna różnica pomiędzy dystrybuantami wyniosła KS= 0,43 (zielona linia). Jeśli model klasyfikowałby do grupy „nielojalnych" osoby, dla których odpowiedź modelu byłaby powyżej 0,27 a do grupy „lojalnych" w pozostałych przypadkach wartość SENS wyniosłaby około 0,77 (za nielojalnych uznaliśmy 77% spośród nielojalnych) natomiast FPR = 0,34 (34% spośród osób lojalnych zostało niepoprawnie zaklasyfikowanych).

Innymi słowy KS to maksymalna różnica pomiędzy wartością SENS oraz FPR. Miara KS jest tożsama z maksymalną wartością indeksu Youdena.

Z definicji, statystyka KS przyjmuje wartości z przedziału [0;1]. Im większa wartość tej miary, tym wyższa zdolność modelu do separowania klientów „lojalnych" i „nielojalnych". Statystyka KS nie zależy od udziału klientów nielojalnych w próbie. Wartości KS można interpretować w następujący sposób [Migut i inni, 2013]:

- Poniżej 0,2 – model do odrzucenia,
- 0,2 – 0,4 – słaba siła dyskryminacyjna,
- 0,4 – 0,5 – akceptowalna siła dyskryminacyjna,
- 0,5 – 0,6 – duża siła dyskryminacyjna,
- 0,6 – 0,75 – bardzo duża siła dyskryminacyjna,
- Powyżej 0,75 – wynik nadmiernie optymistyczny (zbyt wysoki by było prawdziwie).

Statystyka KS jest miarą centralną, czyli nie ocenia całego rozkładu możliwych odpowiedzi modelu[106], a jedynie jeden specyficzny punkt, w którym model posiada największą siłę predykcyjną. Jeżeli punkt ten leży (np. z przyczyn biznesowych) poza przedziałem możliwych punktów odcięcia – KS staje się miarą niewiarygodną. Działanie przedsiębiorstwa na podstawie statystyki KS w przedstawionym przykładzie wiązało by się z koniecznością niepotrzebnego kontaktu z 34% klientów lojalnych, co z punktu widzenia biznesowego byłoby nieracjonalne.

Często używanym wykresem stosowanym w połączeniu z analizą KS jest wykres wartości różnic pomiędzy SENS oraz FPR (tożsamy z wykresem wartości indeksu Youdena – Rysunek 29). Pozwala on określić zakres wartości odpowiedzi modelu, dla których model ma największą siłę dyskryminacyjną.

Ilorazy wiarygodności są kolejnymi miarami, które oparte są na miarach SPEC oraz SENS. Im większa wartość LR+ tym lepsza zdolność do dyskryminacji przypadków pozytywnych. Wartość LR+=1 oznacza losowe działanie modelu, poniżej tej wartości model działa gorzej niż losowy [Biggerstaff, 2000]. Im mniejsza wartość LR- tym lepsza zdolność dyskryminacji przypadków negatywnych. Zatem model A będzie lepszy od modelu B, jeżeli ma większą wartość LR+ oraz mniejszą wartość LR-. Większa wartość obydwóch miar powadzi do nierozstrzygniętych sytuacji, w których jeden model lepiej

[106] W przeciwieństwie na przykład do krzywej ROC.

radzi sobie z przypadkami pozytywnymi a drugi z negatywnymi. Ponieważ obie te miary są oparte na miarach SENS oraz SPEC ich wartości nie zależą od odsetka osób nielojalnych w zbiorze danych. Wydaje się, że z punktu widzenia prognozowania odejść klientów lepszym miernikiem będzie LR+.

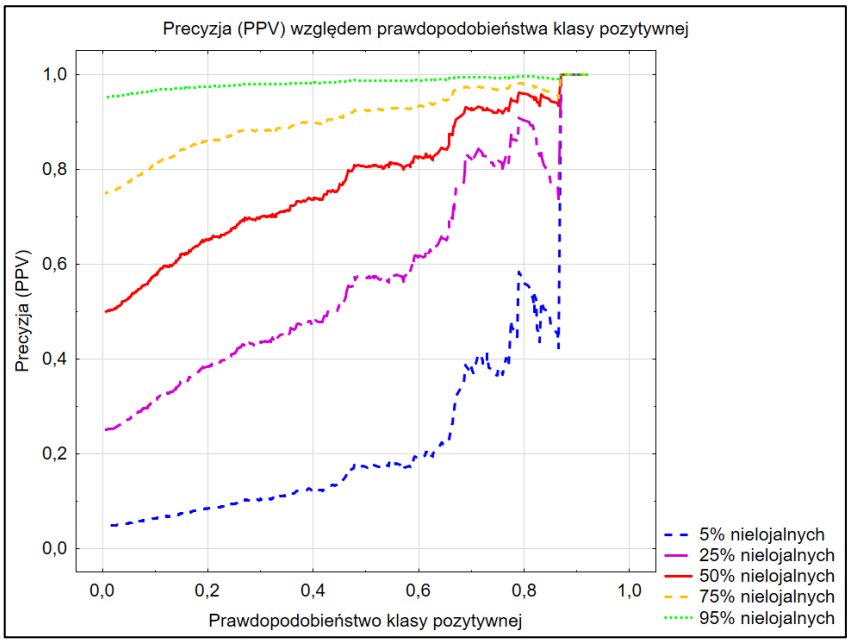
Ostatnią miarą opartą na miarach SENS i SPEC zawartą w niniejszej pracy jest DP (*Discriminant Power*) [Sokolova i inni, 2006].

$$DP = \frac{\sqrt{3}}{\pi} * \left[log\left(\frac{SENS}{1-SENS}\right) + log\left(\frac{SPEC}{1-SPEC}\right)\right]$$

Zgodnie ze swoją nazwą pozwala ocenić siłę dyskryminacyjną modelu. Przyjmuje się następującą interpretację DP [Sokolova i inni, 2006]:

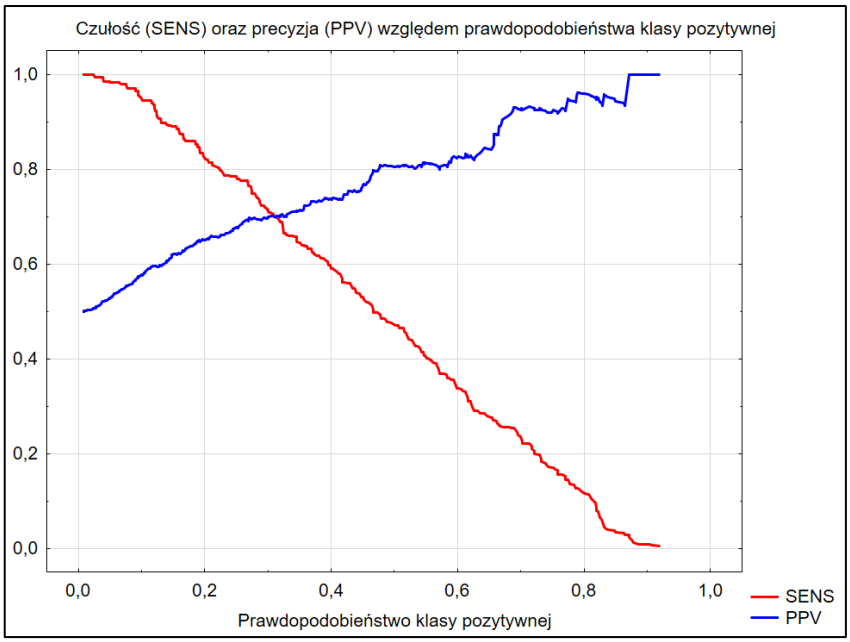
- DP poniżej 1 – słaba siła dyskryminacyjna,
- DP od 1 do 2 – ograniczona siłą predykcyjna,
- DP od 2 do 3 – akceptowalna siła dyskryminacyjna,
- Powyżej 3 – wysoka siła dyskryminacyjna.

Kolejną parą miar, które analizowane łącznie pozwalają na ocenę siły predykcyjnej modelu są PPV oraz SENS. PPV informuje o tym, jaka część osób uznanych przez model za nielojalne rzeczywiście odeszły do konkurencji. Miara ta zależy od odsetka osób nielojalnych w zbiorze danych. W sytuacji klasyfikatora naiwnego, gdy model wskazuje, że wszystkie osoby są nielojalne, miara PPV będzie równa frakcji osób nielojalnych w modelowanym zbiorze. Poniżej przedstawiono wykres PPV dla tego samego modelu przygotowany dla zbiorów o różnych proporcjach klas zmiennej zależnej (Rysunek 33).



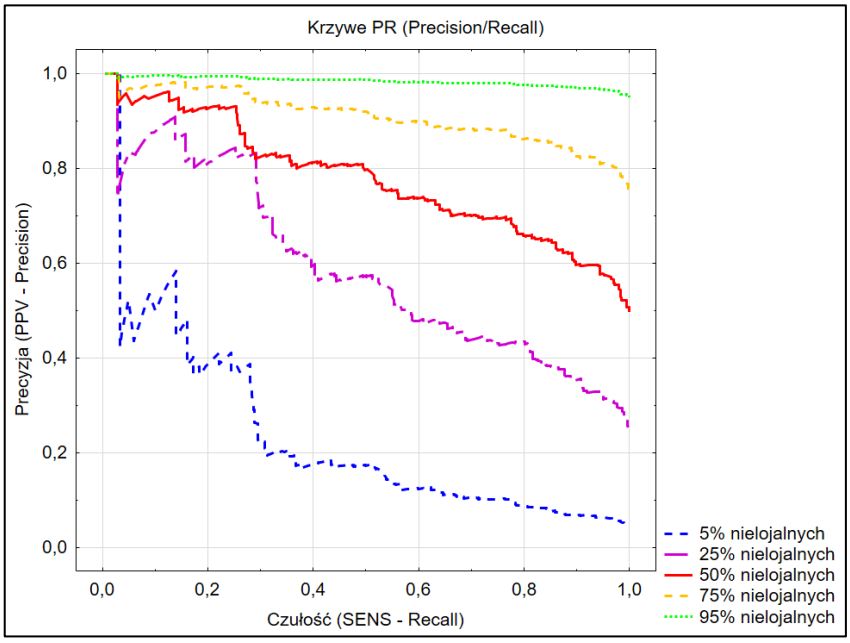
Rysunek 33 Precyzja (PPV) względem prawdopodobieństwa klasy pozytywnej
Źródło: opracowanie własne.

Warto zauważyć, że powyższe miary jakości modelu są w stosunku do siebie antagonistyczne. Wzrost wartości PPV powoduje spadek SENS i na odwrót. Własność tą ilustruje kolejny wykres (Rysunek 34), na którym przedstawiono obie miary na przykładzie zbioru o równych proporcjach klas zmiennej zależnej.



Rysunek 34 Czułość (SENS) oraz precyzja (PPV) względem prawdopodobieństwa klasy pozytywnej
Źródło: opracowanie własne.

Wykres ten umożliwia wybór optymalnego punktu odcięcia. Podobną rolę odgrywa kolejny wykres – krzywa PR (*Precision/Recall Curve*), na której przedstawiono wszystkie możliwe kompromisy pomiędzy PPV oraz SENS (Rysunek 34). Przykładowo dla zbioru o równej proporcji klas zmiennej zależnej przy czułości na poziomie 40% uzyskuje się 80% poprawność klasyfikacji osób nielojalnych.



Rysunek 35 Krzywe PR (Precison/Recall)
Źródło: opracowanie własne.

Krzywa PR może być stosowana w sytuacji, gdy przypadki klasy pozytywnej stanowią niewielką frakcję zbioru danych, bądź gdy dla badacza ważniejsze jest unikanie przypadków fałszywie pozytywnych (FP) niż fałszywie negatywnych (FN). W przeciwnej sytuacji rekomendowana jest krzywa ROC [Geron, 2017].

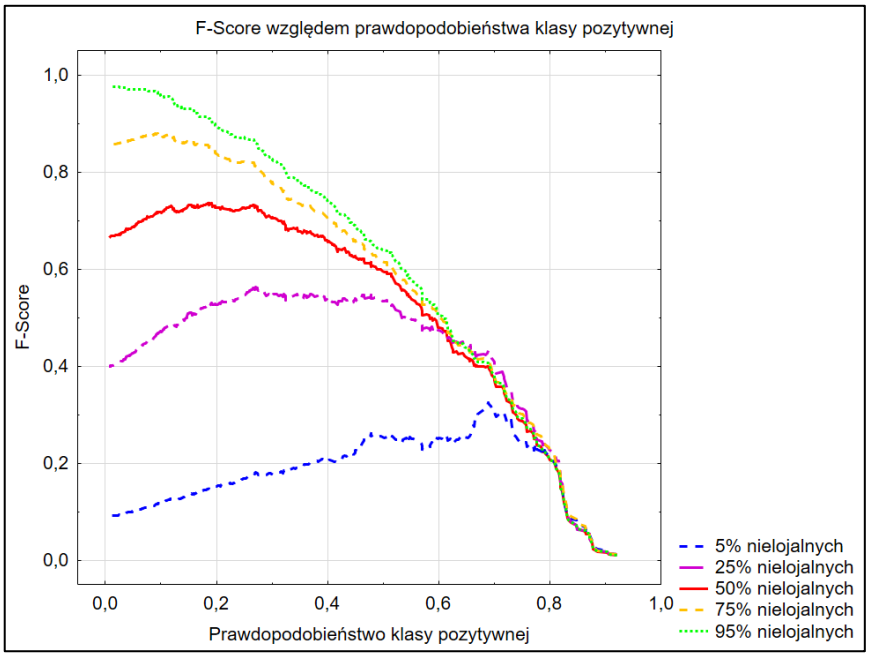
Miara F (*F-Score*) jest syntetyczną miarą siły predykcyjnej modelu łączącą SENS oraz PPV, będącą ich średnią harmoniczną. Jest ona bardzo popularną miarą siły predykcyjnej, przydatną zwłaszcza, gdy celem jest porównanie modeli [Geron, 2017].

$$F\ Score = 2 * \frac{SENS * PPV}{SENS + PPV} = \frac{2}{\frac{1}{SENS} + \frac{1}{PPV}}$$

Miara ta charakteryzuje się wrażliwością na niezbilansowane próby, podobnie jak precyzja (PPV). Jej wartość zależy zarówno od siły predykcyjnej modelu jak i frakcji osób nielojalnych w zbiorze danych. Skutkuje to zaniżonym oszacowaniem jej wartości

w sytuacji, gdy klasa przypadków nielojalnych jest rzadziej reprezentowana w zbiorze danych – szczegóły na Rysunek 36[107].

Warto w tym miejscu przypomnieć, że wartość 0 informuje o braku siły predykcyjnej klasyfikatora natomiast wartość 1 o idealnej klasyfikacji. Kolejną własnością miary *F-Score* jest fakt, iż faworyzuje ona klasyfikatory mające podobny poziom SENS oraz PPV. W przypadku różnicy w wartościach SENS oraz PPV większy wpływ na końcowy wynik ma niższa z tych dwóch wartości.



Rysunek 36 F-Score względem prawdopodobieństwa klasy pozytywnej
Źródło: opracowanie własne.

Aby umożliwić badaczowi subiektywne określenie ważności obydwu miar możliwa jest modyfikacja wzoru poprzez dodanie składnika β. Im większa wartość β tym większe znaczenie miary PPV.

$$F_\beta\ Score = \frac{1}{\beta * \frac{1}{PPV} + (1-\beta) * \frac{1}{SENS}}$$

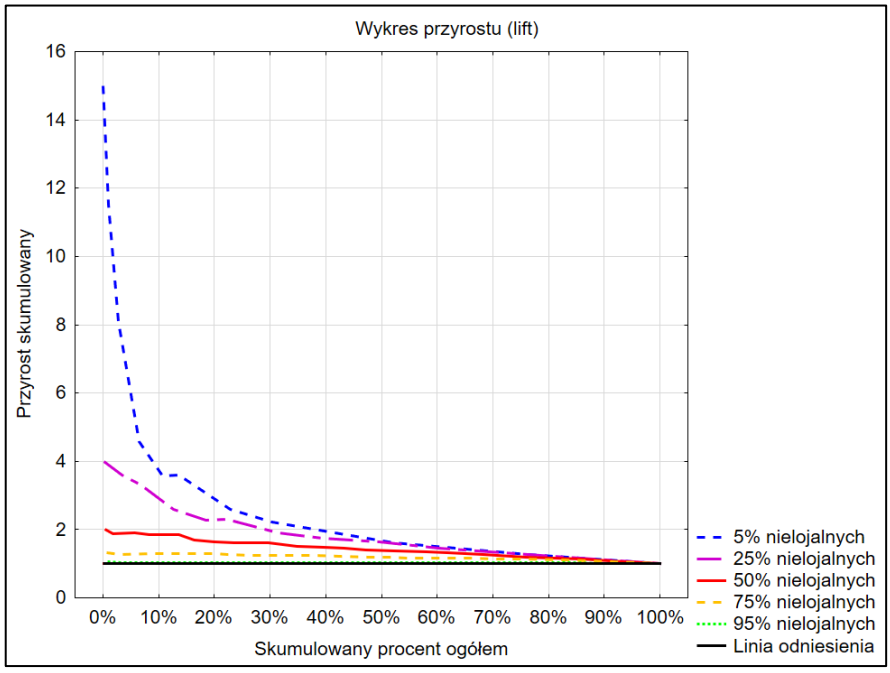
W większości sytuacji modele mają niższą wartość SENS niż PPV. Nadanie im różnych wag może zatem wynikać z chęci zbalansowania wpływu obydwu miar na końcowy wynik. Znalezienie optymalnej wartości β jest kłopotliwe i może zostać uznane

[107] Wykresy miary *G-mean* dają analogiczne wyniki.

za kontrowersyjne. Zdaniem Powersa nie istnieje realne uzasadnienie zastosowania β innego niż 0,5, który nadaje im równą wagę [Powers, 2011]. Należy podkreślić, że niezależnie od wersji wzoru, miara *F-Score* nie bierze pod uwagę komórki TN z macierzy pomyłek. Może to być źródłem obciążenia tej miary.

Przyrost (*lift*) informuje o tym, w jakim stopniu zastosowanie klasyfikatora jest lepsze w stosunku do losowego wyboru przypadków. Jest stosunkiem odsetka osób nielojalnych w grupie wytypowanej przez model jako nielojalni do odsetka nielojalnych osób w całym zbiorze.

Wartość przyrostu jest zależna od frakcji osób nielojalnych w zbiorze danych. Im mniejszy odsetek osób nielojalnych tym wyższe wartości może przyjmować współczynnik *lift*. W przypadku modelu losowego wartość przyrostu wynosi 1. W przeciwieństwie do wcześniej prezentowanych miar, wykres przyrostu nie jest zazwyczaj prezentowany w zależności od wartości prawdopodobieństwa *a posteriori* uzyskanego na podstawie modelu. Najczęściej przed prezentacją wykresu wartości odpowiedzi modelu są rangowane. Na osi X są zatem prezentowane percentyle rozkładu odpowiedzi modelu.

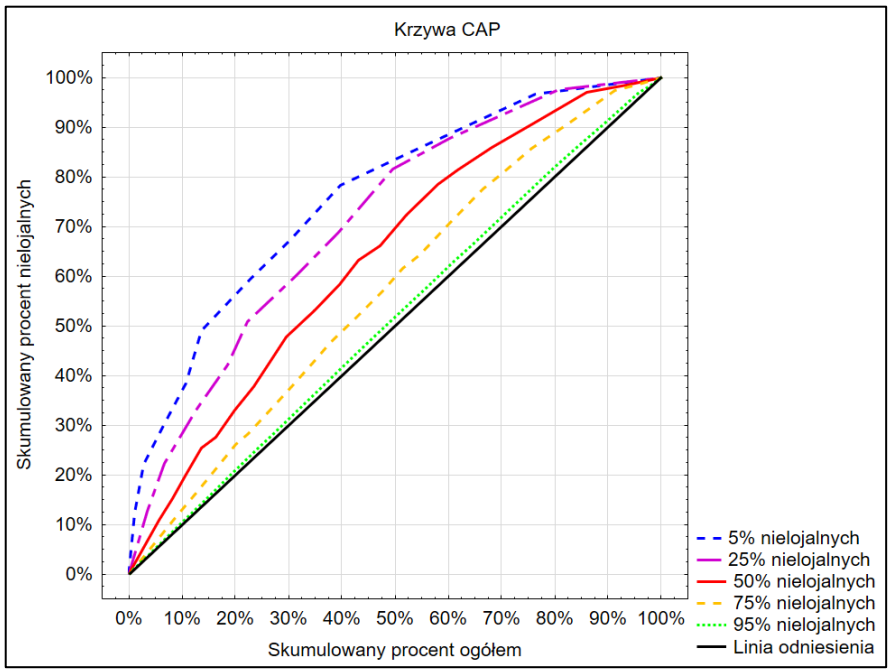


Rysunek 37 Wykres przyrostu (lift)
Źródło: opracowanie własne.

Pomimo widocznej na Rysunek 37 wrażliwości tej miary na niezbilansowanie zbioru danych, jest ona bardzo popularnym narzędziem oceny siły predykcyjnej modelu ze

względu na intuicyjną interpretację zwłaszcza w porównaniu z krzywymi ROC. Na podstawie powyższego wykresu (Rysunek 37) można stwierdzić, że dla zbioru z 5% odsetkiem osób nielojalnych, wskazanie 10% osób z największą według modelu skłonnością do odejścia pozwala osiągnąć wskaźnik lift na poziomie 4. Oznacza to, że w wybranej grupie wskaźnik PPV jest 4 razy większy od analogicznej wartości uzyskanej w grupie dobranej losowo.

W połączeniu z krzywą przyrostu przedstawiana jest zazwyczaj krzywa CAP (*Cummulative Accuracy Profiles*)[108]. Krzywa ta przedstawia czułość (SENS) modelu (oś Y) względem osób ocenionych przez model za nielojalne (PP). Jej wygląd jest zbliżony do krzywej ROC. Kształt krzywej zależy zarówno od mocy predykcyjnej modelu jak i od odsetka osób nielojalnych w zbiorze danych.

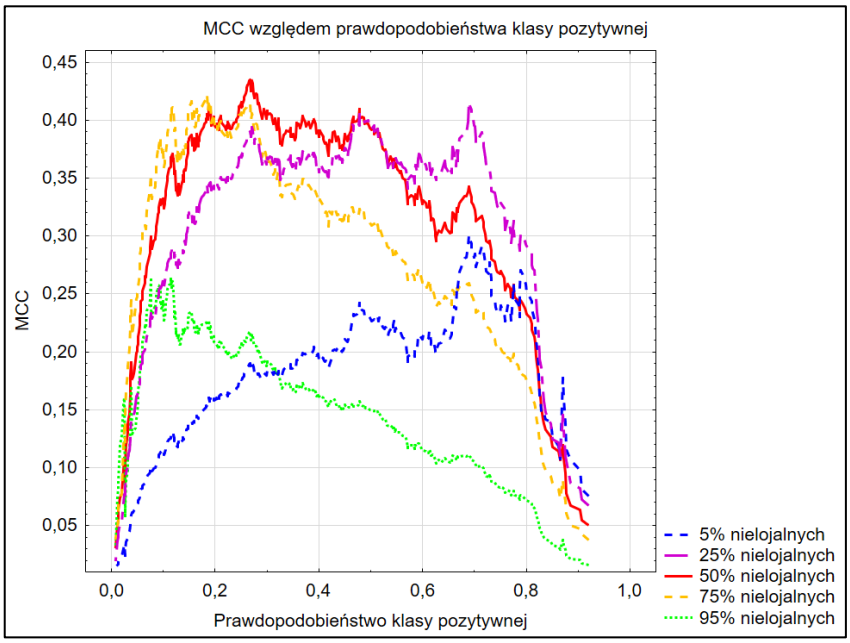


Rysunek 38 Krzywa CAP
Źródło: opracowanie własne.

Na podstawie wykresu widocznego na Rysunek 38 można stwierdzić, że dla zbioru zawierającego 5% osób nielojalnych, dotarcie do połowy z nich wiąże się z koniecznością kontaktu z 14% bazy klientów – osób z najwyższą skłonnością do odejścia. Dla zbioru zawierającego 25% osób nielojalnych podobny efekt uzyskano by po kontakcie z około 22% bazy klientów.

[108] Krzywa ta nazywana jest również krzywą korzyści (*gains chart*)

Kolejną miarą oceniającą dobroć dopasowania jest współczynnik korelacji Matthewsa MCC (*Matthews correlation coefficient*). Miara bierze pod uwagę wszystkie składowe macierzy pomyłek. Jest niewrażliwa na występowanie niezrównoważonych proporcji modelowanych klas w zbiorze danych [Boughorbel i inni, 2017]. Jest średnią geometryczną indeksu Youdena oraz wyrazistości (MK) i zarazem odpowiednikiem współczynnika korelacji liniowej Pearsona dla danych nominalnych [Powers, 2012]. Przyjmuje wartości od -1 do 1. Wartość 1 świadczy o idealnej zgodności pomiędzy przewidywanymi a obserwowanymi, 0 jest równoznaczne z klasyfikatorem losowym, wartość -1 informuje o idealnej niezgodności pomiędzy predykcją a stanem faktycznym [Matthews, 1975]. Na Rysunek 39 przedstawiono wykres MCC dla analizowanych zbiorów. Analiza uzyskanych wyników nie potwierdza całkowitego braku wrażliwości na niezrównoważone proporcje klas zmiennych zależnych. Wartości miary MMC oraz charakterystyka wykresów są bardzo zbliżone do wyników uzyskanych dla Kappy Cohena (Rysunek 22).



Rysunek 39 MCC względem prawdopodobieństwa klasy pozytywnej
Źródło: opracowanie własne.

4.1.3. Miary oparte na probabilistycznej interpretacji odpowiedzi modelu

Wskaźnik Briera (*Brier score, BS)* [Brier, 1950], czyli średni błąd kwadratowy jest definiowany następująco[109]:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2$$

gdzie: n oznacza liczbę przypadków, p_i prawdopodobieństwo zajścia modelowanego zdarzenia, y_i stan faktyczny (zdarzenie=1, brak zdarzenia =0). Im mniejsza wartość wyniku BS tym lepsze dopasowanie. Siłę dyskryminacyjną modelu można również ocenić porównując uzyskany wynik ze wskaźnikiem Briera obliczonego dla prognozy przypadkowej (PBS) [Prusak, 2005]:

$$PBS = \frac{1}{n}[RP\left(\frac{RP}{N} - 1\right)^2 + RN(\frac{RP}{N})^2]$$

Im niższa wartość wskaźnika Briera dla danego modelu w porównaniu z wartością PBS, tym większa siła predykcyjna modelu.

Kolejną miarą opartą na probabilistycznej interpretacji odpowiedzi modelu jest entropia krzyżowa (entropia wzajemna, *cross-entropy, logloss*). Entropia krzyżowa jest miarą wywodzącą się z teorii informacji definiowaną wzorem [Berrar, 2018]:

$$Entropia\ krzyżowa = -\frac{1}{n}\sum_{i=1}^{n} y_i log_2 p_i + (1 - y_i) log_2 (1 - p_i)$$

Im mniejsza wartość entropii tym większa siła predykcyjna modelu. Warto jeszcze raz podkreślić, że powyższe miary mają zastosowanie jedynie w przypadku, gdy uzyskane odpowiedzi modelu można odnieść do szacowanego prawdopodobieństwa wystąpienia modelowanego zdarzenia.

4.2. Strategie walidacji modeli retencji klientów

Jak zostało przedstawione we wcześniejszych rozdziałach, budowa modeli oceniających skłonność klientów do odejścia może być przeprowadzona za pomocą wielu

[109] Spotyka się również wzory, w których odjemna i odjemnik są zamienione pozycjami, co jednak po podniesieniu różnicy do kwadratu nie wpływa na wynik.

metod analitycznych, począwszy od metod liniowych, poprzez metody oparte na drzewach klasyfikacyjnych (decyzyjnych), metody oparte na sieciach neuronowych i im pokrewne, aż po modele hybrydowe. Przedstawiono tam również wiele miar oceny jakości działania modelu. Naturalną konsekwencją poruszanych zagadnień jest oszacowanie jakości modelu, jakiej można oczekiwać podczas stosowania go w praktyce w przyszłości dla nowych klientów. Jakość tę szacuje się bezpośrednio po zbudowaniu modelu (ocena *ex ante*).

Zbiór danych, którym dysponuje badacz jest podstawą zarówno do budowy modelu jak i do jego oceny. Ocena modelu jest wykonywana przy założeniu, że struktura populacji nowych klientów będzie zbliżona do tej reprezentowanej przez analizowany zbiór danych, oraz że wzorce zachowania klientów nie ulegną zmianie w trakcie stosowania modelu[110]. Uzyskane wyniki będzie można uznać za wiarygodne jedynie przy braku naruszenia powyższych założeń.

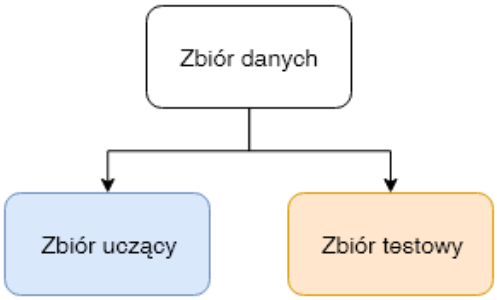
Strategie oceny jakości modelu można podzielić na dwie grupy:

- strategie polegające budowie modelu jedynie na części dostępnych danych, pozostała część przeznaczona jest do jego oceny;
- strategie polegające na budowie modelu na całym dostępnym zbiorze danych, ocena modelu jest następnie przeprowadzana za pomocą szeregu modeli pomocniczych budowanych na odpowiednio dobranych podzbiorach/podpróbach zbioru oryginalnego.

4.2.1. Metody oparte na podziale zbioru danych

Najprostszą, a zarazem najbardziej powszechnie używaną metodą walidacji modelu jest podejście oparte na podziale zbioru danych na dwa podzbiory: uczący oraz testowy. Na podstawie zbioru uczącego budowany jest model, a obliczony błąd nosi miano błędu dopasowania (aproksymacji). Zbiór testowy służy natomiast do oceny skuteczności zbudowanego modelu. Błąd obliczony na jego podstawie nosi nazwę błędu uogólnienia (generalizacji). F. E. Harrell sugeruje [2015], aby w zbiorze testowym znalazło się co najmniej 100 przypadków z mniej licznej klasy (tutaj osób nielojalnych), co w praktyce biznesowej nie stanowi realnego ograniczenia.

[110] Zachodzące w czasie zmiany w strukturze populacji klientów oraz zmiany wzorców ich zachowań wymagają cyklicznego badania skuteczności modelu już w trakcie jego eksploatacji (ocena *ex post*).

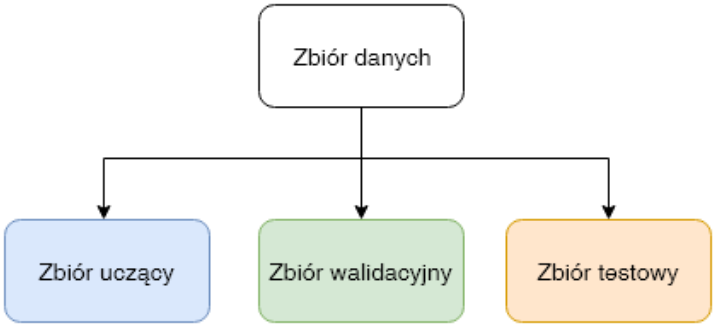


Rysunek 40 Podział zbioru danych na zbiór uczący i testowy
Źródło: opracowanie własne.

Podział zbioru danych do walidacji *ex ante* może być wykonany w następujący sposób:

- *Out of sample* – losowy podział próby wykorzystanej do modelowania na dwie klasy, gdzie zbiór testowy stanowi najczęściej 20-30% zbioru bazowego.
- *Out of time* – wykorzystanie do oceny nowej próby, z ostatniego okresu o takiej samej strukturze jak zbiór uczący.

Wybrane metody analityczne wymagają wykorzystania dodatkowego zbioru danych – zbioru walidacyjnego[111], który służy do określenia optymalnego momentu zatrzymania procesu budowy modelu. Na przykład dla drzew wzmacnianych (*boosted trees*) oraz losowego lasu (*random forests*), dodatkowy zbiór danych pozwala określić optymalną liczbę drzew, jaka tworzyła będzie docelowy model. W przypadku sieci neuronowych zbiór walidacyjny pozwala określić optymalną liczbę epok, przez którą należy uczyć model.



Rysunek 41 Podział zbioru danych na zbiór uczący, walidacyjny i testowy
Źródło: opracowanie własne.

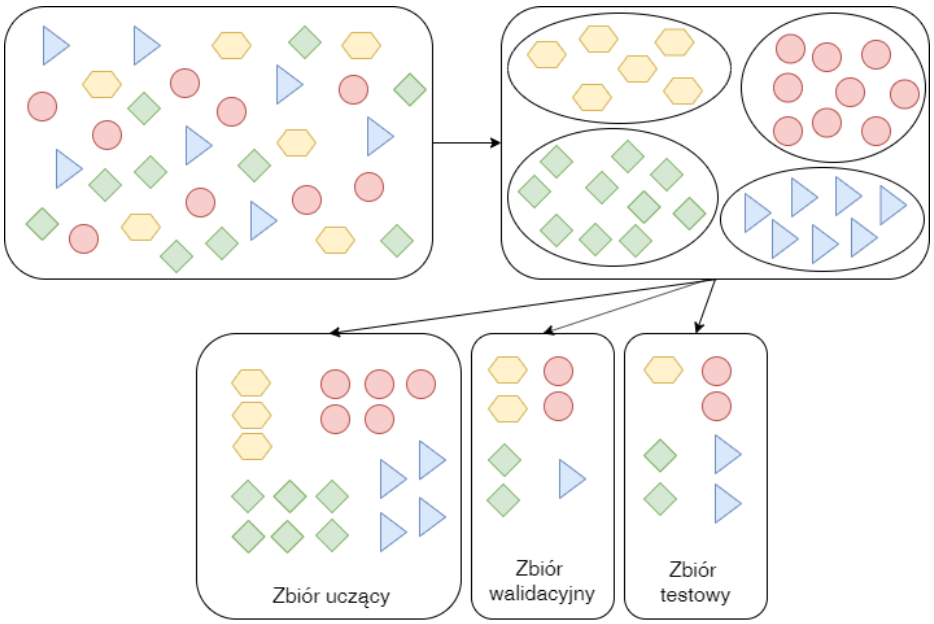
[111] W literaturze oraz narzędziach analitycznych brak zgodności co do terminologii. Nazwy „zbiór walidacyjny" oraz „zbiór testowy" bywają stosowana zamiennie.

Poprawnie przeprowadzona strategia bazująca na podziale zbioru danych na podzbiory wymaga, aby w każdym zbiorze znalazły się przypadki reprezentatywne dla całego zbioru. Aby zapewnić tę własność, po etapie podziału należy dokonać porównania rozkładów zarówno zmiennej zależnej, jak i (o ile to możliwe) kluczowych predyktorów. W przypadku stwierdzenia znaczących różnic należy powtórzyć procedurę losowania. Brak reprezentatywności któregokolwiek ze zbiorów będzie prowadził do błędnych wniosków na temat rzeczywistej siły predykcyjnej zbudowanego modelu. Ryzyko przypadkowego uzyskania niereprezentatywnych podzbiorów jest jedną z wad strategii losowego podziału. Po powtórzonym procesie podziału zbioru można uzyskać inne modele oraz inne oszacowanie dobroci dopasowania.

Strategią zmniejszającą ryzyko powstania niereprezentatywnych zbiorów w trakcie podziału na próby jest technika zaproponowana przez R. Tadeusiewicza i M. Szaleńca [2015]. Polega ona na pogrupowaniu przypadków znajdujących się w analizowanym zbiorze za pomocą wybranej techniki analizy skupień[112]. Po określeniu liczby skupień oraz dokonaniu przypisania każdego z elementów dookreślonego skupienia, proces podziału przypadków do zbiorów odbywa się niezależnie dla każdego skupienia, gwarantując tym samym lepszą reprezentatywność każdego z uzyskanych zbiorów. Wyboru reprezentantów można dokonywać losowo albo kierując się odległością od centrum skupienia.

W przypadku skupień małolicznych i jednocześnie odległych od pozostałych grup (o ile badacz zdecyduje się na uwzględnienie takich przypadków w zbiorze danych), zaleca się, aby wszystkie należące do nich przypadki trafiały do zbioru uczącego.

[112] Na przykład za pomocą metody k-średnich lub sieci neuronowych Kohonena.



Rysunek 42 Podział losowy w oparciu o analizę skupień
Źródło: Opracowanie własne na podstawie [Tadeusiewicz, Szaleniec, 2015]

Strategia podziału zbioru danych na próby, choć niezwykle popularna, nie jest jednak pozbawiona wad. Główna oś krytyki bazuje na fakcie, że podział danych znacznie zmniejsza wielkość próby, jaka służy zarówno do budowy jak i oceny modelu. Rezygnacja z części wzorców może znacząco wpłynąć na końcową postać modelu w porównaniu do modelu budowanego na podstawie pełnej dostępnej informacji. Fakt ten może mieć znaczenie zwłaszcza w sytuacji, gdy zbiór danych zawiera stosunkowo niewielką liczbę przypadków [Harrell, 2015].

4.2.2. Metody oparte na wielokrotnym próbkowaniu

Przedstawiona poniżej grupa strategii walidacji polega na zbudowaniu modelu na oryginalnym zbiorze danych, a następnie przeprowadzeniu jego oceny za pomocą analizy błędów obliczonych na podstawie pomocniczych modeli zbudowanych na podzbiorach zbioru oryginalnego tworzonych za pomocą różnego rodzaju technik losowania (zarówno losowania prostego jak i ze zwracaniem).

Wielokrotny (v-krotny) sprawdzian krzyżowy (*v-fold cross validation*) polega na podziale oryginalnego zbioru danych na *v* w przybliżeniu równolicznych podzbiorów. Wartość *v* jest określana przez badacza, jednak zazwyczaj przyjmuje się, że jest to 10. Każdy przygotowany podzbiór jest sukcesywnie oddzielany od pozostałych tworząc próbę testową. Uzyskany błąd modelu jest następnie uśredniany. Dziesięciokrotny sprawdzian

krzyżowy wymaga zatem zbudowania 10 dodatkowych modeli, w celu oszacowania błędu modelu zbudowanego na oryginalnym zbiorze danych. Strategia ta jest stosowana w przypadku mniejszych zbiorów danych w porównaniu z prostym podziałem dla danej precyzji oszacowania [Harrell, 2015]. Niewątpliwie słabą stroną tego podejścia podobnie jak wszystkich opartych na wielokrotnym próbkowaniu jest wysoki koszt obliczeniowy, który może być kluczowym ograniczeniem w odniesieniu do bardziej złożonych obliczeniowo metod analitycznych. Dodatkowo, jednokrotne wykonanie przedstawionej procedury nie pozwala na precyzyjne oszacowanie błędu modelu. Uzyskanie dokładnego oszacowania błędu zazwyczaj wymaga ponad 20-krotnego przeprowadzenia procedury[113].

Skrajną, rzadziej wykorzystywaną strategią v-krotnego sprawdzianu krzyżowego jest technika LOO (*leave one out*) polegająca na podziale zbioru danych na liczbę podzbiorów równą liczbie przypadków. Zatem za każdym razem podczas budowy modelu pomocniczego próba testowa jest reprezentowana przez jeden przypadek.

W praktyce walidacji modelu można spotkać się z wieloma podejściami opartymi na *bootstrapie*[114]. Najprostsza z nich polega na wylosowaniu (ze zwracaniem) próby przypadków o liczności równej liczności zbioru uczącego. Na podstawie wylosowanego zbioru budowany jest model pomocniczy. Rolę zbioru testowego dla tego modelu pełni zbiór oryginalny. Procedurę losowania i budowy modelu powtarza się wielokrotnie (najczęściej nie rzadziej niż kilkaset razy). Ostatecznie błąd generalizacji jest obliczony jako uśrednienie błędów modeli pomocniczych uzyskanych dla próby oryginalnej.

Ulepszona procedura oparta na *bootstrapie* opiera się na mniej bezpośrednim sposobie szacowania błędu generalizacji. Każda iteracja pozwala na uzyskanie współczynnika optymizmu, który jest różnicą pomiędzy błędem uzyskanym dla próby *bootstrap* a błędem uzyskanym dla próby oryginalnej. Po wielokrotnym powtórzeniu procedury obliczania optymizmu obliczany jest średni poziom optymizmu, który jest odejmowany od wyniku finalnego modelu uzyskanego na całym zbiorze danych. Uzyskana różnica jest nieobciążonym oszacowaniem błędu generalizacji zbudowanego modelu.

Przedstawione powyżej techniki mogą niekiedy prowadzić do nadmiernego oszacowania optymizmu [Harrell, 2015], dlatego zalecaną strategią opartą na *bootstrapie* jest tak zwany *bootstrap .632+*. Na jego podstawie optymizm jest obliczany jako:

$$Optymizm = 0{,}632 * (\epsilon_m - \epsilon_0),$$

[113] Czyli zbudowania ponad 200 modeli pomocniczych.
[114] Podejście zaproponowane przez B Efrona [1979].

gdzie ϵ_m to precyzja oryginalnego modelu, ϵ_m to średnia ważona precyzja obliczona na podstawie przypadków, jakie nie zostały wylosowane do próby *boostrap*. Procedura ta nie ma teoretycznego uzasadnienia i tylko badania empiryczne potwierdziły dobre zachowanie *bootstrapu .632+* [Arlot, Cellise, 2010]. Efektywność walidacji *bootstrap* na podstawie zbioru oryginalnego jest porównywalna z oceną na podstawie zbioru testowego dwukrotnie większego od zbioru oryginalnego [Harrell, 2015].

4.3. Określanie optymalnego punktu odcięcia (*cut-off*)

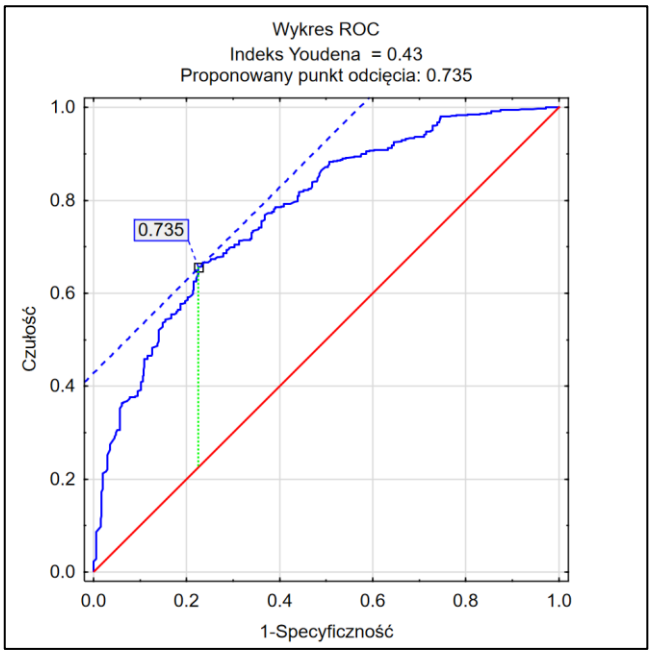
Ustalenie optymalnego punktu odcięcia modelu jest ostatnim kluczowym etapem poprzedzającym jego wdrożenie. Polega na wskazaniu granicznej wartości odpowiedzi modelu, która oddzieliłaby osoby lojalne od nielojalnych. Poniżej przedstawiono wybór metod i strategii używanych w procesie wyboru punktu odcięcia. Podczas ich analizy zwrócono uwagę na ich założenia dotyczące:

- kosztów błędnych klasyfikacji,
- prawdopodobieństwa *a priori* pojawienia się osób nielojalnych w zbiorze danych,
- zyskowności rozumianej w kategoriach ekonomicznych,
- praktycznej możliwości implementacji „optymalnego" punktu odcięcia.

Najprostszą strategią często stosowaną zwłaszcza przez mniej doświadczonych badaczy jest przyjęcie punktu odcięcia na poziomie 0,5 odpowiedzi modelu. Strategia ta jest intuicyjna i wskazuje, że należy prognozować zajście modelowanego zdarzenia, jeśli prawdopodobieństwo *a posteriori* jest większe niż 0,5. Intuicja ta prowadzi jednak do błędnych wniosków w sytuacji niezbalansowanych klas oraz kosztów błędnych klasyfikacji.[115]

Popularnym i prostym do zdefiniowania punktem odcięcia jest wartość odpowiedzi modelu, w której indeks J Youdena osiąga wartość maksymalną. Inna popularna choć nieco trudniejsza w implementacji strategia wskazuje, aby punkt odcięcia został wyznaczony przez styczną do krzywej ROC nachyloną względem osi X pod kątem 45°.

[115] Można spotkać strategie, które dążą do takiego przeskalowania prawdopodobieństwa *a posteriori*, aby punkt odcięcia równy 0,5 był punktem optymalnym. Nie jest ona jednak zalecana, między innymi ze względu na utratę informacji o poziomie prawdopodobieństwa *a priori* zajścia modelowanego zdarzenia.



Rysunek 43 Punkt odcięcia na podstawie indeksu Youdena oraz stycznej do krzywej ROC
Źródło: opracowanie własne.

Na wykresie (Rysunek 43) przedstawiono punkt odcięcia wyznaczony za pomocą indeksu Youdena (pionowa, kropkowana linia łącząca linię modelu losowego z krzywą ROC) oraz metodą stycznej (przerywana niebieska linia styczna do krzywej ROC). Należy zauważyć, że obie strategie sprowadzają się do wskazania tej samej wartości punktu odcięcia. Zakładają zgodnie równe koszty błędnych klasyfikacji oraz równe prawdopodobieństwo *a priori* pojawienia się osoby lojalnej w zbiorze rozpoznawanym[116]. Nie uwzględniają one kwestii zyskowności oraz kwestii praktycznej implementacji modelu. Pomimo swojej popularności nie mogą być zalecane w modelach lojalności ze względu na powyższe własności.

Kolejna reguła wyznacza punkt odcięcia na poziomie równym prawdopodobieństwu *a priori* zajścia modelowanego zdarzenia. Jest ono szacowane na podstawie rozkładu zmiennej zależnej w modelowanej próbie. Według tej reguły osoba jest przypisywana do klasy nielojalnych, jeśli prawdopodobieństwo *a posteriori* odejścia uzyskane na podstawie modelu jest większe od prawdopodobieństwa *a priori*. W praktyce sprowadza się to do wyznaczenia granicy wyrażonej wzorem:

$$punkt\ odcięcia = \frac{RP}{N}$$

[116] Zbiorze klientów, na których model będzie wdrażany.

Reguła ta może być stosowana jedynie pod warunkiem, że odpowiedzi modelu są skalibrowane do rzeczywistego poziomu prawdopodobieństwa rezygnacji klienta z usługi obserwowanego w zbiorze danych. Czynnikami, które wpływają na zaburzenie wartości tych odpowiedzi są:

- stosowanie technik *oversampling* oraz *downsampling* podczas zmiany struktury zbioru uczącego,
- zmiana niektórych hiperparametrów np. prawdopodobieństwa *a priori* w algorytmach drzew klasyfikacyjnych,
- stosowanie metod modelowania nie zapewniających zgodności z bazowym poziomem skłonności do odejścia.

Uzyskanie wyników zgodnych z rzeczywistym poziomem skłonności klientów do odejścia klienta jest możliwe po uprzedniej kalibracji odpowiedzi modelu. Aby przeprowadzić kalibrację wyników, w pierwszym kroku dla zbioru zawierającego rzeczywistą bądź pożądaną proporcję klas zmiennej zależnej wdraża się zbudowany uprzednio model. Następnie na podstawie uzyskanych wyników buduje się model regresji logistycznej [Kuhn, Johnson, 2013], w którym rolę jedynego predyktora pełni uzyskana wcześniej odpowiedź modelu. Odpowiedź modelu regresji logistycznej [117] jest skalibrowanym wynikiem, który umożliwia już stosowanie powyższej formuły.

Metodą, która bierze pod uwagę zarówno rozkład klas zmiennej zależnej, jak i koszty błędnych klasyfikacji jest formuła polegająca na korekcie wartości współczynnika kierunkowego *m* stycznej do krzywej ROC w optymalnym punkcie [Zweig, Campbell, 1993].

$$m = \frac{koszt\ FP}{koszt\ FN} * \frac{1 - P}{P}$$

Wartość P odnosi się do prawdopodobieństwa *a priori* występowania w zbiorze danych modelowanej klasy. Dla równych kosztów błędnych klasyfikacji oraz prawdopodobieństwa *a priori* równego 0,5 reguła ta daje wynik równy wynikowi opartemu na indeksie J Youdena.

Innym sposobem na znalezienie optymalnego punktu odcięcia przedstawionym przez [Kuhn, Johnson, 2013] jest zdefiniowanie PCF (*Probability Cost Function*).

[117] W przypadku regresji logistycznej suma prawdopodobieństw *a posteriori* obliczona dla wszystkich przypadków uczących jest równa liczbie przypadków modelowanej klasy.

$$PCF = \frac{P * koszt\ FP}{P * koszt\ FN + (1 - P) * koszt\ FP}$$

Wartość PCF jest następnie elementem wskaźnika NEC (*Normalized Expected Cost*).

$$NEC = PCF * (1 - TP) + (1 - PCF) * FP$$

Optymalny punkt odcięcia to punkt, w którym PCF przyjmuje wartość minimalną.

Powyższe reguły nie brały pod uwagę możliwości zysku generowanego przez model a jedynie koszty popełnianych pomyłek oraz prawdopodobieństwo *a priori*. Uzupełnieniem reguły przedstawionej w [Zweig, Cambell, 1993] o aspekt zysków, jakie płyną z poprawnych klasyfikacji jest wzór zaproponowany przez [Stein, 2005]. Współczynnik kierunkowy prostej stycznej do krzywej ROC będzie miał wtedy następującą postać:

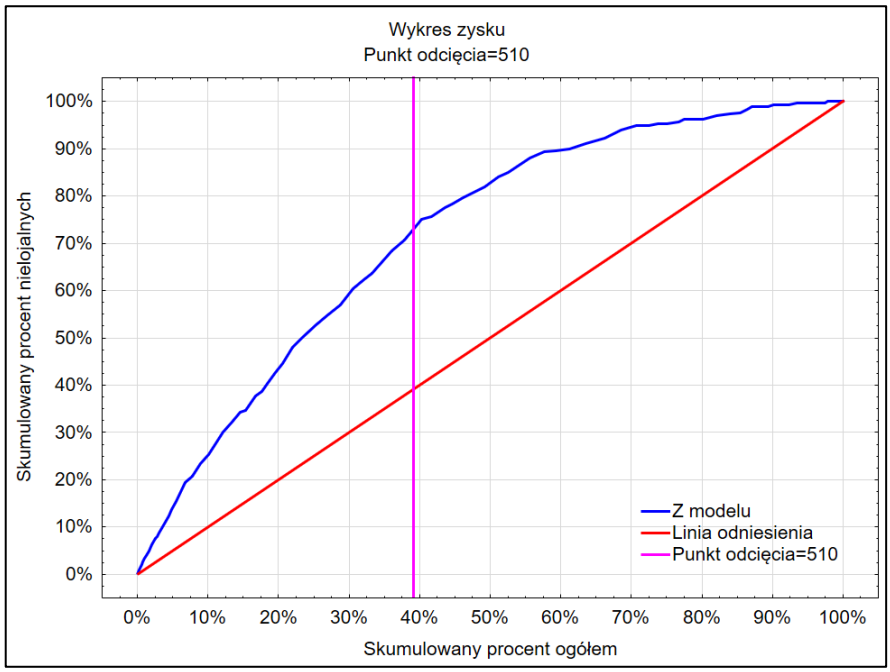
$$m = \frac{koszt\ FP + zysk\ TN}{koszt\ FN + zysk\ TP} * \frac{1 - P}{P}$$

W przypadku modeli lojalności:

- *koszt FP* może być kosztem kontaktu z klientem, który nie miał zamiaru odejść;
- *koszt FN* może być związany z utratą korzyści związaną z odejściem klienta;
- *zysk TP* może wiązać się korzyściami związanymi z zatrzymaniem klientów, których intencją było odejście do konkurencji;
- *zysk TN* nie ma odniesienia w modelach lojalności.

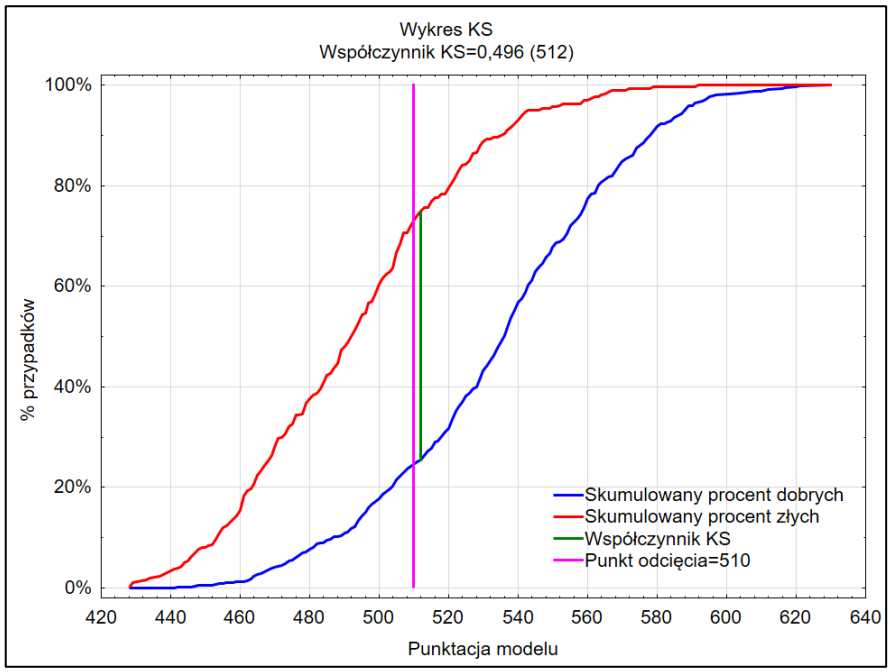
Należy zwrócić uwagę, że stosowanie wszystkich powyższych miar ma sens jedynie w przypadku analizy jednorodnej grupy klientów nie różniących się miedzy sobą poziomem generowanych zysków.

W praktyce biznesowej wybór optymalnego modelu oraz optymalnego punktu odcięcia nie opiera się jedynie na ocenie siły predykcyjnej modelu. Modele o sile predykcyjnej na ogólnie wyższym poziomie mogą zostać odrzucone ze względu na słabszą moc dyskryminacyjną w obszarze stosowalności modelu. W sytuacji, gdy względy biznesowe w sposób jednoznaczny wyznaczają obszar decyzyjny modelu, podstawą wyboru optymalnego punktu odcięcia może być wykres korzyści przedstawiający wszystkie możliwe kompromisy pomiędzy odsetkiem osób wskazanych jako nielojalne a odsetkiem wychwyconych osób nielojalnych. Podobną rolę może pełnić wykres przyrostu.



Rysunek 44 Wykres zysku z punktem odcięcia
Źródło: opracowanie własne.

Po wyznaczeniu punktu odcięcia dobrą praktyką jest zbadanie, na ile wybrany punkt jest z godny z punktem optymalnym z punktu widzenia miary KS. Im punkt odcięcia jest bliższy punktowi optymalnemu, tym w większym stopniu model wykorzystuje swoją siłę dyskryminacyjną.



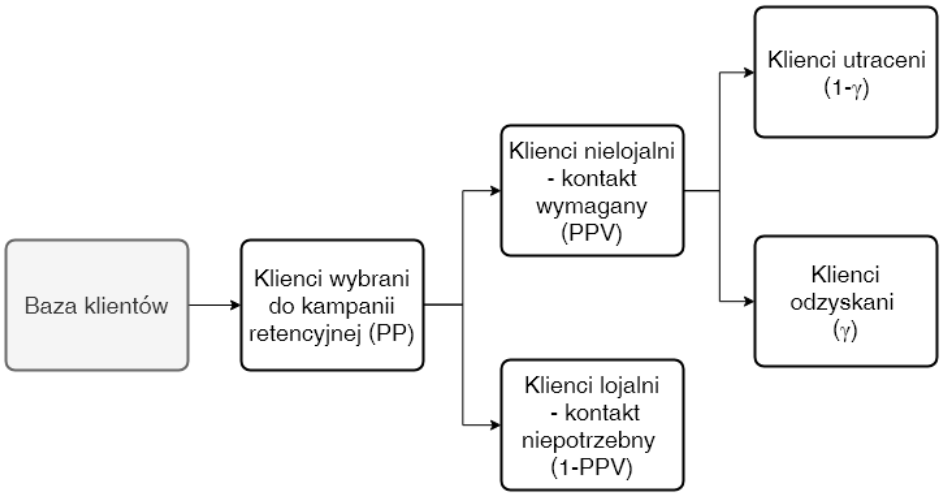
Rysunek 45 Wykres KS z punktem odcięcia
Źródło: opracowanie własne.

Powyższe rozważania przyjmowały założenie, że w praktyce biznesowej model jest podstawą do podjęcia jednoznacznej decyzji o przekroczeniu bądź nie akceptowalnego poziomu ryzyka odejścia. Innym możliwym podejściem stosowanym w praktyce jest wprowadzenie kilku punktów odcięcia dzielących klientów na kilka grup (np. klientów z niskim ryzykiem odejścia, średnim ryzykiem odejścia i wysokim ryzykiem odejścia) i uzależnienie sposobu reakcji od klasy ryzyka oraz poziomu zyskowności klienta.

Wykres zysku pokazuje jaki jest zysk ze stosowania modelu, biorąc pod uwagę koszt ponoszony w wyniku każdego rodzaju pomyłki (*KosztFP*, *KosztFN*) i zysk z trafnie przewidzianej klasyfikacji (*ZyskTN*, *ZyskTP*). Dla każdego punktu odcięcia generowany jest wynik (na przykład wyrażony w danej walucie), jaki jest z nim powiązany. Ogólny wzór na zysk dla wybranego punktu odcięcia przedstawia się następująco:

$$Zysk = TP * ZyskTP + TN * ZyskTN - FP * KosztFP - FN * KosztFN$$

Bazując na powyższej, ogólnej zasadzie, zaproponowano [Verbeke i inni, 2012] kompleksowe podejście do wyboru optymalnego punktu odcięcia w odniesieniu do modeli lojalności. Podejście to bierze pod uwagę grupy klientów przedstawione na Rysunek 46.



Rysunek 46 Odejścia i retencja klientów w ramach bazy klientów
Źródło: opracowanie własne na podstawie [Verbeke i inni, 2012].

Zaproponowany sposób wyboru optymalnego punktu odcięcia wiąże się z implementacją poniższego wzoru:

$$Zysk = PP[PPV * \gamma * (CLV - c - \delta) + PPV * (1 - \gamma) * (-c) + (1 - PPV) * (-c - \delta)] - A$$

gdzie: γ – odsetek osób nielojalnych, jakie zareagowały pozytywnie na zachęty ze strony dostawcy, CLV – średnia wartość życiowa klienta, który zdecydował się na pozostanie, c – koszt kontaktu z klientem, δ – koszt „zachęty" dla klienta w celu podtrzymania relacji, A – stały koszt obsługi zarządzania programem retencyjnym (*anty-churn).*

Parametr PPV jest zależny od siły predykcyjnej zbudowanego modelu oraz od zastosowanego punktu odcięcia. Powyższy wzór jest podstawą do wykreślenia krzywej zysku. Punkt optymalny to punkt, w którym krzywa przyjmuje wartość maksymalną. W przypadku przyjęcia realistycznych ustawień, wartości dodatnie mogą zostać zaobserwowane jedynie dla niewielkiego odsetka klientów o największej skłonności do odejścia ze stosunkowo korzystną relacją PPV w stosunku do (1-PPV).

4.4. Wybór klientów do strategii sprzedażowych i retencyjnych (*uplift modeling*)

Rozważane dotychczas strategie modelowania oraz wyboru optymalnego punktu odcięcia nie brały pod uwagę kwestii zróżnicowania reakcji klienta na ewentualny kontakt ze strony dostawcy. Kontakt z grupą osób najbardziej skłonnych do odejścia nie zawsze

musi zakończyć się zatrzymaniem klienta. Paradoksalnie, działania retencyjne mogą być impulsem do podjęcia przez klienta decyzji o odejściu.

Tabela 10 Korzyści i straty w zależności od kontaktu/braku kontaktu oraz pierwotnego nastawienia klienta

		Jeśli będzie kontakt	
		Nie odejdzie	Odejdzie
Jeśli nie będzie kontaktu	Nie odejdzie	Zdecydowani pozostać (niepotrzebny koszt)	Uśpieni (negatywny wpływ podjętych działań)
	Odejdzie	Podatni na perswazję (korzyści z podjętych działań)	Zdecydowani odejść (niepotrzebny koszt)

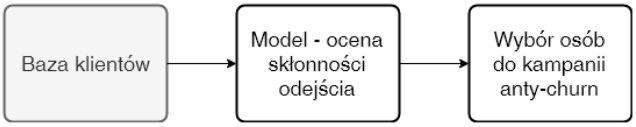
Źródło: opracowanie własne na podstawie [Kane i inni, 2014].

Możliwe zachowania klienta oraz korzyści lub straty w zależności od faktu podjęcia (lub nie) działań mających na celu jego zatrzymanie przedstawia Tabela 10.

Na jej podstawie można zauważyć, że jedynie kierowanie ofert do osób z komórki „Podatni na perswazję" może przynieść korzyści finansowe. Pozostałe warianty generują straty związane bądź z samym kosztem kontaktu („Zdecydowani odejść") dodatkowo powiększonym przez możliwą zbyt hojną ofertę („Zdecydowani pozostać"), lub przez utratę klienta wynikłą z kontaktu („Uśpieni"). Negatywny efekt widoczny jest szczególne w przypadku osób z grupy „Uśpieni", którym oferta przedłużenia umowy może przypomnieć o upływającym terminie jej wygaśnięcia stając się katalizatorem działań związanych z odejściem do konkurencji.

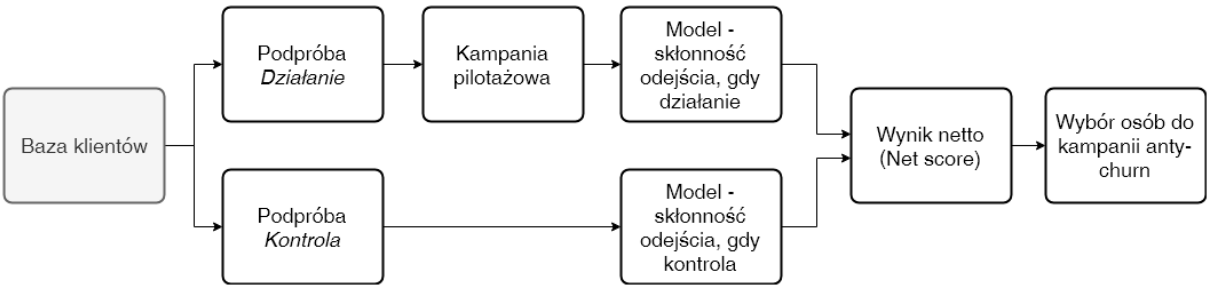
4.4.1. Budowa modeli *uplift*

Cel budowanego modelu powinien koncentrować się zatem na dotarciu jedynie do osób „Podatnych na perswazję" i uniknięciu wysyłania ofert zwłaszcza do osób „Uśpionych" i „Zdecydowanych pozostać". Tradycyjne podejście do budowy modeli lojalności nie jest w stanie dokonać takiego rozróżnienia, koncentruje się bowiem na ocenie skłonności do odejścia nie biorąc pod uwagę wpływu samej kampanii (lub jej braku) na decyzję klienta. Schemat tradycyjnego podejścia do budowy modelu podsumowuje Rysunek 47.



Rysunek 47 Schemat budowy modelu - podejście tradycyjne
Źródło: opracowanie własne.

Uwzględnienie w procesie modelowania wpływu kampanii na sposób zachowania klienta jest wdrażane za pomocą strategii znanych pod zbiorczą nazwą *uplift modeling*. Najprostsza strategia budowy modeli *uplift* polega na zbudowaniu dwóch osobnych modeli zgodnie ze schematem zawartym na Rysunek 48.



Rysunek 48 Modelowanie uplift - budowa niezależnych modeli
Źródło: opracowanie własne na podstawie [Rzepakowski, Jaroszewicz 2012].

Podejście to zakłada zbudowanie dwóch niezależnych modeli. Pierwszego na podstawie zbioru klientów, dla których przeprowadzono kampanię utrzymaniową (zbiór taki można określić mianem „Działanie"). Model taki oceniał będzie skłonność klientów do odejścia pod warunkiem, że otrzymali ofertę utrzymaniową. Drugi model jest budowany na zbiorze klientów, do których nie skierowano oferty utrzymaniowej (zbiór taki można określić mianem „Kontrola"). Model taki będzie oceniał skłonność do odejścia pod warunkiem, że klient nie otrzymał oferty utrzymaniowej. Następnie obydwa modele stosujemy na bazie klientów, spośród których pragniemy wskazać osoby, jakie wezmą udział w planowanej kampanii.

Dla każdego klienta z tej próby otrzymuje się zatem dwa zestawy prawdopodobieństw *a posteriori*:

- *P(Odejdzie/Działanie)* = Prawdopodobieństwo odejścia klienta przy założeniu, że został do niego wykonany kontakt;

– *P(Odejdzie/Kontrola)* = Prawdopodobieństwo odejścia klienta przy założeniu, że nie został do niego wykonany kontakt.

Uzyskane wyniki są podstawą obliczenia wskaźnika *Net Score (Net lift, Lift score)*:

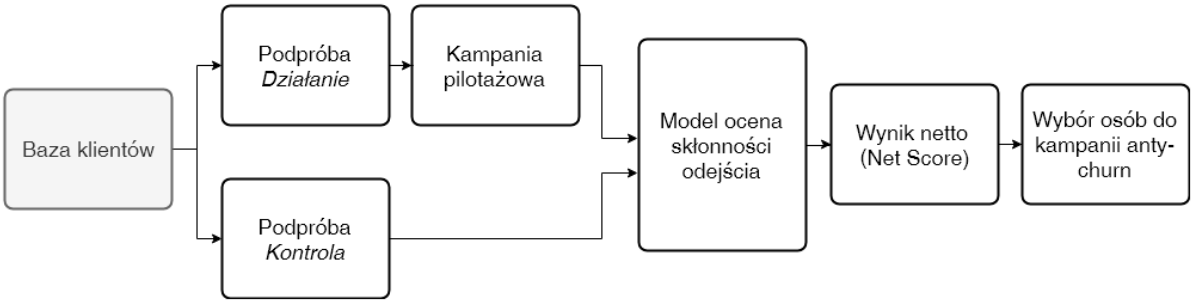
$$Net\ Score = P(Odejdzie/Kontrola) - P(Odejdzie/Działanie)$$

W podejściu „tradycyjnym" ocenia się skłonność klientów do odejścia o ile zostanie podjęte w stosunku do nich działanie. Model *uplift* ocenia zmianę skłonności klientów, gdy podjęto wobec nich działanie w odniesieniu do sytuacji, gdy od tego działania się powstrzymano. Najbardziej interesujący klienci to ci, dla których wartość współczynnika *Net Score* przyjmuje maksymalne dodatnie wartości. W przypadku klientów „uśpionych" wartości tego wskaźnika będą przyjmować wartości ujemne.

Wskaźnik *Net Score* jest podstawą do utworzenia wykresu przyrostu netto (*net lift*), który tworzy się po posortowaniu *Net Score* w porządku malejącym i przedstawieniu ich względem percentyli (oś X). Miara *Net Score* może być podstawą wyznaczania punktu odcięcia na przykład za pomocą krzywej zysku prezentowanej w niniejszej pracy.

Słabą stroną przedstawionego podejścia jest konieczność budowy dwóch modeli. Dodatkowo każdy z modeli bazuje na zmiennych powiązanych ze zmienną zależną Odejdzie/Nie odejdzie i może ignorować zmienne, które są odpowiedzialne za rzeczywisty efekt netto. Dodatkowo odpowiedzi modelu powinny zostać poddane kalibracji, aby uwzględnić efekt różnych frakcji osób nielojalnych w grupie *Działanie* oraz grupie *Kontrola*.

Inna strategia *uplift modeling* zakłada, że budowany jest jedynie jeden model, na podstawie zbioru do którego należą zarówno osoby, które otrzymały ofertę („Działanie"), jak również osób, które jej nie otrzymały („Kontrola") zgodnie ze schematem zaprezentowanym na Rysunek 49.



Rysunek 49 Model Uplift z wykorzystaniem zmiennej informującej o grupach Działanie i Kontrola

Źródło: Opracowanie własne na podstawie [Lo, 2002].

W zbiorze danych umieszcza się dodatkową zmienną informującą czy dany przypadek należy do grupy kontrolnej czy z działaniem (np. o nazwie *DziałanieKontrola*). Etap doboru zmiennych do modelu jest przeprowadzany osobno dla obydwóch grup. Następnie zestaw wybranych predyktorów jest uzupełniany o interakcje[118] ze zmienną informującą o przynależności do grupy kontrolnej lub z działaniem. Po zbudowaniu modelu należy wygenerować na jego podstawie dwa zestawy prawdopodobieństw *a posteriori*:

- *P(Odejdzie/Działanie),* gdy wszystkie wartości zmiennej *DziałanieKontrola* są ustawione na 1.
- *P(Odejdzie/Kontrola),* gdy wszystkie wartości zmiennej *DziałanieKontrola* są ustawione na 0.

Uzyskane wyniki są podstawą obliczenia wskaźnika *Net Score* w sposób analogiczny do prezentowanego powyżej niwelując jednocześnie jego wady. Budowany jest jeden model na większym zbiorze danych zamiast dwóch na mniejszym, a sama wartość *Net Sco*re jest szacowana w sposób bardziej bezpośredni, na podstawie jednego modelu. Pociąga jednak za sobą konieczność utworzenia dużej liczby zmiennych pochodnych i może generować trudności numeryczne na etapie oszacowania modelu[119].

Kolejne podejście w budowie modelu *uplift* również zakłada powstanie jednego modelu [Kane, 2014]. Zmienna zależna w tym przypadku jest zmienną przyjmującą cztery kategorie: Działanie/Odejście (*Treatment Positives* – *TrP*), Działanie/Brak odejścia (*Treatment Negatives* – *TrN*), Kontrola/Odejście (*Control Positives* – CP) oraz Kontrola/Brak odejścia (*Control Negatives* – CN). Schemat budowy modelu jest zatem

[118] Zmienną będącą iloczynem dwóch zmiennych.
[119] Domyślną metodą dla tego podejścia jest regresja logistyczna.

analogiczny do Rysunek 49, chociaż w odmienny sposób oblicza się wskaźnik *Net score*. Warto zauważyć, że osoby „podatne na perswazję", mogą należeć do grupy TrN lub CP, natomiast osoby z grupy „uśpieni" do dwóch pozostałych grup.

$$Net\ score = \left[\frac{P(TrN)}{P(Tr)} + \frac{P(CP)}{P(C)}\right] - \left[\frac{P(TrP)}{P(Tr)} + \frac{P(CN)}{P(C)}\right]$$

Wadą tego podejścia jest niewątpliwie konieczność budowy modelu dla zmiennej nominalnej posiadającej cztery warianty, jednakże, jak podkreśla autor, daje ona najlepsze wyniki w badaniach porównawczych.

Podobna strategia została zaproponowana w pracy [Jaśkowski, Jaroszewicz, 2012]. Autorzy zaproponowali jednak stworzenie binarnej zmiennej zależnej przez połączenie grup TrN i CP („podatni na perswazję") oraz TrP i CN („uśpieni"). Po zbudowaniu modelu *Net score* oblicza się w następujący sposób:

$$Net\ score = 2 * P(Podatni\ na\ perswazję) - 1$$

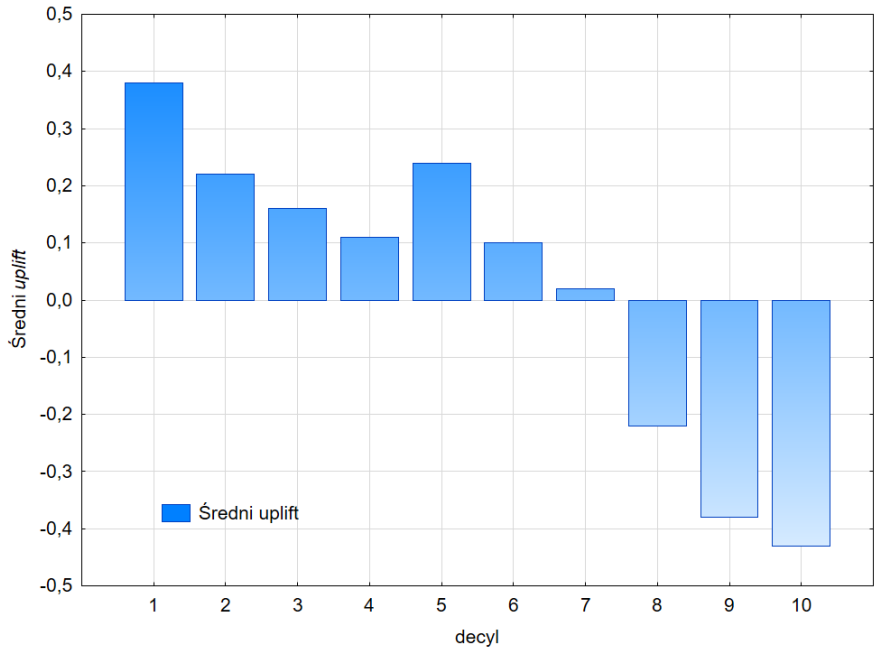
4.4.2. Ocena modeli *uplift*

Badacz oceniający związek przyczynowy napotyka na oczywistą trudność wynikającą z faktu, że dany klient może znaleźć się jedynie w jednej sytuacji (Działanie lub Kontrola). Niemożliwe jest zatem określenie efektu przyczynowego dla pojedynczej osoby, ponieważ nie może być ona jednocześnie w obydwóch grupach. Dzięki podejściu statystycznemu problem można przenieść z poziomu jednostki na poziom populacji, z której pochodzi dana jednostka [Trzciński, 2009]. Ocena modeli opiera się na założeniu, że osoby podobnie ocenione przez model będą zachowywały się w podobny sposób. Można zatem ocenić efekt przyczynowy porównując zachowanie grupy osób podobnych (względem prawdopodobieństwa *a posteriori* modelu) znajdujących się w grupach *Działanie* i *Kontrola*. Przykładową oceną efektu może być różnica w odsetku osób, które odeszły do konkurencji zaobserwowana w obydwóch grupach.

Najprostszym sposobem oceny modelu jest posortowanie wartości *Net Score* malejąco a następnie dyskretyzacja jej wartości na wybraną liczbę równolicznych przedziałów. W ramach wyznaczonych przedziałów oblicza się średnią wartość *uplift* zgodne z następującym wzorem [Kane i inni, 2014]:

$$Uplift = \frac{TrP}{Tr} - \frac{CP}{C}$$

Wyniki można następnie przedstawić na wykresie słupkowym (Rysunek 50).



Rysunek 50 Średni *uplift* względem kolejnych decyli klientów
Źródło: opracowanie własne.

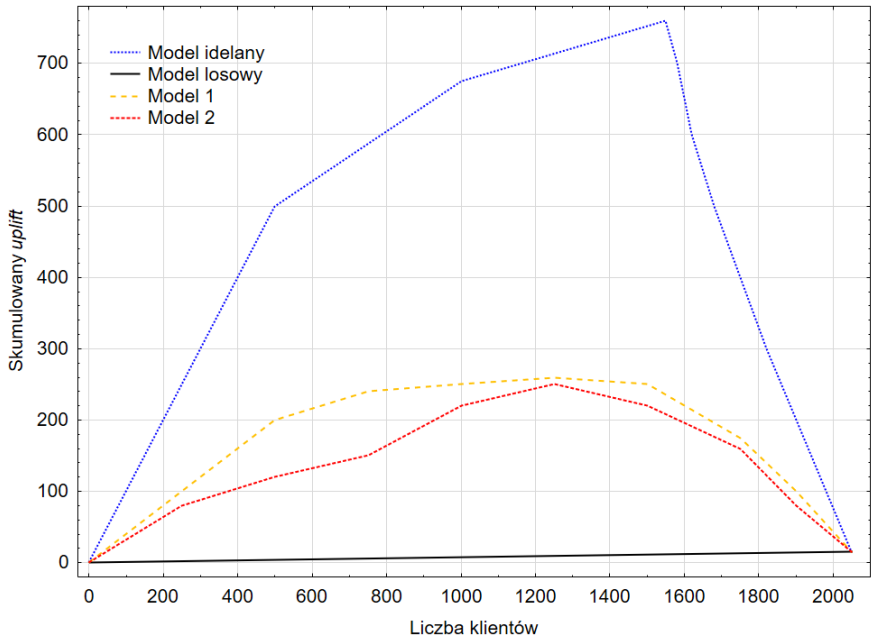
Na jego podstawie możliwa jest ocena siły efektu w poszczególnych przedziałach oraz wskazanie klas, w których obserwuje się możliwy negatywny efekt podjętych działań. Wykres ten daje możliwość oceny stabilności działania modelu, który w idealnym przypadku powinien cechować się monotonicznym liniowym spadkiem średniej wartości *uplift*. K. Kane i inni [2014] zaproponowali oszacowanie modelu regresji liniowej, w którym zmienną zależną jest wartość *uplift*, a predyktorem numer przedziału. Współczynnik determinacji R^2 otrzymany dla tak oszacowanego modelu były podstawą oceny stabilności jego działania[120].

Powyższa metoda nie pozwala na syntetyczną ocenę skuteczności modelu, ani na porównanie kilku modeli równocześnie. Narzędziem umożliwiającym realizację tych celów jest krzywa *uplift* (*uplift curve*). Krzywą tę konstruuje się sortując uprzednio uzyskane wartości *Net Score*, następnie na osi X prezentowana jest (skumulowana) liczba klientów, a na osi Y prezentowane są wartości *uplift* obliczone dla każdej skumulowanej liczby klientów według wzoru [Radcliffe, Surry, 2011]:

[120] Autorzy zwracają jednak uwagę na konieczność łącznej oceny współczynnika R^2 oraz wykresu

$$Uplift = \left(\frac{TrP}{Tr} - \frac{CP}{C}\right) * (Tr + C)$$

Wykres pozwala na porównanie modelu oraz wybór najlepszego w sposób analogiczny do krzywych ROC, porównując pola powierzchni pod krzywymi *uplift* oraz analizując ewentualne punkty przecięcia.

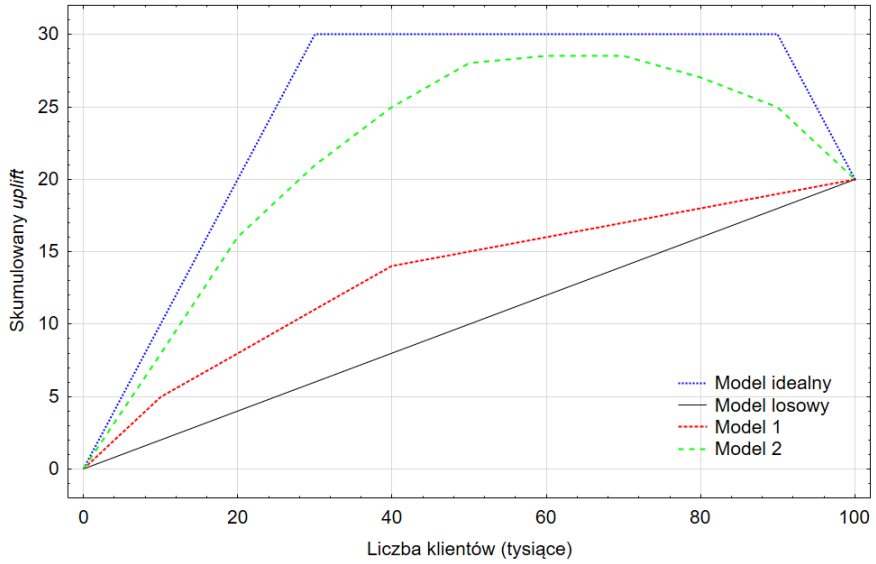


Rysunek 51 Krzywa *uplift*
Źródło: opracowanie własne.

Kolejnym narzędziem służącym do oceny i porównania modeli *uplift* jest krzywa Quini [Radcliffe, Surry, 2011]. Wykres ten jest konstruowany analogicznie do krzywej CAP. Na osi X prezentowane jest liczba klientów posortowanych malejąco względem wartości *Net Score* na osi Y prezentowane są skumulowane wartości *uplift* obliczone według wzoru [Radcliffe, 2007]:

$$Uplift = TrP - \frac{CP * Tr}{C}$$

Na podstawie wykresu możliwe jest obliczenie syntetycznej miary siły predykcyjnej dla modelu *uplift* – Quini Q. Oblicza się ją jako stosunek pola powierzchni krzywej Quini (powyżej diagonalnej reprezentującej losowy model) do pola dla modelu idealnego.



Rysunek 52 Krzywa Quini - krzywa CAP dla modeli uplift
Źródło: opracowanie własne na podstawie [Radcliffe, 2007].

Interpretacja Q jest analogiczna do współczynnika Giniego. Przyjmuje on wartość w zakresie [-1,1], wartość 0 informuje o braku siły predykcyjnej, a 1 o idealnym dopasowaniu, wartości ujemne o działaniu modelu gorszym niż losowe.

Rozdział 5 Identyfikacja optymalnej ścieżki selekcji modelu migracji klientów

5.1. Określenie celu analizy

Budowa modeli retencji klienta realizowana została zgodnie z paradygmatem budowy modeli *data mining* zakładającym wtórne wykorzystanie danych gromadzonych w wyniku realizacji standardowych procesów biznesowych. Implikuje to pracę na zastanych zbiorach danych zgromadzonych w systemach informatycznych przedsiębiorstw. W pracy wykorzystane zostało doświadczenie autora w budowie modeli retencji klienta na rzeczywistych zbiorach danych klientów z branży telekomunikacyjnej, ubezpieczeniowej i usługowej. Ze względu na ograniczenia związane z poufnością danych ich bezpośrednie wykorzystanie nie było możliwe podczas planowanych symulacji i eksperymentów. Wnioski i wiedza płynące ze zrealizowanych projektów zostały uwzględniane w pracy w sposób pośredni. Podstawą analizy był zbiór danych dostępny w domenie publicznej (opisany w dalszej części rozdziału) cechujący się wystarczającą złożonością oraz wolumenem.

Procedura badawcza polegała na wykonaniu badań symulacyjnych oceniających wpływ determinant wpływających na jakość modeli migracji klientów oraz określenie relacji między nimi. W oparciu o dostępne zbiory danych zbudowano szereg modeli zgodnie z metodyką CRISP-DM. Podczas symulacji były brane pod uwagę następujące czynniki:

- ***T**ransformation* – sposób przygotowania predyktorów, dyskretyzacja, standaryzacja itp.,

- ***I**nteraction* – fakt uzupełnienia zbioru danych o zmienne pochodne (*derived variables*),
- ***V**ariables* – sposób doboru zmiennych do modelu,
- ***H**yperparameters* – metody optymalizacji hiperparametrów,
- ***E**nsembles* – dodatkowe strategie uczenia: segmentacja, hybrydyzacja, agregacja modeli.

Aspekty TIVHE zostały uwzględnione w sposób uwzględniający specyfikę wykorzystywanych metod analitycznych.

W powszechnym przekonaniu wielu analityków jedynym kryterium determinującym jakość modelu klasyfikacyjnego jest jego siła predykcyjna. Ten aspekt działania modeli pozwala na ich porównanie i pewną obiektywizację oceny ich działania[121]. Nie ma jednej powszechnie przyjętej jednej miary oceniającej ten wymiar działania modelu, jednak do najczęściej wykorzystywanych należą niewątpliwie pole powierzchni pod krzywą ROC oraz miary z nią związane, a także miara Kołmogorowa-Smirnowa (KS) oraz przyrost (*lift*).

W praktyce siła predykcyjna jest tylko jednym z wielu kryteriów, jakie brane są pod uwagę podczas budowy modelu. Innymi kryteriami biznesowymi są:

- łatwość interpretacji modelu,
- logiczność zależności opisywanych przez model,
- liczba zmiennych w modelu,
- współliniowość predyktorów,
- stabilność działania modelu w czasie,
- wrażliwość na zmiany w strukturze populacji klientów.

Z technicznego punktu widzenia istotnymi kryteriami są także łatwość implementacji i utrzymania oraz szybkość oceny pojedynczego klienta. Wymienione kryteria biznesowe oraz akcent, jaki jest na nie kładziony podczas analizy w dużym stopniu determinują początkowy wybór metod analitycznych tworząc naturalny podział na metody interpretowalne oraz tzw. czarne skrzynki. Kryteria te zostały uwzględnione podczas

[121] Kryterium siły predykcyjnej jest *de facto* jedynym branym pod uwagę w konkursach na najlepszy model.

budowy modeli regresji logistycznej oraz drzew klasyfikacyjnych w dalszej części pracy. Metody „czarnoskrzynkowe" zostaną ocenione jedynie pod kątem siły predykcyjnej. Aspekt techniczny – w największym stopniu zależny od uwarunkowań przedsiębiorstwa oraz narzędzi analitycznych – nie zostanie wzięty pod uwagę.

5.2. Zrozumienie i przygotowanie danych

Analizowany zbiór danych zawiera 100 tys. obserwacji, a każda z nich dotyczy jednego abonenta pewnej sieci telefonicznej. Zbiór zawiera 173 zmienne z czego jedna pełni w modelowaniu rolę zmiennej zależnej, 171 rolę zmiennych niezależnych, a 1 zmienna jest identyfikatorem klienta. Spośród potencjalnych predyktorów 55 to predyktory jakościowe, a 116 to predyktory ilościowe. Dostępne zmienne X opisują cechy demograficzne klientów, sposób korzystania z usługi, preferencje związane z kontaktami z biurem obsługi klienta oraz historię tych kontaktów. Opis poszczególnych zmiennych znajduje się w dodatku zamieszczonym na końcu pracy.

5.2.1. Ocena jakości danych

Pierwszym krokiem analizy jest ocena jakości danych oraz identyfikacja głównych wyzwań związanych z ich przygotowaniem i czyszczeniem. Na początku ocenia się rozkład zmiennej zależnej (Tabela 11). Jak łatwo zauważyć, rozkład osób lojalnych i nielojalnych jest zrównoważony z lekką przewagą osób lojalnych[122].

Tabela 11 Rozkład wartości zmiennej zależnej

Churn	Liczba	Procent
Nie	50438	50,44
Tak	49562	49,56

Źródło: Opracowanie własne.

Obserwowany rozkład zmiennej Y i względnie duża liczebność zbioru obserwacji dają komfort podczas budowy modelu. Liczba ta pozwala na wykorzystanie bardziej zaawansowanych metod uczenia maszynowego cechujących się większymi wymaganiami co do liczby przypadków należących do każdej z klas. Zbalansowany rozkład zmiennej

[122] Obserwowany rozkład jest niezmiernie rzadko obserwowany w rzeczywistości, jego kształt jest najprawdopodobniej efektem doboru przypadków wykonanego przez twórców zbioru.

zależnej nie wymusza stosowania korygujących go metod próbkowania takich jak *down sampling* czy *SMOTE*. W analizie zachowano zatem oryginalną strukturę danych.

Po ocenie rozkładu zmiennej zależnej, kolejnym krokiem jest ocena skali braków danych w zmiennych X. Na podstawie przeprowadzonej analizy, której szczegóły prezentuje Tabela 12 można stwierdzić, że ponad połowa zmiennych jest kompletna, nie zawiera żadnych braków. Ponad 27% zmiennych charakteryzuje się niewielkim odsetkiem braków (w większości są to pojedyncze przypadki). Znaczące problemy związane z brakami danych są widoczne w 32 zmiennych objaśniających, z czego 7 z nich można zaliczyć do tzw. zmiennych rzadkich (*sparse data*). Te ostatnie są praktyce najczęściej pomijane w dalszej analizie.

Tabela 12 Podsumowanie poziomu kompletności danych

Poziom kompletności	Liczba zmiennych	Procent zmiennych
100%	92	53,8
powyżej 95%	47	27,49
od 5% do 95%	25	14,62
poniżej 5%	7	4,09

Źródło: Opracowanie własne.

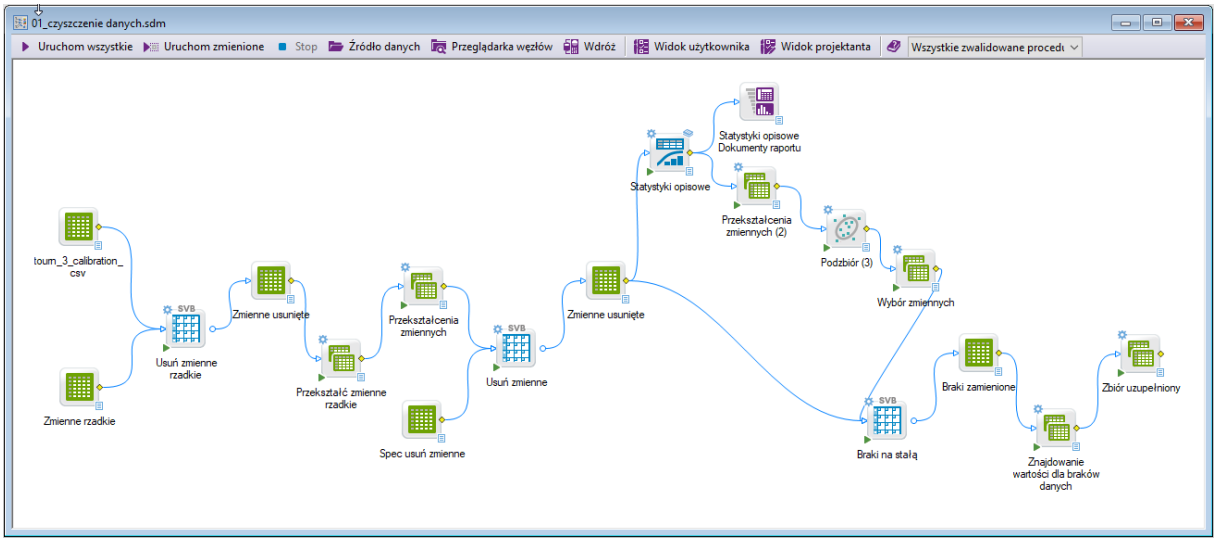
W kolejnym kroku oceniono siłę wpływu braków danych na zmienną zależną. W tym celu zestaw predyktorów przekształcono do zestawu zmiennych binarnych, w których klasa 1 oznacza brak danych, natomiast 0 oznacza kompletną obserwację. Siła predykcyjna braków danych została oceniona za pomocą miar IV (*Information Value*) oraz GINI. Praktycznie we wszystkich zmiennych X nie stwierdzono związku braków danych ze zmienną Y. Jedynie wyniki dla trzech predyktorów wykazują słaby związek z modelowaną zmienną zależną (szczegóły przedstawia Tabela 13). Uzyskany wynik sugeruje możliwość przydatności analizowanych zmiennych w postaci binarnych predyktorów „podano informację/brak informacji". Warto zwrócić uwagę, że wszystkie wyróżnione zmienne należą do grupy zmiennych rzadkich.

Tabela 13 Zmienne rzadkie wykazujące słabą moc predykcyjną

Zmienna	IV	GINI	% Ważnych
retdays	0,02	0,03	3,98
tot_acpt	0,02	0,03	3,98
tot_ret	0,02	0,03	3,98

Źródło: Opracowanie własne.

Sposób uzupełniania braków danych został uzależniony od skali pomiarowej danego predyktora oraz od liczby braków. W przypadku predyktorów jakościowych braki danych dla zmiennych zawierających niewielki ich odsetek (poniżej 5% braków) zostały zamienione wartością najczęstszą (modą). Braki w pozostałych zmiennych jakościowych zostały uzupełnione o dodatkową klasę „brak danych" (BD). W przypadku zmiennych ilościowych z grupy „powyżej 95%" (względnie kompletnych) braki danych zostały zamienione za pomocą mediany[123]. Dla zmiennych z grupy „od 5% do 95%" przyjęto strategię imputacji opartą na dyskretyzacji. Podczas dyskretyzacji zmiennych granice kategorii wyznaczone zostały na podstawie decyli rozkładu, natomiast „brak danych" stanowił jedenastą kategorię przekodowanej zmiennej. Wszystkie zmienne rzadkie zostały usunięte z analizy. Zmienne pochodne oparte na zmiennych rzadkich (Tabela 13) pozostały w zbiorze zgodnie ze wcześniejszym opisem. Proces czyszczenia został wykonany z przestrzeni roboczej programu Tibco Statistica 13.3 (Rysunek 53) umożliwiającej automatyzację i powtarzalność realizowanego procesu. W przestrzeni wykorzystano zarówno węzły dostępne wraz z programem jak i też stworzone od podstaw przez autora do celów niniejszej analizy[124].



Rysunek 53 Przestrzeń robocza do imputacji braków danych

Źródło: opracowanie własne.

[123] Brak związku braków danych ze zmienną zależną pozwala przyjąć hipotezę, że są to braki losowe bądź czysto losowe. Niewielka skala braków uzasadnia wybór prostej metody imputacji. Więcej na temat rodzajów braków danych oraz sposobów ich imputacji można znaleźć w rozdziale 2.

[124] Jak przedstawiono we wcześniejszych rozdziałach, niektóre metody, na przykład drzewa klasyfikacyjne i regresyjne CART oraz budowane na ich podstawie zespoły modeli są odporne na braki danych. Autor przyjął konwencję budowy modeli na podstawie kompletnego zbioru nie korzystając z tych własności wybranych metod.

Po zakończonym procesie imputacji braków danych uzyskano arkusz z kompletnymi obserwacjami, który był podstawą dalszej analizy.

5.2.2. Segmentacja zbioru danych

Kolejnym krokiem analizy była wstępna ocena siły predykcyjnej poszczególnych predyktorów za pomocą miary IV. Jej celem jest ogólna ocena analizowanego zbioru oraz identyfikacja potencjalnych zmiennych segmentacyjnych. Na podstawie uzyskanych wyników (Tabela 14) stwierdzono, że mniej niż 20% predyktorów charakteryzuje się dosyć słabą mocą predykcyjną. Nie zanotowano zmiennych o dużej mocy predykcyjnej Jedynie jedna zmienna wykazywała przeciętną moc predykcyjną.

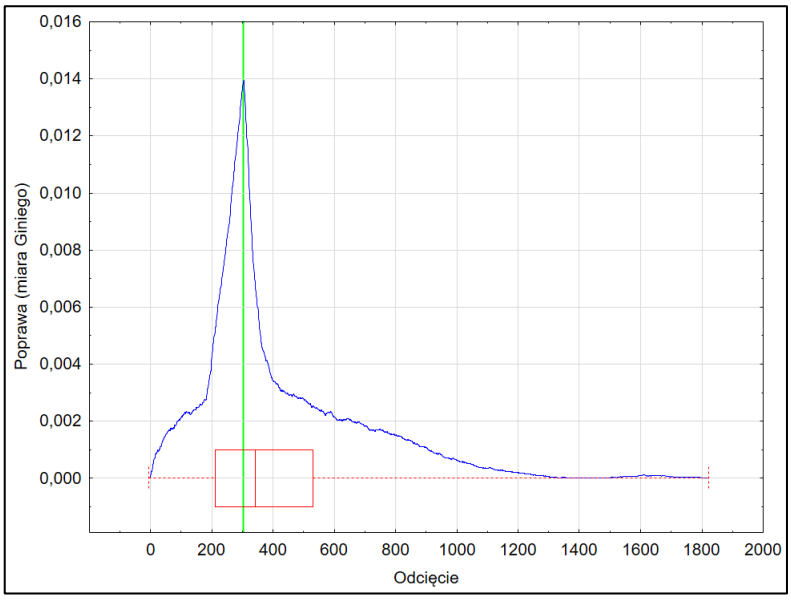
Tabela 14 Podsumowanie mocy predykcyjnej predyktorów dla miary IV

Miara IV	Liczba predyktorów	Procent predyktorów	Interpretacja
[0,3, +∞)	0	0	duża moc predykcyjna
[0,1, 0,3)	1	0,59	przeciętna moc predykcyjna
[0,02, 0,1)	32	18,93	słaba moc predykcyjna
[0,01, 0,02)	54	31,95	brak mocy predykcyjnej
[0, 0,01)	82	48,52	brak mocy predykcyjnej

Źródło: Opracowanie własne.

Zmienną charakteryzującą się przeciętną mocą predykcyjną jest zmienna *equpdays* informująca o liczbie dni korzystania z bieżącego urządzenia. Chęć zmiany operatora związana z tą zmienną może być powiązana z terminem wygaśnięcia umowy terminowej (nieobecnej w zbiorze danych)[125]. Analiza czułości wykonana w za pomocą drzew klasyfikacyjnych CART (Rysunek 54) wskazuje na optymalny punkt podziału tej zmiennej w okolicach 300 (dni).

[125] Przedłużenie umowy bardzo często skutkuje wymianą aparatu na nowy.

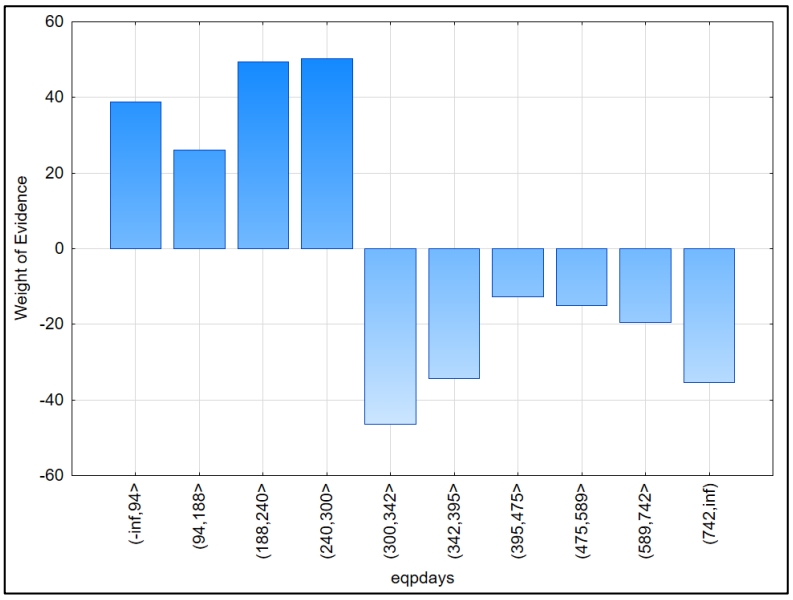


Rysunek 54 Wykres czułości dla zmiennej *eqpdays*

Źródło: opracowanie własne.

Profil ryzyka zmiany operatora dla zmiennej *eqpdays* wskazuje na większą skłonność do odejścia dla osób posiadających urządzenie powyżej 300 dni. Wartości WoE poniżej zera świadczą o większej skłonności klientów do rezygnacji z usługi niż przeciętnie obserwowana w zbiorze[126].

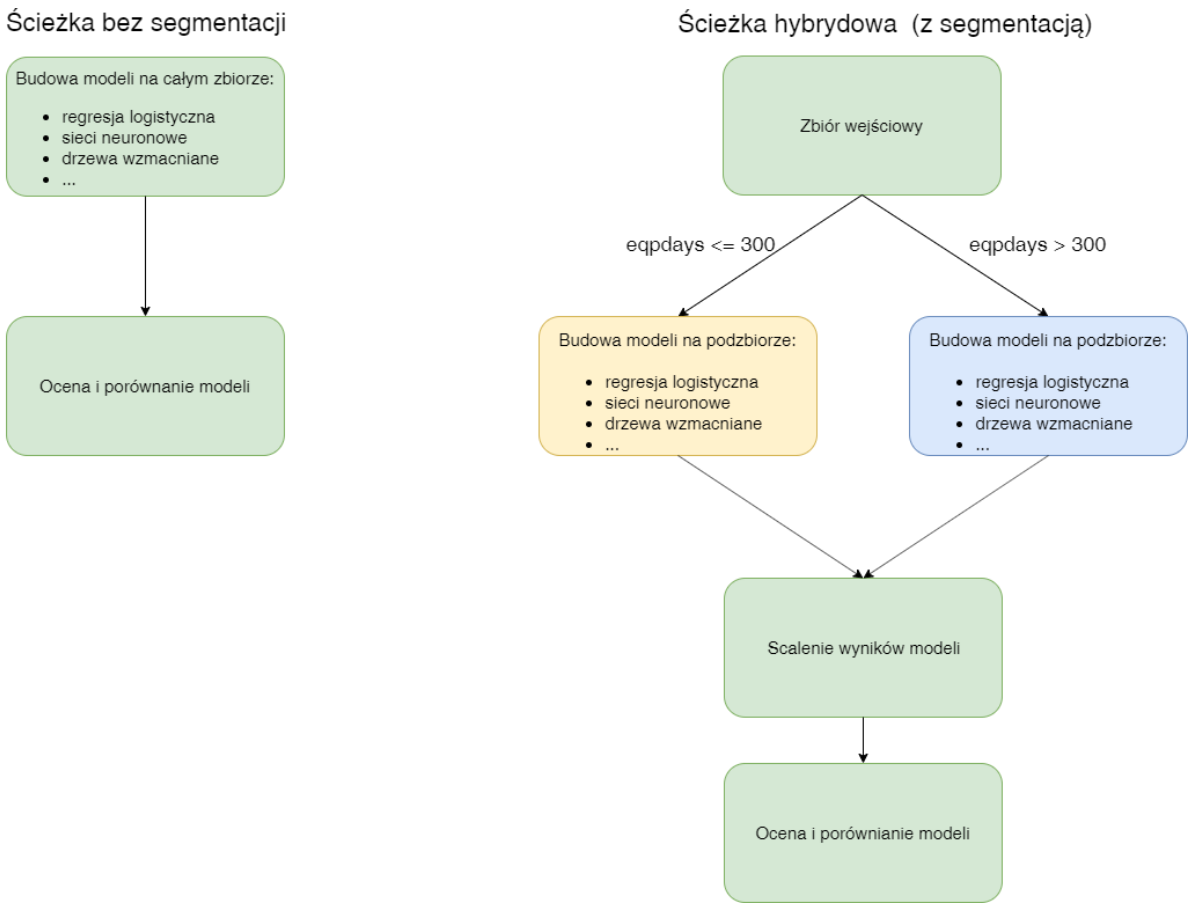
[126] Profil ryzyka zmiennej eqpdays jest analogiczny do profili obserwowanych przez autora podczas komercyjnych projektów dla klientów między innymi z branży telekomunikacyjnej. W analizowanych zbiorach fakt wygaśnięcia umowy terminowej zmieniał zarówno poziom ryzyka odejścia, jak również listę czynników wpływających na to ryzyko. Zazwyczaj jednak tego typu zmienna charakteryzowała się większą siłą predykcyjną. Podobną sytuację można zaobserwować w przypadku zmiennej *eqpdays*, co jest wystarczającą przesłanką, aby zmienna ta pełniła w analizie rolę zmiennej segmentacyjnej.



Rysunek 55 Profil ryzyka zmiennej *eqpdays*

Źródło: opracowanie własne.

Na jej podstawie może zostać dokonany podział (segmentacja) zbioru danych na dwa podzbiory, dla których możliwa będzie budowa odrębnych modeli migracji klientów. Podejście to pozwala zatem na budowę modeli hybrydowych, w zależności od przyjętej metody modelowania jest to na przykład *CART – logit*, bądź *CART – sieć neuronowa*. Poza modelami hybrydowymi do celów porównawczych zostały również zbudowane modele na podstawie całego zbioru danych, bez hybrydyzacji. Niezależne ścieżki analizy przedstawia Rysunek 56.



Rysunek 56 Ścieżki wykonanych analiz

Źródło: opracowanie własne.

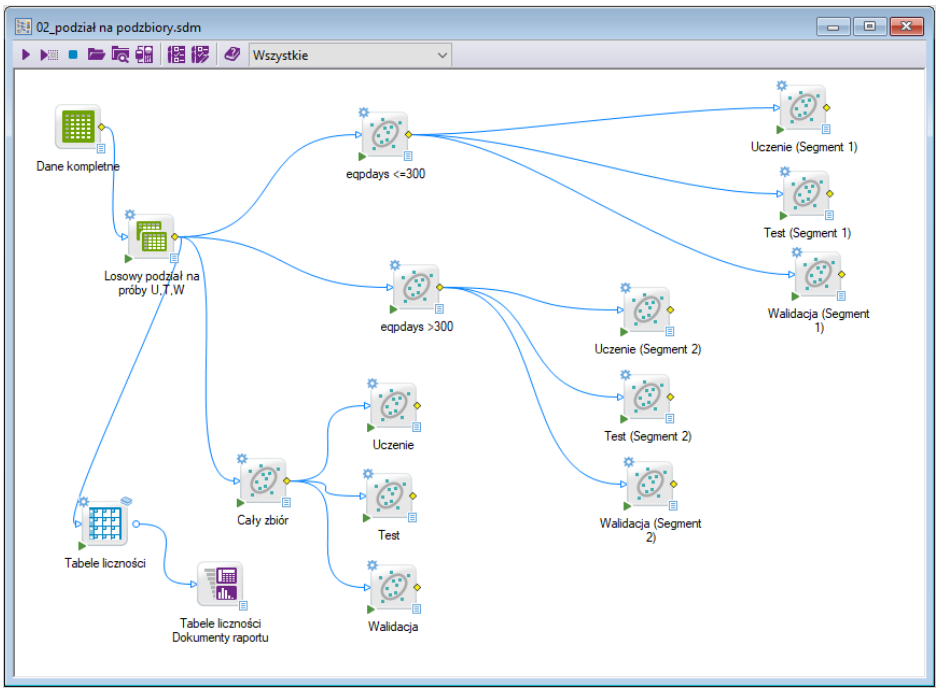
5.2.3. Dodatkowa kategoryzacja zmiennych jakościowych

Analiza rozkładów zmiennych jakościowych wykazała występowanie w zbiorze danych zmiennych posiadających od kilkudziesięciu do kilkuset kategorii. Na etapie doboru zmiennych taka liczba kategorii generuje ryzyko identyfikacji pozornych związków wynikających z artefaktów zawartych w danych. Na etapie modelowania wykorzystanie takich zmiennych może powodować ryzyko nadmiernego dopasowania modelu do danych i spadek jego zdolności do generalizacji. Dla tych zmiennych przeprowadzono zatem ponowną kategoryzację za pomocą drzew CART o głębokości 4 pozwalających na redukcję liczby kategorii do maksymalnie 16 grup. Wyjątkiem była zmienna CSA zawierająca prawie 1000 kategorii z czego wiele z nich odnosiło się jedynie do jednego przypadku. Zmienna ta została usunięta z analizy.

5.2.4. Podział zbioru danych

Przed wykonaniem kolejnych kroków analizy dokonano podziału każdego ze zbiorów danych na trzy podzbiory; uczący, który stanowił około 70% przypadków oraz testowy i walidacyjny zawierające po 15% przypadków [127]. Wszelkie dalsze transformacje zmiennych, czyli: optymalizacja hiperparametrów oraz ostateczny wybór zmiennych dla każdej z użytych metod przeprowadzone zostały na podstawie zbioru uczącego. Zbiór testowy pełnił w analizie pomocniczą rolę umożliwiając na przykład zatrzymanie procesu uczenia czy też ocenę wyborów dokonanych przez algorytmy, bądź przez badacza oraz do wyłonienia najlepszych modeli. Zbiór walidacyjny użyty został jedynie do końcowej oceny zbudowanych modeli pod kątem ich ewentualnego nadmiernego dopasowania.

W celu zapewnienia możliwości porównania wyników modeli bez segmentacji z modelami uwzględniającymi segmentację, podział na próby został wykonany w pierwszej kolejności. Po jego przeprowadzeniu zrealizowany został podział segmentacyjny. Przebieg tego procesu obrazuje Rysunek 57.



Rysunek 57 Przebieg procesu podziału na próby oraz wyodrębnienia segmentów do podejścia hybrydowego

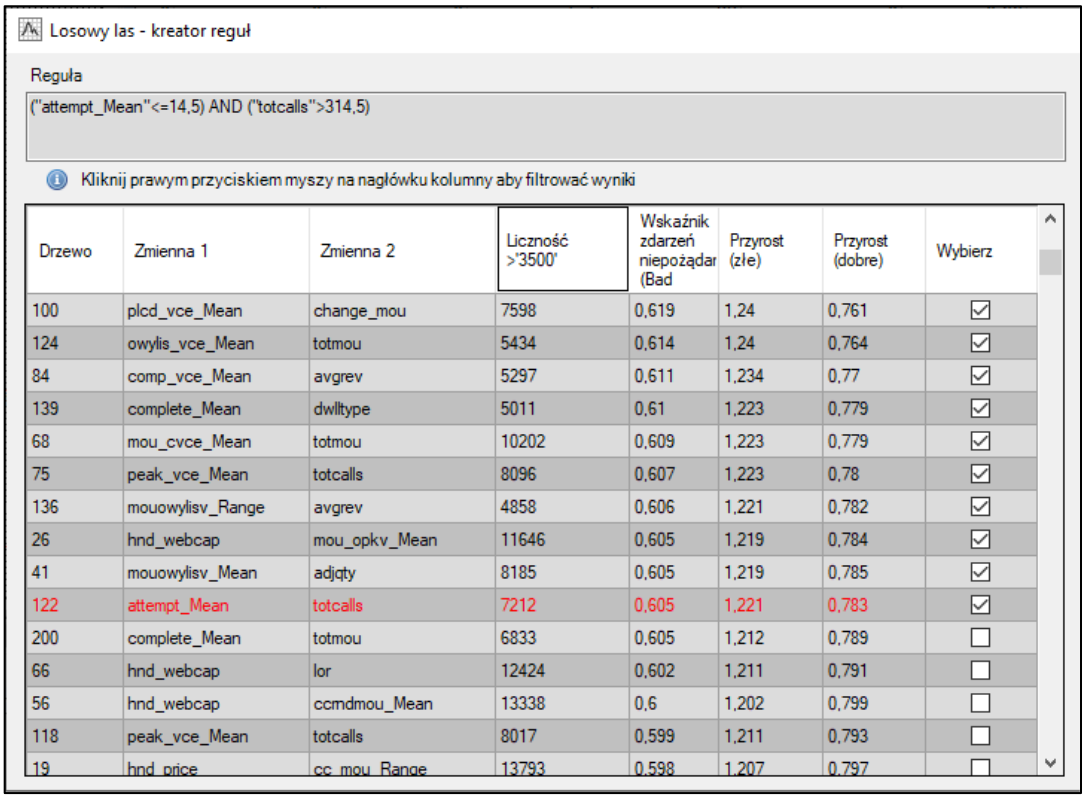
Źródło: opracowanie własne.

[127] Przyjęto konwencję nazw zbiorów zgodną z programem TIBCO Statistica.

5.2.5. Zmienne pochodne w modelu

Niezwykle istotnym aspektem przygotowania danych jest etap definiowania zmiennych pochodnych. Proces ten wymaga połączenia wiedzy biznesowej, jaką dysponuje zazwyczaj ekspert (menedżer) oraz wiedzy z zakresu analizy danych posiadanej przez analityka. W opisywanym tutaj przykładzie efekt synergii był niemożliwy do osiągnięcia, nie mniej jednak na podstawie doświadczeń autora płynących ze zrealizowanych projektów o podobnej charakterystyce zostały przygotowane dodatkowe zmienne pochodne, które uzupełniły zbiór potencjalnych predyktorów. Przykładowo do zbioru danych została dodana zmienna wskaźnikowa *diff_hnd_price* będąca różnicą w cenie bieżącego i poprzedniego urządzenia, jakim dysponował klient.

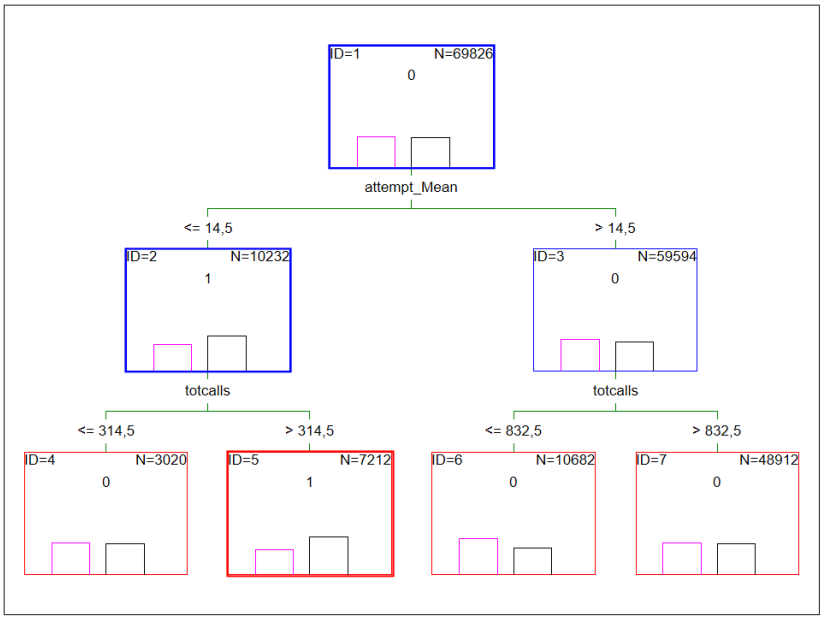
W procesie identyfikacji zmiennych pochodnych wykorzystano również podejście automatyczne oparte na metodzie losowego lasu. Podejście to polega na zbudowaniu zestawu relatywnie płytkich drzew o głębokości 2 lub 3. W przypadku realizowanej pracy badawczej przyjęto głębokość na poziomie 2. Po zbudowaniu zestawu drzew, każdy liść przekształcany jest na niezależną regułę opisującą podziały od korzenia drzewa do liścia. Fakt losowego doboru predyktorów dla każdego z drzew zaimplementowany w losowym lesie pozwala na identyfikację potencjalnie interesujących reguł, niemożliwych do identyfikacji za pomocą tradycyjnych algorytmów drzew klasyfikacyjnych i regresyjnych. Uzyskane reguły filtrowane są pod kątem liczby przypadków (powszechność reguły) oraz na podstawie odsetka nielojalnych klientów spełniających tę regułę i przyrostu (siła reguły). Najbardziej interesujące menedżerów reguły mogą zostać uwzględnione w polityce marketingowej czy też strategii działania zespołów utrzymania klienta. Mogą też zostać przekształcone na zmienne binarne (na zasadzie informacji, że dany przypadek spełnia regułę lub jej nie spełnia) i zasilić zbiór potencjalnych predyktorów.



Rysunek 58 Kreator reguł opartych na metodzie Losowy Las wraz z przykładową regułą
Źródło: opracowanie własne.

W przypadku analizowanego zbioru przyjęto zasadę, że do zbioru danych zostanie dołączonych 20 najsilniejszych reguł, które były spełnione dla co najmniej 5% przypadków. Spośród wybranych reguł połowa odnosiła się do sytuacji podwyższonej skłonności do rezygnacji, połowa do podwyższonego poziomu lojalności. Jeżeli w grupie najsilniejszych reguł pojawiały się reguły o takich samych elementach, wybierano spośród nich tę o największej mocy dyskryminacyjnej.

Przykładową regułę otrzymaną w wyniku analizy zbioru danych bez segmentacji (zbioru uczącego) przedstawia Rysunek 58. Drzewo, na podstawie którego została utworzona reguła przedstawia Rysunek 59. Na rysunku wyróżniono węzły tworzące przykładową regułę.



Rysunek 59 Przykładowe drzewo, będące źródłem reguły

Źródło: opracowanie własne.

5.2.6. Wstępna selekcja zmiennych

Przygotowane zbiory danych zostały poddane wstępnej selekcji zmiennych. Wybór predyktorów został oparty w pierwszej kolejności na filtrach ukierunkowanych a następnie na filtrach nieukierunkowanych. Spośród filtrów ukierunkowanych, do analizy wykorzystano wybrane popularne miary tj. kryterium informacyjne (*Information Value*, IV), współczynnik Giniego (GINI) oraz metodę ReliefF (w trzech odmianach). Próg akceptacji zmiennych dla miary IV oraz GINI określono na poziomie 0,02. W przypadku ReliefF uzyskane dla każdej odmiany wyniki podzielono na decyle i nadano im zgodne z nimi rangi. Zmiennym wykazującym najsłabszy związek ze zmienną zależną nadano rangę 1, zmiennym o najsilniejszym związku rangę 10. W kolejnym kroku wartości rang zostały zsumowane. Dla tak skonstruowanego wskaźnika przyjęto próg akceptacji na poziomie większym bądź równym 7 [128]. Do dalszej analizy zakwalifikowane zostały wszystkie zmienne, które osiągnęły wyznaczony próg dla przynajmniej jednej z wymienionych miar (IV, GINI, ReliefF). Zmienne niespełniające żadnego z powyższych kryteriów zostały usunięte z analizy.

W kolejnym kroku ograniczono nadmiarowość zmiennych wynikającą ze zbyt wysokiej korelacji pomiędzy predyktorami. Na tym etapie analizy pod uwagę wzięte

[128] Do obliczeń miar IV oraz GINI wykorzystano Zestaw Skoringowy Statistica, w przypadku miary ReliefF użyto pakietu CORELearn dostępnego w programie R.

zostały jednie predyktory ilościowe. Identyfikację grup skorelowanych zmiennych objaśniających przeprowadzono za pomocą analizy głównych składowych (PCA). Zastosowano tutaj dodatkowo rotację ładunków za pomocą znormalizowanej metody *Varimax* dążącej do maksymalizacji wariancji kwadratów ładunków w kolumnie macierzy[129]. Kryterium to jest tym lepiej spełnione im bardziej kwadraty ładunków są bliskie 0 lub 1. Metoda ta pozwala zatem na polaryzację kwadratów ładunków w obrębie kolumn umożliwiając identyfikację grup zmiennych o potencjalnie wysokiej korelacji.

Dany czynnik (główna składowa) mógł posiadać reprezentantów (zmienne X), jeżeli jego wartość własna była wyższa od 1. Dana zmienna zostawała reprezentantem czynnika, jeżeli wartość bezwzględna jej ładunku była większa od progu 0,8. Podejście to umożliwiło wyodrębnienie niewielkiej liczby głównych składowych zawierających zmienne silnie ze sobą skorelowane. Przykładowo dla zbioru uczącego bez segmentacji wyodrębniono 9 głównych składowych. Z dalszej analizy wyeliminowane zostały zmienne X skorelowane z innymi zmiennymi X powyżej 0,9 (wartość bezwzględna) i cechujące się niższą od nich siłą predykcyjną. W przypadku triad zmiennych stosowano analogiczną procedurę opierając się dodatkowo na średniej korelacji. Dla bardziej licznych głównych składowych odrzucano nie więcej niż połowę zmiennych[130]. Tabela 15 przedstawia przykładowe wyniki uzyskane za pomocą PCA dla zbioru uczącego bez podziału segmentacyjnego. Cztery zmienne, których ładunki z jednym z czynników miały wartość powyżej progu 0,8 okazały się wysoce skorelowane pomiędzy sobą, co było podstawą do odrzucenia części z nich. Na tym etapie najlepszą metodą eliminacji jest ocena ekspercka. W sytuacji braku możliwości takiej oceny zastosowano procedurę opartą na analizie siły predykcyjnej zmiennych oraz ich średniej korelacji z pozostałymi zmiennymi w grupie. Na tej podstawie z dalszej analizy wyeliminowano zmienne; *plcd_vce_Range* (najniższa siła predykcyjna, nie najwyższa korelacja) oraz *complete_Range* (najniższa korelacja).

[129] Rotacji poddawane są znormalizowane ładunki czynnikowe - surowe ładunki czynnikowe podzielone przez pierwiastki kwadratowe odpowiednich zasobów zmienności wspólnej.

[130] Ostateczna selekcja zmiennych wykonana została w dalszej części analizy za pomocą metod wbudowanych w algorytmy uczące bądź też za pomocą metod je opakowujących.

Tabela 15 Przykładowa macierz korelacji uzupełniona o miary siły predykcyjnej

	plcd_vce Range	attempt Range	comp_vce Range	complete Range	Średnia korelacja	IV	GINI
plcd_vce Range	1,0000	0,9958	0,9551	0,9492	0,9750	0,01	0,01
attempt Range	0,9958	1,0000	0,9511	0,9556	0,9756	0,01	0,01
comp_vce Range	0,9551	0,9511	1,0000	0,9928	0,9747	0,01	0,02
complete Range	0,9492	0,9556	0,9928	1,0000	0,9744	0,01	0,02

Źródło: Opracowanie własne.

Podobną procedurę zastosowano dla wszystkich macierzy korelacji wyodrębnionych za pomocą PCA, prowadząc do kolejnej redukcji zmiennych zawartych z zbiorze danych.

Proces dodawania zmiennych pochodnych oraz wstępnej eliminacji zmiennych podsumowuje Tabela 16. Skrajne wiersze zawierają liczbę zmiennych odpowiednio na początku i na końcu procesu. Wartości w wewnętrznych wierszach oznaczają liczbę zmiennych dodanych lub usuniętych przez daną operację czyszczenia danych.

Tabela 16 Podsumowanie procesu eliminacji oraz dodawania zmiennych pochodnych

Rodzaj operacji	Liczba zmiennych	Zmiana
Zbiór bazowy	173	–
Eliminacja zmiennych rzadkich	169	-4
Zmienne pochodne dodane ekspercko	170	+1
Zmienne pochodne na podstawie reguł	190	+20
Filtry ukierunkowane	161	-29
Filtry nieukierunkowane	151	-10
Zbiór końcowy	**151**	

Źródło: Opracowanie własne.

W zbiorze końcowym 3 zmienne pełniły rolę identyfikatorów, 1 rolę zmiennej zależnej, a pozostałe 147 zmiennych rolę predyktorów.

5.2.7. Budowa modeli za pomocą wybranych narzędzi analitycznych

Dane, które przeszły wstępną selekcję zmiennych zostały poddane transformacjom mogącym potencjalnie ułatwić proces budowy modeli o pożądanych właściwościach. Pierwszym przekształceniem, jakie zastosowano dla pierwotnego zbioru danych była standaryzacja logistyczna Pyle'a opisywana w rozdziale 2. Ten rodzaj standaryzacji

zastosowano dla predyktorów ilościowych, redukując tym samym wartości odstające w analizowanym zbiorze. W analizie przyjęto granicę na poziomie 6 odchyleń standardowych. Powyżej tej granicy wartości predyktorów zostały zredukowane do wartości bliskich 0 lub 1, w zależności od kierunku odchylenia.

Drugim rodzajem zastosowanej standaryzacji była standaryzacja WoE również opisana w rozdziale 2. Wykonano ją dla wszystkich predyktorów; zarówno jakościowych jak i dla poddanych uprzedniej dyskretyzacji predyktorów ilościowych. Dyskretyzacja zmiennych ilościowych została przeprowadzona dwoma sposobami: za pomocą podziału na decyle oraz za pomocą podziału algorytmem CART, przy założeniu, że maksymalna głębokość drzewa wynosi 5. Analizie poddane zostały zatem cztery zbiory danych (w nawiasach podano skróty nazw metod wykorzystywane w dalszym opisie tabeli):

- oczyszczone zmienne surowe (Brak),
- zmienne przekształcone za pomocą standaryzacji Pyle'a (Pyle),
- zmienne WoE na podstawie decyli (WoE-Decyle),
- zmienne WoE na podstawie algorytmu CART (WoE-CART).

Analizowane zbiory danych mogły różnić się zatem ze względu na fakt zastosowania w ich przypadku trzech rodzajów modyfikacji (w nawiasie podano liczbę wariantów):

- standaryzacja zmiennych (4),
- dodanie zmiennych pochodnych (2),
- segmentacja zbioru danych (2).

Daje to 16 wariantów zbiorów danych możliwych do zbadania. W pracy badawczej porównaniu zostaną poddane powyższe kombinacje dla wybranych metod analitycznych. Zestaw wszystkich wariantów zbiorów użytych w analizie przedstawia Tabela 17.

Tabela 17 Warianty zbiorów danych użytych w analizie

Numer zbioru danych	Standaryzacja	Zmienne pochodne	Segmentacja
1.	Brak	Nie	Nie
2.	Pyle	Nie	Nie
3.	WoE – Decyle	Nie	Nie
4.	WoE – CART	Nie	Nie
5.	Brak	Tak	Nie
6.	Pyle	Tak	Nie
7.	WoE – Decyle	Tak	Nie
8.	WoE – CART	Tak	Nie
9.	Brak	Nie	Tak
10.	Pyle	Nie	Tak
11.	WoE – Decyle	Nie	Tak
12.	WoE – CART	Nie	Tak
13.	Brak	Tak	Tak
14.	Pyle	Tak	Tak
15.	WoE – Decyle	Tak	Tak
16.	WoE – CART	Tak	Tak

Źródło: Opracowanie własne.

W pracy zostały uwzględnione cztery metody analizy danych. Podczas ich wyboru kierowano się w pierwszej kolejności powszechnością ich wykorzystania w praktycznych zastosowaniach. Drugim kryterium było zapewnienie odpowiedniej reprezentacji zarówno dla metod „białoskrzynkowych", jak i tych działających na zasadzie czarnej skrzynki. Ostatnim czynnikiem wziętym pod uwagę podczas wyboru metod było wykorzystanie przynajmniej jednej metody ekonometrycznej. Na podstawie powyższych kryteriów wybrano cztery narzędzia analityczne:

- regresję logistyczną, będącą reprezentantem metod ekonometrycznych (oraz „białoskrzynkowych"), powszechnie wykorzystywaną w modelowaniu zagadnień biznesowych, w tym modelowaniu retencji klientów,
- drzewa klasyfikacyjne CART będące jedną z najpopularniejszych metod opartych na rekurencyjnym podziale przestrzeni zmiennych, pozwalającą na łatwą interpretację zbudowanych modeli,
- sieci neuronowe, przeżywające obecnie swój kolejny renesans, w badaniach wykorzystane zostaną jednokierunkowe sieci oparte na perceptronie wielowarstwowym,

– drzewa wzmacniane oparte na bibliotece *XGBoost* wykorzystywanej powszechnie zarówno do rozwiązywania rzeczywistych problemów biznesowych, jak również pozwalającej na zwycięstwa w szeregu konkursach *machine learning* [Chollet, 2019].

Podczas budowy modeli dla każdej z metod analitycznych wprowadzone zostaną dodatkowe, często specyficzne techniki związane z transformacją zmiennych, optymalizacją hiperparametrów, selekcją zmiennych oraz agregacją modeli. Wybór najlepszych modeli zostanie oparty o kryterium pola powierzchni pod krzywą ROC (AUC). W przypadku regresji logistycznej kryterium wyboru modelu zostało uzupełnione o dodatkowe kryteria przyjęte w praktyce biznesowej.

5.3. Model regresji logistycznej

Pierwszą z metod analizy danych wykorzystaną do budowy modelu prognozującego skłonność klientów do rezygnacji z usługi była regresja logistyczna. Metoda ta posiada relatywnie niewielką liczbę hiperparametrów. Dla przykładu implementacja metody dostępna w programie TIBCO Statistica oferuje jedynie trzy opcje: deltę wymiatania, liczbę iteracji oraz zbieżność. Hiperparametry te odnoszą się do dosyć złożonych aspektów obliczeniowych i nie są w praktyce obiektem optymalizacji. Główny wysiłek badaczy koncentruje się na kwestii doboru zmiennych do modelu, odpowiedniej transformacji predyktorów oraz ewentualnego uwzględnienia w procesie modelowania zmiennej (lub zmiennych) segmentującej.

W przypadku regresji logistycznej siła predykcyjna nie jest jedynym czynnikiem branym pod uwagę w procesie poszukiwania optymalnego modelu. Dodatkowymi kryteriami, jakie brane są pod uwagę są:

– liczba zmiennych w finalnym modelu (zazwyczaj liczba ta waha się w granicach od 8 do 15),
– zgodność znaków ocen parametrów regresji z ich biznesową interpretacją,
– niski poziom współliniowości predyktorów.

Podczas procesu identyfikacji optymalnego modelu regresji logistycznej testowanych jest zazwyczaj równolegle kilka metod selekcji zmiennych z różnymi ustawieniami ich hiperparametrów. W wyniku obliczeń tworzony jest zbiór kilkunastu lub kilkudziesięciu

modeli o pożądanej liczbie zmiennych. Z grupy modeli eliminuje się następnie te, dla których stwierdza się niezgodność znaków ocen paramentów regresji z ich biznesową interpretacją.

Spośród pozostałych modeli wybiera się takie, które cechują się niskim poziomem współliniowości predyktorów. W pracy ocenę współliniowości wykonano przy użyciu czynnika inflacji wariancji (*VIF, Variance Inflaction Factor*). Wartość VIF oblicza się dla każdego predyktora w modelu. Przyjęto próg akceptacji dla modeli, w których maksymalny poziom VIF nie przekraczałby poziomu 2,5[131]. Spośród modeli spełniających powyższe kryteria wybiera się model o najwyższej sile predykcyjnej ocenianej na podstawie pola powierzchni pod krzywą ROC oraz niekiedy dodatkowo na podstawie innych miar na przykład statystyki KS. Proces wyboru modelu nierzadko jest uzupełniany biznesową oceną badacza, który spośród kilku porównywalnych modeli może preferować taki, który uwzględniałby określone, preferowane przez niego zmienne.

Podczas budowy modelu skorzystano z dwustopniowej procedury selekcji predyktorów. Wstępny wybór predyktorów wykonano za pomocą metod:

- LASSO[132] dobierając eksperymentalnie parametr lambda, tak aby finalna liczba zmiennych była nie większa niż 40,
- regresji krokowej postępującej,
- regresji krokowej wstecznej.

Uzyskane w wyniku analizy zestawy zmiennych poddano finalnej optymalizacji za pomocą metody *branch and bound (B&B)*[133], zakładając, że finalny model zawierał będzie od 10 do 15 predyktorów.

Powyższa strategia identyfikacji optymalnego modelu zrealizowana została w dla wszystkich 16 wariantów zbioru danych. Kolejne, szczegółowe kroki analizy przedstawiono dla zbioru danych w wariancie numer 7 – zbioru przekodowanego za pomocą WoE na podstawie decyli, zawierającego zmienne pochodne i bez podziału na segmenty. Wybór zbioru został podyktowany rosnącą popularnością tego sposobu przygotowania danych, która w wielu branżach[134] jest ogólnie przyjętym standardem.

[131] Za [Stanisz, 2016].

[132] Wykorzystano implementację algorytmu dostępną w pakiecie *glmnet* programu R obudowaną własnym kodem umożliwiającym generowanie losowych wartości lambda.

[133] Implementacja oparta na pakiecie *leaps* programu R.

[134] Na przykład w obszarze ryzyka kredytowego.

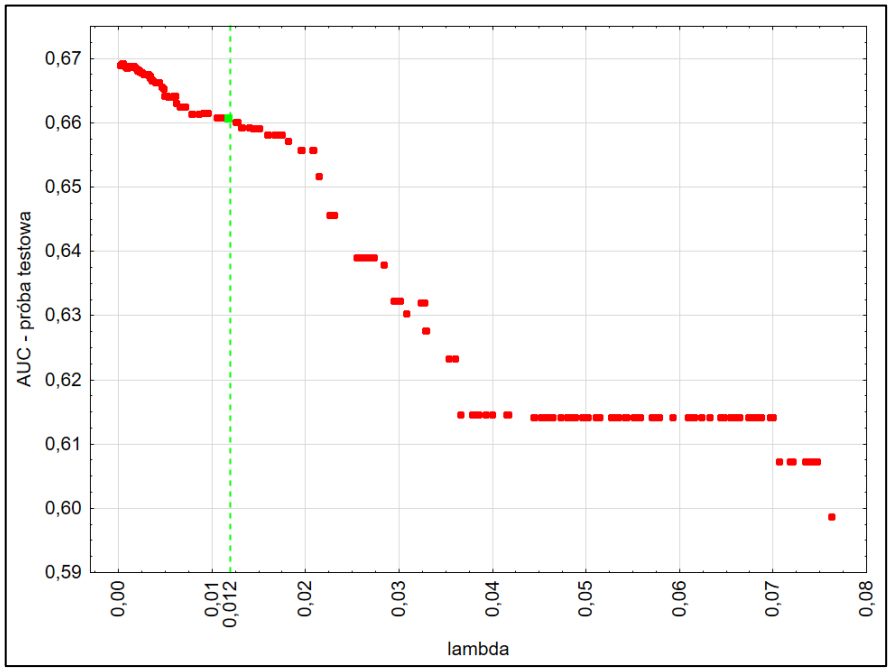
W pierwszym kroku zbudowano model na podstawie danych ze zbioru zawierającego komplet 147 predyktorów. Uzyskane wyniki przedstawia Tabela 18.

Tabela 18 Wyniki wstępne dla zbioru WoE-Decyle

Model	Liczba zmiennych	AUC (próba ucząca)	AUC (próba testowa)
Pełny	147	0,6700	0,6687

Źródło: Opracowanie własne.

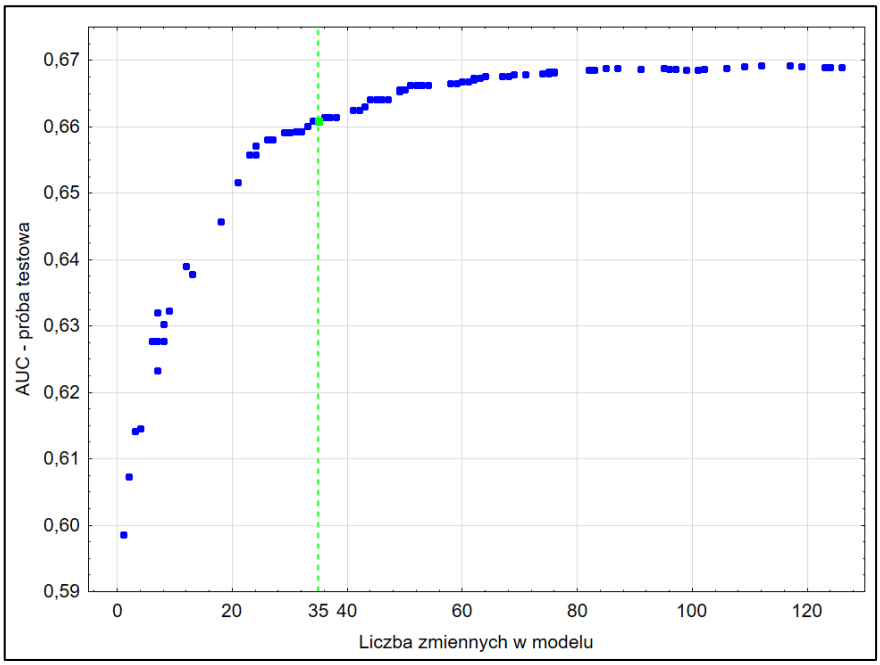
W kolejnym kroku liczba zmiennych została wstępnie ograniczona za pomocą wymienionych powyżej metod selekcji zmiennych. W przypadku metody LASSO eksperymentalnie dobierano wartość hiperparametru lambda. W tym celu wygenerowano losowo 300 wartości z zakresu [0,00001;0,1]. Dla każdej wartości lambda zbudowano osobny model. Uzyskane wyniki pozwoliły wyłonić wartość lambda na poziomie 0,012 (Rysunek 60).



Rysunek 60 Zależność pomiędzy poziomem hiperparametru lambda a siłą predykcyjną modelu

Źródło: Opracowanie własne.

Dla wybranej wartości lambda uzyskany model miał 35 predyktorów (Rysunek 61).



Rysunek 61 Zależność pomiędzy liczbą zmiennych w modelu a siłą predykcyjną modelu
Źródło: Opracowanie własne.

Modele zbudowane metodą krokową postępująca oraz metodą krokową wsteczną miały odpowiednio 65 oraz 69 zmiennych. Podsumowanie tego etapu budowy modelu przedstawia Tabela 19.

Tabela 19 Wyniki pośrednie dla wykorzystanych metod selekcji zmiennych

Model	Liczba zmiennych	AUC (próba ucząca)	AUC (próba testowa)
Lasso	35	0,6598	0,6609
Krokowa postępująca	65	0,6687	0,6689
Krokowa wsteczna	69	0,6689	0,6692

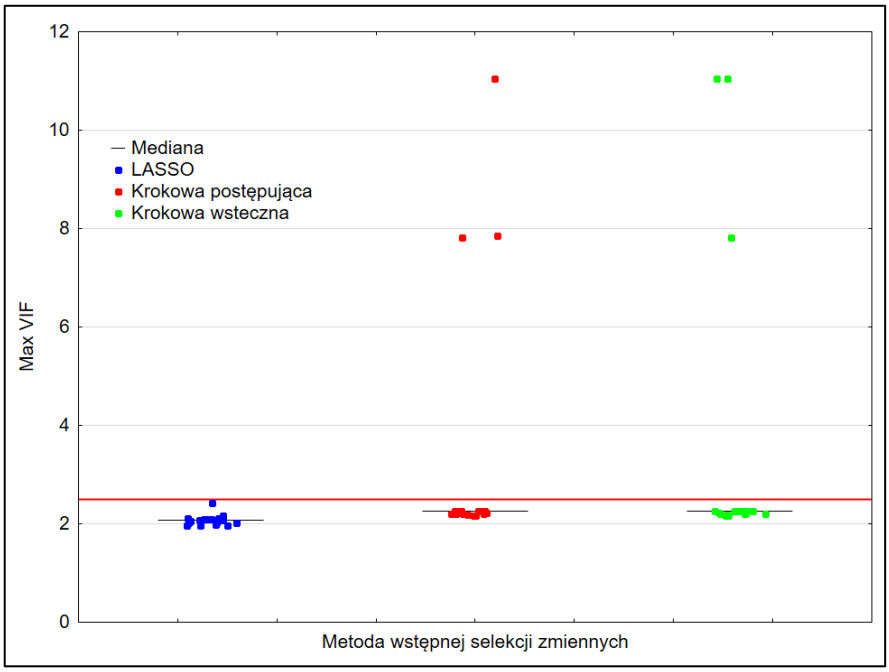
Źródło: Opracowanie własne.

Zmienne wyodrębnione na etapie preselekcji stanowiły zbiór wejściowy dla procesu identyfikacji modeli zawierających docelową liczbę zmiennych (od 10 do 15). Selekcja ta została przeprowadzona za pomocą algorytmu *B&B*. Dla każdego z trzech zestawów zmiennych odnaleziono po trzy modele zawierające dziesięć predyktorów oraz po trzy z odpowiednio jedenastoma, dwunastoma, trzynastoma, czternastoma i piętnastoma zmiennymi. Otrzymano dzięki temu łącznie 54 modeli, spośród których wybrano jeden najlepszy. Proces selekcji na tym etapie był dwustopniowy. W pierwszej kolejności sprawdzano poprawność specyfikacji modelu:

– zgodność znaków ocen parametrów regresji z ocenami uzyskanymi w modelach jednoczynnikowych,

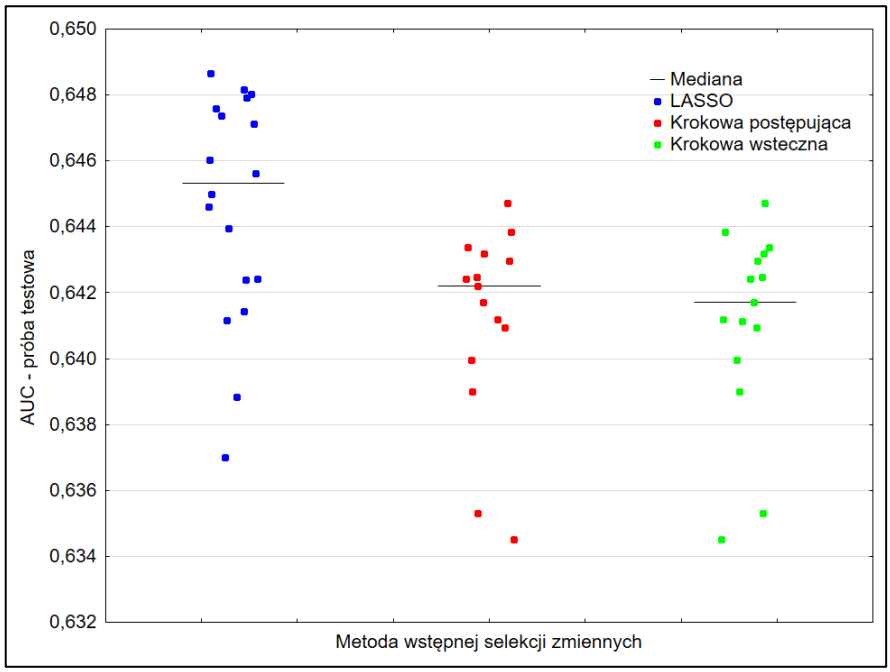
– poziom współliniowości zmiennych nie przekraczający VIF=2,5.

Na tej podstawie wyeliminowano 6 modeli, które nie spełniły kryterium współliniowości, po trzy modele wyłonione spośród zmiennych wskazanych przez metodę krokową postępującą oraz metodę krokową wsteczną (Rysunek 62).



Rysunek 62 Ocena współliniowości zmiennych po algorytmie B&B
Źródło: opracowanie własne.

Pozostałe modele zostały poddane ocenie pod kątem siły predykcyjnej na podstawie statystyki AUC na próbie testowej. Widocznie wyższe wartości siły predykcyjnej odnotowano dla modeli wyłonionych spośród zmiennych wskazanych przez metodę LASSO (Rysunek 63).



Rysunek 63 Ocena siły predykcyjnej modeli o dostatecznym poziomie współliniowości po algorytmie B&B

Źródło: opracowanie własne.

Zwycięski model został zbudowany na podstawie zmiennych wyłonionych metodą LASSO. Szczegółowe wyniki prezentuje Tabela 20.

Tabela 20 Wyniki finalnego modelu

Metoda wstępnej selekcji	Liczba zmiennych	VIF	AUC (próba ucząca)	AUC (próba testowa)
LASSO	15	2,11	0,6470	0,6498

Źródło: Opracowanie własne.

Zaletą kodowania WoE jest możliwość transformacji modelu do karty skoringowej, co następnie umożliwia jego wygodną interpretację biznesową. Fragment zbudowanego modelu w postaci karty skoringowej przedstawia Tabela 21.

Tabela 21 Fragment modelu w postaci karty skoringowej

Zmienna	Zakres	WoE	Ocena	s. Walda	p value	Skoring	Skoring zaokrąglony
age2	(-inf;0>	-3,57	0,0094	56,76691	0	36,5	37
age2	(0;30>	-7,623	0,0094	56,76691	0	35,4	35
age2	(30;38>	1,471	0,0094	56,76691	0	37,9	38
age2	(38;46>	4,362	0,0094	56,76691	0	38,7	39
age2	(46;54>	5,544	0,0094	56,76691	0	39	39
age2	(54;inf)	17,143	0,0094	56,76691	0	42,1	42
age2	Wartość neutralna	-	-			37,5	38
asl_flag	N	-5,683	0,0073	155,58088	0	36,3	36
asl_flag	Y	35,374	0,0073	155,58088	0	44,9	45
asl_flag	Wartość neutralna	-	-			37,5	38
attempt_Range	(-inf;2>	-27,754	0,00223	4,63655	0,0313	35,7	36
attempt_Range	(2;11>	5,679	0,00223	4,63655	0,0313	37,9	38
attempt_Range	(11;21>	5,812	0,00223	4,63655	0,0313	37,9	38
attempt_Range	(21;31>	9,255	0,00223	4,63655	0,0313	38,1	38
attempt_Range	(31;44>	6,238	0,00223	4,63655	0,0313	37,9	38
attempt_Range	(44;59>	5,432	0,00223	4,63655	0,0313	37,8	38
attempt_Range	(59;79>	1,079	0,00223	4,63655	0,0313	37,6	38
attempt_Range	(79;112>	-2,678	0,00223	4,63655	0,0313	37,3	37
attempt_Range	(112;174>	1,316	0,00223	4,63655	0,0313	37,6	38
attempt_Range	(174;inf)	-2,978	0,00223	4,63655	0,0313	37,3	37
attempt_Range	Wartość neutralna	-	-			37,1	37
children	Y	-1,528	0,00645	12,85866	0,00034	37,2	37
children	Missing	-1,507	0,00645	12,85866	0,00034	37,2	37
children	N	13,719	0,00645	12,85866	0,00034	40,1	40
children	Wartość neutralna	-	-			37,5	37

Źródło: Opracowanie własne.

Należy zwrócić uwagę, że wszystkie zmienne zawarte w modelu zostały podzielone na kategorie. Dla każdej kategorii analizowanych zmiennych wyznaczona została punktacja, widoczna w tabeli w kolumnie *Skoring zaokrąglony*. Stosowanie modelu polega na zsumowaniu punktów odpowiadających kategoriom, do których należy dany klient.

Operacja ta pozwala obliczyć końcową punktację klienta, będącą liczbową oceną jego lojalności wyrażoną w postaci szansy pozostania klientem przedsiębiorstwa. Wartości punktacji są idealnie skorelowane z prawdopodobieństwem odejścia obliczonym na podstawie modelu regresji logistycznej. Karta skoringowa nie wnosi zatem nowych wartości w sensie predykcji, pozwala natomiast na lepsze zrozumienie modelu oraz efektywną komunikację pomiędzy analitykami a menadżerami ds. marketingu.

Analogiczną metodykę zastosowano dla pozostałych konfiguracji zmiennych ograniczając się podczas wstępnej selekcji jedynie do metody LASSO. Uzyskane wyniki przedstawia Tabela 22.

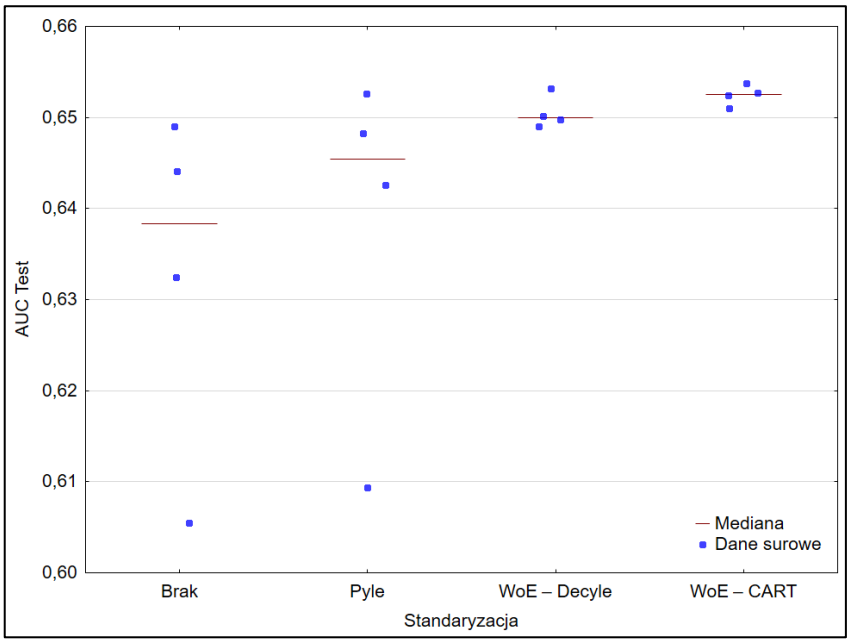
Tabela 22 Zbiorcze wyniki procesu identyfikacji modelu regresji logistycznej

Numer zbioru danych	Standaryzacja	Zmienne pochodne	Segmentacja	AUC (próba ucząca)	AUC (próba testowa)
1.	Brak	Nie	Nie	0,6079	0,6055
2.	Pyle	Nie	Nie	0,6138	0,6094
3.	WoE – Decyle	Nie	Nie	0,6464	0,6502
4.	WoE – CART	Nie	Nie	0,6580	0,6524
5.	Brak	Tak	Nie	0,6327	0,6325
6.	Pyle	Tak	Nie	0,6441	0,6426
7.	WoE – Decyle	Tak	Nie	0,6470	0,6498
8.	WoE – CART	Tak	Nie	0,6589	0,6538
9.	Brak	Nie	Tak	0,6463	0,6441
10.	Pyle	Nie	Tak	0,6460	0,6483
11.	WoE – Decyle	Nie	Tak	0,6505	0,6490
12.	WoE – CART	Nie	Tak	0,6486	0,6510
13.	Brak	Tak	Tak	0,6484	0,6490
14.	Pyle	Tak	Tak	0,6516	0,6526
15.	WoE – Decyle	Tak	Tak	0,6521	0,6532
16.	WoE – CART	Tak	Tak	0,6537	0,6527

Źródło: Opracowanie własne.

W procesie identyfikacji modeli dla wariantów zbiorów bez standaryzacji oraz ze standaryzacją Pyle zaobserwowano problemy ze współliniowością zmiennych modeli proponowanych przez metodę B&B. W takim przypadku konieczna okazała się dodatkowa eliminacja zmiennych i powtórzenie procesu selekcji.

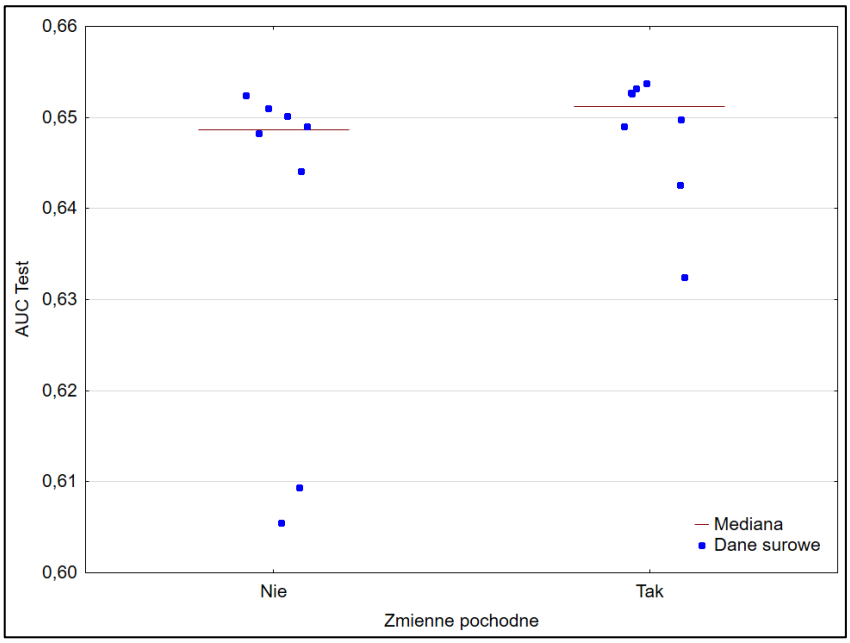
Na podstawie wyników zestawionych w Tabela 22 można stwierdzić, że najlepsze rezultaty uzyskano dla konfiguracji nr 15. Konfiguracja ta odnosi się do zbioru danych, w którym zmienne zostały podzielone na decyle, a następnie wystandaryzowane za pomocą transformacji WoE. Analizowany zbiór został uzupełniony o zmienne pochodne, a także podzielony na dwa segmenty, dla których zbudowano osobne modele.



Rysunek 64 Poziom AUC na próbie testowej w przekroju metod standaryzacji – model regresji logistycznej

Źródło: opracowanie własne.

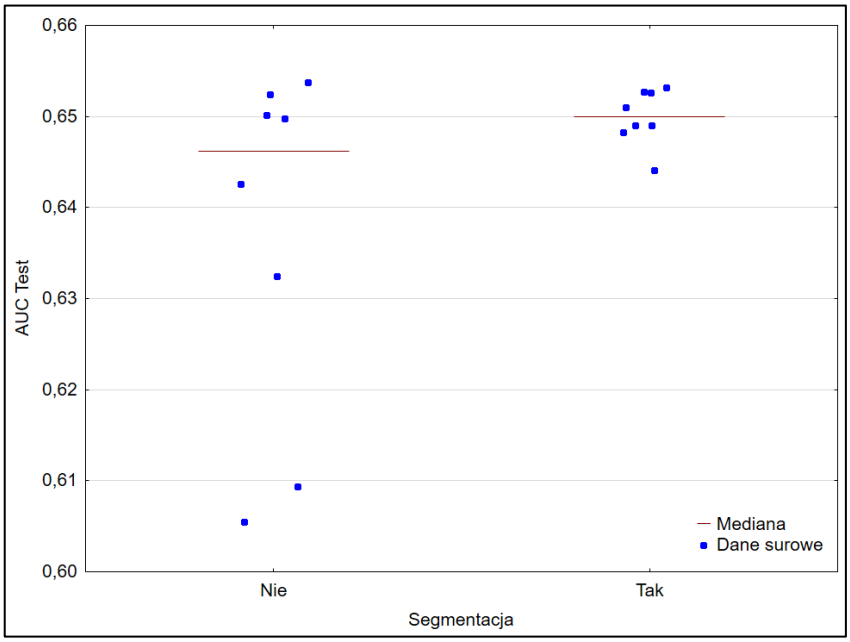
Na podstawie wyników widocznych na Rysunek 64 można stwierdzić, że standaryzacja zmiennych poprawia przeciętny poziom wartości AUC na próbie testowej. Najwyższe przeciętnie wyniki uzyskano na podstawie zbioru WoE-CART. Cechują się one również najmniejszą zmiennością.



Rysunek 65 Poziom AUC na próbie testowej w przekroju zmiennych pochodnych – model regresji logistycznej

Źródło: opracowanie własne.

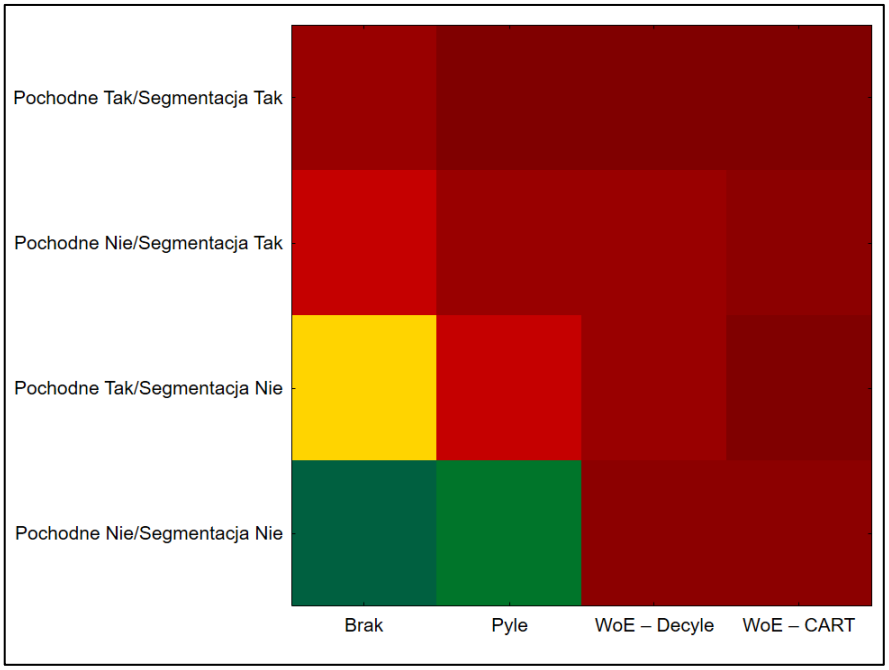
Na podstawie wyników przedstawionych na Rysunek 65 można stwierdzić, że uzupełnienie zbioru danych o zmienne pochodne spowodowało podniesienie się przeciętnego poziom AUC na próbie testowej.



Rysunek 66 Poziom AUC na próbie testowej w przekroju segmentacji – model regresji logistycznej

Źródło: opracowanie własne.

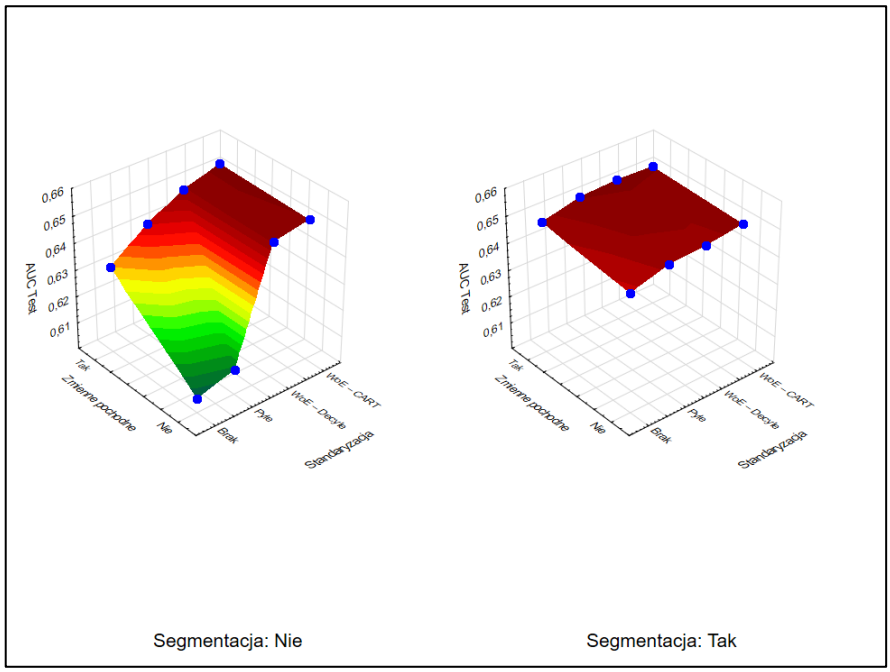
Na podstawie wyników zaprezentowanych na Rysunek 66 można z kolei stwierdzić, że wykonanie segmentacji zbioru danych spowodowało podniesienie się przeciętnego poziom AUC na próbie testowej.



Rysunek 67 Poziom AUC na zbiorze testowym w przekroju wszystkich zmiennych – model regresji logistycznej

Źródło: opracowanie własne.

Rysunek 67 przedstawia poziom AUC w przekroju trzech czynników uwzględnianych w badaniu. Wymiary segmentacja oraz zmienne pochodne zostały połączone (oś pionowa). Na analizowanym wykresie niższe wartości AUC oznaczone zostały kolorem żółtym i zielonym, a najwyższe kolorem brązowym.



Rysunek 68 Wykres warstwicowy w podziale na modele z segmentacją i bez – regresja logistyczna

Źródło: opracowanie własne.

Analiza Rysunek 68 pozwala na zaobserwowanie kluczowego wpływu segmentacji na uzyskane wyniki. Naniesione na wykresie powierzchnie mają charakter poglądowy i w pewnym stopniu ułatwiają interpretację występujących prawidłowości. Modele hybrydowe oparte na segmentacji zapewniają wyższą jakość predykcyjną w przekroju wszystkich wariantów analizy.

W celu ewentualnej poprawy uzyskanych wyników wykorzystano dodatkowo technikę agregacji modeli. Dla zwycięskiej piętnastej konfiguracji przygotowano model zbudowany na podstawie pięciu modeli o największym polu powierzchni pod krzywą ROC na zbiorze testowym. Operację tę wykonano dla obydwu segmentów. W każdym z nich predykcje pięciu najlepszych modeli zostały uśrednione. Wykonany zabieg nie spowodował poprawy jakości predykcji. Na zbiorze uczącym AUC wyniosło 0,6527, natomiast na zbiorze testowym wynik był taki sam jak dla modelu bez uśredniania i wyniósł 0,6532.

5.4. Model drzew klasyfikacyjnych i regresyjnych CART

Drugą po regresji logistycznej metodą wykorzystaną w analizie były drzewa klasyfikacyjne i regresyjne CART. Podobnie jak omawiana wcześniej regresja logistyczna, działają one na zasadzie „białej skrzynki”. Poza dobrocią dopasowania kolejnym kryterium

branym przez badaczy pod uwagę jest jakość uzyskanych reguł. Na jakość tę może składać się zarówno subiektywna ocena badacza jak też kryteria bardziej wymierne, jak liczba elementów poprzednika reguły oraz liczba przypadków wspierających regułę. W niniejszej pracy, ze względu na brak możliwości eksperckiej oceny jakości reguł, główny nacisk położony zostanie na wymiar predykcyjny.

Za punkt wyjścia dla rozważań przedstawionych w niniejszym podrozdziale przyjęto wynik modelu zbudowanego dla domyślnych ustawień hiperparametró[135]. Siła predykcyjna modelu wyrażona w postaci pola powierzchni pod krzywą ROC dla zbioru testowego wyniosła AUC=0,6075. W przypadku zbioru uczącego pole to wyniosło AUC=0,6029.

Tabela 23 Ustawienia domyślne algorytmu CART

Miara dopaso-wania	Koszt błędnej klasyfikacji dla kategorii "0" zmiennej Y	Koszt błędnej klasyfikacji dla kategorii "1" zmiennej Y	Prawdo-podo-bieństwo *a priori*	Głębo-kość drze-wa	Min. Liczeb-ność węzła macie-rzystego	Min. Liczeb-ność węzła poto-mnego	AUC (próba ucząca)	AUC (próba testowa)
GINI	1	1	0,5	10	2125	2125	0,6029	0,6075

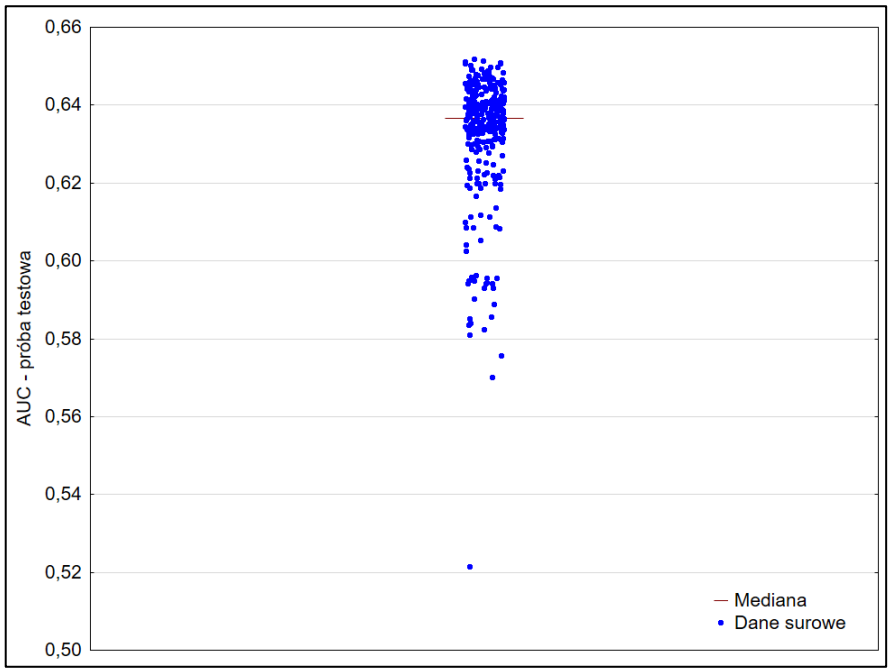
Źródło: Opracowanie własne.

W pierwszym etapie analizy zbiorem wejściowym był zbiór zawierający oczyszczone dane. W celu identyfikacji optymalnych hiperparametrów wygenerowano 300 losowych kombinacji kierując się następującymi regułami (w kolejnych punktach nazwa hiperparametru, w nawiasie zakresy lub listy rozpatrywanych wartości):

- koszty błędnych klasyfikacji (min 1 do 5, maks. 5 do 1),
- prawdopodobieństwo *a priori* (0;1),
- miara dopasowania (GINI, Chi-kwadrat, G-kwadrat),
- głębokość drzewa [5;25],
- minimalna liczność węzła podlegającego podziałowi [6;51],
- minimalna liczność węzła potomnego [6;51].

Dla każdego modelu obliczono miarę AUC, zarówno dla zbioru uczącego jak i testowego. Podczas wyboru najlepszego modelu brano pod uwagę wyniki na zbiorze testowym. Uzyskane wyniki AUC wahały się w granicach od 0,521 do 0,652. Rozkład uzyskanych wartości przedstawia Rysunek 69.

[135] Analizę przeprowadzono w programie Tibco Statistica w module Drzewa interakcyjne.



Rysunek 69 Rozkład wartości AUC na próbie testowej dla 300 losowych konfiguracji hiperparametrów

Źródło: opracowanie własne.

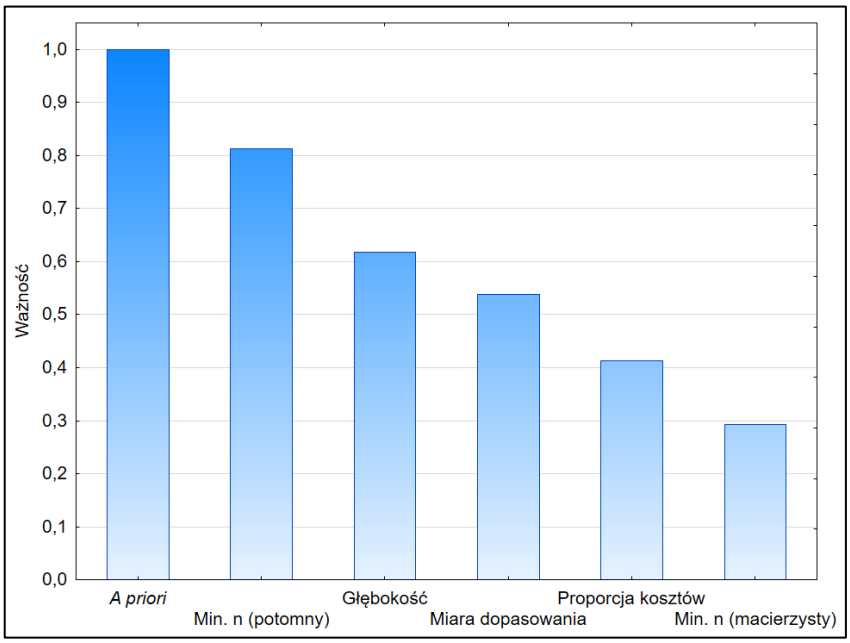
Arkusz zawierający wygenerowane wartości hiperparametrów wraz z odpowiadającymi im wartościami AUC poddano analizie za pomocą drzew regresyjnych CART. Rolę zmiennej zależnej przypisano polu powierzchni pod krzywą ROC dla próby testowej, zmienne zawierające wylosowane ustawienia pełniły rolę predyktorów (koszty błędnych klasyfikacji zastąpiono jedną zmienną wyrażającą ich iloraz). Fragment analizowanego zbioru danych przedstawia Tabela 24.

Tabela 24 Fragment zbioru z hiperparametrami oraz powiązanym z nimi wynikiem działania modelu CART

Miara dopasowania	Koszt błędnej klasyfikacji dla kategorii "0" zmiennej Y	Koszt błędnej klasyfikacji dla kategorii "1" zmiennej Y	Prawdopodobieństwo *a priori*	Głębokość drzewa	Minimalna liczebność węzła macierzystego	Minimalna liczebność węzła potomnego	Proporcja kosztów	AUC (próba ucząca)	AUC (próba testowa)
3	5	3	0,129	17	15	51	1,6667	0,665609	0,651879
1	2	2	0,832	9	24	25	1	0,66377	0,651409
1	5	5	0,825	10	15	24	1	0,659723	0,651143
3	4	1	0,132	25	14	45	4	0,661887	0,651082
3	5	1	0,336	20	26	32	5	0,667779	0,650849
3	3	3	0,043	18	28	28	1	0,662841	0,650661
3	4	1	0,359	11	20	30	4	0,670624	0,650253
1	2	3	0,773	10	43	20	0,6667	0,668185	0,649938
3	5	2	0,167	11	36	18	2,5	0,670212	0,649882
3	2	1	0,067	14	9	50	2	0,661622	0,649289
1	4	5	0,81	16	19	42	0,8	0,667676	0,649119
3	3	2	0,126	21	31	31	1,5	0,667777	0,649107
3	4	2	0,31	11	29	45	2	0,675333	0,648839
3	3	4	0,102	16	12	49	0,75	0,671329	0,648614
3	5	1	0,252	24	20	26	5	0,665264	0,648597
3	4	1	0,006	14	5	7	4	0,664058	0,648561
1	5	3	0,866	11	38	33	1,6667	0,668682	0,648453
2	4	4	0,812	12	30	30	1	0,67672	0,648309
2	1	4	0,551	24	41	42	0,25	0,670307	0,648263
1	2	4	0,865	9	50	33	0,5	0,645475	0,647868

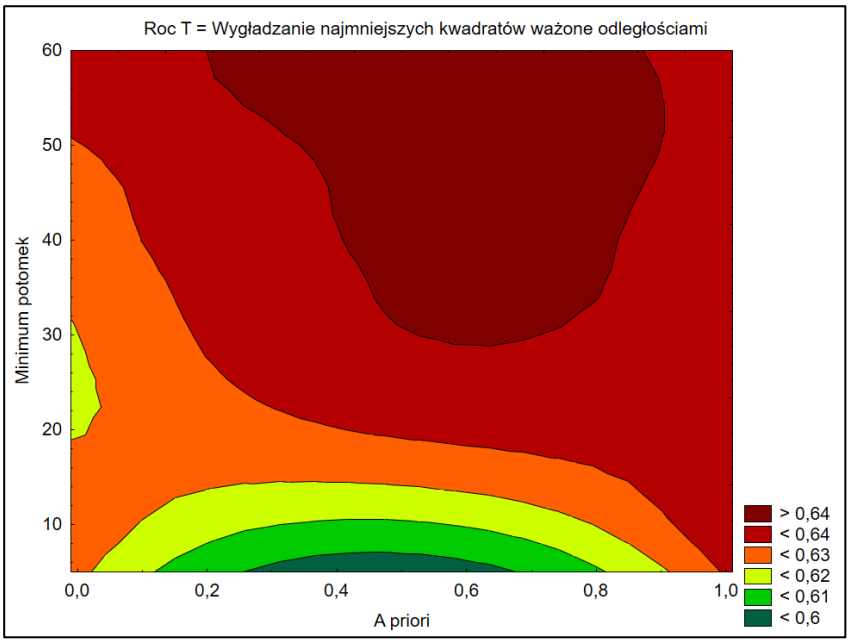
Źródło: opracowanie własne.

Na podstawie powyższego zbioru danych oraz podanej specyfikacji zmiennych zbudowano drzewo regresyjne o głębokości 5 poziomów. Model stał się podstawą do utworzenia rankingu ważności predyktorów przedstawionego na Rysunek 70. Można zauważyć, że największy wpływ na dobroć dopasowania modelu miało prawdopodobieństwo *a priori* (ponad połowa podziałów drzewa została utworzona na podstawie tej zmiennej), w drugiej kolejności minimalna liczność potomka.



Rysunek 70 Wykres ważności predyktorów dla drzewa regresyjnego CART
Źródło: opracowanie własne.

Wpływ najważniejszych hiperparametrów na dobroć dopasowania modelu przybliża Rysunek 71. Obszar najlepszego dopasowana został wyznaczony dla minimalnego potomka powyżej 29, przy równoczesnym prawdopodobieństwie *a priori* w granicach od około 0,2 do około 0,9.



Rysunek 71 Wykres warstwicowy zależności pomiędzy hiperparametrami a dopasowaniem modelu

Źródło: opracowanie własne.

Wnioski płynące z wykonanej analizy stanowiły podstawę do wylosowania kolejnych dwóch zestawów zawierających po 300 kompletów hiperparametrów tym razem ze skorygowanymi ustawieniami. Po korekcie zakres hiperparametrów dla drugiego zestawu kształtował się następująco:

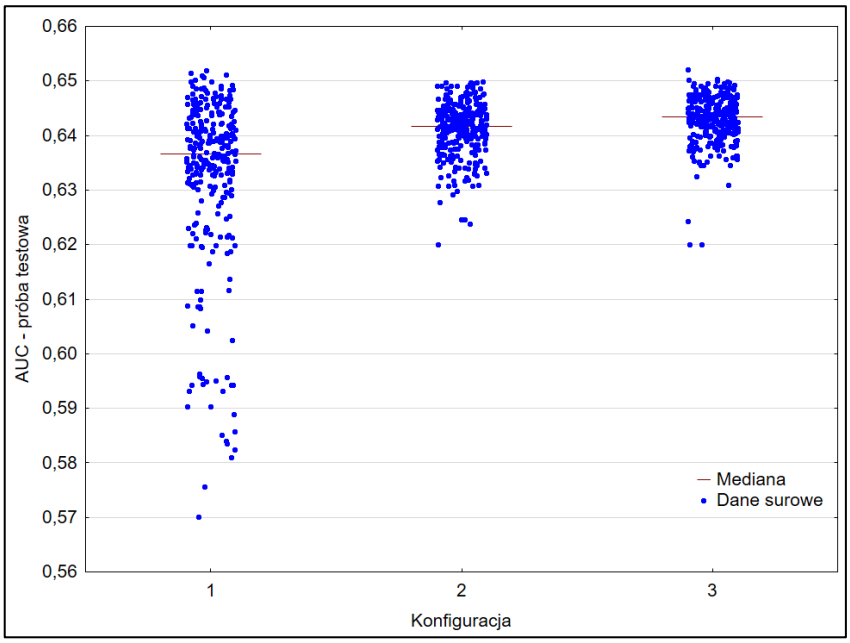
- Prawdopodobieństwo *a priori* [0,2;0,8],
- Minimalna liczność węzła potomnego [30;90].

W przypadku trzeciego zestawu poszerzono zakres wartości prawdopodobieństwa *a priori*. Nowe zakresy wartości hiperparametrów przyjmowały wartości:

- Prawdopodobieństwo *a priori* [0,1;0,9],
- Minimalna liczność węzła potomnego [30;90].

Pozostałe zakresy wartości hiperparametrów pozostały bez zmian. Oznacza to, że łącznie zbudowano 600 dodatkowych modeli. Rozkład uzyskanych wartości dla drugiego i trzeciego zestawu na tle pierwszej konfiguracji hiperparametrów prezentuje Rysunek 72[136].

[136] W celu zapewnienia większej przejrzystości wykresu pominięto skrajny wynik uzyskany dla modelu z pierwszej konfiguracji.



Rysunek 72 Rozkład wartości AUC na próbie testowej dla 3 zestawów konfiguracji hiperparametrów

Źródło: opracowanie własne

Na podstawie uzyskanych wyników (Tabela 25) można zaobserwować, że konfiguracja nr 3 skutkowała najwyższą przeciętną wartością AUC, najmniejszą zmiennością uzyskanych wyników, a także najlepszym modelem.

Tabela 25 Statystyki AUC dla trzech konfiguracji (dane surowe)

Konfiguracja	Średnia	Mediana	Minimum	Maksimum	Odch.std
1	0,6324	0,6366	0,5215	0,6519	0,0170
2	0,6412	0,6417	0,6200	0,6498	0,0049
3	0,6431	0,6434	0,6200	0,6521	0,0043

Źródło: Opracowanie własne.

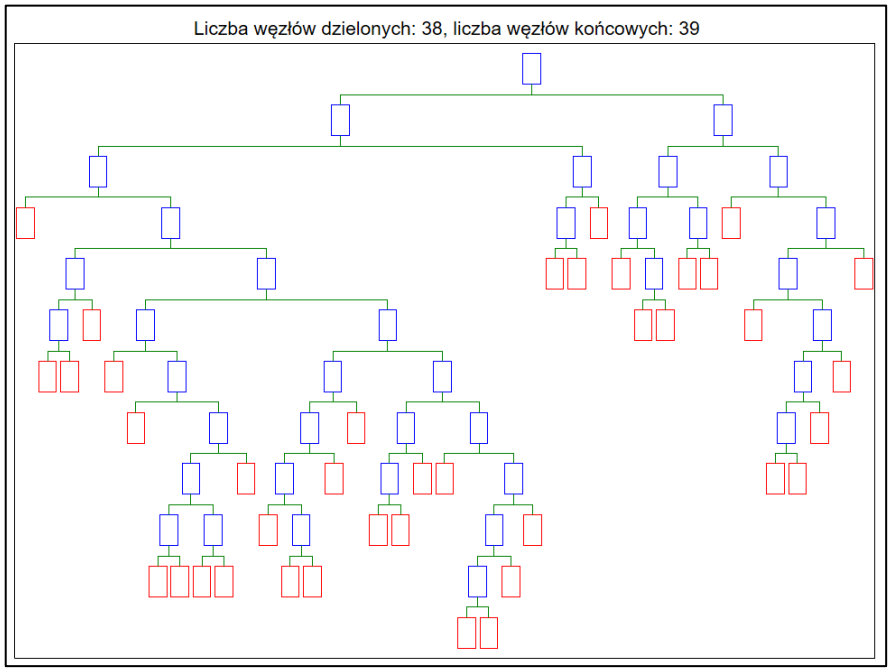
W grupie 10 najlepszych – z punktu widzenia AUC – modeli znalazło się jednak aż 7 modeli pochodzących z losowej konfiguracji, a jedynie 3 z konfiguracji trzeciej. Hiperparametry najlepszego modelu oraz wynik uzyskany na podstawie próby uczącej oraz próby testowej przedstawia Tabela 26.

Tabela 26 Hiperparametry oraz dobroć dopasowania najlepszego modelu

Miara dopaso-wania	Koszt błędnej klasyfikacji dla kategorii "0" zmiennej Y	Koszt błędnej klasyfikacji dla kategorii "1" zmiennej Y	Prawdo-podo-bieństwo *a priori*	Głębo-kość drze-wa	Min. Liczeb-ność węzła macie-rzystego	Min. Liczeb-ność węzła poto-mnego	AUC (próba ucząca)	AUC (próba testowa)
G-kwadrat	5	1	0,31	15	27	40	0,6816	0,6546

Źródło: Opracowanie własne.

Struktura najlepszego modelu drzewa klasyfikacyjnego została przedstawiona na Rysunek 73.



Rysunek 73 Układ drzewa o najlepszym dopasowaniu

Źródło: opracowanie własne.

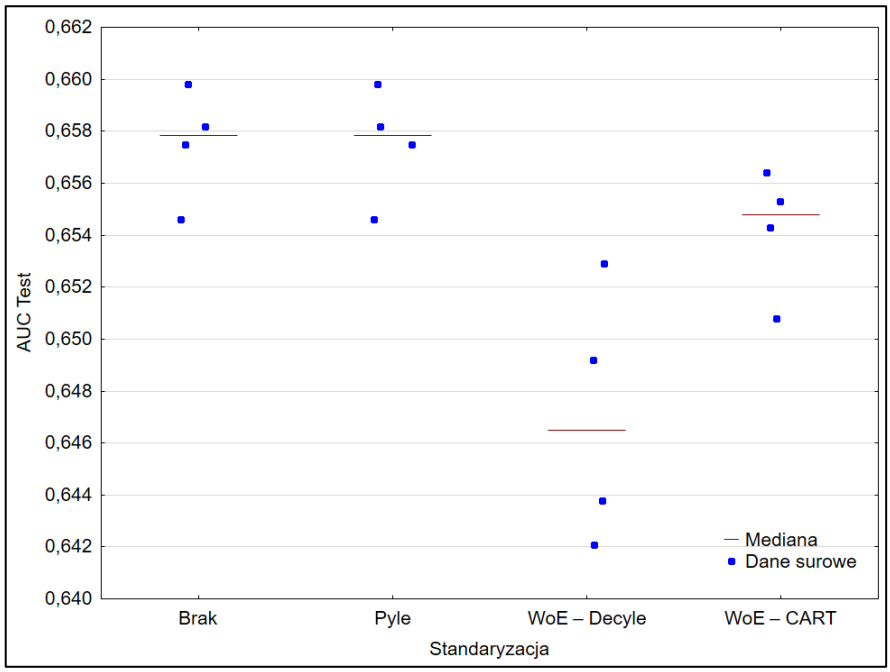
Zabiegi mające na celu przycięcie zbudowanego drzewa prowadziły do zmniejszenia się wartości AUC na próbie testowej. Uzyskaną konfigurację przyjęto zatem jako wersję ostateczną. Dla przygotowanych uprzednio 900 zestawów hiperparametrów zbudowano modele uwzględniając kolejne 15 zbiorów danych opisanych w części metodycznej pracy.

Tabela 27 Zbiorcze wyniki procesu identyfikacji modelu drzew klasyfikacyjnych CART

Numer zbioru danych	Standaryzacja	Zmienne pochodne	Segmentacja	AUC (próba ucząca)	AUC (próba testowa)
1.	Brak	Nie	Nie	0,6816	0,6546
2.	Pyle	Nie	Nie	0,6816	0,6546
3.	WoE – Decyle	Nie	Nie	0,6456	0,6421
4.	WoE – CART	Nie	Nie	0,6738	0,6508
5.	Brak	Tak	Nie	0,6733	0,6582
6.	Pyle	Tak	Nie	0,6733	0,6582
7.	WoE – Decyle	Tak	Nie	0,6755	0,6492
8.	WoE – CART	Tak	Nie	0,6733	0,6564
9.	Brak	Nie	Tak	0,6816	0,6575
10.	Pyle	Nie	Tak	0,6816	0,6575
11.	WoE – Decyle	Nie	Tak	0,6570	0,6438
12.	WoE – CART	Nie	Tak	0,6706	0,6553
13.	Brak	Tak	Tak	0,6825	0,6598
14.	Pyle	Tak	Tak	0,6825	0,6598
15.	WoE – Decyle	Tak	Tak	0,6806	0,6529
16.	WoE – CART	Tak	Tak	0,6706	0,6543

Źródło: opracowanie własne.

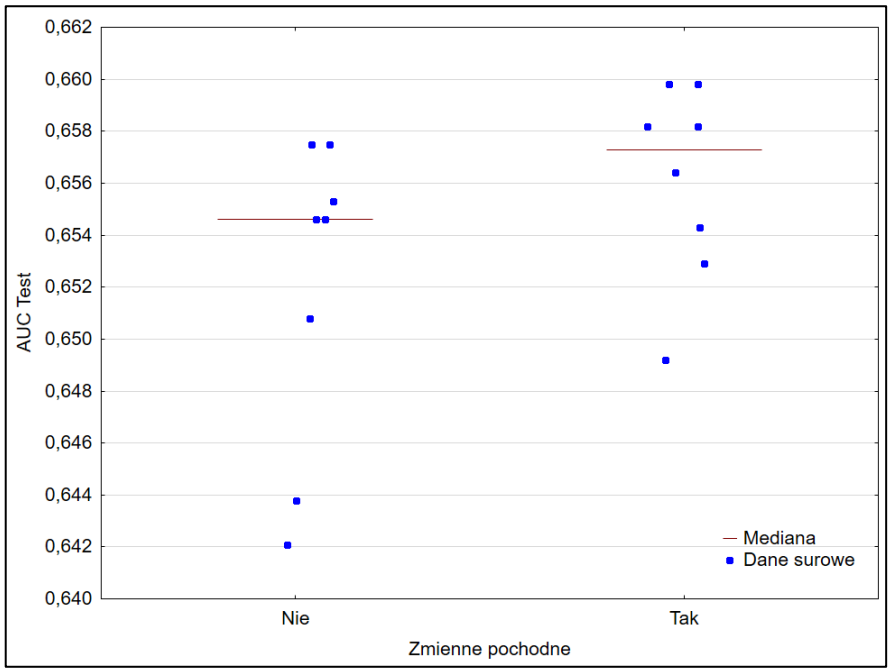
Analizując Rysunek 74 można stwierdzić, że wyniki modelowania na podstawie danych wystandaryzowanych za pomocą przekształcenia Pyle'a są identyczne jak te uzyskane z modeli budowanych na oczyszczonych danych surowych. Standaryzacja Pyle'a nie wprowadza zmian w porządku wartości predyktorów, zmieniając jedynie ich zakres i redukując wartości odstające.



Rysunek 74 Poziom AUC na próbie testowej w przekroju metod standaryzacji – model CART

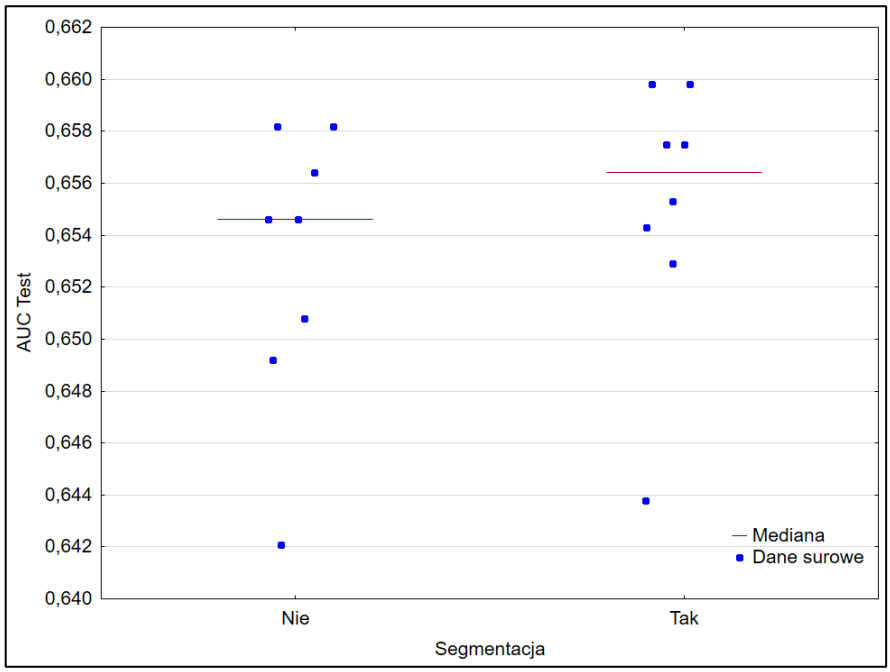
Źródło: opracowanie własne.

Modele budowane na zbiorze standaryzowanym za pomocą metody WoE cechują się mniejszym przeciętnym poziomem pola powierzchni pod krzywą ROC w porównaniu do modeli zbudowanych na podstawie pozostałych zbiorów danych.



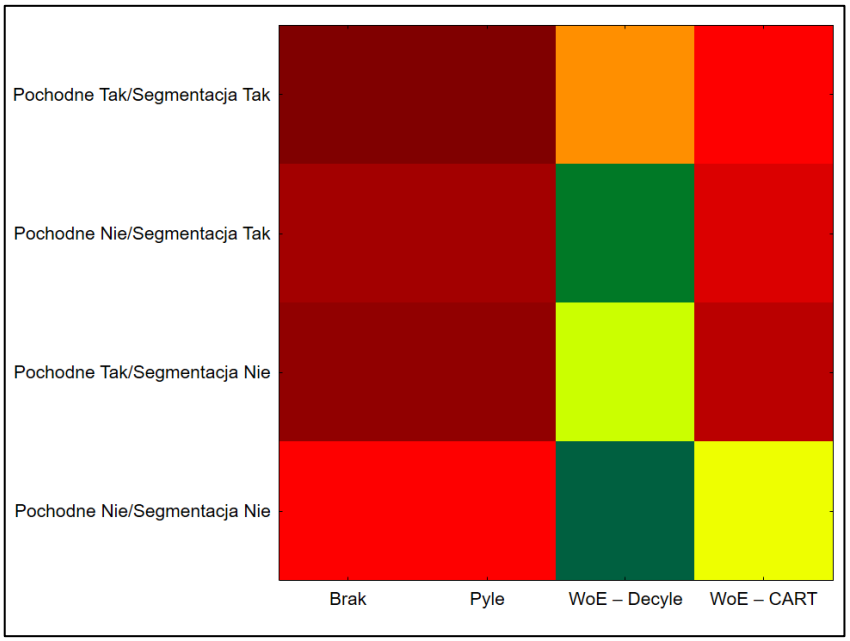
Rysunek 75 Poziom AUC na próbie testowej dla zmiennych pochodnych – model CART
Źródło: opracowanie własne.

Najlepsze modele uzyskane dla zbiorów danych zawierających zmienne pochodne charakteryzowały się zarówno mniejszą zmiennością jak i wyższą medianą. Na podstawie wyników widocznych na Rysunek 75 można stwierdzić, że uzupełnienie zbioru danych o zmienne pochodne spowodowało podniesienie się przeciętnego poziom AUC na próbie testowej.



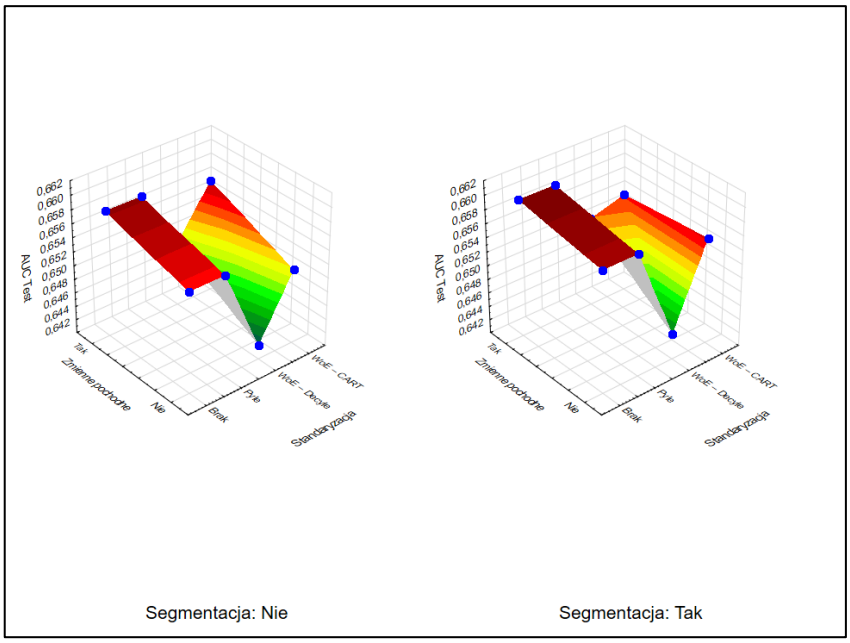
Rysunek 76 Poziom AUC na próbie testowej w przekroju segmentacji – model CART
Źródło: opracowanie własne.

Na podstawie Rysunek 76 można natomiast stwierdzić, że wykonanie segmentacji zbioru danych spowodowało podniesienie się przeciętnego poziom AUC na próbie testowej.



Rysunek 77 Poziom AUC na zbiorze testowym w przekroju wszystkich zmiennych – model CART
Źródło: opracowanie własne.

Analiza łącznego wpływu wszystkich zmiennych (Rysunek 77) pozwala stwierdzić, że najlepsze dopasowanie uzyskanych modeli jest powiązane z wykorzystaniem zbioru danych bez transformacji („Brak") lub ze standaryzacją Pyle'a („Pyle"), ze zmiennymi pochodnymi oraz segmentacją.



Rysunek 78 Wykres warstwicowy w podziale na modele z segmentacją i bez - drzewa CART
Źródło: Opracowanie własne.

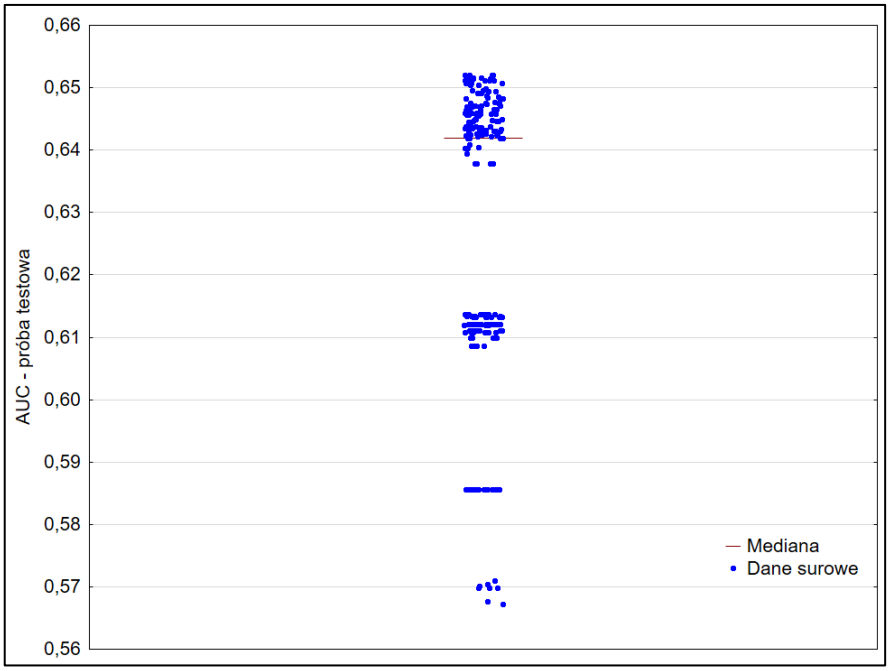
Dodatkowa analiza wyników wykonana za pomocą wykresów warstwicowych (Rysunek 78) potwierdza najlepsze dopasowanie modeli dla zbiorów bez transformacji lub ze standaryzacją Pyle'a. Naniesione na wykresie powierzchnie mają charakter poglądowy i w pewnym stopniu ułatwiają interpretację występujących prawidłowości. Po raz kolejny potwierdzono, że hybrydyzacja połączona z uzupełnieniem zbioru danych o zmienne pochodne poprawia jakość klasyfikacji modeli.

Jako uzupełnienie podstawowej ścieżki badawczej wykonano dodatkowo dwa eksperymenty polegające na:

– zbudowaniu modeli opartych na alternatywnych podziałach drzewa, brano pod uwagę modele wykorzystujące jeden z 10 najlepszych predyktorów w kontekście pierwszego podziału drzewa oraz modele biorące pod uwagę możliwe kombinacje pierwszego oraz drugiego poziomu podziału, w tym przypadku przyjęto kolejkę o długości 7 (kolejka oznacza pozycję w rankingu zmiennych konkurencyjnych algorytmu CART);

– agregacji pięciu najlepszych modeli.

W wyniku ingerencji w naturalny przebieg algorytmu zbudowano 10 modeli, dla których wymuszono pierwszy podział drzewa kolejną zmienną z listy najlepszych predyktorów. Po wykonaniu wymuszonego pierwszego podziału, kolejne przebiegały w sposób automatyczny (z uwzględnieniem współczynnika poprawy – *improvement* - algorytmu CART). Żaden ze zbudowanych w ten sposób modeli nie dostarczył wyniku lepszego od modelu bazowego. Kontynuacją tej procedury była ingerencja zarówno w pierwszy poziom podziału jak również w podziały na drugim poziomie. W tym przypadku rozpatrywano kolejkę 7 najlepszych predyktorów na pierwszym poziomie, a dla nich kombinacje 7 najlepszych podziałów prawego oraz lewego węzła (co dało łącznie ponad 300 modeli). Również dla takiego układu nie stwierdzono wyniku lepszego od bazowego. Duża liczba konkurencyjnych konfiguracji skutkowała uzyskaniem wyniku zbliżonego do wyniku bazowego. Na uwagę zasługuje również szereg znacząco słabszych wyników. Szczegóły przedstawiono na Rysunek 79.



Rysunek 79 AUC dla wymuszonych konfiguracji na pierwszym i drugim poziomie podziału
Źródło: Opracowanie własne.

Ostatnią zastosowaną modyfikacją procesu uczenia było uśrednienie wyników pięciu najlepszych modeli zbudowanych na danych surowych. Uzyskany wynik wyraźnie poprawił uzyskaną prognozę bazową. Przyjęte podejście poprawiło moc predykcyjną

modelu. Strategia ta uniemożliwia jednak jego interpretację. Bardziej zasadne może okazać się zastąpienie jej przez bardziej wyrafinowane metody oparte na podobnym schemacie, na przykład przez metodę losowego lasu. Szczegóły bazowych konfiguracji (modele od M1 do M5) oraz modelu zagregowanego przedstawia Tabela 28**Błąd! Nie można odnaleźć źródła odwołania.**.

Tabela 28 Wyniki dla najlepszych konfiguracji drzew oraz dla modelu zagregowanego

Miara dopaso-wania	Koszt błędnej klasyfikacji dla kategorii "0" zmiennej Y	Koszt błędnej klasyfikacji dla kategorii "1" zmiennej Y	Prawdo-podo-bieństwo *a priori*	Głębo-kość drze-wa	Min. Liczeb-ność węzła macie-rzystego	Min. Liczeb-ność węzła poto-mnego	AUC (próba ucząca)	AUC (próba testowa)
G-kwadrat	5	1	0,31	15	27	40	0,6816	0,6546
G-kwadrat	5	3	0,129	17	15	51	0,6656	0,6519
GINI	2	2	0,832	9	24	25	0,6638	0,6514
GINI	5	5	0,825	10	15	24	0,6597	0,6511
G-kwadrat	4	1	0,132	25	14	45	0,6619	0,6511
Zagregowany							**0,6824**	**0,6618**

Źródło: Opracowanie własne.

5.5. Model drzew wzmacnianych

Kolejnym zastosowanym narzędziem *data mining* były drzewa wzmacniane [137]. W pierwszym kroku zbudowano model na podstawie zbioru reprezentującego pierwszą konfigurację, bazując na domyślnych wartościach hiperparametrów.

Tabela 29 Wynik modelowania bez optymalizacji hiperparametrów

Numer zbioru danych	Standaryzacja	Zmienne pochodne	Segmentacja	AUC (próba ucząca)	AUC (próba testowa)
1.	Brak	Nie	Nie	0,7760	0,6979

Źródło: opracowanie własne.

Dla każdego z wyodrębnionych zbiorów zbudowano 1000 modeli losując uprzednio zestaw hiperparametrów dla każdego z nich. Pod uwagę wzięto następujące ustawienia metody oraz zakresy ich wartości.

[137] Obliczenia zostały wykonane za pomocą biblioteki *xgboost* dostępnej między innymi za pomocą języka Python.

- *Max_depth* – Maksymalna głębokość drzewa. Zwiększenie tej wartości zwiększa złożoność modelu. Przyjęto wartości całkowite z zakresu [1;7].
- *Learning_rate (eta)* – koryguje prędkość uczenia modelu w celu redukcji nadmiernego dopasowania. Zmniejszenie tej wartości spowalnia proces uczenia. Przyjęto wartości z zakresu (0;0,2).
- *Gamma* – Minimalna redukcja błędu wymagana do wykonania kolejnego podziału węzła drzewa. Zwiększenie wartości redukuje liczbę akceptowalnych podziałów. Przyjęto wartości z zakresu (0;10).
- *Scale_pos_weight* – wyrównuje proporcje klas zmiennej zależnej zgodnie z przyjętą wartością. Przyjęto wartości z zakresu (0;6).
- *Lambda* – Składnik regularyzacji wag L2. Zwiększenie tej wartości spowalnia proces nauki. Przyjęto wartości z zakresu (0;8).
- *Alpha* – Składnik regularyzacji wag L1. Zwiększenie tej wartości spowalnia proces nauki. Przyjęto wartości z zakresu (0;8).
- *Subsample* – określa, jaka część zbioru uczącego zostanie wylosowana do realizacji każdej iteracji uczenia. Przyjęto wartości z zakresu (0;1).
- *Colsample_bytree* – określa, jaka część predyktorów zostanie wylosowana podczas budowy każdego z drzew składowych. Przyjęto wartości z zakresu (0;1).
- *Colsample_bylevel* – określa, jaka część predyktorów zostanie wylosowana podczas budowy każdego z poziomów drzew składowych. Losowanie przebiega na podstawie zmiennych dostępnych dla danego drzewa. Przyjęto wartości z zakresu (0;1).
- *Colsample_bynode* – określa, jaka część predyktorów zostanie wylosowana podczas wykonywania każdego z podziałów. Losowanie przebiega na podstawie zmiennych dostępnych dla danego poziomu drzewa. Przyjęto wartości z zakresu (0;1).
- *Booster* – określa rodzaj algorytmu bazowego dla procesu wzmacniania. Losowano jedną z poniższych opcji:
    - *Gbtree* – algorytm drzew klasyfikacyjnych,
    - *Linear* – model liniowy,
    - *Dart* – algorytm drzew z modyfikacją *dropout.*

W wyniku analizy dla każdego z 16 wariantów zbioru uczącego uzyskano zbiór metadanych zawierający 1000 wierszy. Każdy z wierszy reprezentował jeden model opisany przez jego hiperparametry oraz uzyskane wyniki. Przeprowadzono pogłębioną analizę uzyskanych wyników dla pierwszego z analizowanych zbiorów.

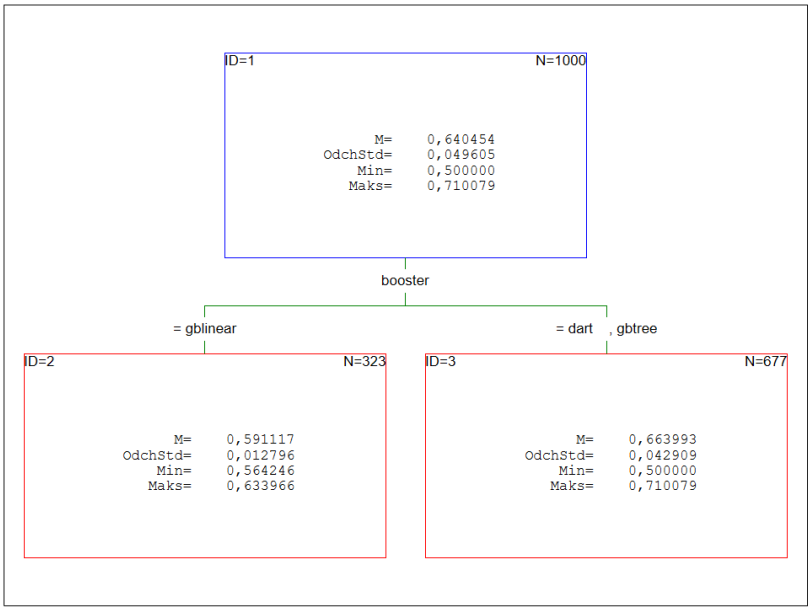
Pierwszą metodą, jaką wykorzystano do analizy uzyskanego zbioru były drzewa regresyjne CART. Za ich pomocą dokonano wstępnej oceny wpływu hiperparametrów na siłę dyskryminacyjną zbudowanego modelu. Ranking predyktorów wykonany dla węzła macierzystego drzewa w sposób jednoznaczny wskazuje na wysoki wpływ wyboru algorytmu bazowego (*booster*) na uzyskany wynik. Statystyka poprawy określa wartość spadku średniego błędu kwadratowego (pomnożonego przez 100) po wykonaniu podziału na postawie danej zmiennej. Kolejne w rankingu hiperparametry mają już zdecydowanie mniejsze znaczenie.

Tabela 30 Ranking predyktorów na podstawie statystyk podziału węzła macierzystego drzewa CART

Hiperparametr	Poprawa
booster	1,1613
colsample_bytree	0,1546
max_depth	0,0968
learning_rate	0,0837
subsample	0,0787
colsample_bynode	0,0640
colsample_bylevel	0,0549
lambda	0,0202
alpha	0,0189
gamma	0,0088
scale_pos_weight	0,0048

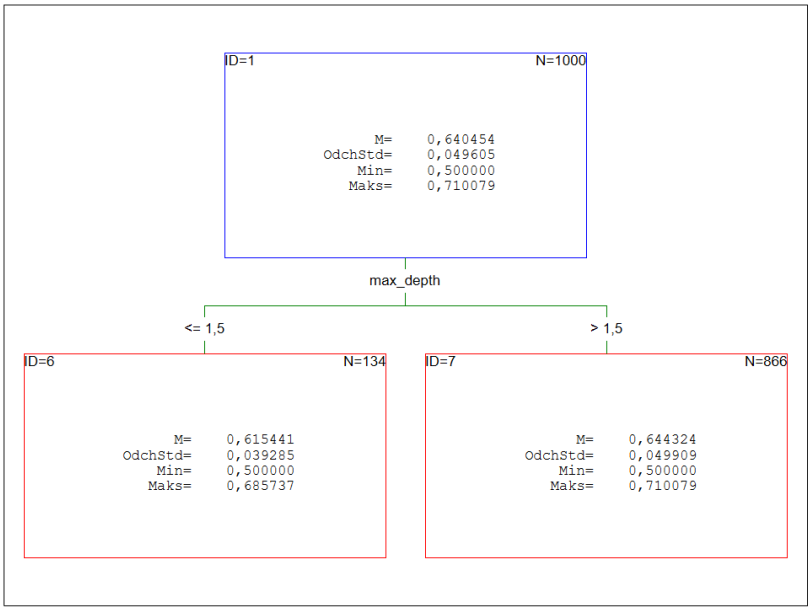
Źródło: Opracowanie własne.

W wyniku wstępnej eksploracji danych wykonano podział węzła macierzystego na podstawie trzech najsilniejszych predyktorów. Zaobserwowano znacząco gorsze od pozostałych wyniki modelu po wyborze opcji *gblinear*. Maksymalna wartość pola powierzchni pod krzywą ROC dla tej opcji wyniosła jedynie 0,634, podczas gdy dla pozostałych metod wyniosła 0,71.

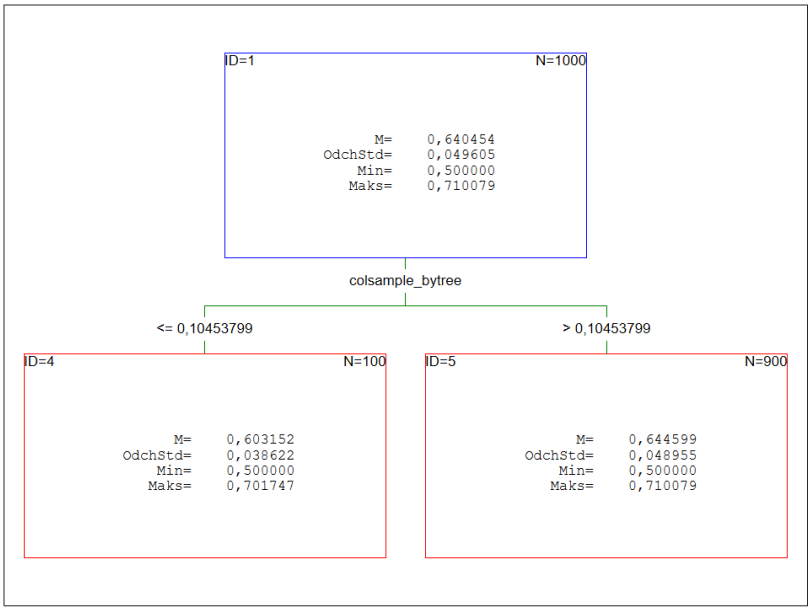


Rysunek 80 Drzewo regresyjne CART – podział drzewa na podstawie zmiennej booster
Źródło: Opracowanie własne.

W przypadku zmiennej *max_depth* oraz colsample_bytree można zauważyć, że maksymalne wartości uzyskanych modeli nie różnią się zasadniczo w zbudowanych liściach.



Rysunek 81 Drzewo regresyjne CART – podział drzewa na podstawie zmiennej max_depth
Źródło: Opracowanie własne.



Rysunek 82 Drzewo regresyjne CART – podział drzewa na podstawie zmiennej colsample_bytree

Źródło: Opracowanie własne.

Wynika stąd, że wybór mniej korzystnej wartości danego hiperparametru mógł być kompensowany przez wartości innych parametrów. Pogłębioną analizę uzyskanego zbioru przeprowadzono za pomocą perceptronu wielowarstwowego z jedną warstwą ukrytą[138]. Podczas budowy wykorzystano opcję automatycznego projektowania sieci umożliwiającą dobór optymalnej architektury sieci. Przed analizą zbiór danych (arkusz zawierający 1000 przypadków z hiperparametrami oraz wynikiem modelowania) został podzielony w sposób losowy na trzy próby:

- uczącą, na podstawie której budowano model – 70% przypadków,
- testową, służącą do monitorowania procesu budowy sieci – 15% przypadków,
- walidacyjną, służącą do końcowej oceny modelu – 15% przypadków.

W wyniku analizy polegającej na losowym doborze hiperparametrów oraz budowie kilkuset modeli zidentyfikowano sieć posiadającą 18 neuronów w warstwie ukrytej. W warstwie tej wykorzystano logistyczne funkcje aktywacji. W warstwie wyjściowej użyto funkcji wykładniczej. Po zidentyfikowaniu optymalnej topologii sieci wykonano 200 analogicznych modeli dla różnych wag początkowych modelu (tzw. metoda multistartu). Spośród nich wybrano 10 modeli kierując się oceną błędu średniokwadratowego

[138] Do analizy wykorzystano program Tibco Statistica.

obliczonego na próbie testowej. Zbudowane modele cechowały się zadowalającą jakością przewidywań. Na próbie testowej zaobserwowano względny błąd procentowy na poziomie od 3,9% do 4,3%. Następnie przeprowadzono analizę wrażliwości jakości modeli na usunięcie z zestawu zmiennych jednego z hiperpararametrów. Uśrednione wyniki analizy wrażliwości przedstawia

Tabela 31 Uśrednione wyniki analizy wrażliwości

Hiperparametr	Wrażliwość
booster	4,0340
max_depth	1,9092
colsample_bytree	1,5095
colsample_bynode	1,2937
learning_rate	1,2693
colsample_bylevel	1,1940
subsample	1,0314
alpha	1,0236
gamma	0,9837
scale_pos_weight	0,9806
lambda	0,9758

Źródło: Opracowanie własne.

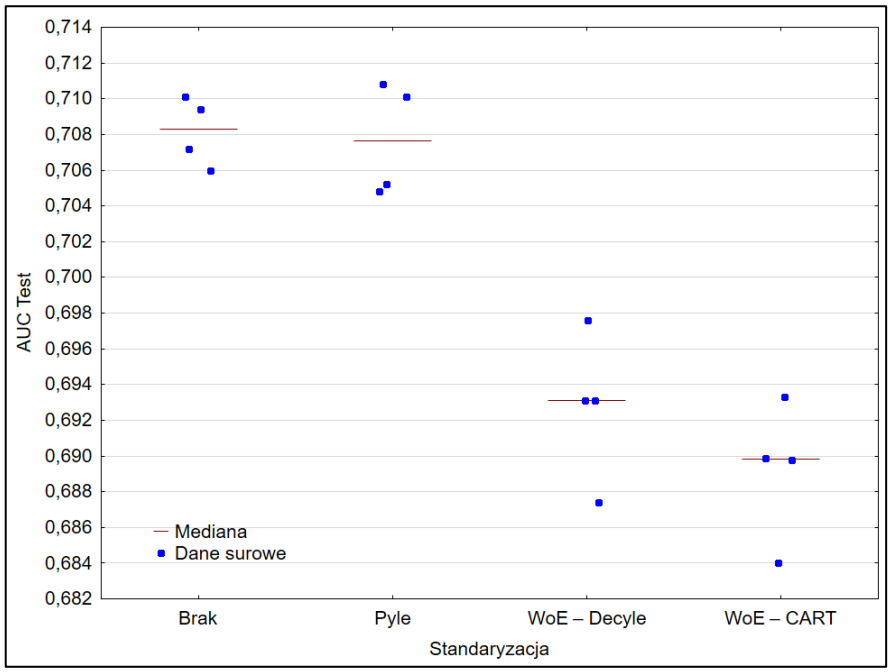
W dalszej kolejności, na podstawie uzyskanych wyników można stwierdzić, że nieuwzględnienie w modelu hiperparametru *booster* spowodowało czterokrotny wzrost błędu predykcji. Usunięcie hiperparametru *max_depth* wiąże się z prawie dwukrotnym wzrostem błędu predykcji. Z drugiej strony uwzględnienie w modelu hiperparametrów *gamma*, *scale_pos_weight* oraz *lambda* wpływało niekorzystnie na jakość prognozy. Usunięcie z modelu każdego z tych hiperparametrów z osobna skutkowało poprawą jakości modelu. Wyniki analiz przeprowadzonych na 16 zbiorach przedstawia Tabela 32.

Tabela 32 Zbiorcze wyniki procesu identyfikacji modelu XGBoost

Numer zbioru danych	Standaryzacja	Zmienne pochodne	Segmentacja	AUC (próba ucząca)	AUC (próba testowa)
1.	Brak	Nie	Nie	0,7550	0,7094
2.	Pyle	Nie	Nie	0,7967	0,7108
3.	WoE – Decyle	Nie	Nie	0,7280	0,6931
4.	WoE – CART	Nie	Nie	0,7572	0,6898
5.	Brak	Tak	Nie	0,7525	0,7101
6.	Pyle	Tak	Nie	0,8021	0,7101
7.	WoE – Decyle	Tak	Nie	0,7595	0,6976
8.	WoE – CART	Tak	Nie	0,7599	0,6933
9.	Brak	Nie	Tak	0,8142	0,7072
10.	Pyle	Nie	Tak	0,7725	0,7048
11.	WoE – Decyle	Nie	Tak	0,7501	0,6874
12.	WoE – CART	Nie	Tak	0,7700	0,6840
13.	Brak	Tak	Tak	0,7981	0,7060
14.	Pyle	Tak	Tak	0,7483	0,7052
15.	WoE – Decyle	Tak	Tak	0,8013	0,6931
16.	WoE – CART	Tak	Tak	0,7564	0,6899

Źródło: Opracowanie własne.

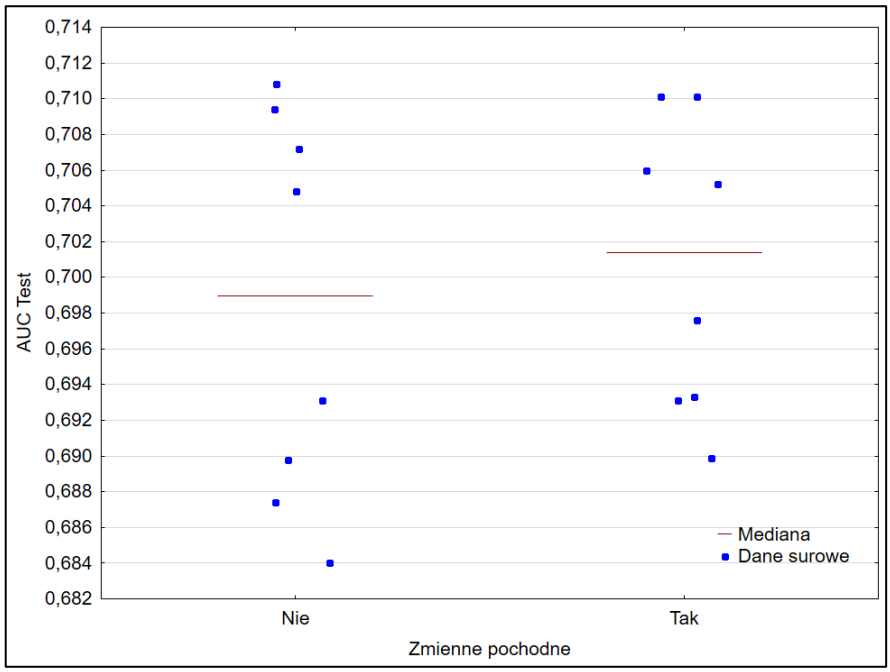
Jak łatwo zauważyć, najlepsze modele uzyskano dla zbiorów danych bez przeprowadzonej standaryzacji. Porównywalne wyniki, choć cechujące się nieco niższą medianą uzyskano dla zbioru wystandaryzowanego za pomocą standaryzacji Pyle'a. Zbiory poddane standaryzacji WoE cechowały się wyraźnie niższym wynikiem (Rysunek 83).



Rysunek 83 Poziom AUC na próbie testowej w przekroju metod standaryzacji – model XGBoost

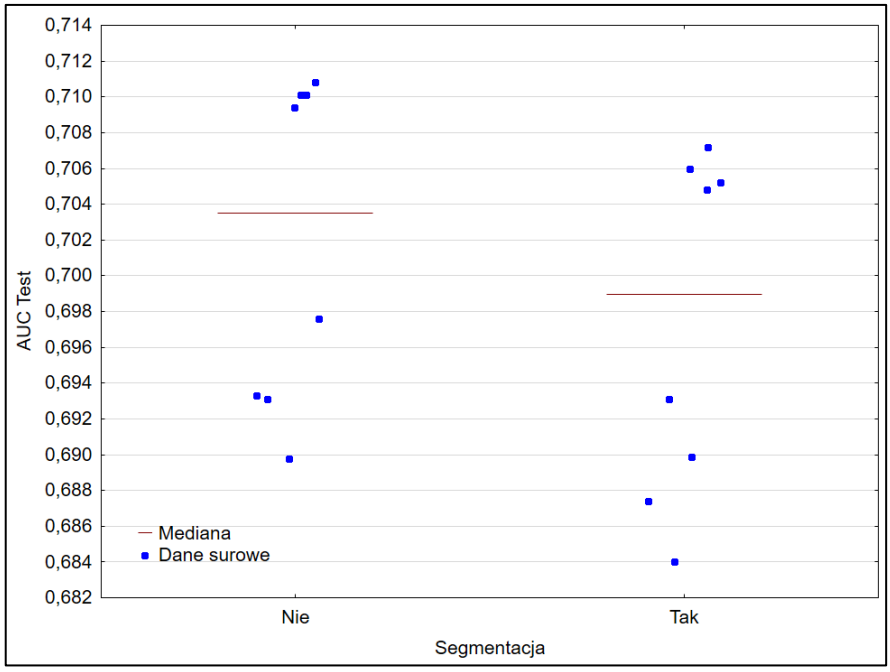
Źródło: Opracowanie własne.

Dodanie do analizowanego zbioru zmiennych pochodnych w nieznaczny sposób poprawia wynik modelowania (Rysunek 84). Wykonanie segmentacji zbioru ma z kolei negatywny wpływ na siłę predykcyjną modelu (Rysunek 85).



Rysunek 84 Poziom AUC na próbie testowej w przekroju zmiennych pochodnych – model XGBoost

Źródło: Opracowanie własne.

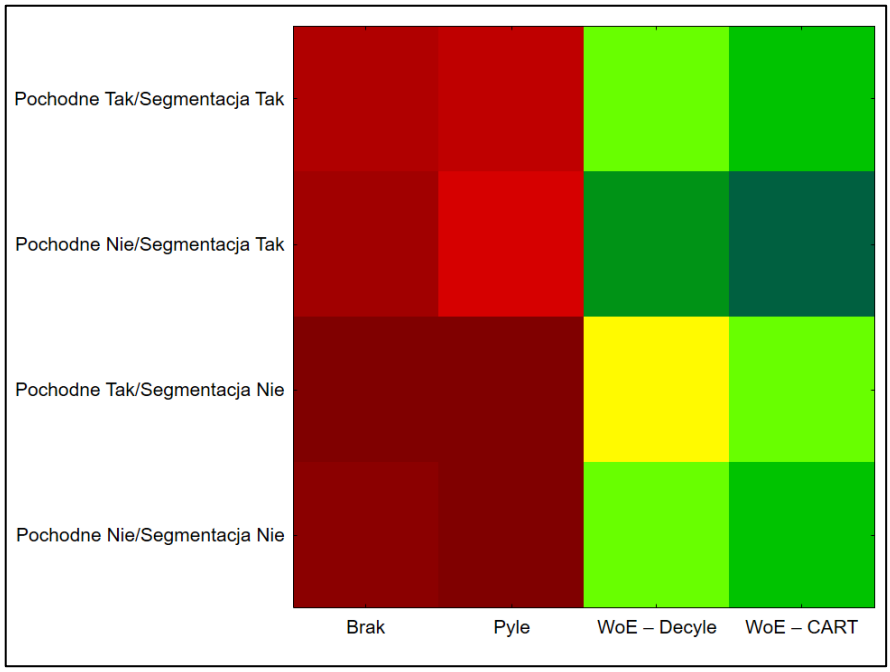


Rysunek 85 Poziom AUC na próbie testowej w przekroju segmentacji – model XGBoost

Źródło: Opracowanie własne.

Analiza wyników prezentowanych na wykresie warstwicowym (Rysunek 86) potwierdza wnioski analiz wykonanych w przekroju jednej modyfikacji zbioru. Najlepsze

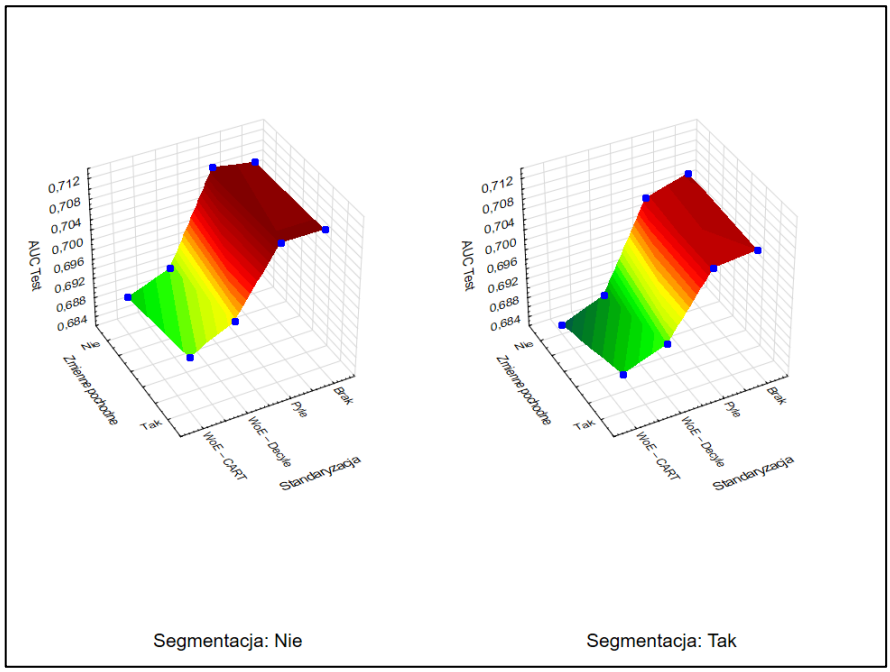
modele obserwowane są dla zbiorów bez segmentacji oraz bez standaryzacji lub ze standaryzacją Pyle'a.



Rysunek 86 Poziom AUC na zbiorze testowym w przekroju wszystkich analizowanych zbiorów – model XGBost

Źródło: Opracowanie własne.

Negatywny wpływ segmentacji na dobroć dopasowania modeli widoczny jest również na Rysunek 87. Naniesione na wykresie powierzchnie mają charakter poglądowy i mają na celu ułatwienie interpretacji występujących prawidłowości. Zaobserwowana tendencja ma odwrotny charakter w porównaniu do modeli regresji logistycznej oraz drzew klasyfikacyjnych CART.



Rysunek 87 Wykres warstwicowy w podziale na modele z segmentacją i bez – model XGBoost
Źródło: Opracowanie własne.

Ostatnim elementem analizy drzew XGBoost była agregacja pięciu modeli o największej sile predykcyjnej. Tabela 33 przedstawia wybrane, najważniejsze hiperparametry najlepszych modeli. Można zauważyć zróżnicowanie ich wartości. Agregacja pozwoliła na uzyskanie niewielkiej poprawy w stosunku do najlepszego modelu.

Tabela 33 Najlepsze konfiguracje, wybrane hiperparametry oraz model zagregowany – wyniki dla XGBoost

Model	Booster	Max depth	Colsample bytree	Colsample bynode	Learning rate	AUC (próba ucząca)	AUC (próba testowa)
M1	dart	6	0,8807	0,4255	0,0502	0,7967	0,7108
M2	gbtree	7	0,3878	0,7789	0,0511	0,8000	0,7096
M3	dart	6	0,6978	0,7812	0,0493	0,7926	0,7085
M4	gbtree	5	0,6078	0,7160	0,0186	0,7493	0,7083
M5	gbtree	6	0,9117	0,5497	0,0243	0,7482	0,7082
Zagregowany.						**0,7798**	**0,7113**

Źródło: Opracowanie własne.

5.6. Model sieci neuronowych (perceptron wielowarstwowy)

Kolejnym narzędziem analitycznym *data mining* były perceptrony wielowarstwowe[139]. Podczas analizy dokonano optymalizacji następujących hiperparametrów modelu:

- *hidden_layer_sizes* – liczba warstw ukrytych oraz ich wielkość. Przyjęto modele z jedną, dwiema lub trzema warstwami ukrytymi. Liczba neuronów w każdej warstwie wahała się od 1 do 256;
- *activation* – rodzaj funkcji aktywacji. Dostępne funkcje to liniowa, logistyczna, tangens hiperboliczny oraz ReLU;
- *solver* – algorytm uczenia sieci. Dostępne opcje to *lbfgs*, *adam* oraz *sgd*;
- *alpha* – wielkość składnika regularyzacji typu L2. Przyjęto wartości z zakresu (0;1);
- *momentum* – wartość bezwładności podczas aktualizacji wag (jedynie dla algorytmu *sgd*). Przyjęto wartości z zakresu (0;1).
- *max_iter* – liczba epok uczenia. Wartość ta była zmieniana w trakcie optymalizacji. Początkowo przyjęto wartość 10 000.
- *early_stopping* – wykorzystanie wczesnego zatrzymania, pozwalającego na zatrzymanie procesu uczenia w momencie wzrostu wartości błędu na próbie testowej. Hiperparametr ten został ustawiony na *True* dla wszystkich budowanych modeli.

Jako punkt odniesienia przyjęto sieć zbudowaną na podstawie pierwszej konfiguracji danych, bez określania ustawień modelu. Wszystkie hiperparametry zostały ustawione automatycznie. W wyniku uczenia modelu uzyskano sieć o własnościach zaprezentowanych w Tabela 34.

Tabela 34 Wynik modelowania bez optymalizacji hiperparametrów

Numer zbioru danych	Standaryzacja	Zmienne pochodne	Segmentacja	AUC (próba ucząca)	AUC (próba testowa)
1.	Brak	Nie	Nie	0,7578	0,6455

Źródło: Opracowanie własne.

[139] Obliczenia wykonane zostały za pomocą implementacji dostępnej bibliotece *sklearn* w języku Python.

Przyjęto kilkuetapowy proces identyfikacji optymalnego zestawu hiperparametrów. W pierwszym kroku wylosowano 1000 zestawów ustawień, na podstawie których zbudowano 1000 modeli. Uzyskane wyniki w postaci arkusza z wartościami hiperparamtrów oraz odpowiadających im wartościom AUC dla próby testowej zostały przeanalizowane za pomocą drzew regresyjnych CART. W drzewach tych predyktorami były zmienne określające ustawienia modelu, zmienną zależną wynik modelu na próbie testowej. Wykonanie analizy pozwoliło wyodrębnić podgrupę sieci o najwyższej sile predykcyjnej. Sieci te charakteryzowały się następującymi ustawieniami:

- liczba warstw ukrytych nie większa niż dwie;
- liczba neuronów w pierwszej warstwie nie większa niż 32;
- liczba neuronów w drugiej warstwie nie większa niż 128;
- algorytm uczenia sieci *lbfgs*;
- funkcja aktywacji tangens hiperboliczny lub *ReLU*.

Wartości składnika regularyzacji oraz bezwładności okazały się nie mieć większego wpływu na uzyskane rezultaty. Wnioski z pierwszej iteracji pozwoliły zawęzić przestrzeń hiperparametrów do zakresów opisanych powyżej. Na tej podstawie zbudowano kolejnych 1000 modeli, zwiększając dodatkowo liczbę epok do 20 000. Spośród uzyskanych wyników wybrano 100 najlepszych ustawień, dla których zwiększono liczbę epok do 200 000. W wyniku analizy pierwszego zbioru danych najlepsza okazała się sieć o następujących ustawieniach:

- algorytm uczenia sieci *lbfgs*;
- funkcja aktywacji tangens hiperboliczny;
- dwie warstwy ukryte;
- liczba neuronów w pierwszej warstwie ukrytej równa 8;
- liczba neuronów w drugiej warstwie ukrytej równa 16.

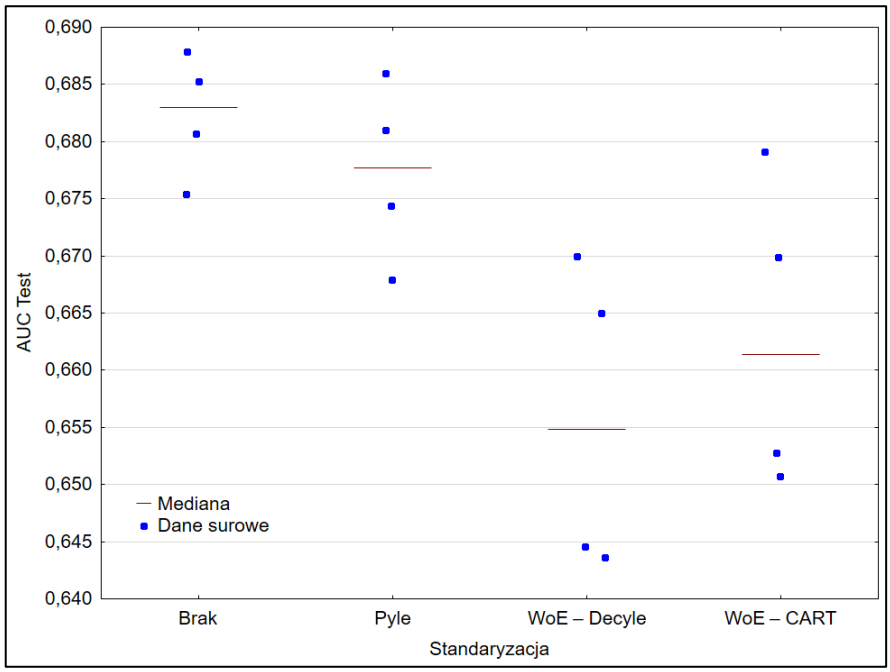
Dla sieci o powyższych ustawieniach uzyskano pole powierzchni pod krzywą ROC dla zbioru uczącego AUC=0,7053, dla zbioru testowego AUC= 0,6853. Wyniki dla wszystkich 16 zbiorów przedstawia Tabela 35.

Tabela 35 Wyniki jakości modeli dla sieci neuronowych MLP

Numer zbioru danych	Standaryzacja	Zmienne pochodne	Segmentacja	AUC (próba ucząca)	AUC (próba testowa)
1.	Brak	Nie	Nie	0,7053	0,6853
2.	Pyle	Nie	Nie	0,6969	0,6860
3.	WoE – Decyle	Nie	Nie	0,6779	0,6700
4.	WoE – CART	Nie	Nie	0.6996	0,6699
5.	Brak	Tak	Nie	0,7009	0,6879
6.	Pyle	Tak	Nie	0,7062	0,6810
7.	WoE – Decyle	Tak	Nie	0,6921	0,6650
8.	WoE – CART	Tak	Nie	0,6995	0,6791
9.	Brak	Nie	Tak	0,7027	0,6754
10.	Pyle	Nie	Tak	0,7132	0,6744
11.	WoE – Decyle	Nie	Tak	0,7117	0,6446
12.	WoE – CART	Nie	Tak	0,7218	0,6528
13.	Brak	Tak	Tak	0,7108	0,6807
14.	Pyle	Tak	Tak	0,7281	0,6679
15.	WoE – Decyle	Tak	Tak	0,7261	0,6436
16.	WoE – CART	Tak	Tak	0,7327	0,6507

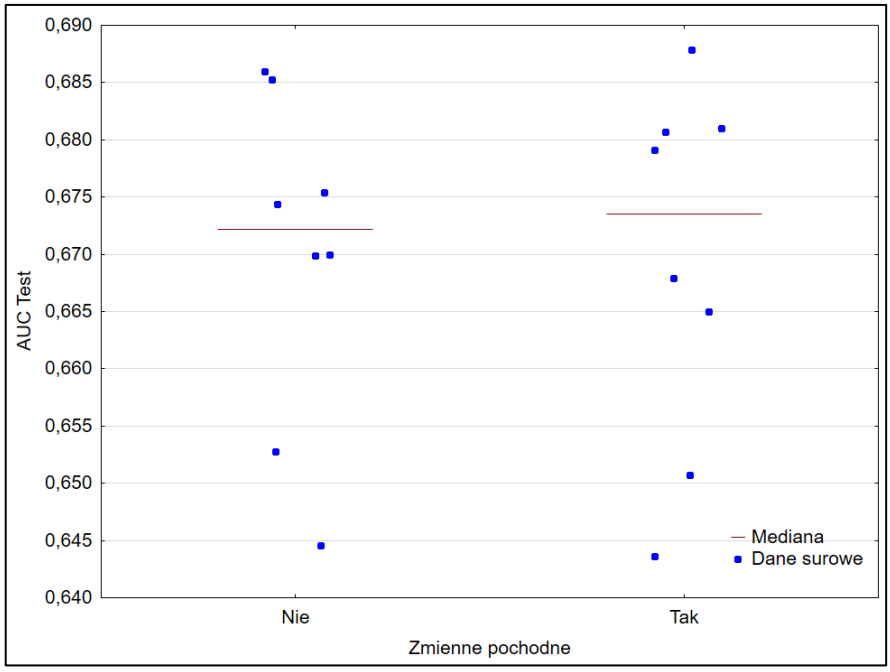
Źródło: Opracowanie własne.

Analiza uzyskanych wyników pozwala stwierdzić, że modyfikacja zbioru polegająca na standaryzacji (Rysunek 88) oraz segmentacji (Rysunek 89) nie powoduje poprawy dobroci dopasowania modelu sieci neuronowych. Nieznaczna poprawa jest obserwowana jedynie w przypadku zmiennych pochodnych (Rysunek 90).



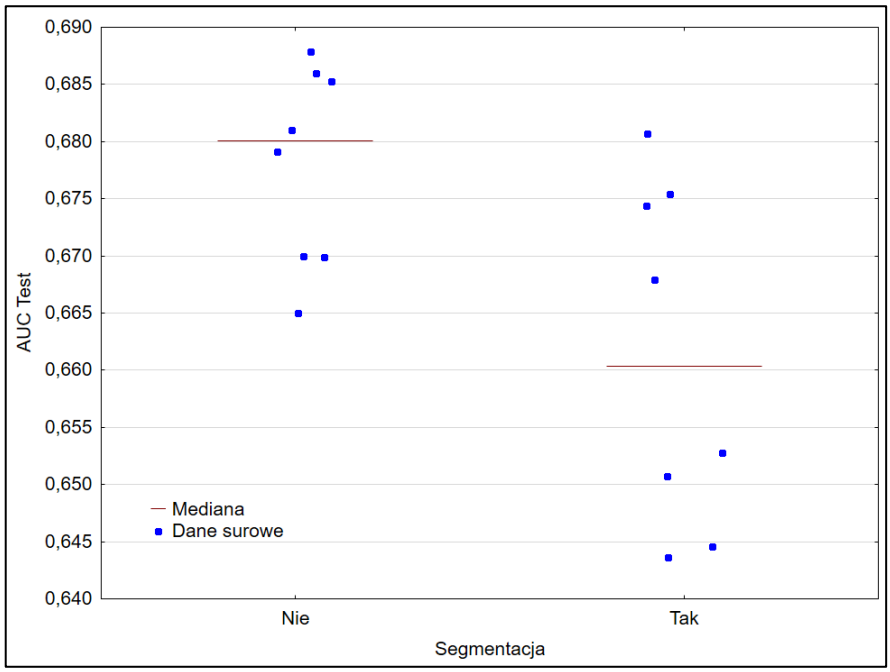
Rysunek 88 Poziom AUC na próbie testowej w przekroju standaryzacji – model sieci neuronowych

Źródło: Opracowanie własne.



Rysunek 89 Poziom AUC na próbie testowej w przekroju zmiennych pochodnych – model sieci neuronowych

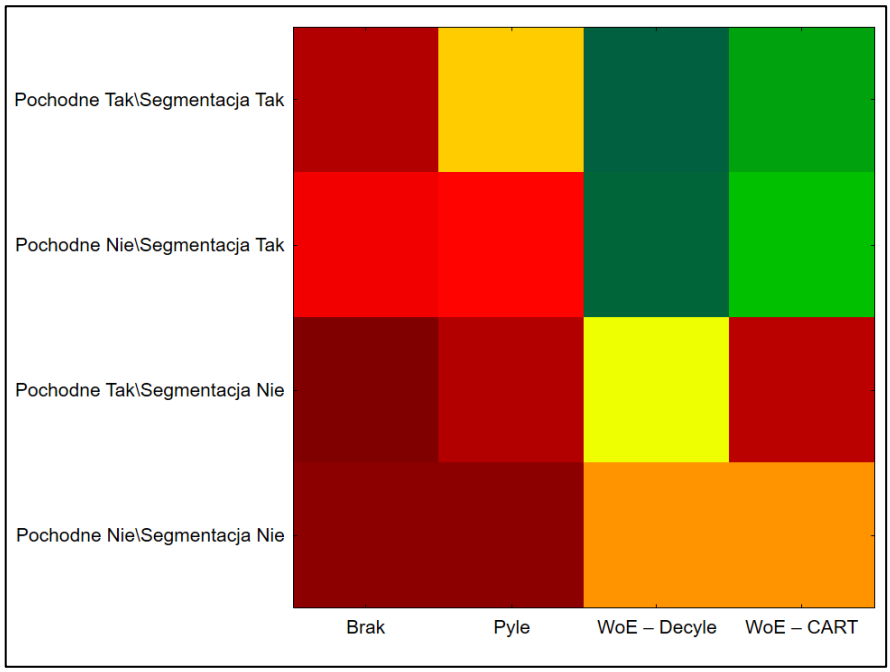
Źródło: Opracowanie własne.



Rysunek 90 Poziom AUC na próbie testowej w przekroju segmentacji – model sieci neuronowych

Źródło: Opracowanie własne.

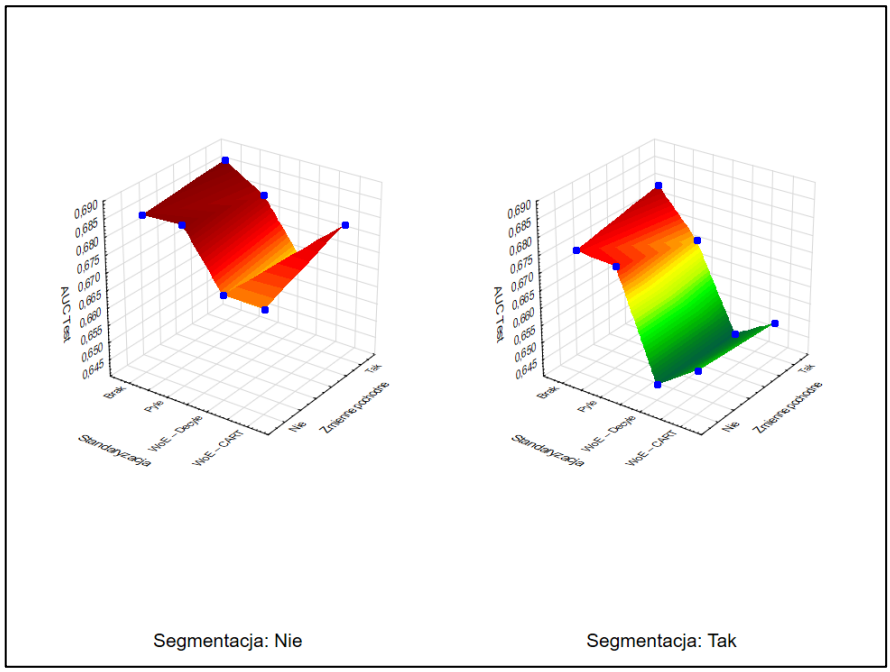
Powyższą obserwację potwierdza analiza wykresów warstwicowych. Rysunek 91 uwidacznia dychotomię pomiędzy modelami budowanymi na zbiorze bez standaryzacji a zbiorami standaryzowanymi za pomocą WoE.



Rysunek 91 Poziom AUC na zbiorze testowym w przekroju wszystkich analizowanych zbiorów – model sieci neuronowych

Źródło: Opracowanie własne

Na Rysunek 92 widoczny jest negatywny wpływ segmentacji na jakość modelowania. Naniesione na wykresie powierzchnie mają charakter poglądowy, a ich celem jest ułatwienie interpretacji występujących prawidłowości. Tendencja ta jest zbliżona do tej zaobserwowanej w przypadku modeli budowanych za pomocą biblioteki *XGBoost*.



Rysunek 92 Wykres warstwicowy w podziale na modele z segmentacją i bez – model sieci neuronowych

Źródło: Opracowanie własne.

Można zauważyć, że dla wszystkich konfiguracji uzyskane wyniki są lepsze od wyników uzyskanych dla ustawień domyślnych. Najlepsze wyniki uzyskano na podstawie konfiguracji piątej, czyli danych surowych, uzupełnionych o zmienne pochodne. Dla zwycięskiej konfiguracji przeprowadzono agregację 5 modeli, dla których uzyskano najlepsze wyniki na podstawie miary AUC obliczonej dla zbioru testowego.

Tabela 36 Najlepsze konfiguracje oraz model zagregowany – wyniki

Model	Funkcja aktywacji	Alpha	Neurony warstwa 1	Neurony warstwa 2	AUC (próba ucząca)	AUC (próba testowa)
M1	Tanh	0,8	4	32	0,7009	0,6879
M2	Tanh	0,6	4	32	0,7007	0,6878
M3	Tanh	0,9	4	32	0,7008	0,6877
M4	Tanh	0,85	4	32	0,7007	0,6877
M5	Tanh	0,9	4	16	0,7005	0,6875
Zagregowany.					**0,7000**	**0,6889**

Źródło: Opracowanie własne.

Najlepsze sieci nie różnią się zasadniczo, jeśli chodzi zarówno o wartości hiperprametrów jak i ich siłę predykcyjną. Wynik uzyskany w wyniku agregacji sieci pozwala uzyskać poprawę dobroci dopasowania sieci.

Na podstawie przeprowadzonych badań można stwierdzić, że czynniki wpływające na jakość modeli retencji klientów działają na nie inaczej w przekroju różnych metod. Opisane w pierwszej części rozdziału aspekty TIVHE przełożyły się w praktyce na pięć czynników, które zostały uwzględnione w sposób uwzględniający specyfikę wykorzystywanych metod analitycznych. W trakcie badań wzięto pod uwagę kwestie:

- transformacji zmiennych,
- wprowadzenia do modelu zmiennych pochodnych,
- optymalizacji hiperparametrów,
- segmentacji zbioru danych,
- agregacji modeli.

Podsumowanie wpływu analizowanych czynników na siłę predykcyjną budowanych modeli przedstawia Tabela 37.

Tabela 37 Wybrane czynniki wpływające na jakość modeli w przekroju analizowanych metod

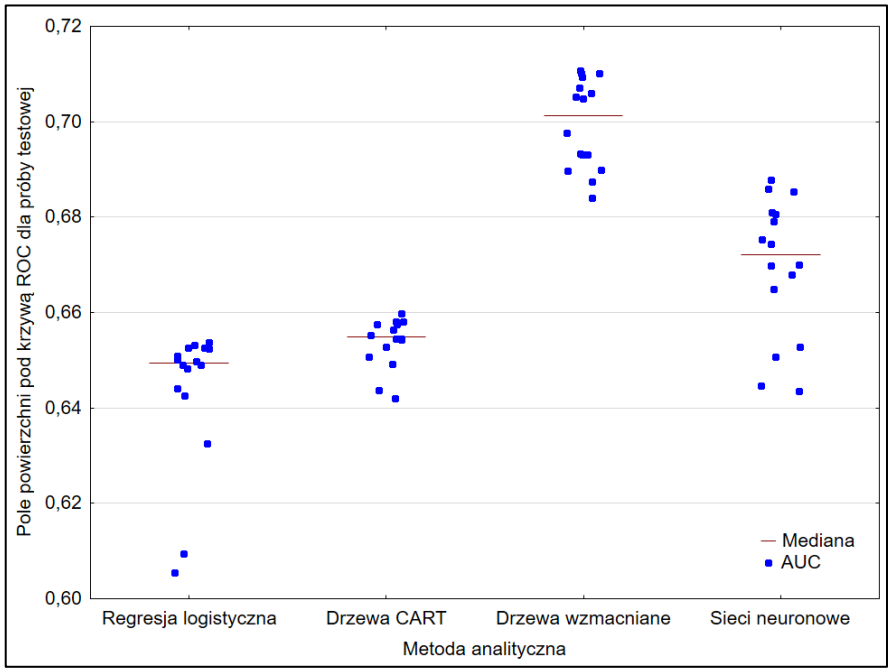
Regresja logistyczna	+	+	–	+	–
Drzewa CART	–	+	+	+	+
Drzewa wzmacniane	–	+	+	–	+
Sieci neuronowe	–	+	+	–	+

Źródło: opracowanie własne.

Analizując uzyskane wyniki można stwierdzić, że aspekt transformacji zmiennych okazał się znaczący jedynie w przypadku regresji logistycznej i miał on dla tej metody relatywnie duży wpływ na jakość modelu. Dla pozostałych metod wpływ ten był neutralny, bądź wręcz pogarszający uzyskane wyniki. Dodanie zmiennych pochodnych było jedynym czynnikiem wpływającym pozytywnie na wyniki wszystkich analizowanych metod. Podobny efekt, z pominięciem regresji logistycznej, dla której nie był on rozważany, zauważono dla optymalizacji hiperparametrów. Segmentacja zbioru danych okazała się czynnikiem poprawiającym jedynie jakość modeli „białoskrzynkowych". W przypadku regresji logistycznej jej wpływ był wręcz dominujący w przekroju badanych czynników.

Dla modeli „czarnoskrzynkowych” stwierdzono negatywny wpływ segmentacji na siłę predykcyjną. Agregacja modeli okazała się czynnikiem poprawiającym jakość wszystkich modeli za wyjątkiem regresji logistycznej. W przypadku tego modelu wpływ agregacji okazał się neutralny tj. nie poprawiał ani nie obniżał trafności predykcji modelu.

Aspektem wartym odnotowania jest porównanie działania poszczególnych metod na analizowanym zbiorze. Uzyskane wyniki przedstawia Rysunek 93. Na jego podstawie można stwierdzić, że najwyższą siłą predykcyjną cechowały się modele budowane za pomocą drzew wzmacnianych. Regresja logistyczna oraz drzewa CART uzyskały porównywalne, słabsze wyniki. W środku zestawienia znajduje się model sieci neuronowych.



Rysunek 93 Porównanie siły predykcyjnej metod wykorzystanych w badaniu
Źródło: opracowanie własne.

Należy dodatkowo zwrócić uwagę na fakt, że modele zwycięskie w swojej klasie narzędzi analitycznych poddano dodatkowej ocenie na walidacyjnym zbiorze danych. Nie stwierdzono znaczących różnic w działaniu zbudowanych modeli w porównaniu do zbioru testowego, co świadczć może o braku nadmiernego dopasowania modeli.

Zakończenie

Proces budowy modelu retencji klientów jest zadaniem wieloetapowym i złożonym, zależnym od wielu czynników będących niejednokrotnie w interakcji. W pracy przedstawiono kolejne etapy procesu budowy modelu *churn* począwszy od identyfikacji biznesowych parametrów modelu takich jak definicja „złego" klienta czy określenie sposobów definiowania okresu obserwacji. Kolejne rozdziały dysertacji stanowią przegląd strategii i technik stosowanych na kolejnych etapach procesu modelowania i odnoszą się do: 1) przygotowania danych podczas budowy modeli retencji klientów; 2) budowy optymalnego modelu klasyfikacyjnego oraz 3) walidacji i wdrażania modeli retencji klientów.

Sposród aspektów przedstawionych w rozdziałach teoretycznych na dodatkową uwagę może zasługiwać część związana z miarami oceny siły predykcyjnej modelu. W części tej poza przedstawieniem miar wykorzystywanych w praktyce badawczej przeprowadzono dodatkowe badania oceny wrażliwości poszczególnych miar na zmiany proporcji klas modelowanej zmiennej zależnej.

Przedstawiona w początkowej części pracy teoria stała się podstawą do przeprowadzenia badań, których głównym celem była identyfikacja determinant wpływających na jakość modeli migracji klientów oraz określenie relacji między nimi. Na podstawie przeprowadzonych badań można stwierdzić, że czynniki wpływające na jakość modeli retencji klientów działają na nie inaczej w przekroju różnych metod. Opisane w pracy aspekty procedury TIVHE przełożyły się w praktyce na pięć czynników, które zostały uwzględnione w sposób uwzględniający specyfikę wykorzystywanych metod analitycznych. W trakcie badań wzięto pod uwagę kwestie: 1) transformacji zmiennych, 2) wprowadzenia do modelu zmiennych pochodnych, 3) optymalizacji hiperparametrów, 4) segmentacji zbioru danych oraz 5) agregacji modeli.

Analizując uzyskane wyniki można stwierdzić, że aspekt transformacji zmiennych okazał się znaczący jedynie w przypadku regresji logistycznej i miał on dla tej metody relatywnie duży wpływ na jakość modelu. Dla pozostałych metod wpływ ten był neutralny,

bądź wręcz pogarszający uzyskane wyniki. Dodanie zmiennych pochodnych było jedynym czynnikiem wpływającym pozytywnie na wyniki wszystkich analizowanych metod. Podobny efekt, z pominięciem regresji logistycznej, dla której nie był on rozważany, zauważono dla optymalizacji hiperparametrów. Segmentacja zbioru danych okazała się czynnikiem poprawiającym jedynie jakość modeli „białoskrzynkowych". Jej efekt był szczególnie widoczny w połączeniu z dodaniem zmiennych pochodnych. W przypadku regresji logistycznej wpływ tego zabiegu był wręcz dominujący w przekroju badanych czynników. Dla modeli „czarnoskrzynkowych" stwierdzono negatywny wpływ segmentacji na siłę predykcyjną. Agregacja modeli okazała się z kolei czynnikiem poprawiającym jakość wszystkich modeli za wyjątkiem regresji logistycznej. W przypadku tego modelu wpływ agregacji okazał się neutralny tj. nie poprawiał ani nie obniżał trafności predykcji modelu.

Kolejnym aspektem analizy uzyskanych wyników łączący się z pobocznymi celami pracy było porównanie działania poszczególnych metod.. Na jego podstawie można stwierdzić, że najwyższą siłą predykcyjną cechowały się modele budowane za pomocą drzew wzmacnianych. Regresja logistyczna oraz drzewa CART uzyskały porównywalne, słabsze wyniki. W środku zestawienia znajduje się model sieci neuronowych.

Badanie nie wykazało zatem możliwości budowy modeli za pomocą metod interpretowalnych przez badacza (biała skrzynka), porównywalnych z zaawansowanymi metodami nieinterpretowalnymi (czarna skrzynka), przy użyciu odpowiednich technik przygotowania danych oraz hybrydyzacji modeli. Metody „białoskrzynkowe" budowane za pomocą najbardziej optymalnej strategii modelowania uzyskały efekt gorszy od wszystkich wyników uzyskanych za pomocą drzew wzmacnianych (niezależnie od strategii modelowania i wariantu zbioru danych) oraz od zdecydowanej większości modeli sieci neuronowych. Regresja logistyczna może do pewnego stopnia konkurować pod względem jakości z modelami drzew CART, jedynie w sytuacji, gdy w procesie modelowania wykorzystana zostanie technika hybrydyzacji oraz dodane zostaną zmienne pochodne pozwalające na uwzględnienie w modelu interakcji.

Kolejnym pobocznym celem pracy była ocena skuteczności modeli drzew klasyfikacyjnych budowanych za pomocą alternatywnych ścieżek podziału. W wyniku ingerencji w naturalny przebieg algorytmu zbudowano modele, dla których wymuszono jedynie pierwszy a także pierwszy i drugi podział drzewa kolejnymi zmiennymi z listy najlepszych predyktorów. Po wykonaniu wymuszonych podziałów, kolejne przebiegały w

sposób automatyczny. Żaden ze zbudowanych w ten sposób modeli nie dostarczył wyniku lepszego od modelu bazowego.

Dodatkowym istotnym aspektem dociekań przeprowadzonych w dysertacji była identyfikacja optymalnej ścieżki budowy modelu ekonometrycznego, na przykładzie regresji logistycznej przy wykorzystaniu szeregu technik selekcji zmiennych opartych zarówno na filtrach, jak również metodach wbudowanych. Dla tej metody przeprowadzono wieloaspektową selekcję zmiennych biorącą pod uwagę zarówno siłę predykcyjną jak również aspekt współliniowości zmiennych oraz zgodności znaków ocen parametrów regresji ze znakami dla modeli uwzględniającymi jedną zmienną. Metodą selekcji zmiennych dającą najbardziej obiecujące wyniki okazała się strategia łącząca technikę LASSO z techniką *Branch&Bound*. Za pomocą techniki LASSO dokonano preselekcji zmiennych ograniczając zbiór potencjalnych predyktorów do kilkudziesięciu wykorzystując optymalizację hiperparametru *lambda*. Końcowa selekcja została przeprowadzona za pomocą techniki *Branch&Bound.* Uzyskany wynik potwierdza wiedzę autora wyniesioną z analogicznych, komercyjnych projektów analitycznych.

W ocenie autora niniejsza praca w zadowalającym stopniu wypełniła lukę badawczą w obszarze wieloaspektowej oceny jakości modeli retencji klientów. Pozwoliła na syntezę wpływu wielu determinant jakości modeli na końcowy rezultat modelowania. Autor dołożył wszelkich starań, aby z dostępnych repozytoriów wybrać zbiór danych o złożoności jak najbardziej zbliżonej do rzeczywistych zbiorów danych, w którym występowałyby cechy statystyczne typowe dla tego typu zjawisk. Autor jest zarazem świadomy ograniczeń wynikających z analizy jednego zbioru danych. Dodatkowa pogłębiona analiza innych zbiorów danych w celu potwierdzenia obserwacji poczynionych w niniejszej pracy może być kierunkiem nowych dociekań naukowych w tym zakresie. Nie mniej jednak uzyskane wyniki mogą przyczynić się do identyfikacji strategii budowy modeli najczęściej prowadzących do uzyskania pożądanych przez badacza rezultatów. Wiedza oparta na wynikach badań może być przyczynkiem do opracowania narzędzi wspierających pracę analityka na przykład w postaci kreatorów w bezpłatnych lub komercyjnych narzędziach analitycznych.

Bibliografia

1. Aggarwal C. C., 2017, *Outlier Analysis. Second Edition*, Springer, New York.
2. Aggarwal C. C., 2018, *Neural networks and deep learning*, Springer, Nowy Jork.
3. Alelyani S., Tang J., Liu H., 2013, *Feature selection for clustering: a review*, Data clustering: algorithms and applications, nr 29, s. 110-121.
4. Allison P. D., 2001, *Missing data (vol. 136)*, Sage publications.
5. Ahmad A.K., Jafar A., Aljoumaa K., 2019, *Customer churn prediction in telecom using machine learning in big data platform*, Journal of Big Data, nr 6(1) s.28.
6. Arlot S., Celisse A., 2010, *A survey of cross-validation procedures for model selection*, Statistics surveys, nr 4, s. 40-79.
7. Baesens B., 2014, *Analytics in a big data world: The essential guide to data science and its applications*, John Wiley & Sons.
8. Bahga A., Madisetti V., 2016, *Big data science & analytics: A hands-on approach*, VPT.
9. Bagnall A., Flynn M., Large J., Line J., Bostrom A., Cawley G., 2018, *Is rotation forest the best classifier for problems with continuous features?*, arXiv preprint arXiv:1809.06705.
10. Bandaragoda T. R., Ting K. M., Albrecht D., Liu F. T., Zhu Y., Wells J. R., 2018, *Isolation-based anomaly detection using nearest-neighbor ensembles*, Computational Intelligence, nr 34(4), s. 968-998.
11. Bergstra J., Bengio Y., 2012, *Random search for hyper-parameter optimization*, Journal of machine learning research, nr 13(2).
12. Berrar D., 2019, *Performance Measures for Binary Classification*, Encyclopedia of Bioinformatics and Computational Biology, vol. 1, s. 546-560.
13. Berry L.L., 1983, *Relationship Marketing* [w:] *Emerging Perspectives in Services Marketing* red. L.L. Berry, G.L. Shostack, G.D. Upah, American Marketing Association, Chicago, s. 25-28.
14. Berry M., Linoff G., 2004, *Mastering data mining: The art and science of customer relationship management*, John Wiley & Sons.
15. Bhattacharya C.B., Bolton, R.N., 2000, *Relationship Marketing in Mass Markets* [w:] *Handbook of relationship marketing* red. A. Parvatiyar, J. N. Sheth, Sage Publications, s. 327-354.
16. Bickel P. J., Götze F., van Zwet W. R., 2012, *Resampling fewer than n observations: gains, losses, and remedies for losses*, Selected works of Willem van Zwet, s. 267-297, Springer, New York.
17. Biggerstaff B. J., 2000, *Comparing diagnostic tests: a simple graphic using likelihood ratios*, Statistics in medicine 19.5, s. 649-663.

18. Blum A., Mitchell T., 1998, *Combining labeled and unlabeled data with co-training* [w:] *Proceedings of the eleventh annual conference on Computational learning theory*, s. 92-100.
19. Błażejczyk-Majka L., 2018, *Zastosowanie wybranych metod taksonomicznych w badaniach historycznych*, Instytut Historii UAM.
20. Boughorbel S., Jarray F., El-Anbari M., 2017, *Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric*, PloS one 12.6.
21. Breiman L., Friedman J. H., Olshen R., Stone C. J., 1984, *Classification and Regression Trees*, Wadsworth Belmont, CA.
22. Breiman L., 2001, *Random forests*, Machine learning, nr 45(1), s. 5-32.
23. Brier G. W., 1950, *Verification of forecasts expressed in terms of probability*, Monthey Weather Review, nr 78(1), s. 1-3.
24. Burkov A., 2019, *The hundred-page machine learning book Vol. 1*., Quebec City, Canada.
25. Castanedo F., Valverde G., Zaratiegui J., Vazquez A., 2014, *Using deep learning to predict customer churn in a mobile telecommunication network*, Wise Athena LLC.
26. Chandola V., Banerjee A., Kumar V., 2009, *Anomaly detection: A survey*, ACM computing surveys (CSUR), nr 41(3), s. 1-58.
27. Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P., 2002, *SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research*.
28. Chen T., Guestrin C., 2016, *Xgboost: A scalable tree boosting system*, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, s. 785-794.
29. Chollet F., 2019, *Deep Learning Praca z językiem Python i biblioteką Keras*, Helion.
30. Chomątowski S., Sokołowski, A., 1978, *Taksonomia struktur*, Przegląd Statystyczny, nr 2.
31. Christopher M., Payne A., Ballantyne D., 2008, *Relationship marketing. Creating Stakeholder Value*, Butterworth Heinemann, Oxford.
32. Clevert D. A., Unterthiner T., Hochreiter S., 2015, *Fast and accurate deep network learning by exponential linear units (elus)*, arXiv preprint arXiv:1511.07289.
33. CRISP-DM, S. P. S. S., 2000, *Step-by-step Data Mining Guide*.
34. Curasi F., Kennedy N., 2002, *From prisoners to apostles: a typology of repeat buyers and loyal customers in service businesses*, Journal of services marketing, nr 16(4), s. 322-341.
35. Doak J., 1992, *An evaluation of feature selection methods and their application to computer security*, Techninal Report CSE, s. 92-18.
36. Drapińska A., 2013, *Pomiar lojalności klientów-wybrane wskaźniki*, Zeszyty Naukowe Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie, Polityki Europejskie, Finanse i Marketing, nr 09 (58).
37. East R., Singh J., Wright M., Vanhuele M., 2016, *Consumer behaviour: Applications in marketing*, Sage.
38. Efron B., 1979, *Bootstrap methods: another look at the jackknife*, The Annals of Statistics, nr 7(1), s. 1-26.
39. Emmott A., Das S., Dietterich T., Fern A., Wong, W. K., 2015, *A meta-analysis of the anomaly detection problem*, arXiv preprint arXiv:1503.01158.

40. Falkner S., Klein A., Hutter F., 2018, *BOHB: Robust and efficient hyperparameter optimization at scale*, International Conference on Machine Learning, s. 1437-1446.
41. Fernández A., Garcia S., Herrera F., Chawla N. V., 2018, *SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary*, Journal of artificial intelligence research 61, s. 863-905.
42. Ferri C., Hernández-Orallo J., Modroiu R., 2009, *An experimental comparison of performance measures for classification*, Pattern Recognition Letters, nr 30.1, s. 27-38.
43. Feurer M., Hutter F., 2019, *Hyperparameter optimization*, Automated Machine Learning, s. 3-33, Springer.
44. Flach P., 2012, *Machine learning: the art and science of algorithms that make sense of data*, Cambridge University Press.
45. Fok D., 2003, *Advanced econometric marketing models*, ERIM Ph.D. Series Research in Management nr 27.
46. Freund Y., Mason L., 1999, *The alternating decision tree learning algorithm*, icml, nr 99, s. 124-133.
47. Freund Y., Schapire R., Abe N., 1999, *A short introduction to boosting*, Journal-Japanese Society For Artificial Intelligence, nr 14, s. 771-780.
48. Furtak R., 2003, *Marketing partnerski na rynku usług*, Polskie Wydawnictwo Ekonomiczne.
49. García D. L., Nebot À., Vellido A., 2017, *Intelligent data analysis approaches to churn as a business problem: a survey*, Knowledge and Information Systems, nr 51(3), s. 719-774.
50. Gatnar E., 2008, *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, PWN, Warszawa.
51. Géron A., 2017, *Hands-On Machine Learning with Scikit-Learn & TensorFlow*, O'Relly, Sebastopol.
52. Glorot X., Bengio Y., 2010, *Understanding the difficulty of training deep feedforward neural networks*, Proceedings of the thirteenth international conference on artificial intelligence and statistics, s. 249-256.
53. Grabmeier J. L., Lambe L. A., 2007, *Decision trees for binary classification variables grow equally with the Gini impurity measure and Pearson's chi-square test*, International journal of business intelligence and data mining, nr 2(2), s. 213-226.
54. Griffin J., 1995, *Customer Loyalty*, Jossey – Bass Publisher, San Francisco.
55. Guidici P., 2003, *Applied data mining. Statistical methods for business and industry*, John Wiley and Sons, Norfolk.
56. Guyon I., Gunn S., Nikravesh M., Zadeh L. A., 2008, *Feature extraction: foundations and applications*, Springer.
57. Haixiang G., Yijing L., Shang J., Mingyun G., Yuanyue H., Bing G., 2017, *Learning from class-imbalanced data: Review of methods and applications*, Expert Systems with Applications, nr 73, s. 220-239.
58. Hall M. A., 1999, *Correlation-based feature selection for machine learning*, Doctoral dissertation, The University of Waikato.
59. Hariri M., Carrasco K., Brunner R. J., 2018, *Extended isolation forest*, arXiv preprint arXiv:1811.02141v3.

60. Harrell Jr. F.E., 2015, *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*, Springer, New York.
61. Hastie T, Tibshirani R., Friedman J., 2009, *The elements of statistical learning. Data mining, Trends and Prediction*, Springer, New York.
62. He H., Bai Y., Garcia E. A., Li S., 2008, *ADASYN: Adaptive synthetic sampling approach for imbalanced learning*, 2008 IEEE international joint conference on neural networks, s. 1322-1328.
63. He K., Zhang X., Ren S., Sun J., 2015, *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*, Proceedings of the IEEE international conference on computer vision, s. 1026-1034.
64. Hill N., Alexander J., 2003, *Pomiar satysfakcji i lojalności klientów*, Oficyna Ekonomiczna.
65. Hosmer Jr D. W., Lemeshow S., Sturdivant R. X., 2013, *Applied logistic regression*, John Wiley & Sons.
66. Huxley T. H., 1854, *On the educational value of the natural history sciences*, J. Van Voorst.
67. Idris A., Khan A., Lee Y. S., 2013, *Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification*, Applied intelligence, nr 39, s. 659-672.
68. Ioffe S., Szegedy C., 2015, *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, International conference on machine learning, s. 448-456.
69. Jajuga K., 1993, *Statystyczna analiza wielowymiarowa*, Wydawnictwo Naukowe PWN.
70. Jaśkowski, M., Jaroszewicz, S., 2012, *Uplift modeling for clinical trial data*, ICML Workshop on Clinical Data Analysis.
71. Jerez J. M., Molina I., García-Laencina P. J., Alba E., Ribelles, N., Martín M., Franco L., 2010, *Missing data imputation using statistical and machine learning methods in a real breast cancer problem*, Artificial intelligence in medicine, nr 50(2), s.105-115.
72. Johnson R., Zhang T., 2014, *Learning nonlinear functions using regularized greedy forest*, IEEE transactions on pattern analysis and machine intelligence, nr 36(5), s. 942-954.
73. Jović A., Brkić K., Bogunović N., 2015, *A review of feature selection methods with applications*, 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), s. 1200-1205.
74. Kane K., Lo V. S., Zheng J., 2014, *Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods*, Journal of Marketing Analytics, nr 2(4), s. 218-238.
75. Kelleher J. D., Mac Namee B., D'arcy A., 2015, *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*, MIT press.
76. Kleiner A., Talwalkar A., Sarkar P., Jordan M. I., 2014, *A scalable bootstrap for massive data*, Journal of the Royal Statistical Society: Series B: Statistical Methodology, s. 795-816.
77. Kotler A.S., Armstrong G., Harris L.C., He H., 2020, *Principles of marketing*, Pearson Education Limited.

78. Kozielski R., 2008, *Wskaźniki marketingowe*, Wolters Kluwer Polska, Warszawa.
79. Krzanowski W., Hand D., 2009, *ROC Curves for Continuous Data*, Chapman & Hall/CRC.
80. Kubus M., 2020, *Evaluation of Resampling Methods in the Class Unbalance Problem*, Econometrics nr 24(1), s. 39-50.
81. Kuhn M., Johnson K., 2013, *Applied predictive modelling*, Springer, New York.
82. Lalwani P., Mishra M. K., Chadha J. S., Sethi P, 2022, *Customer churn prediction system: a machine learning approach*, Computing, s. 1-24.
83. Landwehr N., Hall M., Frank, E., 2005, *Logistic model trees*, Machine learning, nr 59, s. 161-205.
84. Leeflang P., Wieringa J. E., Bijmolt T. H., Pauwels K. H., 2016, *Modeling markets*, Springer-Verlag, Nowy Jork.
85. Leeflang P. S., Wittink, D. R., Wedel, M., Naert, P. A., 2013, *Building models for marketing decisions*, Springer Science & Business Media.
86. Li L., Jamieson K., DeSalvo G., Rostamizadeh A., Talwalkar A., 2017, *Hyperband: A novel bandit-based approach to hyperparameter optimization*, The Journal of Machine Learning Research, nr 18(1), s. 6765-6816.
87. Lilien G. L., Kotler P., Moorthy K. S., 1995, *Marketing models*, Prentice Hall.
88. Lin A. Z., 2018, *Using Information Value, Information Gain and Gain Ratio for Detecting Two-way Interaction Effect*, SAS Global Forum.
89. Lo V. S., 2002, *The true lift model: a novel data mining approach to response modeling in database marketing*, ACM SIGKDD Explorations Newsletter, nr 4(2), s. 78-86.
90. Loyola-González O., Martínez-Trinidad J. F., Carrasco-Ochoa J. A., García-Borroto M., 2016, *Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases*, Neurocomputing 175, s. 935-947.
91. Lula P., 1999, *Jednokierunkowe sieci neuronowe w modelowaniu zjawisk ekonomicznych*, Zeszyty Naukowe/Akademia Ekonomiczna w Krakowie, Seria Specjalna, Monografie (140).
92. Łapczyński M., 2010, *Drzewa klasyfikacyjne w badaniach rynkowych i marketingowych*, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.
93. Łapczyński M., 2016, *Hybrydowe modele predykcyjne w marketingu relacji* Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.
94. Łapczyński M., 2017, *O możliwościach wykorzystania rotacyjnego lasu w badaniach rynkowych i marketingowych*, Ekonometria, nr 55, s. 69-81.
95. Łapczyński M., 2020, *Znaczenie big data w analizie danych marketingowych – obszary, strategie analityczne, perspektywy*, [w:] *Badania marketingowe w gospodarce cyfrowej* K. Mazurek-Łopacińska K., Sobocińska M. (red.), Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, s. 24-37.
96. Łapczyński M., 2021, *Dobór próby w modelach predykcyjnych data mining* [w:] *Dobór próby we współczesnych badaniach marketingowych. Podejścia ilościowe, jakościowe i mieszane*, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin.
97. Ma X., 2018, *Using classification and regression trees: A practical primer*, IAP.
98. Marsland S., 2014, *Machine learning: an algorithmic perspective*, Chapman and Hall/CRC.

99. Matthews B.W., 1975, *Comparison of the predicted and observed secondary structure of T4 phage lysozyme*, Biochimica et Biophysica Acta (BBA)-Protein Structure 405.2, s. 442-451.
100. Mattison R., 2005, *The telco churn management handbook*, XiT Press, Oakwood Hills, Illinois.
101. Mazurek-Łopacińska K., Sobocińska M., 2013, *Ewolucja ram konceptualnych marketingu a zmiany w obszarach, metodach i technikach badawczych*, Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie, nr 909.
102. McDonald M., Dunbar I., 2012, *Market Segmentation How to do it and how to profit from it* (wydanie 4), Wiley.
103. Menardi G., Torelli N., 2014, *Training and assessing classification rules with imbalanced data*, Data Mining and Knowledge Discovery, nr 28(1), s. 92-122.
104. Miguéis V. L., Van den Poel D., Camanho A. S., Falcão e Cunha J., 2012, *Predicting partial customer churn using Markov for discrimination for modeling first purchase sequences*, Advances in Data Analysis and Classification, nr 6, s. 337-353.
105. Migut G., 2004, *Analiza danych i data mining w CRM* [w:] *Analiza danych w CRM, Materiały z seminariów*, StatSoft Polska, Kraków.
106. Migut G., Jakubowski J., Stout D., 2014, *TUTORIAL Developing Scorecards Using STATISTICA Scorecard*, StatSoft Polska/StatSoft Inc., Kraków/Tulsa.
107. Migut G., 2015, *Modele predykcyjne w optymalizacji kampanii sprzedażowych, cross-sellingu i badaniu lojalności klientów – Materiały kursowe*, StatSoft Polska, Kraków.
108. Migut G., 2019, *Metodyki data mining – Materiały kursowe*, StatSoft Polska, Kraków.
109. Migut G., 2020, *Assessment of the Influence of Dependent Variable Distribution on Selected Goodness of Fit Measures Using the Example of Customer Churn Model*, Econometrics 24(1), s. 51-70.
110. Mishra A., Reddy U. S., 2017, *A novel approach for churn prediction using deep learning*, IEEE International Conference on Computational Intelligence and Computing Research (ICCIC).
111. Mitrovic S., Singh G., Baesens B., Lemahieu W., De Weerdt J., 2017, *Scalable RFM-enriched representation learning for churn prediction*, IEEE International Conference on Data Science and Advanced Analytics (DSAA), s. 79-88.
112. Mohri M., Rostamizadeh A., Talwalkar A., 2018, *Foundations of machine learning*, MIT press.
113. Morgan J. N., Sonquist J. A., 1963, *Problems in the analysis of survey data, and a proposal*, Journal of the American statistical association, nr 58(302), s. 415-434.
114. Murphy K. P., 2012, *Machine learning: a probabilistic perspective*, MIT press.
115. Mynarski S., 2003, *Analiza danych rynkowych i marketingowych z wykorzystaniem programu Statistica*, Wydawnictwo Akademii Ekonomicznej.
116. Naveen N., Ravi V., Raghavendra Rao C., 2010, *Data mining via rules extracted from GMDH: an application to predict churn in bank credit cards* [w:]

Knowledge-Based and Intelligent Information and Engineering Systems. Lecture notes in computer science, red R. Setchi i in., vol 6276, Springer–Verlag, Berlin–Heidelberg. s. 80-89.

117. Ooi M. P. L., Sok H. K., Kuang Y. C., Demidenko S., 2017, *Alternating Decision Trees*, Handbook of Neural Computation, Academic Press, s. 345-371.
118. Parvatiyar A., Sheth J.N., 2000, *The domain and conceptual foundations of relationship marketing* [w:] *Handbook of relationship marketing*, red. Sheth J. N., Parvatiyar A., Sage Publications, s. 3-38.
119. Payne A., 2005, *Handbook of CRM Achieving Excellence through Customer Management*, Butterworth-Heinemann, Burlington.
120. Phadke C., Uzunalioglu H., Mendiratta V. B., Kushnir D., Doran D., 2013, *Prediction of subscriber churn using social network analysis*, Bell Labs Technical Journal, nr 17(4), s. 63-76.
121. Powers D.M., 2011, *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*, Journal of Machine Learning Technologies, 2:1, s. 37-63.
122. Powers D.M., 2012, *The problem of Kappa*, Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, s. 345–355., Avignon.
123. Prusak B., 2005, *Nowoczesne metody prognozowania zagrożenia finansowego przedsiębiorstw*, Centrum Doradztwa i Informacji Difin, Warszawa.
124. Pyle D., 1999, *Data preparation for data mining*, Morgan Kaufmann Publishers, San Francisco.
125. Radcliffe N. J., 2007, *Using control groups to target on predicted lift: Building and assessing uplift models*, Direct Market J Direct Market Assoc Anal Council 1, s. 14-21.
126. Radcliffe N. J., Surry P. D., 2011, *Real-world uplift modelling with significance-based uplift trees*, White Paper TR-2011-1, Stochastic Solutions.
127. Rashmi K. V., Gilad-Bachrach R., 2015, *Dart: Dropouts meet multiple additive regression trees*, Artificial Intelligence and Statistics, s. 489-497.
128. Ratner B., 2017, *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data*, Chapman and Hall/CRC.
129. Reinartz W. J, Kumar, V., 2002, *The mismanagement of customer loyalty*, Harvard business review, nr 80(7), s. 86-94.
130. Reinartz W. J., Venkatesan R., 2008, *Decision models for customer relationship management (CRM)* [w:] *Handbook of marketing decision models*, Springer, Boston, s. 291-326.
131. Rodriguez J. J., Kuncheva L. I., Alonso C. J., 2006, *Rotation forest: A new classifier ensemble method*, IEEE transactions on pattern analysis and machine intelligence, nr 28(10), s. 1619-1630.
132. Rosenblatt F., 1958, *The perceptron: a probabilistic model for information storage and organization in the brain*, Psychological review 65(6).
133. Rudawska E., 2005, *Lojalność klientów, Marketing bez tajemnic*, PWE, Warszawa.
134. Ryan D., 2019, *Efficient and Flexible Hyperparameter Optimization*, PyData Miami 2019.

135. Rzepakowski P., Jaroszewicz S., 2012, *Uplift modeling in direct marketing*, *Journal of Telecommunications and Information Technology*, s. 43-50.
136. Sagan A., 2014, *Zmienne ukryte w badaniach marketingowych*, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.
137. Sagan A., 2016, *Modelowanie marketingowe a paradygmat marketingu*, Handel Wewnętrzny 364(5).
138. Samuel A. L., 1959, *Some studies in machine learning using the game of checkers*, IBM Journal of research and development 3(3), s. 210-229.
139. Siddiqi N., 2006, *Credit Risk Scorecards. Developing and Implementing Intelligent Credit Scoring*, John Wiley & Sons.
140. Siddiqi N., 2017, *Intelligent credit scoring: Building and implementing better credit risk scorecards*, John Wiley & Sons.
141. Skowron Ł., Gąsior M., 2017, *Motywacja pracownika a satysfakcja i lojalność klienta*, Difin SA.
142. Skowron S., Skowron Ł., 2012, *Lojalność klienta a rozwój organizacji*, Difin, Warszawa.
143. Sokolova M., Japkowicz N., Szpakowicz S., 2006, *Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation*, Australasian joint conference on artificial intelligence. Springer, Berlin, Heidelberg.
144. Spanoudes P., Nguyen T., 2017, *Deep learning in customer churn prediction: unsupervised feature learning on abstract company independent feature vectors*, arXiv preprint arXiv:1703.03869.
145. Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R., 2014, *Dropout: a simple way to prevent neural networks from overfitting*, The journal of machine learning research, nr 15(1), s. 1929-1958.
146. Srivastava T., 2013, *Trick to enhance power of Regression model*, Dostęp online: https://www.analyticsvidhya.com/blog/2013/10/trick-enhance-power-regression-model-2/ (marzec 2021).
147. Stanisz A., 2016, *Modele regresji logistycznej. Zastosowania w medycynie, naukach przyrodniczych i społecznych*, StatSoft Polska, Kraków.
148. Starmer J., 2019, *Gradient Boost: Part 1-4*, Dostęp online: https://statquest.org/video-index/, (marzec 2021).
149. Starmer J., 2020, *XGBoost: Part 2-3 (Classification, Mathematical Details)*, Dostęp online: https://statquest.org/video-index/, (marzec 2021).
150. StatSoft Inc., 2013, *Statistica Scorecard formula guide*.
151. Stein R. M., 2005, *The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing*, Journal of Banking & Finance nr. 29, s. 1213-1236.
152. Studzińska E., 2015, *Lojalność klienta – pojęcie, podział, rodzaje i stopnie*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, (376), s.195-215.
153. Surma J., 2009, *Business Intelligence Systemy wspomagania decyzji biznesowych*, Wydawnictwo Naukowe PWN, Warszawa.
154. Tadeusiewicz R. (przekład i opracowanie), 2001, *Wprowadzenie do sieci neuronowych*, StatSoft Polska, Kraków.
155. Tadeusiewicz R., Szaleniec M., 2015, *Leksykon sieci neuronowych*, Projekt Nauka, Fundacja na rzecz promocji nauki polskiej.

156. Terlutter R., Weinberg P., 2006, *Relationship Marketing in European Consumer Goods Markets: From Marketing Mix Orientation to Customer Life Cycle Management*. [w:] *Strategic Management — New Rules for Old Europe*, Gabler, s. 123-136.
157. Tesławski M., 2012, *Lojalność konsumenta: jak budować trwałe relacje z klientem*, Helion.
158. Tharwat A., 2018, *Classification assessment methods*, Applied Computing and Informatics.
159. TIBCO Software Inc., 2017, Statistica (data analysis software system), version 13, http://statistica.io.
160. Ting K. M., Aryal S., Washio T., 2018, *Which Outlier Detector Should I use?*, IEEE International Conference on Data Mining (ICDM).
161. Trzciński R., 2009, *Wykorzystanie techniki propensity score matching w badaniach ewaluacyjnych*, Polska Agencja Rozwoju Przedsiębiorczości.
162. Tukey J. W., 1977, *Exploratory data analysis*, Addison-Wesley, Reading, MA.
163. Urban W., Siemieniako D., 2008, *Lojalność klientów, Modele, motywacja i pomiar*, PWN, Warszawa.
164. Urbanowicz R. J., Meeker M., La Cava W., Olson R. S., Moore J. H., 2018, *Relief-based feature selection: Introduction and review*, Journal of biomedical informatics, nr 85, s. 189-203.
165. Verbeke W., Dejaeger K., Martens D., Hur J., Baesens B., 2012, *New insights into churn prediction in the telecommunication sector: A profit driven data mining approach*, European Journal of Operational Research 218(1), s. 211-229.
166. Vittinghoff E., Glidden D. V., Shiboski S. C., McCulloch C. E., 2011, *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*, Springer Science & Business Media.
167. Walesiak M., 2011, *Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R*, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław.
168. Weka, Waikato Environment for Knowledge Analysis, 2022, version 3.8.6.
169. Wierenga B., Van der Lans R., 2008, *Handbook of marketing decision models*, Springer, Nowy Jork.
170. Wübben M., 2008, *Analytical CRM: Developing and maintaining profitable customer relationships in non-contractual settings*, Gabler, Wiesbaden.
171. Youden W.J., 1950, *Index for rating diagnostic tests*, Cancer, 3.1, s. 32-35.
172. Zweig M. H., Campbell G., 1993, *Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine*, Clinical chemistry, nr 39(4), s. 561-577.

Spis rysunków

Spis tabel

Załącznik 1

Opis zmiennych wykorzystywanych w modelowaniu

Zbiór wykorzystywany w trakcie analizy zawierał 115 predyktorów ilościowych. Opis zmiennych przedstawia Tabela 38.

Tabela 38 Opis predyktorów ilościowych wykorzystywanych w analizie

Numer zmiennej	Nazwa zmiennej	Opis zmiennej
1.	ADJMOU	Rozliczenie skorygowane o łączną liczbę minut użytkowania w okresie użytkowania przez klienta
2.	ADJQTY	Billing dostosowany do łącznej liczby połączeń w okresie życia klienta
3.	ADJREV	Billing skorygowany o całkowity przychód w okresie życia klienta
4.	ATTEMPT_MEAN	Średnia liczba prób połączeń
5.	ATTEMPT_RANGE	Zakres liczby prób połączeń
6.	AVG3MOU	Średnia miesięczna liczba minut korzystania z telefonu w ciągu ostatnich trzech miesięcy
7.	AVG3QTY	Średnia miesięczna liczba połączeń w ciągu ostatnich trzech miesięcy
8.	AVG3REV	Średni miesięczny przychód w ciągu ostatnich trzech miesięcy
9.	AVG6MOU	Średnia miesięczna liczba minut użytkowania w ciągu ostatnich sześciu miesięcy
10.	AVG6QTY	Średnia miesięczna liczba połączeń w ciągu ostatnich sześciu miesięcy
11.	AVG6REV	Średni miesięczny przychód w ciągu ostatnich sześciu miesięcy
12.	AVGMOU	Średnie miesięczne minuty użytkowania w okresie życia klienta
13.	AVGQTY	Średnia miesięczna liczba połączeń w okresie życia klienta
14.	AVGREV	Średni miesięczny przychód w okresie życia klienta
15.	BLCK_DAT_MEAN	Średnia liczba zablokowanych (nieudanych) połączeń z danymi
16.	BLCK_DAT_RANGE	Zakres liczby zablokowanych (nieudanych) połączeń z danymi
17.	BLCK_VCE_MEAN	Średnia liczba zablokowanych (nieudanych) połączeń głosowych
18.	BLCK_VCE_RANGE	Zakres liczby zablokowanych (nieudanych) połączeń głosowych

Numer zmiennej	Nazwa zmiennej	Opis zmiennej
19.	CALLFWDV_MEAN	Średnia liczba połączeń z przekierowaniem
20.	CALLFWDV_RANGE	Zakres liczby połączeń z przekierowaniem
21.	CALLWAIT_MEAN	Średnia liczba połączeń oczekujących
22.	CALLWAIT_RANGE	Zakres liczby połączeń oczekujących
23.	CC_MOU_MEAN	Średnia liczba niezaokrąglonych minut korzystania z usług działu obsługi klienta
24.	CC_MOU_RANGE	Zakres niezaokrąglonych minut korzystania z połączeń z obsługą klienta
25.	CCRNDMOU_MEAN	Średnie zaokrąglone minuty korzystania z połączeń z obsługą klienta
26.	CCRNDMOU_RANGE	Zakres zaokrąglonych minut korzystania z połączeń z obsługą klienta
27.	CHANGE_MOU	Procentowa zmiana miesięcznych minut korzystania z usług w porównaniu do średniej z poprzednich trzech miesięcy
28.	CHANGE_REV	Procentowa zmiana miesięcznych przychodów w stosunku do średniej z poprzednich trzech miesięcy
29.	COMP_DAT_MEAN	Średnia liczba zakończonych połączeń z danymi
30.	COMP_DAT_RANGE	Zakres liczby zakończonych połączeń z danymi
31.	COMP_VCE_MEAN	Średnia liczba zrealizowanych połączeń głosowych
32.	COMP_VCE_RANGE	Zakres liczby zrealizowanych połączeń głosowych
33.	COMPLETE_MEAN	Średnia liczba zrealizowanych połączeń
34.	COMPLETE_RANGE	Zakres liczby zrealizowanych połączeń
35.	CUSTCARE_MEAN	Średnia liczba telefonów do obsługi klienta
36.	CUSTCARE_RANGE	Zakres liczby połączeń z obsługą klienta
37.	DA_MEAN	Średnia liczba połączeń wspomaganych przez katalogi
38.	DA_RANGE	Zakres liczby połączeń wspomaganych katalogiem
39.	DATOVR_MEAN	Średni przychód z tytułu przekroczenia limitu danych
40.	DATOVR_RANGE	Zakres przychodów z tytułu przekroczenia limitu danych
41.	DROP_BLK_MEAN	Średnia liczba odrzuconych lub zablokowanych połączeń
42.	DROP_BLK_RANGE	Zakres liczby odrzuconych lub zablokowanych połączeń
43.	DROP_DAT_MEAN	Średnia liczba odrzuconych (nieudanych) połączeń z danymi
44.	DROP_DAT_RANGE	Zakres liczby odrzuconych (nieudanych) połączeń z danymi

Numer zmiennej	Nazwa zmiennej	Opis zmiennej
45.	DROP_VCE_MEAN	Średnia liczba porzuconych (nieudanych) połączeń głosowych
46.	DROP_VCE_RANGE	Zakres liczby porzuconych (nieudanych) połączeń głosowych
47.	EQPDAYS	Liczba dni (wiek) obecnego urządzenia
48.	INONEMIN_MEAN	Średnia liczba połączeń przychodzących trwających krócej niż minutę
49.	INONEMIN_RANGE	Zakres liczby połączeń przychodzących trwających krócej niż jedną minutę
50.	IWYLIS_VCE_MEAN	Średnia liczba przychodzących połączeń głosowych z sieci bezprzewodowej do sieci bezprzewodowej
51.	IWYLIS_VCE_RANGE	Zakres liczby przychodzących połączeń głosowych między sieciami bezprzewodowymi
52.	MONTHS	Całkowita liczba miesięcy korzystania z usługi
53.	MOU_CDAT_MEAN	Średnie niezaokrąglone minuty korzystania z zakończonych połączeń transmisji danych
54.	MOU_CDAT_RANGE	Zakres niezaokrąglonych minut korzystania z zakończonych połączeń transmisji danych
55.	MOU_CVCE_MEAN	Średnie niezaokrąglone minuty korzystania z zakończonych połączeń głosowych
56.	MOU_CVCE_RANGE	Zakres niezaokrąglonych minut korzystania z zakończonych połączeń głosowych
57.	MOU_MEAN	Średnia liczba miesięcznych minut użytkowania
58.	MOU_OPKD_MEAN	Średnie niezaokrąglone minuty korzystania z połączeń danych poza godzinami szczytu
59.	MOU_OPKD_RANGE	Zakres niezaokrąglonych minut korzystania z połączeń danych poza godzinami szczytu
60.	MOU_OPKV_MEAN	Średnia liczba niezaokrąglonych minut korzystania z połączeń głosowych poza godzinami szczytu
61.	MOU_OPKV_RANGE	Zakres niezaokrąglonych minut korzystania z połączeń głosowych poza godzinami szczytu
62.	MOU_PEAD_MEAN	Średnie niezaokrąglone minuty korzystania z połączeń danych w godzinach szczytu
63.	MOU_PEAD_RANGE	Zakres niezaokrąglonych minut korzystania z połączeń danych w godzinach szczytu
64.	MOU_PEAV_MEAN	Średnie niezaokrąglone minuty korzystania z połączeń głosowych w godzinach szczytu
65.	MOU_PEAV_RANGE	Zakres niezaokrąglonych minut korzystania z połączeń głosowych w godzinach szczytu
66.	MOU_RANGE	Zakres liczby minut użytkowania
67.	MOU_RVCE_MEAN	Średnia liczba niezaokrąglonych minut korzystania z odebranych połączeń głosowych
68.	MOU_RVCE_RANGE	Zakres niezaokrąglonych minut korzystania z odebranych połączeń głosowych
69.	MOUIWYLISV_MEAN	Średnie niezaokrąglone minuty korzystania z przychodzących połączeń głosowych między sieciami bezprzewodowymi
70.	MOUIWYLISV_RANGE	Zakres niezaokrąglonych minut korzystania z przychodzących połączeń głosowych między sieciami bezprzewodowymi

Numer zmiennej	Nazwa zmiennej	Opis zmiennej
71.	MOUOWYLISV_MEAN	Średnie niezaokrąglone minuty korzystania z wychodzących połączeń głosowych między sieciami bezprzewodowymi
72.	MOUOWYLISV_RANGE	Zakres niezaokrąglonych minut korzystania z wychodzących połączeń głosowych między sieciami bezprzewodowymi
73.	OWYLIS_VCE_MEAN	Średnia liczba wychodzących połączeń głosowych między sieciami bezprzewodowymi
74.	OWYLIS_VCE_RANGE	Zakres liczby wychodzących połączeń głosowych między sieciami bezprzewodowymi
75.	OPK_DAT_MEAN	Średnia liczba połączeń z danymi poza szczytem
76.	OPK_DAT_RANGE	Zakres liczby połączeń danych poza szczytem
77.	OPK_VCE_MEAN	Średnia liczba połączeń głosowych poza szczytem
78.	OPK_VCE_RANGE	Zakres liczby połączeń głosowych poza szczytem
79.	OVRMOU_MEAN	Średni czas korzystania z usług ponadnormatywnych
80.	OVRMOU_RANGE	Zakres przekroczeń minut użytkowania
81.	OVRREV_MEAN	Średni przychód z tytułu przekroczenia limitu
82.	OVRREV_RANGE	Zakres dochodów z tytułu przekroczenia limitu
83.	PEAK_DAT_MEAN	Średnia liczba szczytowych przesyłów danych
84.	PEAK_DAT_RANGE	Zakres liczby szczytowych przesyłów danych
85.	PEAK_VCE_MEAN	Średnia liczba przychodzących i wychodzących szczytowych połączeń głosowych
86.	PEAK_VCE_RANGE	Zakres liczby przychodzących i wychodzących szczytowych połączeń głosowych
87.	PLCD_DAT_MEAN	Średnia liczba prób nawiązania połączenia z danymi
88.	PLCD_DAT_RANGE	Zakres liczby prób połączeń z danymi
89.	PLCD_VCE_MEAN	Średnia liczba wykonanych prób połączeń głosowych
90.	PLCD_VCE_RANGE	Zakres liczby prób połączeń głosowych
91.	RECV_SMS_MEAN	Średnia liczba odebranych połączeń SMS
92.	RECV_SMS_RANGE	Zakres liczby odebranych połączeń SMS
93.	RECV_VCE_MEAN	Średnia liczba odebranych połączeń głosowych
94.	RECV_VCE_RANGE	Zakres liczby odebranych połączeń głosowych
95.	RETDAYS	Liczba dni od ostatniego połączenia retencyjnego
96.	REV_MEAN	Średni miesięczny przychód (kwota opłaty)
97.	REV_RANGE	Zakres dochodów (kwota opłaty)
98.	RMCALLS	Całkowita liczba połączeń w roamingu
99.	RMMOU	Całkowita liczba minut korzystania z połączeń w roamingu

Numer zmiennej	Nazwa zmiennej	Opis zmiennej
100.	RMREV	Całkowity przychód z połączeń roamingowych
101.	ROAM_MEAN	Średnia liczba połączeń w roamingu
102.	ROAM_RANGE	Zakres liczby połączeń w roamingu
103.	THREEWAY_MEAN	Średnia liczba połączeń trójstronnych
104.	THREEWAY_RANGE	Zakres liczby połączeń trójstronnych
105.	TOTCALLS	Całkowita liczba połączeń w okresie życia klienta
106.	TOTMOU	Łączna liczba minut użytkowania w ciągu życia klienta
107.	TOTMRC_MEAN	Średnia całkowita miesięczna opłata cykliczna
108.	TOTMRC_RANGE	Zakres całkowitej miesięcznej opłaty cyklicznej
109.	TOTREV	Przychody ogółem
110.	UNAN_DAT_MEAN	Średnia liczba połączeń z danymi bez odpowiedzi
111.	UNAN_DAT_RANGE	Zakres liczby połączeń z danymi bez odpowiedzi
112.	UNAN_VCE_MEAN	Średnia liczba połączeń głosowych bez odpowiedzi
113.	UNAN_VCE_RANGE	Zakres liczby połączeń głosowych bez odpowiedzi
114.	VCEOVR_MEAN	Średni przychód z tytułu transmisji głosu
115.	VCEOVR_RANGE	Zakres przychodów z tytułu opłat za usługi głosowe

Źródło: Opracowanie własne.

W analizie wykorzystano również zestaw 56 predyktorów jakościowych lub skategoryzowanych. Tabela 39 przedstawia opis tych zmiennych. W tabeli zamieszczono dodatkowo opis zmiennej zależnej CHURN oraz opis zmiennej identyfikującej przypadki.

Tabela 39 Opis zmiennych jakościowych oraz skategoryzowanych wykorzystywanych w analizie

Numer zmiennej	Nazwa zmiennej	Opis zmiennej
1.	ACTVSUBS	Liczba aktywnych abonentów w gospodarstwie domowym
2.	ADULTS	Liczba osób dorosłych w gospodarstwie domowym
3.	AGE1	Wiek pierwszego członka gospodarstwa domowego
4.	AGE2	Wiek drugiego członka gospodarstwa domowego
5.	AREA	Obszar geograficzny
6.	ASL_FLAG	Limit wydatków na rachunku
7.	CAR_BUY	Kupujący nowy lub używany samochód
8.	CARTYPE	Dominujący rodzaj pojazdu
9.	CHILDREN	Dzieci obecne w gospodarstwie domowym
10.	CHURN	Zmiana stanu posiadania w okresie 31-60 dni od daty obserwacji
11.	CRCLSCOD	Kod klasy kredytowej
12.	CREDITCD	Wskaźnik karty kredytowej
13.	CRTCOUNT	Korekty ratingu kredytowego osoby fizycznej

Numer zmiennej	Nazwa zmiennej	Opis zmiennej
14.	CSA	Lokalny obszar usług komunikacyjnych
15.	CUSTOMER_ID	Unikalny identyfikator klienta
16.	DIV_TYPE	Kod typu działu
17.	DUALBAND	Technologia dwupasmowa
18.	DWLLSIZE	Wielkość mieszkania
19.	DWLLTYPE	Rodzaj jednostki mieszkalnej
20.	EDUC1	Wykształcenie pierwszego członka gospodarstwa domowego
21.	ETHNIC	Pochodzenie etniczne
22.	FORGNTVL	Podróże zagraniczne
23.	HND_PRICE	Aktualna cena urządzenia
24.	HHSTATIN	Wskaźnik statusu gospodarstwa domowego
25.	HND_WEBCAP	Możliwość korzystania z internetu przez telefon
26.	INCOME	Szacowany dochód
27.	INFOBASE	Wystąpienie w bazie InfoBase
28.	KID0_2	Dziecko w wieku 0 - 2 lat w gospodarstwie domowym
29.	KID3_5	Dziecko w wieku 3 - 5 lat w gospodarstwie domowym
30.	KID6_10	Dziecko w wieku 6-10 lat w gospodarstwie domowym
31.	KID11_15	Dziecko w wieku 11-15 lat w gospodarstwie domowym
32.	KID16_17	Dziecko w wieku 16-17 lat w gospodarstwie domowym
33.	LAST_SWAP	Data ostatniej wymiany telefonu
34.	LOR	Długość pobytu
35.	MAILFLAG	Flaga "nie wysyłaj poczty”
36.	MAILORDR	Nabywca wysyłkowy
37.	MAILRESP	Odbiornik poczty
38.	MARITAL	Stan cywilny
39.	MODELS	Liczba wydanych modeli telefonów
40.	MTRCYCLE	Posiadacz motocykla
41.	NEW_CELL	Nowy użytkownik telefonu komórkowego
42.	NUMBCARS	Znana liczba pojazdów
43.	OCCU1	Zawód pierwszego członka gospodarstwa domowego
44.	OWNRENT	Status właściciela domu/najemcy
45.	PCOWNER	Posiadacz komputera
46.	PHONES	Liczba wydanych aparatów telefonicznych
47.	PRE_HND_PRICE	Poprzednia cena urządzenia
48.	PRIZM_SOCIAL_ONE	Grupa społeczna
49..	PROPTYPE	Rodzaj nieruchomości
50.	REF_QTY	Całkowita liczba skierowań
51.	REFURB_NEW	Nowy telefon przy odnowieniu subskrypcji
52.	RV	Wskaźnik RV
53.	SOLFLAG	Zakazu telefonicznego marketingu
54.	TOT_ACPT	Łączne oferty przyjęte od zespołu retencyjnego
55.	TOT_RET	Całkowita liczba połączeń z zespołem ds. retencji
56.	TRUCK	Posiadacz ciężarówki
57.	UNIQSUBS	Liczba unikalnych abonentów w gospodarstwie domowym
58	WRKWOMAN	Flaga – kobieta aktywna zawodowo

Źródło: Opracowanie własne.