dr hab. Marcin Pełka, prof. UE we Wrocławiu
Uniwersytet Ekonomiczny we Wrocławiu
Wydział Ekonomii i Finansów
Katedra Ekonometrii i Informatyki
ul. Nowowiejska 3
58-500 Jelenia Góra
e-mail: marcin.pelka.ue@gmail.com

Jelenia Góra, 07.12.2023 r.

Recenzja rozprawy doktorskiej mgr Grzegorza Miguta

pt.:

„Identyfikacja optymalnej ścieżki budowy modeli data mining w obszarze retencji klientów"

napisanej pod kierunkiem naukowym dr. hab. Mariusza Łapczyńskiego, prof. UEK

1. Uwagi wprowadzające

Podstawą prawną przygotowanej recenzji rozprawy doktorskiej Pana Grzegorza Miguta są wymogi ustawowe stawiane rozprawom doktorskim określone w ustawie *Prawo o szkolnictwie wyższym i nauce* (Dz. U. z 2022 r. poz. 574 z późn. zm.). Recenzja została przygotowana na podstawie pisma Pana Prof. dr hab. Inż. Stanisława Popka, Dyrektora Szkoły Doktorskiej z dnia 30 października 2023 roku.

Za podstawę oceny rozprawy doktorskiej zgodnie ze wcześniej wskazanym aktem prawnym przyjęto następujące kryteria:

a) stopień oryginalności problemu badawczego,

b) ogólną wiedzę Kandydata w dyscyplinie nauk o zarządzaniu i jakości,

c) umiejętność samodzielnego prowadzenia pracy naukowej przez Kandydata.

Przedłożona rozprawa doktorska została przygotowana pod kierunkiem dr hab. Mariusza Łapczyńskiego, prof. UE w Krakowie. Praca licząca 272 strony, ma charakter teoretyczno-empiryczny, a układ jej treści jest przejrzysty oraz tworzy logiczną całość. Pracę tworzą: wstęp, pięć spójnych rozdziałów, zakończenie, bogata bibliografia licząca 172 pozycje (w tym 133 źródeł zagranicznych – głównie w j. angielskim, 3 źródła internetowe), spis 93 rysunków oraz spis 39 tabel, a także załącznik 1 prezentujący opis zmiennych wykorzystywanych w modelowaniu.

2. Ocena trafności sformułowania tematu, wyboru obszaru oraz przedmiotu badań

Dokonując oceny merytorycznej rozprawy doktorskiej mgr Grzegorza Miguta pt. „Identyfikacja optymalnej ścieżki budowy modeli data mining w obszarze retencji klientów" należy podkreślić, że podjęta tematyka oraz obszar badawczy, zarówno z praktycznego jak i teoretycznego punktu widzenia są ważne, aktualne i trafne, a także istotne w odniesieniu nie tylko dla krajowych, ale również zagranicznych trendów badawczych. Sformułowany w rozprawie doktorskiej tytuł wykazuje cechy dysertabilności i jednoznacznie określa rozważany problem, pozostając jednocześnie w zgodzie z treścią opracowania. Istotność podjętego w dysertacji problemu jest poparta licznymi publikacjami krajowymi i zagranicznymi, które zajmują się zwykle jedynie optymalizacją jednego lub co najwyżej kilku parametrów modelu data mining, w tym modelu data mining w kontekście problematyki retencji klientów.

3. Ocena stopnia oryginalności problemu badawczego

Wyznaczonym przez Doktoranta głównym celem pracy jest wypełnienie luki badawczej w obszarze wieloaspektowej oceny jakości modeli retencji klienta (s. 6). Dysertacja zdaniem autora ma syntetyzować wpływ wielu determinant jakości modeli na końcowy rezultat modelowania.

Dodatkowo we wstępnie znalazło się stwierdzenie, że synteza problemów występujących na wielu etapach budowy modelu retencji klientów, czy modelu analizy danych ogólnie, może przyczynić się do zidentyfikowania strategii prowadzących do otrzymania modelu o pożądanych przez badacza własnościach.

Postawiony przez Doktoranta główny problem badawczy ma charakter dysertabilny i w pełni nawiązuje do obecnych osiągnięć nauki i praktyki, co zasługuje na docenienie wkładu Autora. Wybrany problem badawczy nie jest problemem łatwym, ze względu na złożoność problemu jakim jest wybór optymalnego postępowania, ścieżki w analizie danych, mnogość różnorodnych parametrów. Niemniej jednak, zdaniem recenzenta Autor poradził sobie bardzo dobrze z odpowiedzią na tak postawiony cel badawczy. Wybór odpowiedniego sposobu postępowania w danym przypadku (problemie badawczym) nie jest zadaniem łatwym, ponieważ jest wiele ścieżek postępowania, parametrów w metodzie, których wybór ma kluczowe znaczenie dla otrzymanych wyników, ich jakości i w konsekwencji możliwości ich zastosowania. Doktorant słusznie wskazał, a następnie zilustrował problematykę wyboru dokonywanych w trakcie analizy danych ich wpływu na jakość otrzymanego modelu. O oryginalności podjętych przez Doktoranta rozważań świadczy fakt, że w literaturze przedmiotu zwykle autorzy koncentrują się na zagadnieniu doboru oraz optymalizacji ograniczonej liczby parametrów, nie często wskazując

na szerszy aspekt tego zagadnienia. Szczególnie cenne są tu rozważania poświęcone przygotowaniu danych do modelu, zagadnienie budowy optymalnego modelu klasyfikacyjnego oraz walidacja i wdrażanie modeli retencji klientów. Niewątpliwą zaletą jest zaprezentowanie rozważanych początkowo teoretycznie zagadnień na rzeczywistym zbiorze danych.

W świetle powyższych problemów badawczych należy stwierdzić, że Autor podjął się niełatwego, ale kluczowego dla każdej analizy zadania, jakim jest wybór optymalnej ścieżki analizy, począwszy od przygotowania danych, poprzez budowę modelu i dobór jego parametrów po etap walidacji i wdrażania do praktyki.

4. Ocena oryginalności celów badawczych, hipotez i metod badawczych wykorzystanych w pracy do rozwiązania problemu badawczego

Doktorant za cel główny wskazał „identyfikację determinant wpływających na jakość modeli migracji klientów oraz określenie relacji między nimi" (s. 6). W literaturze przedmiotu nie było do tej pory prac, które by próbowały holistycznie podejść do problematyki wyborów dokonywanych przy tworzeniu model migracji klientów, czy szerzej modelu analizy danych, i wpływu dokonanych wyborów na jakość modelu. Tym samym zdaniem recenzenta podjęty przez Autora problem jest oryginalny i dotychczas nie był prezentowany wyczerpująco w literaturze przedmiotu.

W pracy sformułowano także sześć celów pomocniczych, nazywanych w rozprawie pobocznymi (s. 6):

a) „Ocena wpływu wybranych technik czyszczenia i transformacji danych na jakość budowanych modeli klasyfikacyjnych.
b) 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*).
c) Ocena skuteczności modeli drzew klasyfikacyjnych budowanych za pomocą alternatywnych ścieżek podziału.
d) Porównanie skuteczności działania modelu regresji logistycznej z modelami uczenia maszynowego zbudowanymi za pomocą drzew klasyfikacyjnych, perceptronu wielowarstwowego oraz drzew wzmacnianych.
e) Ocena wpływu hybrydyzacji, segmentacji oraz agregacji na jakość budowanych modeli.

f) 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."

Sformułowany przez Doktoranta cel główny oraz cele cząstkowe (poboczne) są przejrzyste, poprawne i wymagały od Doktoranta wykazania pogłębionej wiedzy zarówno teoretycznej jak i praktycznej w dyscyplinie nauk o zarządzaniu i jakości.

Doktorant na s. 6 wskazał, że w pracy ze względu na charakter eksploracyjny hipotez badawczych nie postawiono. Zdaniem recenzenta mimo charakteru pracy, który jest eksploracyjny, Doktorant mógł pokusić się o postawienie hipotezy badawczej.

Wykorzystane przez Autora metody badawcze są adekwatne do treści pracy i sformułowanych celów badawczych. W części empirycznej Doktorant wykorzystał zbiór danych dostępny w domenie publicznej, który zawiera 100 tys. obserwacji, które są opisywane przez 173 zmienne, w tym jedna z nich to zmienna zależna, 171 to zmienne niezależne a 1 to identyfikator klienta. 55 zmiennych to zmienne nazwane przez Doktoranta jakościowymi, a 116 to zmienne nazwane przez Autora ilościowymi.

Przygotowane i przeprowadzone przez Doktoranta eksperymenty z wykorzystaniem zaproponowanego zbioru danych należy uznać za adekwatne do postawionego celu badawczego oraz celów cząstkowych. Autor z dużą dokładnością omówił wyniki przeprowadzonych badań wskazując przy każdym z eksperymentów wnioski jakie można na ich podstawie sformułować nie tylko w kontekście tego zbioru danych, ale także w szerszym kontekście szeroko rozumianej retencji klienta czy jeszcze szerzej w analizie danych. Zdaniem recenzenta dobrym posunięciem było podzielenie badań na poszczególne etapy zgodnie z aspektami THIVE oraz ogólną metodyką wynikającą z podejścia CRISP-DM.

Skrupulatność i sumienność przeprowadzonych badań empirycznych zasługuje na szczególne podkreślenie rzetelności naukowej Doktoranta.

Doktorant w podrozdziale 5.6, na podstawie przeprowadzonych badań zaprezentował w sposób syntetyczny w tabeli 37 (s. 247) płynące z nich wnioski, które dotyczą analizowanych przez Niego modeli. Wnioski te mają charakter na tyle uniwersalny, że dotyczą nie tylko problemu retencji klienta, ale także szeroko rozumianej analizy danych.

5. Ogólna wiedza teoretyczna Doktoranta w dyscyplinie nauk o zarządzaniu i jakości

Po szczegółowej analizie rozprawy doktorskiej mgr Grzegorza Miguta recenzent z pełnym przekonaniem może stwierdzić, że Autor wykazał się nie tylko wiedzą teoretyczną z obszaru

teorii marketingu i badań marketingowych, które to znajdują się w ramach dyscypliny nauk o zarządzaniu i jakości, ale także wiedzą z zakresu szeroko rozumianej analizy danych. W rozprawie zaprezentowane zostały istotne z punktu widzenia nauki zagadnienia związane z retencją klientów, etapy przygotowania danych do budowy modeli retencji, etapy i problemy występujące w trakcie budowy modelu i jego walidacji. Wskazane problemy i aspekty teoretyczne zostały przedstawione w części empirycznej pracy.

Doktorant w rozdziale pierwszym sięga do klasyków związanych z marketingiem i badaniami marketingowymi, nawiązuje do różnorodnych podejść badawczych, które często bazują na różnych, niekiedy sprzecznych założeniach. Rozważania te prowadzą do zaprezentowania szkół i tradycji badawczych, które wykształciły się w ramach badań marketingowych. W rozdziale pierwszym autor wykorzystując własne publikacje naukowe zaprezentował poziomy w ramach CRM oraz omówił obszary analityczne wspierające doskonalenie relacji z klientem. Rozdział ten kończą rozważania poświęcone lojalności klienta, począwszy od definicji tego pojęcia, determinantów kształtujących lojalność, poprzez rodzaje i poziomy lojalności, po sposoby pomiaru i modelowania z wykorzystaniem wybranych podejść, a także rozważania związane z etapami budowy modeli *data mining* na potrzeby analizy retencji klientów. W rozdziale tym Autor zaprezentował rzetelny przegląd literatury przedmiotu wskazując także na najnowsze pozycje w tym względzie.

Rozdział drugi omawia zagadnienie przygotowania danych podczas budowy modeli retencji klientów. Rozważania Doktorant rozpoczyna od wskazania kluczowych parametrów projektu analitycznego. Wykorzystując własne publikacje wskazuje kluczowe parametry i kończy je schematem przygotowania danych i stosowania modelu lojalności klientów. Kolejnym etapem jest etap związany z wiarygodnością danych w którym Doktorant koncentruje się na zagadnieniu uzupełnianiu braków w danych, które w dysertacji nazywane są imputacją. Omawia on najważniejsze metody uzupełniania danych wraz z ich wadami i ograniczeniami. Kolejne istotne zagadnienie wskazane przez Doktoranta to problem niejednorodności zbioru danych, w którym wskazuje na dwa aspekty – występowanie pojedynczych wartości nietypowych praz występowanie grup obserwacji. W rozważaniach z zakresu identyfikacji obserwacji nietypowych Autor prezentuje różne podejścia. Wskazał on różne możliwe rozwiązania w tym względzie w tym takie jak lasy separujące (*isolation forests*) oraz metody pozwalające na radzenie sobie z obserwacjami odstającymi. Podrozdział 2.4 zawiera rozważania dotyczące transformacji zmiennych oraz przygotowania zmiennych pochodnych. Wymienia on wybrane, dobrze znane metody, takie jak standaryzacja czy unitaryzacja zerowana, a następnie prezentuje podejście zaproponowane w pracy Pyle'a i alternatywę kodowania zero-jedynkowego, którą jest

propozycja zawarta w pracy Saddiqui (*weight of evidence* – WoE), a także wyniki analizy RFM do tworzenia zmiennych pochodnych. Niezbilansowany rozkład zmiennej zależnej jest zagadnieniem kończącym rozważania podjęte w tym rozdziale. Rozważania te są ciekawe i prowadzą czytelnika do dalszego etapu, którym jest budowa optymalnego modelu. Rozdział trzeci zawiera rozważania związane z budową optymalnego modelu, takie jak dobór zmiennych do modelu – Doktorant zaprezentował ciekawy i pogłębiony przegląd szeregu metod pozwalających na dobór zmiennych, metody opakowujące oraz metody które stanowią integralną część algorytmu (nazywane w pracy metodami wbudowanymi). Kolejnym istotnym zagadnieniem poruszonym przez Autora jest problematyka wyboru techniki modelowania – w pracy Autor skupił się na następujących metodach: regresji logistycznej, sieciach neuronowych, metodzie wektorów nośnych oraz drzewach decyzyjnych (klasyfikacyjnych i regresyjnych). Rozważania te Doktorant niejako podsumowuje prezentując zagadnienie podejścia wielomodelowego, bazującego w uproszczeniu na łączeniu wielu różnych wyników w jeden model połączony (zagregowany). Podejście to w dysertacji nazwane jest zespołem modeli. W tej części dysertacji zaprezentowano najważniejsze aspekty teoretyczne związane z danym rodzajem modeli.

Optymalizacja parametrów modelu jest zagadnieniem, które podsumowuje rozdział trzeci. Zawarte są tu ważne, z punktu widzenia całości pracy, metody i podejścia pozwalające ustalać jakie początkowe wartości parametrów będą dla danego zagadnienia optymalne. Jest to o tyle ważne, że problematyka doboru wielu parametrów jednocześnie jest zagadnieniem trudnym i zwykle autorzy decydują się na ograniczenie tych wyborów do jednego lub kilku parametrów jednocześnie.

W dysertacji rozdział czwarty poświęcono problematyce walidacji modelu oraz jego wdrożeniu na potrzeby retencji klientów. Zdaniem recenzenta stanowi on istotny wkład Autora w dyscyplinę nauk o zarządzaniu. Rozważania w tym rozdziale Autor rozpoczął od miar dopasowania modelu retencji, poczynając od dobrze znanych miar bazujących na macierzy błędów klasyfikacji, przez miary jakości *a posteriori* takie jak choćby czułość (SENS), specyficzność (SPEC), Kappa Cohena, krzywe ROC, indeks Giniego czy indeks J Youdena. Autor dobrze osadził prezentowane zagadnienia teoretyczne ilustrując je wynikami przeprowadzonych analiz – m.in. zaprezentował tu także wykresy porównujące wybrane miary – jak choćby SPEC i SENS, a prawdopodobieństwo otrzymania klasy pozytywnej (s. 151) czy porównanie indeksu J Youdena, Kappy Cohena, a dokładności dla zbioru niezbilansowanego (s. 156). Kolejny istotny element w tym rozdziale stanowi zagadnienie walidacji modelu retencji klientów. Podobnie jak w przypadku poprzednim, tak i w tym Doktorant przeplata umiejętnie prezentację

zagadnień teoretycznych z wynikami analiz, które umiejętnie ilustrują i objaśniają analizowane problemy.

Kolejnym istotnym wkładem Autora w dyscyplinę nauk o zarządzaniu i jakości jest rozdział piąty, prezentujący identyfikację optymalnej ścieżki selekcji modelu migracji klientów. Rozważania podjęte w tym rozdziale stanowią niejako podsumowanie doczasowych rozważań teoretycznych zawartych w dysertacji. Zdaniem recenzenta doktorant umiejętnie prowadzi czytelnika przez poszczególne etapy, objaśnia ich znaczenie oraz wskazuje na przykładzie jak podjęte decyzje wpływają na wyniki.

Podjęte przez Doktorant zagadnienie ma charakter interdyscyplinarny, ponieważ problematyka budowy modelu jako takiego ma szerokie zastosowanie, chociaż w rozprawie skoncentrowano się na problematyce retencji klientów i umiejscowieniu problematyki w dyscyplinie nauk o zarządzaniu i jakości. Sposób rozwiązania postawionego problemu stanowi istotny i niewątpliwy wkład mgr Grzegorza Miguta w rozwój dyscypliny nauk o zarządzaniu i jakości w zakresie budowy modelu. Problematyka poruszana w dysertacji może stać się impulsem do dalszego rozszerzania i skali badań tego typu nie tylko w ramach badań marketingowych.

Mając na uwadze powyższe rozważania, należy podkreślić, że mgr Grzegorz Migut wykazał się umiejętnością samodzielnego prowadzenia badań naukowych, formułowania i rozwiązywania problemów naukowych przy pomocy odpowiednio dobranych metod i narzędzi badawczych. Mgr Grzegorz Migut udowodnił wszechstronny zakres wiedzy w obszarze badań marketingowych, problemów występujących w trakcie budowy modelu. Zakres wiedzy teoretycznej oraz dobrze dobrany przegląd krajowej i światowej literatury dowodzą posiadania przez Doktoranta szerokiej wiedzy w dyscyplinie nauk o zarządzaniu i jakości.

6. Strona techniczno-edytorska oraz formalna pracy

Pod względem formalnym praca zasługuje na dobrą ocenę, chociaż strona techniczno-edytorska ma pewne niedociągnięcia takie jak:

a) w rozprawie występują licznie błędy w składzie tekstu polegające na pozostawieniu na końcu wiersza krótkich słów, zwłaszcza jednoliterowych – np. s. 5, 6, 7, 8, 9, 10, 12, 15 oraz wiele innych,

b) używanie w rozprawie pojęcia imputacja danych, podczas gdy w polskojęzycznej literaturze przedmiotu dobrze ugruntowane jest pojęcie uzupełniania braków w danych. Zdaniem recenzenta jest to zbytnia kalka z j. angielskiego, gdzie funkcjonuje pojęcie *data imputation techniques*,

c) jakość rysunków 2, 5, 53, 57, 80 i 81 mogłaby być nieco lepsza,

d) na s. 34 pojawia się fragment „Błąd! Nieprawidłowy odsyłacz do zakładki: wskazuje na nią samą.”. Jest to zapewne wynikiem stosowania automatycznych numeracji tabel i wykresów w dysertacji,

e) użycie terminu zespoły modeli zamiast terminu podejście wielomodelowe, który funkcjonuje w zakresie metod analizy danych (por. np. Gatnar E., Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji, PWN, Warszawa 2008),

f) Doktorant nie wprowadził numeracji we wzorach. Zdaniem recenzenta w opracowaniach naukowych należałoby taką numerację stosować. Pozwala to m.in. odwoływać się do wzorów umieszczonych w różnych miejscach pracy czy artykułu,

g) w niektórych wzorach nawiasy wydają się być nie elementem równania (obiektem), ale elementem wpisanym z klawiatury, ponieważ nie obejmują całości równania – zob. np. s. 81, 83, 97, 101, 127.

h) s. 67, akapit 2, brakująca spacja w „W pracy D.Pyle [1999]…”,

i) w pracy brakuje wskazania równania prezentującego sposób obliczania indeksu J Youdena,

j) s. 134, akapit 2 – brakuje orzeczenia (lub części zdania) w zdaniu „Praktycznie każda metoda analityczna ma hiperparametry, których wartości na końcową postać modelu”.

k) recenzent nie znalazł w dysertacji odwołania wskazującego źródła pochodzenia danych. Doktorant ogranicza się do stwierdzenia, że podstawą analizy był zbiór danych dostępny w domenie publicznej. Niemniej jednak, recenzentowi udało się ustalić, że zbiór ten pochodził z Duke University i był on częścią *Churn Modeling Tournament*. Zbiór znajdował się pod adresem http://faculty.fuqua.duke.edu/teradatacenter/ (link obecnie jest już nieaktywny). Zdaniem recenzenta w każdej pracy o charakterze naukowym konieczne jest wskazanie źródeł danych,

l) w tabeli 37 (s. 247) brakuje nagłówków kolumn, co zdaniem recenzenta znacznie ogranicza możliwości jej analizowania,

m) zdaniem recenzenta etap normalizacji danych obejmujący klasyczne podejścia mógłby obejmować także inne dobrze znane metody – poza standaryzacją czy unitaryzacją zerowaną.

Powyższe niedociągnięcia nie ujmują jednakże walorów recenzowanej dysertacji i nie zaniżają pozytywnej oceny merytorycznej rozprawy doktorskiej. Do najważniejszych walorów rozprawy zdaniem recenzenta należy zaliczyć:

a) ważność i aktualność podjętego tematu w tym próba wypełnienia luki badawczej w wymiarze wieloaspektowej oceny jakości modeli retencji klientów,

b) przejrzysty charakter rozprawy doktorskiej oraz logika wywodów,

c) połączenie rozważań teoretycznych z wynikami empirycznymi, które pełniej ilustrują omawiane problemy,

d) właściwy dobór metod badawczych do części empirycznej,

e) umiejętność samodzielnego formułowania wniosków.

Recenzent nie doszukał się w dysertacji zbyt wielu obszarów polemicznych. Jednakże oczekuje od Doktoranta odpowiedzi podczas publicznej obrony na następujące zagadnienia:

1) Dlaczego w rozprawie pominięto zagadnienie wpływu zmiennych zakłócających (*noisy variables*) na model i możliwości jego zastosowania?

2) Jeżeli badacz chciałby podjąć próbę przeanalizowania wszystkich możliwych ścieżek budowy modelu to jaki czas (nawet przybliżony) jest potrzebny na takie zadanie?

3) Czy zdaniem autora podejście wielomodelowe jest podejściem optymalnym, które można by było stosować niezależnie od okoliczności?

7. Konkluzja

Zdaniem recenzenta rozprawa doktorska mgr Grzegorza Miguta jest pracą wartościową, oryginalną oraz świadczącą o znacznym zasobie wiedzy Doktoranta w zakresie dziedzin nauk społecznych – dyscyplinie nauk o zarządzaniu i jakości. Osiągnięte przez Autora rezultaty badawcze wynikają ze znacznego doświadczenia badawczego oraz poziomu erudycji w zakresie krajowej i zagranicznej literatury naukowej w podjętej dyscyplinie badań.

W świetle poczynionych ustaleń stwierdzam, że recenzowana rozprawa doktorska spełnia wymagania stawiane w ustawie. Stanowi ona oryginalne rozwiązanie przez Doktoranta problemu naukowego w sposób samodzielny i poprawny pod względem metodologicznym.

Niniejszym wnoszę o przyjęcie i dopuszczenie recenzowanej rozprawy doktorskiej do publicznej obrony.

Marcin Pełka