ABSTRAKT

mgr Grzegorz Migut

tytuł rozprawy: „Identyfikacja optymalnej ścieżki budowy modeli data mining w obszarze retencji klientów”

Praca dotyczy wieloaspektowej oceny jakości klasyfikacyjnych modeli data mining. Ma na celu identyfikację determinant wpływających na jakość modeli oraz określenie relacji między nimi. Merytoryczny obszar powyższych dociekań został ograniczony do modeli lojalności klienta. Poza głównym celem pracy do celów pobocznych zaliczyć można: ocenę wpływu wybranych technik transformacji danych na jakość budowanych modeli klasyfikacyjnych, identyfikację optymalnej ścieżki budowy modelu ekonometrycznego, na przykładzie regresji logistycznej, ocenę skuteczności modeli drzew klasyfikacyjnych budowanych za pomocą alternatywnych ścieżek podziału, porównanie skuteczności działania modelu ekonometrycznego z modelami uczenia maszynowego, ocenę wpływu hybrydyzacji, segmentacji oraz agregacji na jakość budowanych modeli, porównanie modeli interpretowalnych przez badacza z zaawansowanymi metodami nieinterpretowalnymi.

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; ***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.

Praca składa się z pięciu rozdziałów. Pierwszy rozdział wprowadza w zagadnienie retencji klientów jako obszaru modelowania marketingowego. Rozdział drugi podejmuje tematykę przygotowania danych na potrzeby budowy modeli retencji klientów. Rozdział trzeci przedstawia zagadnienia związane z budową optymalnego modelu klasyfikacyjnego. W rozdziale czwartym omawiane są kwestie odnoszące się do walidacji i stosowania modeli retencji klientów. Spośród aspektów przedstawionych w rozdziałach teoretycznych na dodatkową uwagę może zasługiwać część związana z miarami oceny siły predykcyjnej modelu obejmująca dodatkowe badania oceny wrażliwości miar na zmiany proporcji klas zmiennej zależnej.

Rozdział piąty przedstawia wyniki pracy badawczej związanej z identyfikacją optymalnej ścieżki selekcji modelu migracji klientów. Na podstawie przeprowadzonych badań można stwierdzić, że czynniki wpływające na jakość modeli retencji klientów działają na nie inaczej w przekroju różnych metod. Aspekt transformacji zmiennych okazał się znaczący jedynie w przypadku regresji logistycznej i miał duży wpływ na jakość modelu. Dodanie zmiennych pochodnych było jedynym czynnikiem wpływającym pozytywnie na wyniki wszystkich analizowanych metod. Podobny efekt 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. 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.