Dr hab. Jerzy Surma, prof. SGH
Instytut Informatyki i Gospodarki Cyfrowej
Szkołą Główna Handlowa w Warszawie

Warszawa, 08 stycznia 2024

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

1. Podstawa opracowania recenzji

Podstawą opracowania recenzji jest pismo z dnia 30 października 2023 Dyrektora Szkoły Doktorskiej Uniwersytetu Ekonomicznego w Krakowie prof. dr hab. inż. Stanisława Popeka w sprawie powierzenia mi obowiązków recenzenta rozprawy doktorskiej mgr Grzegorza Miguta.

2. Ocena wyboru tematyki badawczej, celu i hipotez badawczych

Autor podjął w rozprawie próbę opisania problematyki budowy modeli retencji klientów z wykorzystaniem metod eksploracji danych. Jest to niezwykle ważne zagadnienie zarówno w zakresie naukowym jaki i zastosowań praktycznych w biznesie. W tym kontekście wybór tematyki rozprawy uważam za bardzo dobry.

We Wstępie rozprawy zdefiniowano jej cel główny jako identyfikację determinant wpływających na jakość modeli migracji klientów oraz określono 6 celi dodatkowych. Te zadania zrealizowano poprzez budowę modelu retencji w wykorzystaniem rzeczywistych danych empirycznych udostępnionych przez nie określone z nazwy przedsiębiorstwo. Zastosowana procedura badań polegała na wykonaniu szereg badań symulacyjnych, w ramach których każdy z modeli był budowany z godnie z metodyką CRISP-DM. Jako punkt odniesienia przyjęto model zbudowany z wykorzystaniem regresji liniowej, który był porównywany z modelami zrealizowanymi z wykorzystaniem algorytmów drzew decyzyjnym (w tym drzew wzmacnianych zgodnie z podejściem XGBoost, co jest bardzo dobrym wyborem) oraz sieci neuronowych (wielowarstwowy perceptron).

Rozprawa doktorska miała na celu znalezienia metody rozwiązania problemu naukowego (wynik metodyczny). Dobór celów badawczych jest poprawny. Na szczególne wyróżnienie zasługuje cel numer:

- 2 dotyczący m.in. poprawnego doboru technik selekcji zmiennych.
- 4 dotyczący analizy porównawczej wybranych metod budowy modelu retencji klientów.
- 6 tj. wykazanie możliwości budowy modeli o zadowalającej jakości za pomocą metod, w których wynikowy model może podlegać interpretacji, czyli objaśnianiu swoich decyzji. Jest to ważne zagadnienie badawcze, które obecnie implikuje szereg badań naukowych związanych z próba wydobywania interpretowalnej wiedzy z modeli typu black box. Wykazanie, że możliwe jest budowanie akceptowalnych jakościowo modeli retencji z wykorzystanie modeli interpretowalnych byłoby istotnym osiągnięciem badawczym.

3. Ocena poprawności struktury rozprawy oraz jej zawartości merytorycznej

Recenzowana rozprawa doktorska liczy 251 stron, a łącznie z bibliografią oraz stosownymi spisami i załącznikami 272 strony. W pracy omówiono kolejne etapy procesu budowy modelu retencji: formalizację zmiennych zależnych i niezależnych, zebranie i przygotowanie zbiorów danych, budowa modelu, oraz proces walidacji i testowania modelu. Struktura pracy jest poprawna. Na szczególne uznanie zasługuje zwrócenie uwagi i omówienie:

- Systemów klasy CRM i w tym kontekście szczegółowe omówienie modeli CLV (Customer Livetime Value (CLV) oraz modelu ACURA w Rozdziale 1.
- Problemu niejednorodnych zbiorów danych oraz niezbilansowanych rozkładów zamiennej zależnej w Rozdziale 2.
- Zagadnienia doboru zmiennych i strategii wybory hiperparametrów w Rozdziale 3. W tym rozdziale autor podkreśla też istotny problem zanikającego gradientu w sieciach neuronowych wielowarstwowych.
- Strategii walidacji oraz metod punktów odcięcia cut-off w Rozdziale 4.
- Identyfikacja czynników wpływających na jakość modeli w Rozdziale 5.

W ramach rozprawy autor dokonał następujących oryginalnych odkryć zasługujących na podkreślenie:

- Brak wpływu transformacji zmiennych na jakość modelu za wyjątkiem modeli opartych na regresji liniowej.
- Istotność zmiennych pochodnych.
- Ważność doboru hiperparametrów co jest zgodne z badaniami naukowymi w tym zakresie.
- Pozytywny (na modele intepretowane) i negatywny (na modele nieintepretowane) wpływ segmentacji zbioru danych.
- Pozytywny wpływ agregacji na jakość modeli z wyjątkiem regresji liniowej.
- Identyfikacja adekwatnej metody selekcji zmiennych dla modeli regresji logistycznej, która wykorzystuje podejście LASSO w połączeniu z techniką Branch & Bound.

4. **Uwagi merytoryczne do treści rozprawy**

Uwagi krytyczne i polemiczne:

- We Wstępie określono jeden z celów jako wypełnienie luki badawczej w obszarze jakości modeli retencji klientów. Aby określić lukę badawczą należy dogłębnie określić aktualny stan wiedzy w danym obszarze, zidentyfikować problemy i w tym kontekście zaproponować własne rozwiązanie. To zadanie nie zostało zrealizowane z należytą dociekliwością naukową.

- Autor używa w swojej dysertacji pojęcia retencji (np. w tytule), migracji (np. str. 52) oraz churn (np. str. 249). Nie jest klarowne czy autor traktuje te pojęcia jako synonimy, czy też w zależności od kontekstu te pojęcia mają różne interpretacje. Ten brak precyzji nomenklaturowej bardzo utrudnia zrozumienie pracy.

- Istotny kontekst Sztucznej Inteligencji (SI) (Rozdział 1.6) został opisany bardzo powierzchownie i kontrowersyjnie. Autor twierdzi, że SI wywodzi czy też jest inspirowana statystyką matematyczną, co wskazuje na brak dogłębnej znajomości tej tematyki. Podobna powierzchowność dotyczy tematyki powstania Deep Learning. Niezrozumiałe jest, że w tak kluczowe pojęcie w dysertacji jak Data Mining nie jest omawiane i definiowane w oparciu o pionierskie artykuły takich kluczowych autorów jak: Usama Fayyad, Gregory Piatetsky-Shapiro oraz Padhraic Smyth.

- Fundamentalnym zagadnieniem w omawianej pracy jest definicja, sposób formalizacji oraz dyskusja na temat zmiennej zależnej w modelach retencji klientów, czyli opisana na stronie 55 zmienna binarna: lojalny/nielojalny klient. Całe to kluczowe zagadnienie jest omówione pobieżnie w jednym paragrafie w odwołaniu do własnej publikacji. Brakuje tutaj omówienia jak historycznie w literaturze naukowej jest ta zmienna byłą definiowana i jak ja poprawnie formalizować w praktycznych zastosowaniach, w kontekście analizy wad i zalet różnych sposobów formalizacji.

- Rozdział 3 dotyczy budowy optymalnego modelu klasyfikacji i według autora model jest optymalny, jeżeli dobór zmiennych, hiperparametrów, itp. będzie optymalny. Jest to bardzo nieformalne rozumienie pojęcia optymalność (tego zagadnienia dotyczą też moje uwagi w następnym punkcie) zakładające, że docelowy model będzie optymalny jeżeli w trakcie jego budowy i wyboru najlepszego modelu wykonywano quasi optymalne działania. Jest to o tyle rażące, że ramach w teorii rozpoznawania obrazów jest jednoznacznie zdefiniowany klasyfikator optymalny i znana jest jego postać, jeśli tylko posiadamy (co w praktyce jest wątpliwe) rozkłady prawdopodobieństw warunkowych wystąpienia obiektu pod warunkiem danej klasy.

- Tytuł pracy odwołuje się do pojęcia „identyfikacji optymalnej ścieżki budowy modeli" i jest to tematyka Rozdziału 5. Autor nie zdefiniował pojęcia „optymalnej ścieżki" i używa pojęcia optymalizacja w sposób nieformalny. Formalnie zadanie optymalizacji wymaga sformułowania ilościowej funkcji celu i jej maksymalizację/minimalizację przy określonych ograniczeniach. Używanie tego pojęcia jest zatem niepotrzebnym nadużyciem, nadającym pozorny charakter większej „naukowości" rozprawy.

- Celem dodatkowym numer 6 pracy jest mówienie możliwości budowy modeli objaśnianych (tzw. white vox), które byłyby konkurencyjne do modeli trudnych do

interpretacji typu black box. W tym kontekście (oraz także czynników określających jakość modeli) zadziwiająca jest powierzchowna i skrótowa analiza. Ta kluczowa część rozprawy nie jest opisana w osobnym podrozdziale tylko w ramach rozdziału 5.6 o sieciach neuronowych! I tak:

- Na stronie 247 w tabeli 37, której wiersz z nazwami kolumn jest nieczytelny, przedstawiono na zasadzie binarnej (+/-) czynniki wpływające na jakość modeli.
- Na stronie 248 przedstawiono analizę porównawczą modeli (rysunek 93), która jest ograniczono tylko do krzywych ROC (brak klasycznych wskaźników takich jak accuracy, recall, itp.) i nie wykazano czy te różnice są faktycznie statystycznie istotne, co byłoby przesłanką do wyciągania wiarygodnych konkluzji.

Faktem jest, że to zagadnienie jest także podsumowane w Zakończeniu, niemniej brak jednoznacznego odwoływania się do celów pracy utrudnia ocenę pracy badawczej autora.

- Cel dodatkowy numer 6 jest też dobry przykładem braku dojrzałej metodologii badawczej. Autor na stronie 250 stwierdza, że „badanie nie wykazało zatem możliwości budowy modeli za pomocą metod interpretowalnych przez badacza, porównywalnych z zaawanasowanymi metodami nieinterpretowalnymi". Faktem jest, że ten wynik jest zgodny z aktualnymi badaniami naukowymi w tym zakresie. Niemniej taki wniosek jest wyciągnięty na podstawie jednego badania i faktycznie jest prawdziwy dla tego konkretnego badania. Niemniej praca naukowa ma zwykle na celu podjęcie próby generalizacji, czy zatem to badanie w zakresie dostępnego (jednego) zbioru danych i użytych metod daje podstawy do jakiś wniosków ogólniejszych na temat jakimi metodami budować modele retencji klientów? Niestety autor nie podejmuje krytycznej dyskusji na ten temat ponad stwierdzenie o „dołożeniu wszelkich starań, aby wybrać zbiór danych o złożoności jak najbardziej zbliżonej do rzeczywistych zbiorów danych, w którym występowałyby cechy statystyczne typowe dla tego typu zjawisk".

- W pracy brakuje odwołania ważnych trendów rozwojowych i problemów naukowych związanych z rozwojem metod maszynowego uczenia i bezpośrednio związanych z tematyką rozprawy tj.
  - Alligment problem – czyli budowa modelu zgodnego z oczekiwaniem projektanta i w tym kontekście np. nieoczekiwanych zachowań systemu w przypadku użycia złożonych funkcji kosztu.
  - MLOps – czyli pełnowymiarowe zarządzanie cyklem życia modeli eksploracji danych i w tym kontekście na przykład problem aktualizacji modeli, który jest całkowicie pominięty w rozprawie.
  - Bias–variance tradeoff – czyli problem nadmiernego dopasowania, kluczowy w kontekście wspomnianej w rozprawie generalizacji modelu (strona 5)

5. Syntetyczna ocena rozprawy i konkluzja końcowa

Dokładnie tak jak jest wspomniane w Zakończeniu „proces budowy modelu retencji klientów jest zadaniem wieloetapowym i złożonym". To trudne wyzwanie zostało poprawnie

przedstawione w recenzowanej rozprawie, przedstawiono adekwatną metodykę i zaproponowano oryginalne autorskie rozwiązania. Jest to istotne nie tylko w kontekście wartości naukowej, ale też, co należy podkreślić, stanowi olbrzymią wartość dla praktyki zarządzania relacjami z klientami w realnym biznesie, zwłaszcza w takich branżach jak bankowość, ubezpieczenia i telekomunikacja. Podając zatem ocenie całość pracy uważam, że Doktorant:

1. wykazał się zadawalającą ogólną wiedzą teoretyczną w dyscyplinie nauki o zarządzaniu i jakości,
2. udowodnił, że posiada akceptowalny poziom umiejętność samodzielnego prowadzenia pracy naukowo-badawczej.
3. przedstawił w swojej rozprawie elementy oryginalnego rozwiązania w zakresie wykorzystania własnych badań naukowych.

Całościowa ocena recenzowanej rozprawy doktorskiej jest pozytywna. Recenzowana praca doktorska dotyczy ważnego problemu o istotnym znaczeniu poznawczym. Doktorant wykazał się dobrą znajomością badanej problematyki. Dysertacja jest, jak już wspomniałem, szczególnie istotna z praktycznego punktu widzenia.

Reasumując stwierdzam jednoznacznie, że pomimo zgłoszonych zastrzeżeń i uwag, rozprawa doktorska mgr Grzegorza Miguta spełnia ustawowe wymogi stawiane pracom doktorskim. Konkludując wnoszę o przyjęcie recenzowanej rozprawy doktorskiej dopuszczenie jej do publicznej obrony.

Jerzy Surma

Podpisano przez /Signed by:
JERZY SURMA
Data /Date: 2024.01.08 10:05

mSzafir