wpł. RDITI-28.08.2023
M. Jkom

Dr hab. inż. Przemysław Głomb
Instytut Informatyki Teoretycznej i Stosowanej PAN

Gliwice, 28 sierpnia 2023 r.

Recenzja rozprawy doktorskiej mgr inż. Joanny Badury pt. „Solutions for selected problems of demand forecasting based on machine learning methods and domain knowledge"

1 Wstęp

Rozprawa doktorska mgr inż. Joanny Badury dotyczy zastosowania metod uczenia maszynowego, w szczególności algorytmów typu regresyjnego, w zagadnieniach związanych z prognozowaniem popytu. W ujęciu przedstawionym w rozprawie, zagadnienie to dotyczy przede wszystkim predykcji wartości liczbowych związanych z konkretnymi sprzedawanymi produktami, takich jak wielkość sprzedaży, obrotu, czy liczba zwrotów. W tym obszarze tematycznym występuje szereg problemów powiązanych ze sobą od strony biznesowej, ale istotnie różniących się z perspektywy konstrukcji algorytmów – prognoza może być wymagana dla nowego produktu, bez danych historycznych sprzedaży, albo dla produktu znanego, ale sprzedawanego w nowym miejscu. Jest to znany obszar problemowy, obecny od początku rozwoju metod uczenia maszynowego – a w zasadzie od początku informatyki – ale wciąż aktualny. Dotyka on bezpośrednio wielu obszarów gospodarki, i z perspektywy ekonomicznej jest bardzo istotny; „być albo nie być" wielu gałęzi przemysłu.

Tezą pracy jest[1] jest, że „rozbicie zadania prognozowania konkretne zagadnienia oraz wykorzystanie wiedzy dziedzinowej pozwala, dzięki zastosowaniu metod uczenia maszynowego, uzyskać przydatne wartości prognoz". Wymienione w tezie wykorzystanie wiedzy dziedzinowej jest doprecyzowane jako[2] „wprowadzenie nowych atrybutów lub rekordów z podobnych problemów do modelu, co powoduje poprawę wyników prognozowania". Uzupełnione jest to definicją dwóch celów pracy: wprowadzenia taksonomii problemów prognozowania popytu a następnie zaproponowanie metod uczenia maszynowego do ich rozwiązania. Ten dosyć obszerny, szeroko zakrojony plan prac został zrealizowany ze zdecydowanym naciskiem na analizę i propozycję metod dla czterech wybranych problemów. Uzyskiwaniu dobrych prognoz w wybranych sytuacjach poświęcona jest znacząca większość pracy, w tym szereg różnorodnych praktycznych i ciekawych eksperymentów.

[1] W oryginale: „By breaking down the forecasting task into specific issues and taking into account domain knowledge, useful forecast values can be obtained using machine learning methods."

[2] W oryginale: „The domain knowledge can be included by introducing new attributes or new data records from similar problems into the model, resulting in an improvement in the final forecast results."

2 Zakres pracy

Przedstawiona rozprawa jest napisana w języku angielskim, składa się z siedmiu rozdziałów, liczy 208 stron razem z bibliografią.

Rozdział pierwszy jest wstępem i zawiera: krótkie wprowadzenie do problemu, tezę, cele oraz opis pozostałych rozdziałów rozprawy.

Rozdział drugi rozpoczyna się omówieniem zagadnień problemowych związanych z prognozowaniem popytu. Nadrzędny ich podział dotyczy okresu czasu prognozy – krótko-, średnio- i długoterminowe prognozowanie popytu – w ramach których zostały wymienione i opisane poszczególne zadania prognozowania. Przykładowe zadania to: prognozowanie popytu sprzedawanego wcześniej produktu w nowym miejscu sprzedaży; prognozowanie dla grupy produktów i w drugim kroku uściślanie jej dla produktów rzadkich lub o niewielkiej sprzedaży; prognozowanie nowych produktów na podstawie danych historycznych podobnych towarów. Z przedstawionego wyliczenia wyłania się obraz różnorodnych, praktycznych problemów, w których dane historyczne mogą pełnić różne role w zależności od celu analizy. W szczególności, problemy te znacznie przekraczają klasyczne sformułowanie problemu predykcji. Wymienionych jest 15 problemów, każdy przedstawiony w postaci krótkiego, dosyć ogólnego opisu słownego. Tak zaprezentowana „taksonomia" – a właściwie podział systematyczny – jest niestety niekompletny; brakuje m.in. powiązania z literaturą, uzasadnienia, powiązania ze sobą zadań podobnych na poziomie opisu słownego, a także dyskusji wiążącej go z wyborem problemów poruszanych dalej w pracy. Jest to źródło mojej pierwszej głównej uwagi do pracy (3.1.1). W dalszej części rozdziału znajduje się ogólne wprowadzenie do prognozowania szeregów czasowych – m.in. standardową dekompozycję na trend, sezonowość i cykle – a także przedstawienie kilku wybranych, popularnych algortymów – ARIMA, Prophet, XGBoost, LSTM.

Rozdział trzeci, najobszerniejszy i w mojej opinii jeden z dwóch najciekawszych w pracy, dotyczy prognozy skuteczności promocji. Skuteczność tą określa sześć wskaźników parametrycznych, zaproponowanych przez autorkę rozprawy – dotyczą one zarówno promowanego produktu (np. średnia liczba sprzedanych jednostek lub kg na dzień) ale też całego procesu zakupowego (np. średnia liczba produktów zakupionych lub wartość koszyka zawierającego promowany produkt). W badaniach posłużono się zbiorem danych sprzedaży trzech grup produktów w ramach sieci 500 sklepów, w okresie trzech lat. Elementem przygotowawczym dla zbioru danych było opracowanie wektora cech oraz identyfikacja odpowiadających sobie par okresów bez i z promocją. Następnie wykonano eksperyment szacujący skuteczność na tak przygotowanym zbiorze algorytmu XGBoost oraz wybranych architektur sieci neuronowych z automatycznie optymalizowaną architekturą; algorytm XGBoost uzyskał niewielką przewagę. Następnie przeprowadzono kolejny eksperyment, tym razem dla wybranych pojedynczych produktów oraz innego zestawu algorytmów; tu najlepsze wyniki osiągnął algorytm drzew losowych (ang. „Random Forest") oraz sieci neuronowe (MLP). Ten zestaw eksperymentów

miał na celu przeanalizowanie zbioru danych pod kątem przygotowania modeli prognostycznych, pozwala też uzyskać informacje na temat charakterystyki zbioru danych. Szkoda, że w pracy nie poświęcono więcej uwagi spójności konstrukcji obu eksperymentów (uwaga 3.1.2) i pogłębionej analizie ich rezultatów, co pozwoliłoby m.in. na odniesienie wyników do oczekiwań biznesowych. Natomiast istotnym ich wynikiem było przygotowanie modeli, które umożliwiły dalsze prace nad komponentami symulacji wyniku promocji.

W dalszej części rozdziału trzeciego, modele prognostyczne są wykorzystane do budowy trzech komponentów programowych pozwalających na zastosowanie uzyskanych wyników badawczych w warunkach operacyjnych. Pierwszym jest symulator wpływu promocji na sprzedaż i zysk. Po zadaniu parametrów, odpowiadających cechom wykorzystywanym w poprzednim eksperymencie, symulator wyświetla wykresy prognozy zysku i sprzedaży w zależności od promocyjnej zmiany ceny. Drugim jest generator promocji, który po zadaniu parametrów (produkt, zakresy dat itp) pozwala zaproponować zoptymalizowane parametry promocji (datę rozpoczęcia, cenę itp). Dla tego komponentu został opracowany specjalny algorytm, który wykorzystuje m.in. hierarchicznie przeszukiwanie według atrybutów (m.in. dla sklepów, dla dat, dla opcji reklamy), szukając najlepszej kombinacji ich wartości. Trzecim komponentem jest rozszerzony symulator prognozy zysku i sprzedaży, uzupełniony o pozostałe wskaźniki parametryczne określające skuteczność promocji. Zaprezentowane komponenty stanowią bardzo dobry przykład praktycznego zastosowania wyników wcześniej przeprowadzonych eksperymentów i wyznaczonych modeli uczenia maszynowego. Autorka jednak nie poprzestaje na tym; mając możliwość prognozowania wyników i dostarczając do tego narzędzia, skupia się następnie ma możliwości monitorowania i reakcji na parametry trwającej aktualnie promocji w kolejnej części rozdziału.

Najciekawszą, a także najbardziej rozbudowaną częścią rozdziału 3 jest mechanizm aktywnej rekomendacji działań podczas trwającej promocji, kiedy jej parametry spadają poniżej określonych wartości granicznych. W tym celu autorka posługuje się metodą „reguł działań dla przeżycia" (ang. „survival action rules"). Za pomocą oryginalnego algorytmu wyznaczane są reguły, wiążące wartości atrybutów w celu uzyskania poprawy parametrów wydajnościowych. Dzięki połączeniu algorytmu uczącego – indukcji reguł – oraz semantycznego znaczenia poszczególnych atrybutów uzyskujemy nie tylko rekomendacje działania – np. zwiększenie liczby promocji dla utrzymania wysokiej wartości koszyka produktów – ale też uzyskujemy wgląd w dane, wyjaśniający część wzorców za nimi stojących. Ta część pracy jest w mojej ocenie szczególnie wartościowa; zaproponowany algorytm jest ciekawy od strony teoretycznej, a jednocześnie stanowi ukoronowanie pracy badawczej zmierzającej do praktycznego wykorzystania możliwości uczenia maszynowego dla problemu zarządzania promocjami. Całość rozdziału trzeciego uznaję za godny naśladowania przykład wnikliwości i konsekwencji prowadzenia badań przy zetknięciu się z trudnym praktycznym problemem.

Rozdział czwarty zajmuje się problemem prognozowania sprzedaży produktu wewnątrz grupy (np. ubrania konkretnego koloru i rozmiaru w ramach kolekcji). Ten problem nazwany jest „top down", i sformalizowany jako szacowanie wartości ułamka sprzedaży całej grupy przypadającego na konkretny produkt. Jako dane posłużył zbiór historii sprzedaży ubrań, dostarczony przez partnera biznesowego wraz z prognozami odniesienia (ang. „baseline"). Jako pierwsza badana jest „data cube approach", której centralnym elementem jest tensor indeksowany atrybutami rozróżniającymi produkt w grupie, którego elementy to udziały cząstkowe w grupowej sprzedaży. W oparciu o tą stukturę danych badano algorytmy predykcji tzw. „naiwnej" (poprzedniej wartości); kNN, dla którego wprowadzono dodatkowy algorytm preselekcji przykładów do zbioru treningowego; a także model zmieszania liniowego (ang. „linear mixing model", LMM) i XGBoost. Ten rozdział jest znacznie krótszy w stosunku do poprzedniego i poprzestaje na porównaniu wyników predyktorów, dodatkowo zweryfikowanym testami statystycznymi. Pozostawia to pewien niedosyt; wprawdzie dwie z metod – LMM i kNN – dały lepszy wynik niż prognozy odniesienia, to jednak jak wynika z rysunku 4.5 poprawa nie jest istotna statystycznie, a w rozdziale brakuje dyskusji szczegółowej wyników i odniesienia ich do wyjściowego problemu.

Rozdział piąty zajmuje się problemem predykcji szeregów czasowych wykorzystując dodatkowe serie danych. Jest to druga, obok rozdziału 3, obszerna i ciekawa część pracy, cechująca się pogłębioną i szczegółową analizą problemu. Rozdział składa się z dwóch części, opisujących dwa oddzielne eksperymenty. Pierwszy z nich dotyczy możliwości poprawy predykcji, przy uwzględnieniu danych z innych lokalizacji – np. predykcja sprzedaży na stacji paliw uwzględniając dane z innych stacji. Aby porównać predykcję „pojedynczą" i „grupową", dobrano wiele metod, różniących się zarówno algorytmami predykcji, jak i sposobem agregacji różnych serii danych. Dobór metod jest obszerny – jest ich aż 13 – i zasługuje na podkreślenie zarówno ze względu na różnorodność, jak i czytelność przedstawienia; zdecydowanie takie podejście pozwala wyczerpująco zbadać problem. Dodatkowym elementem prac było opracowanie cech dla metod dostosowanych do tabelarycznych danych (XGBoost). Wynik dla dwóch zbiorów danych – sprzedaży paliwa na stacji oraz zużycia gazu LPG z przydomowych zbiorników – wyraźnie pokazuje przewagę metod wykorzystujących grupowanie lokalizacji, i bezpośrednio odnosi się do tezy pracy.

W drugiej części rozdziału piątego przedstawiony jest scenariusz predykcji również dla wielu serii danych, ale tym razem dodatkowe serie dotyczą składowych konsumpcji, której sumaryczna wartość jest szacowana w podstawowym problemie. Tym razem zbiór dotyczy zużycia prądu elektrycznego, gdzie warunki zastosowania określiły nieco inne parametry skuteczności – błąd MAPE predykcji poniżej 25%. Porównanie metod ponownie wskazało, że metody łączące dane z wielu źródeł uzyskują lepsze rezultaty niż te działające na pojedynczej serii danych. Dla tego scenariusza, w końcowej części rozdziału znajduje się bardzo ciekawa kontynuacja. Ze względu na określony cel

– MAPE < 25% – można było określić predykcję jako poprawną lub nie, a następnie przeprowadzić analizę wyjaśniającą, skąd biorą się błędy. Wykorzystano do tego drzewa decyzyjne, które wskazują jakie konkretne atrybuty – cząstkowe punkty pomiaru zużycia prądu – mogą być odpowiedzialne za błąd. Dodatkowo, ze względu na to że monitorowanie każdego punktu zawiera pewien koszt, przeprowadzone zostało studium optymalizacji – redukcji punktów monitoringu w taki sposób, aby utrzymać błąd w określonych parametrach celu. Podsumowując, ten rozdział w elegancki sposób traktuje problem predykcji w warunkach posiadania wielu serii danych, odnosząc się zarówno do aspektu badawczego, w tym tezy, jak i do praktycznych zagadnień wdrożenia metod uczenia maszynowego.

Rozdział szósty związany jest z prognozowaniem zwrotów zakupionych towarów. Celem było przygotowanie klasyfikatora, który dla danej transakcji zakupowej określi, czy była ona przedmiotem zwrotu czy nie. Jest to zagadnienie biznesowo istotne, co w sposób zwarty i skuteczny jest przedstawione we wstępie do rozdziału. Jako dane posłużyły rzeczywiste dane sprzedaży i zwrotów otrzymane w trakcie projektu wdrożeniowego. Brak w danych informacji łączących transakcję zakupu i zwrotu wymusił przygotowanie przez autorkę algorytmu dopasowania, wykorzystującą wiedzę dziedzinową i analizę problemu. Zaproponowany algorytm jest ciekawy i obiecujący, chociaż na tym etapie nie była możliwa weryfikacja jego skuteczności, i dzięki niemu możliwe było zrealizowanie analizy skuteczności prognozowania zwrotów. Okazało się że problem jest trudny; najwyższy uzyskany wynik zrównoważonej skuteczności (ang. „balanced accuracy") to 66% – w klasyfikacji dwuklasowej! – co pokazuje, jakim wyzywaniem może być próba zastosowania metod uczenia maszynowego dla takich problemów. W eksperymencie porównano trzy metody (regresja logistyczna, XGBoost oraz losowe drzewa) oraz uzupełniono zbiór cech o ręcznie przygotowane cechy. Pewnym mankamentem rozdziału jest niestandardowa i niespójna z poprzednimi rozdziałami prezentacja wyników, m.in. wprowadzone do dyskusji są klasyfikatory bez optymalizacji hiperparametrów i wyniki na zbiorze treningowym, które nie wprowadzają istotnych informacji, a jednocześnie pominięto zastosowanie testów statystycznych, obecnych w poprzednich rozdziałach.

Rozdział siódmy zawiera podsumowanie wyników osiągniętych w poszczególnych rozdziałach, wyliczenie i podsumowanie autorskiego wkładu, a także krótkie przedstawienie możliwych dalszych kierunków prac.

3 Ocena i uwagi do pracy

Najważniejszym rezultatem badawczym w mojej ocenie jest zaproponowanie, przeanalizowanie i zweryfikowanie propozycji algorytmicznych rozwiązania czterech problemów związanych z nadrzędnym zagadnieniem prognozowania popytu. Autorka wykazała się przy tym znajomością algorytmów, wiedzą z zakresu projektowania i realiza-

cji eksperymentów, w tym analizy i przygotowania zbiorów danych.

Mocną stroną pracy są wyniki przedstawione w rozdziałach trzecim i piątym. Stanowią one przykład dogłębnej analizy, idącej poza standardowy eksperyment uczenia maszynowego, rozwijając dalej rezultaty badań w stronę stworzenia na ich podstawie narzędzi stosowanych w praktyce, wyjaśnienie wartości wychodzących z algorytmów, ujawnienia wzorców obecnych w danych – przejścia z informacji do wiedzy. Pozytywnie należy ocenić również sukces konfrontacji ze złożonym praktycznym problemem, jakim jest prognozowanie popytu, a także wkład pracy włożony zaprojektowanie i wykonanie tak wielu eksperymentów, opracowanie algorytmów, przygotowanie danych. Istotne znaczenie ma, że prace, których wyniki prezentowane są w recenzowanej rozprawie, zostały zrealizowane w ramach dofinansowanych projektów wdrożeniowych – oznacza to że istnieje konkretne zapotrzebowanie na te wyniki, a równocześnie pokazuje dobry przykład synergii między pracą badawczą, nastawioną na osiąganie nowej wiedzy i pracą wdrożeniową, nastawioną na praktyczne rozwiązania dla konkretnych odbiorców.

Słabą stroną pracy są, przede wszystkim: skrótowe potraktowanie pierwszego (z dwóch!) celu pracy i związany z tym brak dyskusji wyników w kontekście literatury oraz brak spójności i konsekwencji w przygotowaniu kolejnych eksperymentów i analizie ich wyników. Oznacza to, że o ile pojedyncze ścieżki analizy w poszczególnych rozdziałach uznaję za poprawne – a niektóre oceniam bardzo dobrze, patrz rozdziały trzeci i piąty – to bardzo trudno odnieść je do siebie, a przede wszystkim umieścić w kontekście wyników innych autorów, prezentowanych w literaturze. Te dwa punkty są rozwinięte poniżej, jako uwagi główne. Poza tym, niestety, praca nie została należycie zredagowana; tekst jest często niejasny, co w wielu momentach utrudnia zrozumienie, a w skrajnych wypadkach wymusza odgadywanie treści. Związane z tym najważniejsze uwagi szczegółowe są zaprezentowane w dalszej części recenzji.

3.1 Uwagi główne

3.1.1 Skrótowe potraktowanie pierwszego celu pracy, brak powiązanej literatury

Jako pierwszy cel pracy zostało zadeklarowane wprowadzenie podziału systematycznego zagadnień lub zadań prognozowania popytu. Równolegle, teza pracy rozpoczyna się od „rozbicie zadania prognozowania konkretne zagadnienia [...]" (ang. „By breaking down the forecasting task into specific issues [...]"). O ile wyliczenie przedstawione w rozdziale 2.1 pracy można potratować jako wspomniane w tezie „rozbicie", to jednak w zaprezentowanej postaci ma ono szereg mankamentów:

1. Wprowadzony podział nie odnosi sie w ogóle do literatury – brak referencji do innych prac, porównania z nimi; trudno uwierzyć, żeby inne podziały nie były obecne w literaturze?

2. Wprowadzony podział nie posiada argumentacji ani uzasadnienia – dlaczego taki a nie inny podział?

3. Wprowadzony podział zawiera jedynie opis słowny, przy użyciu dosyć ogólnych sformułowań; brakuje sformalizowanego (np. matematycznego) zapisu, który umożliwiłby odniesienie do siebie poszczególnych zadań – np. „demand forecasting for a product with a history" dla długoterminowej i średnioterminowej prognozy są z opisu bardzo podobne, co je łączy a co dzieli?

4. Brakuje powiązania wprowadzonego podziału ze szczegółowymi problemami poruszanymi w kolejnych rozdziałach pracy – czy wybrane problemy pokrywają w sumie całość zagadnień – co sugeruje teza – czy są jedynie podzbiorem – na co wygląda po analizie poszczególnych rozdziałów?

W mojej opinii takie przedstawienie jest niewystarczające, i przekłada się na pobieżne potraktowanie połowy z celów pracy – przez co niestety tytuł znacznie trafniej oddaje jej zakres niż teza.

Dodatkowym problemem tak sformułowanej tezy jest trudność w odniesieniu się do „przydatnych wartości prognoz" (ang. useful forecast values); o ile wyniki przedstawione w pracy są przydatne – w mojej opinii zdecydowanie – to jednak ta przydatność nie jest w pracy dyskutowana. A powinna być koniecznie, jako że jest to element tezy. Jest to o tyle istotne, że pozostałe rozdziały pracy zawierają ciekawe, oryginalne i wartościowe wyniki, które w ten sposób są niewyeksponowane. Być może lepszym pomysłem było pozostawienie samej drugiej części tezy – dotyczącej wiedzy domenowej (ang. „domain knowledge"), która jest wyczerpująco oparta o rozdziały trzeci i piąty – i potraktowanie rozdziału 2, po uzupełnieniu literatury, jako części wstępu.

Brak odniesień do literatury widoczny też jest w braku dyskusji wyników rozdziałów, które ograniczone są właściwie wyłącznie do prezentacji i analizy statystycznej. Żaden z rozdziałów nie ma klasycznej części „dyskusji" – w sensie struktury IMRaD, „introduction–method–results–discussion"; o ile rozdziały trzeci i piąty bronią się w jakimś stopniu wspomnianą wcześniej kilkustopniową analizą wyników, to w rozdziałach czwartym i szóstym – i oczywiście drugim – jest to bardzo widoczne.

3.1.2 Brak spójności i konsekwencji w przygotowywaniu kolejnych eksperymentów i analizie ich wyników

Teza i cele pracy określiły szeroko zestaw problemów, będących tematem rozprawy. W tych warunkach jest zrozumiałe, że kolejne rozdziały oparte są o różniące się od siebie zbiory danych, mają również osobne definicje eksperymentów, analizę wyników, i dyskusję. Jednak w wielu wypadkach jest możliwe, i ma dużą wartość dodaną, uwspólnienie części narzędzi badawczych, co ułatwia zestawienie ze sobą wynków poszczególnych eksperymentów. Jest to tym ważniejsze, jeżeli poruszane problemy – a tak jest

w tej pracy – mają reprezentować jedną dziedzinę. Niestety, w prezentowanej rozprawie nie tylko nie widać próby wprowadzenia elementów spójności i konsekwencji, ale wręcz kolejne rozdziały redefiniują założenia analizy, często bez podania uzasadnienia. Poniżej przedstawione zostało to w wybranych aspektach budowy eksperymentów.

Dobór metod W rozdziale drugim, w ramach wstępu, jest wprowadzenie do algorytmów: XGBoost, ARIMA, LSTM. Trzeci rodział zaczyna się od wyników dla XGBoost i wybranych architektur sieci neuronowych (nie LSTM). W kolejnym eksperymencie trzeciego rozdziału są już inne metody: drzewa losowe, „gradient boosting trees", kNN, modele liniowe GLM i sieci neuronowe (MLP, podobne, ale nie identyczne z poprzednią wersją eksperymentu). W rozdziale 4 jest metoda „data cube", kNN, LMM oraz XGBoost. W piątym – pierwszy zestaw (średnia krocząca, ARIMA, LSTM, XGBoost, kNN) też różni się od drugiego (predykcja wartości poprzedniej, regresja liniowa, LSTM, Prophet); to jest jedyny (!) rozdział, w którym ta zmiana jest uzasadniana i dyskutowana. W szóstym – regresja logistyczna, drzewa losowe, XGBoost.

Miary wydajności W rozdziale trzecim wprowadzone są miary jakości: MAE, RMSE, MAPE oraz WMAPE. Jednak ta definicja jest tylko na potrzeby pierwszej części rozdziału; kolejne wyniki są dla RMSE i MAE, a jeszcze dalej już tylko dla RMSE – za to pojawia korelacja jako miara wydajności. W rozdziale czwartym z kolei wykorzystywana jest tylko miara MAE, za to dodatkowo pojawia się ocena rangowa (ang. „rank") jako uzupełnienie. Z kolei w rozdziale piątym – najpierw RMSE, MAEi WMAPE, i dodatkowo osobno definiowany błąd względny; a w drugiej części – tylko MAPE.
(W rozdziale 6 zmiana miar – zrównoważona skuteczność i AUC – jest uzasadniona, ponieważ w odróżnieniu do poprzednich rozdziałów jest to problem typu klasyfikacyjnego, nie regresyjnego).

Analiza statystyczna W rozdziale trzecim istotność statystyczna jest oceniana przez test Wilcoxona – tylko w drugim eksperymencie (!). W rozdziale czwartym pojawia się test Friedmana z analizą post-hoc w postaci diagramów CD. W rozdziale piątym analiza zaczyna się od testu Friedmana z diagramami CD, ale potem ponownie wracamy do testu Wilcoxona – ale tylko w pierwszej części, bo w drugiej części rozdziału analizy istotności nie ma; podobnie jak brakuje jej w rozdziale szóstym.

Konstrukcja eksperymentu W rozdziale 3 stosowany jest klasycznych schemat uczenia maszynowego: definicja cech – wspólnych dla metod – wybór metod, optymalizacja hiperparametrów. W czwartym rozdziale metody mają różniące się zestawy cech. W stosunkowo prostym eksperymencie rozdziału szóstego pojawią się klasyfikatory bez optymalizacji hiperparametrów, zmienne definicje cech dla różnych metod, wyniki na

zbiorach treningowych. (Rozdział piąty pomijam, odrębność zastosowanego tam schematu jest uzasadniona i dyskutowana)

Ten labirynt podejść i metod, stosowany w kolejnych rozdziałach, znacznie utrudnia analizę wyników i wspólne ich zestawienie. Moim zdaniem, w sytuacji kiedy zmieniają się zbiory danych – szczególnie że są one niedostępne dla czytelnika – i tworzone są nowe algorytmy, jak najwięcej ze standardowych procedur (optymalizacja hiperparametrów, walidacja krzyżowa, testy statystyczne, miary wydajności) powinny być stosowane w sposób konsekwentny i spójny. W przeciwnym wypadku, i tak jest niestety w tej pracy, trudna jest zarówno analiza i ocena rezultatów, jak i odniesienie do wyników innych autorów.

3.2 Uwagi szczegółowe

1. Rozdział 3 – Wskaźniki parametryczne w tabelach 3.2 i 3.3 mają inne nazwy niż wcześniej zdefiniowane w rozdziale, czy nowe nazwy odpowiadają poprzednim? Do jakich modeli odnosi się tabela 3.2?

2. Rozdział 3, rysunek 3.2 – jaka jest motywacja zastosowania zaprezentowanego tam skomplikowanego algorytmu optymalizacji hiperparametrów XGBoost, zamiast klasycznego przeszukiwania kraty parametrów (ang. „grid search"), zresztą wspomnianego w tekście?

3. Rozdział 3 – Tabela 3.4 podaje wartości hiperparametrów dla sieci neuronowych; z niej wynika że są tam trzy warstwy. Jeżeli tak, to dlaczego to podejście jest określone jako „Deep Learning", skoro zwyczajowo przyjmuje się, że modele głębokie charakteryzują się znacznie większą liczbą warstw?

4. Rozdział 3 – jak mają się się pojedyncze produkty 101–117 wykorzystane w drugim eksperymencie do trzech grup wykorzystanych w pierwszym?

5. Rozdział 3, str. 52 – uzyskanie w teście p-wartości powyżej krytycznej ($0.109 > 0.05$) oznacza, że nie możemy powiedzieć że serie danych są takie same (nie możemy potwierdzić H0) – skąd dalszy wniosek że nie ma istotnych różnic?

6. Rozdział 3, str. 52 – dlaczego przy wielokrotnych porównaniach nie zastosowano poprawki Bonferroniego?

7. Rozdział 3, str. 52 – dlaczego w opisie, odwołującym się do wykresu 3.7, podana jest wartość korelacji 0.62, skoro na wykresie wyraźnie widoczne są wartości wokoło i powyżej 0.8?

8. Rozdział 3, str. 73 – dlaczego na „ocenzurowanych" danych (and. „censored observations") jest niemożliwe zastosowanie klasycznych metod regresji? O jakie metody regresji chodzi??

9. Rozdział 3, str. 73 – cały fragment opisujący estymator Kaplan-Meir jest napisany skrótowo, i trudny do zrozumienia; wymaga obszerniejszego wytłumaczenia i powiązania z resztą pracy. Co oznacza r_j, „number of observations at risk" we wzorze 3.1? O jakie ryzyko chodzi? Jaki zbiór danych jest na wykresie 3.25? Do czego stosuje się ten estymator w kontekście pracy? Co oznacza zdanie „The log rank test is a χ^2 test"? Jakie, w kontekście pracy, ma znaczenie że krzywe KM się różnią?

10. Rozdział 3, str. 75 – akapit „The premise of an action rule [...] target range of these values" jest bardzo trudny do zrozumienia, wymaga obszerniejszego wyjaśnienia.

11. (*) Rozdział 3, sekcja 3.5.2 – zaprezentowany algorytm jest złożony i trudny do zrozumienia z przytoczonego opisu, zdecydowanie wymaga przykładu przed przejściem do jego zastosowania w kolejnej sekcji.

12. Rozdział 4, str. 94 – czy „data cube approach" jest oryginalną propozycją doktorantki? Tekst sugeruje zastosowanie metody, ale brak cytowania.

13. Rozdział 4, sekcja 4.2.1 – w jaki sposób jest liczona prognoza dla grupy? Przedstawiony algorytm wyznacza tylko prognozę ułamka dla produktu w ramach grupy.

14. Rozdział 4, str. 96 – w jaki sposób w algorytmie „data cube" unikamy przetrenowania, skoro de facto występuje tam wykorzystanie tylko zbioru treningowego, bez mechanizmów regularyzacji lub walidacji które przed przetrenowaniem zabezpieczają?

15. Rozdział 4, str. 101 – co oznacza „we tried to optimize the range of conditional attributes", wobec definicji atrybutów na str. 98? Które z nich są „conditional"?

16. Rozdział 4, str. 100/101 – dla jakiego zakresu parametru k były wykonane eksperymenty? Obok siebie w tekście podane są $k = 3$, $k \in \{1, 2, 3\}$ oraz $k \in \{1, \ldots, 8\}$.

17. Rozdział 4, str. 101 – jaki był wynik selekcji cech zastosowanej dla algorytmu kNN? Jakie cechy zostały usunięte i jakie wartości miary błędów były uzyskane w tym procesie?

18. Rozdział 4, str. 101 – jakie jest wytłumaczenie „the permutation test was used to prove the significance of random effects"? O jaki test chodzi, jakie statystyki (p-wartości?) zostały uzyskane, jak zdefiniowane są „random effects"?

19. Rozdział 5, sekcja 5.1 – jakie jest znaczenie słowa „dataset" – czy chodzi o całość zbioru danych, czy dane dla pojedynczej lokalizacji? (Słowo pojawia się zamiennie w obu kontekstach)

20. Rozdział 5, wykres 5.19 – powinny być oznaczone wyniki poszczególnych metod.

21. Rozdział 6, sekcja 6.2 – dlaczego w ostatnim kroku algorytmu dopasowania zwrotów towaru do zakupów występuje pasowanie losowe, zamiast pominięcia tych rekordów? Jak można szacować skuteczność algorytmu?

22. Rozdział 6, str. 168 – jaka była struktura zbioru danych, skoro na etapie dopasowania zwrotów (sekcja 6.2) było dostępnych prawie 192 tys. zwrotów, a na etapie ich prognozowania tylko 7315?

23. Rozdział 6, str. 172 – dlaczego dla klasyfikatorów dopasowywany jest parametr „random state", czyli de facto ziarno generatora liczb losowych, służące w zasadzie do uzyskiwania powtarzalnych wyników eksperymentów?

24. Rozdział 6, sekcja 6.3.3 – dla wykorzystanych klasyfikatorów, jakie konkretnie szacowane parametry były użyte jako „cut off" do wyznaczania później krzywych ROC? Jak wyglądały wyznaczone krzywe?

25. Rozdział 6, sekcja 6.3 – dlaczego jako jeden z wariantów badane są klasyfikatory bez dopasowania hiperparametrów, skoro ich dopasowanie jest standardową procedurą wykonywaną praktycznie w każdej sytuacji treningu algorytmów uczenia maszynowego? Dlaczego akurat w tym przypadku podawane są wyniki dla zbioru treningowego?

26. Rozdział 6, sekcja 6.4 – jak należy rozumieć propozycję zastąpienia miary AUC do WHM jako „optimization metrics", skoro AUC w poprzednim rozdziale była użyta do oceny wyników, a nie do optymalizacji?

27. Rozdział 6, sekcja 6.3 – dlaczego w eksperymentach nie zastosowano schematu walidacji krzyżowej, skoro wiadomo że daje bardziej wiarygodne wyniki niż jednokrotny podział danych na część treningową, walidacyjną i testową?

28. Rozdział 6 – dlaczego zamiast klasyfikować transakcje binarnie – podlegające zwrotom lub nie – nie zdecydowano się na analizę prawdopodobieństwa zwrotu?

4 Wniosek końcowy

Przeprowadzone przez mgr inż. Joanny Badurę badania stanowią istotny i wartościowy wkład w zakresie zastosowania uczenia maszynowego dla wybranych problemów zagadnienia prognozowania popytu; istotnego i praktycznego problemu badawczego. Rozprawa prezentuje oryginalne rozwiązanie tego problemu. Autorka wykazała się wiedzą z zakresu algorytmów uczenia maszynowego, planowania i realizacji eksperymentów, a także analizy ich wyników i propozycji kolejnych etapów przetwarzania zorientowanych na zastosowanie; prezentując tym samym zarówno ogólną wiedzę teoretyczną jak i umiejętność samodzielnego prowadzenia pracy naukowej. Teza pracy została wykazana; pokazane zostało, że zarówno dekompozycja problemu jak i wykorzystanie wiedzy dziedzinowej prowadzi do przydatnych i wartościowych wyników;otrzymane wyniki mają duże znaczenie praktyczne. Mimo wymienionych wyżej uwag krytycznych oceniam, że cel rozprawy został osiągnięty a wyniki zostały zaprezentowane zadowalająco.

Na podstawie powyższej recenzji stwierdzam, że w mojej ocenie **rozprawa doktorska mgr inż. Joanny Badury pt. „Solutions for selected problems of demand forecasting based on machine learning methods and domain knowledge" spełnia wymogi stawiane pracom doktorskim przez obowiązującą aktualnie w Polsce Ustawę o stopniach i tytule naukowym oraz o stopniach i tytule w zakresie sztuki, i wnioskuję o dopuszczenie jej do publicznej obrony**.