

Dr hab. Wiesław Szczesny, prof. nadzw. SGGW
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
Wydział Zastosowań Informatyki i Matematyki
Katedra Informatyki
ul. Nowoursynowska 159, 02-776 Warszawa

Recenzja rozprawy doktorskiej

mgr Damiana Przekopa

pt. Algorytmiczna budowa predyktorów w modelowaniu zdarzeń rzadkich na przykładzie wykrywania oszustw kredytowych

I. Ogólna charakterystyka rozprawy

Przedstawiona do oceny praca doktorska mgr Damiana Przekopa pt. *Algorytmiczna budowa predyktorów w modelowaniu zdarzeń rzadkich na przykładzie wykrywania oszustw kredytowych* została przygotowana pod kierunkiem prof. dr hab. Marka Gruszczyńskiego jako promotora oraz dr Marcina Owczarczuka jako promotora pomocniczego.

Recenzowana rozprawa podejmuje bardzo ważny i aktualny temat dotyczący detekcji obserwacji nietypowych. Zagadnienie to zostało zaprezentowane w pracy na przykładzie wykrywania oszustw kredytowych.

Praca liczy 128 stron, składa się z wprowadzenia, pięciu rozdziałów, wniosków oraz spisu literatury. We wprowadzeniu przedstawiono dość pobieżnie i zbyt krótko wkład pracy do dorobku nauki (ok. pół strony). W pozostałej części liczącego prawie 5 stron wprowadzenia zaprezentowano strukturę pracy. W pierwszym rozdziale opisano różnego typu ryzyka związane z funkcjonowaniem banku, natomiast w drugim rozdziale, na podstawie analizy literatury przedmiotu, zaprezentowano różne metody analityczne wykrywania nadużyć w tym sektorze. Trzeci rozdział zawiera prezentację obszarów, w których przeprowadzane są działania mające na celu wykrywanie nadużyć. Rozdział czwarty i piąty stanowią empiryczną część pracy. W pierwszym z nich przedstawiono własną propozycję zestawu procedur służących wykrywaniu rzadko występujących zdarzeń, w drugim zaś zaprezentowano wyniki badań empirycznych wykorzystujących autorską metodę oraz inne, znane i powszechnie używane metody detekcji obserwacji nietypowych. W końcowej części pracy, nazwanej Wnioskami, zawarto – zapewne dla wygody czytelnika - ponownie informacje o celu i strukturze pracy. W części tej zamieszczono także konkluzje dotyczące potwierdzenia

postawionych w pracy hipotez badawczych, ocenę wkładu do rozwoju wiedzy i sugestie na temat kierunków przyszłych badań.

II. Ocena merytoryczna rozprawy

Rozprawa dotyczy problematyki ważnej z punktu widzenia funkcjonowania współczesnego sektora bankowego. Po zapoznaniu się z treścią pracy stwierdzam, że przedstawiona do recenzji praca doktorska mgr Damiana Przekopa niewątpliwie stanowi wartościowe opracowanie. Jest to poprawne i kompletne (w ramach przyjętych założeń oraz udostępnionych do badania danych z banku detalicznego) studium dotyczące budowy narzędzi przeznaczonych do wsparcia działań w zakresie ograniczenia zjawiska nadużyć w procesie udzielania kredytów gotówkowych. Nie ulega wątpliwości, iż w sektorze finansowym problematyka dotycząca wykrywania nadużyć jest aktualnie jednym z ważniejszych wyzwań, jakie stoją przed kierownictwem departamentów ryzyka w poszczególnych instytucjach finansowych. Ten obszar ryzyka w coraz większym stopniu znajduje odzwierciedlenie w wynikach poszczególnych instytucji finansowych, a co za tym idzie staje się także widoczny w wynikach całego sektora. Dlatego obecnie ważnym zagadnieniem jest poszukiwanie technik wspomagających wykrywanie tego typu działań. Odrębnym zagadnieniem jest sprawdzenie skuteczności proponowanych rozwiązań w oparciu o rzeczywisty materiał empiryczny wybranej instytucji. Przedstawiona przez autora propozycja wpisuje się w ten nurt, w którym jest jeszcze wiele miejsca na badania. Warto dodatkowo podkreślić, iż literatura dotycząca problematyki ryzyka w sektorze finansowym jest bardzo bogata, ale niezbyt łatwo jest dotrzeć do opracowań w sposób przystępny koncentrujących się na tej części ryzyka, które dotyczy obszaru nadużyć w poszczególnych autonomicznych obszarach sektora finansowego. Przedstawiona do recenzji praca – według mojej oceny – pomaga także w tym zadaniu. Doktorant umiejętnie wykorzystuje istniejące w literaturze metody teoretyczne przenosząc je na grunt prowadzonych analiz empirycznych. Rozprawa została napisana – poza niezbyt licznymi wyjątkami – poprawnie językowo z uwzględnieniem zasad przygotowania prac naukowych.

Opiniowana rozprawa zyskuje moją pozytywną ocenę. Praca jest przykładem autorskiego wartościowego wykorzystania metod i narzędzi informatycznych i analitycznych w obszarze ich zastosowania. Postawiony w rozprawie problem naukowy definiuje cel o dużych walorach praktycznych, ale równocześnie posiada aktualne znaczenie naukowe. W literaturze z zakresu modelowania matematycznego bardzo często proponowane są rozwiązania, których właściwości nie pozwalają na wdrożenie ze względu na specyfikę

danych. Dlatego utarło się przekonanie, że lepsze są proste metody, które nie wykorzystują nowoczesnych osiągnięć naukowych. Niniejsza praca, stanowi próbę pokazania, że możliwa jest adaptacja zaawansowanych metod matematycznych w warunkach rzeczywistych. Warto zwrócić uwagę na rozdźwięk pomiędzy szybko rozwijającymi się koncepcjami matematycznymi, a ich absorpcją przez osoby pracujące w tzw. praktyce gospodarczej. Stąd praca dotycząca efektywnej metody klasyfikacji rzadkich zdarzeń jest niezmiernie cenna. Odpowiada ona również stopniem trudności problemom jakie formułowane są jako tezy rozpraw doktorskich z zakresu zastosowań. Narzędziami stosowanymi w rozprawie są złożone algorytmy numeryczne oparte na teoriach z zakresu Data Mining. Ich zastosowanie do problemu będącego przedmiotem rozprawy wymagały od Doktoranta wykorzystania nowoczesnych technologii komputerowych.

Reasumując, **za główne osiągnięcia Doktoranta** uważam:

1. Syntetyczne opracowanie opisu problemu oszustw kredytowych w sektorze bankowym.
2. Uporządkowanie rozproszonego w literaturze krajowej i zagranicznej dorobku publikacyjnego na temat analitycznych metod wykrywania nadużyć finansowych.
3. Zaproponowanie własnego zestawu procedur określania predyktorów mających zastosowanie w przypadku zdarzeń rzadkich na przykładzie wykrywania oszustw kredytowych.
4. Wykonanie badań empirycznych na podstawie rzeczywistych danych jednego z polskich banków mających na celu:
 - porównanie efektywności różnych algorytmów detekcji obserwacji nietypowych na przykładzie procesu udzielania kredytów gotówkowych przez bank detaliczny oraz
 - zbadanie stabilności zaproponowanych przez Autora procedur ze względu na kilka aspektów dotyczących wykorzystanej próby danych empirycznych oraz przyjętej definicji funkcji celu.

Mimo pozytywnej całościowej oceny, praca zawiera wiele niedociągnięć, których usunięcie znacząco poprawiłoby jakość dysertacji.

Uwagi do pracy

1. Pierwsze trzy akapity Streszczenia i Wprowadzenia są identyczne. O ile w streszczeniu powinna znaleźć się adnotacja na temat wkładu pracy doktorskiej do dorobku nauki, to zamieszczenie tego typu informacji w drugim akapicie Wprowadzenia może powodować pewną konsternację. We Wprowadzeniu zawarto

ponadto inne sformułowania budzące podobne odczucia czytelnika, np. *Stanowi to o unikalnym wymiarze tej dysertacji* (str. 12).

2. Zwyczajowo przyjmuje się, że Wprowadzenie powinno zawierać informacje o celu (celach) pracy oraz hipotezach. W recenzowanej pracy w tej części pracy przedstawiono jedynie mglistą i mało precyzyjną informację: *Celem pracy jest sformułowanie i przedstawienie nowego podejścia do budowy modeli predykcyjnych, wykorzystywanych w modelowaniu zdarzeń rzadkich, które są charakterystyczne dla zjawisk nietypowych* (pierwsze zdanie pracy!, str. 8). Cel ten został dokładniej określony dopiero w podrozdziale 4.2 (str. 62-63). Treść hipotez (str. 9-10) została – według mojej oceny - niepotrzebnie powtórzona na str. 68 i wyodrębniona w specjalnym podrozdziale 4.5, którego zawartość liczy mniej niż jedną stronę.
3. O ile struktura pracy w zasadzie nie budzi poważnych zastrzeżeń, to w treści pracy brakuje w pewnym sensie konsekwencji w prowadzeniu wywodu naukowego. W szczególności:
 - 3.1. Autor często wyprzedza czytelnika podając wyniki opisane w dalszych podrozdziałach, np.
 - na str. 70 w podrozdziale 4.7 napisano: *zaproponowane w podrozdziale 2.2 algorytmy okażą się w podrozdziale 5.2 być niedopasowane do rozpatrywanego problemu;*
 - na str. 83 w punkcie 5.2.1 prezentując wyniki uzyskane na podstawie algorytmu INFLO stwierdzono: *Algorytm INFLO nie jest w stanie konkurować z regresją logistyczną (przedstawioną w sekcji 5.3.2, dla której ocena Lift kształtuje się na poziomie powyżej 3);* podobnie na str. 89 w punkcie 5.2.3 napisano: *Algorytm LOF nie jest w stanie konkurować z regresją logistyczną (przedstawioną w sekcji 5.3.2, dla której ocena Lift kształtuje się na poziomie powyżej 3);*
 - miara LIFT jest wykorzystywana do oceny jakości algorytmów wykrywania obserwacji nietypowych (str. 81-90, podrozdział 5.2), jednak dokładniejszy jej opis przedstawiono dopiero na str. 94-95 (podrozdział 5.3).
 - 3.2. Opis prowadzony jest niejednolicie, np.:
 - w trzeciej osobie liczby pojedynczej, np. *autor dokonuje próby osadzenia tematu w literaturze* (str. 9), w pozyskaniu takiej nowej informacji *autor*

upatruje możliwości poprawy jakości predykcji modeli antyfraudowych (str. 41),

- w pierwszej osobie liczby mnogiej, np. *mamy tu na myśli funkcję F* (str. 37); *otrzymujemy kolejne potwierdzenie przypuszczeń* (str. 89); *... na próbach niezbilansowanych, czyli takich, z jaką mamy do czynienia w niniejszym badaniu* (str. 90),
 - w formie bezosobowej, np. *W przypadku algorytmów klasteryzujących dokonano próby zasilenia modelu regresji czternastoma zmiennymi* (str. 87).
4. W drugim rozdziale powinien znaleźć się opis metod efektywności algorytmów wykrywania obserwacji nietypowych oraz charakterystyka miary Lift.
 5. Wskazane byłoby zastosowanie oprócz miary Lift innych mierników informujących o jakości predykcji wszystkich zastosowanych metod (dla ODMAT oprócz miary LIFT wykorzystano dodatkowo skorygowany zliczeniowy R kwadrat; LASSO w autorskiej metodzie).
 6. Opis „autorskiej metody algorytmicznej budowy predyktorów” budzi zastrzeżenia. W szczególności wyjaśnienia co do „eksperymentalnych decyzji badacza” (str. 71 i 74,) są niewystarczające.
 7. Charakterystyka wykorzystanych danych empirycznych przedstawiona w podrozdziale por. 5.1 jest bardzo uboga. W literaturze jest wiele dobrych wzorców w tym zakresie, które autor mógłby wykorzystać. Za przykład może służyć opracowanie A. Matuszyk, A. Ptak-Chmielewska, „Profile of the Fraudulent Customer”, *Bezpieczny Bank*, Nr 2(59) 2015.
 8. Przybliżając problematykę oszustw kredytowych w podrozdziale 1.1 Autor stwierdza „Istotą funkcjonowania banku, jak ujmuję to w swojej książce Iwanicz-Drozdowska (2012), powinno być długofalowe osiągnięcie rentowności.” Przy takim sformułowaniu jest to dla wielu czytelników co najmniej teza dyskusyjna. Banki bowiem są jednymi z najważniejszych instytucji finansowych w gospodarce, a rentowność to tylko jeden z celów (por. P. Rose „Zarządzanie Bankiem Komercyjnym”, ZBP, Warszawa 1997).
 9. Przystawiając ryzyko rynkowe Autor nie wykorzystuje najnowszych rekomendacji/trendów w zakresie pomiaru tego ryzyka. Mianowicie powołuje się na VaR pomijając fakt, iż Komitet Bazylejski ds. Nadzoru Bankowego (BCBS) w 2013 r. zaproponował przeniesienie ilościowego systemu miar ryzyka z wartości zagrożonej VaR na oczekiwany niedobór ES (ang. Expected Shortfall). BCBS zauważył, że „zidentyfikowano wiele niedociągnięć z wykorzystaniem VaR do

określania wymogów kapitałowych” (cyt. “a number of weaknesses have been identified with using VaR for determining regulatory capital requirements, including its inability to capture tail risk”, pochodzi z: Ch.-L. Chang, J. Jiménez-Martín, E. Maasoumi, M. McAleer, T. Pérez-Amaral, “A Stochastic Dominance Approach to the Basel III Dilemma: Expected Shortfall or VaR?”, TI 2015-056/III, Tinbergen Institute Discussion Paper).

10. W przypadku opisu ryzyka operacyjnego można także odnieść wrażenie, iż Autor nie zapoznał się w pełni z najnowszą literaturą przedmiotu. Aktualnie w części literatury przedmiotu prezentowany jest pogląd, że ryzyko operacyjne jest dobrze zdefiniowanym typem ryzyka, które doczekało się bardzo szczegółowych metodologii oraz narzędzi informatycznych, o czym świadczą poniższe przykłady publikacji oraz rysunek prezentujący światowych dostawców (według stanu w roku 2010) narzędzi/produktów informatycznych wspierających zarządzanie ryzykiem operacyjnym na tle wartości dwu cech: stopnia pokrycia wymagań biznesowych (ang. completeness of vision) oraz właściwości technologicznych (ang. ability to execute).

- F. Aue, M. Kalkbrener, “LDA at work: Deutsche Bank’s approach to quantifying operational risk”, *Journal of Operational Risk* (49–93) Volume 1/Number 4, Winter 2006/07
- F. Cheng, N. Jengte, W. Min, B. Ramachandran, D. Gamarnik, “Modeling operational risk in business processes”, *Journal of Operational Risk* (73–98) Volume 2/Number 2, Summer 2007.
- M. Górniewicz, „Wymogi jakościowe dotyczące rozwiązań w zakresie metody AMA”, Dep. Insp. Bankowych i Instytucji Płatniczych KNF, CEDUR 2012 r.
- Uchwała KNF nr 76/2010 zał. nr 14 § 55, rekomendacja M??
- P. V. Shevchenko, “Modelling Operational Risk Using Bayesian Inference”, Springer-Verlag Berlin Heidelberg 2011.



Rysunek 1. “Gartner Magic Quadrant for Operational Risk Management”. Źródło: Gartner, Ic., 2010

11. W Sekcji 5.2.1 Autor przesadnie zdawkowo opisuje zastosowanie algorytmu INFLO. Wśród 213 zmiennych były przecież zmienne jakościowe. Czytelnikowi w związku z

tym nasuwają się następujące pytania: Jaki algorytm wykorzystany był do wyznaczenia kierunków głównych? Czy analiza bazowała na macierzy kowariancji czy korelacji oraz jak estymowane były te macierze dla tak dużej liczby zmiennych, w szczególności czy wystąpił problem degeneracji powyższych macierzy?

12. W dalszej części sekcji 5.2.1 Autor opisuje trzy wersje przeprowadzonego eksperymentu obliczeniowego. Jednakże ten opis też jest mocno skrócony i trudno dociec jaka jest różnica pomiędzy metodą pierwszą i drugą – obie oparte są na idei centyli. Nie ma informacji na temat tego, jak poradzono sobie z transformacją w przypadku powtarzających się wartości? Brakuje szerszego komentarza w jaki sposób wybrnięto z problemu braków danych, natomiast należy podkreślić, że uzupełnianie (imputacja) braków danych, wspomniana w trzeciej transformacji, stanowi element całego zespołu różnych metod i technik, począwszy od prostych imputacji określonymi wartościami poprzez zaawansowane algorytmy (por. np. J. Graham, "Missing Data Analysis and Design", Springer, New York, 2012).
13. W końcowej części sekcji 5.3.1 Autor napisał, że „próba w szacowanych modelach budowana była z wykorzystaniem techniki undersamplingu w stosunku 10:90”, w zasadzie bez komentarza pozostawiając tę informację. Natomiast sugerowanie użycia zrównoważonej próbki jako panaceum w przypadku, gdy klasa jest wyjątkowo rzadka jest co najmniej dyskusyjne (por. B. Scarpal, N. Torelli, "Selecting the Training Set in Classification Problems with Rare Events", w: "New Developments in Classification and Data Analysis", 2005, gdzie stwierdzono, że taka sugestia nie ma podstaw ani teoretycznych, ani empirycznych). Z przedstawionych w pracy badań wynika, że w przeciwieństwie do przypadku logistycznego, prosta strategia pobierania próbek z rozwarstwieniem według zmiennej klasyfikacyjnej nie jest tak dobra, jak się spodziewano, a jego możliwe korzyści są nadal niejasne nawet po zastosowaniu rozsądnych korekt.
14. W podrozdziale 5.4, na samym wstępie nasuwają się pytania: Co autor miał na myśli przeprowadzając analizę stabilności rozwiązania? Jaka jest definicja stabilności użyta w tej pracy? Niestety tego można dowiedzieć się dopiero po przeczytaniu całej sekcji.

Wypada podkreślić, że większość przedstawionych uwag ma charakter sugestii, które w znaczący sposób nie wpływają na ocenę całej pracy. Należałoby je jednak uwzględnić przy przygotowaniu pracy do druku, opracowaniu artykułów naukowych oraz w dalszych badaniach naukowych prowadzonych przez Autora.

Wykorzystana literatura naukowa

Autor powołał się na ponad 100 pozycji źródłowych, wśród których dominują artykuły zamieszczone w renomowanych czasopismach naukowych. Doktorant wykazał się dobrą znajomością literatury anglojęzycznej w zakresie zastosowania systemów uczących się w wykrywaniu nadużyć i oszustw finansowych. Jednakże Autor często przy opisie wykrywania obserwacji nietypowych powoływał się na źródła wtórne a nie pierwotne, np. definiując odległość Mahalanobisa na (Walesiak, 2016) zamiast na (Mahalanobis, 1936)¹.

Ocena strony formalnej pracy

Rozprawa napisana jest poprawnie pod względem redakcyjnym. Cytowania zostały opatrzone stosownym odwołaniami do źródeł, tabele oraz rysunki zostały starannie opracowane. W liczącej 128 stron pracy znaleźć jednak można uchybienia natury technicznej, m.in.:

1. Powołując się na literaturę Doktorant często podaje nazwisko autora oraz skrót et al. stosowany w tekstach anglojęzycznych (por. np. (Liao et al., 2015) – str. 17; (Carminati et al., 2015) – str. 21). Jest to praca w języku polskim, a zatem należało użyć polskiego skrótu: (Liao i in., 2016); (Carminati i in., 2015).
2. Sposób powoływania się na literaturę nie jest ujednolicony. W pracy zwykle wykorzystywano konwencję: (Nazwisko autora, rok publikacji), jednak czasami stosowano inne zapisy, np. (Konikiewicz 2012, [za: Frączak, 2012]).
3. Na str. 14 Autor powołuje się na opracowanie (Matuszyk, 2012) zawarte w książce (Iwanicz-Drozdowska, 2012). W bibliografii zabrakło jednak tej pierwszej pozycji - należało uwzględnić pozycję Matuszyk A. (2012), Ryzyko kredytowe, [w:] Iwanicz-Drozdowska M. (red.), Zarządzanie ryzykiem bankowym, Wydawnictwo Poltext, Warszawa. Podobna uwaga dotyczy opracowania (Konikiewicz, 2012) zawartego w książce (Frączak, 2012).
4. Definicja obserwacji nietypowej zaproponowana przez Hawkinsa (1980) przytoczona na str. 41 jest niepotrzebnie powtarzana na str. 63 i 69.
5. Praca zawiera wiele niezręcznych językowo oraz stylistycznie sformułowań, np.:
Ta część pracy pozwala stworzyć wyobrażenie na temat problemu oszustw, na jakie narażony jest sektor bankowy. Przedstawia ona pomysłowość oszustów (...) – str. 9.

¹ Mahalanobis P. C. (1936), On the generalised distance in statistics, Proceedings of the National Institute of Sciences of India. 2 (1), 49–55.

6. W Polsce separatorem dziesiętnym jest przecinek, natomiast w pracy wykorzystano do tego celu kropkę (por. np. tabele 3-14).
7. W sytuacji gdy tabela lub rysunek jest opracowany przez Autora na podstawie innego rysunku lub tabeli zamieszczonego w literaturze, zamiast „Źródło: Opracowanie własne” należy napisać: „Źródło: opracowanie własne na podstawie ...”. W szczególności, zamieszczając w sekcji 2.2.3 Rysunek 1 podano jako źródło opracowanie własne, natomiast bardzo podobne ilustracje zawiera wiele publikacji, na które warto powołać się (np. (i) Duan L., Xu L., Liu Y., Lee J. (2009): “Cluster-based outlier detection”, *Annals of Operations Research*, Vol. 168, pp. 151-168; (ii) Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; Sander, J. (2000): “LOF: Identifying Density-based Local Outliers”, *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. SIGMOD*, Dalles, pp. 93–104).

Wnioski końcowe

Lista uchybień (niedociągnięć) przedstawiona w poprzedniej Sekcji jest stosunkowo długa, ale ich znaczenie nie jest duże i nie mają one na tyle znaczącego wpływu, aby zmienić końcową pozytywną ocenę pracy. Przeprowadzone badania i wnioskowanie na podstawie ich wyników oceniam pozytywnie. Uważam, iż Doktorant wykazał się umiejętnością zarówno formułowania problemów badawczych, jak i właściwym doбором narzędzi i metod badawczych. W rezultacie oceniam, iż przedłożona praca doktorska świadczy o poprawnym warsztacie badawczym Doktoranta i dowodzi znajomości problematyki zastosowań metod analitycznych w wykrywaniu nadużyć, a zaproponowane przez Autora algorytmy wykrywania zdarzeń rzadkich za oryginalne propozycje które mogą znaleźć zastosowanie w realnych procesach udzielania kredytów gotówkowych. Uznaję, że rozprawa w pełni spełnia wymogi Ustawy z 14 marca 2003 r. o stopniach naukowych i tytule naukowym, dlatego wnoszę o jej dopuszczenie do publicznej obrony.

Warszawa 2017-09-18

Wiesław Szczesny

