Szkoła Główna Handlowa w Warszawie
Kolegium Analiz Ekonomicznych
Zakład Ekonometrii Stosowanej

Damian Przekop

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

Autoreferat rozprawy doktorskiej
przygotowanej pod kierunkiem naukowym
prof. dr. hab. Marka Gruszczyńskiego (promotor)
oraz
dr. Marcina Owczarczuka (promotor pomocniczy)

Dziedzina: nauki ekonomiczne
Dyscyplina: ekonomia

Warszawa 2017

1. O pracy

Celem pracy jest sformułowanie i przedstawienie nowego podejścia do budowy modeli predykcyjnych, wykorzystywanych w modelowaniu zdarzeń rzadkich, które są charakterystyczne dla zjawisk nietypowych. Badanie, w którym zaproponowano ich zastosowanie oparte zostało na biznesowym zjawisku oszustw kredytowych.

Wkładem pracy do znanego obecnie dorobku nauki jest wyprowadzenie dwóch algorytmów, wspomagających rozwiązanie problemu identyfikacji zdarzeń rzadkich. W ramach dysertacji autorowi udało się ustalić, że wzbogacenie modeli skonstruowanymi w ramach proponowanych algorytmów predyktorami prowadzi do poprawy ich mocy predykcyjnych i zrozumiałej interpretacji otrzymanych wyników. Uniwersalna konstrukcja oraz założenia leżące u podstaw proponowanych algorytmów sprawiają, iż możliwe jest ich zastosowanie do rozwiązywania problemów rzadkich innych niż oszustwa kredytowe.

Praca składa się z dwóch części, z pierwszej: szerokiej analizy źródeł literaturowych poświęconych wykrywaniu nadużyć oraz z drugiej: prezentacji metody, będącej punktem centralnym dysertacji wraz z badaniem empirycznym przedstawiającym zastosowanie proponowanego rozwiązania na rzeczywistych danych pochodzących z jednego z polskich banków, działającego w sferze obsługi klientów detalicznych.

2. Przedmiot badania i motywacje

Tematyce wykrywania oszustw poświęcono wiele publikacji, jednak oszustwa kredytowe będące przedmiotem opisanego w pracy badania należą do zagadnień bardzo ubogo komentowanych w literaturze. Dorfleitner i Jahnes (2014) przyczyny takiego stanu rzeczy upatrują w braku powszechnego dostępu do danych oraz rygorystycznej cenzurze wyników badań nakładanej przez instytucje dysponujące takimi danymi. Działania te mają na celu ograniczenie ryzyka zyskania przez przestępców przewagi nad bankiem w staraniach o nieuprawnione przejęcie środków znajdujących się do dyspozycji klientów banku. Należy tu dodatkowo podkreślić, za Hand (2007), że sektor bankowy, niezależnie od czynionych wysiłków pozostaje zawsze, w najlepszym razie, krok za oszustami.

Trzema głównymi rodzajami ryzyka specyficznymi dla sektora bankowego według Nowej Umowy Kapitałowej (*Basel II*) są: ryzyko kredytowe, rynkowe i operacyjne (Basel Committee 2006). Z tych trzech rodzajów ryzyka najmniej rozpoznane i określone jest ryzyko operacyjne, które według różnych szacunków odpowiadać może nawet za 35% zmienności wyników finansowych banków uniwersalnych (Cruz 2002). Definicja przyjęta przez Komitet Bazylejski traktuje ryzyko operacyjne jako *ryzyko strat wynikających z niedostosowania lub zawodności wewnętrznych procesów, ludzi i systemów technicznych lub ze zdarzeń zewnętrznych* (Basel Committee 2006). Standard *Bazylea 3* włącza ponadto zagadnienie ryzyka płynności, które od roku 2019 ma zostać poddane ścisłym regulacjom (Iwanicz-Drozdowska, 2012).

Jednym z elementów ryzyka operacyjnego są oszustwa kredytowe, popełniane przez klientów banku na etapie składania wniosku kredytowego. Choć Sanusi et al. (2015) podają, iż ta grupa nadużyć może stanowić aż ponad połowę łącznej wartości wszystkich nadużyć dotykających sektor bankowy, zagadnienie to nie jest jeszcze należycie zbadane. Poważnym utrudnieniem jest wyznaczenie granicy pomiędzy przypadkami nadużyć aplikacyjnych (które są realizacją ryzyka operacyjnego) oraz przypadkami niedotrzymania umowy znanymi w żargonie ryzyka jako *default* (które są realizacją ryzyka kredytowego). Należy podkreślić, iż choć w przypadku zarówno nadużycia aplikacyjnego jak i działania ryzyka kredytowego może dojść do niespłacenia zobowiązania, to motywy kierujące klientami banku w momencie jego zaciągania pozostają odmienne. Tematyką tą zajmowali się m.in.: Hartmann-Wendels et al. (2009), Mählmann (2010), Dorfleitner i Jahnes (2014).

Banki, na miarę swoich możliwości, budują systemy przeciwdziałające zjawisku wyłudzania kredytów przez oszustów kredytowych. Mogą to być proste reguły eksperckie, ale i również zaawansowane modele statystyczne oraz na przykład proponowane przez Van Vlasselaer et al. (2015) sieci powiązań pozwalające na identyfikację potencjalnych relacji wiążących poddawanych weryfikacji klientów z osobami, które w przeszłości dopuściły się już nadużycia. Oszuści z biegiem czasu wykształcają coraz bardziej wyszukane techniki omijania regulacji bezpieczeństwa opracowywanych przez banki. W związku z tym banki sięgają po zaawansowane narzędzia analityczne, które pozwalają na wykrycie zachowań trudnych do prostej weryfikacji.

Z badania przeprowadzonego wśród instytucji finansowych działających na rynku polskim wynika, iż skala zjawiska nadużyć z roku na rok stale się zwiększa[1]. Respondenci każdorazowo spodziewają się wzrostu skali tego zjawiska. W 2012 roku 58% respondentów oceniło, że liczba nadużyć będzie rosła, w 2013 było to już 60%, a w 2014 blisko 75% ankietowanych wyraziło obawy, że skala zjawiska w przyszłości będzie się zwiększać.

W roku 2015, 85% respondentów wszczęło postępowania wyjaśniające w związku z potencjalnymi przypadkami nadużyć, z czego 5% ankietowanych wszczęło ponad 10 000 tego typu postępowań. Skutkiem tego 43% respondentów złożyło zawiadomienia o podejrzeniu popełnienia przestępstwa do właściwych organów ścigania, z czego 8% złożyło ponad 1 000 takich wniosków.

W roku 2016, większość badanych (60%), których dotknął problem nadużyć, zadeklarowała poniesienie łącznej straty w wysokości nieprzekraczającej 500 tys. PLN. W przypadku co trzeciej badanej instytucji deklarowana strata wyniosła powyżej miliona PLN. Dwa rodzaje instytucji, które najczęściej ponosiły straty powyżej 500 tys. PLN to banki uniwersalne i firmy leasingowe.

Wyłudzenia kredytów stanowią 7% wszystkich wyłudzeń w ramach sektora finansowego w Polsce.

Motywacją do podjęcia problemu był deficyt metodyki modelowania, którą można wykorzystać do rozwiązania zagadnienia predykcji zdarzenia rzadkiego, a do tego dynamicznego i zmiennego w czasie, takiego jak oszustwo kredytowe.

3. Problem badawczy

Przedmiotem badań w pracy doktorskiej jest segment klientów *nowych*, tj. nieposiadających historii produktowej w danym banku. Choć charakteryzuje się on kilkukrotnie wyższą częstotliwością występowania nadużyć w stosunku do reszty populacji klientów (posiadających historię produktową w banku), to przypadki nadużyć stanowią jedynie 3%

[1] Raport z badania nadużyć na rynku finansowym, będący podsumowaniem badania przeprowadzonego w ramach cyklicznego projektu badawczego Konferencji Przedsiębiorstw Finansowych oraz firmy audytowej EY. Edycja 2015: https://kpf.pl/pliki/raporty/raport_naduzycia_2015.pdf (stan na 02.02.2017) oraz edycja 2016: https://kpf.pl/pliki/raporty/raport_naduzycia_2016.pdf (stan na 02.02.2017).

rozpatrywanej populacji. Według Hawkinsa (1980) obserwacja nietypowa to *obserwacja odbiegająca tak bardzo od pozostałych, że rodzi się podejrzenie, że wygenerowana została przez mechanizm inny niż pozostałe obserwacje*. Ze względu na charakter zjawiska, niewielki odsetek zdarzeń będących przedmiotem modelowania w rozpatrywanym problemie prognostycznym sprawia, że zdarzenia te postrzegać można jako obserwacje nietypowe dla standardowego schematu postępowania jednostek w badanej populacji.

Problematyka ta jest ściśle powiązana z rozpoznanym w ekonometrii zagadnieniem modelowania zdarzeń rzadkich. Znane w tym zakresie rozwiązania niwelują obciążenie estymatorów wynikające z niezbilansowanej próby uczącej. Nie pomagają one jednak objaśnić większej części zmienności badanego zjawiska.

Proponowanym w literaturze podejściem do zagadnienia poprawy właściwości predykcyjnych modeli jest koncepcja rozwiązań hybrydowych zakładająca wykorzystanie informacji pozyskanej z działania algorytmów uczenia bez nadzoru w ramach szerszych badań, opartych na rozwiązaniach korzystających z uczenia z nauczycielem. Podejście to stosowane było między innymi w pracach Thornton et al. (2014), Farvaresh i Sepehri (2011) czy Krivko (2010). Stanowią one próby wzbogacenia zbioru zmiennych pozostających w dyspozycji badaczy o nowe, kluczowe dla rozwiązania problemu zmienne.

Z punktu widzenia trafności predykcji, metody uczenia bez nadzoru nie tylko nie są w stanie konkurować z najprostszymi nawet metodami uczenia pod nadzorem, ale również pozyskana z nich informacja nie poprawia jakości predykcji modeli opartych na koncepcji uczenia z nauczycielem.

Dla rozpatrywanego zagadnienia, uchwycenie nietypowości wybranych obserwacji wiąże się z poszukiwaniem specyficznych kombinacji cech użytecznych z perspektywy rozwiązania problemu predykcyjnego. Pierwszym ze sposobów rozumienia specyficznego układu cech jest dobór takiej konfiguracji zmiennych (charakterystyk klienta), która z dużym prawdopodobieństwem towarzyszy oszustom kredytowym. Drugim sposobem rozumienia specyficznego układu cech jest rozkład zmiennych ciągłych cechujących danego klienta w ramach wybranej grupy odniesienia, rozumianej jako określona grupa ludzi, względem której wybrany klient charakteryzuje się wyraźnym podobieństwem.

4. Hipotezy badawcze

Tezą badawczą stawianą w pracy jest możliwość poprawy jakości prognoz dostarczanych przez modele predykcyjne w wyniku wzbogacenia ich o dodatkowe predyktory, które – poprzez zwiększenie różnorodności reprezentowanej w modelu – pozwalają na lepszą identyfikację zmienności natury zjawiska rzadkiego – w tym przypadku: oszustwa kredytowego.

W ramach tak postawionej tezy weryfikowane są następujące hipotezy badawcze:

- Hipoteza 1: Zdarzenia rzadkie można prognozować trafniej za pomocą nowego algorytmu konstrukcji zmiennych opartego na koncepcji rozkładu cech ciągłych w ramach zdefiniowanych grup odniesienia.
- Hipoteza 2: Zdarzenia rzadkie można prognozować trafniej za pomocą nowego algorytmu konstrukcji zmiennych opartego na koncepcji wykorzystania informacji mającej swe źródło w logice reguł decyzyjnych.
- Hipoteza 3: Modele wzbogacone o zmienne zaproponowane w oparciu o nowe algorytmy charakteryzują się obciążeniem w postaci przeuczenia, które można redukować przez ograniczenie liczby wykorzystanych w nich predyktorów, bez szkody dla poprawy właściwości predykcyjnych modeli opartych na zaproponowanych zmiennych.

5. Autorska metoda algorytmicznej budowy predyktorów

W pracy prezentowane są dwa algorytmy konstrukcji zmiennych, które poprawiają właściwości predykcyjne modeli dedykowanych zdarzeniom rzadkim.

W ramach pierwszego algorytmu proponowane są zmienne odzwierciedlające specyficzny układ cech rozumiany jako kombinacja wybranych charakterystyk, opierająca się na logice reguł generowanych przez drzewa decyzyjne. Algorytm ten pozwala ująć w jednej zmiennej informację zawartą w kilku zmiennych pierwotnego zbioru danych. Pozwala to na oszczędność wykorzystywanych w modelowaniu stopni swobody.

Włączenie otrzymanych zmiennych do zbioru zasilającego modele stanowi swego rodzaju połączenie dwóch klasyfikatorów. Taki zabieg daje w efekcie model znany w literaturze jako

multi-inducer (Rokach, 2010). Połączenie dwóch paradygmatów klasyfikacji pozwala na otrzymanie efektu synergii, który może dostarczać bardziej precyzyjne wyniki niż składowe takiego rozwiązania.

Istotą pierwszego algorytmu jest identyfikacja unikalnych kombinacji cech, które mogą pozostawać w związku ze zdarzeniem rzadkim jakim jest oszustwo kredytowe. Jego konstrukcja zakłada możliwość stworzenia zmiennych, które w szczególności mogą być każdą możliwą kombinacją zmiennych pochodzących z inicjalnego zbioru danych i są w stanie uchwycić mniej lub bardziej oczywiste predyktory, które reprezentują ukryte w danych trendy *fraudowe* (ang. *fraud* - oszustwo). W wybranych przypadkach nadużyć, wyłącznie kombinacje cech mogą okazać się być przydatne w predykcji ich wystąpienia.

Drugi algorytm opiera się na logice rozkładu zmiennych ciągłych wyznaczanego w ramach wybranej grupy odniesienia, będącej kombinacją wybranych cech nominalnych. Jest to podejście opracowane przez autora w oparciu o znajomość specyfiki badanego zjawiska.

Uzasadnienie logiki wypracowanego algorytmu leży w koncepcji analizy grup odniesienia (ang. *Peer Group Analysis*) przedstawionej przez Kima i Sohna (2012). Informacją cenną z perspektywy rozwiązania problemu predykcyjnego jest nie samo zachowanie badanej jednostki, ale ocena jego nietypowości. Taka konstrukcja zmiennej pozwala dodatkowo mieścić w ramach jednej cechy ładunek informacyjny obejmujący nie tylko konkretną jednostkę, ale i pozostałe obserwacje ujęte w badaniu (poprzez umiejscowienie wartości danej cechy na dystrybuancie rozkładu w ramach wybranej grupy odniesienia).

Drugi algorytm ma na celu uchwycenie informacji użytecznej w śledzeniu nadużyć, która może być dostrzeżona jedynie przez pryzmat najbardziej podobnej grupy odniesienia. Taka sama wartość bezwzględna wybranej zmiennej (np. zarobki) może nieść za sobą dwie różne informacje, gdy interpretowana jest w odniesieniu do dwóch różnych grup odniesienia. Celem algorytmu jest zaakcentowanie znaczenia względnej wartości zmiennych ciągłych.

6. Wyniki badania

W celu weryfikacji wpływu opisywanych algorytmów na poprawę jakości predykcji dostarczanej przez modele *antyfraudowe*, wykorzystano regresję regularyzowaną typu LASSO (ang. *least absolute shrinkage and selection operator*) zaproponowaną przez Tibshirani (1996). Główną zaletą tej metody jest możliwość jej zastosowania do problemów wielowymiarowych, w przypadku których badacz ma do czynienia z wieloma potencjalnymi predyktorami, których liczba w szczególności może przekraczać liczbę dostępnych stopni swobody. Jest to własność, która w przypadku rozpatrywanego problemu jest jedną z kluczowych dla jego rozwiązania.

Eksperyment zakłada zastosowanie regresji logistycznej typu LASSO odpowiednio: dla zbioru danych zawierających wyłącznie zmienne inicjalne oraz zbiorów wzbogaconych o zmienne pochodzące z zaproponowanych wyżej algorytmów.

Za miarę odzwierciedlającą skalę poprawy jakości predykcji dostarczanej przez modele przyjęto poziom miary Lift dla 5 górnych percentyli rozkładu prognoz. Miara ta odzwierciedla stopień spełnienia wymagań biznesowych stawianych przed modelami *antyfraudowymi* wykorzystywanymi w instytucjach bankowych.

Wyniki przeprowadzonego badania potwierdzają wszystkie trzy sformułowane wyżej hipotezy badawcze.

Wyniki badania wskazują, iż możliwa jest poprawa jakości prognoz dostarczanych przez modele predykcyjne nawet o 25% w świetle przyjętego kryterium pomiaru ich jakości. Odbywa się to przez włączenie zmiennych wspomnianych w obu pierwszych hipotezach badawczych, będących rezultatem działania obu zaprezentowanych w pracy algorytmów.

Proponowane w pracy modele rozszerzone o dodatkowe zmienne charakteryzują się pewnym stopniem przeuczenia. Na niezależnym zbiorze testowym dają jednak lepsze prognozy niż standardowe modele oparte na inicjalnym zbiorze zmiennych. Analiza stabilności rozwiązania ze względu na redukcję liczby zmiennych dowodzi możliwości niwelowania tego obciążenia przez ograniczenie liczby predyktorów. Zabieg ten nie tylko nie

pogarsza jakości dostarczanej przez modele predykcji, ale wręcz stanowi on element sprzyjający jej poprawie. Stanowi to potwierdzenie hipotezy trzeciej.

Przeprowadzone dodatkowo analizy stabilności rozwiązania względem doboru próby oraz modyfikacji postaci zmiennej celu potwierdzają niewrażliwość rozwiązania na oba te czynniki. Ostatnia ze wspomnianych analiz potwierdza możliwość wykorzystania proponowanych w podrozdziale algorytmów w celu rozwiązywania problemów rzadkich odmiennych niż zdefiniowane w pracy oszustwa kredytowe.

7. Wkład do rozwoju wiedzy

Wkładem pracy do znanego obecnie dorobku nauki jest wyprowadzenie dwóch algorytmów, wspomagających rozwiązanie problemu identyfikacji zdarzeń rzadkich. W przypadku przeprowadzonego badania są nimi oszustwa kredytowe popełniane przez klientów banku. W ramach badania autorowi udało się ustalić, że wzbogacenie modeli skonstruowanymi w ramach proponowanych algorytmów predyktorami prowadzi do poprawy ich mocy predykcyjnych i zrozumiałej interpretacji otrzymanych wyników. Uniwersalna konstrukcja oraz założenia leżące u podstaw proponowanych algorytmów sprawiają, iż możliwe jest ich zastosowanie do rozwiązywania problemów rzadkich innych niż oszustwa kredytowe.

Podstawową trudnością, z którą musiał zmierzyć się autor pracy jest ubogi stan literatury w zakresie modelowania zjawiska oszustw kredytowych. Choć poruszane zagadnienie znajduje się w obszarze przecięcia się dwóch bardzo dobrze rozpoznanych w literaturze obszarów wiedzy: modelowania ryzyka oraz detekcji nadużyć, to modelowanie ryzyka oszustw kredytowych nie jest jeszcze dobrze opisane w literaturze.

Główną motywacją do podjęcia problemu poruszanego w pracy był deficyt metodyki modelowania, którą można wykorzystać do rozwiązania zagadnienia predykcji zdarzenia rzadkiego charakterystycznego dla zjawisk nietypowych. Na jej skutek, w przedkładanej pracy rozszerzono zbiór możliwych podejść.

Modelowanie zdarzeń rzadkich jest znane w literaturze, jednak najczęściej proponuje się rozwiązania, które jedynie niwelują obciążenie estymatorów wynikające z niezbilansowania

próby uczącej. Nie pomagają jednak objaśnić większej części zmienności badanego zjawiska, co czynią algorytmy prezentowane w prezentowanej pracy doktorskiej.

Proponowane rozwiązanie jest w zamyśle metodą ogólnego przeznaczenia, mającą zastosowanie nie tylko w przypadku modelowania oszustw kredytowych, ale także i innych zdarzeń rzadkich charakteryzujących się silną zmiennością. Jako, że nie jest ona techniką samego modelowania, a doboru zmiennych objaśniających stanowić może ona element będący częścią badań opartych na szerokiej gamie klas modeli predykcyjnych – poczynając od regresji liniowej, a kończąc na rozwiązaniach złożonych takich jak klasyfikatory zintegrowane (ang. *ensemble methods*).

Metoda prezentowana w pracy jest skuteczna, a jej działanie stabilne i niewrażliwe na zakłócenia.

Dodatkowo, rozszerzenie zakresu dostępnych w procesie modelowania predyktorów, pozwala na lepsze poznanie źródeł zmienności badanego zjawiska – w tym przypadku: nadużycia kredytowego. Z uwagi jednak na wrażliwość wykorzystanych danych, temat ten nie został jeszcze podjęty. Niemniej jednak zaprezentowana w pracy metoda budowy predyktorów z powodzeniem może być wykorzystywana przez praktyków i stosowana do wspomagania rzeczywistych problemów decyzyjnych.

8. Literatura

- Basel Committee on Banking Supervision (2006), International Convergence of Capital Measurement and Capital Standards a Revised Framework, Bank for International Settlements, Basel
- Cruz M. (2002), Modeling, Measuring and Hedging Operational Risk, J. Wiley & Sons, New York
- Dorfleitner G., Jahnes H. (2014), What factors drive personal loan fraud? Evidence from Germany, Review of Managerial Science, 1/8, 89-119
- Farvaresh H., Sepehri M. (2011), A data mining framework for detecting subscription fraud in telecommunication, Engineering Applications of Artificial Intelligence, Volume 24, Issue 1, 182–194
- Hand D.J. (2007), Mining Personal Banking Data to Detect Fraud, Selected Contributions in Data Analysis and Classification, 377-386
- Hartmann-Wendels T., Mählmann T., Versen T. (2009), Determinants of banks' risk exposure to new account fraud – Evidence from Germany, Journal of Banking & Finance,33:347–357
- Hawkins D. (1980), Identification of Outliers, Chapman and Hall Hawkins, London
- Iwanicz-Drozdowska M. (2012), Zarządzanie ryzykiem bankowym (redaktor i współautor), Wydawnictwo Poltext, Warszawa
- Kim Y., Sohn S. (2012), Stock fraud detection using peer group analysis, Expert Systems with Applications, 39:8986–8992
- Krivko M. (2010), A hybrid model for plastic card fraud detection systems, Expert Systems with Applications, 37:6070–6076
- Mählmann T. (2010), On the correlation between fraud and default risk, Zeitschrift für Betriebswirtschaft, December, Volume 80, Issue 12, 1325-1352
- Rokach L. (2010), Ensemble-based classifiers, Artificial Intelligence Review, 33:1-39
- Sanusi Z., Rameli M., Isa Y. (2015), Fraud Schemes in the Banking Institutions: Prevention Measures to Avoid Severe Financial Loss, Procedia Economics and Finance, 28:107 – 113
- Thornton D., Capelleveen G., Poel M., Hillegersberg J., Müller R. (2014), Outlier-based Health Insurance Fraud Detection for U.S. Medicaid Data, 16th International

Conference on Enterprise Information Systems, ICEIS 2014, 27-30 April 2014, Lisbon, Portugal, 684-694

- Tibshirani T. (1996), Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society. Series B (Methodological), 58:267-288
- Van Vlasselaer V., Eliassi-Rad T., Akoglu L., Snoeck M., Baesens B. (2015), Gotcha! Network-based Fraud Detection for Social Security Fraud, Management Science (submitted)