Warszawa, 30 listopada 2021 r.

dr hab. Piotr Wójcik, prof. ucz.

Katedra Finansów Ilościowych

Wydział Nauk Ekonomicznych

Uniwersytet Warszawski

Ul. Długa 44/50

00-241 Warszawa

Recenzja

rozprawy doktorskiej mgr. Wojciecha Starosty

pt. Modeling the Loss Given Default of Retail Contracts

napisanej pod kierunkiem dr. hab. Pawła Baranowskiego, prof. UŁ

oraz dr. Mariusza Górajskiego

sporządzona na zlecenie Komisji ds. stopni naukowych w dyscyplinie ekonomia i finanse

Uniwersytetu Łódzkiego

Ocena ryzyka związanego z działalnością kredytową jest jednym z najważniejszych wyzwań nowoczesnych instytucji finansowych. Zgodnie z wymogami regulacyjnymi (Bazylea II) instytucje finansowe muszą wdrożyć zaawansowane podejście do pomiaru ryzyka kredytowego oparte na ratingach wewnętrznych (ang. *advanced internal rating-based*, AIRB). Metoda ta wymaga, aby wszystkie elementy ryzyka były obliczane wewnętrznie w instytucji finansowej. Zgodnie ze standardem AIRB wymagane jest szacowanie trzech parametrów ryzyka: prawdopodobieństwo niewykonania zobowiązania (ang. *probability of default*, PD), strata z tytułu niewykonania zobowiązania (ang. *loss given default*, LGD) oraz ekspozycja w momencie niewykonania zobowiązania (ang. *exposure at default*, EAD). Badania empiryczne poświęcone ryzyku kredytowemu w przeważającej większości skupiają się na indywidualnej zdolności kredytowej (PD) oraz jej czynnikach. Analizy naukowe poświęcone pozostałym wskaźnikom są znacznie mniej liczne. Autor ocenianej rozprawy skupił się w prezentowanych badaniach na modelowaniu i prognozowaniu straty z tytułu niewykonania zobowiązania (LGD) i poszukiwaniu czynników na nią wpływających. Tym samym Autor dołączył do olbrzymiej rzeszy naukowców i praktyków próbujących modelować ryzyko kredytowe, skupiając się jednak na mniej wyeksploatowanym empirycznie wskaźniku. Można zatem stwierdzić, że przedstawione w rozprawie badania mieszczą się w głównym nurcie ekonomii i finansów empirycznych. Głównym narzędziem stosowanym przez Doktoranta są modele regresyjne, służące do modelowania ciągłej zmiennej objaśnianej. Autor stosuje w przedstawionych badaniach zarówno modele parametryczne (regresję liniową oraz regresję ułamkową), jak i nieparametryczne (drzewa regresyjne), a także wybrane nowoczesne algorytmy uczenia maszynowego (np. maszyna wektorów nośnych, ang. *suport vector machine*, SVM).

Przedstawiona mi do recenzji rozprawa dotyczy więc aktualnego i ważnego tematu, a w przeprowadzonych badaniach wykorzystane zostały stosunkowo nowe, zaawansowane narzędzia, które

pozwoliły na uzyskanie oryginalnych i ważnych wyników empirycznych. Jest to więc niewątpliwie bardzo wartościowa praca. Jej szczegółową ocenę przedstawiam poniżej.

Ocena wartości merytorycznej rozprawy

Rozprawa doktorska pana mgr. Wojciecha Starosty składa się ze wstępu i czterech artykułów opublikowanych w latach 2020-2021 w punktowanych czasopismach naukowych:

1. Starosta Wojciech (2020),„Modelling Recovery Rate for Incomplete Defaults Using Time Varying Predictors, *Central European Journal of Economic Modelling and Econometrics*, 70 pkt MNiSW, IF: brak.
2. Starosta Wojciech (2021), "Beyond the Contract: Client Behavior from Origination to Default as the New Set of the Loss Given Default Risk Drivers", *Journal of Risk Model Validation*, Vol. 15, No. 1, 40 pkt MNiSW, IF: 0.357 (2021).
3. Starosta Wojciech (2021), "Loss given default decomposition using mixture distributions of in-default events", *European Journal of Operational Research*, 2021, vol. 292, issue 3, 1187-1199, 140 pkt MNiSW, IF: 5.334 (2021).
4. Starosta Wojciech (2021), "Forecast combination approach in the loss given default estimation", *Applied Economics Letters*, vol. 28, pp. 1813-1817, 40 pkt MNiSW, IF: 1.157 (2020).

W dalszej części recenzji będę przywoływał numery artykułów zgodnie z podanym powyżej porządkiem ustalonym przez Autora.

Głównym celem prezentowanej rozprawy sformułowanym we **wstępie**, choć nieco ukrytym w tekście na stronie 3, jest zaproponowanie efektywnych form estymacji LGD, co obejmuje właściwe obliczenie tej miary, uwzględnienie odpowiednich czynników oraz formy funkcyjnej modelu, a także wykazanie, że zastosowana metoda jest właściwa i daje precyzyjne oszacowania. We wstępie Autor omawia podstawowe pojęcia związane z modelowaniem ryzyka kredytowego, nietypowy (dwumodalny) rozkład wartości LGD oraz kwestie regulacyjne. Następnie przedstawia kolejne kroki procesu modelowania, w tym czynniki ryzyka oraz najpopularniejsze modele parametryczne (regresja liniowa, regresja ułamkowa, regresja beta) stosowane do estymacji LGD. Omawia również krótko dwa wybrane modele nieparametryczne stosowane w literaturze przedmiotu – drzewa regresyjne i maszynę wektorów nośnych, a także wspomina o stosowanym w ostatnich latach modelowaniu dwustopniowym powiązanym z dekompozycją estymacji LGD na etap modelowania prawdopodobieństwa wystąpienia straty, a następnie modelowania oczekiwanej warunkowej wartości straty. W kolejnej części wstępu omówione zostały kwestie walidacji i monitorowania modelu wraz ze wskazaniem odpowiednich miar siły dyskryminacyjnej czy precyzji kalibracji modelu. Autor podkreśla, że ważnym elementem budowania, walidacji i monitorowania modelu jest pełne zrozumienie jego działania, transparentność, możliwość zrozumienia uzyskanych zależności. W tym kontekście przywołuje zalety tradycyjnych modeli parametrycznych, w których uzyskane wartości parametrów mogą być interpretowane w odniesieniu do kierunku i siły wpływu poszczególnych czynników na badane zjawisko. W kontraście do modeli parametrycznych stawia modele uczenia maszynowego, będące „czarnymi skrzynkami", które bezpośrednio takiej interpretacji nie umożliwiają. Wspomina jednak jednym zdaniem o tzw. interpretowalnym uczeniu maszynowym (nazywanym także wytłumaczalną sztuczną inteligencją, ang. explainable artificial intelligence, XAI) – dynamicznie rozwijających się w ostatnich latach narzędziach, które pozwalają zajrzeć w głąb dowolnej „czarnej skrzynki" i zwizualizować odkryty przez model kształt relacji między zmienną objaśnianą a wybranym predyktorem (np. Partial Depencency Profile, PDP) czy uszeregować zmienne wg ich ważności mierzonej w analogiczny sposób niezależnie od rodzaju zastosowanego modelu (np. permutacyjne miary ważności).

Ostatnim elementem wstępu jest krótkie omówienie poszczególnych artykułów oraz ich wkładu (ang. *contribution*) do istniejącego stanu wiedzy na temat modelowania LGD. Przy każdym artykule Autor zamieścił powiązaną z nim hipotezę badawczą (przy pierwszym artykule nawet dwie). Przy czym hipotezy sprawiają wrażenie wymyślonych *ex-post* – już po publikacji omawianych artykułów, na potrzeby przygotowania wstępu do pracy doktorskiej. Nie są one przywołane w opublikowanych artykułach. Co więcej, w omówieniu pracy brak też podsumowania wyników ich weryfikacji na podstawie przeprowadzonych badań. Dodatkowo, sposób sformułowania hipotez badawczych nie zawsze jest precyzyjny. Hipoteza naukowa powinna być sformułowana w sposób umożliwiający jej sfalsyfikowanie (potwierdzenie lub odrzucenie) przy pomocy przeprowadzonego badania. Tymczasem np. hipoteza druga brzmi:

„Pożyczki zabezpieczone i niezabezpieczone mają różne wzorce, które mogą być odzwierciedlone kolejno metodą nieparametryczną i parametryczną" [ang. „*Secured and non-secured loans include different patterns, which can be reflected by non-parametric and parametric method consecutively*."]

Trudno bez dodatkowego wyjaśnienia domyślić się, co konkretnie Autor miał na myśli.

Hipoteza trzecia została sformułowana następująco:

„Zachowanie klienta po udzieleniu pożyczki staje się ważnym elementem zbioru czynników ryzyka LGD" [ang. „*Client behaviour after loan granting becomes important part of loss given default risk drivers set*"].

Trudno powiedzieć na podstawie sformułowania hipotezy, co Autor rozumie przez „zachowanie" oraz przez „staje się"? Lepsze byłoby stwierdzenie, że zachowanie po prostu jest ważnym czynnikiem ryzyka i jego pominięcie może skutkować niedoszacowaniem ryzyka, oraz krótkie omówienie, jakiego rodzaju zachowanie Autor ma na myśli.

Podobnie przy sformułowaniu hipotezy czwartej:

„Dekompozycja LGD na podstawie zdarzenia niewykonania zobowiązania prowadzi do wzrostu precyzji" [ang. „*LGD decomposition based on in-default event leads to precision uplift*"]

oraz piątej:

„Uśrednianie prognoz z modeli opartych na zmiennych idiosynkratycznych i systematycznych oddzielnie prowadzi do poprawy precyzji prognoz długoterminowych dla parametru LGD" [ang. „*Forecast averaging from models based on idiosyncratic and systematic variables separately leads to precision improvement of long-term forecasts for LGD parameter*"]

zabrakło informacji względem czego Autor oczekuje poprawy wspomnianych wskaźników. Pytanie też co Autor rozumie przez „długoterminowe prognozy dla parametru LGD"?

Generalnie właściwe byłoby szersze omówienie i bardziej szczegółowe uzasadnienie poszczególnych hipotez, czego w przygotowanym wstępie niestety zabrakło.

Razi też nieco brak wyraźnego określenia w poszczególnych artykułach ich głównego celu. Zamiast tego Autor omawia ich wkład do istniejącego stanu wiedzy na temat modelowania LGD.

Niemniej wstęp jest dobrym wprowadzeniem do przedmiotu badania i tematyki poszczególnych artykułów, których wspólnym mianownikiem jest modelowanie LGD. Jednak refleksja nad celami badania, pytaniami czy hipotezami badawczymi powinna być bardziej spójna i pełniejsza.

W **pierwszym artykule** Autor podejmuje ważny problem potencjalnego obciążenia próbki badawczej i uwzględniając wytyczne regulacyjne, szacuje wielkość straty z tytułu niewykonania zobowiązania bazując nie tylko na zamkniętych przypadkach niewykonania zobowiązania (ang. *default*), ale także uwzględniając częściowe odzyski z niezakończonych przypadków. Autor proponuje metodę szacowania częściowych stóp odzysku dla spraw otwartych opartą na modelowaniu odzysków w przedziałach, w których zmienną objaśnianą są wszystkie przepływy pieniężne obserwowane od początku przedziału do końca okna procesu odzyskiwania. W badaniu wykorzystał rzeczywiste dane z polskiego banku komercyjnego stosującego system AIRB do obliczania LGD. Modele budowane są na danych z lat 2003-2015 i walidowane na niewykonaniach zamkniętych w okresie 2015-2017. Na próbce stosowane są dwie metody parametryczne (regresja ułamkowa i regresja beta) i dwie nieparametryczne (drzewo regresyjne i regresja wektorów nośnych). Na podstawie przeprowadzonego badania Autor wnioskuje, że odzyski są determinowane przez inny zestaw cech w kolejnych okresach po momencie niewykonania zobowiązania. Okazuje się również, że drzewa regresyjne dają lepsze wyniki od innych metod dla produktów zabezpieczonych, natomiast regresja ułamkowa okazuje się najlepsza dla kredytów niezabezpieczonych.

Opis wykorzystanych metod (zwłaszcza nieparametrycznych) jest dość lakoniczny. Nie chodzi o opis jak działają, tylko o dobór niezbędnych hiperparametrów. Nie ma też w artykule przekonującego uzasadnienia dlaczego ograniczono się wyłącznie do tych modeli. W tekście brakuje informacji w jaki sposób dokonano parametryzacji algorytmu SVR (parametr kosztu + funkcja jądra i jej hiperparametry)? Dlaczego wykorzystano radialną funkcję jądra i dlaczego wyłącznie tę? Czy przy znajdowaniu optymalnych wartości hiperparametrów SVR także wykorzystano 10-krokową walidację krzyżową, podobnie jak przy optymalizacji parametru kosztu złożoności dla drzewa regresyjnego? Czy wykorzystano ten sam podział na „folds”? Jeśli nie, dlaczego optymalny model wyznaczono w inny sposób? Co Autor rozumie przez „plain version” modelu SVR (s. 214)? Zapewne chodzi o domyślne wartości hiperparametrów dostępne w wykorzystanej implementacji tego algorytmu – jakie i dlaczego takie? Dlaczego zastosowano wyłącznie pojedyncze drzewo regresyjne, a nie również metody oparte na wielu drzewach (np. las losowy czy wzmacniane drzewa regresyjne)? Powszechnie znaną wadą drzewa regresyjnego jest ograniczona liczba unikalnych prognozowanych wartości zmiennej objaśnianej. Autor wspomina o wykorzystaniu regularyzacji L1 w przypadku metod parametrycznych ze względu na silną korelację między predyktorami. Nie informuje jednak, w jaki sposób wybrano parametr wagi przy regularyzacji. Brak też informacji, dlaczego analogicznie nie wykorzystano regularyzacji w wykorzystanych metodach nieparametrycznych. Czy skorelowane zmienne zostały pomięte przy estymacji modeli nieparametrycznych? Jeśli nie – dlaczego? Inna sprawa, że SVR jest de facto modelem parametrycznym, w którym szacowane są parametry wyznaczające optymalną hiperpłaszczyznę. Podsumowując – opis metodologii jest skrótowy, a przez to nieprecyzyjny.

Z artykułu wynika, że modele były porównywane na próbie out-of-time – pożyczek mających status otwartych w roku 2015 i zamkniętych w 2017. Pytanie, czy wydzielono także próbę walidacyjną dla tego samego okresu i struktury pożyczek, które były uwzględnione w próbie uczącej? Na przykład czy porównywano modele na próbie walidacyjnej wykorzystanej w ramach 10-krokowej walidacji krzyżowej dla drzewa regresyjnego? Jeśli nie, dlaczego? Zwykle modele ryzyka kredytowego porównywane są zarówno na próbie *out-of-sample*, jak i *out-of-time*.

Autor wspomina w końcowej części artykułu o innych metodach nieparametrycznych stosowanych w literaturze przedmiotu (lasy losowe, wzmacnianie gradientowe), dających potencjalnie lepsze dopasowanie modelu i prognozy kosztem zmniejszenia interpretowalności samego modelu. Pytanie dlaczego nie wykorzystał ich w swoim badaniu? Zwłaszcza że jednym z celów badania jest właśnie porównanie skuteczności modeli parametrycznych i nieparametrycznych. Poza tym wykorzystany w artykule model SVR (choć w pewnym sensie jest jednak parametryczny), również nie daje możliwości łatwej interpretacji wyników. Poza

tym brak interpretowalności modeli zwanych „czarnymi skrzynkami" jest do pewnego stopnia mitem, ze względu na dynamicznie rozwijający się w ostatnich latach zestaw narzędzi nazywanych interpretowalnym uczeniem maszynowym (lub wytłumaczalną sztuczną inteligencją) – czego Autor jest zresztą świadom, wskazując na możliwość wykorzystania wykresów *Partial Dependency Profile* lub *Individual Conditional Expectation*.

W **drugim artykule** Autor wskazuje na potrzebę uwzględnienia w modelu nowych czynników ryzyka, które mogłyby zwiększyć precyzję estymacji wartości LGD. Podejmuje więc bardzo istotny temat doboru właściwych zmiennych objaśniających. Uwzględnienie odpowiednich czynników modelowanego zjawiska, co więcej uwzględnienie ich w odpowiedniej formie (tzw. *feature engineering*), jest znacznie ważniejsze dla jakości modelu niż wykorzystanie skomplikowanych „modnych" modeli analitycznych. Dodatkowe zmienne dotyczą zachowania klienta na rachunku depozytowym, stopnia relacji klienta z bankiem (częstości logowania w serwisie transakcyjnym, kanałów komunikacji, wykorzystania aplikacji mobilnej), informacji o innych posiadanych produktach kredytowych oraz składanych wnioskach o nowe produkty. W większości przypadków Autor uzasadnia w przejrzysty sposób oczekiwany kierunek ich wpływu na wartość LGD. Rozszerzenie zestawu predyktorów skutkuje redukcją wartości analizowanych miar błędów i poprawia jakość prognoz z modeli. Poprawa jest większa w podejściu parametrycznym (regresja ułamkowa) niż w nieparametrycznym (drzewo regresyjne). Efektem badania może być rekomendacja dla instytucji regulacyjnych sugerująca włączenie informacji o zachowaniach klienta do modeli LGD.

Autor szacuje model wykorzystujący „standardowe" predyktory jako punkt odniesienia (benchmark). Pytanie w jakim sensie te predyktory są standardowe? Czy jest to wynik jakiegoś konsensusu? Czy są one określone regulacjami czy badaniami empirycznymi? Uzasadnienie jest dość lakoniczne – opiera się jedynie na sześciu wcześniejszych badaniach, w tym pięciu z lat 2010-2013. Przydałoby się oparcie listy „standardowych" predyktorów na szerszym przeglądzie literatury z zestawieniem listy analizowanych czynników ryzyka i wskazaniem częstości ich wykorzystania w badaniach. Być może sensowne byłoby wariantowe zdefiniowanie kilku „koszyków" standardowych predyktorów – np. występujących we wszystkich badaniach, w większości badań, itp.

W artykule brak jest precyzyjnej informacji w jaki sposób mierzona jest ważność zmiennych (podawana dla drzewa regresyjnego) – „sum of goodness of split measures for each split for which it was the primary variable" – czy to oznacza sumę wartości statystyki ANOVA? Co kluczowe, ocena ważności zmiennych wykorzystanych w drzewie regresyjnym NIE jest porównywalna z oceną istotności zmiennych w modelu regresji liniowej na podstawie ich istotności statystycznej. Tymczasem Autor wydaje się równorzędnie traktować oba podejścia („Seven of the eight new variables were found to be statistically significant, which supports the hypothesis that the contract owner behavior is connected to the RR (Table 8). The same can be stated for the regression tree model (Table 9)" – s. 83 oraz "In the regression tree model the new variables have less influence […]" – s. 85). Aby móc bezpośrednio porównać ważność zmiennych w modelach parametrycznych i nieparametrycznych, należy wykorzystać metody agnostyczne względem modelu – np. permutacyjne miary ważności zmiennych z arsenału metod interpretowalnego uczenia maszynowego, którego istnienia Autor jest świadom, gdyż wspominał o tym w pierwszym opublikowanym artykule oraz we wstępie do dysertacji. W kontekście braku takiego porównania, zacytowane powyżej stwierdzenia zamieszczone w artykule są zbyt mocne. Sama istotność statystyczna nie mówi też nic o rankingu ważności zmiennych w regresji liniowej. Tu przydatne byłoby choćby proste porównanie wartości (absolutnych) standaryzowanych współczynników regresji.

Autor stwierdza, że wszystkie dodatkowe zmienne (dotyczące zachowań klienta) w modelu regresji ułamkowej są istotne statystycznie, co nie dziwi w sytuacji, gdy model szacowany jest na bardzo dużej próbie

(150 tys. obserwacji). Wyniki testów indywidualnej istotności powinny zostać uzupełnione o test istotności łącznej. Co ważniejsze, kwestia wnioskowania na podstawie istotności statystycznej w erze *big data* powinna być traktowana nieco inaczej niż w mniej licznych próbach. Należy spodziewać się, że im większa próba badawcza, tym mniejsze różnice w szacowanych na jej podstawie wskaźnikach będą istotne statystycznie. Pytanie **czy każda istotna statystycznie różnica jest znacząca** (ang. *meaningful*)? Ta ocena zależy w dużym stopniu od wiedzy eksperckiej o analizowanym problemie (której Autorowi nie brakuje). Również w kontekście tego problemu zasadne byłoby inne podejście do oceny sensowności wykorzystania dodatkowych zmiennych w modelu – oceny ich mocy predykcyjnej, np. za pomocą permutacyjnych miar ważności zmiennych. Biorąc powyższe pod uwagę, **zbyt mocna jest konkluzja** artykułu mówiąca, że „incorporating information about the contract owner's behavior plays a **crucial** [podkreślenie PW] role in the predictive accuracy of LGD modeling" (s. 89).

W zdaniu „Taking fractional regression into consideration, the challenger model can be characterized by a material upgrade, and adding new variables **significantly** [podkreślenie PW] boosts the precision and discrimination" (s. 87) użyte sformułowanie sugeruje weryfikację statystyczną różnic między miarami dopasowania i precyzją prognoz modeli, co nie miało miejsca, a przynajmniej nie zostało omówione w przedstawionym artykule.

Na s. 83 Autor nieprecyzyjnie używa sformułowania „interakcje" zmiennych w modelu, sugerującego wykorzystanie w formie funkcyjnej oprócz indywidualnych predyktorów także ich iloczynów, które pozwalają ocenić, czy wpływ wybranego predyktora na zmienną zależną jest funkcją innego z predyktorów. Tymczasem chodzi po prostu o łączne wykorzystanie wielu predyktorów w modelu. Co więcej, wykorzystanie interakcji w modelu regresji ułamkowej byłoby jak najbardziej wskazane i mogłoby podnieść moc predykcyjną modelu. Założenie, że wpływ niektórych zmiennych behawioralnych albo dotyczących posiadanych produktów bankowych na LGD/RR jest różny np. w różnych grupach wiekowych, zależy od stażu klienta w banku albo od płci, wydaje się racjonalne i warte empirycznego zweryfikowania. Pytanie, czy niewykorzystanie w modelu faktycznych interakcji między zmiennymi wynika z regulacji (np. zabraniających dyskryminacji różnego rodzaju), czy ma inne uzasadnienie?

Nasuwa się również pytanie, dlaczego, mimo wysokiej korelacji między niektórymi predyktorami, Autor nie zastosował regularyzacji L1, którą wykorzystywał w poprzednim badaniu (artykuł 1), tylko wykonywał „ręczne" usuwanie wybranych zmiennych?

Autor wnioskuje o nieliniowym wpływie liczby logowań do serwisu internetowego na zmienną zależną („At the beginning, there is indeed a positive relationship, but a negative relationship develops as the number of log-ins increases" – s. 84-85) błędnie interpretując zmianę współczynnika przy tej zmiennej z dodatniego na ujemny po dodaniu większej liczby predyktorów do modelu. Stwierdzenie Autora byłoby łatwo zweryfikować dodając do modelu kwadrat wspomnianej zmiennej, o czym Autor zresztą napisał kilka zdań później, ale z niewiadomych przyczyn nie zastosował tego w swoim badaniu. De facto taka prosta weryfikacja występowania potencjalnie nieliniowych relacji ze zmienną celu mogłaby zostać zastosowana dla wszystkich predyktorów.

Dodatkowo kierunek i siłę wpływu poszczególnych predyktorów na zmienną celu można by bezpośrednio porównać dla obu modeli (regresja ułamkowa i drzewo regresyjne) za pomocą dostępnych narzędzi interpretowalnego uczenia maszynowego (np. PDP) – nie do końca zrozumiałe jest czemu Autor z nich nie skorzystał, wiedząc o ich możliwościach.

Brak oczekiwań co do spodziewanego kierunku wpływu predyktora na zmienną zależną nie zwalnia badacza z próby interpretacji uzyskanych wyników analizy empirycznej. W tym kontekście stwierdzenie: „The parameter sign for the last two variables is negative, but we did not make any initial assumptions about their

influence, which implies the need to confirm the direction in further research" jest unikiem, który powinien być poprzedzony próbą zrozumienia tych rezultatów.

W **trzecim artykule**, który został opublikowany w wysoko punktowanym czasopiśmie, Autor stosuje modelowanie dwustopniowe – dekompozycję estymacji LGD na etap modelowania prawdopodobieństwa wystąpienia straty, a następnie modelowania oczekiwanej warunkowej wartości straty. Ponadto, w każdym komponencie wykorzystuje dane zagregowane na różnych poziomach, aby odzwierciedlić charakterystykę zjawiska w każdym etapie procesu windykacyjnego. Ponownie Autor stosuje jedynie metody modelowania wykorzystywane we wcześniejszych artykułach (regresja liniowa/ułamkowa, regresja logistyczna, drzewa regresyjne i SVM). Jako punkt odniesienia zastosowano jednostopniowy model regresji liniowej. Ostatecznie wyestymowane zostały cztery modele: (1) jednoetapowy model regresji liniowej, (2) dwuetapowy model łączący maszynę wektorów nośnych (SVM/LS-SVC) i regresję liniową, (3) dwuetapowy model z regresją logistyczną i regresją liniową, (4) dwuetapowy model łączący drzewo klasyfikacyjne z drzewem regresyjnym. Uzyskane wyniki sugerują, że dekompozycja jest bardziej efektywna niż podejście jednostopniowe, a jednocześnie model zachowuje wysoki poziom interpretowalności.

Ten artykuł sprawia wrażenie najbardziej dopracowanego (co zapewne jest powiązane ze złożonym procesem recenzyjnym w wysoko punktowanym czasopiśmie), dlatego mam do niego najmniej uwag. Niemniej Autor nie ustrzegł się pewnych błędów. Przede wszystkim ponownie porównuje rzeczy nieporównywalne – ważności zmiennych w drzewie regresyjnym (przeskalowane tak, aby w sumie dawały 100) z wartościami współczynników regresji logistycznej (nieprzeskalowanymi). W przypadku regresji logistycznej sensowniejsza byłaby ocena ważności zmiennych na podstawie wartości absolutnych standaryzowanych efektów krańcowych (Tabela 4 i 5). Najlepszym rozwiązaniem byłoby porównanie obu modeli za pomocą analogicznej metryki ważności zmiennych bazującej na podobnej ocenie ważności niezależnie od rodzaju zastosowanego modelu (agnostycznej względem modelu).

Dodatkowo, jak rozumiem Autorowi chodziło o oparcie zdekomponowanego modelu na dwóch komponentach tego samego typu – parametrycznym (regresja logistyczna + regresja liniowa) i opartym na drzewie (drzewo klasyfikacyjne + drzewo regresyjne). Ale naturalny dodatkowy model obejmowałby maszynę wektorów nośnych i regresję wektorów nośnych lub potencjalnie mieszankę różnych typów modeli. Co więcej, wiele innych nowoczesnych algorytmów uczenia maszynowego (również tych opartych na łączeniu wielu drzew) umożliwia analizę zarówno problemów klasyfikacyjnych, jak i regresyjnych oraz ocenę ważności zmiennych. Czy nie zostały one wykorzystane z jakiegoś konkretnego powodu?

Hiperparametry modeli CART i SVM zostały dobrane za pomocą 10-krokowej walidacji krzyżowej – pytanie czy z analogicznym podziałem na części (folds) w przypadku obu modeli. Brak również choćby krótkiego uzasadnienia do (krótkiej) listy rozważanych wartości hiperparametrów.

W **czwartym artykule** Autor proponuje procedurę łączenia prognoz z modeli bazujących osobno na zmiennych dotyczących kontraktu oraz wskaźnikach makroekonomicznych. Wykorzystuje tu jedynie regresję liniową. Łączenie prognoz wykonane jest na trzy sposoby: z wykorzystaniem równych wag, metodą Grangera-Ramanathana oraz uśrednianiem metodą Mallowsa. Testując jakość prognoz na dodatkowej próbie Autor stwierdza, że precyzja prognozy kombinowanej jest wyższa niż prognoz pochodzących z poszczególnych modeli. Nie wykonano jednak formalnych testów porównujących dokładność prognoz z różnych podejść, jak Autor uczynił choćby w artykule 3 (test Diebolda-Mariano). Czy porównywano jakość prognoz z poszczególnych modeli cząstkowych? Jakie dały rezultaty?

Badanie zaprezentowane w tym krótkim artykule to dość proste ćwiczenie empiryczne, które warto byłoby uzupełnić o większą liczbę dodatkowych analiz wrażliwości. Informacja o wykorzystanym zbiorze danych jest bardzo lakoniczna, w przeciwieństwie do pozostałych artykułów. Jakie było źródło danych? Czy analizowane obserwacje zostały wybrane z większej próbki czy stanowią pełną populację kredytów hipotecznych z pewnego polskiego banku? Autor wykorzystuje zmienne makroekonomiczne (produkt krajowy brutto, CPI, przeciętne płace, itp.), które były także uwzględnione we wcześniejszym badaniu (artykuł 3) oraz w innych badaniach przywoływanych w tym artykule (np. Yao et al., 2017). Autor nie omawia jednak szczegółów dotyczących wykorzystanych wskaźników makroekonomicznych – czy były to dane roczne/kwartalne/dzienne, różne dla różnych wskaźników? Na jaki moment w czasie? Koniec roku? Ogólnokrajowe czy regionalne? Brak tych informacji utrudnia ocenę adekwatności wykorzystanych wskaźników.

Czy poszczególne modele cząstkowe szacowane są na pełnej próbie czy na mniejszych podpróbach o tej samej liczebności *n*? Nie wynika to jasno z opisu w artykule. Autor pisze o podzieleniu zbioru danych na dwie podpróby (ang. *sub-sets*) i szacowaniu każdego z modeli na próbie wielkości *n*, nie definiując jednak co ta wielkość oznacza.

Problemem w przypadku tego badania może być obciążenie uzyskanych estymatorów. Jeśli w modelu regresji liniowej wykorzystana zostanie ograniczona postać funkcyjna modelu (pominięte zostaną ważne zmienne objaśniające), uzyskane oszacowania parametrów będą obciążone (jest to znany w literaturze problem zmiennych pominiętych), a tym samym model będzie niepoprawny. Obciążenie estymatorów nie wystąpi jedynie wtedy, gdy pominięte czynniki nie są skorelowane z predyktorami uwzględnionymi w modelu. W artykule zabrakło choćby krótkiego odniesienia do tego problemu. Czy analizowano korelacje między predyktorami? Dlaczego nie zastosowano regularyzacji L1, a jedynie eliminację wsteczną, która niekoniecznie wyeliminuje silnie skorelowane zmienne?

To jedyny artykuł, w którym Autor wykorzystuje jedynie prostą regresję liniową – uzyskane wnioski byłyby silniejsze, gdyby większa precyzja prognoz kombinowanych została potwierdzona przy wykorzystaniu innego rodzaju modelu bazowego – np. stosowanych przez Autora we wszystkich poprzednich artykułach drzew regresyjnych. Czy taka próba została podjęta? Jeśli tak, jaki dała efekt?

Formą weryfikacji sensowności oddzielenia zmiennych dotyczących kontraktu od wskaźników makroekonomicznych i oszacowania na ich podstawie osobnych modeli mógłby być losowy podział wszystkich zmiennych na dwa podzbiory, oszacowanie na nich osobnych modeli i weryfikacja jakości prognoz kombinowanych uzyskanych w ten sposób. Czy taka próba została podjęta? Jeśli tak, jaki dała efekt?

W prognozach kombinowanych lepsze rezultaty uzyskano w przypadku zróżnicowania wag dla modeli cząstkowych. Prognozy z modelu opartego na wskaźnikach makroekonomicznych miały niższą (jak bardzo? Brak zestawienia wag prognoz cząstkowych) wagę. Autor wnioskuje z tego, że zmienne makroekonomiczne mają ograniczony wpływ na zmienność LGD. Jest to wniosek dość oczywisty, skoro zmienne makroekonomiczne przyjmują tę samą wartość dla wielu obserwacji, przez co znacznie słabiej różnicują wartości LGD od siebie. Sensowna wydaje się w tym przypadku weryfikacja użyteczności wskaźników makroekonomicznych dostępnych na poziomie regionalnym czy lokalnym (np. stopa bezrobocia, przeciętne wynagrodzenie, udział budżetów gmin w przychodach z PIT/CIT, regionalne indeksy cen, itp.) – przypisanym do miejsca zamieszkania osoby, której umowa podlega ocenie.

Słabością rozprawy jest, moim zdaniem, zastosowanie ograniczonego, w zasadzie w każdym artykule tego samego, zestawu modeli, mimo świadomości Doktoranta, uzewnętrznionej w przeglądach literatury zamieszczonych w artykułach, o wykorzystywaniu w podobnych badaniach także innych nowoczesnych modeli

uczenia maszynowego, w tym algorytmów *baggingu* (np. lasy losowe) czy *boostingu* (liczne algorytmy wzmacnianych drzew). Za słabość można również uznać porównywanie rzeczy nieporównywalnych (współczynniki regresji vs miary ważności zmiennych z drzewa regresyjnego), mimo świadomości istnienia metod interpretowalnego uczenia maszynowego. Na podkreślenie zasługuje to, że w każdej z ocenianych prac znajduje się bardzo dobry przegląd literatury oraz omówienie nowatorskich elementów zawartych w danym artykule. Otrzymane wyniki empiryczne są w każdym przypadku starannie omówione i zinterpretowane. Badania zaprezentowane przez Doktoranta są bardzo obszerne, a przy tym dobrze zaprojektowane i przeprowadzone. Dlatego niezależnie od przedstawionych powyżej uwag czy wątpliwości, uważam przedstawione badania za wartościowe i wnoszące nowy wkład do wiedzy na temat modelowania i prognozowania wartości LGD. Składają się na oryginalne rozwiązanie problemu modelowania wartości LGD.

Ocena rozprawy od strony formalnej i redakcyjnej

Przedstawiona rozprawa składa się z czterech artykułów opublikowanych w języku angielskim w punktowanych czasopismach naukowych. Strona edycyjna artykułów nie budzi zastrzeżeń. Natomiast konstrukcja wstępu (również napisanego w języku angielskim) powoduje pewien niedosyt. Wskazane byłoby wyodrębnienie osobnej części poświęconej celom badania, hipotezom badawczym i ich omówieniu (umotywowaniu), czego w pracy zabrakło. Zabrakło także krótkiego łącznego podsumowania dysertacji. Nieco rażą również liczne drobne błędy językowe pozostałe we wstępie.

Chciałbym wyraźnie podkreślić, że żadne ze wskazanych w recenzji potknięć ani elementów dyskusyjnych nie miało wpływu na moją ogólną ocenę tej bardzo dobrej rozprawy doktorskiej. Jestem pod wrażeniem obszerności i zaawansowania przeprowadzonych badań. Uważam, że wyniki są interesujące i znacznie rozszerzają naszą wiedzę na temat modelowania wielkości straty z tytułu niewykonania zobowiązania. O wartości rozprawy decydują aktualność i znaczenie tematyki, zakres badania oraz zaawansowane narzędzia ekonometryczne.

Konkluzja

Podsumowując przedstawione oceny i uwagi, stwierdzam, że recenzowana rozprawa doktorska spełnia warunki ustawowe. Pomimo sformułowanych przeze mnie uwag i zastrzeżeń mogę stwierdzić, że Autor dobrze poradził sobie z podjętym wyzwaniem naukowym. Zaprezentowane wyniki są ważne dla rozumienia opisywanych przez niego zjawisk. Recenzowana rozprawa jest opracowaniem oryginalnym, dojrzałym, świadczącym o dobrym opanowaniu warsztatu ekonomisty.

W związku z tym wnoszę o przyjęcie rozprawy i dopuszczenie mgr. Wojciecha Starosty do publicznej obrony.

Piotr Wójcik