METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XXIV/4, 2023, s. 229 – 241

SELEKCJA ZMIENNYCH METODAMI STATYSTYCZNYMI I UCZENIA MASZYNOWEGO. PORÓWNANIE PODEJŚĆ NA PRZYKŁADZIE DANYCH FINANSOWYCH

Urszula Grzybowska https://orcid.org/0000-0001-7342-5382
Marek Karwański https://orcid.org/0000-0001-5192-7920
Instytut Informatyki Technicznej
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
e-mail: urszula_grzybowska@sggw.edu.pl; marek_karwanski@sggw.edu.pl

Streszczenie: Zgodnie z nowymi dyrektywami międzynarodowego nadzoru finansowego (MSSF9) banki powinny przyjrzeć się nowemu zestawowi narzędzi analitycznych, takich jak uczenie maszynowe. Wprowadzenie tych metod do praktyki bankowej wymaga przeformułowania celów biznesowych, zarówno w zakresie trafności przewidywań, jak i definicji czynników ryzyka. W artykule porównano metody selekcji zmiennych i przypisania „ważności" w modelach statystycznych i algorytmicznych. Obliczenia przeprowadzono na przykładzie klasyfikacji danych finansowych. Na wybranych zbiorach zmiennych porównano skuteczność różnych algorytmów uczenia maszynowego. Wyniki analiz wskazują na potrzebę rewizji koncepcji „ważności" zmiennej, tak aby nie była ona zależna od struktury modelu.

Słowa kluczowe: selekcja zmiennych, uczenie maszynowe, ważność zmiennych

JEL classification: C45, C52, C55

WSTĘP

Analizy ilościowe w finansach bardzo często korzystają z modeli regresyjnych, zarówno w podejściu indywidualnym, jak i zagregowanym. Trzy główne cele modeli to: dokładne przewidzenie wartości wyniku w oparciu o zestaw predyktorów, wyjaśnienie różnic w wartościach wyników poprzez różnice w zmiennych objaśniających, opisanie związku między zmiennymi zależnymi i niezależnymi.

https://doi.org/10.22630/MIBE.2023.24.4.18

Zmienne niezależne są wybierane tak, aby reprezentowały spodziewane wpływy na podstawie teorii, wcześniejszych badań i kontekstu lokalnego (w czasie i przestrzeni). W tradycyjnym podejściu przy interpretacji regresji nacisk kładzie się na wielkość i istotność statystyczną współczynników dla zmiennych niezależnych. W modelach algorytmicznych wprowadza się pojęcie ważności zmiennych (importance). Ważności zmiennych to wartości liczbowe przypisywane zmiennym w oparciu o to, jak przydatne są one w przewidywaniu zmiennej docelowej. Istnieje wiele sposobów pomiaru ważności cech. Inaczej wyznaczamy ważność zmiennych w oparciu o metody statystyczne (testowanie hipotez statystycznych), a inaczej w modelach uczenia maszynowego. Osobnym zagadnieniem jest ocena własności predykcyjnych zmiennych w modelach sieci neuronowych. Różne podejścia do tego problemu przedstawiono np. w [Olden 2004]. Uwzględnienie nadmiernej liczby zmiennych w modelu nieznacznie zwiększa dopasowanie modelu. Niestety, takie dopasowanie nie jest pożądane z punktu widzenia biznasowego ponieważ prowadzi do modeli bardziej złożonych. Wprowadzenie modeli algorytmicznych przedstwia inną filozofię, ponieważ miary dopasowania modeli statystycznych i algorytmicznych opierają się na różnych założeniach.

W pracy przedstawiamy zastosowanie różnych metod do selekcji zmiennych na przykładzie zbioru danych z 174 zmiennymi. Celem naszym jest przedstawienie i zastosowanie w kontekście ekonomicznym współczesnych metod selekcji zmiennych. Na danych finansowych pokazujemy zastosowanie metod takich jak LASSO[1], miara informacji wzajemnej czy rekurencyjna eliminacja cech oraz wyznaczanie ważności zmiennych w modelach algorytmicznych: Lasów losowych, Boostingu gradientowego, XGBooost oraz stosunkowo mało jeszcze popularnej metody ważności cech w siecich neuronowych. Dla otrzymanych zbiorów zmiennych porównujemy skuteczność metod algorytmicznych oraz regresji logistycznej. W pierwszym rozdziale pracy przedstawiamy różne sposoby doboru zmiennych do modelu wraz z przeglądem literatury dotyczącej tego zagadnienia. W drugiej części opisujemy dane. W trzecim rozdziale przedstawimy wyniki obliczeń dotyczące zarówno wyboru zmiennych jak i sprawdzenia działania modeli uczenia maszynowego na wyróżnionych zbiorach zmiennych. Ostatni rozdział poświęcony jest wnioskom i podsumowaniu.

Obliczenia wykonano w Python ver. 3.9.

SELEKCJA ZMIENNYCH

Selekcja zmiennych związana z redukcją wymiaru danych ma kluczowe znaczenie. Pozwala na wyeliminowanie zmiennych nieistotnych, zbędnych, pozwala uniknąć przetrenowania modelu, zwiększa prędkość obliczeniową i umożliwia lepszę interpretację wyników. Jest wiele metod selekcji zmiennych a ich przegląd

[1] Metoda wprowadzona w latach 80-tych, zwana także regularyzacją L1. Skrót oznacza Least Absolute Shrinkage and Selection Operator.

można znaleźć w wielu publikacjach (zob. [Li i in. 2017; Pudjihartono i in. 2022; Jia i in. 2022; Zebari i in. 2020]). Zwyczajowo dzielimy je na metody związane z modelem danych tzw. metody wrapper (np. rekurencyjna eliminacja cech, metody heurystyczne) lub metody embedded (Lasy losowe, LASSO, Regresja grzbietowa) [Lal i in. 2006] oraz niezależne od modelu metody filtrowania [Sánchez-Maroño i in. 2007; Hopf 2021] np. oparte na korelacji zmiennych czy mierze informacji wzajemnej (MI) [Vergara 2014; Gajowniczek i in. 2022].

Modelowanie statystyczne

Testy hipotez są najpopularniejszym kryterium doboru zmiennych w praktycznych problemach modelowania statystycznego. Testowanie iteracyjne modeli wykonywane jest poprzez algorytmy selekcji zmiennych do przod (forward selection) lub eliminacji wstecz (backward selection), w zależności od tego, czy zaczyna się od modelu pustego, czy od modelu ze wszystkimi zmiennymi, które mogą być brane pod uwagę. Podczas gdy kryteria istotności są zwykle stosowane w celu włączenia lub wyłączenia zmiennych z modelu, kryteria informacyjne koncentrują się na wyborze modelu z zestawu wiarygodnych modeli. Uwzględnienie większej liczby zmiennych w modelu zwiększa dopasowanie modelu. Niestety takie dopasowanie nie zawsze jest pożądane, gdyż prowadzi do zwiększenia błędu dopasowania. Opracowano kryteria informacyjne w celu uniknięcia tego pozornego efektu dopasowania prowadzącego do wyboru bardziej złożonych modeli. Jako kryteria informacyjne używane są np. statystyki AIC bądź BIC.

Wybór modelu można również przeprowadzić, stosując strategię opartą na tzw. operatoratorze regularyzacji (LASSO), która polegają na nałożeniu dodatkowych warunków na funkcję błędu przy wyliczaniu współczynników regresji [Hastie i in 2008; Hastie i in 2015, str. 32]. Modele LASSO są szeroko stosowane w wielowymiarowych problemach. Współczynniki regresji oszacowane w procedurach LASSO są obciążone, ale mogą mieć mniejszy średni kwadratowy błąd całkowity niż przy konwencjonalnym oszacowaniu. Ze względu na obciążenie ich interpretacja w modelach wyjaśniających lub opisowych jest trudna, a przedziały ufności oparte na procedurach ponownego próbkowania, takich jak bootstrap, nie osiągają deklarowanego poziomu nominalnego [Taylor, Tibshirani 2015].

Wielu autorów podkreśla znaczenie wykorzystywania wiedzy merytorycznej przy doborze zmiennych. W niniejszej pracy, sposób tworzenia generycznego zestawu zmiennych powoduje, że nie musi się uwzględniać empirycznego powiązania zmiennych ze zmienną celu. Innymi słowy można założyć, że merytoryczna wartość wszystkich zmiennych jest taka sama.

Modelowanie algorytmiczne

W przypadku modeli algorytmicznych tworzone są dedykowane sposoby przypisywania ważności zmiennym predykcyjnym. [Elith i in 2008; Adler, Painsky 2022]. Wartości mierzące wagi zmiennych pomagają w interpretacji danych, a także

pozwalają na dokonanie rankingu zmiennych i ułatwiają dobór zmiennych do modelu. W ramach modeli uczenia maszynowego takich jak Drzewa decyzyjne, Lasy losowe, Boosting gradientowy czy XGBoost miary ważności zmiennych są oparte na liczbie przypadków, w których zmienna jest wybierana do podziału, ważonej kwadratową poprawką do modelu dla każdego podziału i uśrednioną dla wszystkich drzew [Elith i in 2008; Ben Jabeur i in. 2023]. W ramach modelowania algorytmicznego możemy także stosować jedną z metod opakowanych (wrapper), rekurencyjną eliminację cech (RFE). W tej metodzie uwzględnianiane są coraz mniejsze podzbiory zmiennych, najmniej ważne zmienne są usuwane z bieżącego zestawu na podstawie miary ważności, aż do osiągnięcia zadanej z góry liczby zmiennych [Kohavi, John 1997]. Dla sieci neuronowych jednym ze sposobów wyboru zmiennych jest metoda VIANN oparta na modyfikacj algorytmu Welforda [De Sa 2019]. W celu ocenienia ważność zmiennej x_s używamy miary opartej na uśrednionej wariancji zmian wag-parametrów sieci pierwszej warstwy ukrytej podłączonej do zmiennej wejścia x_s w całym procesie propagacji wstecznej. Oznacza to, że końcowy wynik ważności zmiennej będzie zależał zarówno od wag-parametrów końcowych, jak i od odchylenia ich podczas treningu. Zakłada się, że im bardziej waga $w_{(a,b)}$ połączenia (a, b) zmienia się w fazie uczenia, tym większe znaczenie węzła a w procesie predykcji. Korzystając z VIANN, musimy określić, na jakich etapach uczenia aktualizujemy wariancję. Można rozważyć kilka opcji, ze względu na iterację (po każdej partii), na epokę lub interwał zdefiniowany przez użytkownika. Dla uproszczenia w pracy aktualizujemy wariancję wag w każdej epoce.

Obok sposobu wyboru zmiennych w ramach metod uczenia maszynowego, istnieją także metody hybrydowe, które np. łączą metody filaracyjne z metodami opartymi na uczeniu maszynowym. W ostatnim czasie dla dużych zbiorów stosowane są metody heurystyczne wyboru zmiennych do modelu [Jia i in. 2022].

W pracy zastosujemy model hybrydowy. Osobno dokonamy selekcji zmiennych kategorycznych i ciagłych. Pierwszym krokiem będzie filtacja zmiennych. Dla zmiennych kategorycznych stodujemy miarę χ^2 oraz miarę informacji wzajemnej (MI), która jest znana także jako przyrost informacji lub entropia krzyżowa. Dla zmiennych ciagłych dokonamy eliminacji zmiennych silnie skorelowanych oraz zmiennych z dużym VIF[2]. Dla pozstałych zmiennych ciagłych dokonamy wyboru zmiennych metdą LASSO, Regresji grzbietowej oraz obliczając ważność zmiennych w modelach uczenia maszynowego.

OPIS DANYCH

Kredyty w rachunku bieżącym dla segmentu małych i średnich przedsiębiorstw (MŚP) są kredytami odnawialnymi ze stałym limitem pieniężnym.

[2] Wskaźnik inflacji wariancji służący do badania współliniowości zmiennych.

Informacje potrzebne do budowy modeli analitycznych zbierane są w kilku systemach informatycznych. Dla celów niniejszego artykułu wykorzystano dane jednego z polskich banków. Dane zostały poddane modyfikacjom tak aby niemożliwa była ich identyfikacja a jednocześnie zachowane relacje pomiędzy atrybutami. W tym celu stworzony został „model danych kredytowych" – zdefiniowano atrybuty transakcji/klienta, a następnie przeliczono sztuczne atrybuty stworzone na podstawie pierwotnych danych.

Na potrzeby budowy modeli parametrów ryzyka dokonano wstecznego przeglądu defaultów i oznaczono defaulty zgodnie z powszechnie stosowanymi procedurami. Zastosowano metodologię stałego horyzontu czasowego [Engelmann i Rauhmeier 2011]. Zaletami stosowania tej metodologii są: (1) rozproszenie dat referencyjnych po całym okresie zbierania danych, (2) zastosowanie wspólnego horyzontu T przyczynia się do jednorodności realizowanej straty ekonomicznej, (3) łatwość uzyskania informacji z RDS (Reference Data Set) dostępnego w banku. Metoda stałego horyzontu czasowego ma również wady, które należy rozważyć: (1) ustalony horyzont czasowy T nie pozwala na kalkulację w wielu horyzontach, (2) nie można uwzględnić bezpośrednio transakcji przeterminowanych, gdy wiek transakcji jest mniejszy niż T, (3) nie uwzględnia wszystkich istotnych informacji, ponieważ w przypadku niewykonania zobowiązania w okresie obserwacji, wykorzystuje się obserwację tylko z tego okresu, (4) nie uwzględnia się możliwości, że bieżące ekspozycje mogą nie zostać zrealizowane w dowolnym momencie następnego okresu.

W procesie zbierania danych wykorzystano informacje z kilku źródeł wykorzystywanych w banku w ciągu długiego okna czasowego (modele TTC). Aby stworzyć Tabelę Analityczną (Analytical Base Table) wykorzystano dane z następujących systemów:

1. Oceny wniosków kredytowych (Applicatins evaluations)
2. Historia klienta (Customer relationship history
3. Spłat (Repayment history)
4. Inne informacje kredytowe (Other loan informations)
5. Informacje o wykorzystaniu produktów (General information on product usage)
6. Segmentacja klienta (Customer segmentation)
7. Operacje gotówkowe (Cash based transactions)
8. Rozliczenia (Settelments)
9. Zabezpieczenia (Collarerals)
10. Ratingi BIK (Credit Bureau ratings)
11. Informacje o właścicielach (SME owners information)

Model danych kredytowych zebrano w tzw. *modelu generycznym* – zaproponowanym „z góry" układzie atrybutów, który zawierał definicje czynników ryzyka podzielone na dwie grupy: podstawowe informacje o klientach i informacje dodatkowe o kontach. Model ten został zbadany przez ekspertów biznesowych. Główną ideą modelu generycznego był wybór jak największej liczby atrybutów -

„słabych uczniów" i stworzenie odpowiedniej reguły klasyfikacji opartej na wnioskach statystycznych, a nie biznesowych. Zgodnie z tym podejściem procedura budowy modelu polegała na określeniu modelu maksymalnego, tj. modelu z największą liczbą zmiennych, a następnie utworzenie finalnego modelu poprzez usunięcie zmiennych z modelu maksymalnego.

Ostatecznie zebrane dane zawierały 6590 obserwacji wśród których było 1014 defaultów. Każda obserwacja opisywana była przez 174 atrybuty (zmienne). W zbiorze pierwszych 14 zmiennych stanowiły zmienne mierzone w skali nominalnej pozostałe mierzono w skali interwałowej. Zmienną targetową była zmienna binarna default = 1, non-default = 0.

WYNIKI OBLICZEŃ

Selekcja zmiennych kategorycznych

W pierwszym kroku wyboru zmiennych usunięto zmienne o zbyt dużej liczbie kategorii (X_4- 31 kategorii, X_7- 55 kategorii), zmienną o 2 kategoriach, gdzie druga liczyła 1 obserwację (X_6). Dla pozostałych zmiennych kategorie z małą liczbą obserwacji zostały scalone. Ostatecznie wykorzystano 11 zmiennych nominalnych. Dla tego zbioru zbudowano model regresji logistycznej z selekcją zmiennych metodą krokową w przód. Trzy zmienne nie zostały włączone do modelu. Istotne okazały się zmienne X_{11}, X_2, X_{12}, X_{13}, X_5 i X_{10}. Następnie dla 11 zmiennych obliczono miarę informacji wzajemnej, która wyróżniła zmienne X_{11}, X_5, X_{12} i X_1. Z kolei miara oparta na teście chi2 wskazała, że nie ma zależności między zmienną y a zmiennymi X_3, X_8 i X_{13}. Największy wpływ na y mają zmienne: X_{11}, X_{14} i X_{12}. Metody uczenia maszynowego pozwalają na wyznaczenie ważności zmiennych (zobacz tabela 1).

Tabela 1. Ranking zmiennych kategorycznych uzyskany metodami uczenia maszynowego

	RF		XGBoost		GB	
1	X_{11}	0,4684	X_{11}	0,5005	X_{11}	0,7232
2	X_2	0,0964	X_2	0,1249	X_2	0,1280
3	X_3	0,0880	X_{12}	0,0751	X_{12}	0,0515
4	X_5	0,0801	X_1	0,0428	X_{13}	0,0191
5	X_8	0,0620	X_{14}	0,0418	X_5	0,0187
6	X_9	0,0479	X_{10}	0,0410	X_{10}	0,0173
7	X_{10}	0,0446	X13	0,0384	X_9	0,0158
8	X_{12}	0,0440	X_9	0,0382	X_{14}	0,0141
9	X_{13}	0,0387	X_5	0,0365	X_8	0,0056
10	X_{14}	0,0210	X_8	0,0309	X_1	0,0041
11	X_1	0,0089	X_3	0,0300	X_3	0,0028

Źródło: opracowanie własne

Wszystkie 3 rozważane metody wskazały zmienne x_{11} i x_2 jako najważniejsze. Kolejną pozycję zajmowała zmienna X_{12} dla Boostingu gradientowego (GB) i algorytmu XGBoost, zaś Lasy losowe (RF) wskazały jako najważniejszą zmienną X_{11}.

Wybór zmiennych ciągłych

W bazie występowało 160 zmiennych typu ciągłego. W pierwszym kroku wyeliminowano zmienne skorelowane, przyjmując jako wartość progową 0,9. Usunięto także zmienne z wysokim VIF, przyjmując jako próg wartość 20. Pozostały 73 zmienne. Usunięto zmienne quasi-stałe. Do dalszej analizy pozostało 68 zmiennych. Dalszej selekcji zmiennych dokonano metodami uczenia maszynowego a także za pomocą regularyzacji L1 (LASSO) i L2 (regresja grzbietowa, Ridge) w regresji logistycznej, które są metodami typu osadzonego (embedded methods). Wyznaczono ważność zmiennych metodami uczenia maszynowego tj. przy użyciu Lasów losowych (RF), algorytmu XGBoost (XGB), Boostingu gradientowego (GB) oraz sztucznych sieci neuronowych (ANN). Wartości miar ważności dla najlepszych 11 zmiennych przedstawiono w tabeli 2.

Tabela 2. Ranking zmiennych uzyskany metodami uczenia maszynowego

	RF		XGBoost		GB		ANN	
1	X_{152}	0,29	X_{152}	0,879	X_{152}	0,993	X_{152}	1,000
2	X_{153}	0,11	X_{74}	0,009	X_{103}	0,004	X_{150}	0,721
3	X_{149}	0,08	X_{50}	0,008	X_{153}	0,001	X_{147}	0,398
4	X_{161}	0,05	X_{125}	0,007	X_{50}	0,001	X_{38}	0,252
5	X_{147}	0,05	X_{103}	0,007	X_{66}	0,001	X_{153}	0,175
6	X_{118}	0,05	X_{39}	0,007	X_{38}	0,000	X_{97}	0,163
7	X_{154}	0,03	X_{47}	0,006	X_{44}	0,000	X_{145}	0,142
8	X_{129}	0,03	X_{90}	0,005	X_{47}	0,000	X_{143}	0,132
9	X_{103}	0,03	X_{153}	0,005	X_{150}	0,000	X_{57}	0,108
10	X_{39}	0,03	X_{38}	0,005	X_{46}	0,000	X_{120}	0,107
11	X_{76}	0,03	X_{91}	0,005	X_{147}	0,000	X_{160}	0,105

Źródło: opracowanie własne

Każda ze stosowanych metod uczenia maszynowego wyróżniła inne zmienne, choć niektóre zmienne np. X_{152} i X_{153} powtarzają się w każdym rankingu. Obliczenie korelacji rang dla wszystkich 68 zmiennych zamieszczone w tabeli 3 pokazuje brak korelacji między rangami zmiennych, których kolejność, poprzez przyporządkowanie ważności, została wyznaczona w poszczególnych modelach uczenia maszynowego. Wizualizacja związku między rangami zamieszczona została na rysunku 1.

Tabela 3. Współczynniki korelacji rang Kendalla wraz z odchyleniami standardowymi

	FR	XGBoost	GB	ANN
FR	1,0000	0,1624	-0,1229	0,1484
		0,0502	0,1383	0,0736
XGBoost	0,1624	1,0000	0,2388	-0,1062
	0,0502		0,0040	0,2002
GB	-0,1229	0,2388	1,0000	-0,0667
	0,1383	0,0040		0,4211
ANN	0,1484	-0,1062	-0,0667	1,0000
	0,0736	0,2002	0,4211	

Źródło: opracowanie własne

Następnie zbudowano model Regresji logistycznej (LR) dla wszystkich 68 zmiennych i dla tego modelu zastosowano regularyzację L1 (LASSO) z parametrem c = 0,01. Regularyzacja wyróżniła następujące zmienne jako istotne: X_{24}, X_{25}, X_{32}, X_{43}, X_{44}, X_{45}, X_{46}, X_{49}, X_{59}, X_{66}, X_{74}, X_{76}, X_{84}, X_{87}, X_{103}, X_{114}, X_{115}, X_{116}, X_{118}, X_{129}, X_{133}, X_{134}, X_{136}, X_{143}, X_{147}, X_{149}, X_{150}, X_{152}, X_{153}, X_{160}, X_{162}, X_{171}.

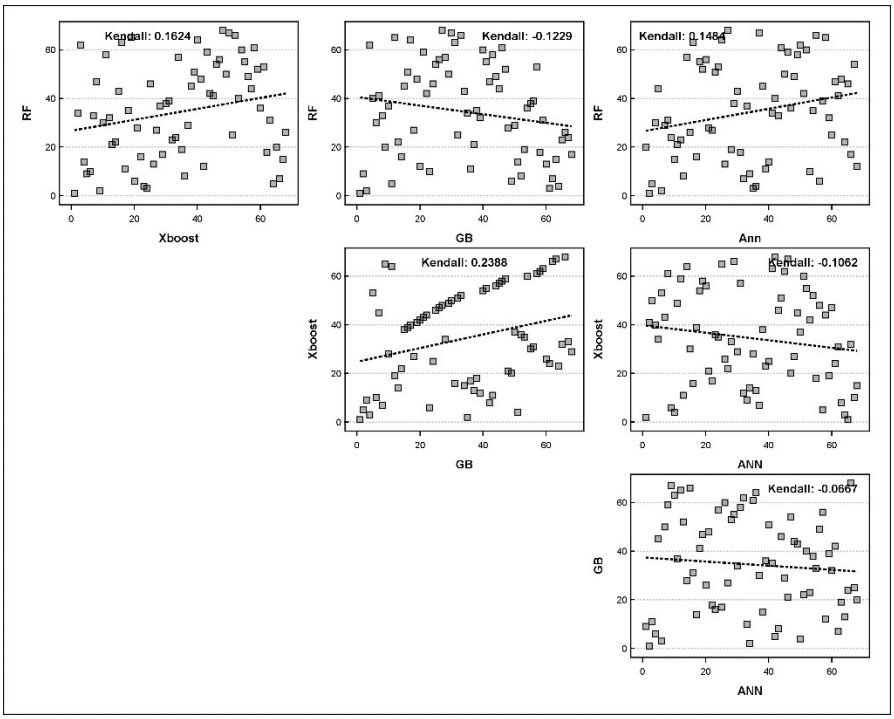
Regularyzacja L2 (grzbietowa) z parametrem c = 0,05 wskazała następujące zmienne jako istotne: X_{44}, X_{45}, X_{46}, X_{59}, X_{66}, X_{87}, X_{103}, X_{114}, X_{118}, X_{129}, X_{147}, X_{149}, X_{150}, X_{152}, X_{153}, X_{162}. Te same zmienne uzyskano dla regularyzacji L2 z parametrem c = 0,01.

Przeprowadzono także rekurencyjną eliminację cech (RFE). Jest to metoda kosztowana obliczeniowo. W pierwszej kolejności przeprowadzono rekurencyjną eliminację cech biorąc jako wyjściowy zbiór wszystkie 68 zmiennych. Przy wyborze 5 zmiennych stosując jako algorytm drzewo decyzyjne, otrzymano zbiór zmiennych X_{38}, X_{50}, X_{103}, X_{152}, X_{153}. Las losowy wskazał zbiór X_{118}, X_{149}, X_{152}, X_{153} oraz X_{161}. W sytuacji, gdy RFE przeprowadzono dla zbioru 16 zmiennych wyróżnionych przez regularyzację L2, Las losowy wskazał zmienne X_{118}, X_{129}, X_{149}, X_{152} i X_{153}, zaś Boosting gradientowy (GB) wyróżnił zmienne X_{87}, X_{103}, X_{118}, X_{152}, X_{153}. Zauważmy, że choć wyróżnione przez różne sposoby selekcji zbiory są różne, to niektóre zmienne powtarzają się, np. X_{103}, X_{152}, X_{153}. Powtarzające się zmienne oznaczają odpowiednio: liczbę dni do pierwszego zamknięcia (zapadalności) ACA na rachunku OVERDRAFT, wpływy /zaangażowanie kredytowe (%) - wartość średnia z ostatnich - 2 lat, wzrost sprzedaży (%) - wartość średnia z ostatnich - 2 lat[3].

W tabeli 4 przedstawiono zbiory zmiennych uzyskane różnymi metodami, na których następnie przeprowadzono klasyfikację z wykorzystaniem wybranych metod uczenia maszynowego (LR, RF, GB oraz XGB). Zbiory w prawej kolumnie różnią się od zbiorów w kolumnie lewej dodaniem zmiennych nominalnych X_2 i X_{11}.

[3] Dokładny opis zmiennych wykorzystanych w analizie jest dostępny u autorów.

Rysunek 1. Wykres zależności między rangami zmiennych w różnych modelach



Źródło: opracowanie własne

Tabela 4. Wyróżnione zbiory zmiennych użyte w analizach porównawczych

A	X_{38}, X_{50}, X_{103}, X_{152}, X_{153}	A1	X_2, X_{11}, X_{38}, X_{50}, X_{103}, X_{152}, X_{153}
B	X_{118}, X_{149}, X_{152}, X_{153}, X_{161}	B1	X_2, X_{11}, X_{118}, X_{149}, X_{152}, X_{153}, X_{161}
C	X_{118}, X_{129}, X_{149}, X_{152}, X_{153}	C1	X_2, X_{11}, X_{118}, X_{129}, X_{149}, X_{152}, X_{153}
D	X_{87}, X_{103}, X_{118}, X_{152}, X_{153}	D1	X_2, X_{11}, X_{87}, X_{103}, X_{118}, X_{152}, X_{153}
E	X_{118}, X_{147}, X_{149}, X_{152}, X_{153}, X_{161}	E1	X_2, X_{11}, X_{118}, X_{147}, X_{149}, X_{152}, X_{153}, X_{161}
F	X_{39}, X_{50}, X_{74}, X_{103}, X_{125}, X_{152}	F1	X_2, X_{11}, X_{39}, X_{50}, X_{74}, X_{103}, X_{125}, X_{152}
G	X_{50}, X_{66}, X_{103}, X_{152}, X_{153}	G1	X_2, X_{11}, X_{50}, X_{66}, X_{103}, X_{152}, X_{153}
H	X_{38}, X_{97}, X_{143}, X_{145}, X_{147}, X_{150}, X_{152}, X_{153}	H1	X_2, X_{11}, X_{38}, X_{97}, X_{143}, X_{145}, X_{147}, X_{150}, X_{152}, X_{153}

Źródło: opracowanie własne

W tabeli 5 przedstawiono miarę jakości klasyfikacji, trafność, która oznacza odsetek poprawnych klasyfikacji dokonywanych przez model LR, RF, GB lub XGB i wyróżnionych zbiorów zmiennych ciągłych. W tabeli 6 przedstawiono trafność wyznaczoną na zbiorach zmiennych uzupełnionych o zmienne nominalne X_2 i X_{11}.

Tabela 5. Wyniki trafności modeli wraz z odchyleniami standardowymi

	A	B	C	D
LR	0,9116 (0,0312)	0,9512 (0,0100)	0,9267 (0,0097)	0,9021 (0,0103)
RF	0,9994 (0,0008)	0,9994 (0,0008)	0,9992 (0,0009)	0,9992 (0,0009)
GB	0,9990 (0,0009)	0,9990 (0,0009)	0,9990 (0,0009)	0,9992 (0,0009)
XGB	0,9990 (0,0009)	0,9991 (0,0009)	0,9992 (0,0009)	0,9988 (0,0012)
	E	F	G	H
LR	0,9531 (0,0099)	0,7813 (0,0364)	0,8810 (0,0097)	0,8467 (0,0129)
RF	0,9982 (0,0012)	0,9990 (0,0009)	0,9992 (0,0009)	0,9990 (0,0009)
GB	0,9990 (0,0009)	0,9992 (0,0009)	0,9994 (0,0008)	0,9990 (0,0009)
XGB	0,9992 (0,0009)	0,9990 (0,0009)	0,9992 (0,0009)	0,9992 (0,0009)

Źródło: opracowanie własne

Tabela 6. Wyniki trafności modeli wraz z odchyleniami standardowymi

	A1	B1	C1	D1
LR	0,9387 (0,0304)	0,9673 (0,0090)	0,9558 (0,009488)	0,9472 (0,0080)
RF	0,9990 (0,0009)	0,9994 (0,0008)	0,9990 (0,000948)	0,9992 (0,0009)
GB	0,9994 (0,0008)	0,9992 (0,0009)	0,9992 (0,000929)	0,9992 (0,0009)
XGB	0,9994 (0,0008)	0,9992 (0,0009)	0,9992 (0,000929)	0,9994 (0,0008)
	E1	F1	G1	H1
LR	0,9669 (0,0085)	0,7813 (0,0364)	0,9275 (0,0224)	0,8467 (0,0129)
RF	0,9988 (0,0009)	0,9996 (0,0007)	0,9994 (0,0008)	0,9992 (0,0009)
GB	0,9992 (0,0009)	0,9994 (0,0008))	0,9994 (0,0008)	0,9992 (0,0009)
XGB	0,9992 (0,0009)	0,9990 (0,0009)	0,9996 (0,0007)	0,9992 (0,0009)

Źródło: opracowanie własne

Otrzymane wyniki pokazują, że wybór zestawu zmiennych nie wpływa znacząco na jakość klasyfikacji mierzoną trafnością dla takich metod uczenia maszynowego jak Lasy losowe (RF), Boosting gradientowy (GB), czy algorytm XGB. Dla tych metod włączenie do zbioru zmiennych kategorycznych nie poprawiło jakości klasyfikacji, która jest bardzo wysoka. Dla Regresji logistycznej (LR) obserwujemy zależność jakości klasyfikacji od użytego zbioru zmiennych. Dodatkowo, dla wszystkich zestawów zmiennych z wyjątkiem F oraz H, włączenie zmiennych kategorycznych poprawiło znacząco jakość klasyfikacji dla Regresji logistycznej.

PODSUMOWANIE

W podejściu statystycznym modele regresyjne w naturalny sposób pozwalają definiować ważność zmiennych poprzez wartości parametrów/wag (przy zastosowaniu transformacji standaryzującej) oraz wykorzystują statystyczną istotność pozwalającą wyeliminować parametry, których ważność jest bliska zeru. Dzięki temu interpretacja ważności jest intuicyjna i zgodna z fizyczną interpretacją teoretyczną. Coraz częściej zaczyna się wprowadzać do modelowania procedury regularyzacji, które są mniej oczywiste, ale pozwalają na bardziej elastyczne podejście do procesu selekcji zmiennych zastępując regułę 0/1 (jest/nie ma) regułą włączania/wyłączania częściowego. Warto zauważyć, że takie podejście ma na celu identyfikację związku przyczynowego będącego głównym przedmiotem badań. Poznanie i zrozumienie przyczyny pozwala na opracowanie skutecznego sposobu reagowania w oparciu o model.

W modelach algorytmicznych proces selekcji zmiennych zależy od wewnętrznej struktury modelu. I tak np. w przypadku Lasów losowych powiązany jest z funkcją zanieczyszczenia co powoduje, że interpretacja wyników zależy bardzo silnie od wewnętrznych algorytmów stosowanych w modelu i nie musi odpowiadać znaczeniu jakie intuicyjnie nadajemy pojęciu ważności zmiennej. Jeszcze silniej widać to w przypadku modeli sztucznych sieci neuronowych, w których ważność zmiennej powiązana jest ze ścieżką zmian wewnętrznych parametrów w procesie trenowania modelu.

Wyniki przeprowadzonych analiz dowodzą, że badając strukturę zależności między czynnikami przy pomocy modeli regresyjnych bardzo silnie uzależniamy jej interpretację od używanych narzędzi.

Należy podkreślić, że nie ma jednej optymalnej metody selekcji zmiennych. Każdy z przedstawionych sposobów wyróżnił inny zestaw. Pewne zmienne pojawiały się w każdym z nich. Niestety, korelacja (Kendalla) pomiędzy wynikami różnych metod nie pozwala na potwierdzenie, że ważność zmiennej jest cechą zewnętrzną przypisaną do niej. Widać potrzebę dokonania rewizji pojęcia „ważność" tak, aby nie była ona zależna od struktury modelu.

Wyniki prezentowane w niniejszej pracy wskazują, że powinniśmy zmienić punkt widzenia: zamiast opisywać ważność poszczególnych zmiennych należy interpretować grupy zmiennych czyli inaczej rating zmiennych. Oczywiście zmienia to sposób interpretacji i wykorzystania modeli, ale pozwala lepiej uchwycić związek między zmiennymi zależnymi i niezależnymi.

BIBLIOGRAFIA

Adler A. I., Painsky A. (2022) Feature Importance in Gradient Boosting Trees with Cross-Validation Feature Selection. Entropy, 24(5), 687. https://doi.org/10.3390/e24050687.

Ben Jabeur S., Stef N., Carmona P. (2023) Bankruptcy Prediction using the XGBoost Algorithm and Variable Importance Feature Engineering. Comput Econ, 61, 715-741. https://doi.org/10.1007/s10614-021-10227-1.

Breiman L. (2001) Statistical Modeling: The Two Cultures. Statistical Science, 16(3), 199-215.

De Sa C.R. (2019) Variance-Based Feature Importance in Neural Networks. [in:] Kralj Novak P., Šmuc T., Džeroski S. (eds) Discovery Science, Lecture Notes in Computer Science, 11828, Springer, Cham. https://doi.org/10.1007/978-3-030-33778-0_24.

Engelmann B., Rauchmeier R. (2011) The Basel II: Risk Parameters. Estimation, Validation, Stress Testing - with Applications to Loan Risk Management. Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-16114-8.

Elith J., Leathwick J. R. and Hastie T. (2008) A Working Guide to Boosted Regression Trees. Journal of Animal Ecology, 77, 802-813. https://doi.org/10.1111/j.1365-2656.2008.01390.x.

Gajowniczek K., Wu J., Gupta S., Bajaj C. (2022) HOFS: Higher Order Mutual Information Approximation for Feature Selection in R. SoftwareX, 19, 1-9. https://doi.org/10.1016/j.softx.2022.101148.

Hastie T., Tibshirani R., Friedman J. (2008) The Elements of Statistical Learning (2nd ed.), Springer.

Hastie T., Tibshirani R., Wainwright M. (2015) Statistical Learning with Sparsity: The Lasso and Generalizations. New York Chapman & Hall/CRC. https://doi.org/10.1201/b18401.

Hopf K., Sascha R. (2021) Filter Methods for Feature Selection in Supervised Machine Learning Applications - Review and Benchmark. arXiv preprint arXiv:2111.12140, 2021.

Jia W., Sun M., Lian J. et al. (2022) Feature Dimensionality Reduction: A Review. Complex Intell. Syst., 8, 2663-2693. https://doi.org/10.1007/s40747-021-00637-x.

Kohavi R, John G. H. (1997) Wrappers for Feature Subset Selection. Artificial Intelligence, 97(1-2), 273-324. https://doi.org/10.1016/S0004-3702(97)00043-X.

Lal T. N., Chapelle O., Weston J., Elisseeff A. (2006) Embedded Methods. [in:] Guyon I., Nikravesh M., Gunn S., Zadeh L. A. (eds) Feature Extraction. Studies in Fuzziness and Soft Computing, 207, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-35488-8_6.

Li J., Cheng K., Wang S., Morstatter F., Trevino R. P., Tang J., Liu H. (2017) Feature Selection: A Data Perspective. ACM Computing Surveys, 50(6), Article 94, 1-45. https://doi.org/10.1145/3136625.

Olden J., Joy M., Death R. (2004) An Accurate Comparison of Methods for Quantifying Variable Importance in Artificial Neural Networks using Simulated Data. Ecological Modelling, 178(3-4), 389-397. https://doi.org/10.1016/j.ecolmodel.2004.03.013.

Pudjihartono N., Fadason T., Kempa-Liehr A. W., O'Sullivan J. M. (2022) A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Front. Bioinform, 2(927312). doi: 10.3389/fbinf.2022.927312.

Sánchez-Maroño N., Alonso-Betanzos A., Tombilla-Sanromán M. (2007) Filter Methods for Feature Selection – A Comparative Study. [in:] Yin H., Tino P., Corchado E., Byrne W., Yao X. (eds) Intelligent Data Engineering and Automated Learning - IDEAL 2007. IDEAL 2007. Lecture Notes in Computer Science, 4881, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-77226-2_19.

Taylor J., Tibshirani R. J. (2015) Statistical Learning and Selective Inference. Proc Natl Acad Sci U S A, 112(25), 7629-34. doi: 10.1073/pnas.1507583112.

Vergara J. R., Estévez P.A. (2014) A Review of Feature Selection Methods Based on Mutual Information. Neural Comput & Applic, 24, 175-186. https://doi.org/10.1007/s00521-013-1368-0.

Zebari R., Abdulazeez A., Zeebaree D., Zebari D., Saeed J. (2020) A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. Journal of Applied Science and Technology Trends, 1(2), 56-70. https://doi.org/10.38094/jastt1224.

VARIABLE SELECTION BY STATISTICAL AND MACHINE LEARNING METHODS. COMPARISON OF APPROACHES USING FINANCIAL DATA AS AN EXAMPLE

Abstract: In line with new international financial supervision directives (IFRS9), banks should look at a new set of analytical tools, such as machine learning. The introduction of these methods into banking practice requires reformulation of business goals, both in terms of the accuracy of predictions and the definition of risk factors. The article compares methods for selecting variables and assigning "importance" in statistical and algorithmic models. The calculations were carried out using the example of financial data classification. The effectiveness of various machine learning algorithms on selected sets of variables was compared. The results of the analyzes indicate the need to revise the concept of the "importance" of a variable so that it does not depend on the structure of the model.

Keywords: variable selection, machine learning, variable importance

JEL classification: C45, C52, C55