Uniwersytet Warszawski
Wydział Nauk Ekonomicznych

Modelowanie ryzyka kredytowego dla klientów indywidualnych w oparciu o dane aplikacyjne

Praca zaliczeniowa
z przedmiotu Modelowanie ryzyka kredytowego -
- budowa kart scoringowych w R

Warszawa, luty 2021

Abstrakt

Głównym celem poniższej pracy jest przedstawienie procesu budowania modelu ryzyka kredytowego zgodnie z najlepszymi praktykami zaprezentowanymi na zajęciach. Do badania wykorzystano dane aplikacyjne klientów detalicznych, które pochodzą ze strony Kaggle. Z bazy składającej się początkowo z 122 zmiennych objaśniających, po przeprowadzeniu wstępnej analizy pozostawiono 17 najlepszych. Na ich podstawie zbudowano model regresji logistycznej oraz model lasu losowego, które następnie poddano ewaluacji i porównano ze sobą wzajemnie.

Wstęp

Ryzyko kredytowe jest pojęciem oznaczającym prawdopodobieństwo niewypełnienia przez kredytobiorcę zobowiązań i warunków umowy, co prowadzi do straty finansowej kredytodawcy. W celu maksymalnego zredukowania tego zagrożenia kluczową kwestią jest odpowiednie zarządzanie ryzykiem kredytowym. Jednym z takich działań jest ocena zdolności kredytowej, którą bada się pod względem formalno-prawnym oraz merytorycznym. Przez lata banki stworzyły wiele sposobów na ocenę zdolności kredytowej, które można podzielić na (G. Migut 2003):

- Metody opisowe - opierają się na sytuacji ekonomiczno-finansowej klienta. Pod uwagę bierze się wiele wskaźników ekonomicznych (bilans, stan zadłużenia itp.) uzyskanych z analizy oraz z oceny personalnej. Następnie banki wykorzystując metodę punktową, która polega na określeniu wartości danego wskaźnika pewną oceną liczbową starają się określić poziom ryzyka dla danego kredytobiorcy.
- Metody statystyczne - polegają na ocenie zdolności kredytowej danego klienta na podstawie danych opisujących wcześniejszych klientów. Ważną sprawą w tym podejściu jest odpowiedni dobór zmiennych i zbudowanie najlepszego modelu, który będzie prawidłowo klasyfikował, czy dany klient spłaci kredyt.

Proces budowania modelu składa się z kilku etapów. Przede wszystkim istotną czynnością jest odpowiednie przygotowanie danych. Początkowo należy zrozumieć jak posiadane zmienne powinny wpływać na nasz wynik, co ułatwi interpretację rezultatów. Inną ważną kwestią jest poradzenie sobie z podstawowymi problemami takimi jak: braki danych, obserwacje odstające, czy uwzględnienie nieliniowych zależności. Jedną z bardziej optymalnych

metod jest dyskretyzacja, wykorzystująca techniki fine-classingu oraz coarse-classingu, które zostały użyte w przypadku tej analizy. Poprzez wydzielanie poziomów zmiennych metoda ta pozwala utrzymać jakość predykcyjną, czy rozwiązać problem braków danych. Kolejnym krokiem jest selekcja zmiennych. Ograniczenie liczby predyktorów można dokonać analizując między innymi ich zdolność dyskryminacyjną, czyli jak dokładnie dana zmienna separuje złych i dobrych klientów. Gdy powyższe kroki zostaną wykonane w należyty sposób można przejść do prognozowania prawdopodobieństwa defaultu z wykorzystaniem różnych modeli. Ze względu, że większość danych do analizy ryzyka kredytowego charakteryzuje się niezbilansowaną próbą, w której dużo więcej klientów spłaca kredyt, najczęściej wykorzystuje się techniki prognostyczne uwzględniające ten problem. Jedną z takich metod jest regresja logistyczna, która jest traktowana jako jedna z najlepszych praktyk. Jej ogromną zaletą w porównaniu do większości modeli uczenia maszynowego jest jej interpretowalność, czyli oprócz samej prognozy, można uzyskać informacje co spowodowało, że te wyniki się tak kształtują. Problem z modelami takimi jak SVM, sieci neuronowe, czy drzewa decyzyjne wiążę się z ich złożonością, która doprowadza, że mimo bardzo dobrych rezultatów prognostycznych, trudno jest wytłumaczyć danemu klientowi dlaczego jego aplikacja została odrzucona. W poniższej analizie oprócz modelu logitowego został stworzony las losowy, głównie w celu porównania otrzymanych rezultatów. Ponadto, skupiając się na regresji logistycznej, oceniono jakość uzyskanych wyników oraz na podstawie najlepszego modelu zbudowano kartę scoringową, która ułatwia interpretację biznesową. Cała analiza została przeprowadzona w programie Rstudio.

Opis danych

Dane aplikacyjne, na podstawie których przeprowadzona została analiza pochodzą z serwisu Kaggle. Baza ta składa się z 122 predyktorów oraz jednej zmiennej mówiącej o tym, czy kredyt ostatecznie został spłacony, czy nie. Plik ten posiada informacje na temat 307 511 aplikacji o kredyt, co wymaga dokładnego przeprowadzenia wstępnej analizy w celu uzyskania jak najlepszych wyników prognostycznych. W celu lepszego zrozumienia tej bazy warto przyjrzeć się poszczególnym zmiennym:

- DEF - zmienna binarna, która mówi o tym, czy dany klient spłacił kredyt(0), czy nie(1). W bazie znajduje się 8,07% defaultów, czyli 24825 osób, które nie spłaciły swojego kredytu
- LOAN_TYPEF - zmienna przyjmująca dwa poziomy CASH - aplikacja dotyczy kredytu gotówkowego, lub REVOLVING - aplikacja dotyczy kredytu odnawialnego
- SEXF - predyktor wskazujący płeć klienta, gdzie male - oznacza mężczyznę, a female - kobietę. Cztery obserwacje były brakami danych, które w tym przypadku po prostu usunięto

SEXF	**male**	**female**
Liczebność	105 059	202 448
% defaultów	10,14%	7,00%

- CARF - zmienna mówi o tym, czy klient posiada samochód(YES), czy nie (NO)

CARF	**NO**	**YES**
Liczebność	202 922	104 585
% defaultów	8,5%	7,24%

- PROPERTYF - czy klient posiada nieruchomość, czy nie?
- NUM_CHILDRENF - predyktor opisujący liczbę posiadanych dzieci. Klientów, którzy posiadają trójkę dzieci, bądź więcej połączono w jedną grupę.
- INCOME - dochód klienta
- AMT_CREDIT - wielkość zaciągniętego kredytu przez daną osobę
- AMT_ANNUITY - wielkość raty jaką klient musi spłacać
- GOOD_PRICE - cena dobra, na jaką został zaciągnięty kredyt konsumencki
- JOINT_LOANF - zmienna mówiąca kto jest współkredytobiorcą. Zamieniona została na zmienną binarną informującą, czy istnieje taka osoba czy nie.

- SOURCE_INCOMEF - źródło podstawowego dochodu klienta. Ze względu na małą liczebność połączono grupy Businessman oraz Commercial associate, a także w drugą grupę połączono zmienne Student, Unemployed, Maternity leave, oraz Pensioner.

SOURCE_INCOMEF	**Commercial associate**	**State servant**	**Unemployed**	**Working**
Liczebność	71 626	21 703	55 407	158 771
% defaultów	7,48%	5,75%	5,4%	9,59%

- EDUCATIONF - zmienna charakteryzująca wykształcenie klienta. Połączono poziom academic degree z higher education ze względu na małą liczebność tej pierwszej grupy.

EDUCATIONF	**Higher**	**Incomplete higher**	**Secondary**	**Lower secondary**
Liczebność	75 026	10 276	218 389	3 816
% defaultów	5,35%	8,48%	8,94%	10,93%

- CIVIL_STATEF - określa stan cywilny klienta. Usunięto dwie obserwacje, które były brakami danych. Posiada 5 poziomów: Civil marriage, Married, Separated, Single, Widow

CIVIL_STATEF	**Civil marriage**	**Married**	**Separated**	**Single**	**Widow**
Liczebność	29 774	196 429	19 770	45 444	16 088
% defaultów	9,94%	7,56%	8,19%	9,81%	5,82%

- HOUSING_TYPEF - zmienna określająca typ mieszkania klienta, która przyjmuje następujące poziomy: House, Rented, With parents, Municipal apartment, Office apartment, Co-op apartment

- REGION - zmienna ciągła znormalizowana opisującą zaludnienie regionu, w którym klient mieszka
- DAYS_BIRTH - wiek klienta wyrażony w dniach w sposób ujemny
- DAYS_EMPLOYED - zmienna wyrażona w dniach, mówiąca o tym jak długo poszczególna osoba jest zatrudniona na danym stanowisku.
- DAYS_REGISTRATION - ile dni przed aplikacją klient zmieniał swoją rejestrację
- DAYS_ID_CHANGE - liczba dni od ostatniej zmiany dokumentu
- CAR_AGE - zmienna opisująca wiek samochodu klienta
- EMP_MOBILE_PHONEF - informacja o tym czy dana osoba posiada telefon służbowy
- HOME_PHONEF - informacja o tym czy dana osoba posiada telefon stacjonarny
- FLAG_PHONEF - zmienna mówiąca o tym czy klient posiada telefon domowy
- EMAILF - zmienna mówiąca o tym czy dany klient posiada adres e-mail
- JOBF - predyktor opisujący stanowisko zajmowane przez aplikanta. Poziom IT staff został połączony z High skill tech staff, natomiast braki danych oznaczone jako Missing

JOBF	**Liczebność**	**% defaultów**
Accountants	9 812	4,83%
Cleaning staff	4 653	9,6%
Cooking staff	5 946	10,44%
Core staff	27 569	6,3%
Drivers	18 603	11,33%
High skill tech staff	11 906	6,17%
HR staff	563	6,39%
Laborers	55 186	10,58%
Low-skill Laborers	2 092	17,16%
Managers	21 370	6,21%
Medicine staff	8 537	6,7%
Missing	96 389	6,51%
Private service staff	2 652	6,6%

Realty agents	751	7,86%
Sales staff	32 102	9,63%
Secretaries	1 305	7,05%
Security staff	6 721	10,74%
Barmen staff	1 348	11,28%

- CNT_FAM_MEMBERSF - zmienna mówiąca o ilości osób w rodzinie. W przypadku 5 i więcej osób utworzony jeden poziom FIVE_PEOPLE+
- REGION_RATEF - ocena regionu przez klienta

REGION_RATEF	**1**	**2**	**3**
Liczebność	32 197	226 979	48 329
% defaultów	4,82%	7,89%	11,1%

- WEEKDAY_STARTF - zmienna mówiąca o tym, w którym dniu został złożony wniosek
- HOUR_START - godzina złożenia wniosku o kredyt
- ORGANIZATION_TYPEF - rodzaj organizacji, w której klient pracuje. Zmienna została pogrupowana: połączone zostały wszystkie poziomy Industry type w jedno, podobnie postąpiono z Business Entity Type, Trade: type, Transport: type. Braki zastąpiono poziomem Other

ORGANIZATION_TYPEF	**Liczebność**	**% defaultów**
Advertising	429	8,16%
Agriculture	2 454	10,47%
Bank	2 507	5,19%
Business Entity	84 527	9,12%
Cleaning	260	11,15%
Construction	6 721	11,68%

Culture	379	5,54%
Electricity	950	6,63%
Emergency	560	7,14%
Government	10 404	6,98%
Hotel	966	6,42%
Housing	2 958	7,94%
Industry	14 310	8,6%
Insurance	596	5,7%
Kindergarten	6 879	7,04%
Legal services	305	7,86%
Medicine	11 192	6,59%
Military	2 634	5,13%
Mobile	317	9,15%
Other	72 057	5,92%
Police	2 341	5,00%
Postal	2 157	8,44%
Realtor	396	10,61%
Religion	85	5,88%
Restaurant	1 811	11,71%
School	8 893	5,92%
Security	3 247	9,98%
Security Ministries	1 974	4,86%
Self-employed	38 412	10,17%
Services	1 575	6,60%
Telecom	577	7,63%
Trade	14 315	9,07%
Transport	8 990	9,66%

University	1327	4,90%

- EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3 - znormalizowana punktacja klienta pochodząca z zewnętrznych baz
- OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE - zmienne mówiące o ilości osób z otoczenia klienta, które nie spłaciły należności od 60 dni.
- DAYS_PHONE_CHANGE - liczba dni jaka minęła od ostatniej zmiany telefonu przez klienta
- DOCUMENTF - zmienna opisująca, czy klient dostarczył dokumenty czy nie. W bazie znajduje się 20 takich zmiennych binarnych.
- REG_REGION_NOT_LIVE_REGIONF – zmienna przyjmuje wartość YES, jeśli region składania wniosku jest inny od regionu zamieszkania
- REG_REGION_NOT_WORK_REGIONF – predyktor, który mówi czy region składania wniosku jest inny od regionu, w którym kredytobiorca pracuje
- LIVE_REGION_NOT_WORK_REGIONF – zmienna mówi o tym, czy region zamieszkania oraz pracy jest taki sam
- REG_CITY_NOT_LIVE_CITYF – informacja o tym, czy miasto składania wniosku jest też miastem zamieszkania
- REG_CITY_NOT_WORK_CITYF – zmienna przyjmująca wartość YES, jeżeli miasto składania wniosku jest też miejscem pracy klienta
- LIVE_CITY_NOT_WORK_CITYF – zmienna mówi o tym, czy miasto zamieszkania oraz pracy jest to samo
- AMT_REQ_CREBIT_BUREAU_HOURF, AMT_REQ_CREBIT_BUREAU_DAYF, AMT_REQ_CREBIT_BUREAU_WEEKF, AMT_REQ_CREBIT_BUREAU_MONF, AMT_REQ_CREBIT_BUREAU_QRTF, AMT_REQ_CREBIT_BUREAU_YEARF - liczba zapytań o klienta do biura kredytowego odpowiednio w ciągu ostatniej godziny, dnia, tygodnia, miesiąca, kwartału i roku.

Ponadto przed przeprowadzeniem dalszej analizy usunięto kilka zmiennych:

- MOBILE_PHONE - zmienna opisująca czy dana osoba ma telefon komórkowy. Usunięta ze względu na brak zróżnicowania gdyż w bazie znajdowała się tylko jedna obserwacja bez telefonu.
- FLAG_CONT_MOBILE - podobny przypadek jak powyżej - słabe zróżnicowanie wyników.
- REGION_RATING_CLIENT_W_CITY - zmienna określająca to samo co REGION_RATE. Poziom korelacji jest wysoki (0,96), dlatego zmienna została wyrzucona.
- Wszystkie zmienne informujące o budynku i okolicy, w którym mieszka kredytobiorca. Zmienne te posiadały liczne braki danych oraz ich interpretacja była trudna.
- OBS_30_CNT_SOCIAL_CIRCLE oraz DEF_30_CNT_SOCIAL_CIRCLE - zmienne mówiące o ilości osób z otoczenia klienta, które nie spłaciły należności od 30 dni. Zmienne te są silnie skorelowane ze swoimi odpowiednikami w przypadku 60 dni. Dlatego postanowiono zostawić tylko jedną parę takich predyktorów

Coarse-classing

Klasyfikacja coarse-classing to proces podziału na grupy o drobnej wielkości w celu połączenia grup o podobnym ryzyku i utworzenia mniejszej liczby grup. Celem jest osiągnięcie uproszczenia poprzez utworzenie mniejszej liczby przedziałów, z których każdy charakteryzuje się odmiennymi czynnikami ryzyka, przy jednoczesnym zminimalizowaniu utraty informacji. Jednakże, aby stworzyć solidny model, odporny na przepełnienie, każdy z przedziałów powinien zawierać wystarczającą liczbę obserwacji z całego rachunku. Te przeciwstawne cele mogą być osiągnięte poprzez optymalizację w postaci optymalnego grupowania - binowania, które maksymalizuje moc predykcyjną zmiennej podczas procesu klasyfikacji zgrubnej.

Optymalne grupowanie wykorzystuje te same miary statystyczne, które są używane podczas selekcji zmiennych, takie jak wartość informacyjna (IV), statystyka Gini. Najbardziej popularną miarą jest, ponownie, wartość informacyjna (IV), chociaż kombinacja dwóch lub więcej miar jest często korzystna. Brakujące wartości, jeśli zawierają informację predykcyjną,

powinny stanowić oddzielną klasę lub zostać połączone w grupę o podobnych czynnikach ryzyka.

Klasyfikacja coarse-classingu umożliwiła utworzenie bazy składającej się z 17 zmiennych objaśniających, wraz ze zmienną defaultu oraz zmiennej o odwróconym znaczeniu defaultu.

VAR	**IV**	**Gini**
EXT_SOURCE_2	0.3057	0.30564499
EXT_SOURCE_3	0.3287	0.27490438
DAYS_EMPLOYED	0.1013	0.17729311
GOOD_PRICE	0.092	0.16743590
DAYS_BIRTH	0.0841	0.16385272
JOBF	0.0829	0.15305976
ORGANIZATION_TYPEF	0.06	0.13307062
SOURCE_INCOMEF	0.0574	0.12166476
EXT_SOURCE_1	0.1514	0.12123804
AMT_CREDIT	0.0451	0.11897923
DAYS_PHONE_CHANGE	0.0467	0.11569268
DAYS_ID_CHANGE	0.0385	0.11042234
REGION_RATEF	0.0484	0.09612849
SEXF	0.0386	0.09524731
EDUCATIONF	0.0505	0.09355449
AMT_ANNUITY	0.0266	0.08855807
DAYS_REGISTRATION	0.0269	0.08562732

Analizując poszczególne zmienne po wartości IV można zauważyć, że większość zmiennych objaśniających ma charakter słabego predyktora IV < 0.1. Tylko dwie zmienne określające zewnętrzne informacje na temat klienta (EXT_SOURCE_2, EXT_SOURCE_3) są silnym predyktorem. Te same zmienne wykazują największe wartości współczynnika Giniego, a więc są również zmiennymi wykazującymi największą nierówność rozkładu.

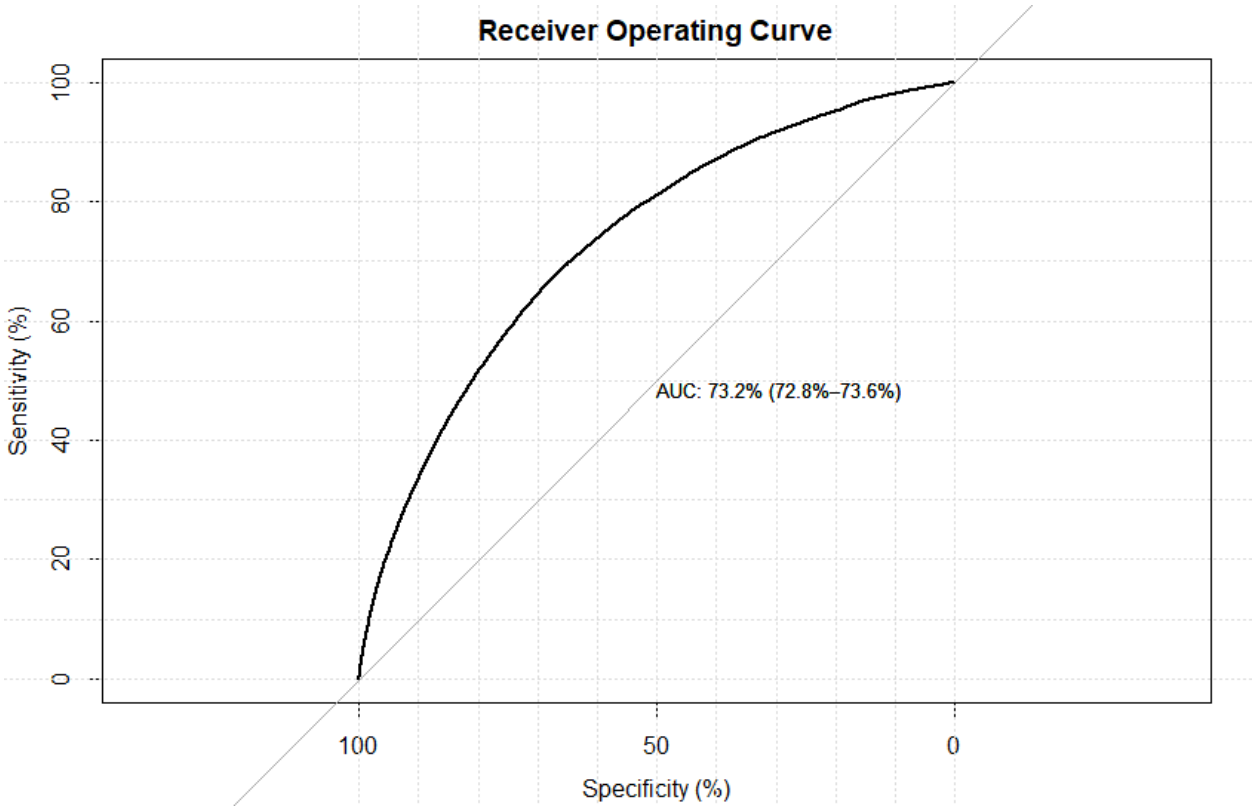
Indeks Giniego w badanym zbiorze wyniósł 9,293%, co na statystykę określoną przedziałem [0,1], ukazuje pewną bliskość 0 co oznacza stosunkową obecność równomierności rozkładu. Wartość IV została wyliczona na poziomie 0,0536, a więc zmienne wykazują wartość predykcyjną.

Model regresji logistycznej

Regresja logistyczna jest jednym z najważniejszych modeli dla danych zawierających odpowiedzi kategoryczne. Jest to przykład uogólnionego modelu liniowego, którego głównym zastosowaniem jest oszacowanie prawdopodobieństwa wystąpienia odpowiedzi binarnej w oparciu o szereg zmiennych predykcyjnych. Regresja logistyczna jest wykorzystywana w wielu różnych zastosowaniach. Jednym z przykładów tych ostatnich jest wykorzystanie modeli binarnej regresji logistycznej do oceny zdolności kredytowej, czyli: modelowanie prawdopodobieństwa, że klient ma zdolność kredytową (tj. jest w stanie terminowo wywiązać się ze zobowiązania finansowego) przy użyciu szeregu zmiennych predykcyjnych. Predyktory te mogą obejmować wielkość kredytu, jak również inne dane osobowe, takie jak roczne dochody klienta, jego zawód, inne niespłacone długi, jego wcześniejsze zachowania związane z brakiem spłaty i historia kredytowa.

Im lepiej model opisuje dany problem, tym krzywa ROC jest bardziej odgięta w kierunku górnego rogu rysunku. Pole pod krzywą używane jest jako miara jakości dopasowania modelu. Na podstawie danych jakość dopasowania do modelu ROC AUC score wynosi 73,2%. Taki poziom statystyki pozwala sądzić o dobrej trafności dopasowania modelu do danych wyjściowych.

Statystyka Giniego opisująca jakość regresji logistycznej przyjmuje poziom 46,4%, co również jest wynikiem dobrym dla badanego modelu.



Wyniki testu Hausmana – Lemeshow'a, to: chi = 15,48, a p-value = 0,0504, zatem w tym teście na poziomie istotności 0,05 nie odrzucamy hipotezy zerowej, wskazującej na poprawność formy funkcyjnej.

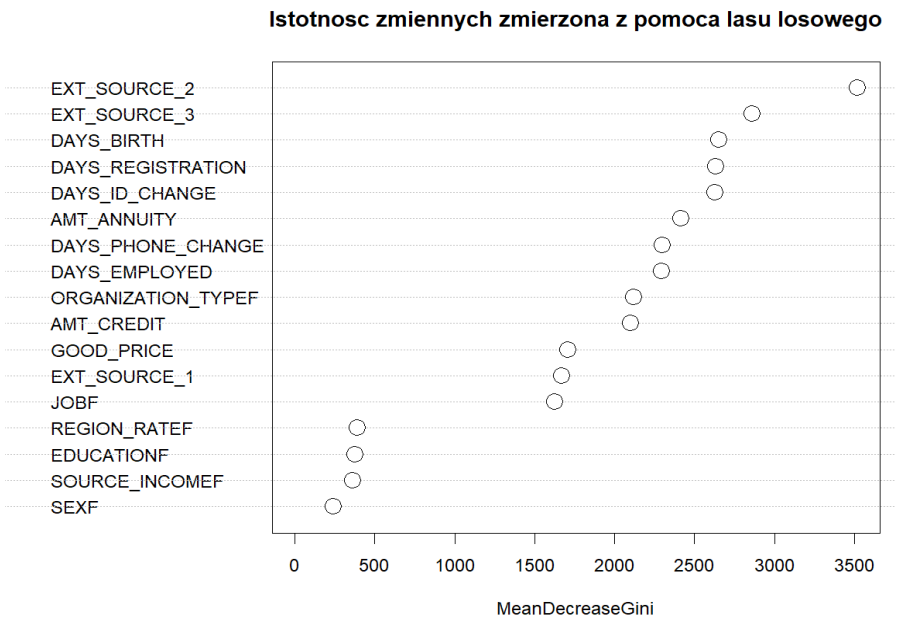
Testy badające stabilność zmiennej w czasie wykazały odmienne wyniki, zmienna PSI przyjęła wartość 0,02% co jest wartością stosunkowo małą, natomiast statystyka K-S wykazała poziom 52,537%, co charakteryzuje stabilność zmiennych w czasie.

Model lasu losowego

Metoda lasu losowego stanowi jedną z efektywniejszych technik uczenia maszynowego w rozwiązywaniu problemów klasyfikacyjnych. Algorytm polega na konstruowaniu wielu drzew decyzyjnych w procesie uczenia i wygenerowaniu klasy, która stanowi dominantę klas poszczególnych drzew. Samo drzewo decyzyjne konstruuje się w sposób rekurencyjny od korzenia do liścia. Zaczynając od pojedynczego węzła, metoda wykorzystuje miarę entropii (miara rozproszenia informacji) do wyboru cechy, która najlepiej podzieli zbiór. Następnie dla każdego rezultatu dane treningowe są odpowiednio dzielone do nowych gałęzi. Algorytm

powtarza się aż do momentu, gdy spełnione jest dane kryterium (np. maksymalna liczba liści). Podstawowym problemem drzewa decyzyjnego jest nadmierne dopasowywanie się do zbioru treningowego. Natomiast las losowy poprzez uśrednianie wielu głębokich drzew decyzyjnych, wyszkolonych na różnych częściach tego samego zestawu treningowego radzi sobie z tą kwestią. Oprócz poprawienia jakości las losowy jest odporny na rozmaite problemy z danymi oraz daje możliwość uzyskania dokładniejszych zależności niż są to w stanie zrobić drzewa decyzyjne.

Przechodząc do problemu modelowania ryzyka kredytowego przeprowadzono pierwszą estymację lasu losowego na niezmodyfikowanych danych. Wybrano te same zmienne objaśniające, co do modelu logitowego oraz zmniejszono liczbę drzew decyzyjnych z 500 do 300, ze względu na rozmiar bazy i czas wykonywania algorytmu. Jakość otrzymanej regresji mierzona współczynnikiem Giniego wyniosła 43%, zaś statystyka ROC AUC Score była równa 71,5%, co jest zadowalającym wynikiem. W celu polepszenia jakości tego modelu warto byłoby przeprowadzić fine - tuning, czyli metodę stopniowego dobierania optymalnych parametrów modelu w celu poprawienia wyniku. Jednakże z powodu, że proces ten jest zasobożerny oraz trwa bardzo długo pozostawiono parametry na domyślnym poziomie. Oprócz samej jakości warto przyjrzeć się jak kształtuje się wykres istotności poszczególnych predyktorów, czyli w jakim stopniu dana charakterystyka klienta ułatwia przewidywanie, że klient nie spłaci kredytu. Poniższy rysunek ukazuje, że trzema zmiennymi, które najlepiej różnicują klientów jest EXT_SOURCE_2, EXT_SOURCE_3 oraz DAYS_BIRTH, natomiast słabą informację na temat prawdopodobieństwa defaultu przekazują REGION_RATEF, EDUCATIONF, SOURCE_INCOMEF oraz SEXF.



Przechodząc dalej warto przeprowadzić ewaluację skonstruowanego modelu na zbiorze testowym. Zarówno współczynnik Giniego oraz wartość ROC AUC Score są zbliżone do tych otrzymanych na zbiorze treningowym i wynoszą odpowiednio 43,4% oraz 71,72%. Jednakże bardzo przydatnym narzędziem do oceniania trafności modelu jest tablica pomyłek (z angielskiego confusion matrix). Na jej podstawie można określić kilka statystyk: dokładność, która wynosi 91,93%, czułość - 99,9% oraz swoistość (z angielskiego specificity) równa 1,2%. Ta ostatnia miara jest szczególnie niepokojąca, ponieważ jest to ilość prawidłowo wykrytych defaultów w stosunku do wszystkich defaultów. Oznacza to, że model ten nie osiąga dobrych wyników w przypadku identyfikowania złych klientów, co należałoby poprawić.

		Klasy rzeczywiste	
		0	1
Klasy Prognozowane	0	84732	7355
	1	72	92

W tym celu w analizie sprawdzono również, czy może wartości woe w przypadku lasu losowego dadzą lepsze wyniki niż dane surowe. Jednakże wartość Giniego dla zbioru treningowego wyniosła zaledwie 24%, zaś ROC AUC Score 62%, co jest dużo gorszym rezultatem niż dla danych niemodyfikowanych. Dla zbioru testowego otrzymano zbliżone wyniki, co oznacza, że wartości woe nie sprawdzają się w estymacji modelu lasu losowego.

Podsumowując, na podstawie współczynnika Giniego można dojść do wniosku, że model lasu losowego nie różni się znacząco od regresji logistycznej pod względem zdolności predykcyjnej. Jednakże niepokojący jest jednak fakt, że swoistość powyższego modelu jest tak niska i może prowadzić do błędnego wykrywania złych klientów. Wówczas warto pozostać przy regresji logistycznej, która ponadto jest łatwiejsza do interpretacji i istnieje możliwość prześledzenia skąd biorą się konkretne wyniki.

Bibliografia

1. Baza danych z serwisu Kaggle "https://www.kaggle.com/gauravduttakiit/loan-defaulter?select=columns_description.csv"
2. Migut Grzegorz "Modelowanie ryzyka kredytowego" Statsoft Polska, 2003