Zastosowanie modelowania matematycznego w bankowości - lista 4

Łukasz Rębisz

2023-06-14

Zadanie 1

Otrzymane dane zostały oczyszczone w pliku Dane.xlsx.

Dane zapisane w arkuszu zostały oczyszczone w następujący sposób:

- WNIOSKOWANA_KWOTA: filtr - usunięcie putych wartości
- WOJEWODZTWO: filtr - usunięcie wartości X
- MIESIACE_ZATRUDNIENIA: filtr - pozostawienie tylko wartości dodatnich
- SEKTOR: filtr - usunięcie pustych wartości
- BIK_Liczba_zap_ost_mies: filtr - usunięcie pustych wartości.

Uzyskaliśmy 18 585 obserwacji (wierszy) z 22 061 wierszy wyjściowego arkusza.

Dane przedstawiają informacje o tym, czy dany klient banku przestał spłacać kredyt w ciągu 12 miesięcy od daty aplikacji (zmienna *DEFAULT*: 1 - tak, 2 - nie).

Zadanie 2

Celem raportu jest **zamodelowanie PD (Probability of Default)**, to znaczy prawdopodobieństwa, że klient przestanie spłacać kredyt w ciągu 12 miesięcy od daty aplikacji. Produkt kredytowy przechodzi w stan default, jeżeli spełniony jest co najmniej jeden z poniższych warunków:

- rachunek znajduje się w windykacji,
- liczba dni z przeterminowaniem > 400 PLN i $> 1\%$ wartości ekspozycji przekracza 90 dni,
- nastąpiła śmierć wszystkich kredytobiorców.

Zastosujemy model **regresji logistycznej**, w którym zmienna wynikowa (tu: DEFAULT) przyjmuje dwie wartości: 0 oraz 1.

W celu sprawdzenia dokładności modelu podzielimy analizowany zbiór danych na dwie próby: treningową oraz testową. Model oprzemy na danych treningowych, natomiast dokładność predykcji sprawdzimy zarówno na danych treningowych jak i testowych. Zasotosujemy bucketowanie polegające na podzieleniu wartości danej zmiennej na kilka kategorii (czynników). W kolejnym kroku wykluczymy z modelu nieistotne zmienne. Każdemu bucketowi przypiszemy wynik wpływający na p-stwo zdarzenia default.

Na koniec sprawdzimy moc i stabilność modelu w czasie.

Zadanie 3

Dokonajmy podziału danych na **próbę treningową** oraz **testową**.

Dysponujemy 18585 obserwacjami. Wybierzmy losowo 3000 obserwacji do próby testowej.

Zadanie 4

Dla danego klienta opisano następujące dane (dokonaliśmy podziału danych na czynniki - typ danych *factor*):

- CUSTOMER_CODE: numer klienta w systemie;
- APPLICATION_DATE: data aplikacji o produkt kredytowy: "I poł. 2019", "II poł. 2019", "I poł. 2020", "II poł. 2020";
- PRODUKT: CL - kredyt gotówkowy, CC - karta kredytowa, OV - limit w koncie;
- WNIOSKOWANA_KWOTA: "poniżej 20 tys.", "20-40 tys.", "40-60 tys.", "60-80 tys.", "80-100 tys.", "powyżej 100 tys.";
- WOJEWODZTWO;
- STAN_CYWILNY: "Stan wolny", "Związek" obejmujący małżeństwa z oraz bez rozdzielności majątkowej, a także wolne związki, "Rozwód", "Wdowa/wdowiec";
- STATUS_MIESZKANIOWY: "Właściciel" (domu lub mieszkania), "U rodziny", "Wynajem" (komercyjny lub komunalny), "Służbowe";
- WYKSZTAŁCENIE: "Podstawowe, zawodowe", "Średnie", "Wyższe" (zarówno pierwszego jak i drugiego stopnia);
- ZAWOD_WYKONYWANY: "Student", "Fizyczny", "Umysłowy" (w tym samodzielne stanowisko), "Zarząd", "Właściciel" (w tym przedsiębiorca);
- Dochód: "< 2,5 tys.", "2,5-5 tys.", "5-7,5 tys.", "7,5-10 tys.", "10-15 tys.", "> 15 tys.";
- MIESIACE_ZATRUDNIENIA: "< 50", "50-100", "100-150", "150-200", "200-300", "> 300";
- RODZAJ_ZATRUDNIENIA: "Umowa czas nieokreślony", "Umowa dzieło/czas określony", "Przedsiębiorca", "Emerytura", "Inny";
- SEKTOR: "Przemysł", "Sektor publiczny", "Usługi", "Usługi specjalistyczne";
- TYP_PRACODAWCY: "Państwowe", "Działalność gospodarcza", "Firma komercyjna", "Emerytura";
- WIELKOSC_ZATRUD: "[0 ; 3]", "[4 ; 29]", "[30 ; 59]", "[60 ; 119]", "[120 ; max]";
- BIK_Liczba_zap_ost_mies: liczba wniosków o produkt kredytowy w bieżącym miesiącu;
- BIK_Liczba_zap_poprz_mies: liczba wniosków o produkt kredytowy w poprzednim miesiącu;
- liczba_mies_aktywny_produkt: liczba miesięcy z aktywnym produktem oszczędnościowym w ostatnim roku;
- liczba_m2_na_osobe: "do 20 m2", "20-40 m2", "powyżej 40 m2".

Zadanie 5

Dla otrzymanego podziału wyznaczmy dla każdej zmiennej (dla wszystkich kategorii zmiennej osobno) wartości następujących statystyk:

WOE (*Weight of Evidence*)

Statystyka WOE mierzy jakość grupy o danym atrybucie (wartości) cechy (zmiennej) relatywnie do całej populacji:

- wartość ujemne - atrybut pozytywny, ryzyko mniejsze od średniej,
- wartość ujemne - atrybut negatywny, ryzyko większe od średniej,
- wartość zerowa - cecha neutralna, ryzyko równe średniej.

Wartość WOE wylicza się zgodnie ze wzorem:

$$WOE(A) = \log\left[\frac{P(A|G)}{P(A|B)}\right],$$

gdzie A oznacza badany atrybut, G - *good*: obserwacje dobre (bez zdarzenia default w ciągu 12 miesięcy), natomiast B - *bad* to obserwacje złe.

W rachunkach zastosujemy nieobciążone estymatory:

$$P(A|G) = \frac{\text{liczba obserwacji o atrybucie A w grupie G} + 0.5}{\text{liczebność grupy G} + 0.5},$$

$$P(A|B) = \frac{\text{liczba obserwacji o atrybucie A w grupie B} + 0.5}{\text{liczebność grupy B} + 0.5}.$$

IV (*Information Value*)

Statystyka WOE jest miarą na poziomie atrybutu (wartości). Sprawdźmy **istotność zmiennych**, wyznaczając statystykę **IV** (*Information Value*), która jest miarą na poziomie całej cechy (zmiennej):

$$IV(G,B) = AVG_G(WOE) - AVG_B(WOE) = \sum_i (P(A_i|G) - P(A_i|B)) \cdot WOE(A_i).$$

Im większa wartość informacyjna IV, tym dana cecha (zmienna) jest ważniejsza w odróżnianiu kredytów dobrych od złych. Do modelu wybieramy zmienne o wysokim IV:

- 0-10% zmienna o niskiej mocy
- 11-25% „średnia" moc zmiennej
- powyżej 25% zmienna o dużej mocy.

Otrzymaliśmy następujące wartości informacyjne IV dla poszczególnych zmiennych (za istotne zmienne uznajemy te, dla których IV $> 11\%$):

Zmienna	Wartość IV	Zmienna istotna
APPLICATION_DATA	0.24	1
PRODUKT	0.37	1
WNIOSKOWANA_KWOTA	0.18	1
WOJEWÓDZTWO	0.09	0
STAN_CYWILNY	0.08	0
STATUS_MIESZKANIOWY	0.11	0
WYKSZTAŁCENIE	0.07	0
ZAWOD_WYKONYWANY	0.10	0
Dochód	0.02	0
MIESIACE_ZATRUDNIENIA	0.17	1
RODZAJ_ZATRUDNIENIA	0.11	0
SEKTOR	0.06	0
TYP_PRACODAWCY	0.19	1
WIELKOSC_ZATRUD	0.07	0
BIK_Liczba_zap_ost_mies	0.39	1
BIK_Liczba_zap_poprz_mies	0.46	1
liczba_mies_aktywny_produkt	0.04	0
liczba_m2_na_osobe	5.46	?

Wniosek: za istotne uznajemy zmienne: *APPLICATION_DATA, PRODUKT, WNIOSKOWANA_KWOTA, MIESIACE_ZATRUDNIENIA, TYP_PRACODAWCY, BIK_Liczba_zap_ost_mies, BIK_Liczba_zap_poprz_mies.*

Wartość otrzymana dla zmiennej *liczba_m2_na_osobe* jest nieracjonalna. Nie włączamy zmiennej do modelu w celu uniknięcia ewewntualnych błędów.

Zadanie 6

Skonstruujmy finalny **model regresji logistycznej** ze statystycznie istotnymi zmiennymi wyznaczonymi w poprzednim zadaniu.

Zakładana w modelu zależność pomiędzy zmiennymi:

$$\mu = \frac{1}{1 + exp(-\eta)},$$

gdzie μ oznacza średnie prawdodobieństwo osiągnięcia statusu *default* w ciągu roku, natomiast η to kombinacja liniowa zmiennych objaśniających.

Uzyskaliśmy następujące współczynniki kombinacji liniowej dla poszczególnych zmiennych:

- -0.98543: APPLICATION_DATE
- -0.83399: PRODUKT
- -0.13902: WNIOSKOWANA_KWOTA
- -0.66421: MIESIACE_ZATRUDNIENIA
- -0.70255: TYP_PRACODAWCY
- -1.39618: BIK_Liczba_zap_ost_mies
- -1.16943: BIK_Liczba_zap_poprz_mies.

Zbadajmy **korelację** pomiędzy zmiennymi w modelu.

Macierz korelacji pomiędzy zmiennymi objaśniającymi

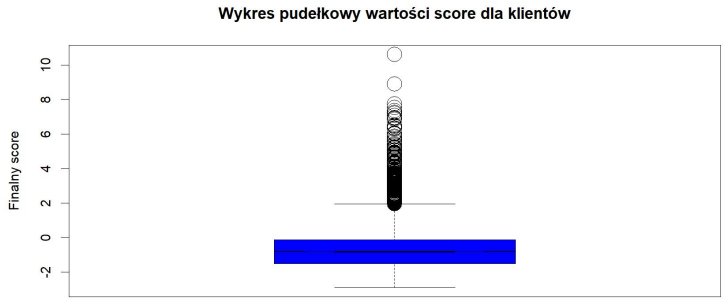
$$\begin{array}{c} \\ 1. \\ 2. \\ 3. \\ 4. \\ 5. \\ 6. \\ 7. \end{array} \begin{array}{c} \begin{array}{ccccccc} 1. & 2. & 3. & 4. & 5. & 6. & 7. \end{array} \\ \begin{pmatrix} 1.00 & -0.06 & 0.13 & 0.04 & 0.07 & 0.02 & -0.02 \\ -0.06 & 1.00 & 0.21 & -0.02 & -0.06 & 0.11 & 0.07 \\ 0.13 & 0.21 & 1.00 & -0.03 & 0.10 & 0.15 & 0.16 \\ 0.04 & -0.02 & -0.03 & 1.00 & 0.10 & 0.06 & 0.05 \\ 0.07 & -0.06 & 0.10 & 0.10 & 1.00 & 0.06 & 0.07 \\ 0.02 & 0.11 & 0.15 & 0.06 & 0.06 & 1.00 & 0.14 \\ -0.02 & 0.07 & 0.16 & 0.05 & 0.07 & 0.14 & 1.00 \end{pmatrix} \end{array}$$

Analiza powyższych wyników wskazuje na **niską korelację** pomiędzy wykorzystanymi w modelu zmiennymi objaśniającymi. Wartość korealcji pomiędzy różnymi zmiennymi nie przekracza wartości 0.21 (maksymalna korelacja wynosi 1 - przykład: korelacja zmiennej z samą sobą). Dla większości zmiennych wartość korelacji nie przekracza 0.15.

Zadanie 7

Przypiszmy każdemu bucketowi wartość **score** równą iloczynowi odpowiedniego współczynnika modelu liniowego i wartości WOE danego bucketu.

Następnie przypiszmy „finalny score" każdemu klientowi, tzn. sumę wartości score wszystkich badanych zmiennych dla danego klienta.



Analiza powyższego wykresu pokazuje, że:

- mediana wartości osiąga wartości bliską zero (wartość nieznacznie ujemna). Oznacza to, że dla dużej liczby klientów uzyskany wynik *score* nie wskazuje istotnego prawdopodobieństwa wystąpienia zdarzenia default.
- Z kolei grupa klientów z ujemną wartością *score* charakteryzuje się bardzo małym p-stwem zdarzenia default.
- Duża część klientów z dodatnim *scorem* została ujęta na powyższym wykresie w postaci obserwacji odstających (czarne okręgi). Klienci ci charakteryzują się bardzo dużym p-stwem wystąpienia zdarzenia i default, zatem szczególnie tę grupę klientów warto poddać dalszej analizie.

Zadanie 8

Sprawdźmy **moc modelu** na próbie treningowej oraz testowej za pomocą **testu Kołmogorowa-Smirnowa** oraz **miary Giniego**.

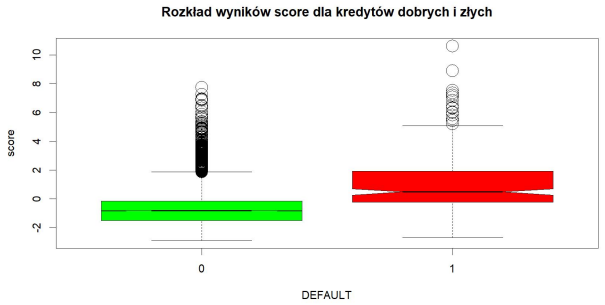
Test Kołmogorowa-Smirnowa

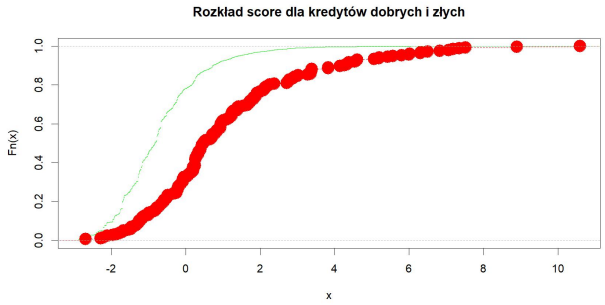
Test Kołmogorowa-Smirnowa bada równość rozkładów dwóch prób (kredyty dobre i złe) poprzez wyznaczenie maksymalnej odległości pomiędzy dystrybuantami rozkładów kredytów dobrych i złych

Próba treningowa

Otrzymana p-wartość dla testu Kołmogorowa-Smironowa nie przekracza 0.00000000000000022, zatem zdecydowanie odrzucamy równość rozkładów wartości *score* dla kredytów dobrych i złych. Oznacza to, że przypisane wartości *score* dobrze charakteryzują ryzyko kredytowe.

Sprawdźmy, jak wyglądają rozkłady wartości *score* dla kredytów dobrych i złych.





Analiza wykresu po lewej stronie wskazuje na **różne rozkłady** wartości *score* dla kredytów dobrych i złych (zgodnie z wynikiem testu Kołmogorowa-Smirnowa). W szczególności mediany rozkładów osiągają różne wartości. Ponadto, porównanie wykresów z pominięciem obserwacji odstających (okręgów) pokazuje, że większa część kredytów, które znalazły się w stanie default osiąga wyższe wartości *score*. Niemniej jednak dla takich kredytów również obserwujemy małe wartości *score*.

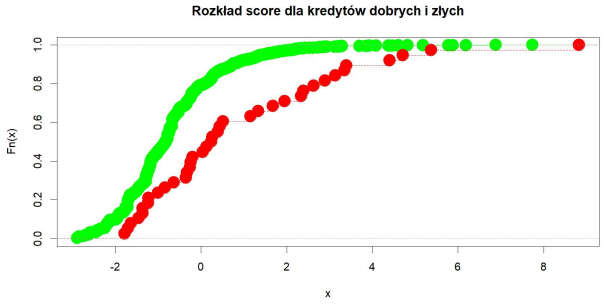
Ponadto wykres po prawej stronie obrazujący rozkłady wyników *score* dla kredytów dobrych i złych wyraźnie pokazuje, że **rozkłady** dla tych dwóch grup kredytów są **różne** (zgodnie z wynikiem testu KS).

Próba testowa

Przeprowadźmy test Kołmogorowa-Smirnowa dla próby testowej. W tym celu dla każdej z badanych w modelu zmiennych objaśnianych przypiszmy wartość *score* wyznaczoną w próbie treningowej.

Otrzymana p-wartość dla testu Kołmogorowa-Smironowa jest równa 0.00001131, zatem odrzucamy równość rozkładów wartości *score* dla kredytów dobrych i złych. Oznacza to, że przypisane wartości *score* dobrze charakteryzują ryzyko kredytowe również dla próby testowej. Warto zauważyć, że otrzymana p-wartość jest o kilka rzędów większa niż p-wartość otrzymana w teście KS dla próby treningowej.

Rozkład wartości *score* dla kredytów dobrych i złych w próbie testowej potwierdza wynik testu KS (różne rozkłady):



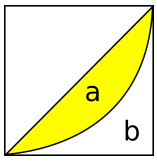
Miara Giniego

Zbadajmy następnie moc testu dla próby treningowej i testowej, wykorzytsując **miarę Giniego**. Współczynnik Giniego najbardziej zależy od mocy cechy w skrajnych wartościach. Natomiast statystyka KS mierzyła moc cechy dla wartości zbliżonych do średniej.

Graficznie współczynnik Giniego stanowi dwukrotność pola obszaru pomiędzy krzywą Lorenza a przekętną kwadratu jednostkowego (Hasło: Współczynnik Giniego, [w:] Wikipedia. Wolna encyklpedia):

$$G = \frac{a}{a+b},$$

gdzie a to pole opisane powyżej obszaru, a b to pole jego dopełnienia do trójkąta.



Własności:

- Miara Giniego przyjmuje wartości z przedziału [0, 1]. Często wyrażana jest w procentach.
- Wartość zerowa wskazuje na pełną równomierność rozkładu.
- Wzrost wartości współczynnika oznacza wzrost nierówności rozkładu.

Otrzymujemy następujące wartości miary Giniego:

- 0.889 dla próby treningowej,
- 0.893 dla próby testowej.

Powyższe wyniki świadczą o dużej nierówności rozkładu, czyli o dobrym dopasowaniu modelu do danych.

Porównanie miary Giniego w czasie (2019 vs 2020):

$$\begin{array}{l} \\ \text{Próba treningowa} \\ \text{Próba testowa} \end{array} \begin{array}{c} \begin{array}{cc} 2019 & 2020 \end{array} \\ \begin{pmatrix} 0.929 & 0.799 \\ 0.887 & 0.893 \end{pmatrix} \end{array}$$

Powyższą wyniki pokazują, że w każdej z analizowanych sytuacji **miara Giniego** osiąga **wysoką wartość** świadczącą o dobrym dopasowaniu modelu.

Największa wartość została osiągnieta dla próby treningowej w 2019 roku. Dla tych danych obserwujemy spadek wartości miary Giniego w kolejnym roku.

W przypadku próby testowej obserwujemy podobną wartość miary Giniego na poziomie ok. 0.89 w obu badanych latach. Wynik przemawia na korzyść dużej mocy modelu w czasie.

Zadanie 9

Sprawdźmy następnie **stabilność modelu** w czasie.

W tym celu wyznaczmy **miarę PSI** (*Population Stability Index*) pozwalającą ocenić różnice pomiędzy dwoma populacjami względem badanej cechy. Wyznaczymy wartość miary PSI dla kredytów dobrych i złych względem wartości *score*, biorąc pod uwagę próbę testową względem treningowej. Interpretacja wyników:

- PSI poniżej 10% - brak istotnych różnic,
- PSI w zakresie 10-20% - „podejrzenie" różnic,
- PSI powyżej 20% - istotne różnice.

Otrzymujemy następujące wyniki PSI dla wartości *score*:

$$\begin{array}{l} \\ \text{PSI} \end{array} \begin{array}{c} \begin{array}{ccc} \text{cała próba} & 2019 & 2020 \end{array} \\ \begin{pmatrix} 0.026\% & 1.17\% & 4.14\% \end{pmatrix} \end{array}$$

Powyższe wyniki świadczą o bardzo **małych różnicach** pomiędzy badanymi próbami względem wartości *score*. Oznacza to **stabilność modelu**: w próbie treningowej i testowej badana wartość *score* ma podobny rozkład. Ponadto, obserwujemy nieznaczną różnicę wartości PSI w czasie pomiędzy 2019 a 2020 rokiem, która również świadczy na korzyść stabilności modelu.