Dokumentacja projektu PD lookup zrealizowanego w ramach zajęć „Inżynieria oprogramowania” w semestrze zimowym roku akademickiego 2023/2024

wyk. Aleksandra Chmarzyńska,

Agata Buksińska,

Karolina Mucha

Uniwersytet Gdański, 2024

1. Charakterystyka oprogramowania:

Model PD (Probability of Default) jest używany w dziedzinie finansów, szczególnie w ocenie ryzyka kredytowego. Do zbudowania modelu PD wykorzystana została regresja logistyczna. Jest to powszechnie akceptowana i obecnie najbardziej powszechna metoda oceny prawdopodobieństwa wystąpienia zdarzenia. Daje możliwość stworzenia karty scoringowej (scorecard), która stanowi syntetyczną charakterystykę badanej populacji - w tym przypadku klientów banku.

Za pomocą regresji logistycznej można stworzyć model deskryptywny oraz predykcyjny.

Model deskryptywny służy do poznania i opisu badanej populacji. Model predykcyjny, którego stworzenia podjęto się w niniejszym projekcie, służy do oceny wiarygodności kredytowej klientów w procesie aplikacji o kredyt i zbudowany jest na danych historycznych.

Celem budowy modelu predykcyjnego jest uzyskanie oceny pochodzącej z karty scoringowej, która została skalibrowana do prawdopodobieństwa wystąpienia zdarzenia niewypłacalności (probability of default).

Możliwe zastosowania wyników modelu regresji logistycznej w banku:

- ustalanie adekwatności kapitałowej,
- ustalanie wysokości rezerw na oczekiwane straty kredytowe,
- testy warunków skrajnych,
- regulowanie apetytu na ryzyko,
- optymalizacja kampanii pre-akceptowalnych,
- wyznaczenie wartości godziwej aktywów i inne.

Oprogramowanie związane z modelem PD wykonuje szereg zadań i funkcji, spełniając różnorodne potrzeby użytkowników:

1. Funkcjonalność:
    - Obliczanie i prezentowanie prawdopodobieństwa niewypłacalności (PD) dla danego kredytu lub portfela kredytowego umożliwiając instytucjom finansowym oszacowanie ryzyka związane z udzieleniem kredytów
    - Wykonywanie analiz historycznych danych kredytowych i określa wzorce, które wpływają na prawdopodobieństwo niewypłacalności wykorzystując analizę statystyczną, trendów i czynników wpływających na ryzyko
    - Implementacja skomplikowanych modeli matematycznych i statystycznych do przewidywania prawdopodobieństwa niewypłacalności na podstawie różnych czynników

- Oprogramowanie wspiera procesy doskonalenia modeli PD poprzez analizę wyników, dostosowywanie parametrów i uwzględnianie nowych danych
- Generowanie raportów i wizualizacji danych związanych z prawdopodobieństwem niewypłacalności (użytkownicy, tacy jak analitycy kredytowi, potrzebują czytelnych informacji do podejmowania decyzji)
- Wspieranie zarządzania portfelem kredytowym, umożliwiając monitorowanie i analizę ryzyka na poziomie całego portfela
- W przypadku instytucji finansowych, oprogramowanie mogłoby być zintegrowane z systemami do automatyzacji decyzji kredytowych, co pozwala na szybkie reagowanie na zmiany w ryzyku
- Wsparcie w spełnianiu wymogów regulacyjnych dotyczących oceny ryzyka kredytowego, zapewniając zgodność z przepisami i standardami branżowymi.
- Jeśli instytucja finansowa podlega nadzorowi organów regulacyjnych, oprogramowanie mogłoby wspierać proces raportowania danych związanych z ryzykiem kredytowym.

2. Wydajność:
    - Optymalizacja kodu
    - Minimalne zużycie zasobów systemowych
    - Szybkie odpowiedzi oprogramowania
    - Użycie wbudowanych funkcji R
    - Użycie odpowiednich pakietów
3. Niezawodność:
    - Oprogramowanie działa zgodnie z oczekiwaniami w różnych warunkach
    - Oprogramowanie zostało napisane prostym, przejrzystym kodem, co w przypadku błędów będzie gwarantowało łatwość naprawy
4. Użyteczność:
    - Oprogramowanie jest intuicyjne i łatwe w użyciu dla użytkowników końcowych
    - Dokumentacja jest przejrzysta i łatwo dostępna
    - Autorki gwarantują wsparcie użytkowników przy ewentualnych problemach
5. Utrzymanie:
    - Oprogramowanie można łatwo utrzymać i modyfikować
    - Zastosowany został czytelny kod
    - Zastosowane zostały standardy programowania oraz dokumentacja
6. Bezpieczeństwo:
    - Nie zostały zapewnione odpowiednie środki bezpieczeństwa, takie jak kontrola dostępu, szyfrowanie danych i zabezpieczenia przed atakami

7. Zgodność:
    - Oprogramowanie spełnia określone normy, przepisy i standardy branżowe
8. Łatwa instalacja:
    - Odwiedź oficjalną stronę R na adresie https://www.r-project.org/.
    - Kliknij na "CRAN" w sekcji "Download"
    - Wybierz odpowiednią wersję R dla swojego systemu operacyjnego
9. Dokumentacja została sporządzona w sposób klarowny, kompletny i zrozumiały.
10. Interoperacyjność:
    - Oprogramowanie jest zdolne do współpracy z innym oprogramowaniem i systemami.

Ostatecznym celem oprogramowania związanego z modelem PD jest umożliwienie instytucjom finansowym skutecznego zarządzania ryzykiem kredytowym, podejmowania informowanych decyzji kredytowych i spełniania wymagań regulacyjnych. W ten sposób odpowiada na potrzeby użytkowników, którzy pracują w obszarze oceny i zarządzania ryzykiem finansowym.

2. Prawa autorskie

Prawa autorskie do kodu napisanego w języku R przysługują jego autorkom zgodnie z ogólnymi zasadami prawa autorskiego. Prawa autorskie obejmują różne aspekty, takie jak kopiowanie, modyfikowanie, rozpowszechnianie i publiczne wykonywanie kodu.

- Przy tworzeniu kodu współpracował zespół trzech studentek: Agata Buksińska, Aleksandra Chmarzyńska oraz Karolina Mucha
- Od chwili stworzenia kodu, zespół autorek automatycznie posiada prawa autorskie do swojej pracy
- Autorki mogą udzielić licencji na korzystanie z jego kodu innym osobom
- Autorki zdecydowały się opublikować kod jako otwarty źródłowy
- Przed rozpoczęciem projektu ustalone zasady dotyczące praw autorskich i licencji oraz została zawarta umowa o współtworzeniu projektu
- Zgodnie z prawami autorskimi, jedynie autorki oryginalnego kodu mają prawo wprowadzać modyfikacje. Inni użytkownicy mogą to robić jedynie na podstawie licencji udzielonej przez ww. zespół autorek
- Prawa autorskie obejmują nie tylko sam kod, ale również dokumentację i komentarze

3. Specyfikacja wymagań

Nie wieeeem co tuuu daac. Nie pozdrawiam Jerzego.

4. Architektura systemu/oprogramowania

Jerzy jakieś stosy tu chce, kto go tam wie.

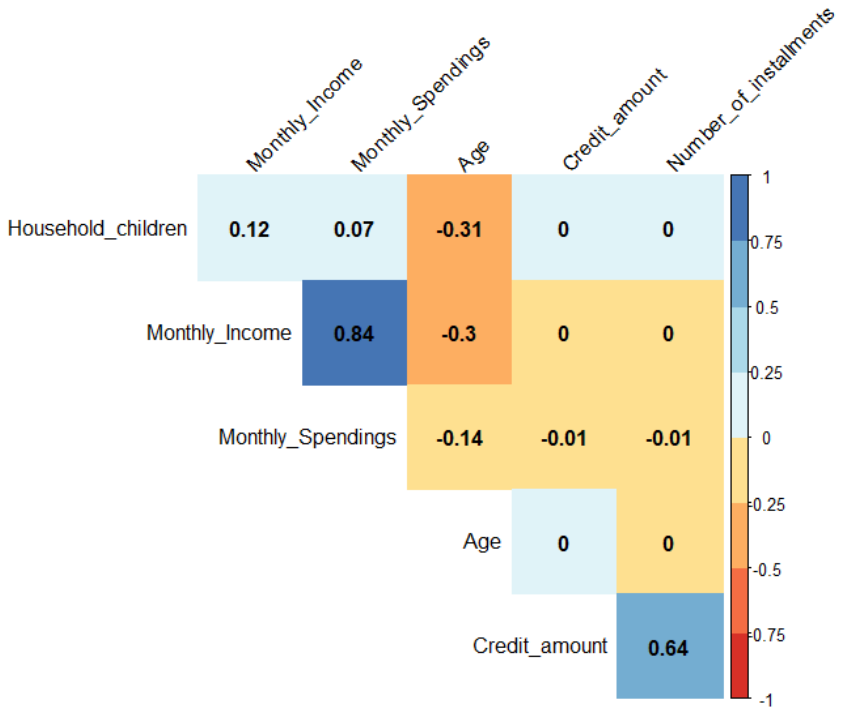
5. Testy

Punkt 10 w poniższym opisie

Opis prac

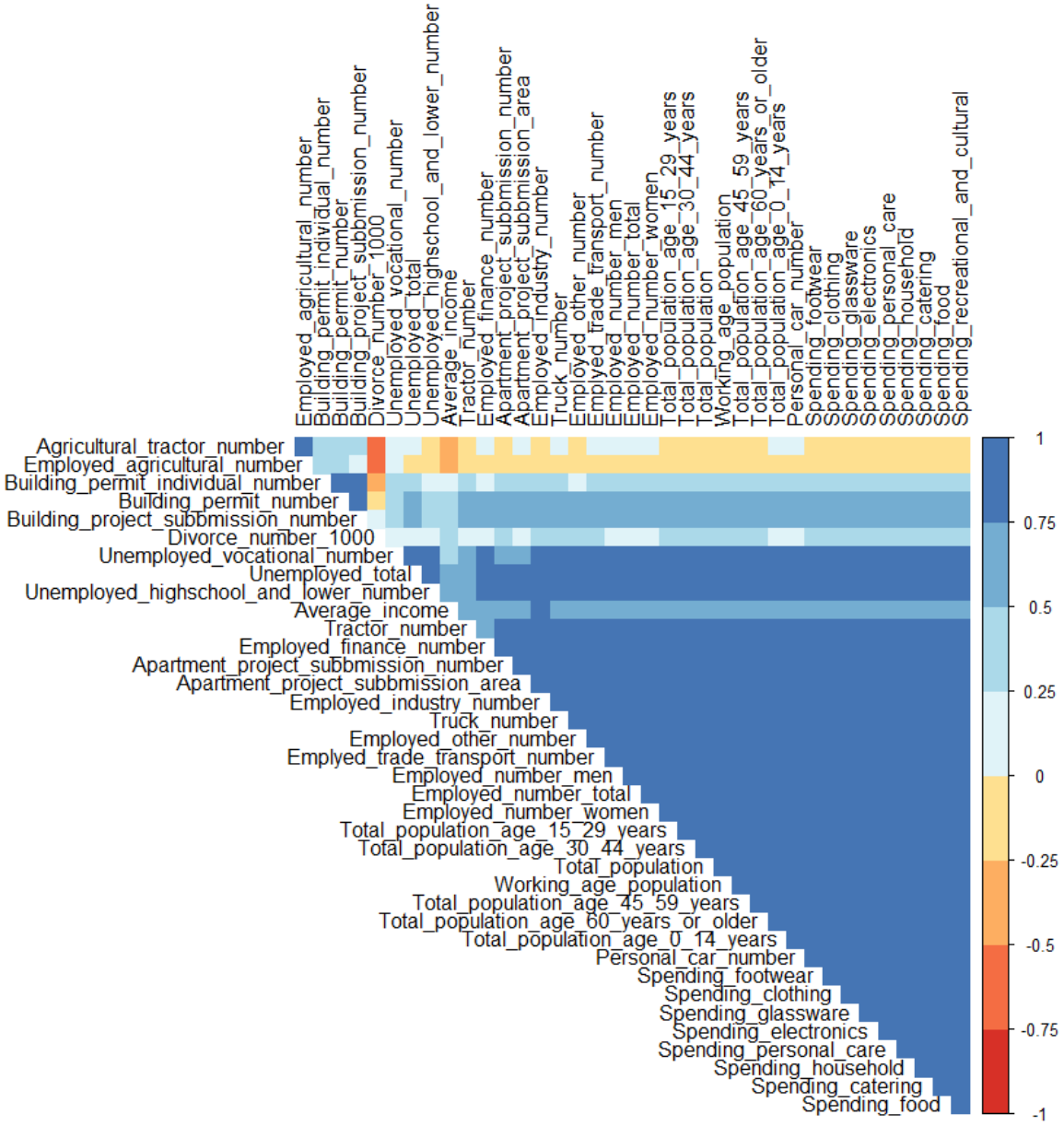
1. **Podstawowe statystyki – obserwacje**
    - AGE: minimalny wiek to 17, jest to wartość odstająca, patrząc na wykres.
    - MONTHLY_SPENDING: minimalne wydatki to 0.
    - MONTHLY_AVG_INCOME: średnie miesięczne zarobki są bardzo wysokie w porównaniu do MONTHLY_INCOME, mogą zakłamywać obraz rzeczywistości ze względu na wysokie wartości odstające i nierówność w zarobkach, lepszym odniesieniem byłaby mediana, a nie średnia (widać to też po utworzeniu zmiennej Income_to_avg_income). Zmienne będą standaryzowane.
2. **Analiza braków danych**
    - Braki danych występują dla zmiennej: IntoDefFlag. Dla zmiennej IntoDefFlag występuje 30% braków danych, wynika to z charakteru samego zbioru danych, ponieważ 30% obserwacji zmiennej celu zostało zakrytych w celu wykonania późniejszej oceny modelu na tym zbiorze.
3. **Analiza obserwacji odstających za pomocą Z-SCORE**
    - Przeprowadzono analizę obserwacji odstających za pomocą Z-score.
    - W regresji logistycznej przy niskiej liczbie zdarzeń usunięcie znacznej liczby obserwacji odstających, które często są odrębnymi zdarzeniami, może prowadzić do zjawiska quasi-separacji, a tym samym powodować problem z oszacowaniem parametrów modelu.
4. **Macierze korelacji dla zmiennych numerycznych**
    - Wykonano je w podziale na trzy grupy zmiennych (aplikacyjne, geolokalizacyjne oraz behawioralne).

Macierz korelacji dla zmiennych numerycznych z kategorii aplikacyjnych:



- Widać silną korelację zmiennej miesięczne wydatki i miesięczne dochody, co ma
- również uzasadnienie ekonomiczne (korelacja na poziomie 84%).
- Silna korelacja (64%) widoczna też jest dla zmiennej przedstawiającej liczbę rat oraz kwotę kredytu.
- W dalszej części analizy wykorzystane zostaną modyfikacje tych zmiennych.

Macierz korelacji dla zmiennych geolokalizacyjnych:

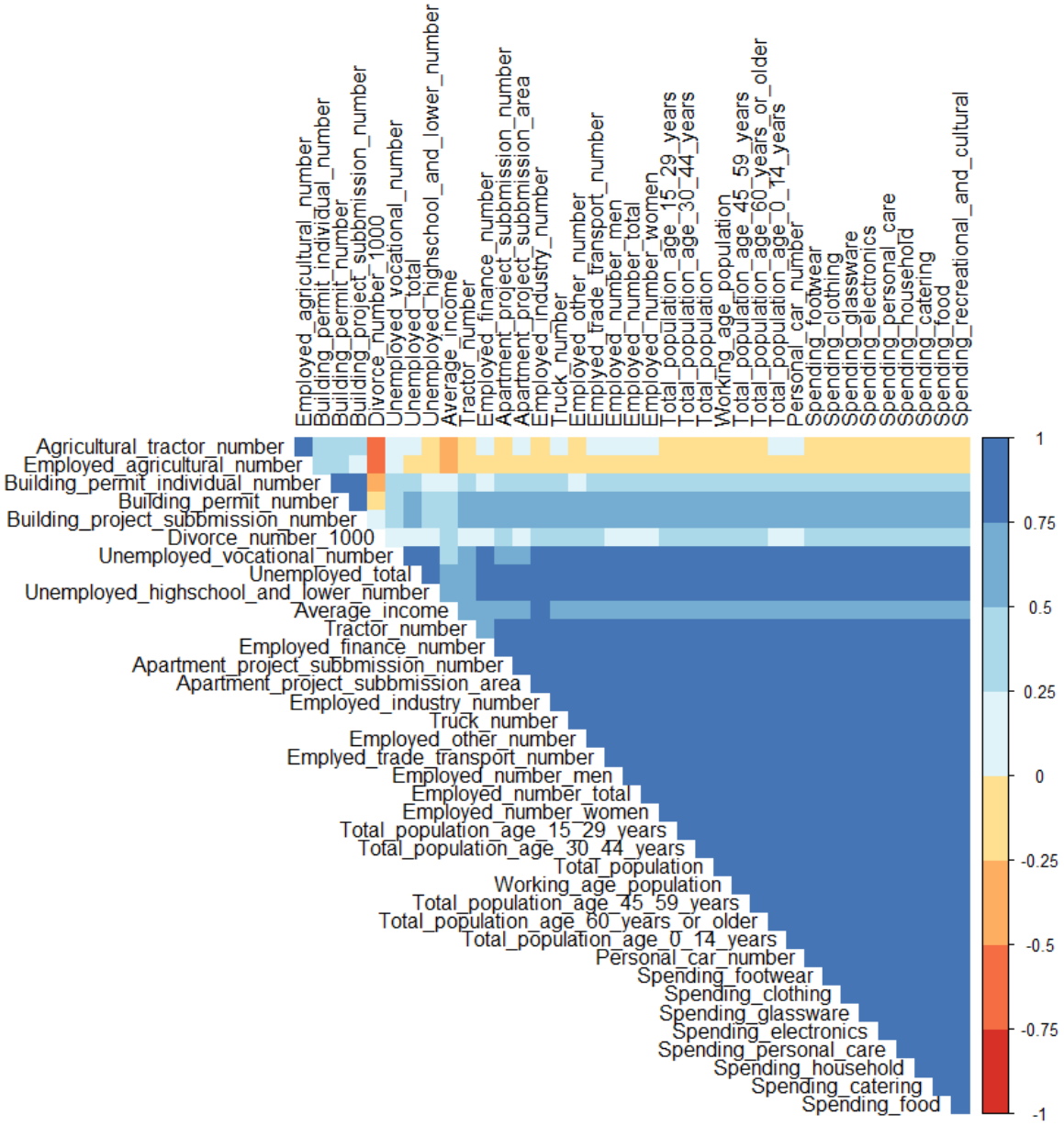


Dla zmiennych z kategorii geolokalizacyjnych widać dużą korelację. Jedynie dla zmiennych:

- “Liczba zarejestrowanych ciągników rolniczych”
- “Liczba zatrudnionych w rolnictwie”
- “Liczba rozwodów na 1000 osób”
- “Liczba pozwoleń na budowę, budownictwo indywidualne”

występuje słaba korelacja z innymi zmiennymi.

Macierz korelacji dla zmiennych behawioralnych:

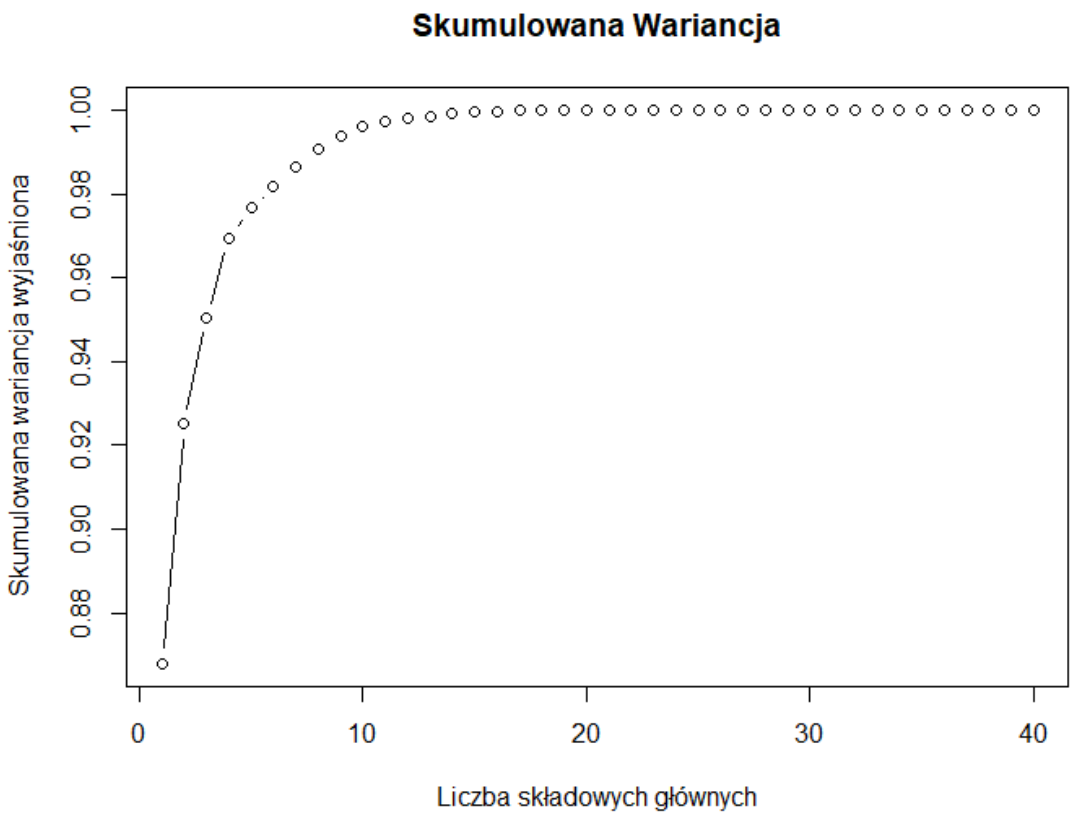


- Macierz korelacji dla zmiennych behawioralnych wykazała, że zmienne DPD, Notional Value oraz Notional Overdue są skorelowane ze swoimi własnymi opóźnieniami, co jest też zgodne z intuicją.

Przyjęto, że o modelowania nie przyjmuje się skorelowanych zmiennych, dla których wsp. korelacji Pearsona > 0,5. Aby wyłonić bardziej pożądaną zmienną objaśniającą do modelu ze skorelowanej pary brano pod uwagę poniższe kryteria:

- Relevancy - wybór zmiennej która jest bardziej relewantna dla problemu. Jeśli jedna zmienna ma większy związek z problemem lub jest bardziej logicznie powiązana z analizą PD zostaje w wzięta do analizy,
- Interpretowalność - wybór zmiennej, która jest łatwiejsza do zrozumienia i interpretacji,
- Mniejsza korelacja z innymi zmiennymi,
- Cel modelu - wybierano zmienną, która bardziej przyczyni się do osiągnięcia celu.

5. **Analiza wielowymiarowa (pokazuje zależności zmiennych)**
    - Wykonano analizę głównych składowych dla ustandaryzowanych i znormalizowanych zmiennych numerycznych.



Wykres przedstawia skumulowaną wariancję wyjaśnioną przez kolejne składowe główne w analizie PCA. Oś pionowa reprezentuje skumulowany procent wyjaśnionej wariancji, a oś pozioma wskazuje liczbę składowych głównych.

Można zauważyć, że pierwsza składowa główna wyjaśnia znaczącą część wariancji, ponieważ krzywa szybko się wznosi. Następnie przyrost skumulowanej wariancji zwalnia przy dodawaniu kolejnych składowych, co widać po stopniowym płaskim zbliżaniu się linii do wartości 1, co oznacza 100% wyjaśnionej wariancji.

Pierwsza składowa ma odchylenie standardowe około 1.31, co oznacza, że wyjaśnia ona większość wariancji w danych. Kolejne składowe mają coraz mniejsze wartości, co wskazuje na to, że każda kolejna składowa wnosi coraz mniej do wyjaśnienia wariancji.

Zmienne: Personal_car_number, Truck_number, Tractor_number, i Monthly_Income mają stosunkowo duże wartości bezwzględne, co wskazuje na to, że te zmienne mają silny wpływ na kształtowanie tej składowej.

```
Standard deviations (1, .., p=40):
 [1] 1.312593e+00 3.368224e-01 2.243189e-01 1.941247e-01 1.194172e-01 1.015490e-01 9.610520e-02 8.972616e-02 7.931731e-02 6.956150e-02
[11] 4.558243e-02 3.874366e-02 3.388680e-02 3.386711e-02 3.017687e-02 1.636606e-02 1.370215e-02 1.119660e-02 9.997884e-03 8.476615e-03
[21] 7.084085e-03 4.723233e-03 4.634300e-03 4.389136e-03 3.822961e-03 3.001084e-03 2.347384e-03 1.768753e-03 9.117193e-04 7.197679e-04
[31] 4.228521e-04 5.916812e-06 4.790088e-06 1.154376e-06 3.513938e-07 1.349979e-07 1.944186e-08 3.084185e-15 5.829083e-16 1.198957e-16
```

	PC1
Monthly_Income	-6.912894e-05
Monthly_Spendings	-6.445159e-05
Credit_amount	-2.941526e-04
Personal_car_number	-1.781859e-01
Truck_number	-1.784564e-01
Tractor_number	-9.595707e-02
Agricultural_tractor_number	2.142412e-04
Building_permit_number	-8.752144e-02
Building_permit_individual_number	-4.494059e-02
Building_project_subbmission_number	-7.362161e-02
Apartment_project_subbmission_number	-1.877103e-01
Apartment_project_subbmission_area	-1.854389e-01
Employed_number_total	-1.783248e-01
Employed_number_men	-1.781154e-01
Employed_number_women	-1.785090e-01
Employed_agricultural_number	2.090461e-02
Employed_industry_number	-1.798586e-01
Emplyed_trade_transport_number	-1.792398e-01
Employed_finance_number	-1.742926e-01
Employed_other_number	-1.815776e-01
Average_income	-9.732788e-02
Total_population_age_0_14_years	-1.788971e-01
Total_population_age_15_29_years	-1.847582e-01
Total_population_age_30_44_years	-1.824001e-01
Total_population_age_45_59_years	-1.834639e-01

6. **Wybór zmiennych do modelu i badanie współliniowości**

   Zmienne numeryczne:

   # Installment_amount

   # IIR

   # SIR

   # Available_income_amount

unemployed_to_working_age
Average_income
Divorce_number_1000
Building_permit_individual_number
DPD_t0
DPD_lag_2
NotionalValue_t0
NotionalOverdue_t0
NotionalOverdue_lag12

Zmienne kategoryczne, przekształcone na dummy variables:
zakres_wiekowy
liczba_dzieci
Job_type
Marital_status
Home_status
Car_status
Credit_purpose

Potencjalne interakcje między zmiennymi:
Age~Maritial_status
Age~Credit_purpose
Age~Household_children
Job_type~Credit_amount
Monthly_income~Credit_purpose
Maritial_status~Household_children
Monthly_income~Monthly_spendings
Monthly_income~Credit_amount
Monthly_spendings~Credit_amount
Job_type~Employed_number_total
Car_status~Personal_car_number
Age~NotionalValue_lag12
Age~DPD_lag12
Age~NotionalOverdue_lag12
Job_type~NotionalValue_lag12
Job_type~DPD_lag12
Job_type~NotionalOverdue_lag12

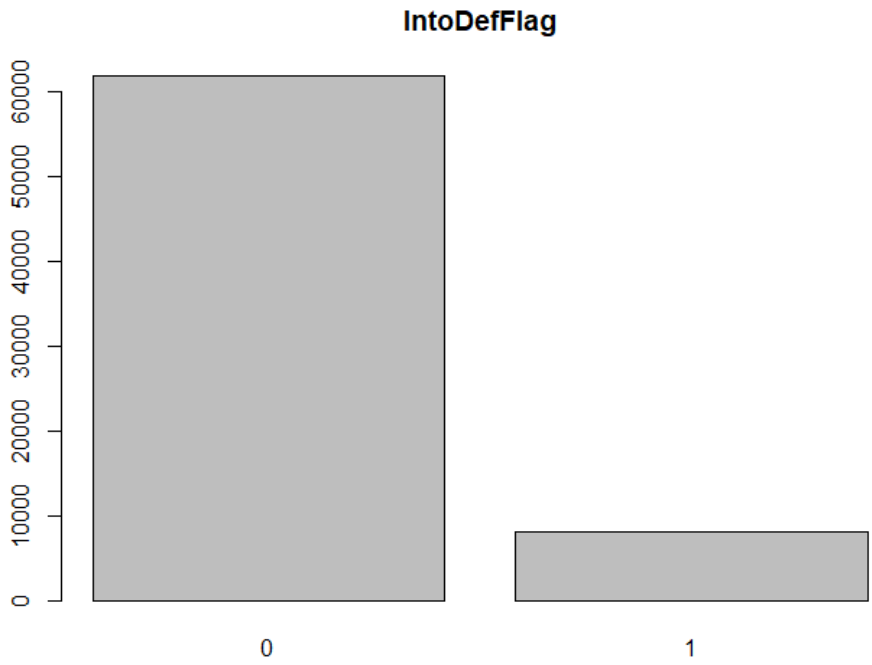
SIR~NotionalValue_lag12

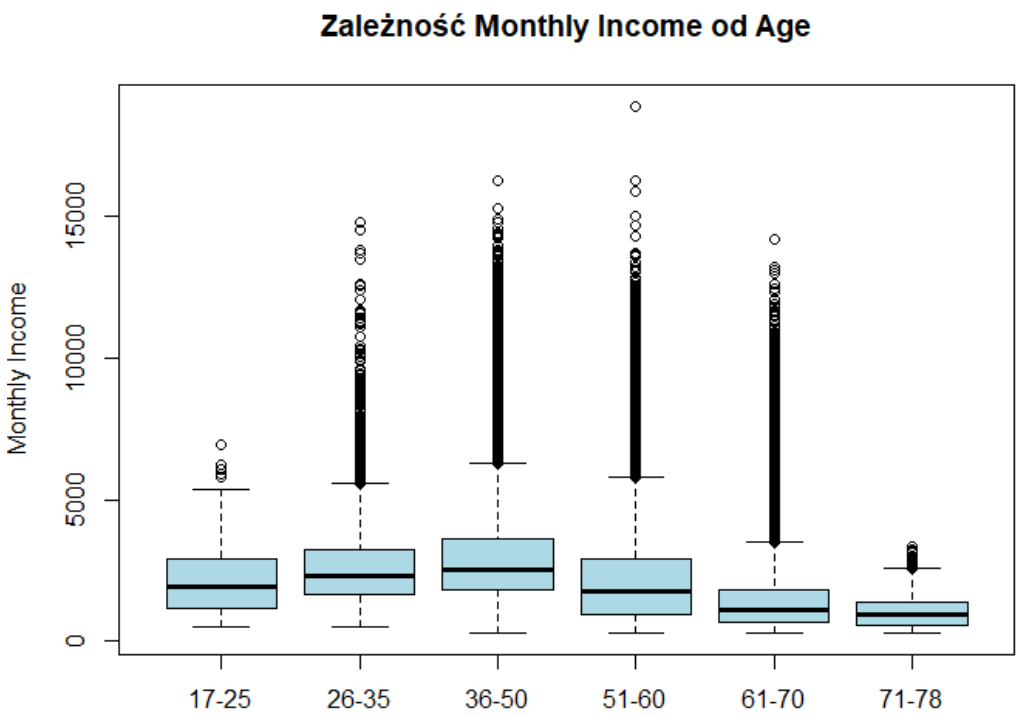
Badanie współliniowości dla zmiennych numerycznych:

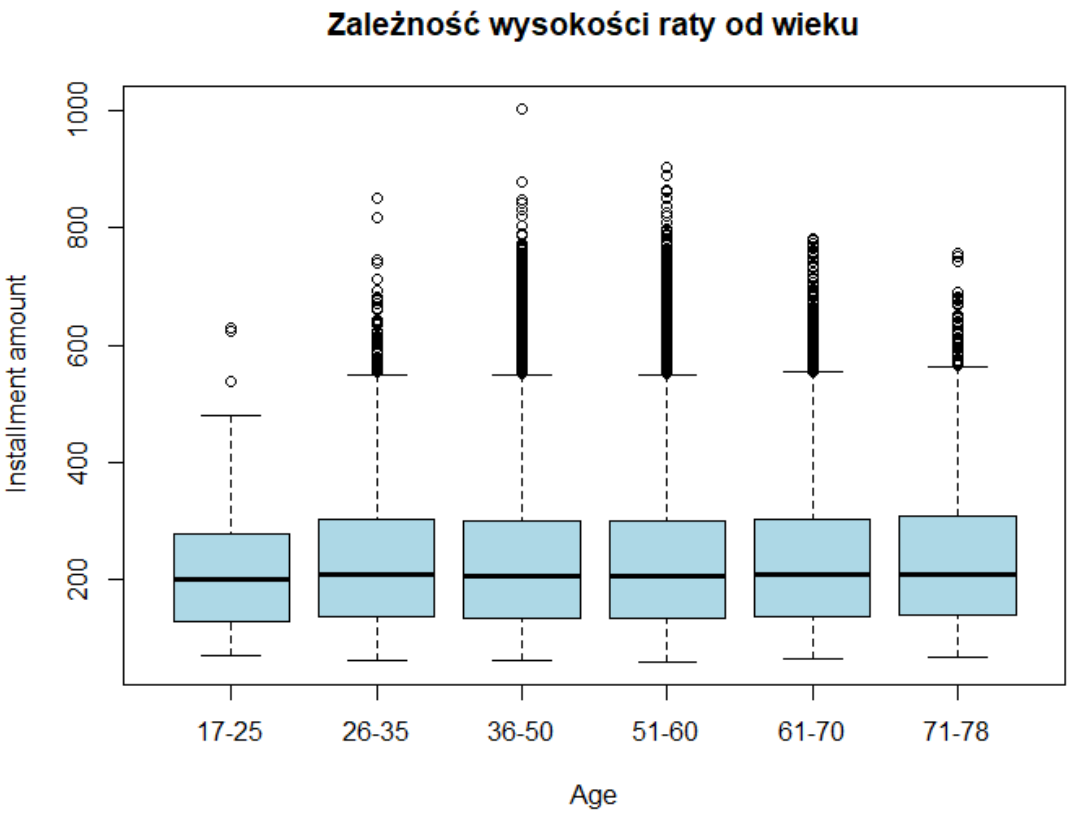
	vif_values
data1$Installment_amount	1.448399
data1$IRR	1.938547
data1$SIR	1.148179
data1$Available_income_amount	1.645974
data1$unemployed_to_working_age	1.408714
data1$Average_income	1.628078
data1$Divorce_number_1000	1.519996
data1$Building_permit_individual_number	1.229269
data1$DPD_t0	1.024450
data1$DPD_lag12	1.064397
data1$NotionalValue_t0	1.015505
data1$Notionaloverdue_t0	1.031384
data1$Notionaloverdue_lag12	1.058884

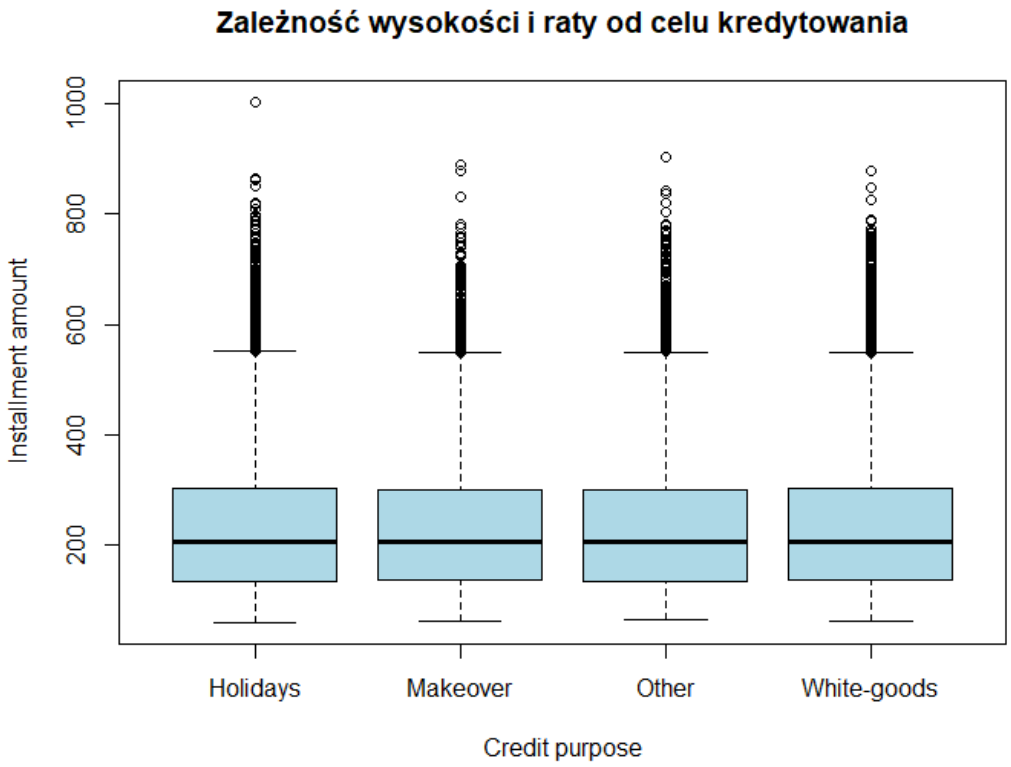
W literaturze przedmiotu przyjmuje się wartości VIF > 2,5, jako przesłanka współliniowości zmiennych w tym przypadku ona nie występuje, bo współczynniki VIF dla wszystkich zmiennych są mniejsze od 2.

7. Analiza zmiennej celu i wykresy zależności między zmiennymi
    - Wykres iledeafultów vs ileniedefaultów




Zależność Monthly Income od Age
Monthly Income
0
5000
10000
15000
17-25
26-35
36-50
51-60
61-70
71-78


Zależność wysokości raty od wieku
Installment amount
200
400
600
800
1000
17-25
26-35
36-50
51-60
61-70
71-78
Age



8. Podział zbioru na uczący oraz testowy w stosunku 70% do 30%.
9. Zbudowano trzy modele regresji logistycznej, z czego wszystkie zostały poddane eliminacji zmiennych metodą krokową, która bazuje na kryterium informacyjnym Akaike. Zbudowano następujące modele:
    - Model „klasyczny”, zawierający zmienne ze zbioru,
    - Model „z interakcjami”, uwzględniający zmienne z modelu klasycznego oraz interakcje między zmiennymi ze zbioru danych,
    - Model „pca”, który w zbiorze zmiennych objaśniających ma dziesięć głównych składowych wyodrębnionych w analizie pca.

    Dokonano oceny modeli bazując na kryterium AIC (im wyższe tym lepszy model) oraz na miarach pseudo R-kwadrat (im niższe tym lepszy model):

	kryterium_AIC	McFadden	Cragg_Uhler
logit_basic_final	32008.08	0.09940506	0.13469540
logit_interactions_final	31981.77	0.10116060	0.13698836
logit_pca_final	33680.96	0.05153874	0.07104235

Dla trzech finalnych modeli najlepsze wyniki ma model, który uwzględnia główne składowe zamiast zwykłych zmiennych.

Model z PCA jest przydatny, gdy chcemy zredukować liczbę zmiennych i usunąć współliniowość. Jest to szczególnie pomocne w dużych zbiorach danych z wieloma zmiennymi,

gdzie tradycyjne modele regresji mogą być przeciążone lub zawierać wiele skorelowanych predyktorów. Jak pokazała analiza VIF zmienne użyte w modelu nie są współliniowe, a dodatkowo ich ilość pozwala na swobodną interpretację. Interpretacja modelu PCA wymaga uwzględnienia informacji, że każda główna składowa to kombinacja wielu oryginalnych zmiennych, co zdecydowanie komplikuje jego stosowanie, zwłaszcza w kontekście modelu scoringowego. Dlatego do dalszego porównania zdecydowano wybrać się dwa pozostałe modele.

10. Weryfikacja modelu klasycznego oraz modelu z interakcjami na zbiorze testowym
    - Wyznaczono punkt odcięcia, który będzie stanowił graniczny punkt klasyfikacji klientów na „dobrych" i „złych" na poziomie 50%.
    - Wyznaczono tablice trafności dla obu modeli

      Dla modelu bez interakcji:

obserwowane	przewidywane 0	1
0	43120	264
1	5518	237

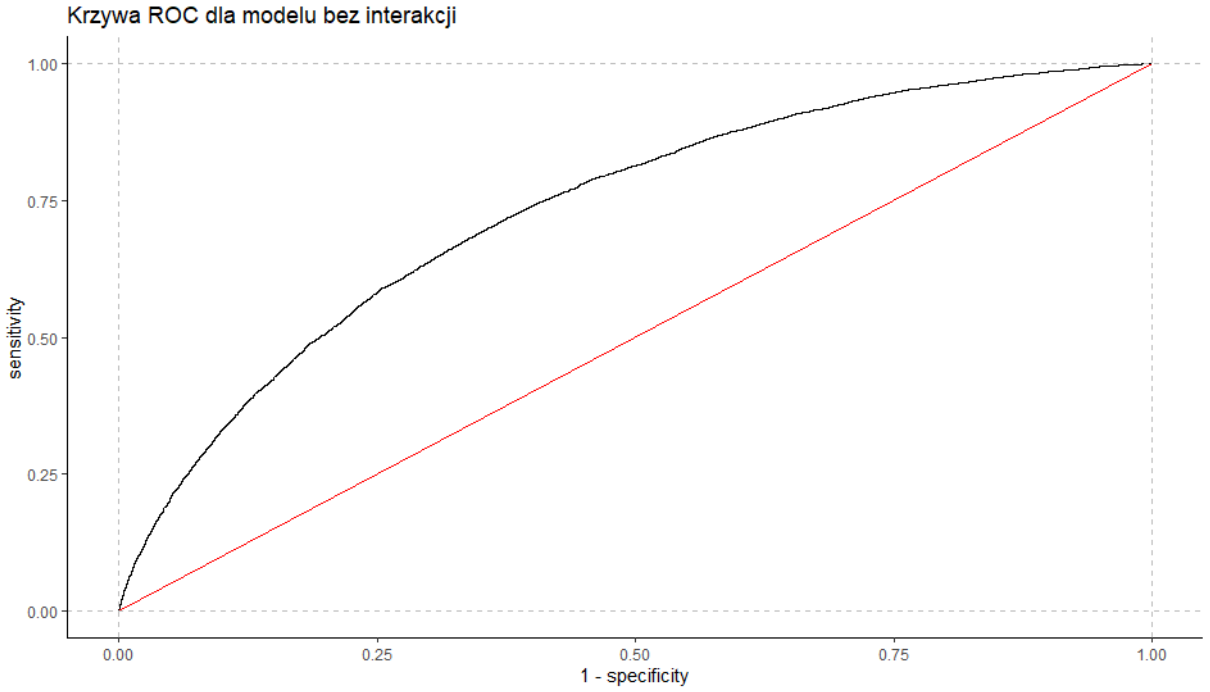
      Dla modelu z interakcjami:

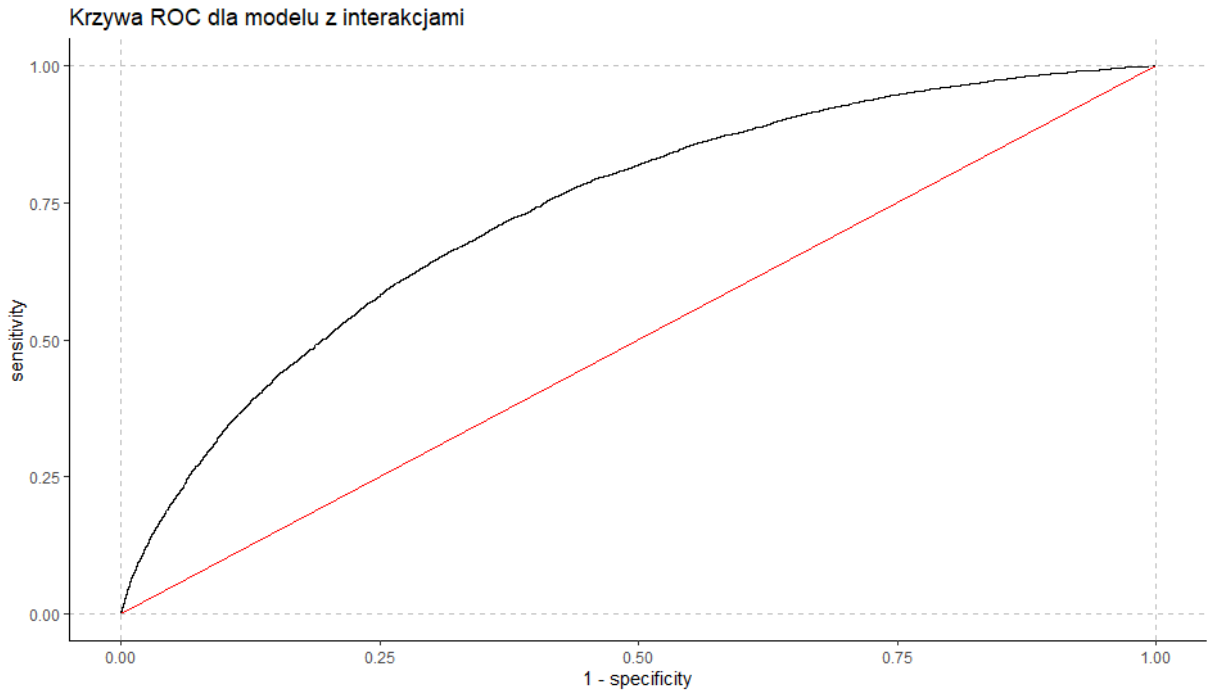
obserwowane	przewidywane 0	1
0	43120	264
1	5518	237

    - Na podstawie tabeli trafności wyznaczono miary jakości predykcji:

	ACC	ER	SENS	SPEC	PPV	NPV
logit_basic_final	0.8834188	0.1165812	0.03257732	0.9953352	7.56701	0.004285094
logit_interactions_final	0.8835147	0.1164853	0.03340206	0.9953352	7.56701	0.004393578

Przy obranym punkcie odcięcia p, miary jakości predykcji nieznacznie się różnią dla obu modeli, z tego też powodu przeanalizowana zostanie krzywa ROC i uzyskane na na jej podstawie miary.





Oba wykresy wskazują, że krzywa ROC jest powyżej linii losowości, co wskazuje na to, że modele mają dobrą zdolność rozróżniania między zdarzeniami pozytywnymi a negatywnymi. W celu weryfikacji który jest lepszy zastosowana zostanie miara AUC i wsp. Giniego.

```
                                        [,1]
pole_AUC_logit_final              0.7331042
pole_AUC_logit_interactions_final 0.7350674
```

Miara AUC zawiera się między 0,5 a 1. Nie ma jednoznacznego poziomu AUC, który określałby, że predykcja modelu jest na odpowiednim poziomie. Jest to zależne często od kontekstu analizowanego problemu. W tym przypadku zostanie przyjęte, że poziom 0,75, oznacza odpowiednie dopasowanie modelu, ale zostanie też policzony wskaźnik Giniego, który również jest miarą predykcji modelu.

```
                                        [,1]
Wsp_Ginniego_logit_basic_final        0.4662083
Wsp_Ginniego_logit_interactions_final 0.4701347
```

Wskaźnik Giniego na poziomie ok. 47% oznacza że model ma pewną zdolność do rozróżniania między dobrymi a złymi kredytobiorcami, ale nie jest to zdolność wyjątkowo wysoka. Żeby model lepiej przewidywał "dobrych" i "złych" klientów należałoby spróbować go zmodyfikować.

Oto możliwe ulepszenia:

- Dodanie nowych zmiennych,
- regularyacja
- zbalansowanie zmiennej celu,
- dodatkowe analizy/metody

Podsumowanie testów:

Kryterium AIC wskazało na to, że model bez interakcji jest lepszy, tak samo jak miary pseudo R-kwadrat, a dodatkowo pozwala on na dużo łatwiejszą interpretację zmiennych.

Zatem jako model finalny wybrany został model "logit_basic_final", ponieważ weryfikacja na zbiorze testowym nie wykazała istotnych różnic między dwoma modelami.