

Ranking istotności cech aplikacyjnych w prognozie skuteczności sprawy

Zuzanna Nogala

Zbiór danych i przedstawienie problemu

- Cechy aplikacyjne - cechy spraw zakupionych przez firmę Kruk (19 atrybutów przypisanych do każdej sprawy)

Zmienne aplikacyjne

LoanAmount
TOA
Principal
Interest
Other } Kwota pożyczki i składowe zadłużenia

ExternalAgency
Bailiff
ClosedExecution } Zmienne związane z działalnością windykacyjną

Zmienne aplikacyjne

PopulationInCity	Zmienne ankietowe
MeanSalary	
GDPPerCapita	
Land	
Age	
Gender	
Product	Aktywność sprawy przed zakupem
DPD	
D_ContractDateToImportDate	
LastPaymentAmount	
M_LastPaymentToImportDate	

Zbiór danych i przedstawienie problemu

- Zbiór spraw zakupionych przez firmę Kruk z 19 atrybutami przypisanych do każdej sprawy (zmienne aplikacyjne)
- Skuteczność - **zmienna binarna**, która będzie określać rokuję sprawa na podstawie 12 miesięcy obsługi, połączenie dwóch miar skuteczności:

$$Suc_1 = \begin{cases} 1 & \text{jeżeli } \frac{\text{Suma Wpłat}}{\text{Zadłużenie początkowe}} \geq 0.05 \\ 0 & \text{w.p.p} \end{cases}$$

$$Suc_2 = \begin{cases} 1 & \text{jeżeli liczba wpłat} \geq 2 \\ 0 & \text{w.p.p} \end{cases}$$

Zbiór danych i przedstawienie problemu

Skuteczność sprawy w 12 miesiącach to połączenie dwóch wskaźników skuteczności.

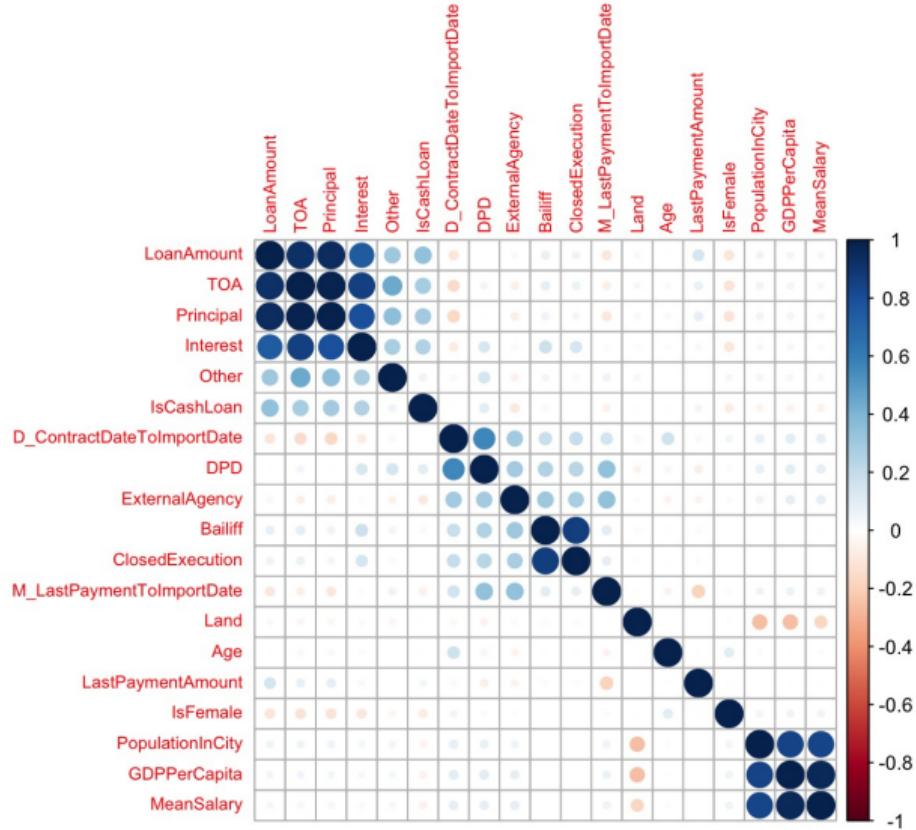
$$Suc = \begin{cases} Suc_2 & \text{jeżeli } Suc_2 > Suc_1 \\ Suc_1 & \text{jeżeli } Suc_2 \leq Suc_1 \end{cases}$$

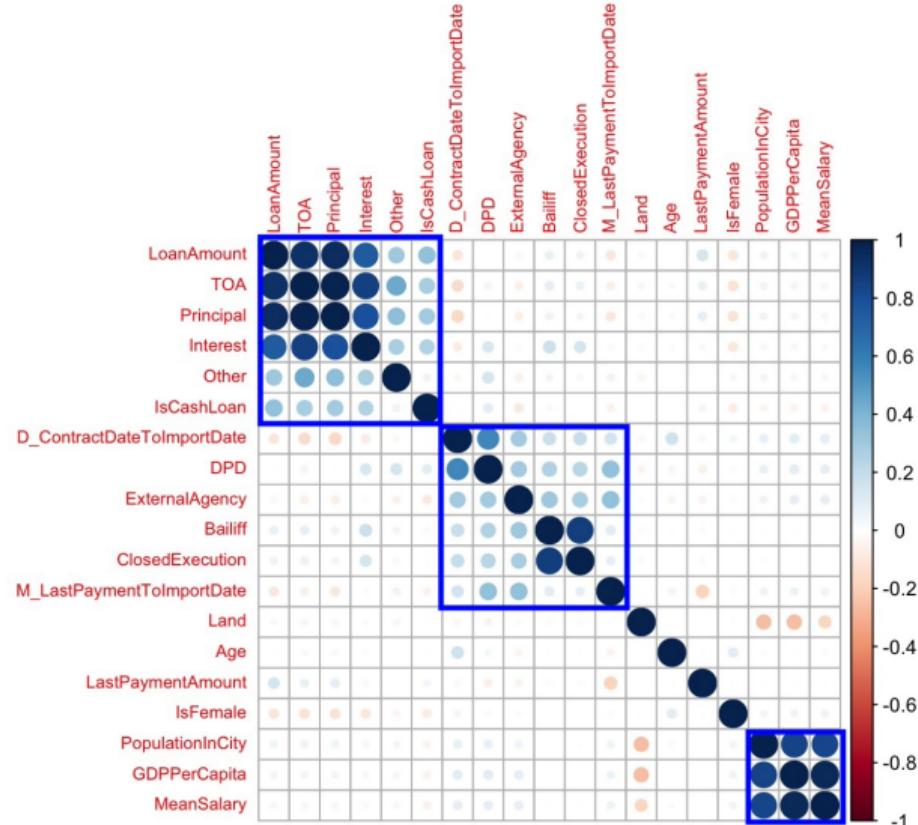
Zbiór danych i przedstawienie problemu

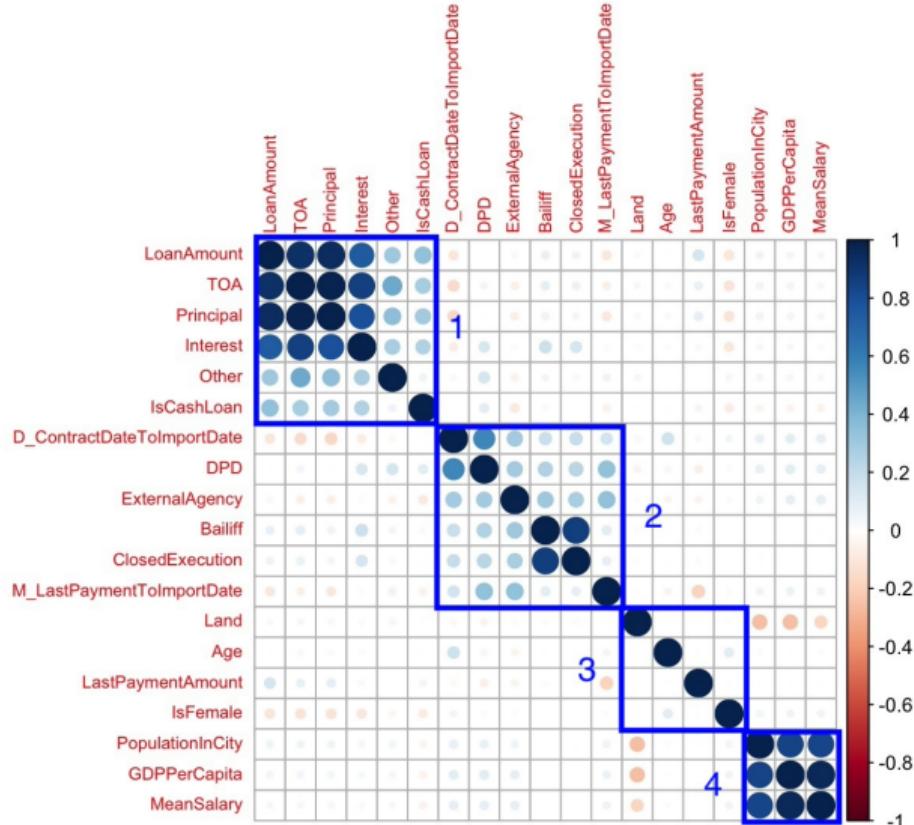
Skuteczność sprawy w 12 miesiącach to połączenie dwóch wskaźników skuteczności.

$$Suc = \begin{cases} Suc_2 & \text{jeżeli } Suc_2 > Suc_1 \\ Suc_1 & \text{jeżeli } Suc_2 \leq Suc_1 \end{cases}$$

- Ranking zmiennych (najlepsza cecha dostaje 1 miejsce)
 - *Korelacja*
 - *Boosting*
 - *KNN*
 - *Model logistyczny*
- Ranking koszyków korelacji cech







Ocena Istotności cech w modelach

1. Z funkcji *summary*

Ocena Istotności cech w modelach

1. Z funkcji *summary*
2. Porównywanie błędu na zbiorze testowym L z błędem na zbiorze testowym z permutowaną/ usuniętą jedną zmienną/ jednym koszykiem C - $L^{(C)}$

$$\frac{L^{(C)}}{L} = \begin{cases} \text{Zmienna } C \text{ istotnie wpływa na model} & \text{jeżeli } \frac{L^{(C)}}{L} > 1 \\ \text{Zmienna } C \text{ nie wpływa na model} & \text{jeżeli } \frac{L^{(C)}}{L} \leq 1 \end{cases}$$

Ocena Istotności cech w modelach

1. Z funkcji *summary*
2. Porównywanie błędu na zbiorze testowym L z błędem na zbiorze testowym z permutowaną/ usuniętą jedną zmienną/ jednym koszykiem C - $L^{(C)}$

$$\frac{L^{(C)}}{L} = \begin{cases} \text{Zmienna } C \text{ istotnie wpływa na model} & \text{jeżeli } \frac{L^{(C)}}{L} > 1 \\ \text{Zmienna } C \text{ nie wpływa na model} & \text{jeżeli } \frac{L^{(C)}}{L} \leq 1 \end{cases}$$

3. Analogiczne porównanie wartości AUC (*ang. Area under ROC*)

$$\frac{AUC^{(C)}}{AUC} = \begin{cases} \text{Zmienna } C \text{ istotnie wpływa na model} & \text{jeżeli } \frac{AUC^{(C)}}{AUC} < 1 \\ \text{Zmienna } C \text{ nie wpływa na model} & \text{jeżeli } \frac{AUC^{(C)}}{AUC} \geq 1 \end{cases}$$

Uzupełnienie braków danych

Bailiff	Uzupełnienie zgodnie z rozkładem
ClosedExecution	
ExternalAgency	
Land	
Age	
Gender(=IsFemale)	
M_LastPaymentToImportDate	Uzupełnienie z uwzględnieniem innej zmiennej
MeanSalary	
GDPPerCapita	
LoanAmount	

Uzupełnienie braków danych

D_ContractDateToImportDate - Uzupełnienie medianą

Others - Uzupełnienie z zależności zmiennej z innymi składowymi zadłużenia początkowego

LastPaymentAmount

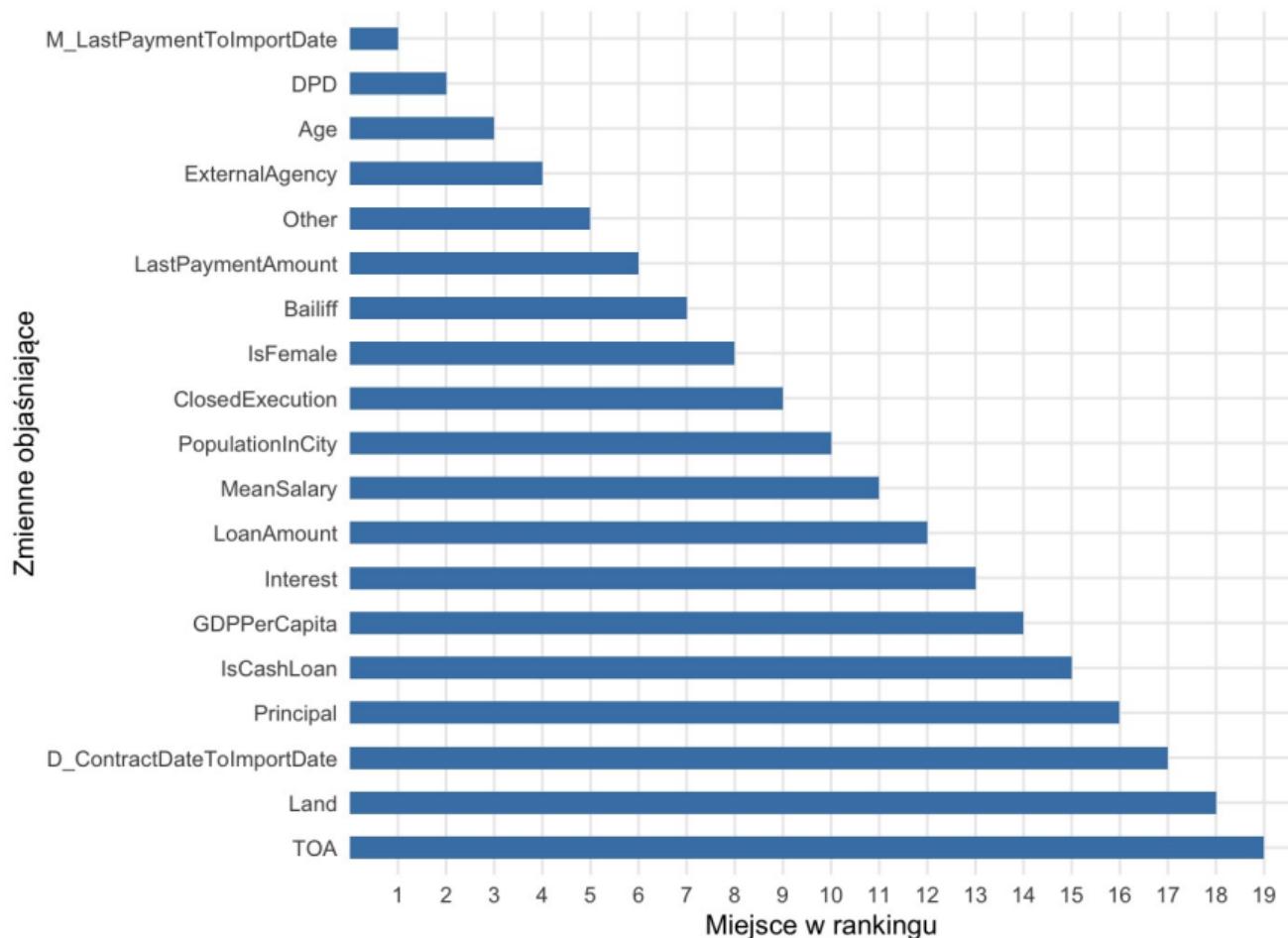
DPD

PopulationInCity

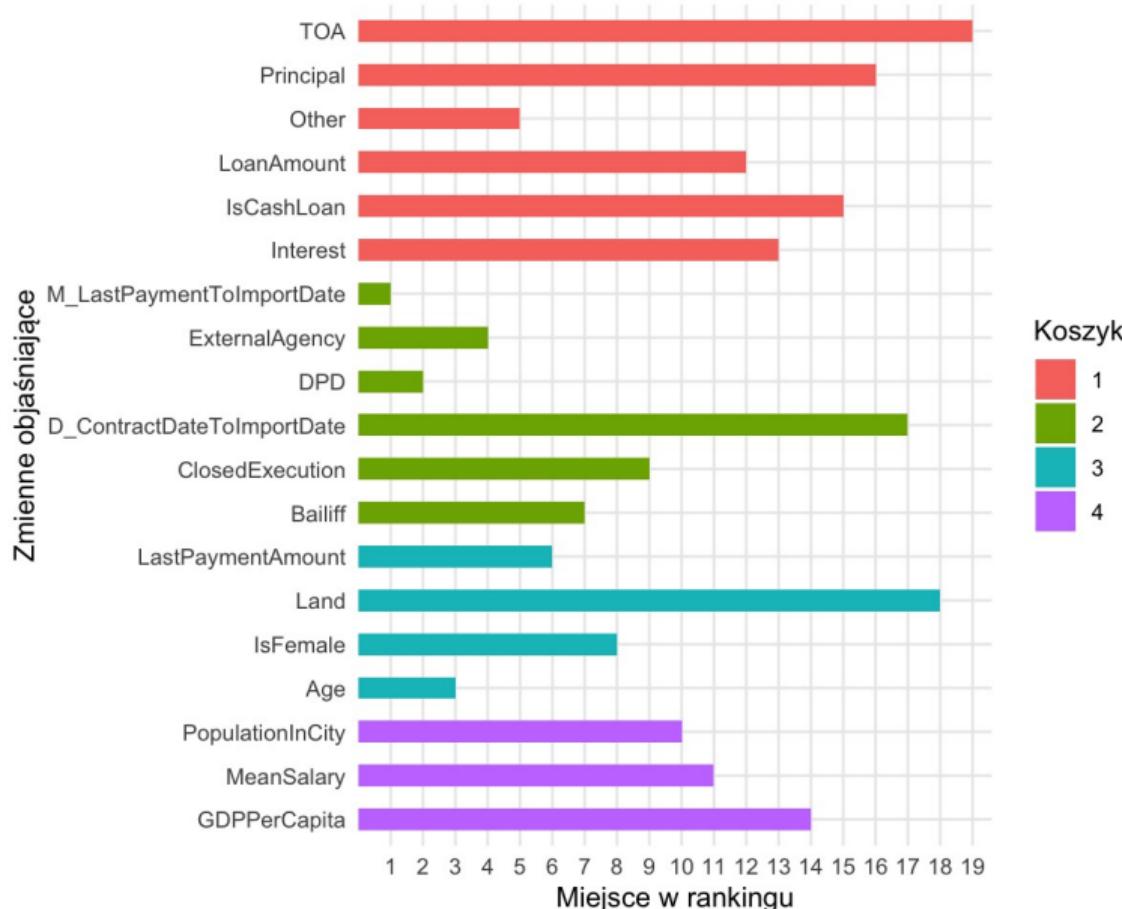
} Uzupełnienie modelem na mocnych cechach

Korelacja zmiennych aplikacyjnych ze skutecznością

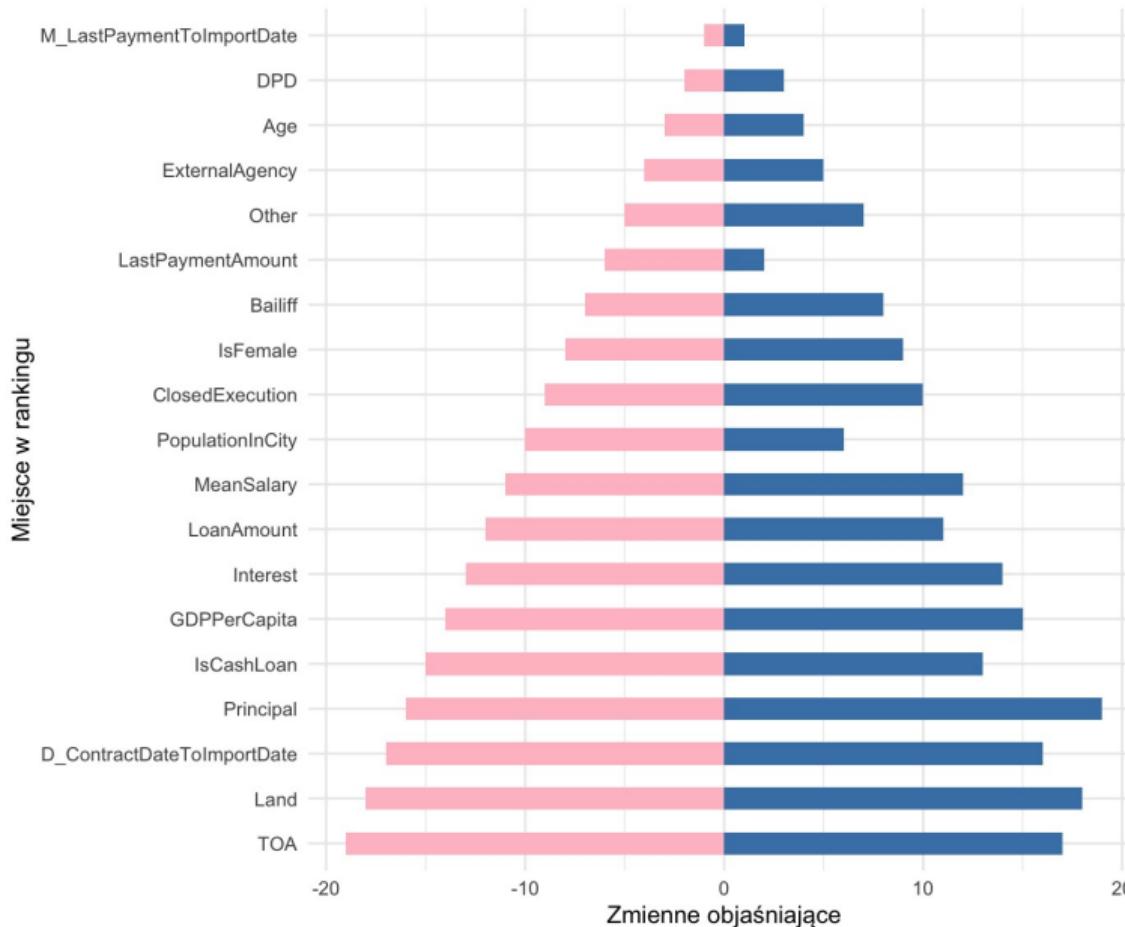
Pojedyncze zmienne - korelacja Pearsona



Koszyki - korelacja Pearsona

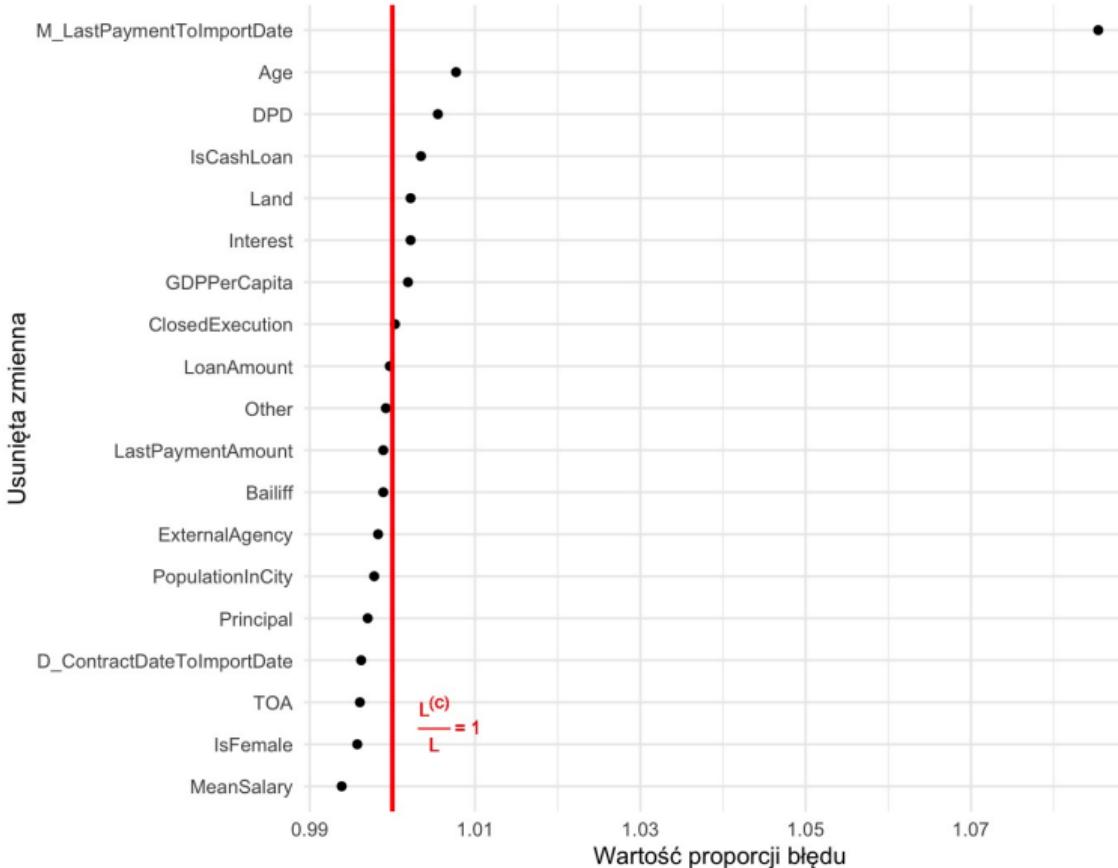


Korelacja Pearsona i Spearmana - porównanie

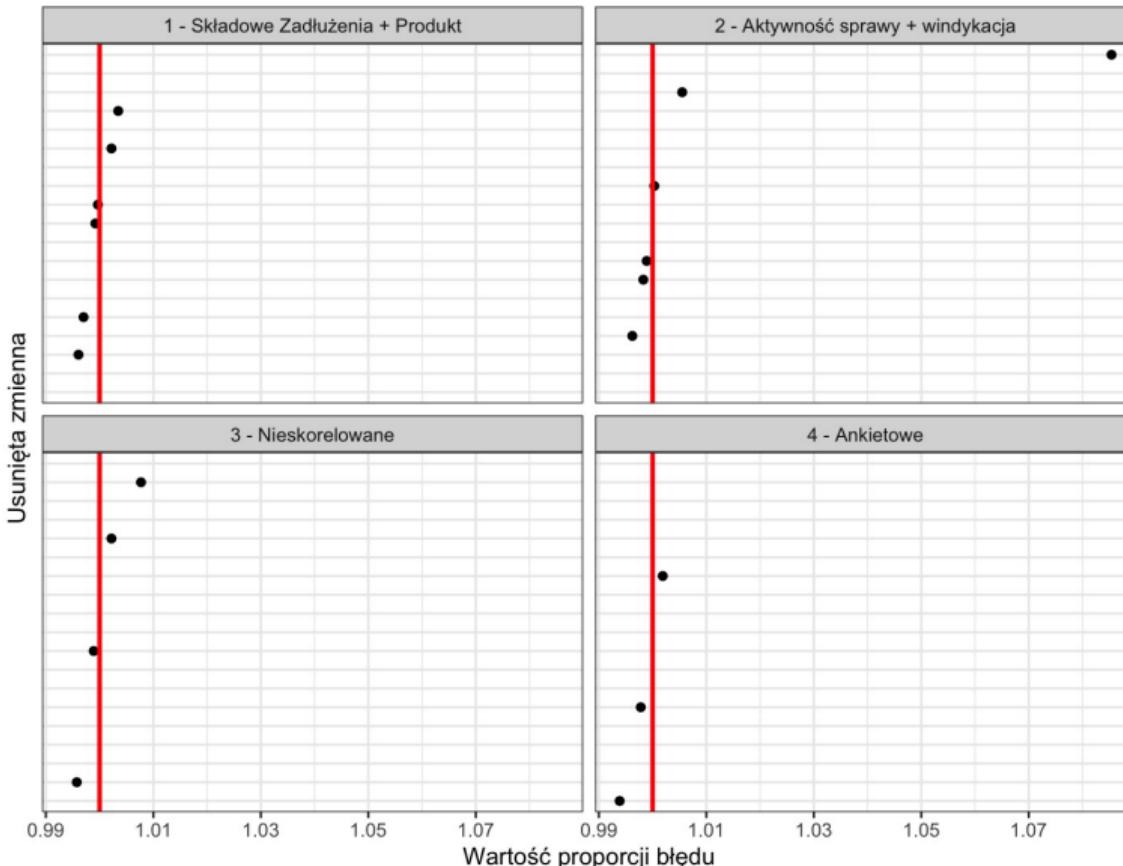


Model k najbliższych sąsiadów

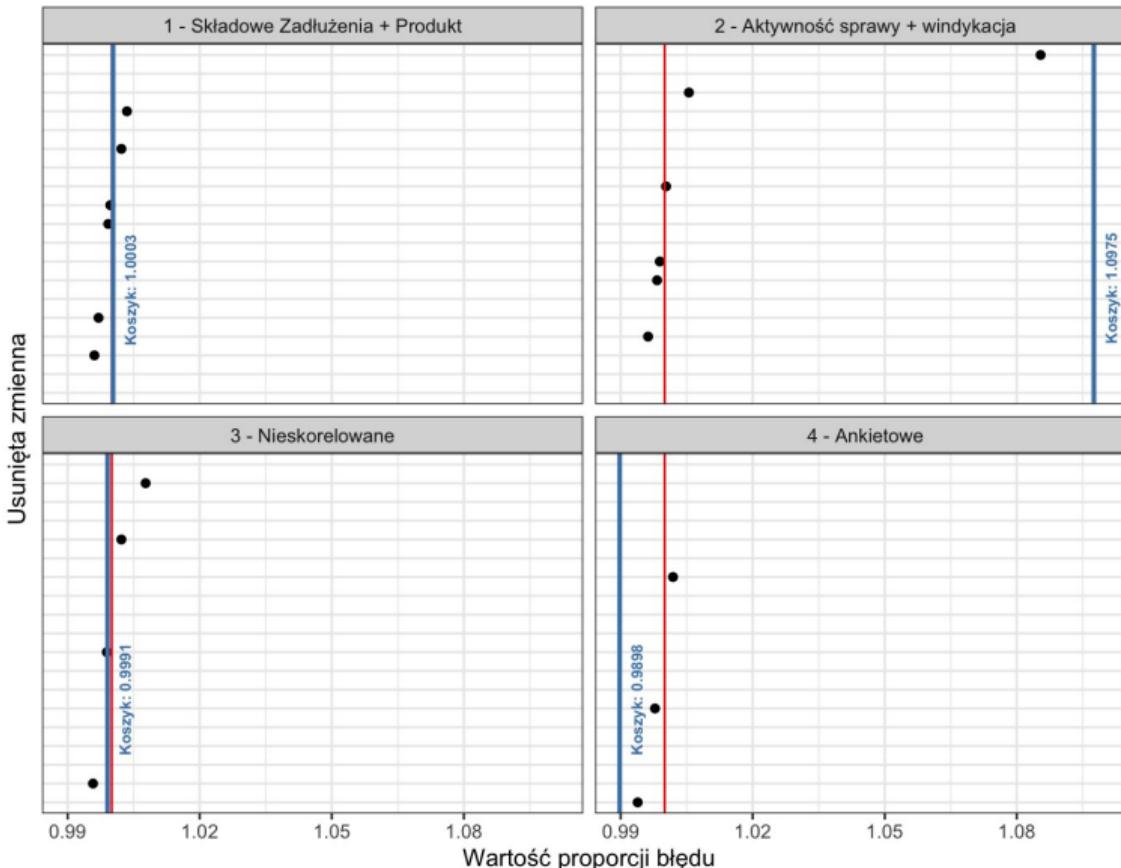
Proporcja błędu $\frac{L^{(c)}}{L}$ - usunięcie pojedynczej zmiennej



Proporcja błędu $\frac{L^{(c)}}{L}$ - usunięcie pojedynczego koszyka

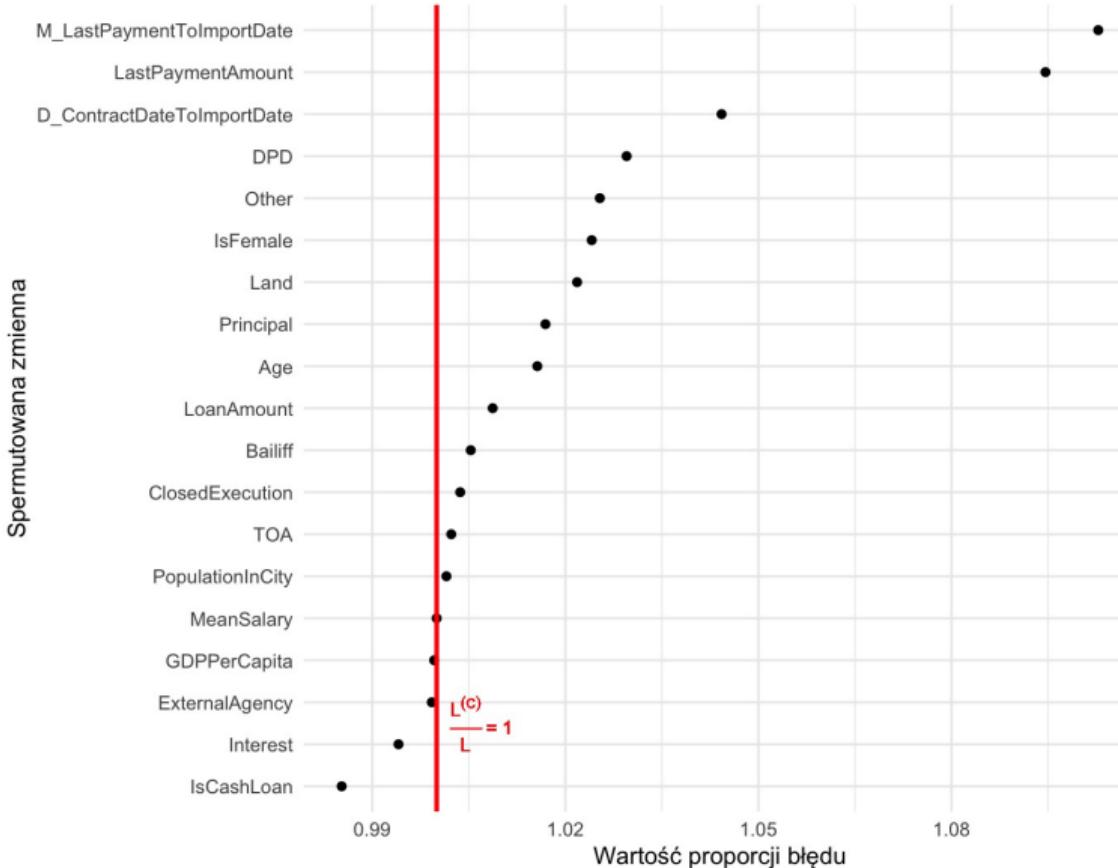


Proporcja błędu $\frac{L^{(c)}}{L}$ - usunięcie pojedynczego koszyka

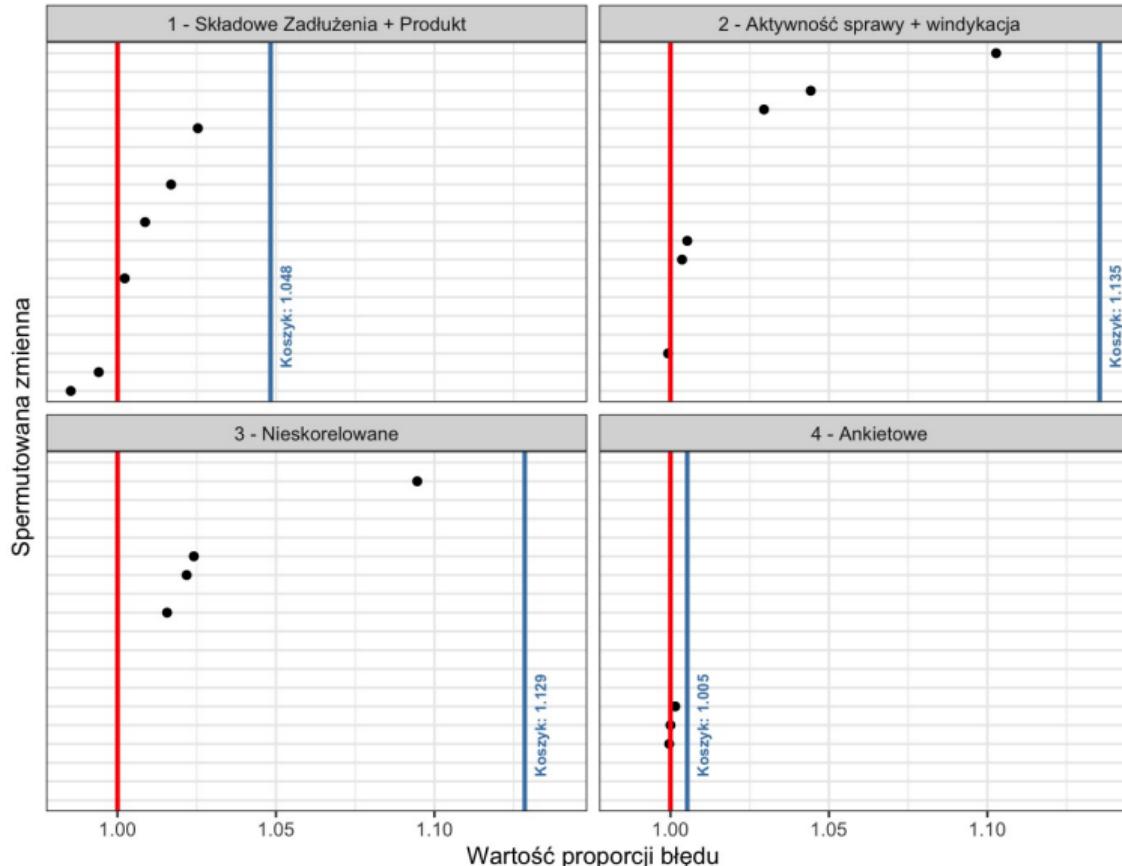


Model Boosting

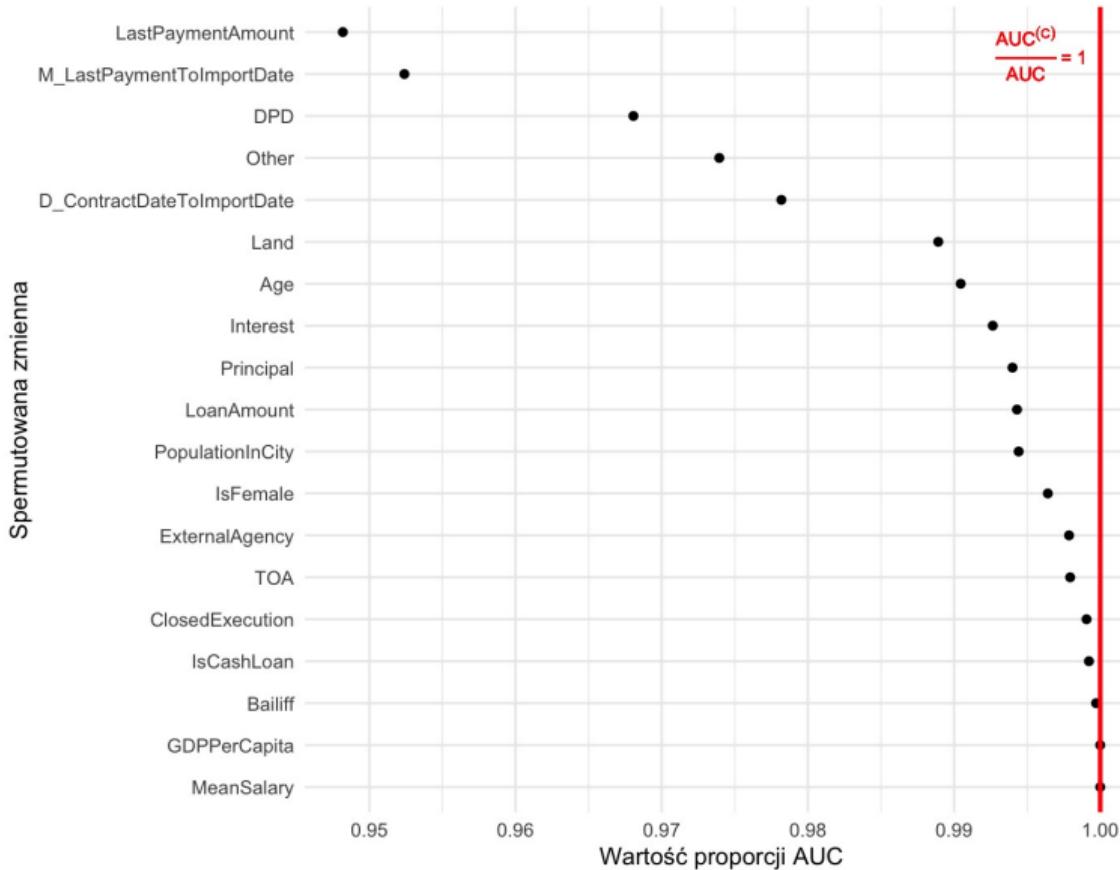
Proporcja błędu $\frac{L^{(c)}}{L}$ - permutacja pojedynczej zmiennej



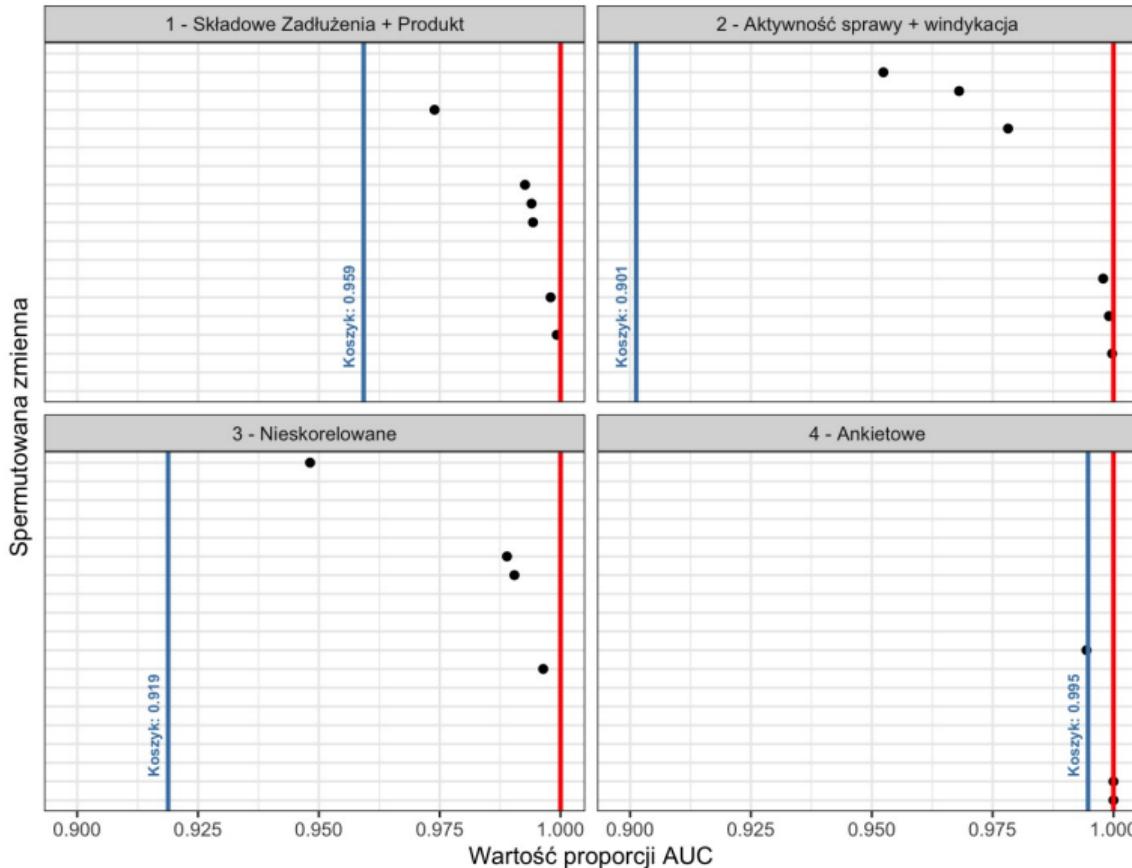
Proporcja błędu $\frac{L^{(c)}}{L}$ - permutacja pojedynczego koszyka



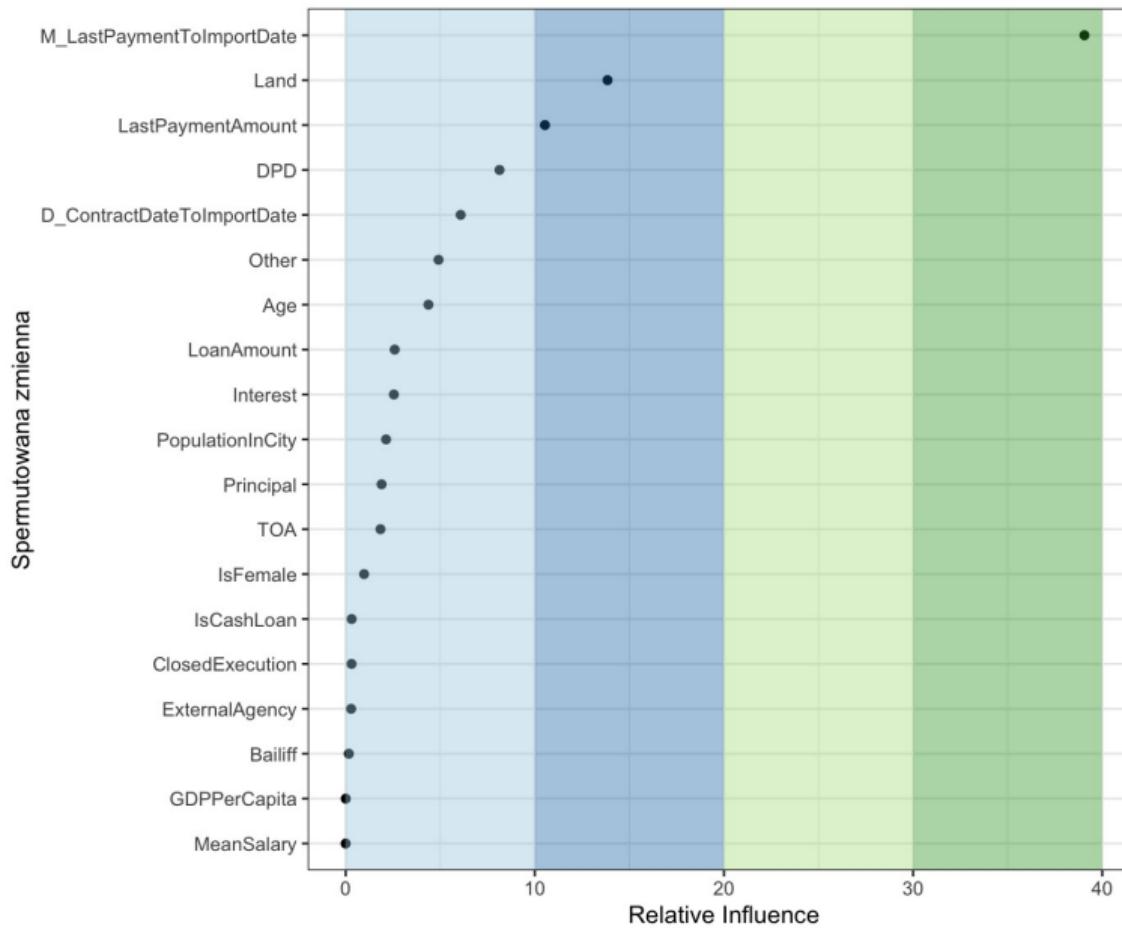
Proporcja AUC $\frac{AUC^{(c)}}{AUC}$ - permutacja pojedynczej zmiennej



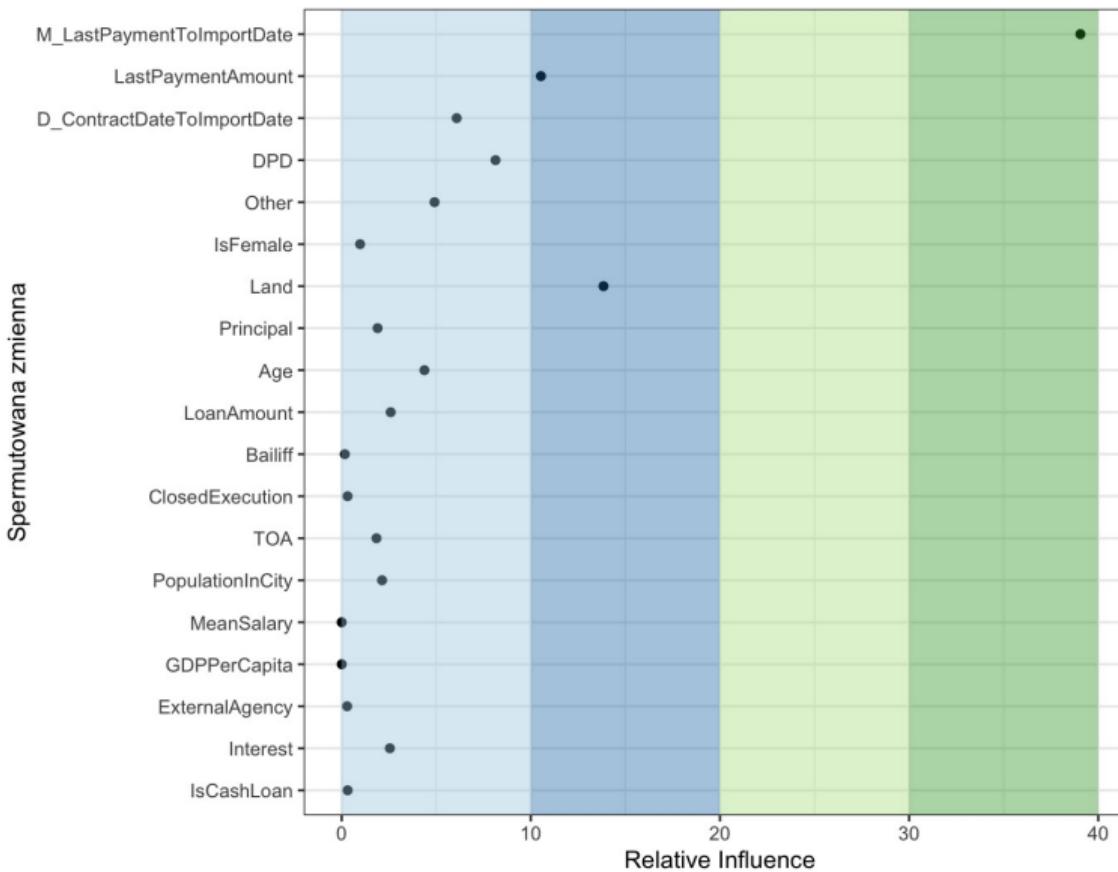
Proporcja AUC $\frac{AUC^{(c)}}{AUC}$ - permutacja pojedynczego koszyka



Istotność cechy z summary

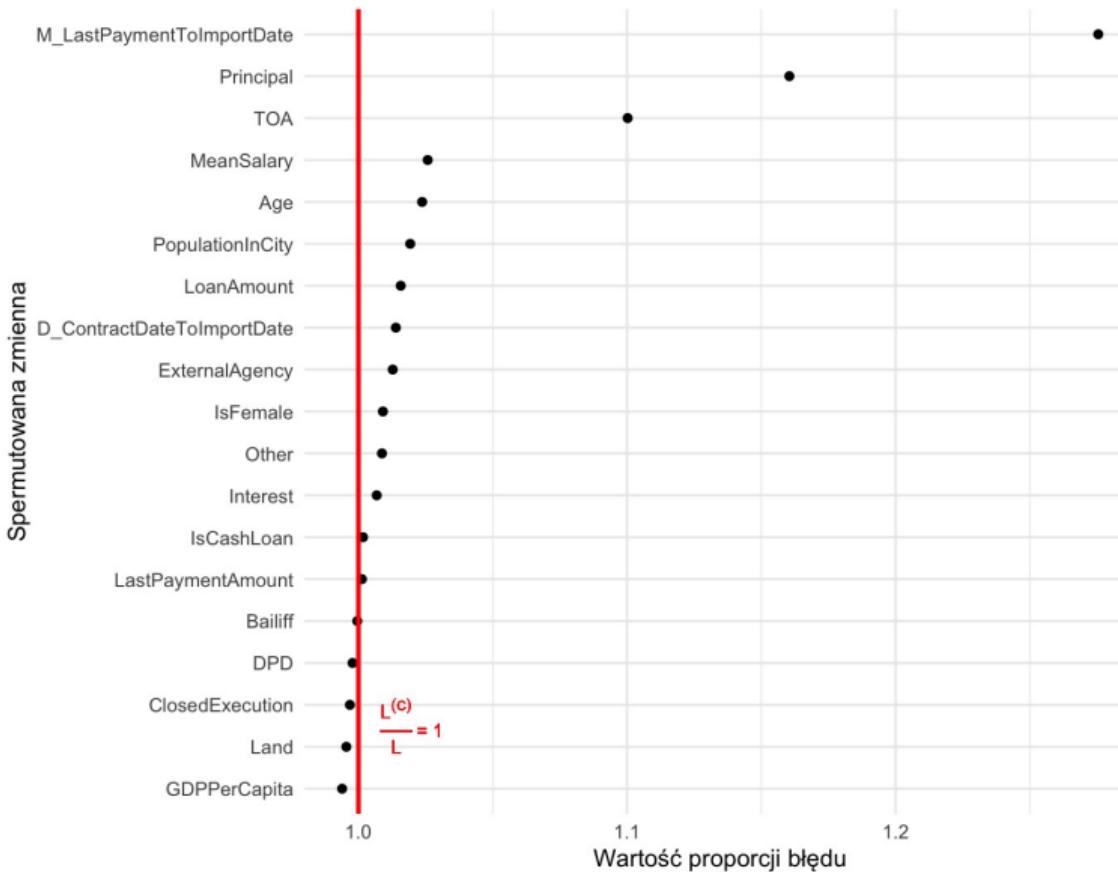


Istotność cechy z summary - sortowanie po proporcji błędu $\frac{L^{(c)}}{L}$

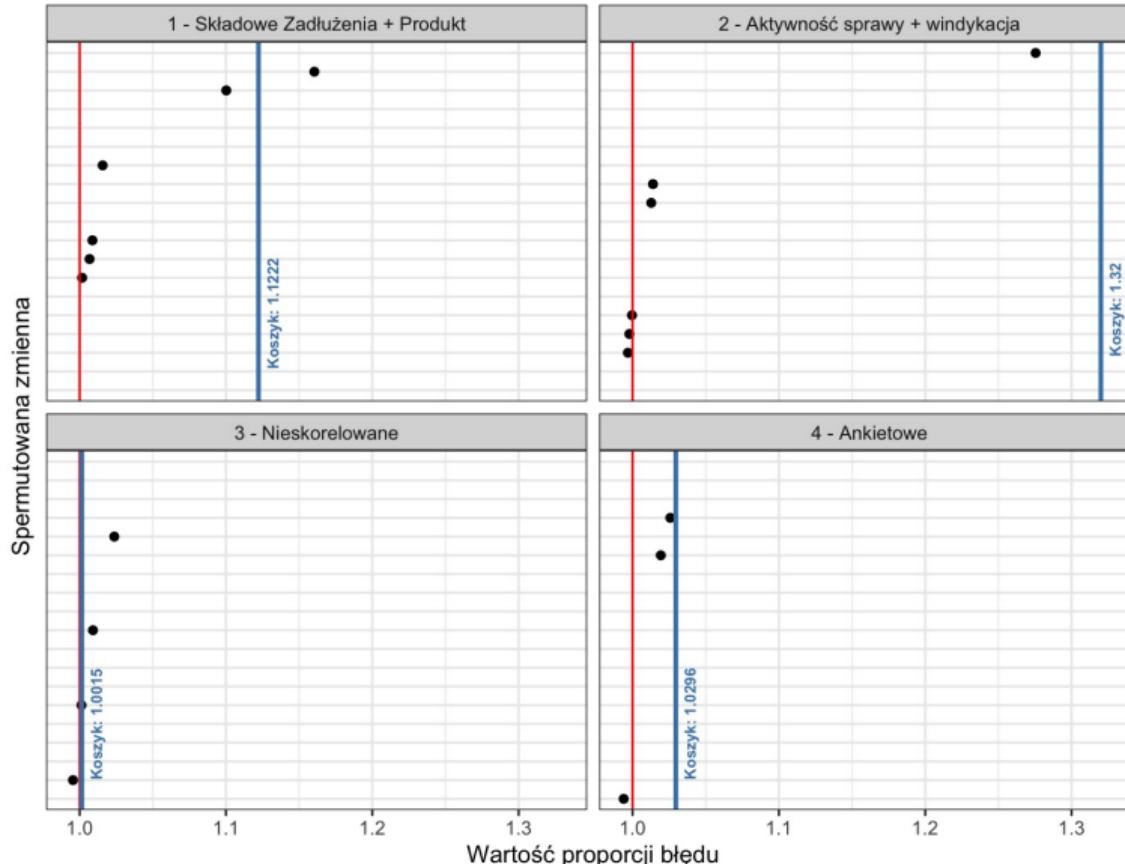


Model Logistyczny

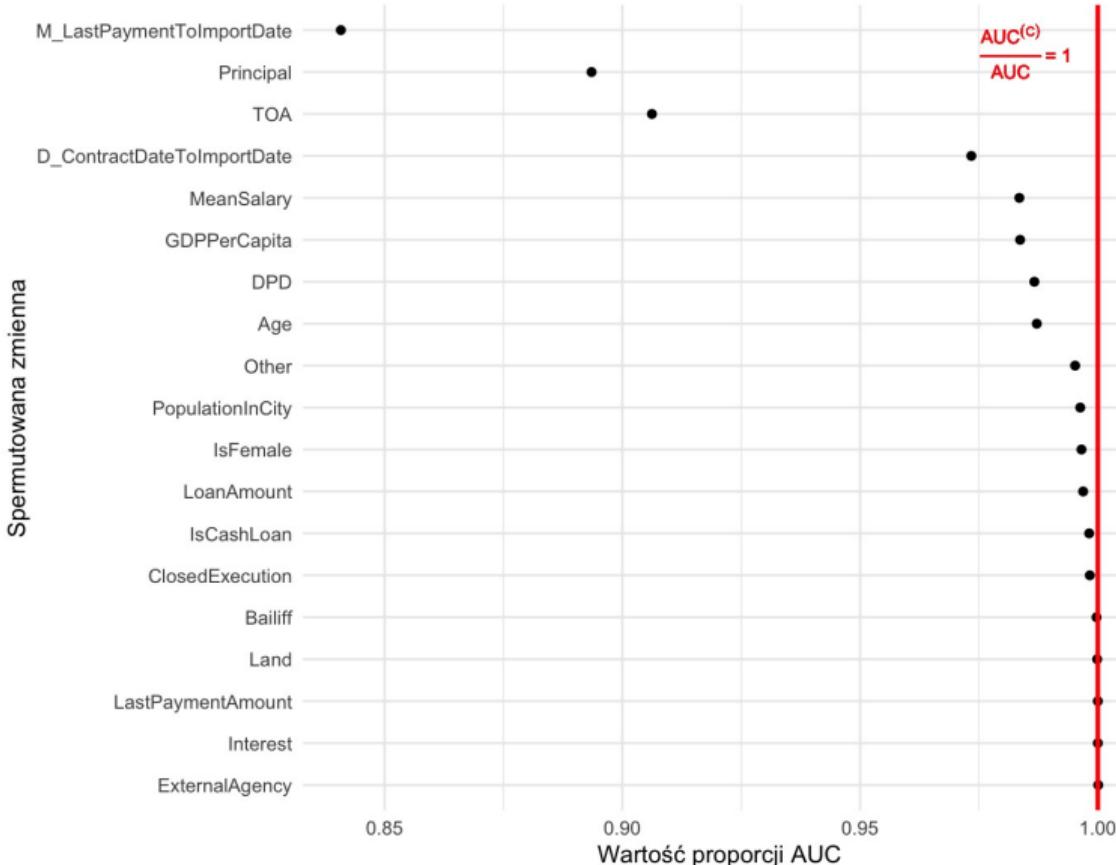
Proporcja błędu $\frac{L^{(c)}}{L}$ - permutacja pojedynczej zmiennej



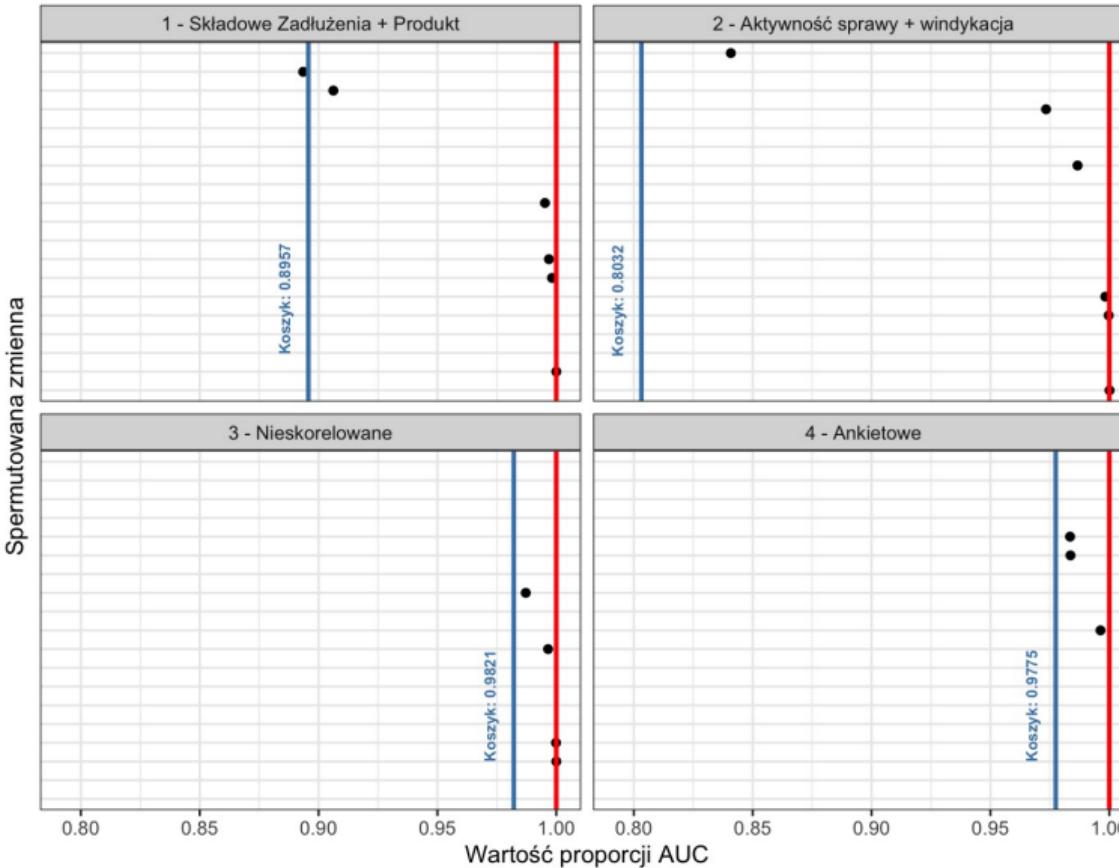
Proporcja błędu $\frac{L^{(c)}}{L}$ - permutacja pojedynczego koszyka



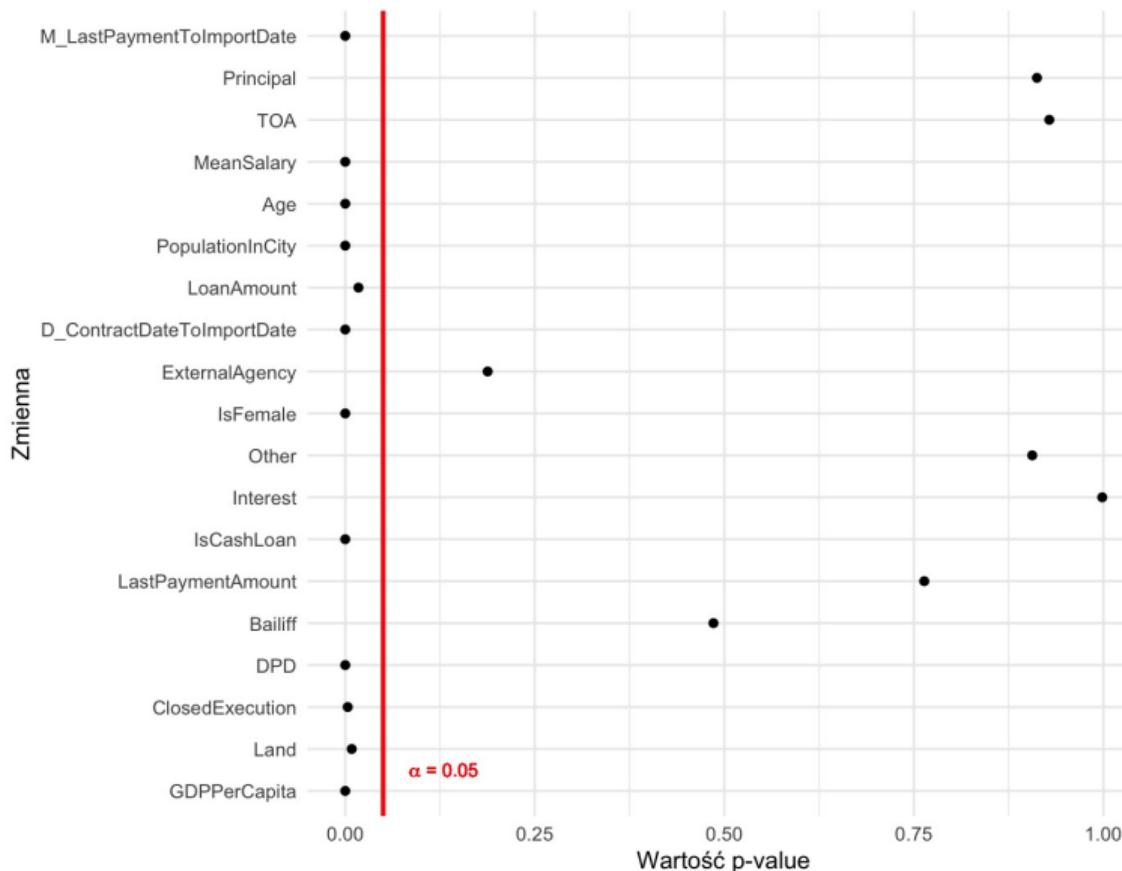
Proporcja AUC $\frac{AUC^{(c)}}{AUC}$ - permutacja pojedynczej cechy



Proporcja AUC $\frac{AUC^{(c)}}{AUC}$ - permutacja pojedynczego koszyka



Wartość p-value dla zmiennej - sortowanie po proporcji błędu $\frac{L^{(c)}}{L}$



Wyniki

Ranking

Ranking zmiennych:

1. M_LastPaymentToImportDate
2. DPD
3. Age
4. LastPaymentAmount
5. Other
6. D_ContractDateToImportDate

Ranking

Ranking koszyków:

1. Koszyk nr 2 (Aktywność sprawy z działalnością windykacyjną)
2. Koszyk nr 1 (Składowe zadłużenia z pierwotną kwotą pożyczki)
3. Koszyk nr 3 (Zmienne słabo skorelowane)
4. Koszyk nr 4 (Informacyjny)

Wnioski

- Zmienne, które są w czołówce mają bardziej charakter behawioralny niż ankietowy

Wnioski

- Zmienne, które są w czołówce mają bardziej charakter behawioralny niż ankietowy
- Statystyki oceny istotności zmiennych w modelu:
 - Proporcja błędu $\frac{L^{(C)}}{L} >$ proporcja AUC $\frac{AUC^{(C)}}{AUC}$

Wnioski

- Zmienne, które są w czołówce mają bardziej charakter behawioralny niż ankietowy
- Statystyki oceny istotności zmiennych w modelu:
 - Proporcja błędu $\frac{L^{(C)}}{L} >$ proporcja AUC $\frac{AUC^{(C)}}{AUC}$
 - Summary z Boostingu - podobne wyniki, Summary z logistycznego - ryzykowne podejście

Co można więcej?

- Jak zmienne w modelu regresji logistycznej reagują na metody krokowe?

Co można więcej?

- Jak zmienne w modelu regresji logistycznej reagują na metody krokowe?
- Porównanie statystyki proporcji błędu, gdy zmienna jest usuwana/permutowana - wybrać lepszą

Co można więcej?

- Jak zmienne w modelu regresji logistycznej reagują na metody krokowe?
- Porównanie statystyki proporcji błędu, gdy zmienna jest usuwana/permutowana - wybrać lepszą
- Boosting - Porównanie oceny zmiennych z summary i z permutacją błędów

Dziękuję za uwagę!