

Metody selekcji zmiennych do modeli skoringowych

Współczesne metody i praktyki w ryzyku kredytowym 2021

Sebastian Zając, Karol Przanowski

Warsaw School of Economics

18.05.2021

Prosty przykład

ZAŁOŻENIA:

- 20 tys. Klientów
- 348 kampanii marketingowych rocznie
- ~7 mln decyzji – wysłać czy nie?
- Koszt jednostkowy: 5
- Zarobek przy zakupie: 800
- Średnia szansa zakupu: 0,5%



Wysyłamy wszystkim

- Przychody: 28 000 000
- Koszty: 35 000 000
- Zysk: -7 000 000

Całkowicie nieopłacalne



Reguły eksperckie

- Przychody: 15 895 139
- Koszty: 12 250 000
- Zysk: 3 645 139

Zauważalne zyski



Wyniki Finansowe

Występują zauważalne zyski, ale czy można je poprawić?

Prosty przykład

Model_A – dane bazowe

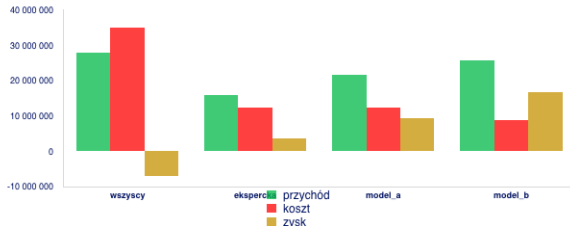
- Przychody: 21 531 163
- Koszty: 12 250 000
- Zysk: 9 281 163

Wzrost zysków o ponad 5,5 mln vs reguły eksperckie

Model_B – specjalistycznie rozbudowany zbiór danych

- Przychody: 25 599 340
- Koszty: 8 750 000
- Zysk: 16 849 340

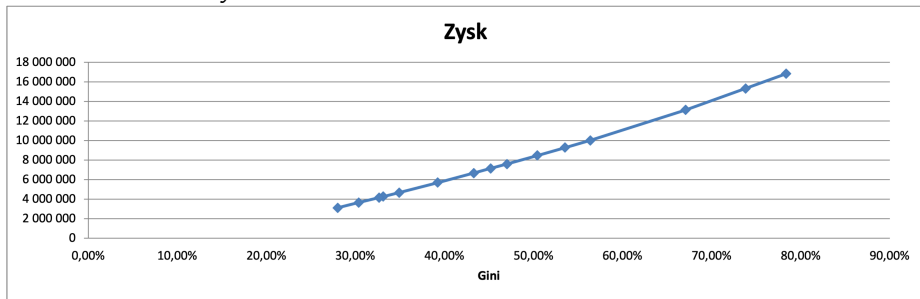
Jeszcze większy zysk oraz obniżenie kosztów wstępnych



Wyniki Finansowe

Prosty przykład

$$\Delta Gini\ 5\% = \Delta Zysk\ 1.36MLN$$



Po co selekcja zmiennych ?

Aurelien Geron

„**Z gipsu tortu nie ulepisz**”.

System zawsze uczy się jedynie za pomocą danych zawierających wystarczającą liczbę **istotnych cech** i niezaśmieconych nadmiarem cech nieistotnych.

Elementem krytycznym jest wybór dobrego zbioru cech uczących (feature engineering).”

Składa się on z :

- dobór cech (feature selection)

Po co selekcja zmiennych ?

Aurelien Geron

„**Z gipsu tortu nie ulepisz**”.

System zawsze uczy się jedynie za pomocą danych zawierających wystarczającą liczbę **istotnych cech** i niezaśmieconych nadmiarem cech nieistotnych.

Elementem krytycznym jest wybór dobrego zbioru cech uczących (feature engineering).”

Składa się on z :

- dobór cech (feature selection)
- odkrywanie cech (feature extraction)

Po co selekcja zmiennych ?

Aurelien Geron

„**Z gipsu tortu nie ulepisz**”.

System zawsze uczy się jedynie za pomocą danych zawierających wystarczającą liczbę **istotnych cech** i niezaśmieconych nadmiarem cech nieistotnych.

Elementem krytycznym jest wybór dobrego zbioru cech uczących (feature engineering).”

Składa się on z :

- dobór cech (feature selection)
- odkrywanie cech (feature extraction)
- nowe cechy z nowych danych

Po co selekcja zmiennych ?

Aurelien Geron

„Z gipsu tortu nie ulepisz”.

System zawsze uczy się jedynie za pomocą danych zawierających wystarczającą liczbę **istotnych cech** i niezaśmieconych nadmiarem cech nieistotnych.

Elementem krytycznym jest wybór dobrego zbioru cech uczących (feature engineering)."

Składa się on z :

- dobór cech (feature selection)
- odkrywanie cech (feature extraction)
- nowe cechy z nowych danych

o czym nie będzie ? feature extraction / streaming feature selection

PCA oraz autoencondery czyli liniowe i nieliniowe kombinacje zmiennych.
wybieranie zmiennych w czasie rzeczywistym

Wygenerowane dane przedstawiają informacje zbierane podczas procesu udzielania kredytów w bankach. Modele zbudowane na ich podstawie prognozują zajście zdarzenia default – wejścia w opóźnienia więcej niż 3 raty (inaczej więcej niż 90 dni opóźnień) od punktu obserwacji w ciągu następnych 12 miesięcy.

Zmienna celu jest zależna od predyktorów w sposób silnie nieliniowy.

Szczegółowy opis algorytmu zamieszczono w książce K. Przanowski Credit Scoring w Erze Big Data.

Zbiór	L. obserwacji	L. dobrych	L. złych	L. nieok.	P. dobrych [%]	P. złych [%]	P. nieok. [%]
ABT BEH	52 841	31 010	15 378	6 453	58,7	29,1	12,2

(Pre)Selekcja

- Random

(Pre)Selekcja

- Random
- Gini, Information Value

(Pre)Selekcja

- Random
- Gini, Information Value

Metody rekurencyjne

- Forward, Backward (RFE)

(Pre)Selekcja

- Random
- Gini, Information Value

Metody rekurencyjne

- Forward, Backward (RFE)

Metody modelowe

- Lasso, Ridge

(Pre)Selekcja

- Random
- Gini, Information Value

Metody rekurencyjne

- Forward, Backward (RFE)

Metody modelowe

- Lasso, Ridge
- drzewa decyzyjne, lasy losowe

(Pre)Selekcja

- Random
- Gini, Information Value

Metody rekurencyjne

- Forward, Backward (RFE)

Metody modelowe

- Lasso, Ridge
- drzewa decyzyjne, lasy losowe
- Xgboost, sieci neuronowe, SGDClassifier

Wybrane metody selekcji zmiennych

(Pre)Selekcja

- Random
- Gini, Information Value

Metody rekurencyjne

- Forward, Backward (RFE)

Metody modelowe

- Lasso, Ridge
- drzewa decyzyjne, lasy losowe
- Xgboost, sieci neuronowe, SGDClassifier

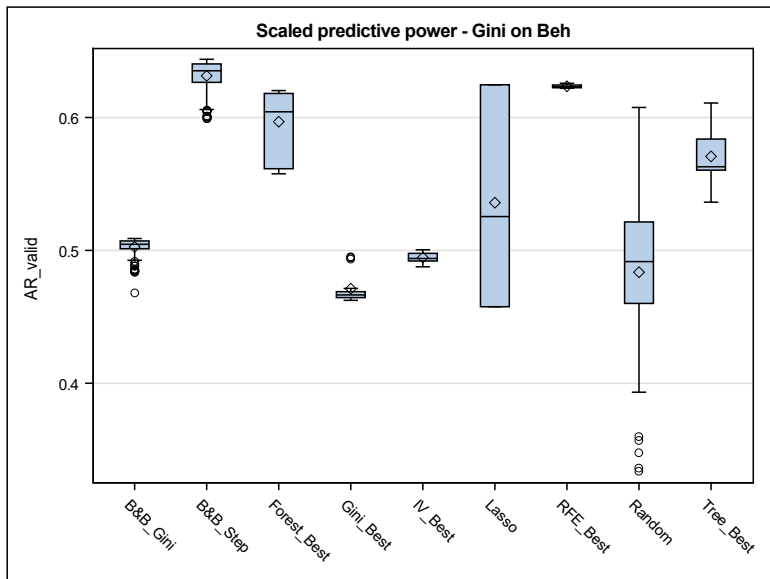
Metody zaawansowane

- Branch and bound w SAS

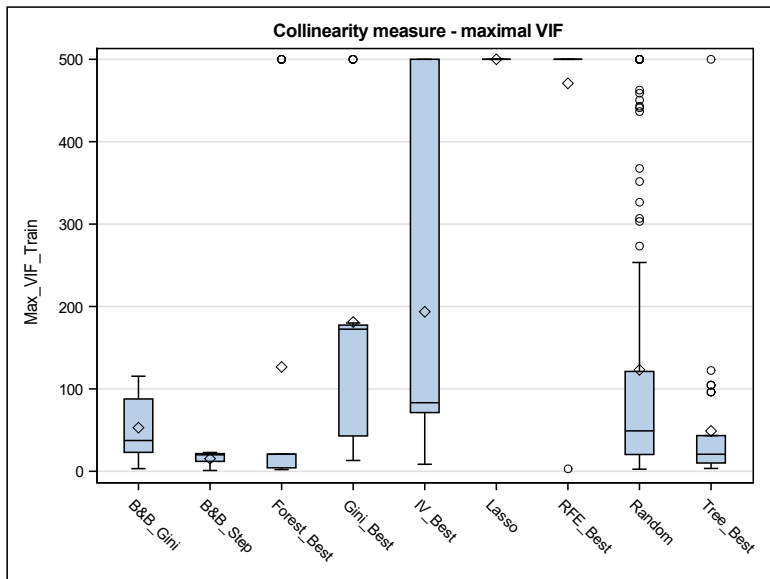
Podejście klasyczne do modelowania scoringowego

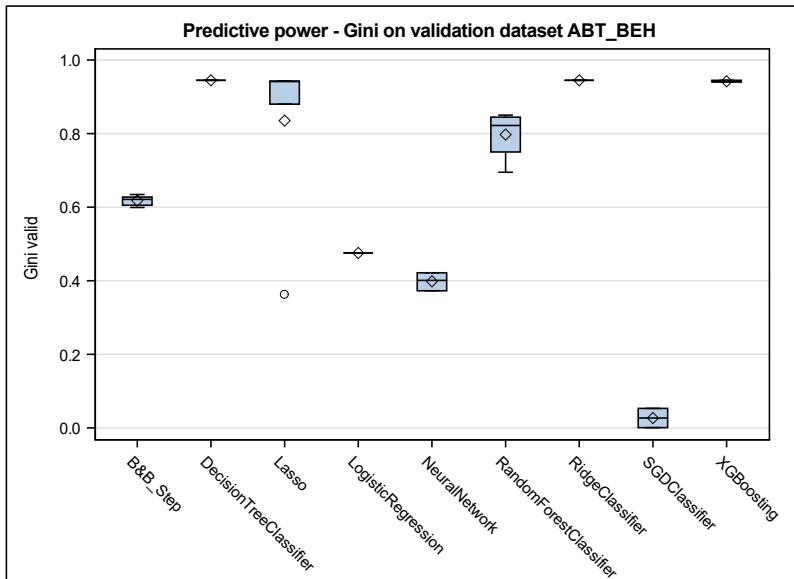
- Preselekcja
usuwanie dużej korelacji, niski Gini, duże deltaGini, małe IV
- Dyskretyzacja zmiennych
- Transformacja zmiennych do WOE
- **Selekcja zmiennych**
- Diagnostyka współliniowości
- Finalny model regresji logistycznej

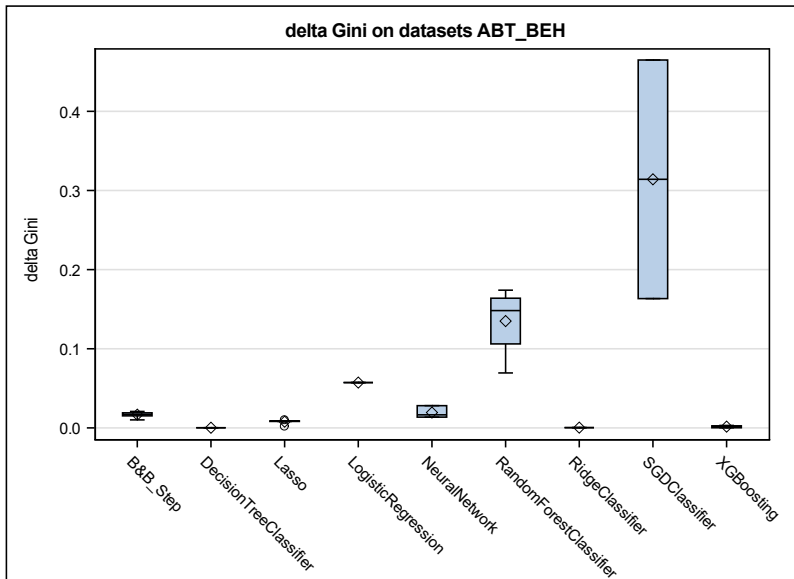
Selekcja zmiennych - porównanie



Selekcja zmiennych - współliniowość







Dziękujemy za uwagę!

kprzan@sgh.waw.pl, sebastian.zajac@sgh.waw.pl