Metody selekcji zmiennych w modelach skoringowych – klasyka kontra AI/ML

SAS dla Administratorów i Praktyków 2020

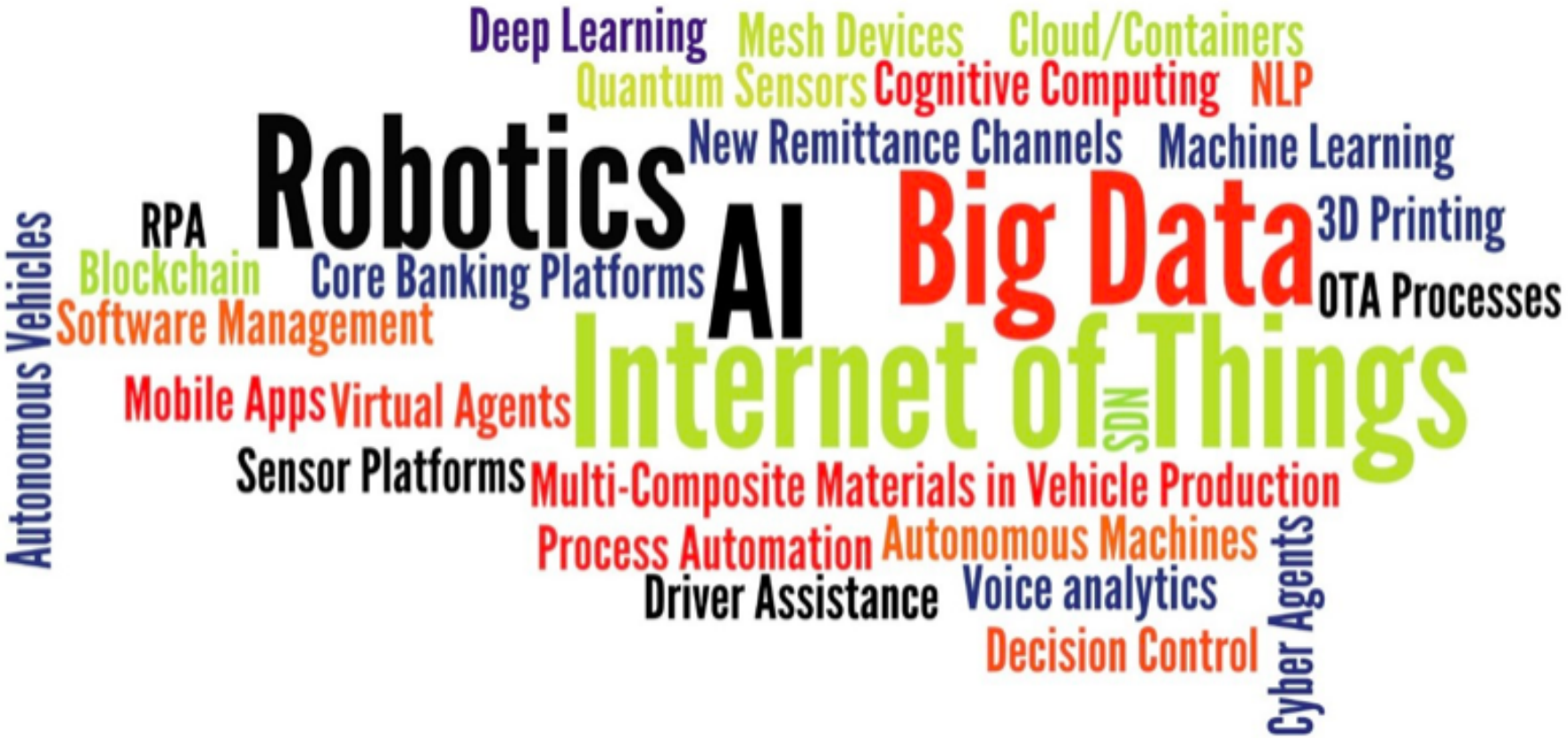
Sebastian Zając, Karol Przanowski

Szkoła Główna Handlowa

27.10.2020

Analiza danych

Nie bój się – to tylko synonimy analityki



Zastosowanie ... danych

PRZYKŁADY BRANŻ, W KTÓRYCH MODELE PREDYKCYJNE MOGĄ BYĆ WYKORZYSTYWANE

FINANSE

- Fundusze inwestycyjne i gwarancyjne
- Ubezpieczenia
- Kredyty / Leasing/ Faktoring
- Windykacja
- Ochrona przed nadużyciami

MARKETING

- Częstotliwości i rodzaj kontaktu z klientem
- Programy lojalnościowe
- Retencja w usługach abonamentowych
- Promocje cenowe
- Sprzedaż internetowa

INNE

- Centra usług wspólnych
- Punkty masowej obsługi klienta
- Domy wysyłkowe
- Logistyka
- Firmy windykacyjne

NAUKA

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

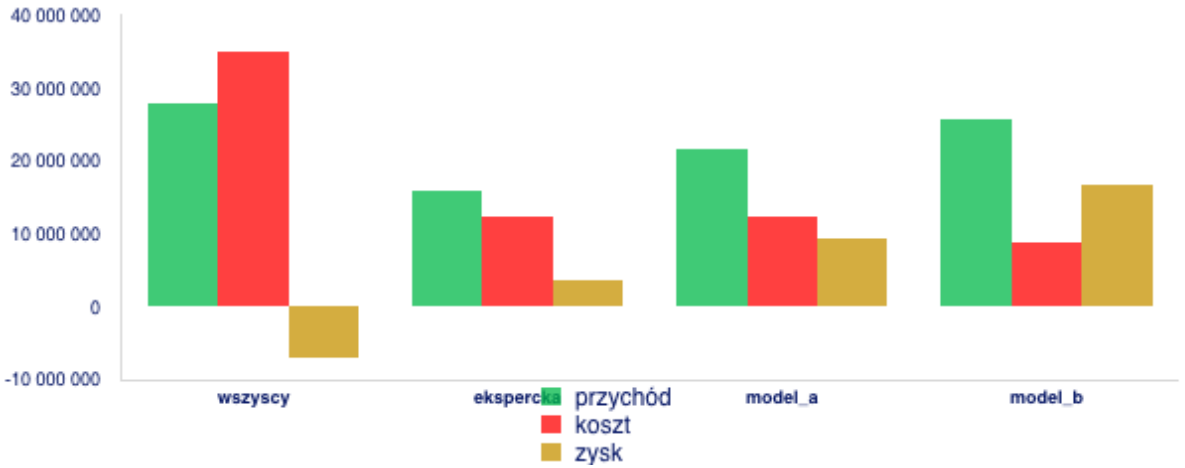
- Przychody: 21 531 163
- Koszty: 12 250 000
- Zysk: 9 281 163

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

Motywacja - 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 enginering).”
Składa się on z :

- dobór cech (feature selection)

Motywacja - 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 enginering).”
Składa się on z :

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

Motywacja - 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 enginering).”
Składa się on z :

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

Motywacja - 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 enginering).”
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

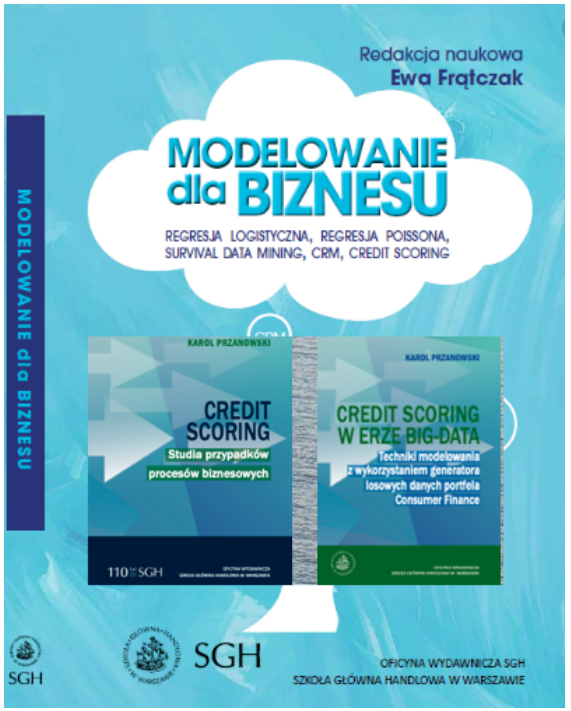
Przygotowanie danych

Python

```
from random import choice
import pandas as pd
import numpy as np
from sklearn.datasets import make_classification
import os

class DataOptions(object):

    n_samp = 50000
    n_feat = 50
    n_infor = [10,11,12,13,14,15]
    n_red = [0,1,2,3,4,5,6,7]
    w_weights = []
    flip_y =[0,0.01,0.02,0.03]
    names = ['zm'+str(x) for x in range(n_feat)]


    def __init__(self):

        self.n_informative = choice(self.n_infor)
        self.n_redundant = choice(self.n_red)
        self.flipy = choice(self.flip_y)
```

20 zestawów danych:
50 zmiennych po 50.000 przypadków.

SAS - generator danych



ABT: 1600 zmiennych, 700.000 przypadków.

Wybrane metody selekcji zmiennych

(Pre)Selekcja

- Wariancja, testy statystyczne dla zmiennych - univariate methods

Wybrane metody selekcji zmiennych

(Pre)Selekcja

- Wariancja, testy statystyczne dla zmiennych - univariate methods
- Gini, Information Value

Wybrane metody selekcji zmiennych

(Pre)Selekcja

- Wariancja, testy statystyczne dla zmiennych - univariate methods
- Gini, Information Value

Metody rekurencyjne

- Forward, Backward (RFE), Stepwise selection (SAS)

Wybrane metody selekcji zmiennych

(Pre)Selekcja

- Wariancja, testy statystyczne dla zmiennych - univariate methods
- Gini, Information Value

Metody rekurencyjne

- Forward, Backward (RFE), Stepwise selection (SAS)

Metody modelowe

- Regularyzacja lasso

Wybrane metody selekcji zmiennych

(Pre)Selekcja

- Wariancja, testy statystyczne dla zmiennych - univariate methods
- Gini, Information Value

Metody rekurencyjne

- Forward, Backward (RFE), Stepwise selection (SAS)

Metody modelowe

- Regularyzacja lasso
- drzewa decyzyjne, lasy losowe

Wybrane metody selekcji zmiennych

(Pre)Selekcja

- Wariancja, testy statystyczne dla zmiennych - univariate methods
- Gini, Information Value

Metody rekurencyjne

- Forward, Backward (RFE), Stepwise selection (SAS)

Metody modelowe

- Regularyzacja lasso
- drzewa decyzyjne, lasy losowe
- Xgboost, sieci neuronowe, SGDClassifier

Wybrane metody selekcji zmiennych

(Pre)Selekcja

- Wariancja, testy statystyczne dla zmiennych - univariate methods
- Gini, Information Value

Metody rekurencyjne

- Forward, Backward (RFE), Stepwise selection (SAS)

Metody modelowe

- Regularyzacja lasso
- drzewa decyzyjne, lasy losowe
- Xgboost, sieci neuronowe, SGDClassifier

Metody zaawansowane

- Branch and bound w SAS

Podejście klasyczne vs AI/ML

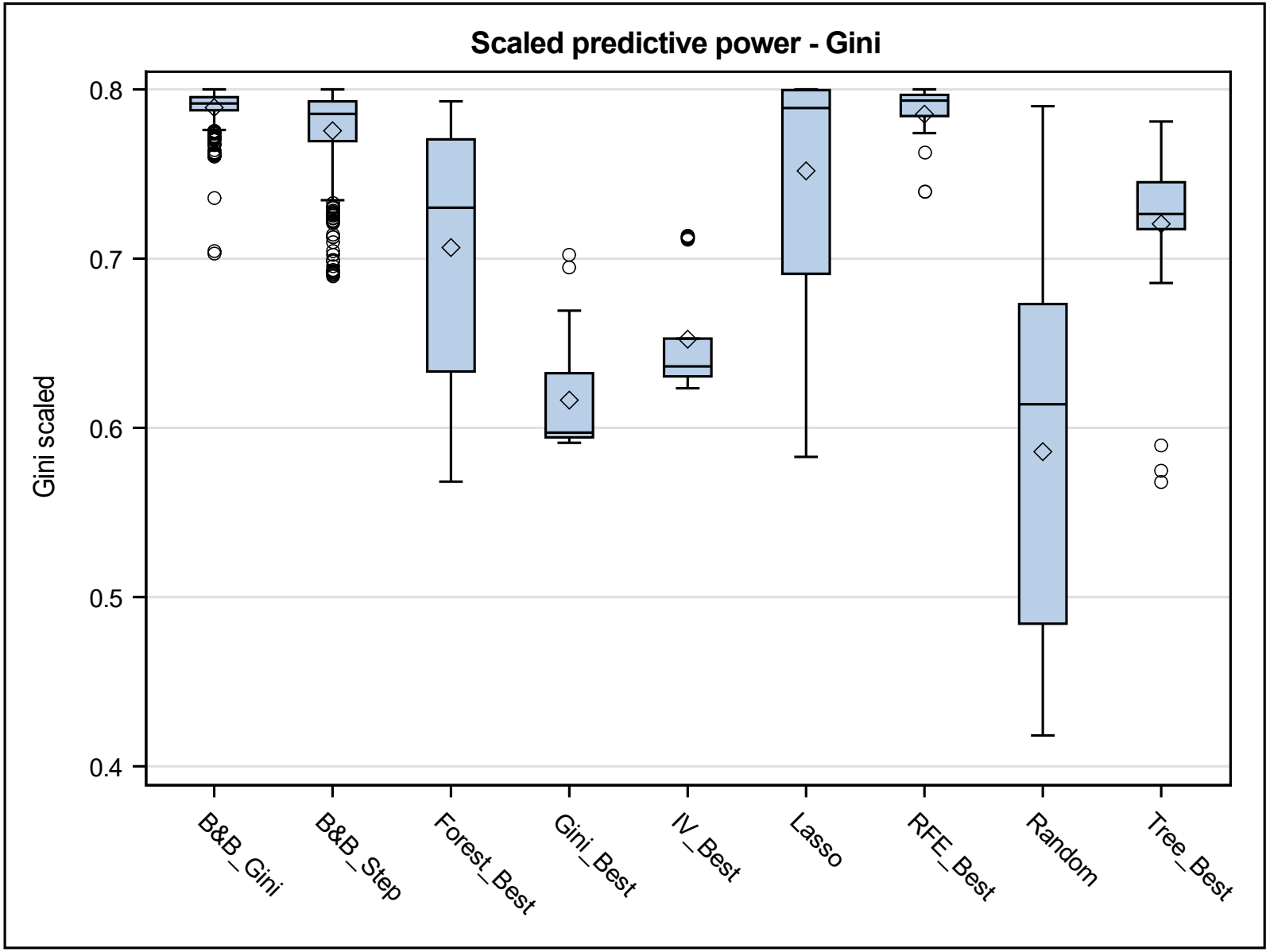
Przygotowanie danych - podejście klasyczne

- Preselekcja (usuwanie dużej korelacji)
- Dyskretyzacja zmiennych
- Transformacja zmiennych do WOE
- Finalny model regresji logistycznej

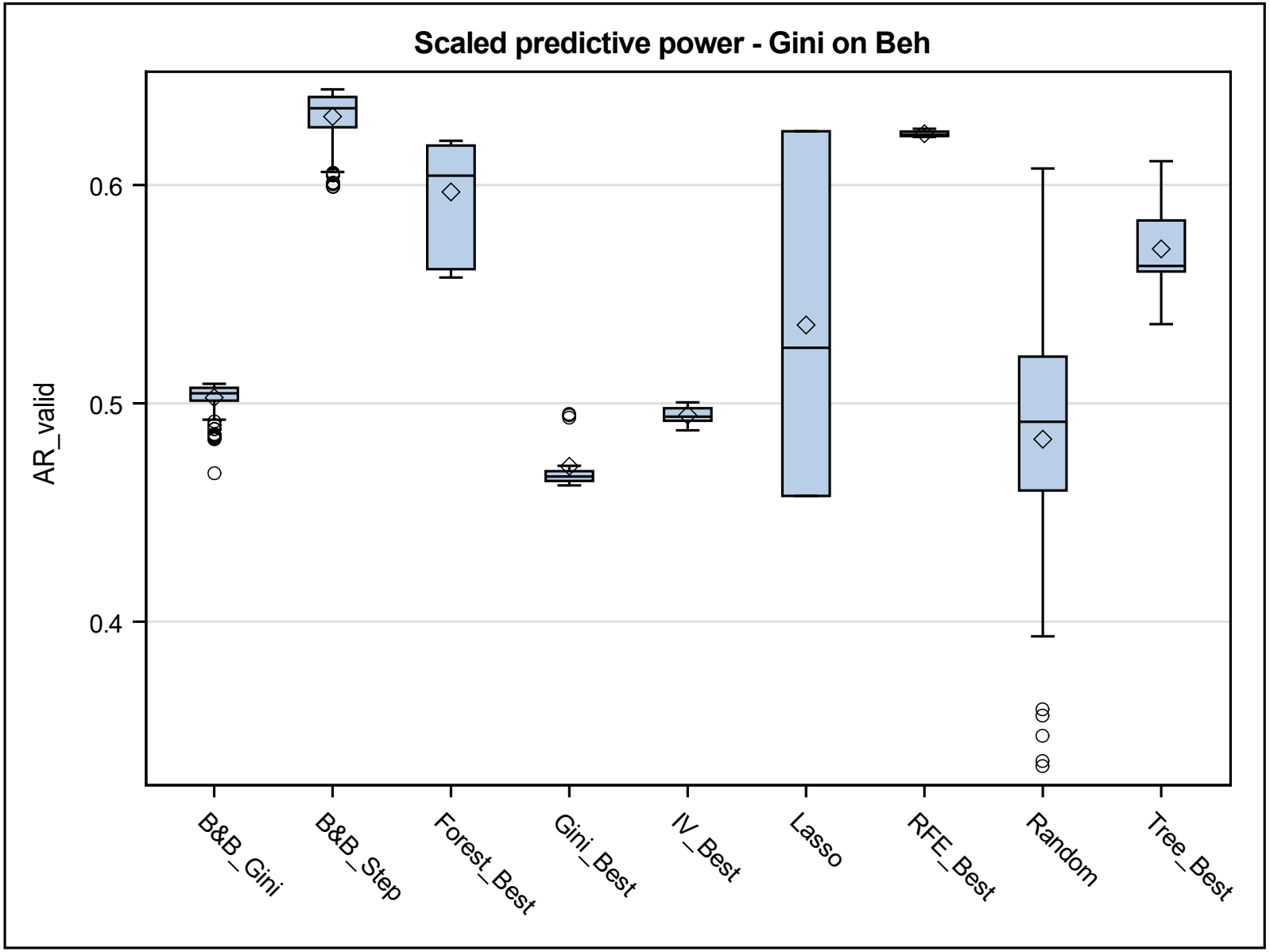
Podejście AI/ML

- Zmienne indykatorowe dla braków danych
- Uzupełnienie braków danych przez średnią
- Finalne modele AI/ML

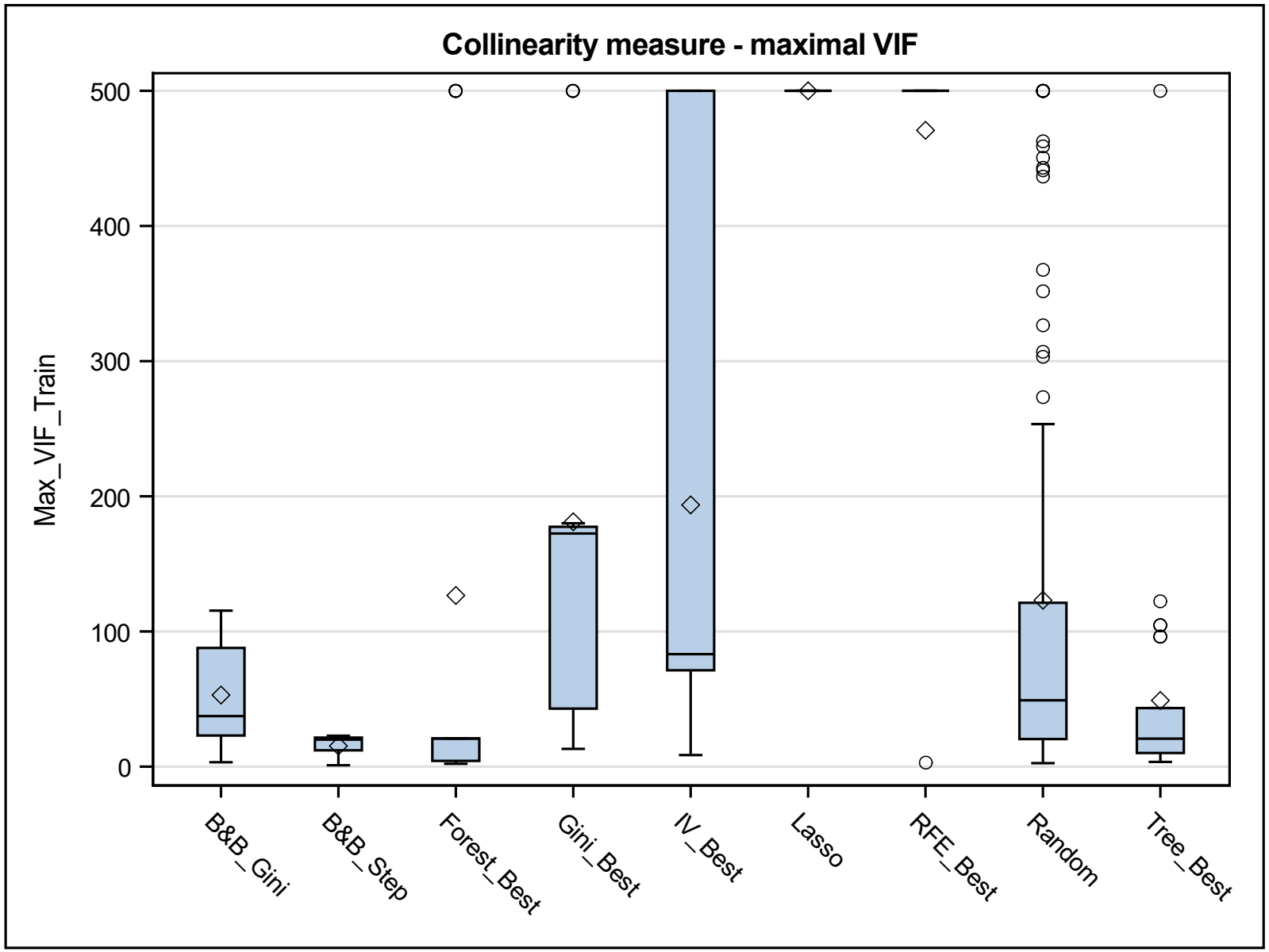
Metody klasyczne



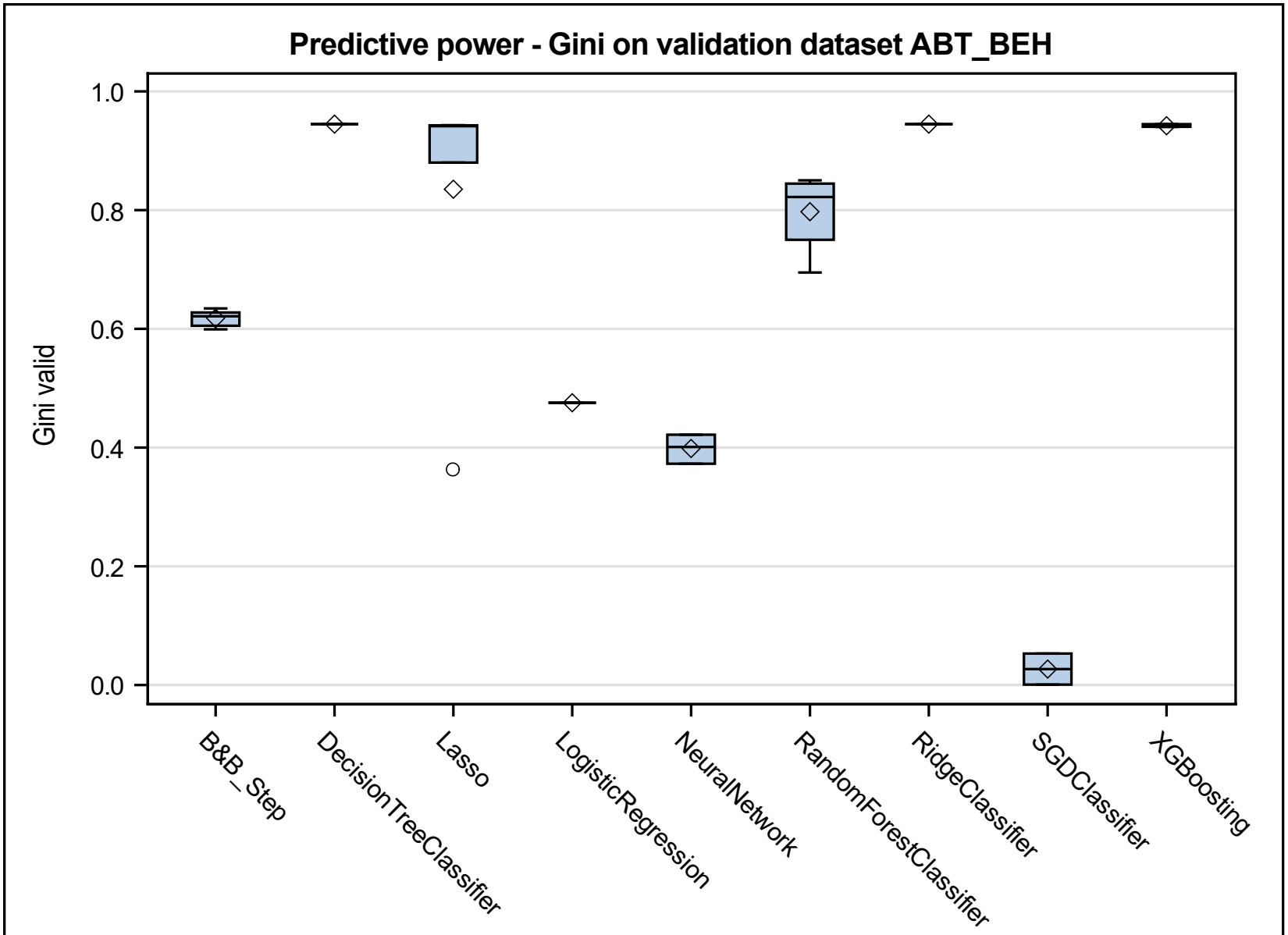
Metody klasyczne



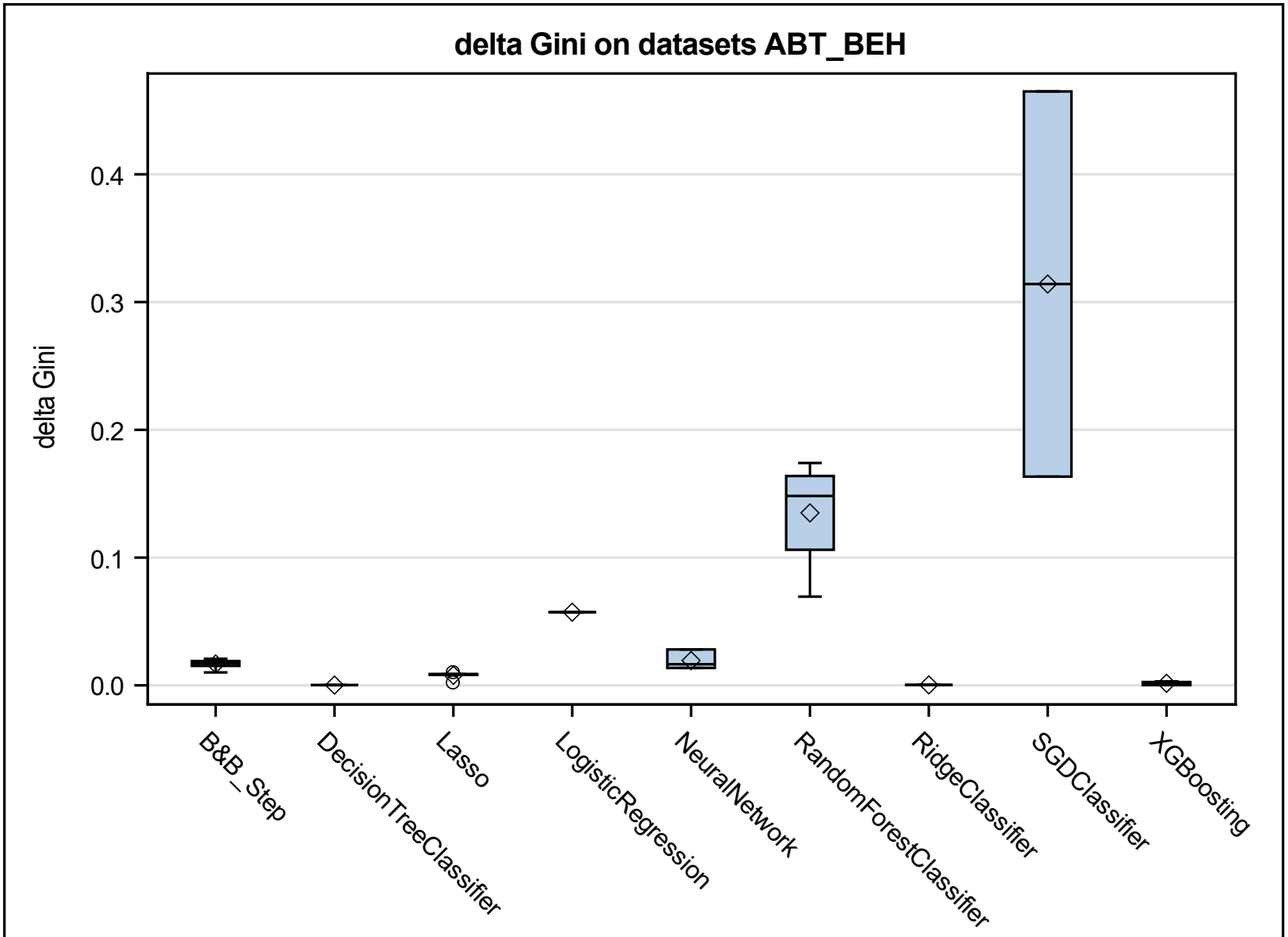
Metody klasyczne



Metody AI/ML


Predictive power - Gini on validation dataset ABT_BEH
Gini valid
1.0
0.8
0.6
0.4
0.2
0.0
B&B_Step
DecisionTreeClassifier
Lasso
LogisticRegression
NeuralNetwork
RandomForestClassifier
RidgeClassifier
SGDClassifier
XGBoosting

Metody AI/ML


delta Gini on datasets ABT_BEH
delta Gini
0.4
0.3
0.2
0.1
0.0
B&B_Step
DecisionTreeClassifier
Lasso
LogisticRegression
NeuralNetwork
RandomForestClassifier
RidgeClassifier
SGDClassifier
XGBoosting

Podsumowanie

Dziękujemy za uwagę!

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