

# Metody selekcji zmiennych w modelach skoringowych

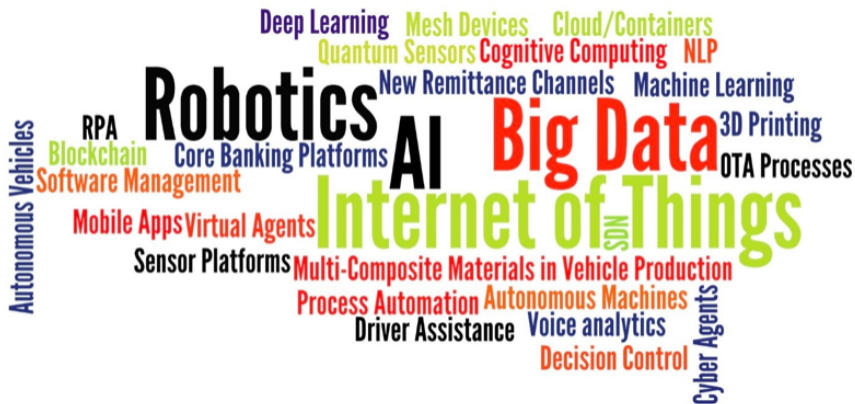
## Modelowanie dla Biznesu 2019

dr Sebastian Zając, dr Karol Przanowski

Instytut Statystyki i Demografii  
Zakład Analizy Historii Zdarzeń i Analiz Wielowymiarowych

28.11.2019

Nie bój się – to tylko synonimy analityki



## 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

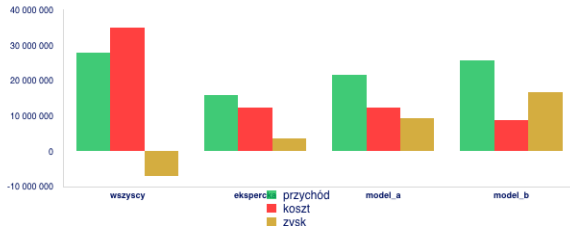
- Przychody: 21 531 163
- Koszty: 12 250 000
- Zysk: 9 281 163

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

# Prosty przykład

Number of cases	7 000 000
Average income on responded case	800
Average cost of contact, offer, campaign	5

Gini global	78,36%
-------------	--------

Global response rate	0,5%
Accepted response rate	1,83%
Acceptance rate	25,00%
Cummulative lift on accepted	3,66
Captured percent (Gains)	91,43%
Global cost	35 000 000
Global income	28 000 000
Global profit	-7 000 000

Accepted cost	8 750 000
Accepted income	25 599 340
Accepted profit	16 849 340
Number of offers	1 750 000
Number of expected responders	31 999

Number of campaigns	29
Number of months	12
Number of customers	20 115
Number of cases	7 000 000

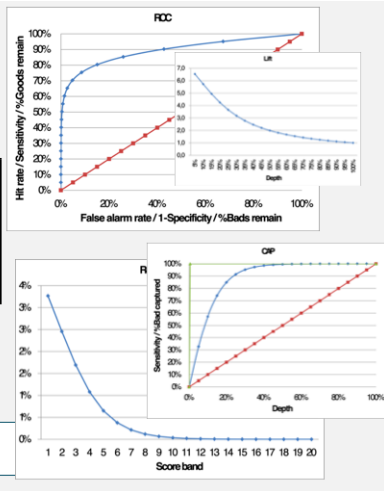
Delta Gini	Delta Profit
1%	272 569
5%	1 362 844
10%	2 725 687

Przykład studium przypadku w Excelu:



[http://administracja.sgh.waw.pl/pl/OW/publikacje/Strony/](http://administracja.sgh.waw.pl/pl/OW/publikacje/Strony/2015.aspx)

[2015.aspx](http://administracja.sgh.waw.pl/pl/OW/publikacje/Strony/2015.aspx)



## Aurelien Geron - Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow

„**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)

## Aurelien Geron - Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow

„**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)



## Aurelien Geron - Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow

„**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

## Aurelien Geron - Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow

„**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

# 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):
```

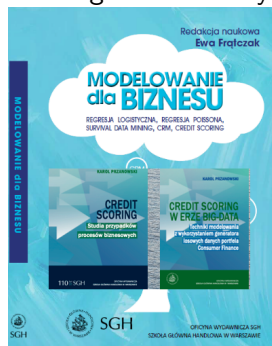
```
    n_samp = 50000
    n_feat = 50
    n_infor = [10,11,12,13,14,15]
    n_red = [0,1,2,3,4,5,6,7]
    w_weights = []
    flip_y = [0,0.01,0.02,0.03]
    names = ['zm'+str(x) for x in range(n_feat)]
```

```
    def __init__(self):
```

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

20 zestawów danych:  
50 zmiennych po 50.000 przypadków.

## SAS - generator danych



ABT: 212 zmiennych, 68500 przypadków.

## (Pre)Selekcja

- Wariancja, testy statystyczne dla zmiennych - univariate methods

## (Pre)Selekcja

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

## (Pre)Selekcja

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

## Metody rekurencyjne

- Forward, Backward (RFE), Stepwise selection

## (Pre)Selekcja

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

## Metody rekurencyjne

- Forward, Backward (RFE), Stepwise selection

## Metody modelowe

- Regularyzacja lasso

## (Pre)Selekcja

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

## Metody rekurencyjne

- Forward, Backward (RFE), Stepwise selection

## 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

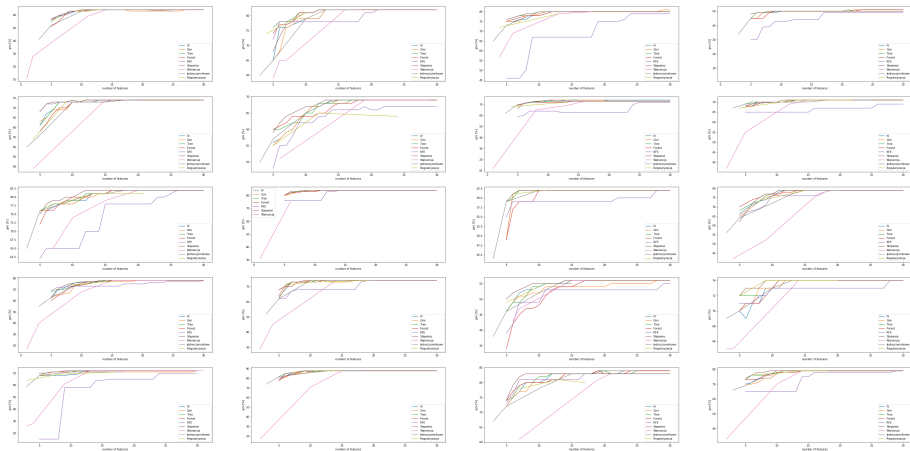
## Metody modelowe

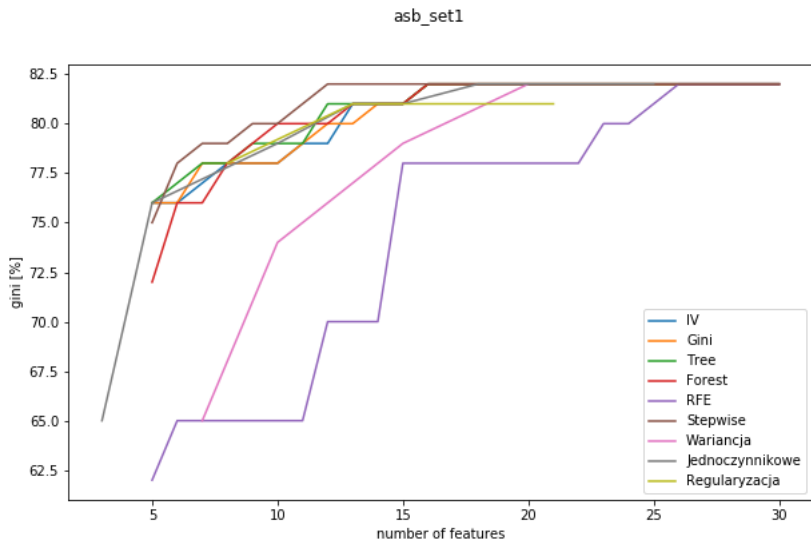
- Regularyzacja lasso
- drzewa decyzyjne, lasy losowe

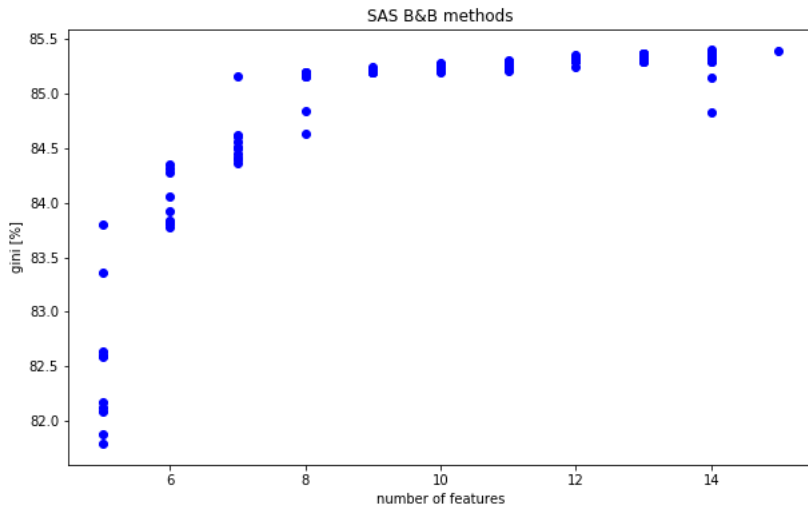
## Metody zaawansowane

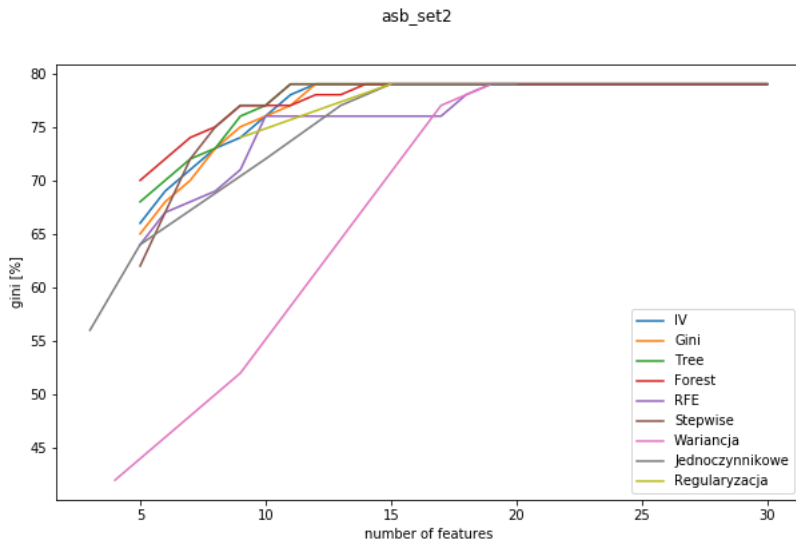
- Branch and bound w SAS

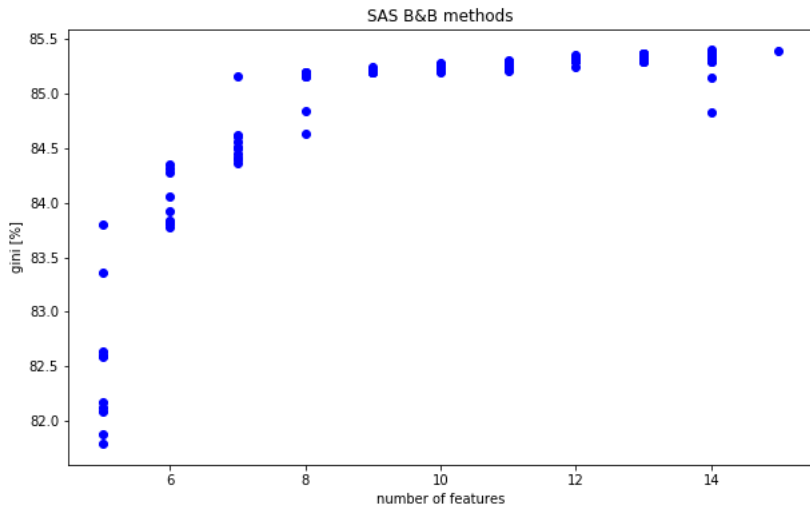
# Wyniki Python

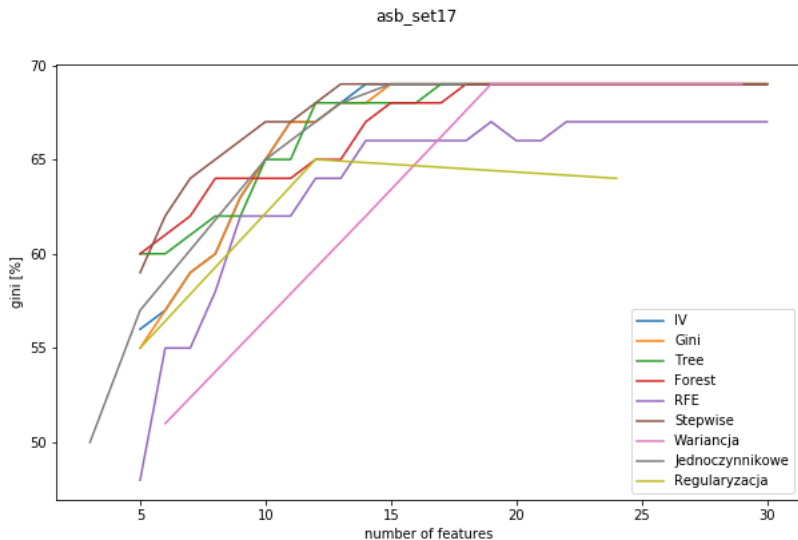


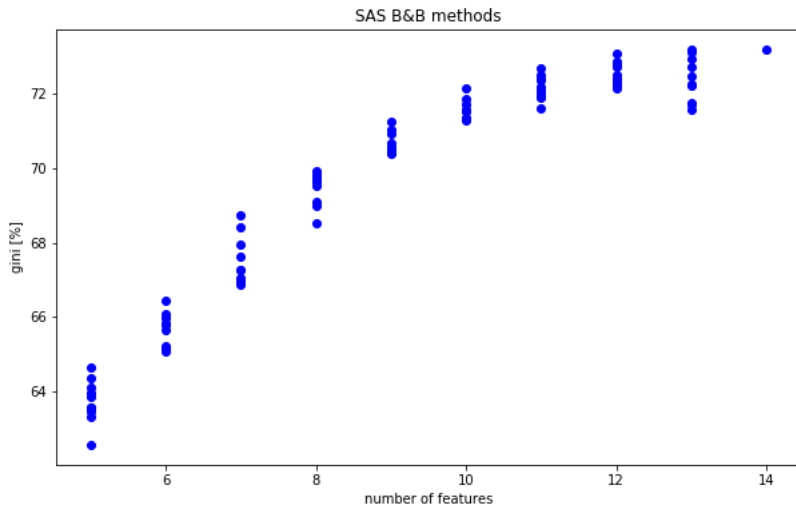




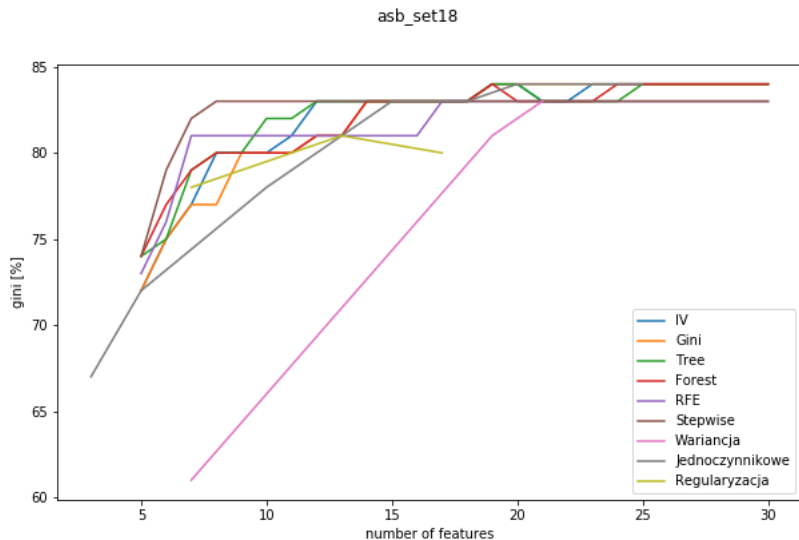


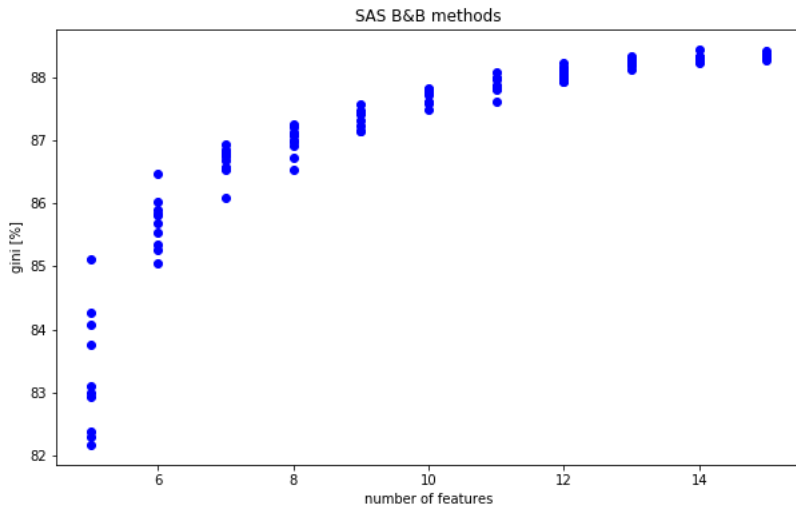













# Plany na przyszłość

## Co chcielibyśmy jeszcze przetestować ?

- Algorytmy genetyczne,
- Persistent topology,
- mieszanie metod ML z uczeniem głębokim.


## Quantum Computing



Facebook group: <https://www.facebook.com/groups/qpoland>

Facebook page: <https://www.facebook.com/QPoland-110308580421373>

Twitter: [QPolandCousin](#)



Dziękujemy za uwagę !  
kprzan@sgh.waw.pl, szajac2@sgh.waw.pl

Dziękujemy za udział w konferencji  
„Modelowanie dla Biznesu”  
Zapraszamy na rozdanie certyfikatów SAS