



# W dwadzieścia memów dookoła modelowania

czyli teoretyczne wprowadzenie do warsztatów  
modelowanie w Pythonie

opracowała:

Patrycja NAUMCZYK

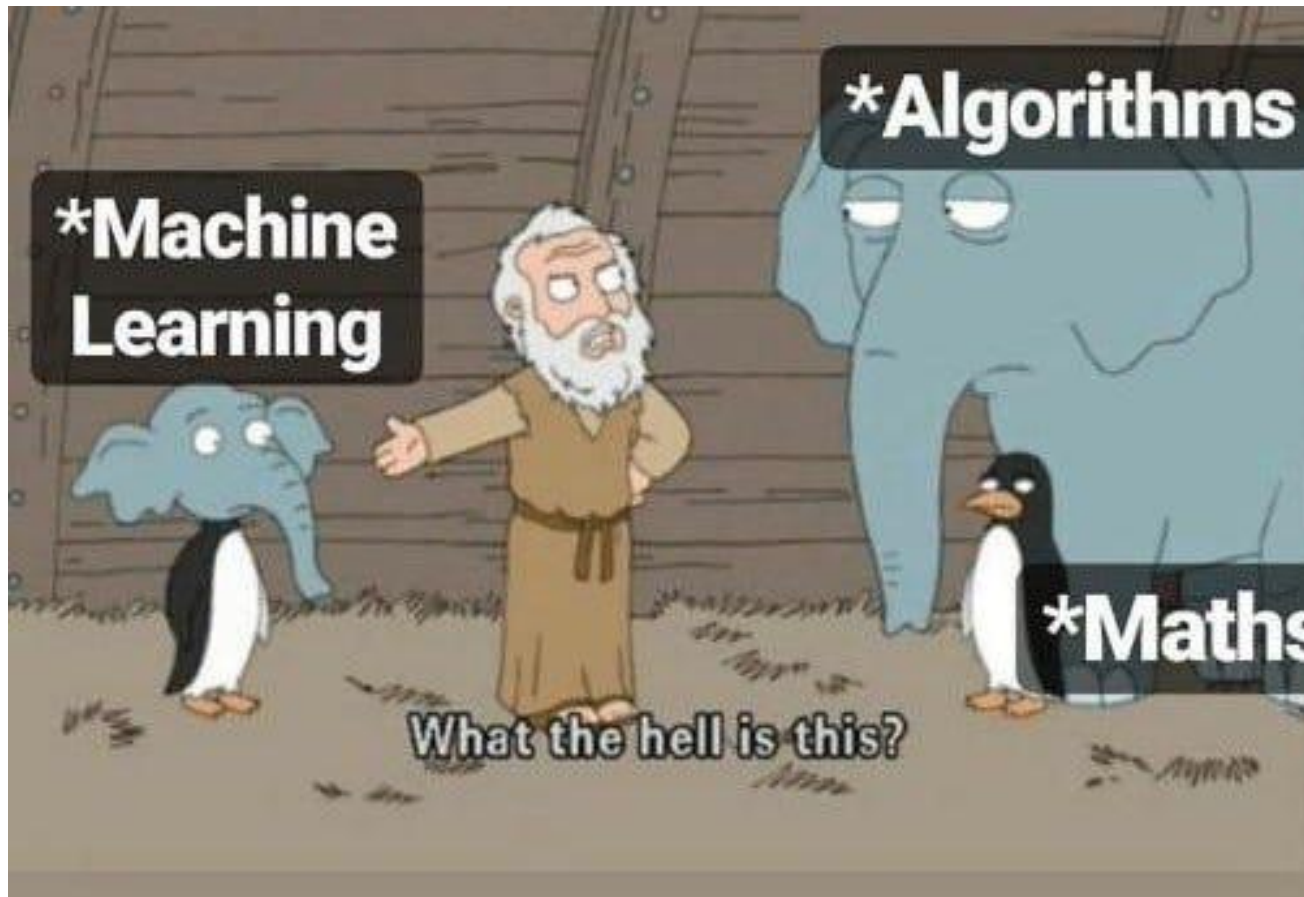
# O czym będzie?

1. Uczenie maszynowe – czym jest i czym nie jest?
2. Modelowanie
  - a) Wybór modelu
    - i. Ogólne modele liniowe
    - ii. Zmienna wyjaśniana
    - iii. Podział zmiennych wyjaśniających
  - b) Analiza eksploracyjna
    - i. Typy danych
    - ii. Braki danych
    - iii. Duplikaty
    - iv. Statystyki opisowe
    - v. Rozkłady zmiennych
  - c) Czyszczenie danych:
    - i. Braki danych
    - ii. Wartości odstające
    - iii. Usuwanie vs zamiana
  - d) Dobór zmiennych (feature engineering)
    - i. Przekształcenia danych
    - ii. Dobór zmiennych
  - e) Uczenie i optymalizacja parametrów modelu
    - i. Sety testowy i treningowy
    - ii. Przekształcenia danych
    - iii. Regularyzacja
    - iv. Walidacja krzyżowa
  - f) Ewaluacja modelu
    - i. Model „zero”
    - ii. Metryki
    - iii. Shap



# **UCZENIE MASZYNOWE – Z CZYM TO SIĘ JE?**

# Uczenie maszynowe – czym jest i czym nie jest?

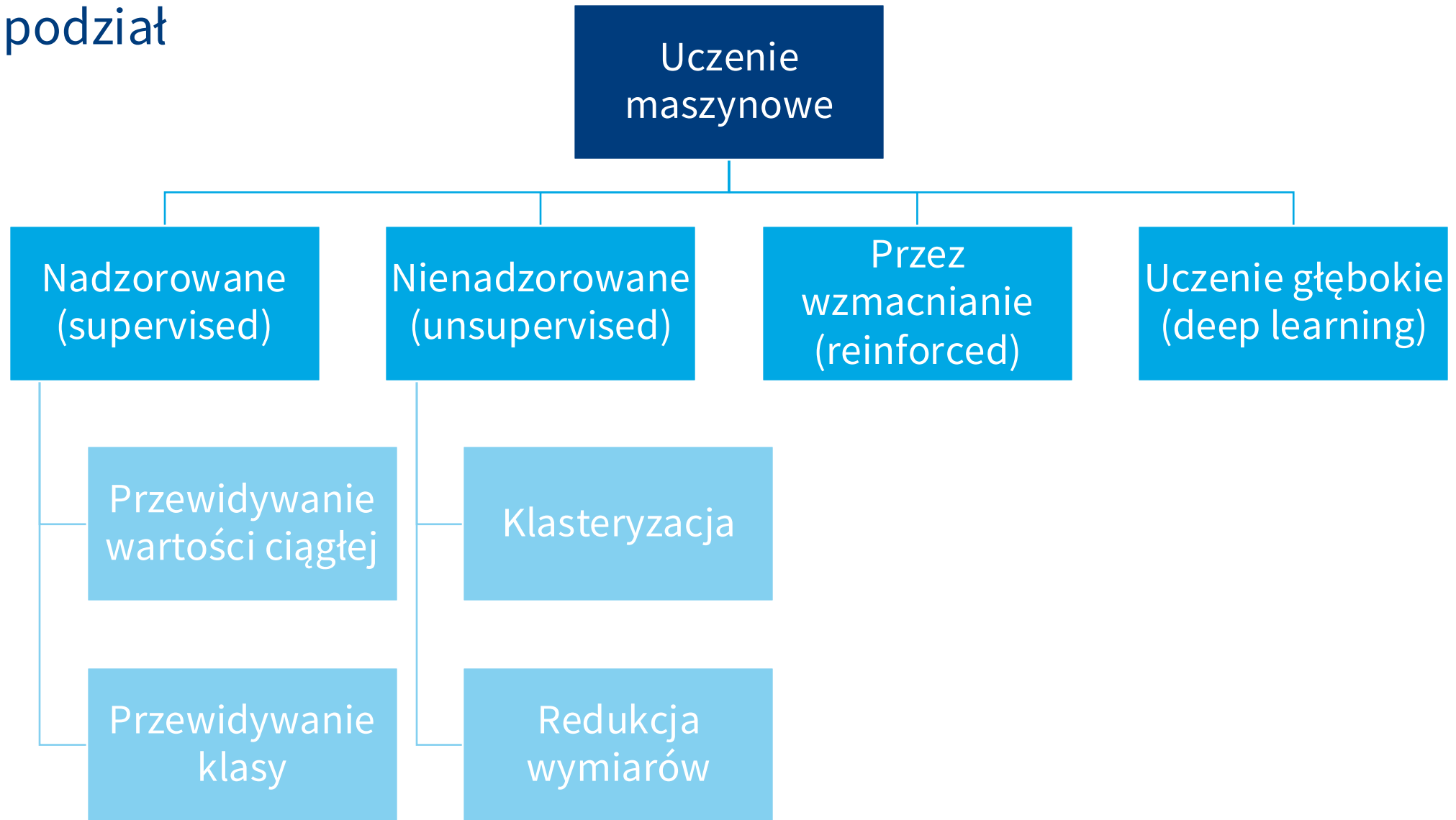


Dziedzina sztucznej inteligencji

DANE → PROGNOZA

DANE → DECYZJA

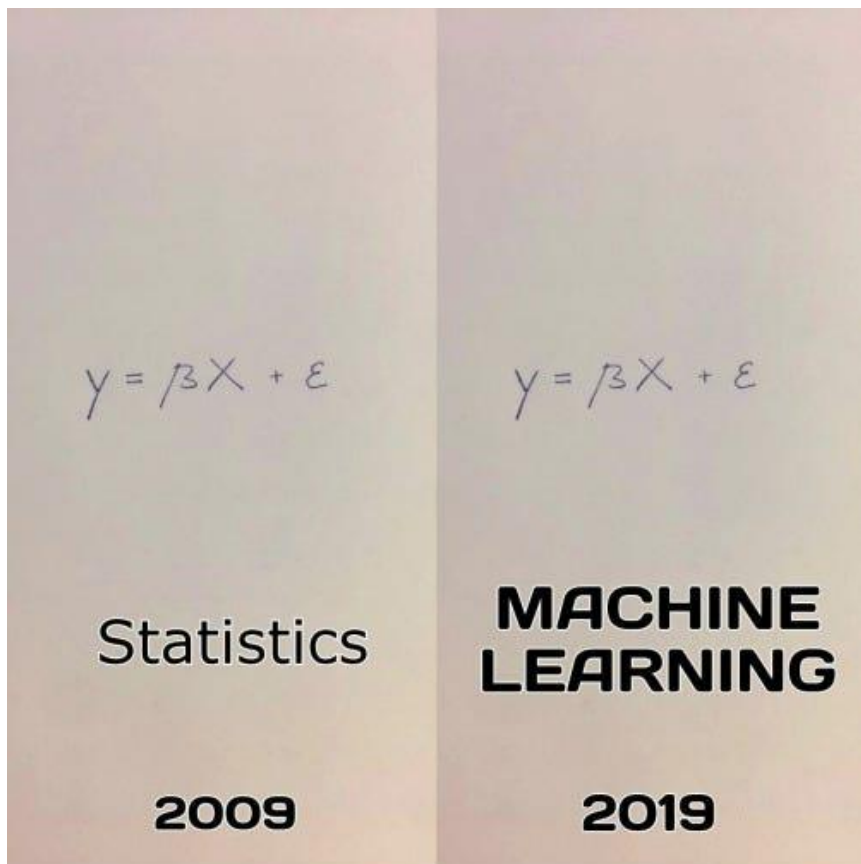
# ML – podział



## Regresja liniowa – wszyscy znają, prawda?



# Regresja liniowa – podsumowanie



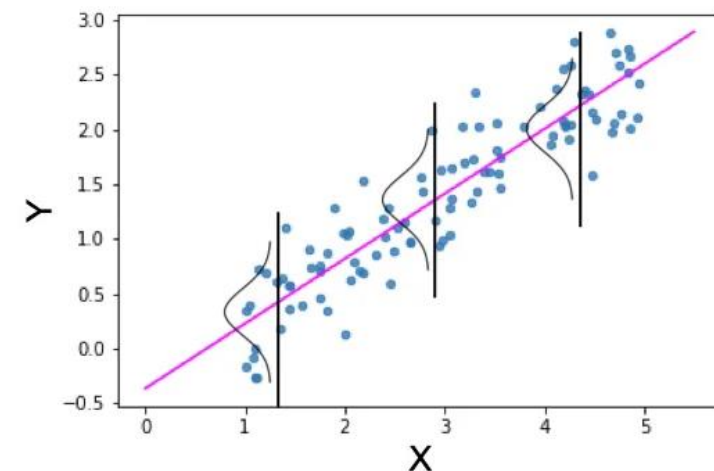
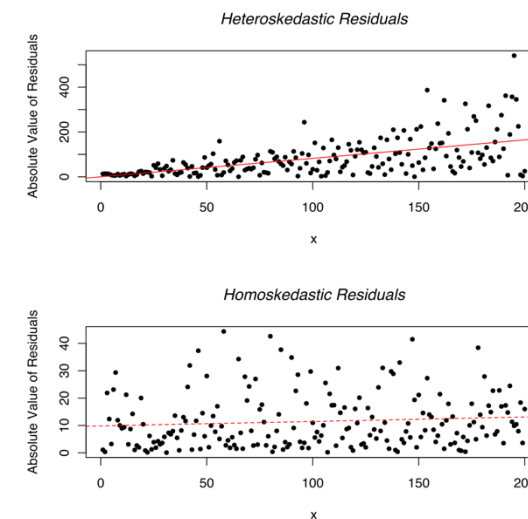
#10yearchallenge

Założenia:

- Liniowa zależność między zmienną wyjaśnianą a wyjaśniającymi
- Homoscedastyczność (homogeniczność wariancji reszt)
- Normalność rozkładu reszt
- Wzajemna niezależność zmiennych wyjaśniających

Czasem dodatkowo:

- Brak autokorelacji reszt



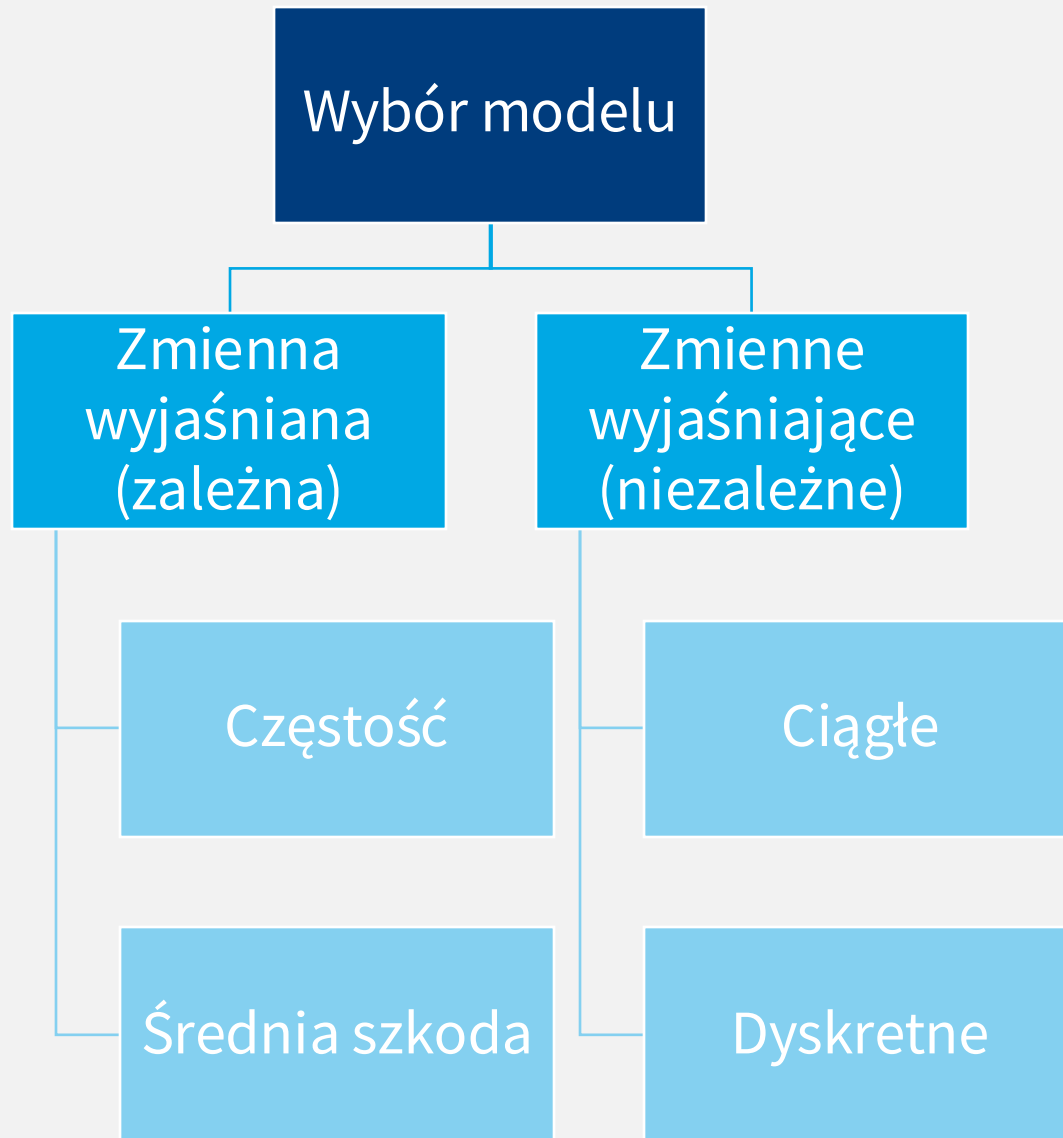
Linear regression illustrated





# **WYBÓR MODELU**





„Rób , jak uważasz,  
tylko uważaj, co robisz!”

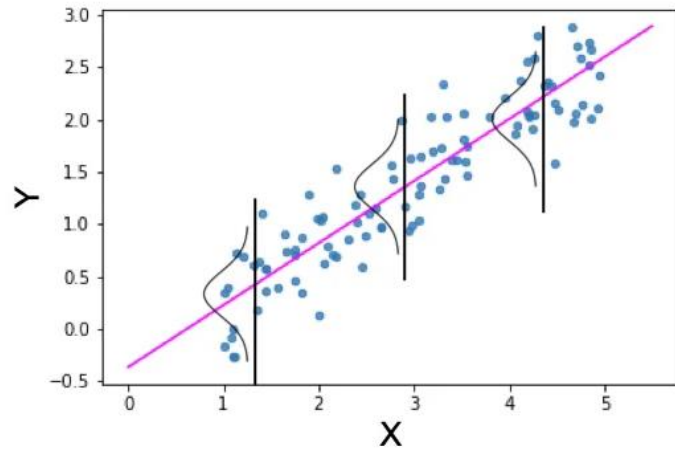
*mądrość ludowa*

# Generalized Linear Models

## REGRESJA LINIOWA

$$\mu_i = b_0 + b_1 x_i$$

$$y_i \sim \mathcal{N}(\mu_i, \varepsilon)$$



Linear regression illustrated

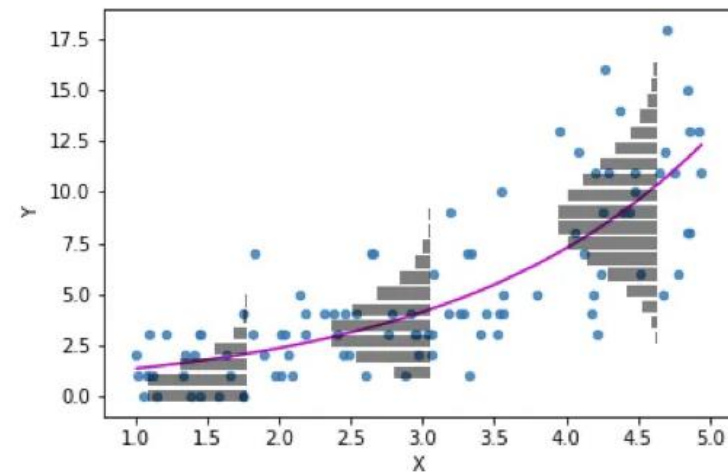
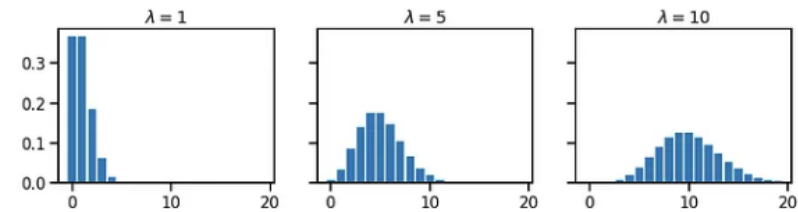
## UOGÓLNIENIE – REGRESJA POISSONA

Link function Linear predictor

$$\ln \lambda_i = b_0 + b_1 x_i$$

$$y_i \sim \text{Poisson}(\lambda_i)$$

Probability distribution



2a. Modelowanie – wybór modelu

# Założenia modeli ryzyka

Regresja Poissona – modelowanie częstości:

1. Zmienna wyjaśniana jest policzalna (nieujemna liczba całkowita)
2. Zmienna wyjaśniana ma rozkład Poissona
3. Wzajemna niezależność indywidualnych obserwacji
4. Liniowość (w relacji do funkcji łączącej)
5. Wariancja zmiennej wyjaśnianej i średnia z predykcji modelu wynikowego są do siebie zbliżone (wówczas brak overdispersion i underdispersion)

Regresja Gamma – modelowanie średniej szkody:

1. Zmienna wyjaśniana jest dodatnią liczbą rzeczywistą
2. Zmienna wyjaśniana ma rozkład Gamma
3. Wzajemna niezależność indywidualnych obserwacji
4. Liniowość (w relacji do funkcji łączącej)
5. Wariancja proporcjonalna do kwadratu średniej modelu wynikowego (wówczas brak overdispersion i underdispersion)

Przyczyny over- i underdispersion

- a) Brak ważnej zmiennej wyjaśniającej
- b) Wartości odstające
- c) Brak ważnej interakcji
- d) Brak przekształcenia wartości surowych
- e) Za mała moc statystyczna (za mało danych, dane rzadkie)
- f) Nielosowe braki w danych
- g) Zły model (!)

## A jeśli nie Poisson, to co?

1. Generalized Poisson Regression models – korekta dyspersji
  - a) Consul's Generalized Poisson Regression (GP-1)
  - b) Famoye's Restricted Generalized Poisson Regression (GP-2)

$$P_y(y = k) = \frac{e^{-\lambda} * \lambda^k}{k!}$$

$$P_y(y = k) = \frac{e^{-(\lambda + \alpha * k)} * (\lambda + \alpha * k)^{k-1}}{k!}$$
$$Mean(y) = \frac{\lambda}{(1 - \alpha)}$$
$$Variance(y) = \frac{\lambda}{(1 - \alpha)^3}$$

$$P_y(y = k) = \left(\frac{\lambda}{(1 + \alpha * \lambda)}\right) \frac{(\lambda + \alpha * k)^{k-1}}{k!} e^{\left(\frac{-\lambda(1 + \alpha * k)}{1 + \alpha * \lambda}\right)}$$
$$Mean(y) = \lambda$$
$$Variance(y) = \lambda * (1 + \alpha * \lambda)^2$$

2. Negative Binomial model – brak założenia *średnia = wariancja*

$$Variance = mean + \alpha * mean^p$$

$$Variance = mean + \alpha * mean$$
$$= (1 + \alpha) * mean$$

$$Variance = mean + \alpha * mean^2$$

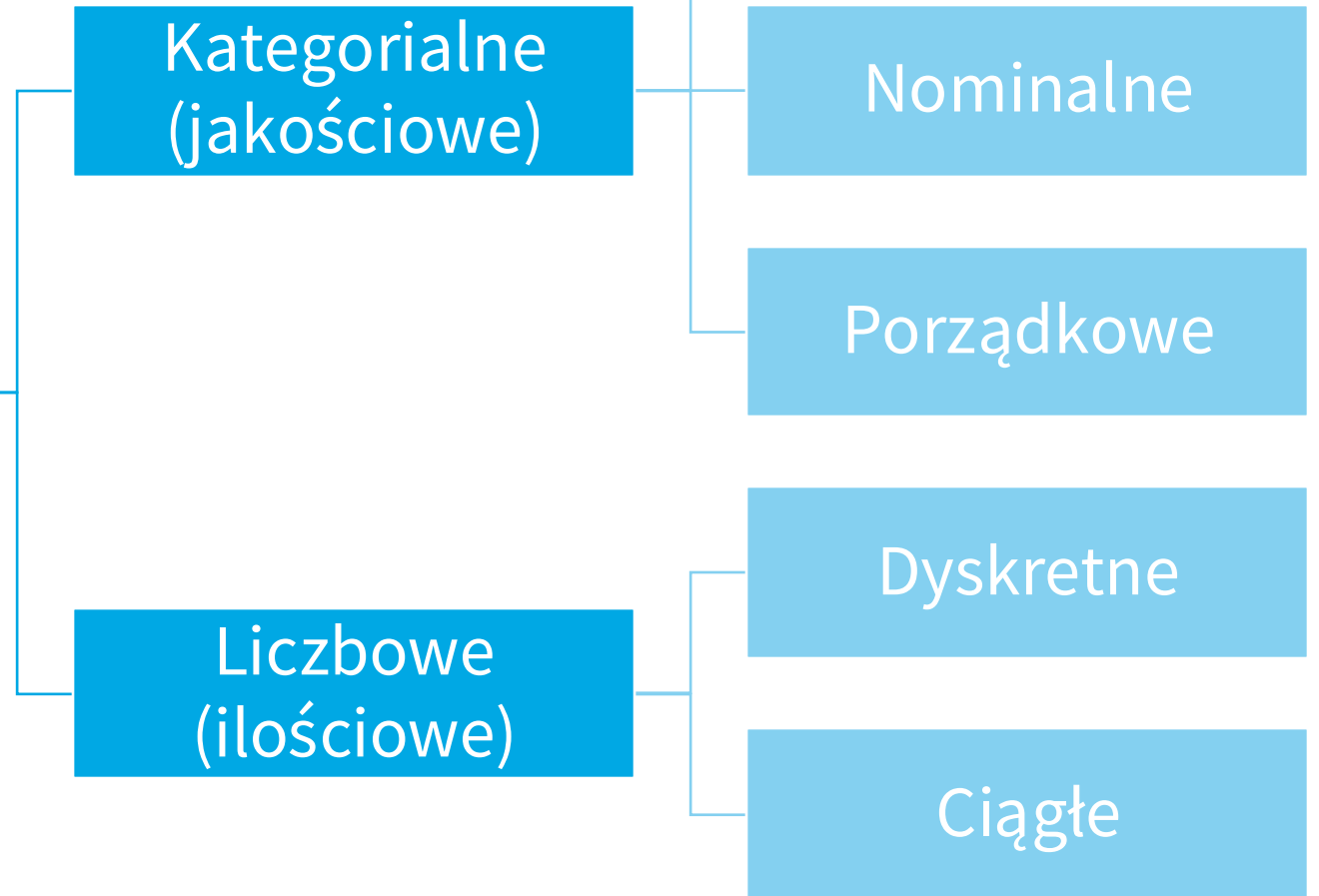
3. Zero-Inflated Poisson model – nadmiar wartości zerowych w danych
4. Tweedy model (częstość i średnia szkoda w jednym)
5. Drzewa decyzyjne

# Zmienne wyjaśniające

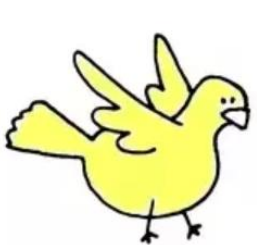
Skale pomiarowe:

1. Skala nominalna
2. Skala porządkowa
3. Skala przedziałowa
4. Skala ilorazowa

Zmienne



CATEGORICAL DATA:



I am a bird.  
I am yellow.  
I am awesome.



I am a seahorse.  
I am orange.  
I am super awesome.



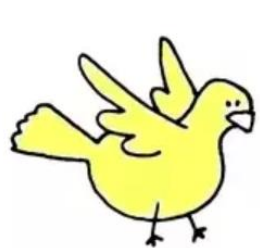
I am a T-rex.  
I am green.  
I am extinct.

# Zmienne wyjaśniające

Skale pomiarowe:

1. Skala nominalna
2. Skala porządkowa
3. Skala przedziałowa
4. Skala ilorazowa

Zmi



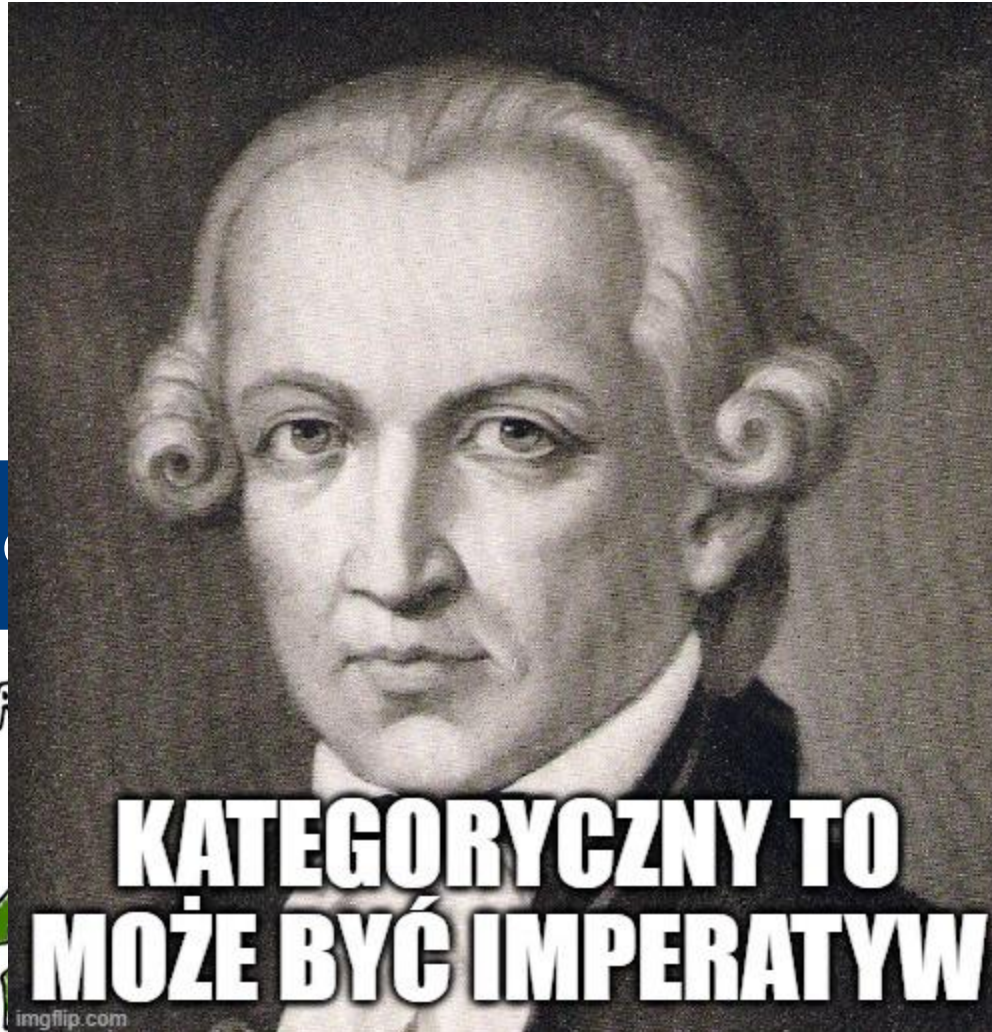
I am a bird.  
I am yellow.  
I am awesome.



I am a seahorse.  
I am orange.  
I am super awesome.



I am a T-rex.  
I am green.  
I am extinct.



Dychotomiczne

Nominalne

Porządkowe

Dyskretne

Ciągłe



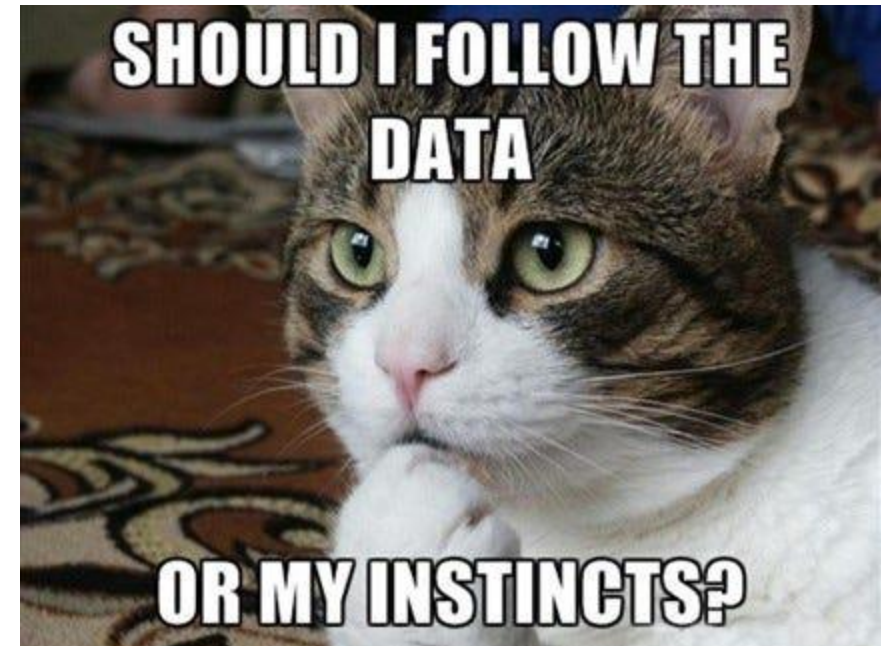


# **ANALIZA EKSPLORACYJNA**




# Analiza eksploracyjna (EDA)

1. Typy danych
  - a) Czy odzwierciedlają skalę pomiarową zmiennej?
2. Braki danych
  - a) Czy związane ze zmienną, czy z konkretnymi obserwacjami?
3. Duplikaty
  - a) Usuwamy, ale...
  - b) ... skąd one się wzięły?
4. Statystyki opisowe (adekwatne do skali pomiarowej zmiennej)
5. Rozkłady zmiennych




# Analiza eksploracyjna (EDA)

- 1.
- 2.
- 3.
- 4.
- 5.

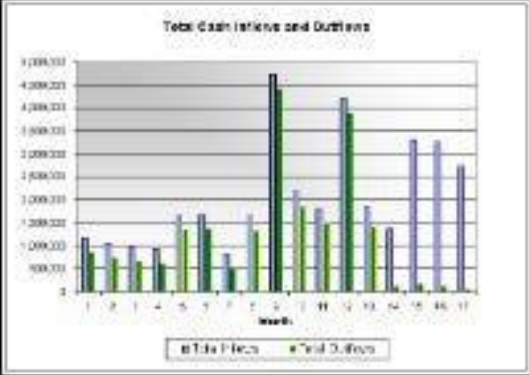


**WHEN DATA IS IN TABLE FORM**

ID	NAME	CLASS	MARK	SEX
1	John Doe	Four	75	female
2	Max Rubin	Three	85	male
3	Arnold	Three	95	male
4	Kristi Star	Four	80	female
5	John Mlee	Four	60	female
6	Alex Johns	Four	55	male
7	My John Rob	Fifth	78	male
8	Arnold	Five	85	male
9	Tee Cry	Six	75	male
10	Big John	Four	55	female



**WHEN DATA IS IN PLOT**



czymi?  
nej)

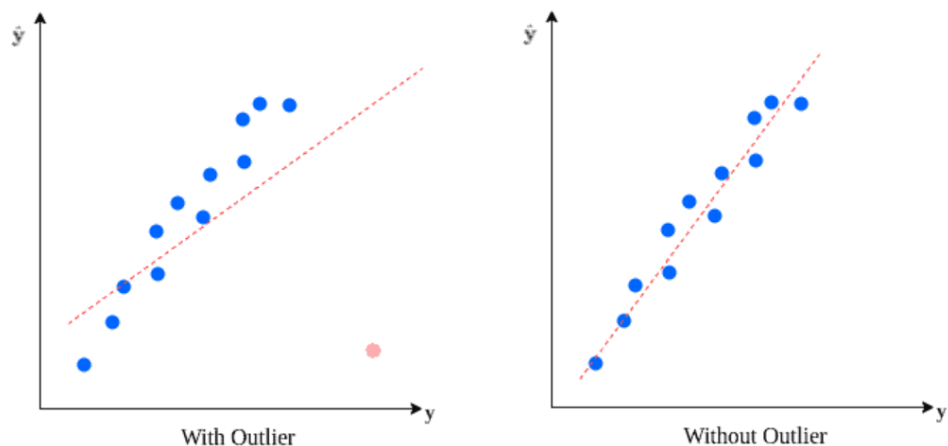




**CZYSZCZENIE DANYCH**

# Czyszczenie danych

1. Braki danych
2. Wartości odstające
3. Usuwanie vs zamiana
  - a) Wartość a priori
  - b) Wartość graniczna
  - c) Miara tendencji centralnej (dominanta, mediana, średnia)





# Czyszczenie danych

1. Braki danych
2. Wartości odstające
3. Usuwanie vs zamiana

When they asked you to clean the data and you cleaned all of it.

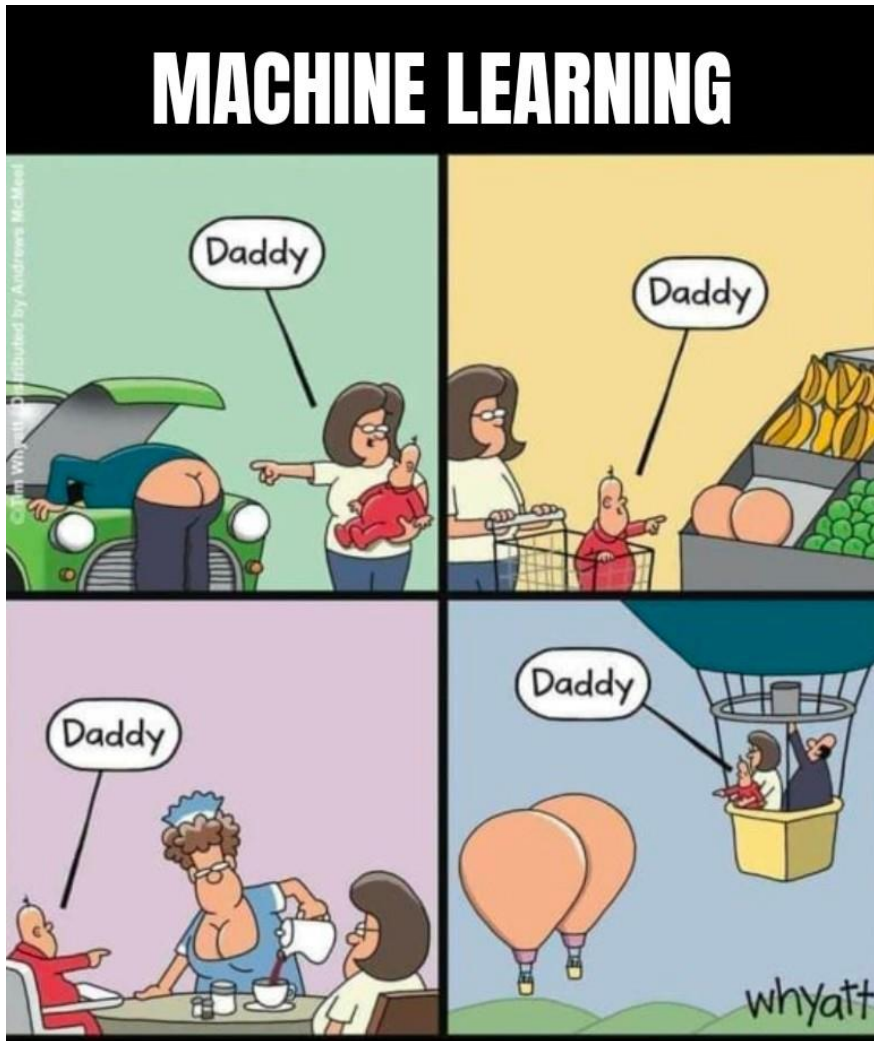
a, mediana, średnia)



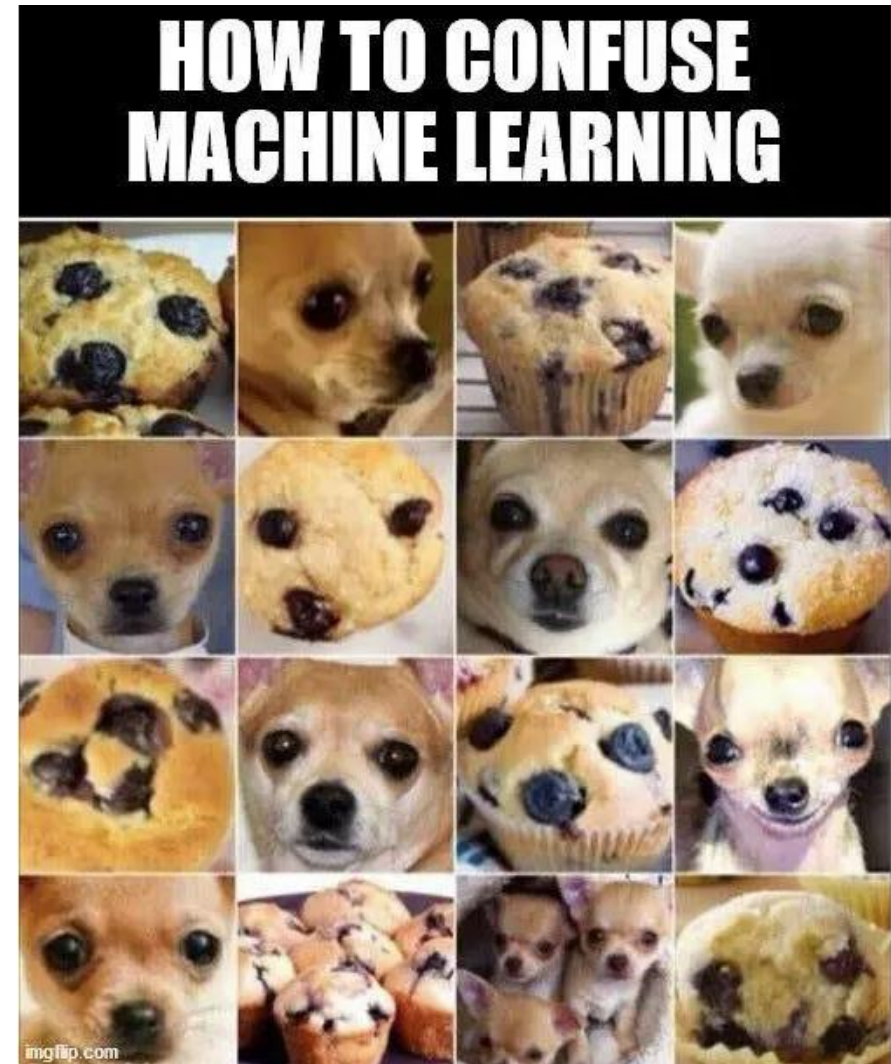




# MODEL JEST TAK DOBRY, JAK DANE NA KTÓRYCH SIĘ NAUCZYŁ !!!



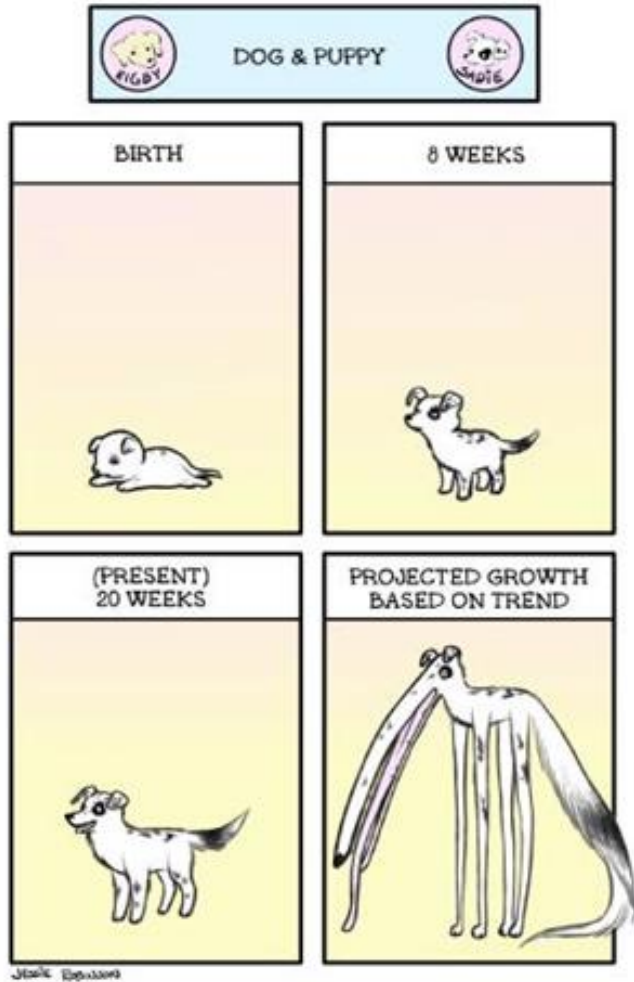
Brak usunięcia wartości odstających



Potencjalne błędy w klasyfikacji zmiennej wyjaśnianej



# MODEL JEST TAK DOBRY, JAK DANE NA KTÓRYCH SIĘ NAUCZYŁ !!!



Przyjęty za krótki  
przedział obserwacji



Dane z HD

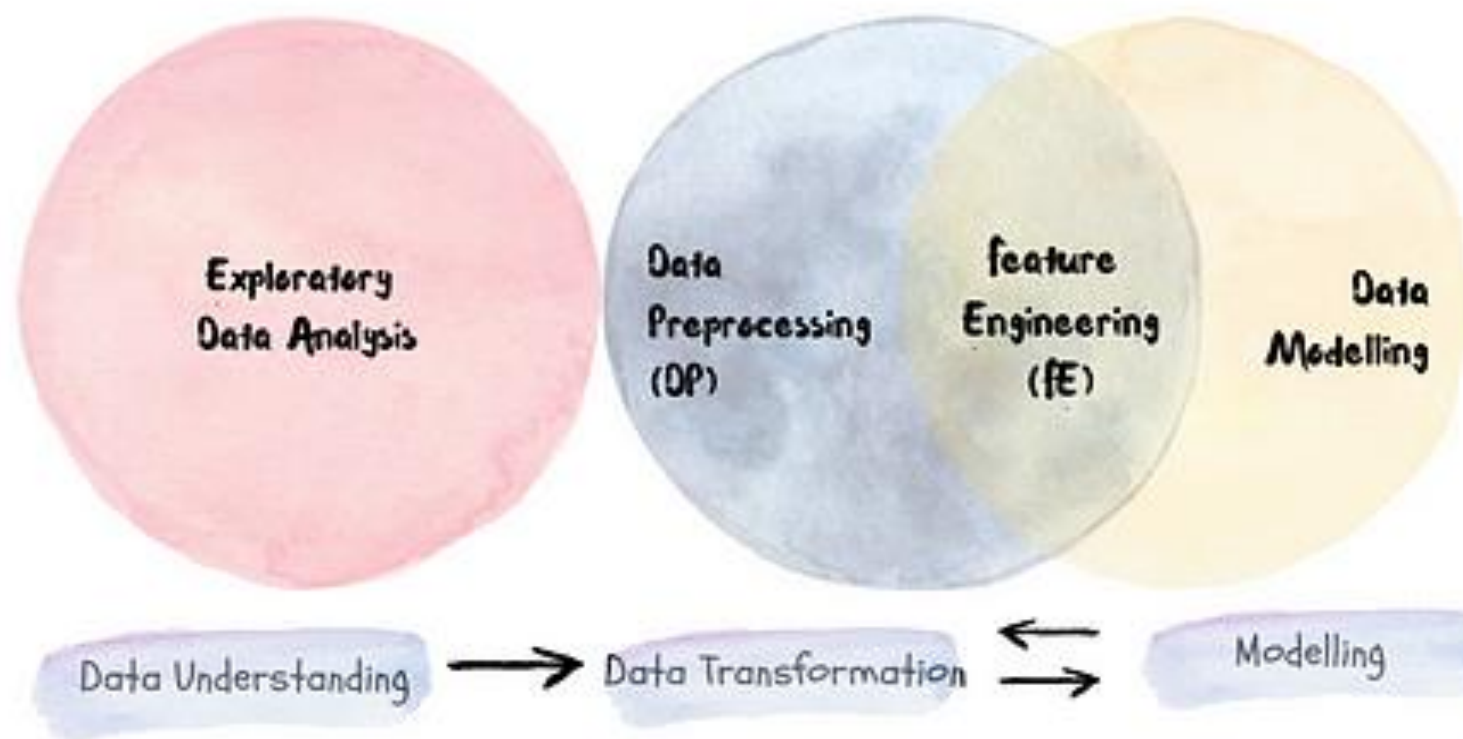




# **DOBÓR I PRZEKSZTAŁCENIA ZMIENNYCH**

# EDA — DP — FE

## DIFFERENCES



LEAH NGUYEN



# Dobór i przekształcanie zmiennych wyjaśniających (feature engineering)

1. Przekształcenia danych
  - a) Nowe zmienne
  - b) Zmiana rozkładu
  - c) Kwantyzacja



## Categorizing Continuous Variables<sup>†</sup>

Douglas G. Altman

First published: 29 September 2014 | <https://doi.org/10.1002/9781118445112.st>

<sup>†</sup> This article was originally published online in 2005 in Encyclopedia of Biostatistics Ltd and republished in Wiley StatsRef: Statistics Reference Online, 2014.



f2harrell

## Problems Caused by Categorizing Continuous Variables

> [Can J Psychiatry](#). 2002 Apr;47(3):262-6. doi: 10.1177/070674370204700307.

## Breaking up is hard to do: the heartbreak of dichotomizing continuous data

David L Streiner<sup>1</sup>

*Statistics Notes*

### The cost of dichotomising continuous variables

Douglas G Altman, Patrick Royston

<http://psych.colorado.edu/~mcclella/MedianSplit/>

Debate | [Open access](#) | Published: 29 February 2012

## Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents

[Caroline Bennette](#) & [Andrew Vickers](#) ✉

[BMC Medical Research Methodology](#) 12, Article number: 21 (2012) | [Cite this article](#)



## Categorizing continuous variables resulted in different predictors in a prognostic model for nonspecific neck pain

[Jasper M. Schellingerhout](#)<sup>a</sup> ✉, [Martijn W. Heymans](#)<sup>b c</sup>, [Henrica C.W. de Vet](#)<sup>b</sup>, [Bart W. Koes](#)<sup>a</sup>,  
[Arianne P. Verhagen](#)<sup>a</sup>

Modelowanie – dobór zmiennych wyjaśniających



# Dobór i przekształcanie zmiennych wyjaśniających (feature engineering)

1. Przekształcenia danych
  - a) Nowe zmienne
  - b) Zmiana rozkładu
  - c) Kwantyzacja
2. Dobór zmiennych
  - a) Związek ze zmienną wyjaśnianą
  - b) Współzmiennność i interakcje
  - c) Metody redukcji cech
    - i. Analiza komponentów głównych (PCA)
    - ii. Grupowanie (klasteryzacja)
    - iii. Regresja krokowa (RFECV i SFS)







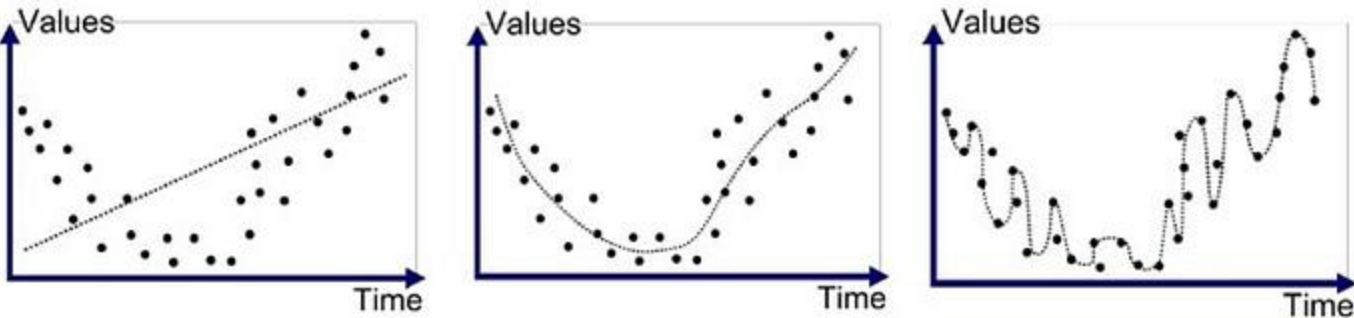
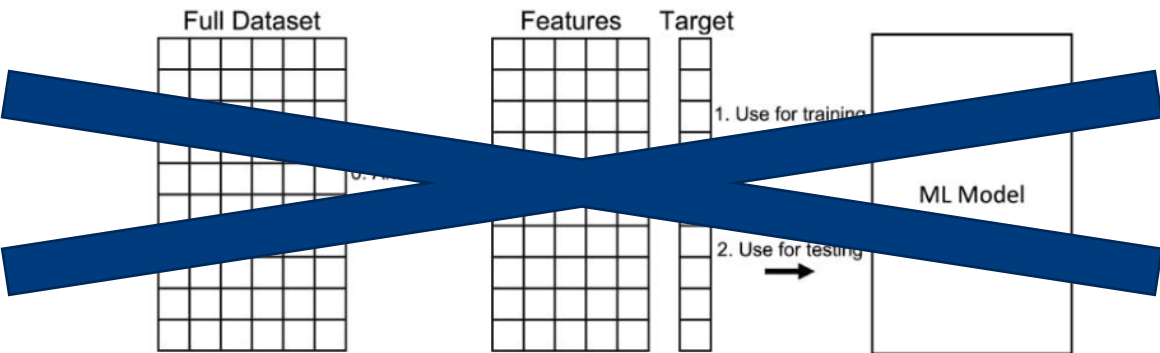


**PRZERWA !!!**



**MODELOWANIE !!!**

# Uczenie i optymalizacja modelu



Underfitted

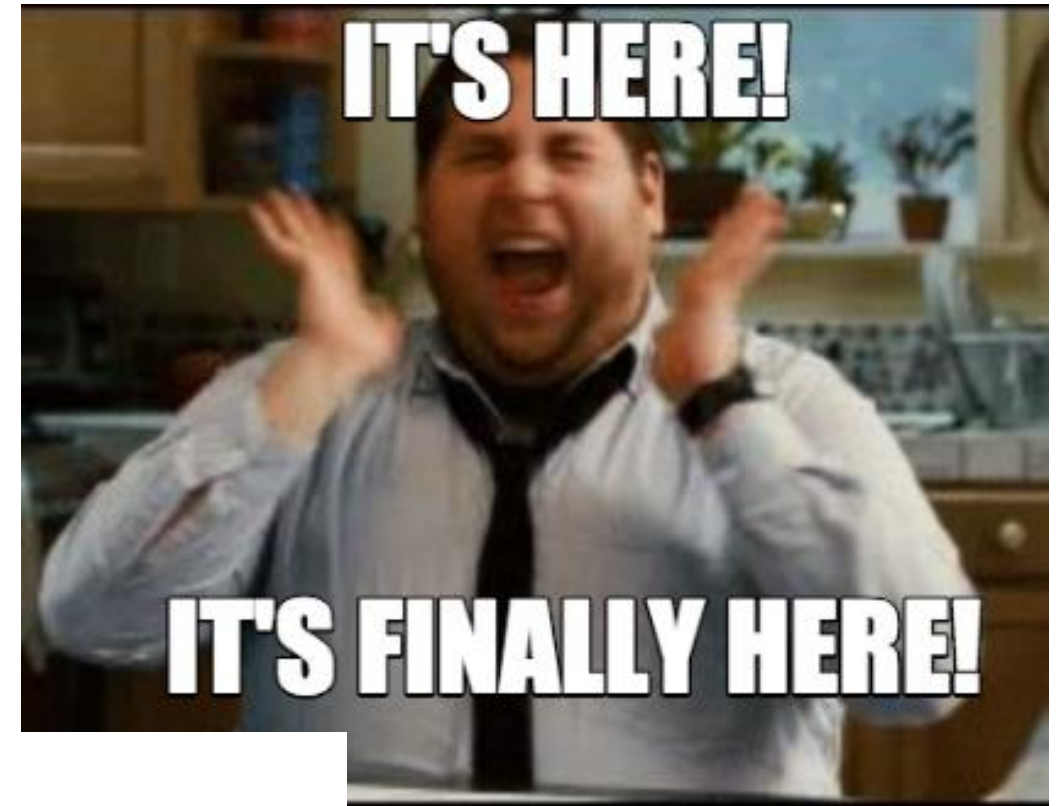
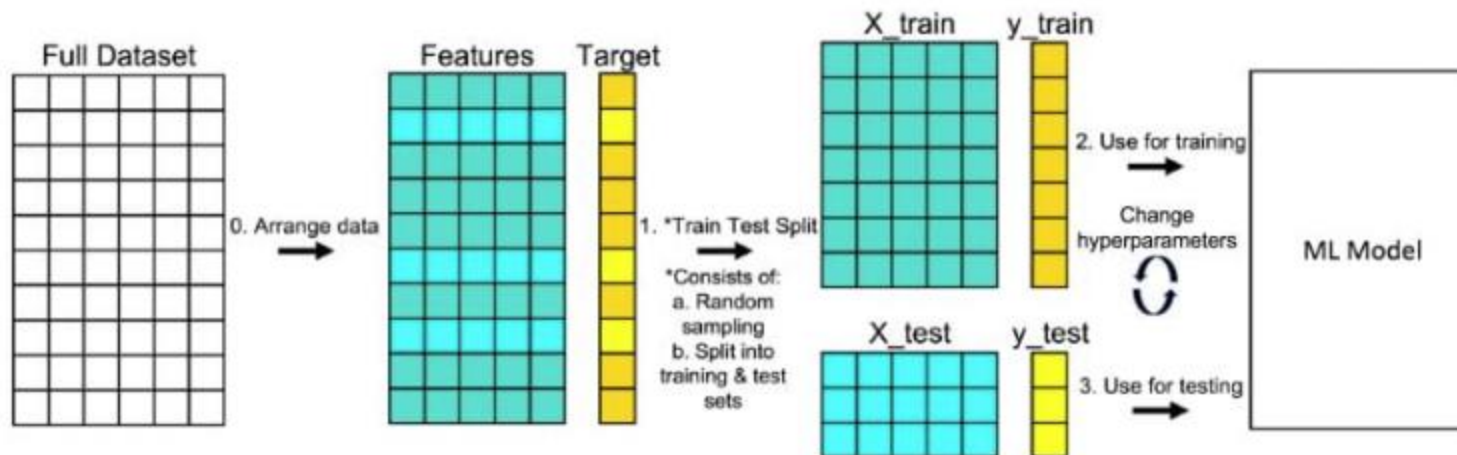
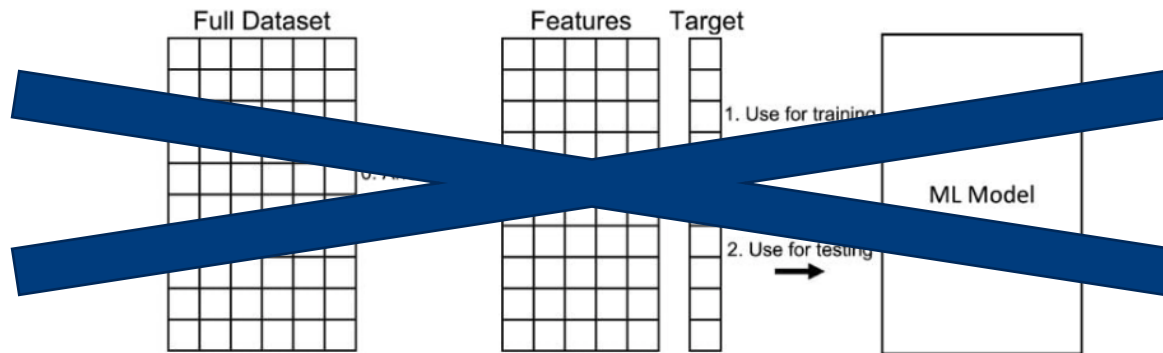
Good Fit/Robust

Overfitted



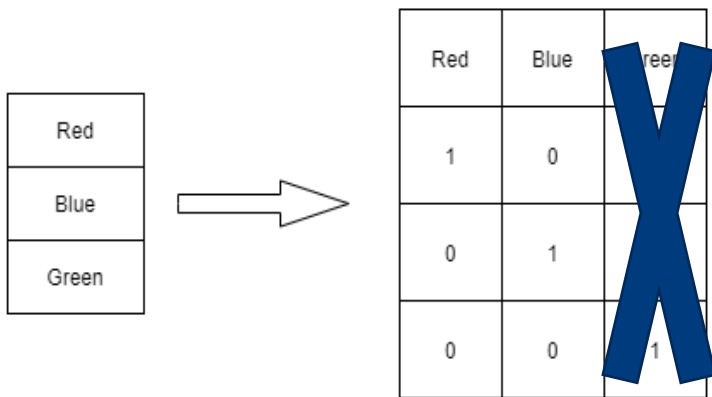
# Uczenie i optymalizacja modelu

1. Podział na sety treningowy i testowy
  - a) „Przeciekanie” danych
  - b) Podział proporcjonalny do zmiennej

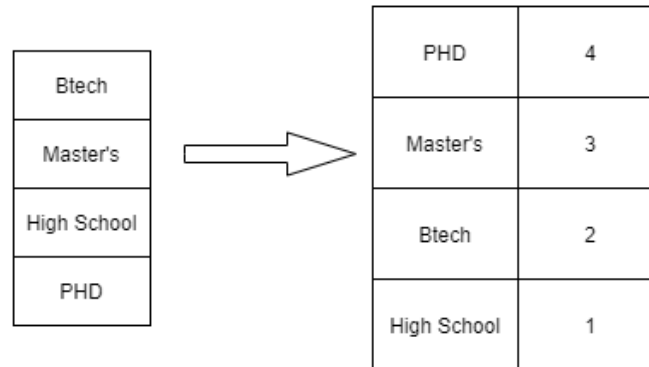


# Uczenie i optymalizacja modelu

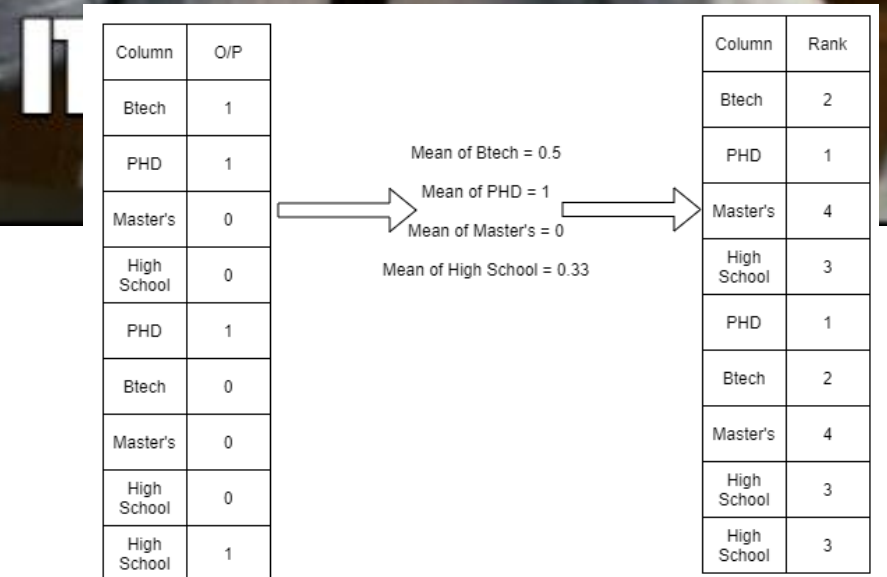
1. Podział na sety treningowy i testowy
  - a) „Przeciekanie” danych
  - b) Podział proporcjonalny do zmiennej
2. Przekształcenia danych raz jeszcze – dane kategoryjne
  - a) One-hot encoding
  - b) Label encoding
  - c) Target encoding



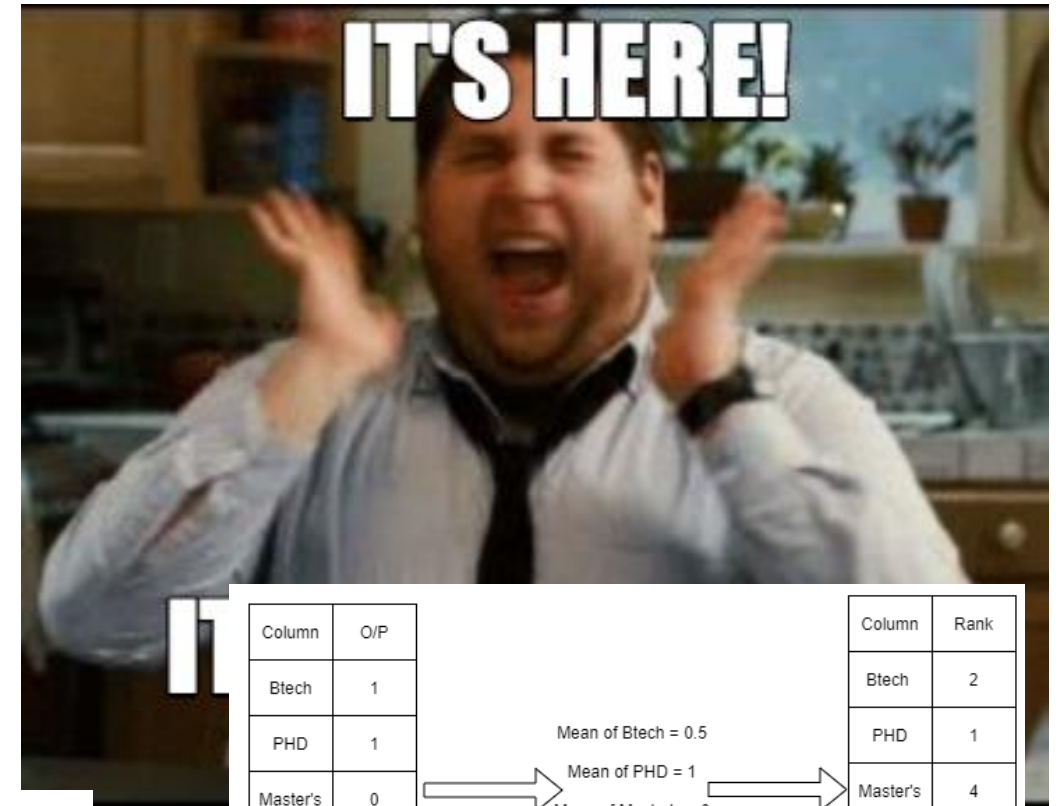
One Hot Encoding



Label Encoding



Target Encoding



# Uczenie i optymalizacja modelu

1. Podział na sety treningowy i testowy
  - a) „Przeciekanie” danych
  - b) Podział proporcjonalny do zmiennej
2. Przekształcenia danych raz jeszcze
  - a) One-hot encoding
  - b) Label encoding
  - c) Target encoding
3. Uczenie modelu i optymalizacja parametrów
  - a) Regularyzacja – konieczne przekształcenie danych ciągłych!
    - i. Skalowanie min-max
    - ii. Normalizacja (standardization / Z-score)

$$RSS = \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2.$$

$$RSS + \lambda \sum_{j=1}^p \beta_j^2$$

**RIDGE**

$$RSS + \lambda \sum_{j=1}^p |\beta_j|.$$

**LASSO**



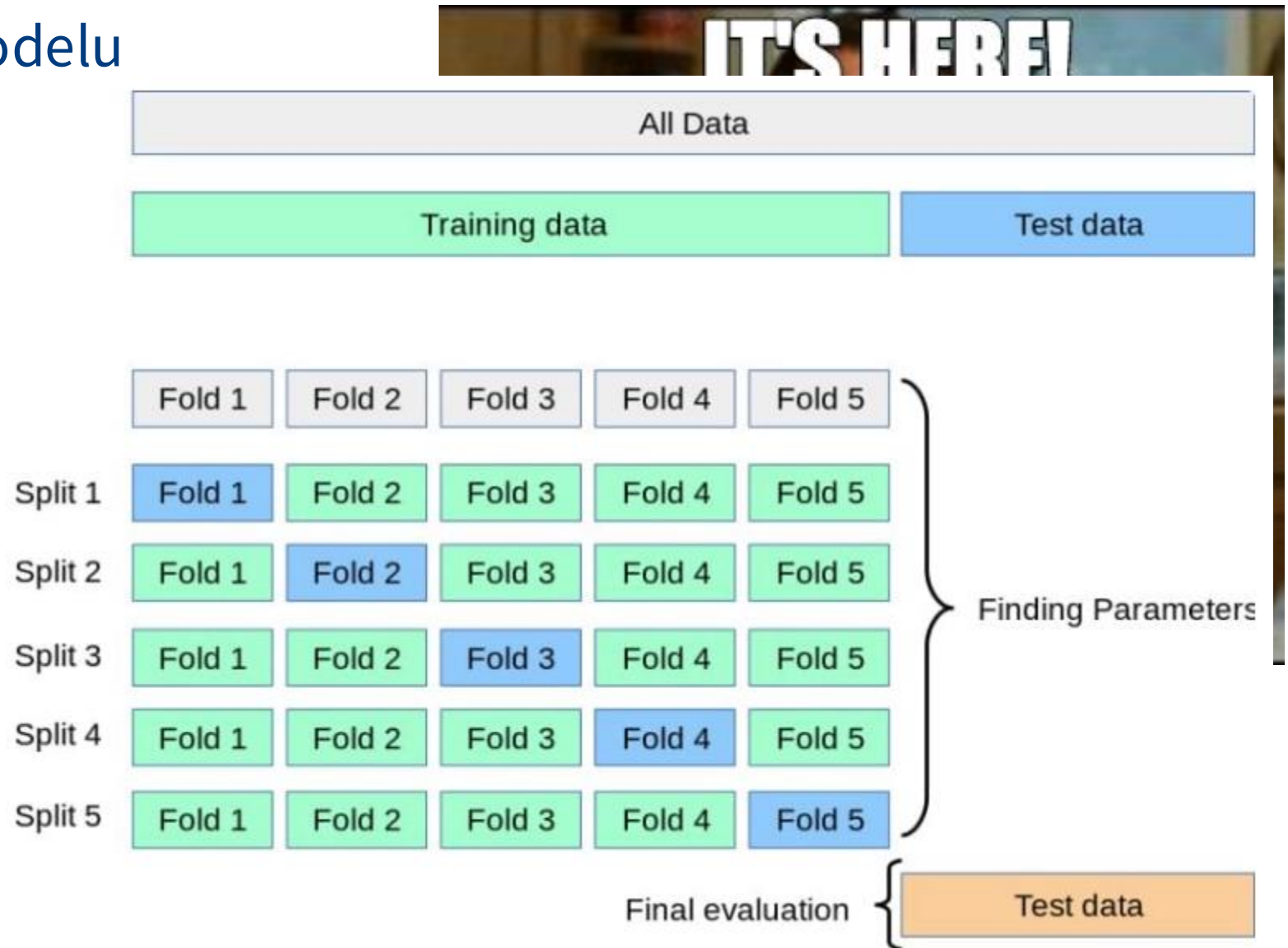
$$X_{\text{new}} = (X - X_{\text{min}}) / (X_{\text{max}} - X_{\text{min}})$$

$$X_{\text{new}} = (X - \text{mean}) / \text{Std}$$

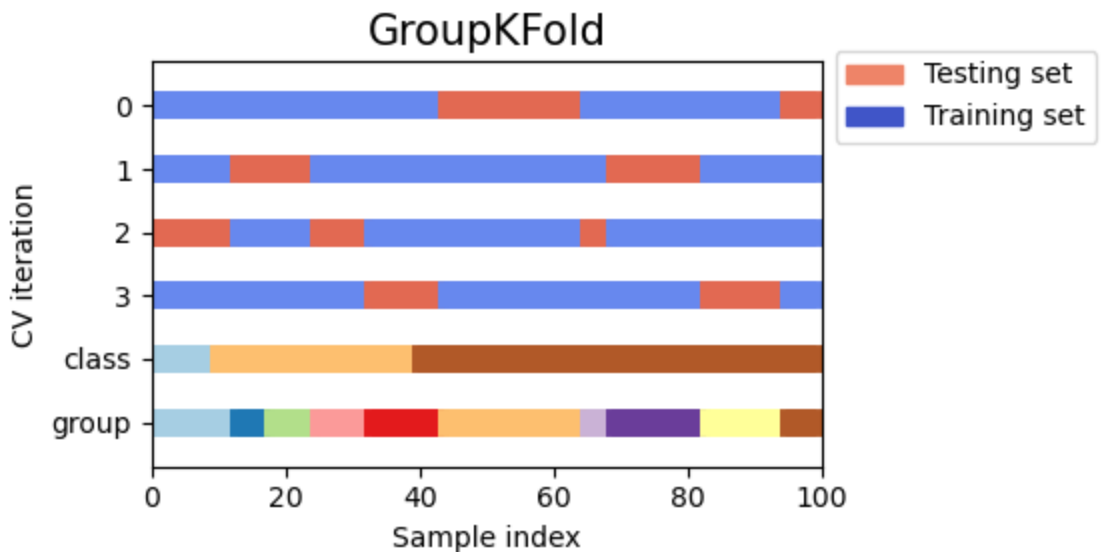
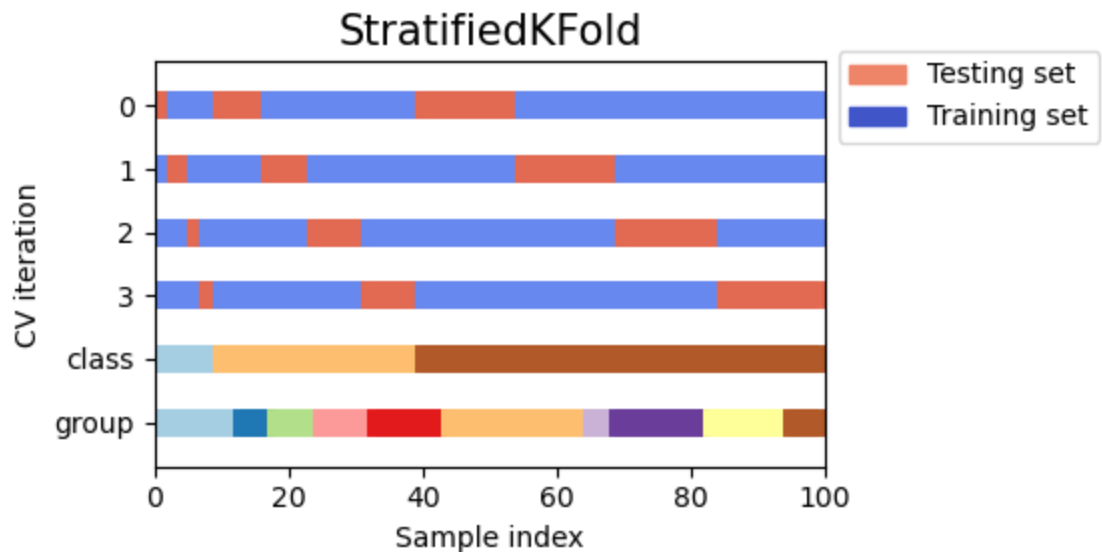
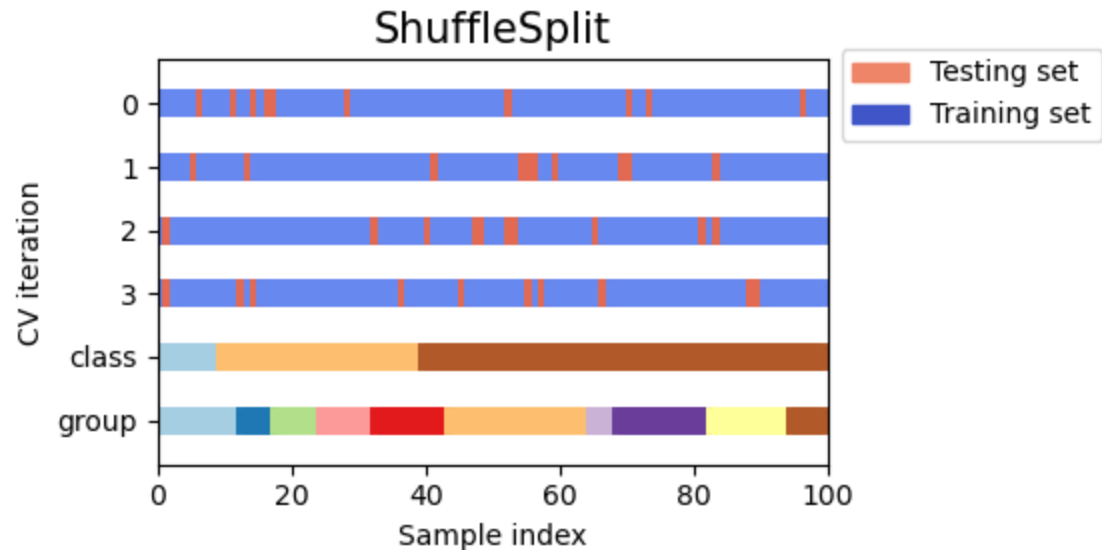
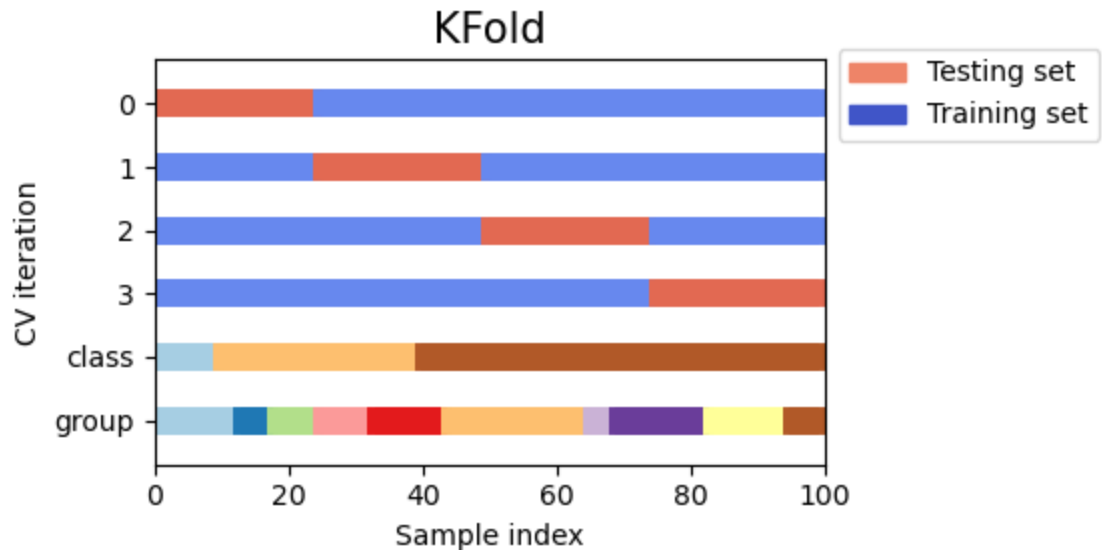


# Uczenie i optymalizacja modelu

1. Podział na sety treningowy i testowy
  - a) „Przeciekanie” danych
  - b) Podział proporcjonalny do z
2. Przekształcenia danych raz jeszcze
  - a) One-hot encoding
  - b) Ordinal encoding
  - c) Target encoding
3. Uczenie modelu i optymalizacja
  - a) Regularyzacja
  - b) Walidacja krzyżowa



# Uczenie i optymalizacja modelu – walidacja krzyżowa





**EWALUACJA MODELU**

# Ewaluacja modelu

1. Model „zero”

2. Metryki

a) Metryki regresji

i.  $R^2$

ii. Mean Absolute Error

iii. Mean Squared Error

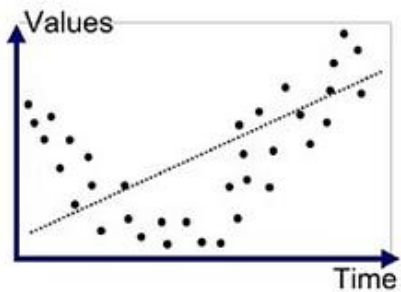
iv. Mean Poisson / Gamma / Tweedie Deviance

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

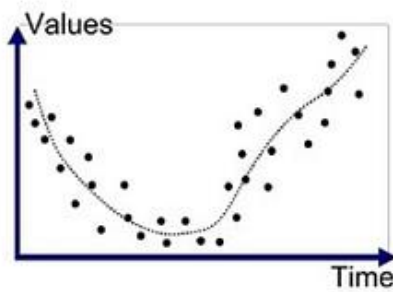
$$\text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i|.$$

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2.$$

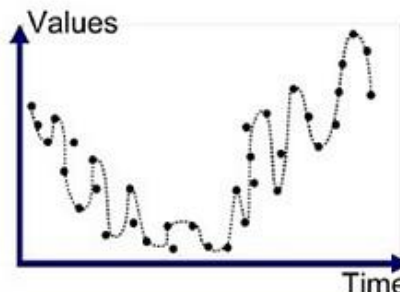
$$D(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \begin{cases} (y_i - \hat{y}_i)^2, & \text{for } p = 0 \text{ (Normal)} \\ 2(y_i \log(y_i/\hat{y}_i) + \hat{y}_i - y_i), & \text{for } p = 1 \text{ (Poisson)} \\ 2(\log(\hat{y}_i/y_i) + y_i/\hat{y}_i - 1), & \text{for } p = 2 \text{ (Gamma)} \\ 2 \left( \frac{\max(y_i, 0)^{2-p}}{(1-p)(2-p)} - \frac{y_i \hat{y}_i^{1-p}}{1-p} + \frac{\hat{y}_i^{2-p}}{2-p} \right), & \text{otherwise} \end{cases}$$



Underfitted



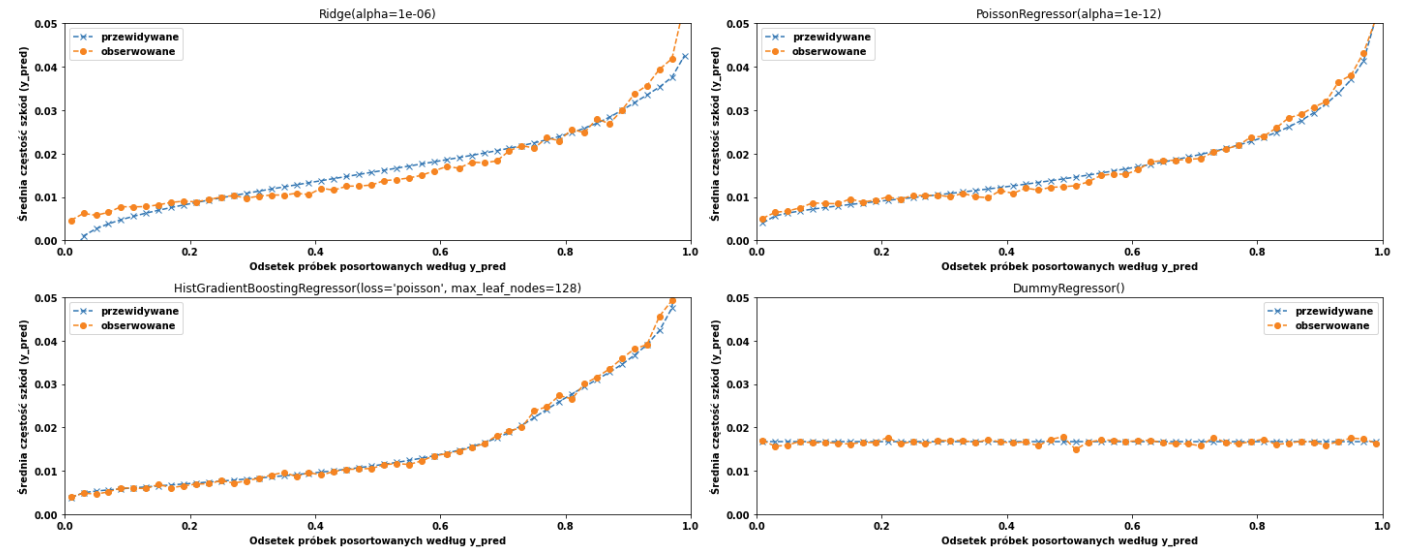
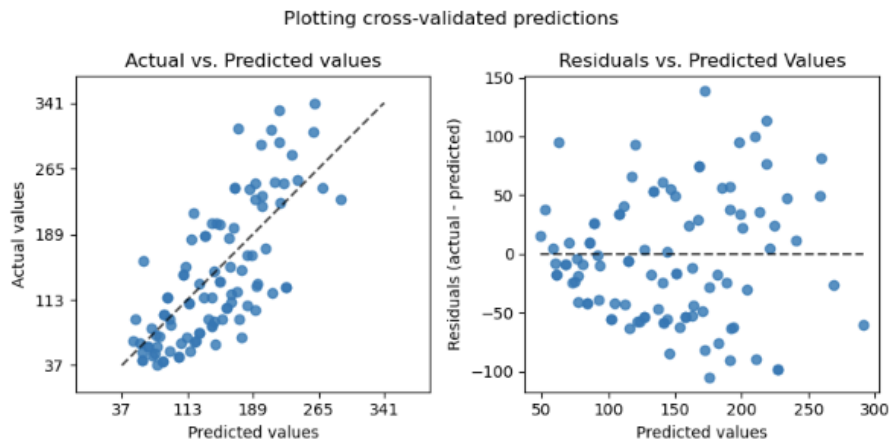
Good Fit/Robust



Overfitted

# Ewaluacja modelu

1. Model „zero”
2. Metryki
  - a) Metryki regresji
    - i.  $R^2$
    - ii. Mean Absolute Error
    - iii. Mean Squared Error
    - iv. Mean Poisson / Gamma / Tweedie Deviance
  - b) Wizualizacje

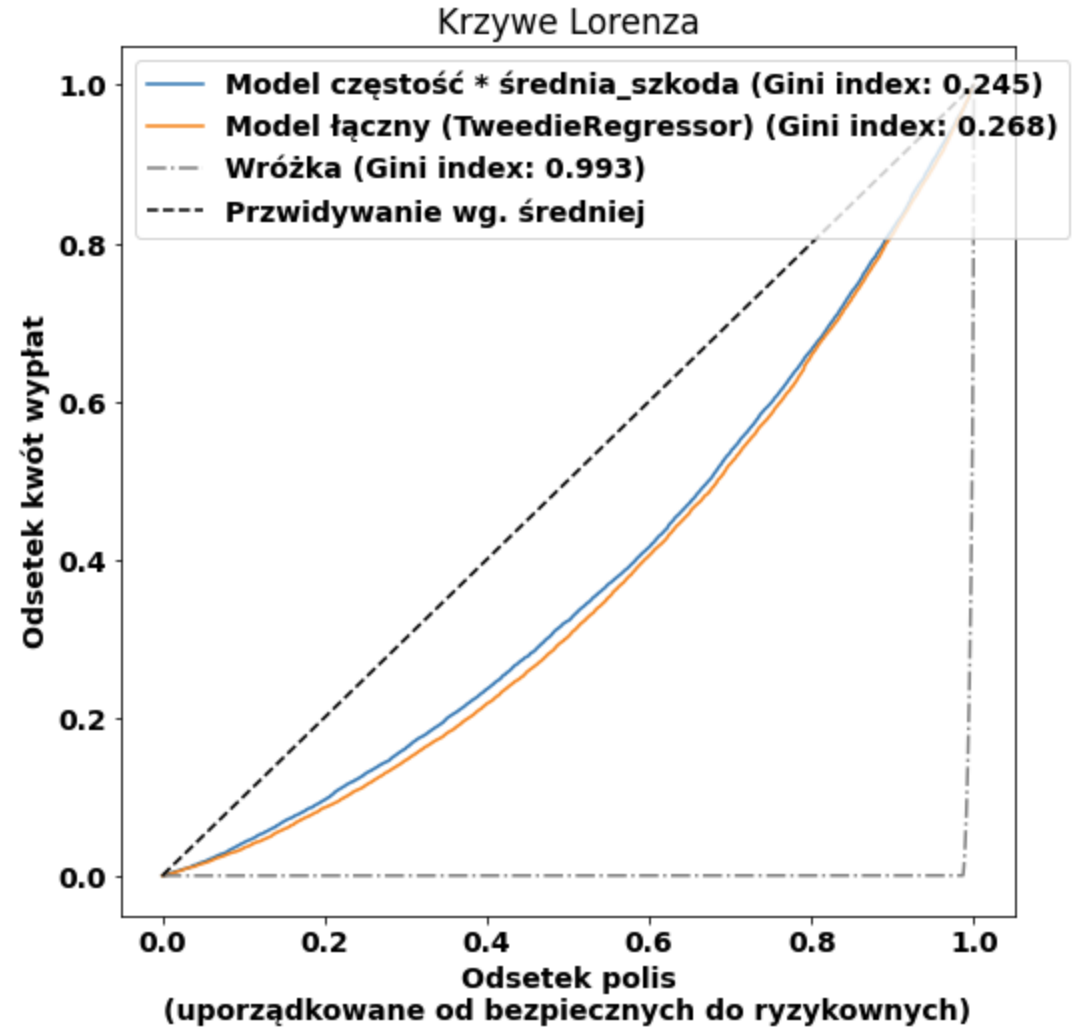
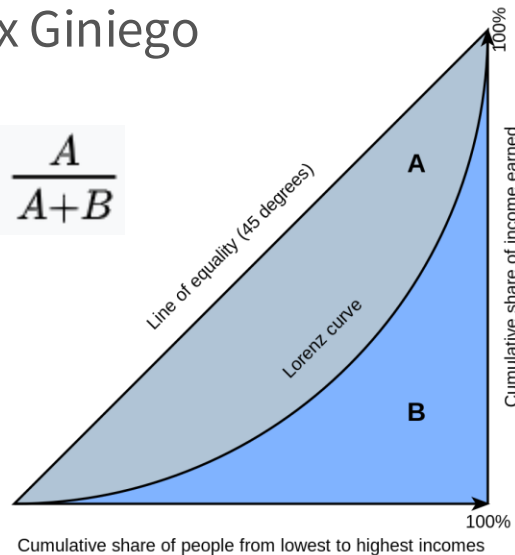




# Ewaluacja modelu

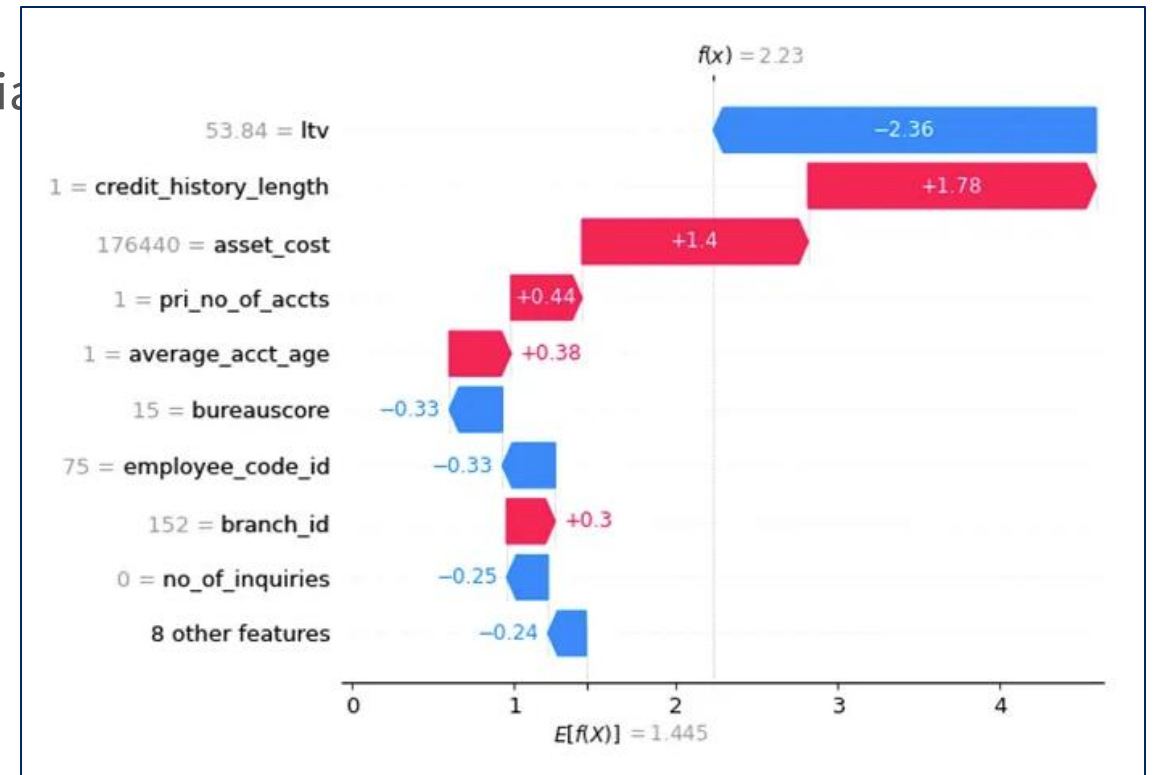
1. Model „zero”
2. Metryki
  - a) Metryki regresji
    - i.  $R^2$
    - ii. Mean Absolute Error
    - iii. Mean Squared Error
    - iv. Mean Poisson / Gamma / Tweedie Deviance
  - b) Wizualizacje
  - c) Krzywe Lorenza
    - i. Index Giniego

$$\text{Gini} = \frac{A}{A+B}$$



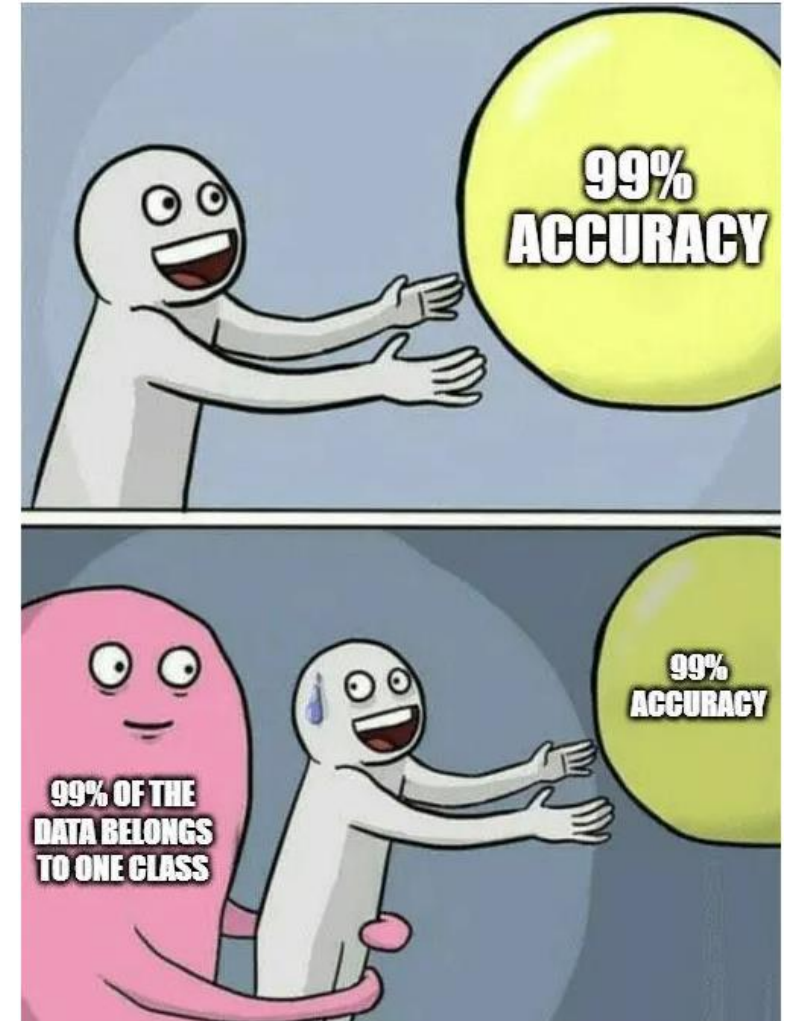
# Ewaluacja modelu

1. Model „zero”
2. Metryki
  - a) Metryki regresji
    - i.  $R^2$
    - ii. Mean Absolute Error
    - iii. Mean Squared Error
    - iv. Mean Poisson / Gamma / Tweedie Deviance
  - b) Wizualizacje
  - c) Krzywe Lorenza
    - i. Index Giniego
  - d) Wartości SHAP



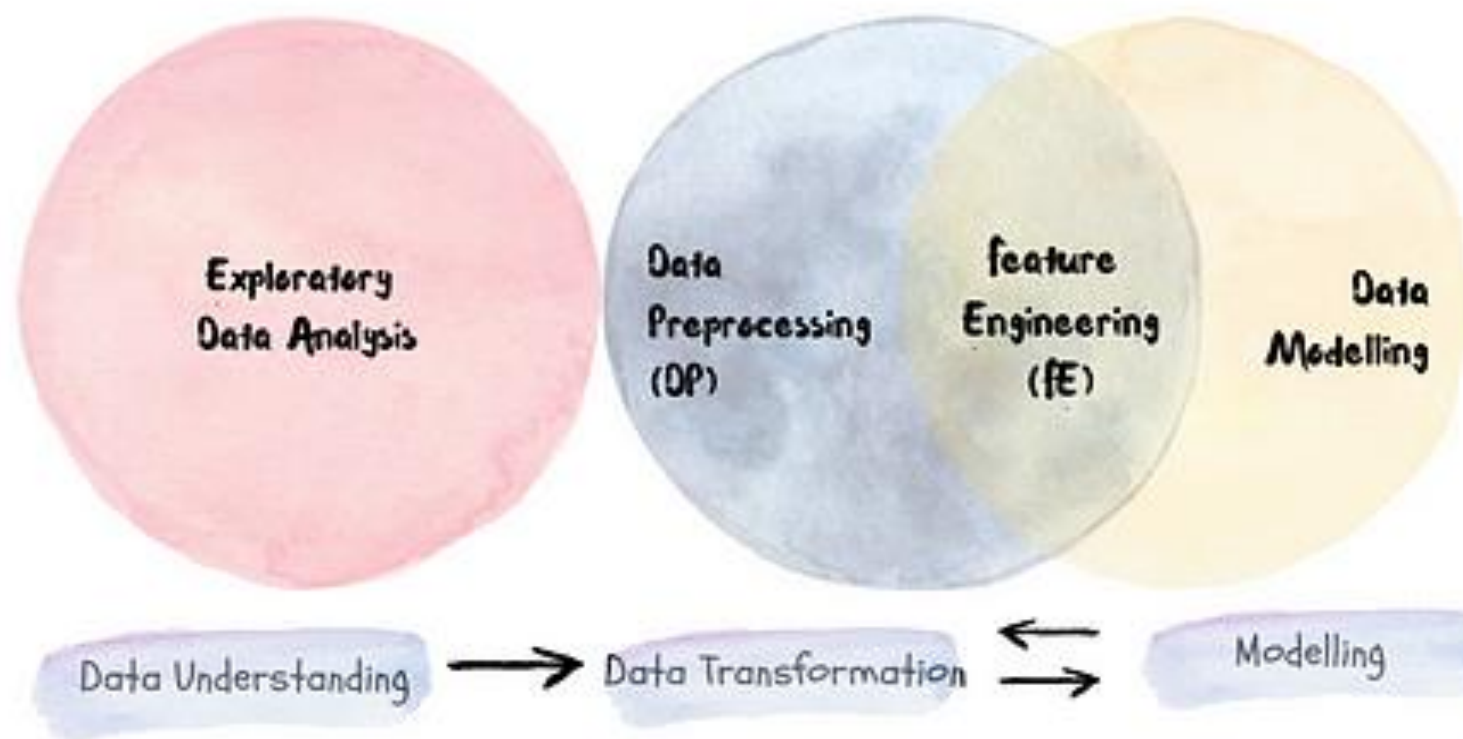
# Ewaluacja modelu – NIE DZIAŁA !!!

1. Źle dobrana metryka optymalizacji parametrów modelu
2. Niewłaściwy model
3. Nie uwzględnienie ważnej zmiennej
4. Nie uwzględnienie ważnej interakcji
5. Niewłaściwe wyczyszczenie danych
6. Niewłaściwe przekształcenie danych wejściowych
7. Niewłaściwa walidacja krzyżowa
8. „Przeciekanie danych”



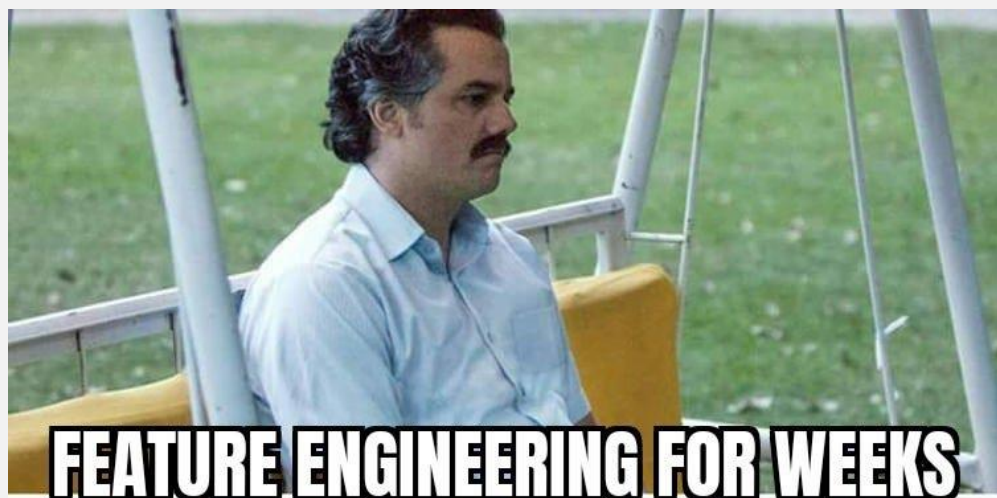
# EDA — DP — FE

## DIFFERENCES



LEAH NGUYEN





**FEATURE ENGINEERING FOR WEEKS**



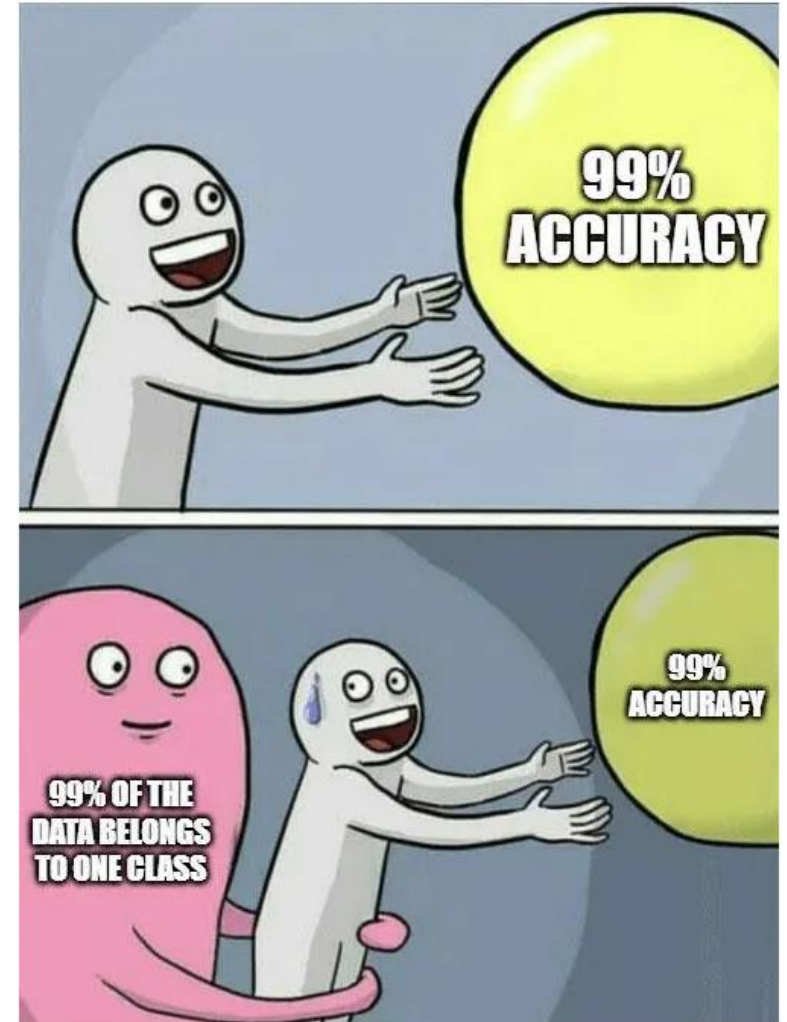
**NO IMPROVEMENT IN PERFORMANCE**





# Ewaluacja modelu – NIE DZIAŁA !!!

1. Źle dobrana metryka optymalizacji parametrów modelu
2. Niewłaściwy model
3. Nie uwzględnienie ważnej zmiennej
4. Nie uwzględnienie ważnej interakcji
5. Niewłaściwe wyczyszczenie danych
6. Niewłaściwe przekształcenie danych wejściowych
7. Niewłaściwa walidacja krzyżowa
8. „Przeciekanie danych”
9. ...dane nie umożliwiają predykcji





**PODSUMOWANIE**

# O czym było?

1. Uczenie maszynowe – czym jest i czym nie jest?
2. Modelowanie
  - a) Wybór modelu
    - i. Ogólne modele liniowe
    - ii. Zmienna wyjaśniana
    - iii. Podział zmiennych wyjaśniających
  - b) Analiza eksploracyjna
    - i. Typy danych
    - ii. Braki danych
    - iii. Duplikaty
    - iv. Statystyki opisowe
    - v. Rozkłady zmiennych
  - c) Czyszczenie danych:
    - i. Braki danych
    - ii. Wartości odstające
    - iii. Usuwanie vs zamiana

- d) Dobór zmiennych (feature engineering)
  - i. Przekształcenia danych
  - ii. Dobór zmiennych
- e) Uczenie i optymalizacja parametrów modelu
  - i. Sety testowy i treningowy
  - ii. Przekształcenia danych
  - iii. Regularyzacja
  - iv. Walidacja krzyżowa
- f) Ewaluacja modelu
  - i. Model „zero”
  - ii. Metryki
  - iii. Shap

# Gdzie szukać inspiracji?

1. Kaggle <https://www.kaggle.com/>
2. Medium / TowardsDataScience <https://medium.com/>
3. StackOverflow <https://stackoverflow.com/>
4. StackExchange <https://stackexchange.com/>
5. Dokumentacja bibliotek:
  - a) pandas <https://pandas.pydata.org/>
  - b) numpy <https://numpy.org/>
  - c) seaborn <https://seaborn.pydata.org/>
  - d) matplotlib <https://matplotlib.org/>
  - e) statsmodels <https://www.statsmodels.org/stable/index.html>
  - f) scikit-learn <https://scikit-learn.org/stable/>
  - g) explainer dashboard <https://explainerdashboard.readthedocs.io/en/latest/>





**DZIĘKUJĘ ZA UWAGĘ!**