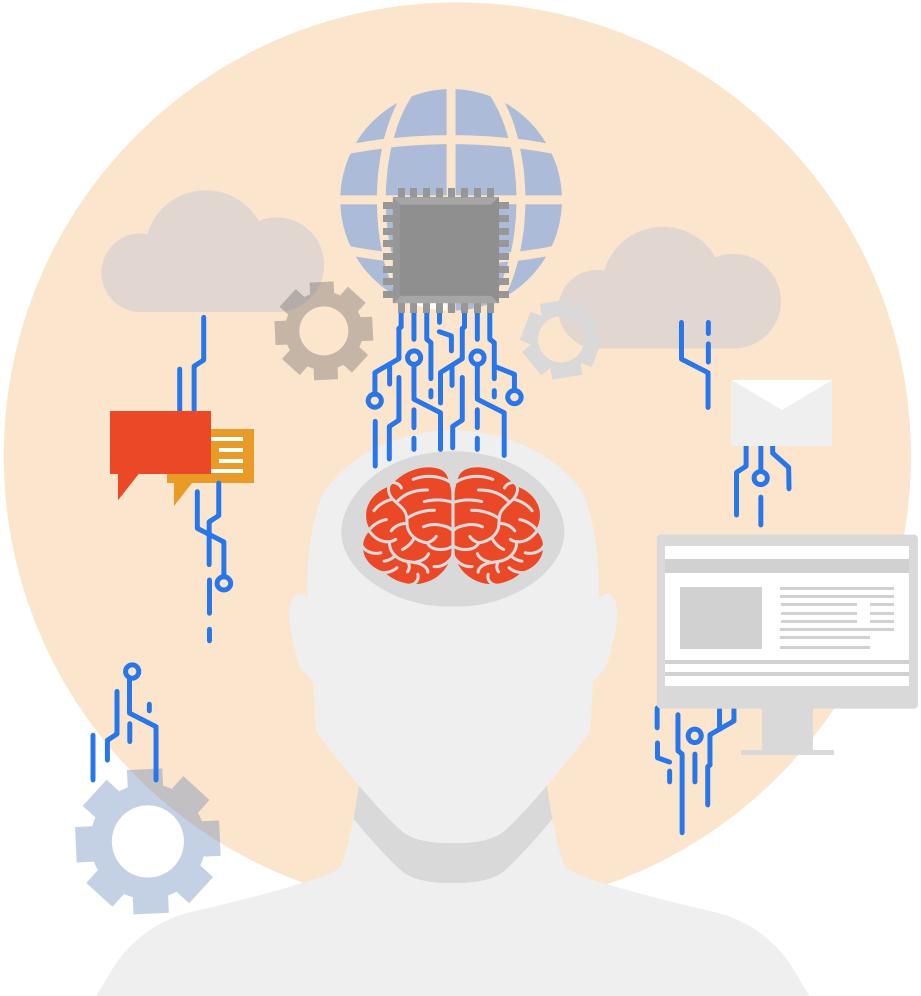


Machine Learning Final Project

Karolina Solarska
Zuzanna Kostecka

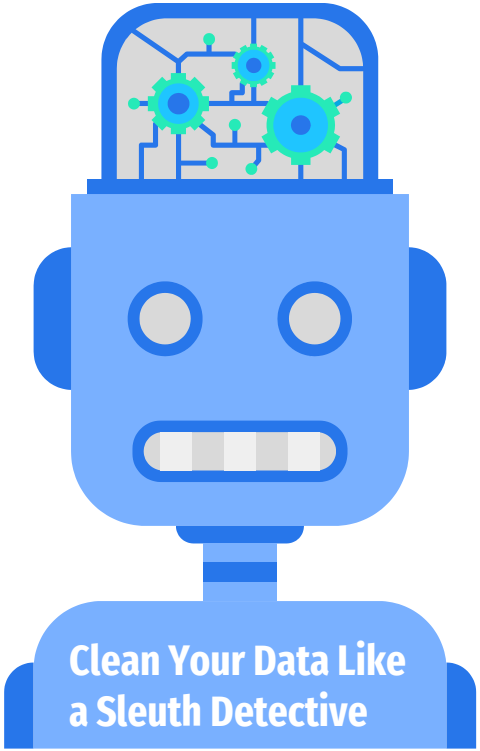
WNE UW 2023

Regression

Data preparation

01 Convert columns to the appropriate data type

02 Dealing with missing values



03 Dealing with high correlation

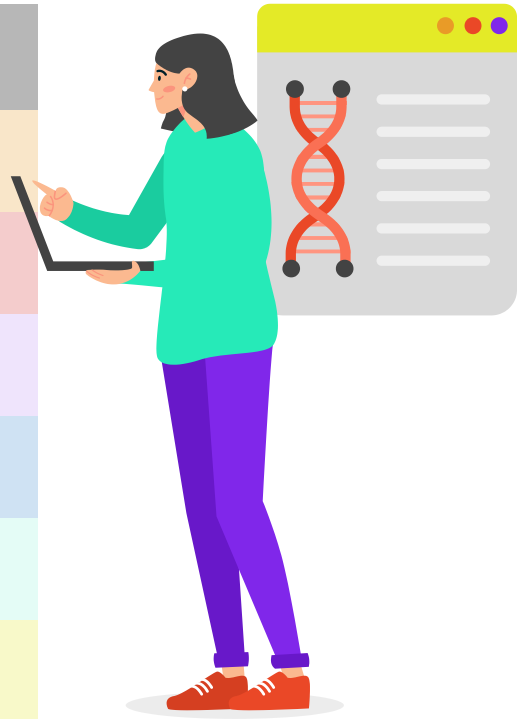
04 Encoding Categorical Variables

05 Feature Scaling

06 Feature engineering

Machine Learning Infographics

Models	MAPE
Linear Regression	0.1571
Lasso	0.1577
Elastic Net	0.1581
XGBoost	0.1542
Decision Tree	0.1917
Random Forest	0.1560



Classification

Data preparation

01

NA's handling

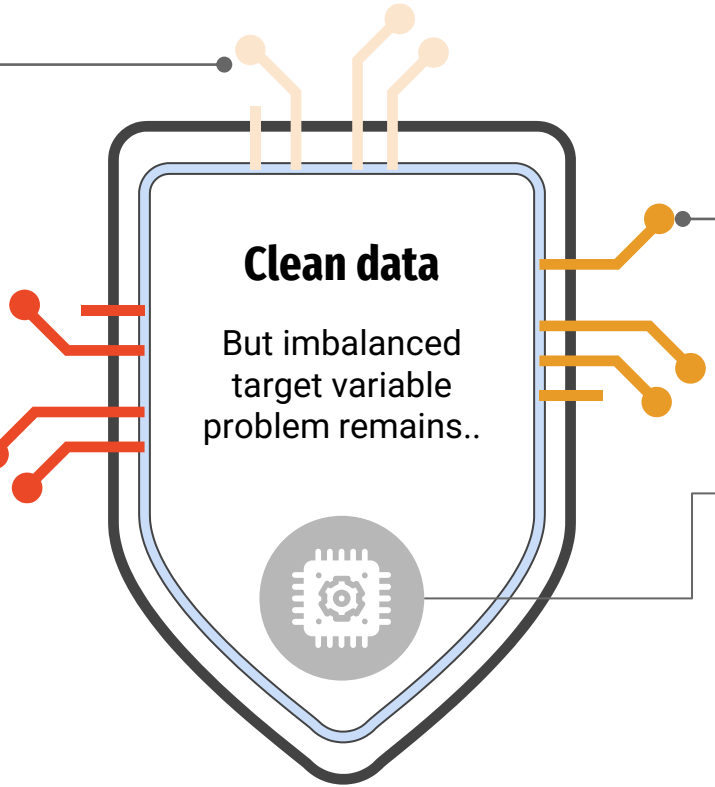
Categorical: replace randomly form non-missing

Continuous: random from assumed distribution

02

Outliers

Winsorization, scaling (standarization)



03

Drop columns

Drop not meaningful columns

04

Train and test

Internal split of training database into train and test

Various algorithms and their crucial hyperparameters

KNN

n = 13

SVM

C=100.0
Kernel = linear

Elastic net

Alpha = 0.01
L1_ratio = 0.25

Ridge

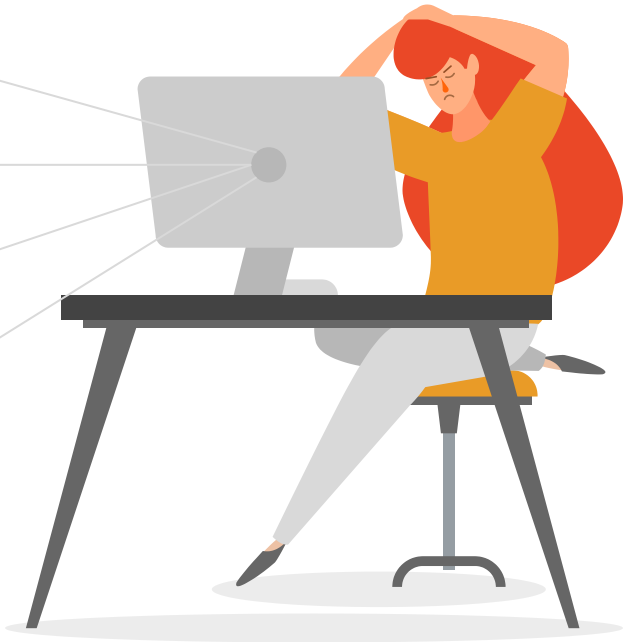
The Earth is the third planet from the Sun

01

02

03

04



Choosing the best algorithm

We still have in mind our imbalanced data problem

Evaluation metrics

	KNN	SVM	EN	Ridge
f1	0.88	0.95	1	1
AUC	0.88	0.88	1	1

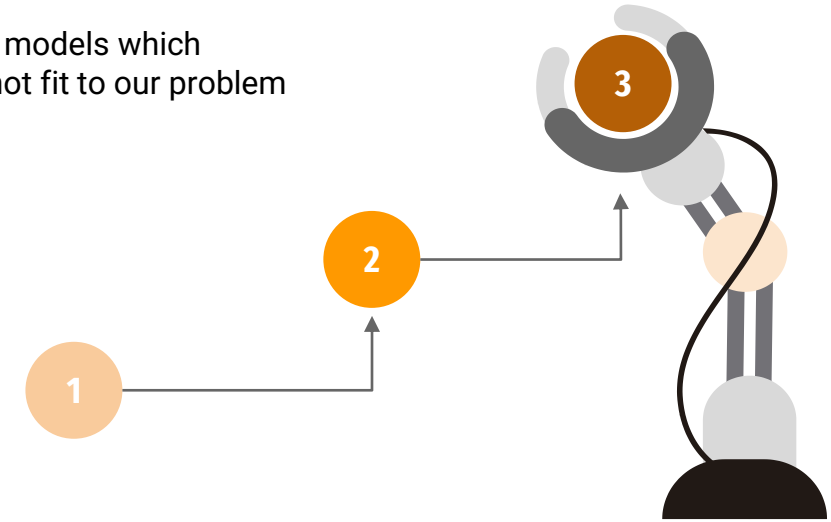
Converted to binary

Rejection

We reject the models which primarily do not fit to our problem

Prediction

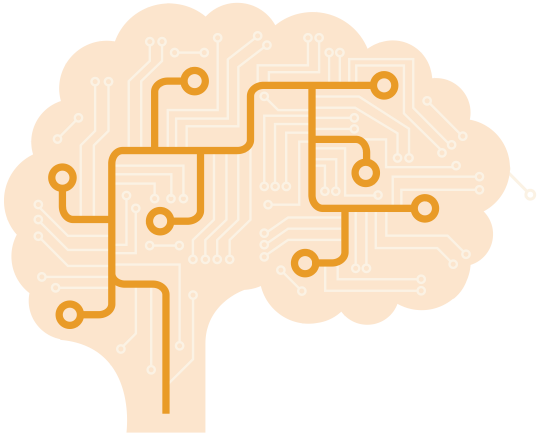
We conduct a prediction on the appropriate data



Rejection step

SVM

C=100.0
Kernel = linear



Evaluate the model’s performance

↑

Choose the model

↑

Compare the overall performance between models

↑

Reject the inappropriate models (Ridge and Elastic Net)

Thank you