Investment Campaign Case Study

Vojtěch Flídr
Zuzana Shivram Sami
Daniela Pinkasová

Data Science Coding Bootcamp Team

Case Study Definition & Objectives

- Banking campaign for investment product
- First round of campaign already performed, results are available
  - Result for each client: success = client invested, failure = client did not invest
- Objective of the case study is to target additional 3,000 clients for second round of campaign who are most likely to invest and therefore maximize bank's revenue
- Clients will be offered a cash-back bonus of 1,000 CZK if they decide to invest in the product

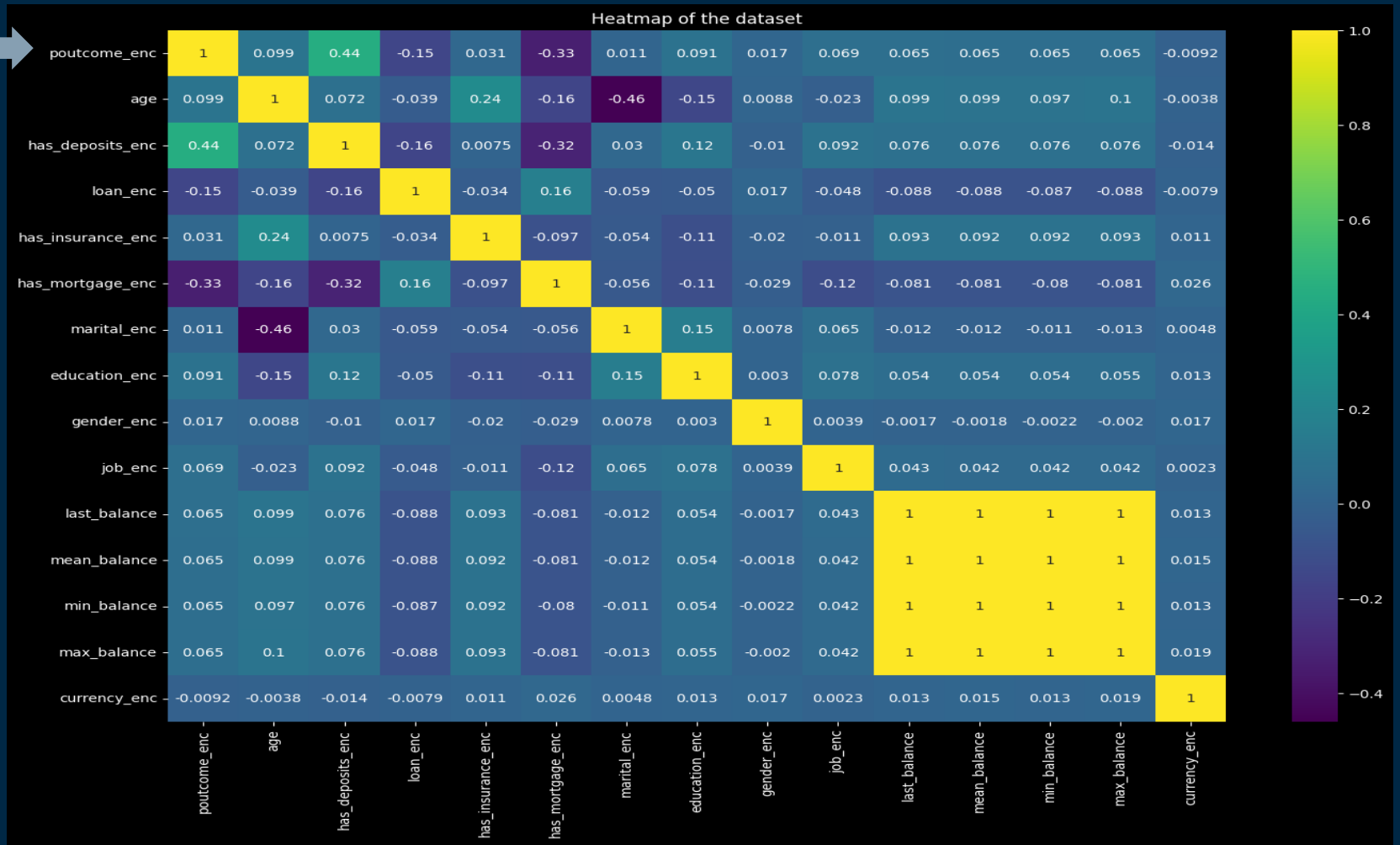
Workflow


Data Preprocessing
Feature Engineering
Model Selection, Hyperparameter Tuning
Selection of the Best Model
Evaluating the Results
Conclusion

Data Preprocessing

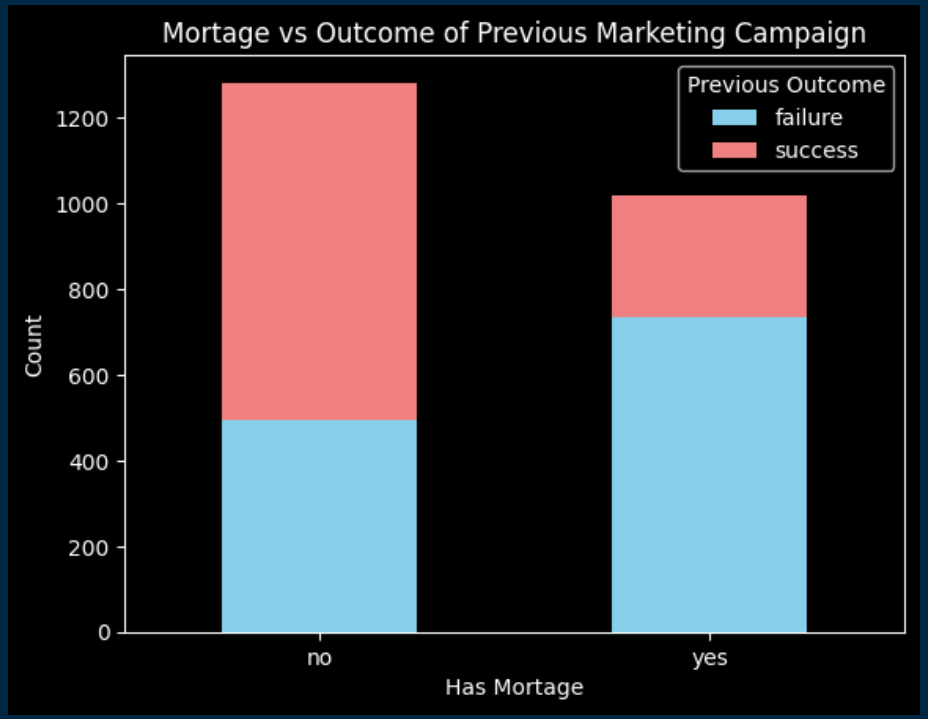
- Dataset of 11,162 clients, for 2,299 results from first round are available
- Database of 4 tables
  - Client's personal info
  - Client's bank product info
  - Client's balances
  - Results from first round of investment campaign
- Data cleaning
  - Replacing missing values
  - Label encoding of categorical variables
  - Client balances recalculated to CZK using daily FX rates for EUR and USD

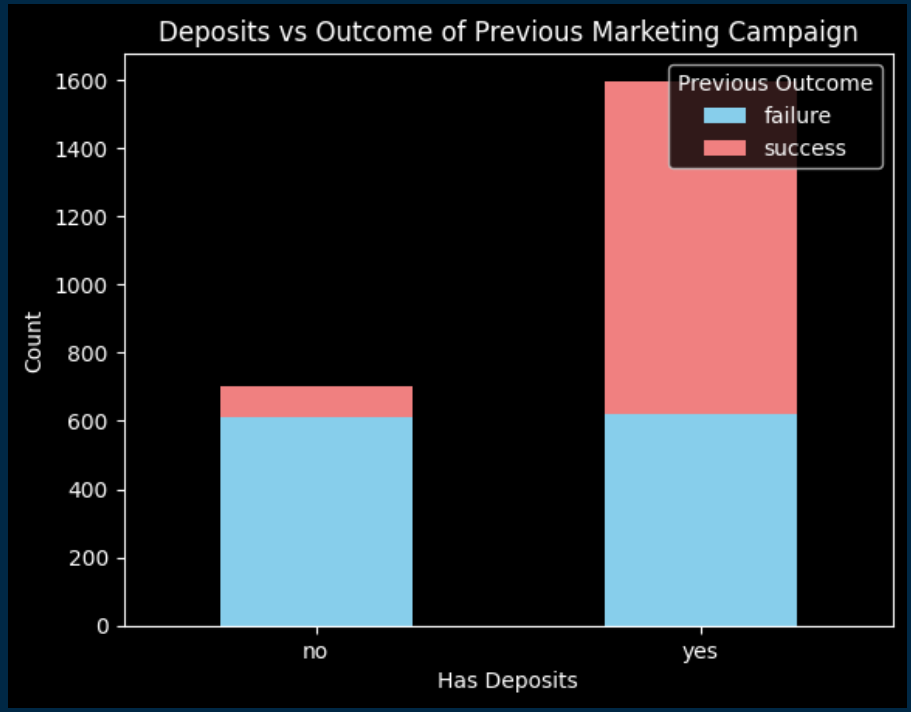
Visualization of variable relationships



Exploratory Analysis of First Round

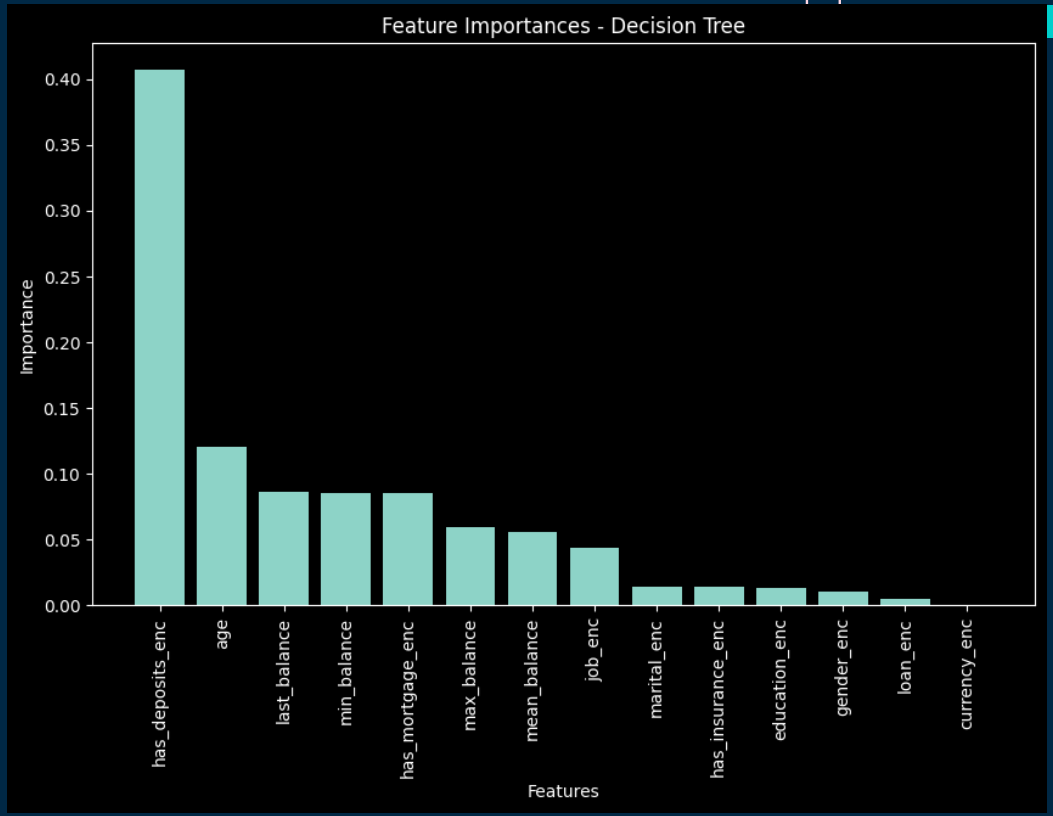
- Clients who do not have mortgage and have deposits were more likely to invest in first round of campaign





Feature Engineering

- Developing new variables
  - Mean, Last, Min, Max balances
  - Currency
- Selecting most important features based on a Decision Tree algorithm



Model Selection

- Random Forest
- Logistic Regression
- K-Nearest Neighbors (KNN)
- Adaptive Boosting (AdaBoost)
- Neural Network

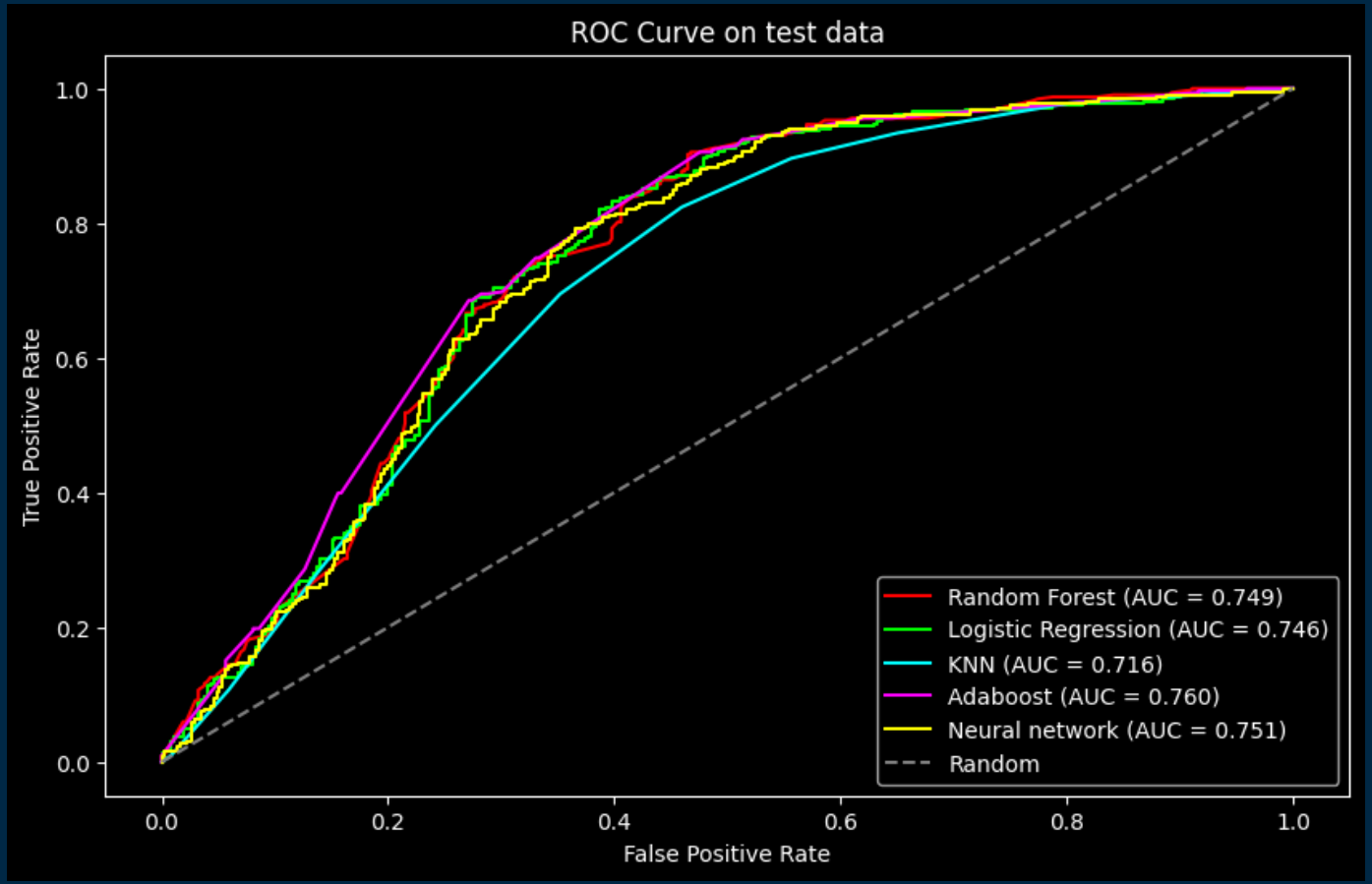
Hyperparameter tuning

- Finding the best combination of parameters to achieve the maximum performance of the models
- E.g. What should be the number of trees in a random forest or number of layers in neural network?
- GridSearch - build model for each combination of parameters, evaluating each model and selecting the architecture which produces the best results

Metrics Used for Model Evaluation

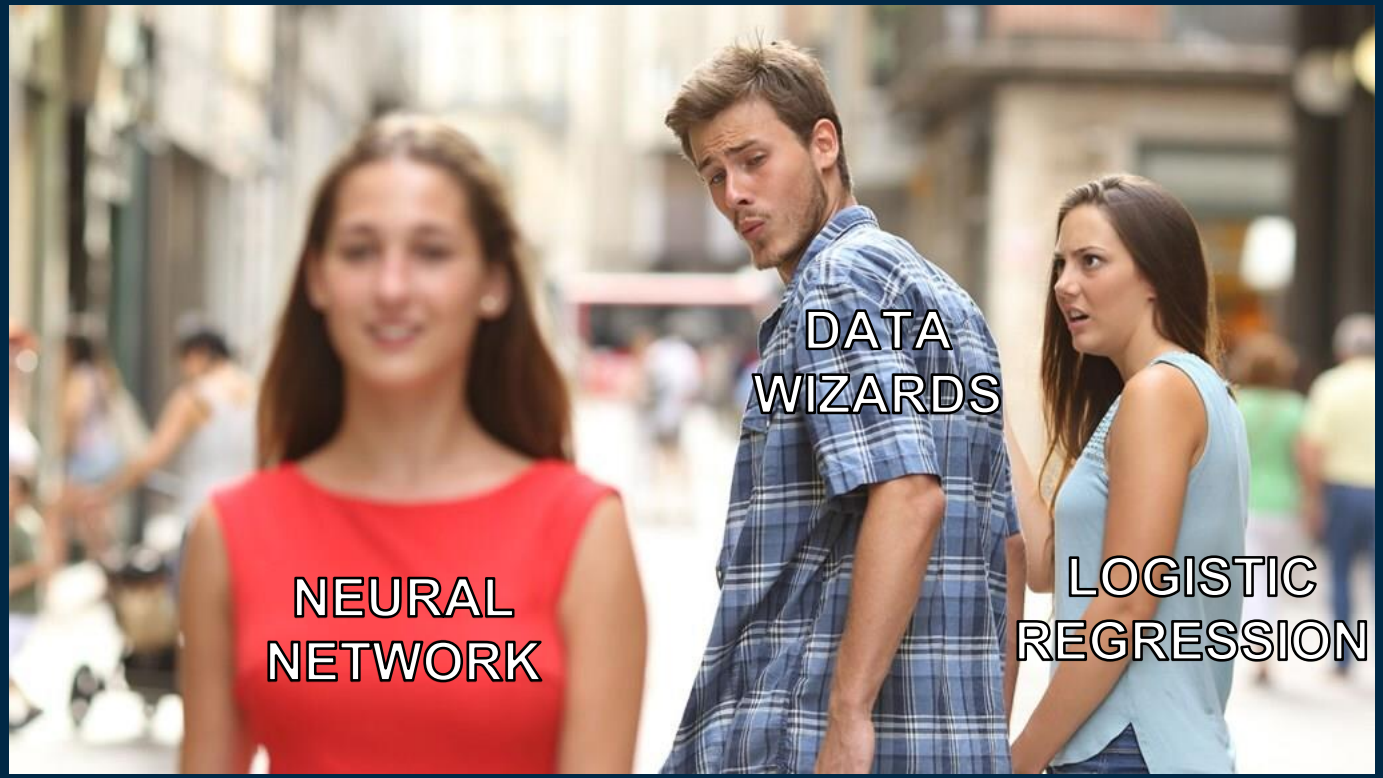
	Random Forest	Logistic Regression	KNN	Adaboost	Neural network
F1 Score	0.672	0.679	0.660	0.681	**0.693**
AUC-ROC	0.749	0.746	0.716	**0.760**	0.751
AUC-PR	0.655	0.646	0.610	**0.681**	0.642
Accuracy	0.699	**0.700**	0.670	0.699	0.696

ROC Curve


ROC Curve on test data
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest (AUC = 0.749)
Logistic Regression (AUC = 0.746)
KNN (AUC = 0.716)
Adaboost (AUC = 0.760)
Neural network (AUC = 0.751)
Random

Selection of The Best Model

- Random Forest
  - Metric scores are lower compared to others, computationally inefficient
- KNN
  - Overall lowest performance
- AdaBoost
  - Highest AUC scores but does not provide default probabilistic output
- Neural Network
  - Highest F1 Score but no interpretability
- Logistical Regression
  - Highest Accuracy, Highest Robustness between train and test data, Model Simplicity


DATA
WIZARDS
NEURAL
NETWORK
LOGISTIC
REGRESSION

Evaluating the Results

- Run the logistic regression model with best hyperparameters
- Output is probability of success (investment) for each client
- Based on the outputs 3,000 clients were chosen
- Our data-driven solution predicted 71.4% of the investments
- Comparison of data-driven solution and random choice
- If each client invests on average 5,000 CZK
  - Based on our model: total amount invested by clients would be 6,890,000 CZK
  - Based on random choice: total amount invested by clients would be 3,595,000 CZK

3,295,000 CZK

Potential increase in the clients
investments with data-driven solution

THANK YOU FOR YOUR ATTENTION

Do you have any questions?