Faculty of Electrical Engineering and Informatics

Predictive Analytics for Default of Credit Card Clients

Alžbeta Bačová, František Babič

Department of Cybernetics and Artificial Intelligence

Faculty of Electrical Engineering and Informatics, Technical University of Košice

IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI 2021)

Outline

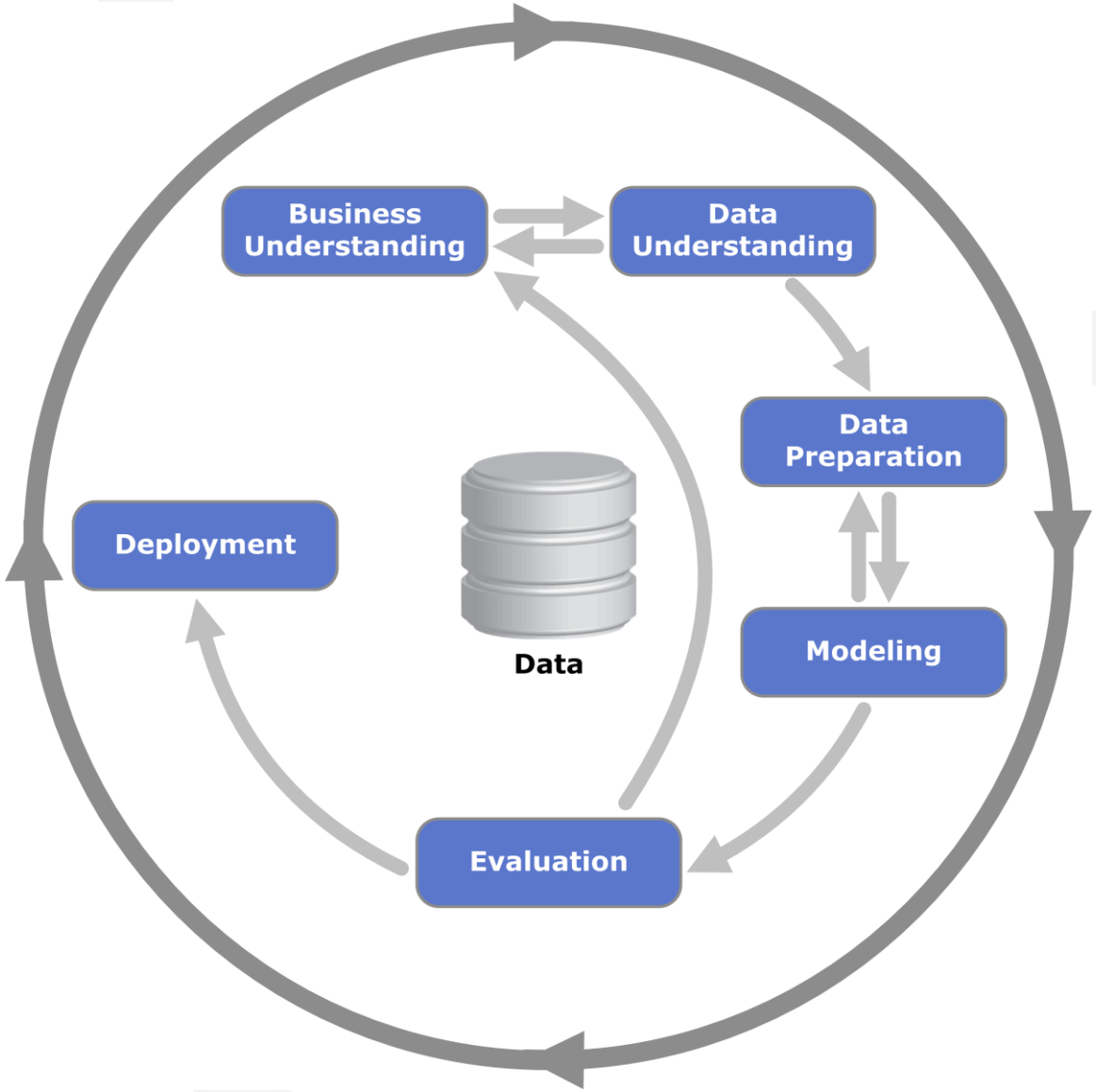
- Introduction
- CRISP-DM
  - Business understanding
  - Data understanding
  - Modelling
  - Evaluation
- Conclusion

Introduction

- The banking industry produces a significant amount of data every day that holds valuable information.
- The analysts can apply predictive methods to bank transaction data, clients ‘data, credit card history, customer experience, and stock market data.
- The Accenture study proposes a scenario that Artificial intelligence will transform financial service providers into data- and AI-based businesses.
- Artificial intelligence and machine learning offer a wide range of methods and algorithms. It is necessary to test, verify and select the most suitable ones based on the type of task.

CRISP-DM

- Cross-industry process for data mining as a process methodology for successful data analysis, modeling, and knowledge discovery.

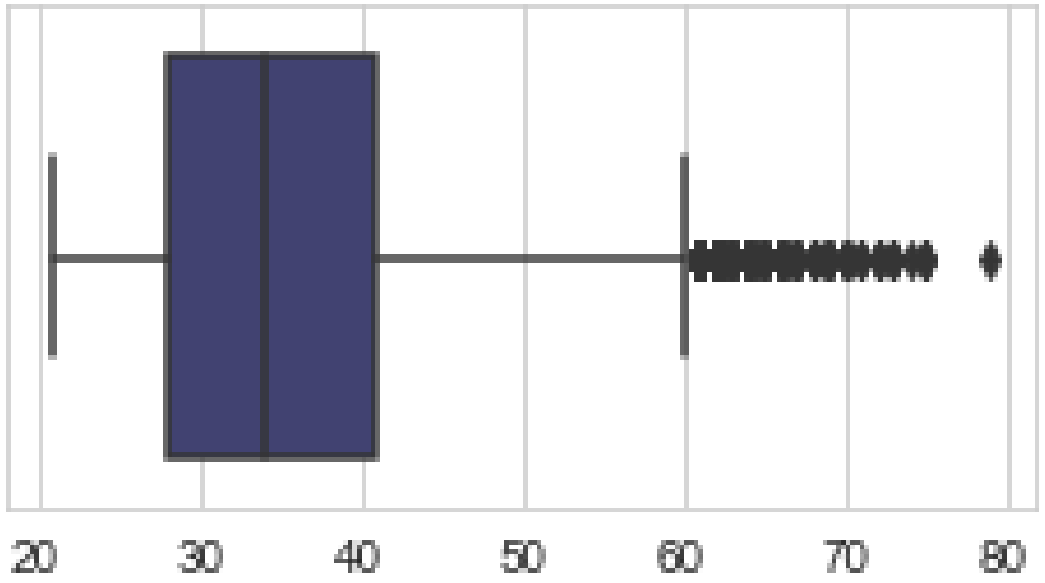


Business and Data understanding

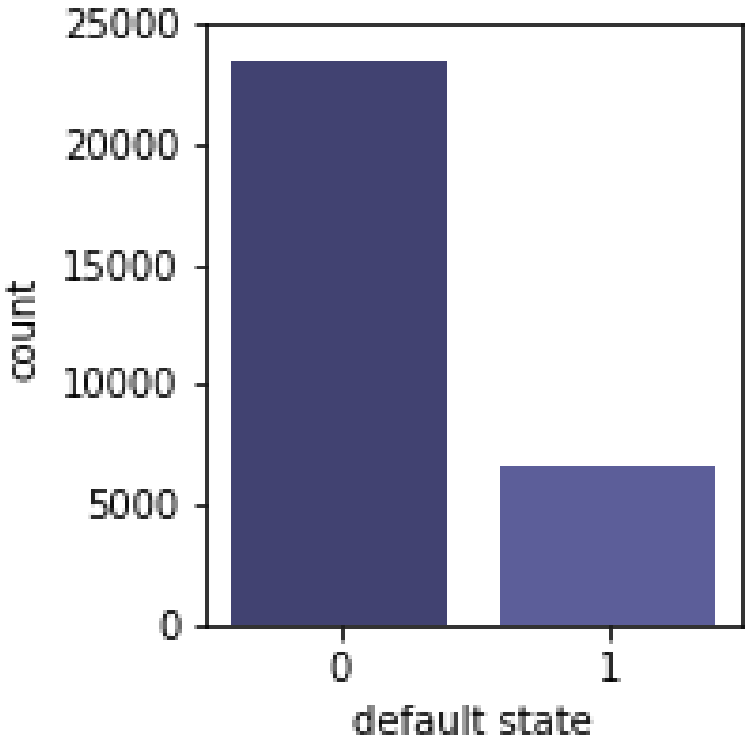
- The default risk of bank customers and the decision-making process => a classification task.
- Public data of credit card holders from Taiwan from April to September 2005.
- 30 thousand records, 25 attributes.
- Target attribute (binary) = defaulter or non-defaulter.
- Examples:
  - EDUCATION: (0 = unknown, 1 = graduate school, 2 = university, 3 = high school, 4 = others, 5 = unknown, 6 = unknown).
  - MARRIAGE: marital status of individual (0 = unknown, 1 = married, 2 = single, 3 = others).

Data understanding

- The age structure



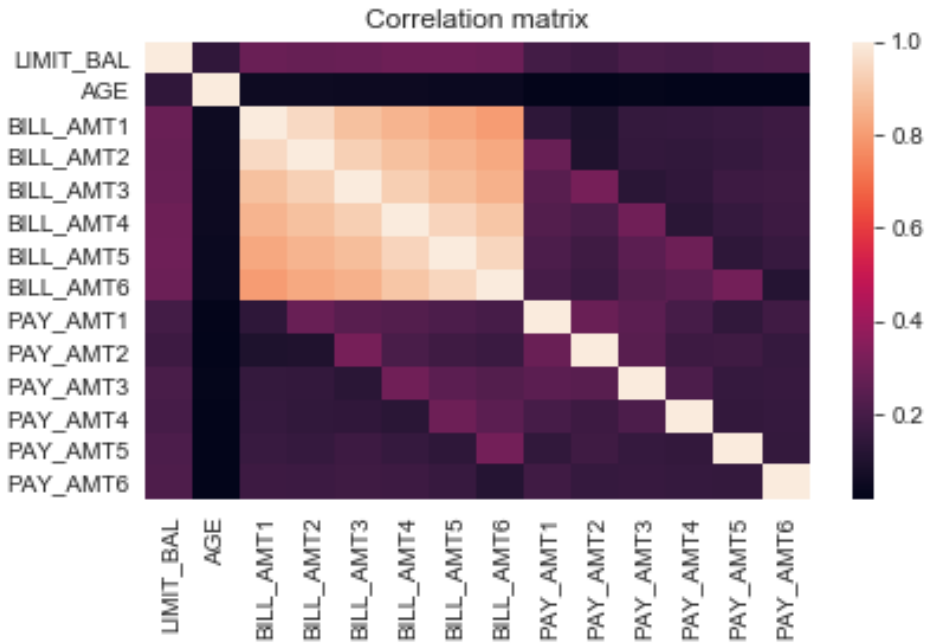
- The target attribute histogram



Faculty of Electrical Engineering and Informatics

Data preparation

- Correlation matrix

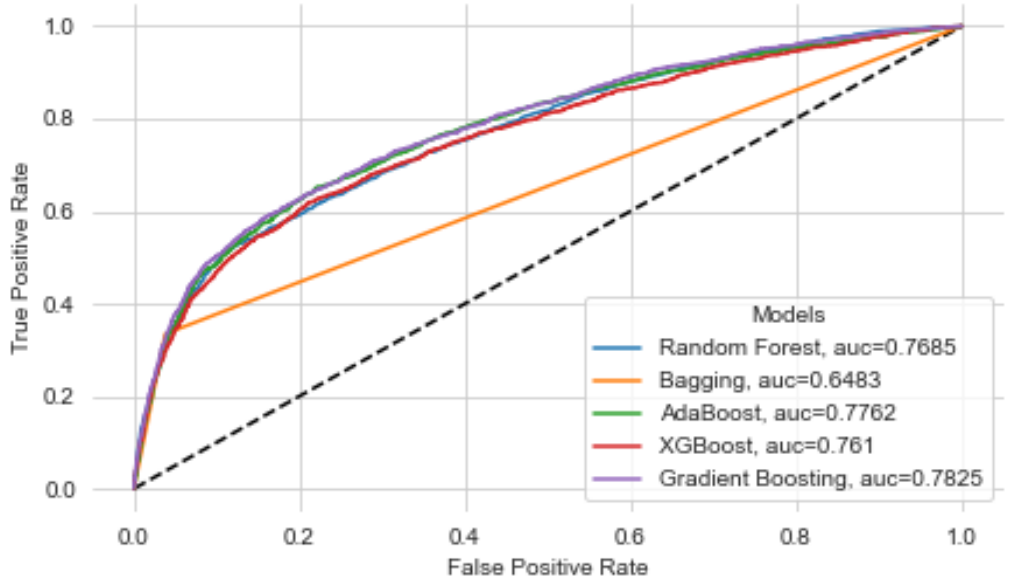


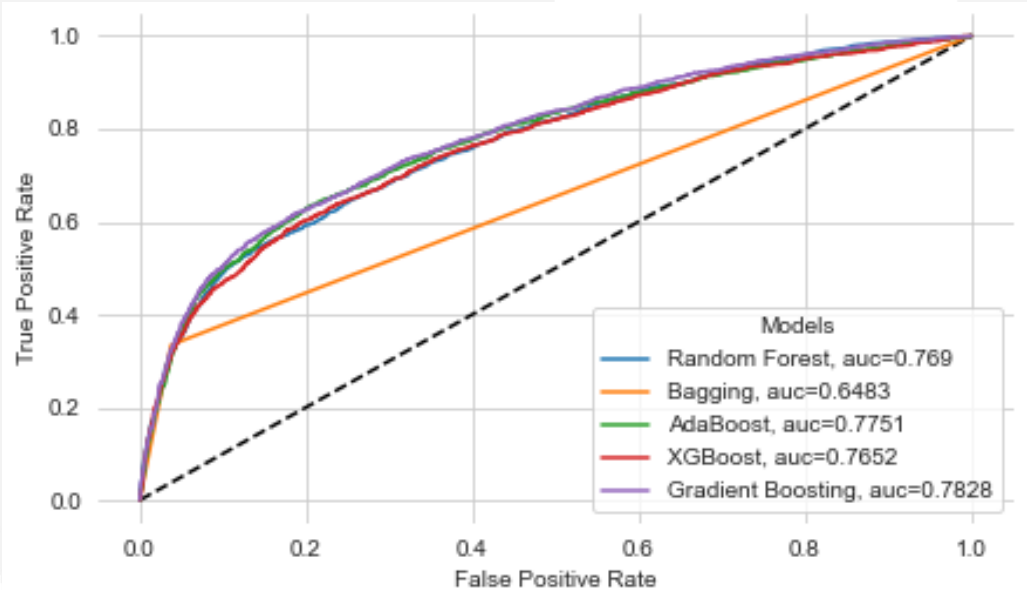
- Original data without any preprocessing => baseline model.
- Preprocessed data <= inconsistencies removing, data standardisation.
- Data division => stratified splitting 70% for training and 30% for testing; 80:20 and 60:40.

Faculty of Electrical Engineering and Informatics

Modelling and evaluation

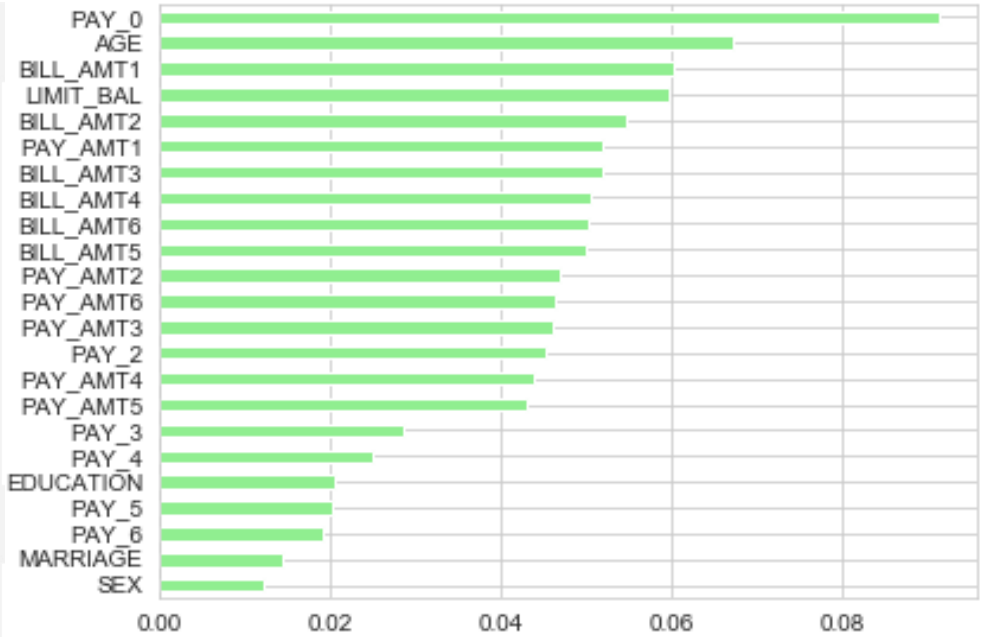
- Random Forest, AdaBoost, XGBoost, and Gradient Boosting algorithm.
- Accuracy, precision, recall, ROC, and AUC.





Original / preprocessed dataset

Feature importance:



Conclusion

Matthews correlation coefficient:

- The original data => Gradient Boosting (0.4111), Bagging (0.4044), and Random forest (0.3918).
- Minimal or no improvement on preprocessed data.

- According to the precision score and ROC, the best algorithms are AdaBoost and Gradient Boosting (defaulter).
- We can state that our study's algorithms are valuable for identifying clients' default state and producing a good performance.

Faculty of Electrical Engineering and Informatics

Faculty of Electrical Engineering and Informatics

Thank you for your attention ☺