BANKING CASE STUDY

ALEXANDRE SARAIVA MOREIRA
JOÃO LUÍS SOARES ISAÍAS

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

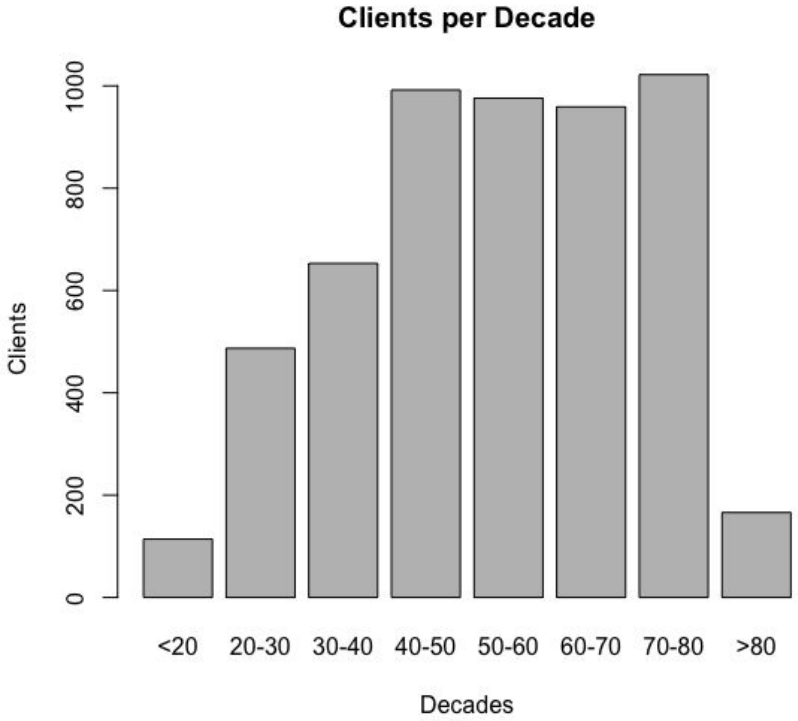
1.
DOMAIN DESCRIPTION

DOMAIN DESCRIPTION

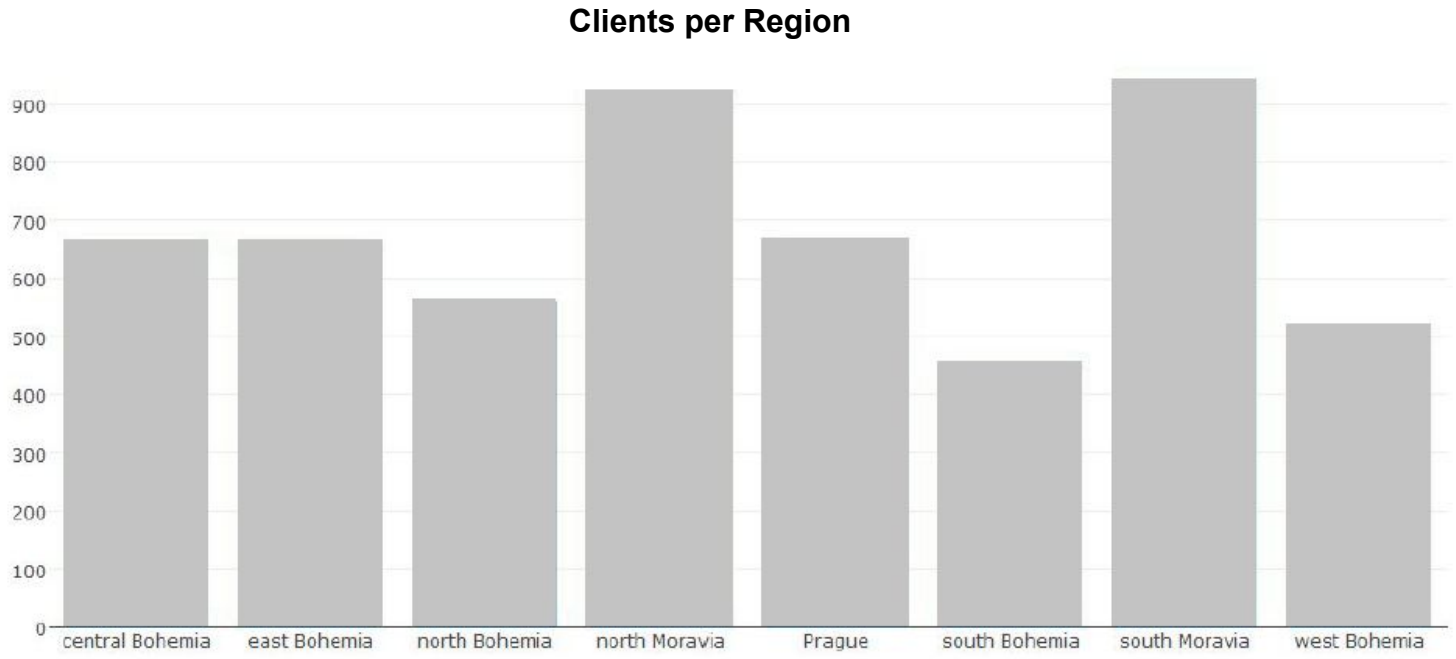
- Develop a data mining case study (banking loans) with the given dataset
- Descriptive task: describe the clients' profile
- Predictive task: predict if a loan will be paid back to the bank or not
- Tools used: RStudio (for preparing the dataset) and RapidMiner (to perform the main descriptive and predictive tasks)

2. EXPLORATORY DATA ANALYSIS

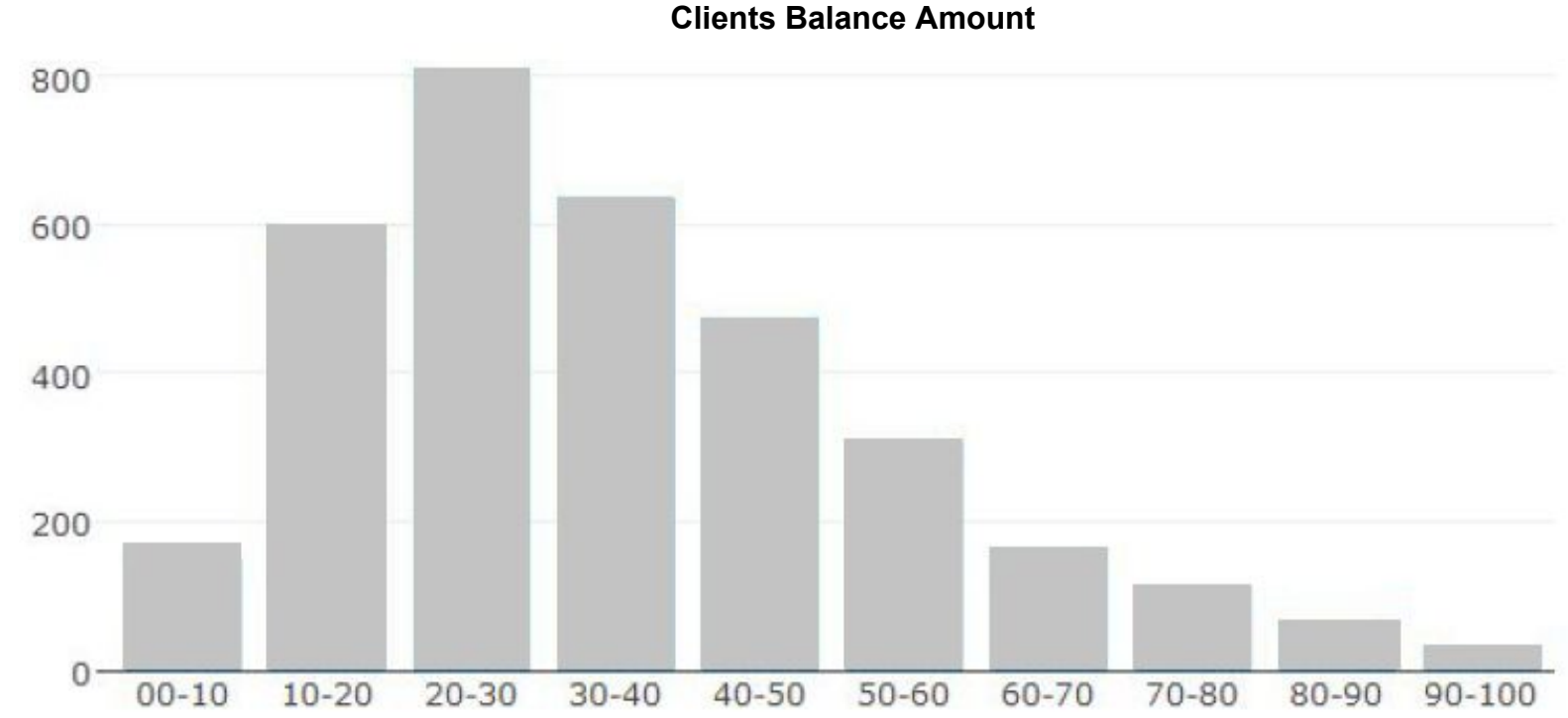
EXPLORATORY DATA ANALYSIS



EXPLORATORY DATA ANALYSIS



EXPLORATORY DATA ANALYSIS



3.
DESCRIPTIVE TASK

PROBLEM DEFINITION

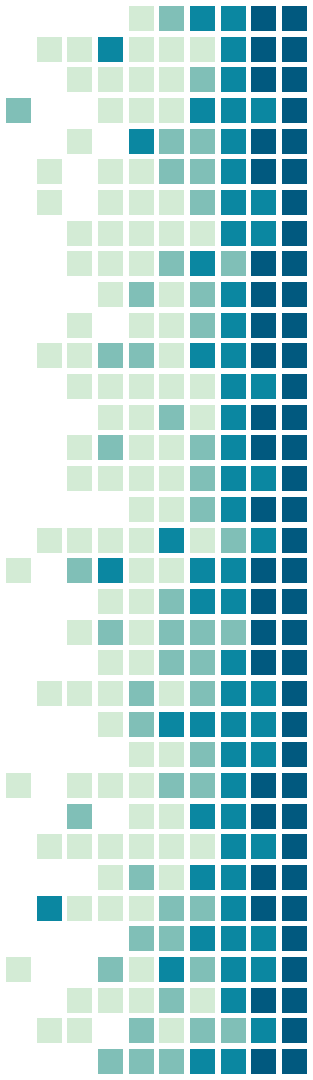
The analysed information varied from personal details, such as gender, age and district, to information from their accounts, based on the transactions they made, such as average balance and number of transactions.

The main focus of the descriptive problem was to describe the relation between the clients' balance and the clients' age.

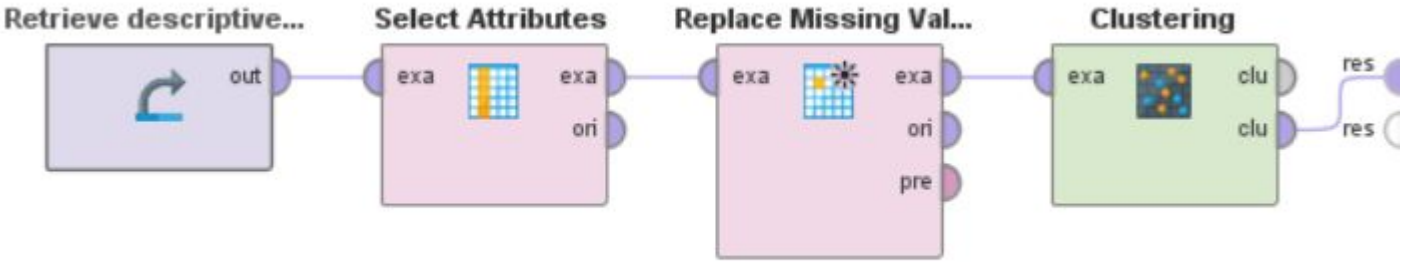
The available data was filtered and aggregated to contain only relevant data for this analysis.

DATA PREPARATION

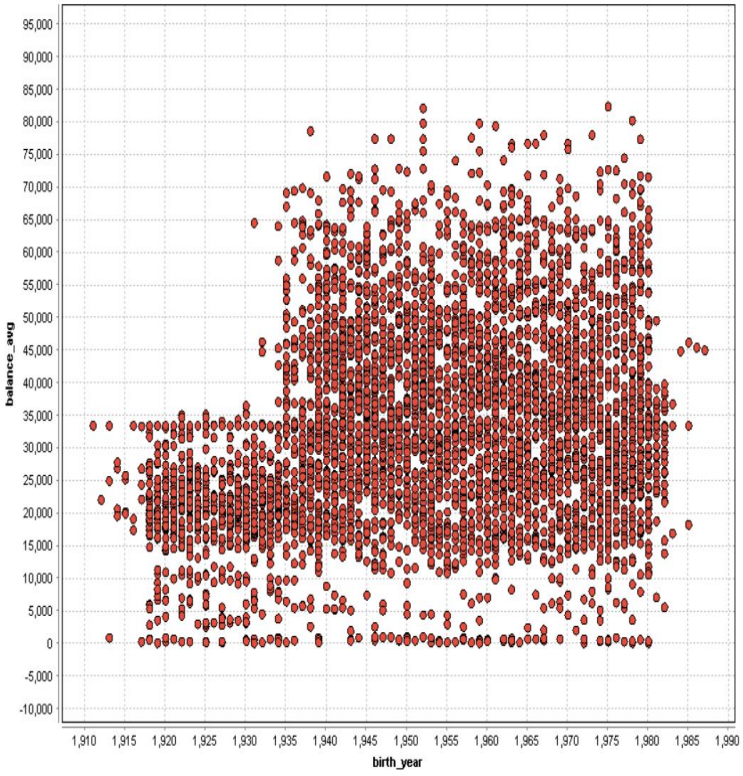
- Filter clients by gender
- Fix women birth dates
- Get clients' birth year
- Get clients' balance average and standard deviation

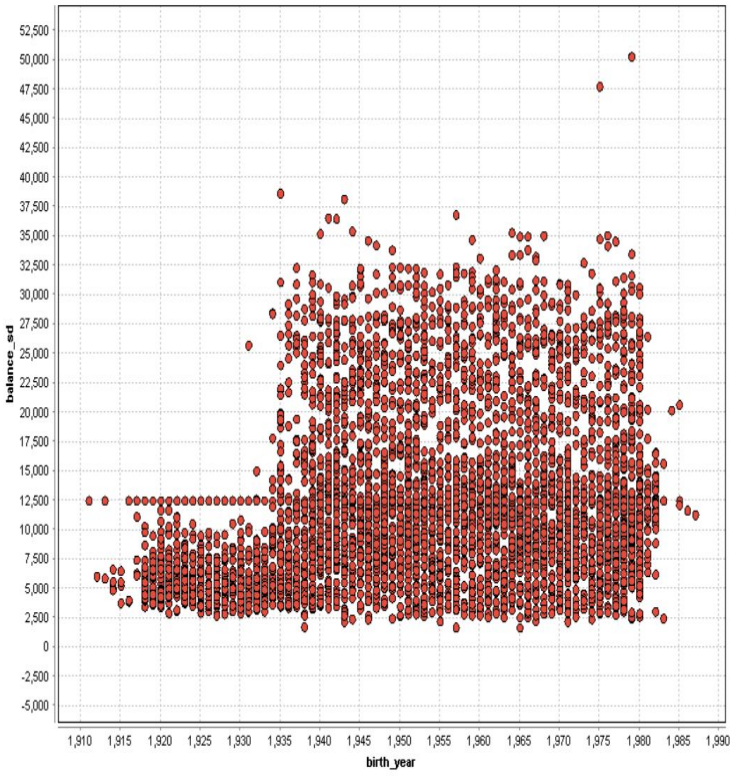


EXPERIMENTAL SETUP



RESULTS


balance_avg
95,000
90,000
85,000
80,000
75,000
70,000
65,000
60,000
55,000
50,000
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
-5,000
-10,000
1,910 1,915 1,920 1,925 1,930 1,935 1,940 1,945 1,950 1,955 1,960 1,965 1,970 1,975 1,980 1,985 1,990
birth_year


balance_sd
52,500
50,000
47,500
45,000
42,500
40,000
37,500
35,000
32,500
30,000
27,500
25,000
22,500
20,000
17,500
15,000
12,500
10,000
7,500
5,000
2,500
0
-2,500
-5,000
1,910 1,915 1,920 1,925 1,930 1,935 1,940 1,945 1,950 1,955 1,960 1,965 1,970 1,975 1,980 1,985 1,990
birth_year

4.
PREDICTIVE TASK

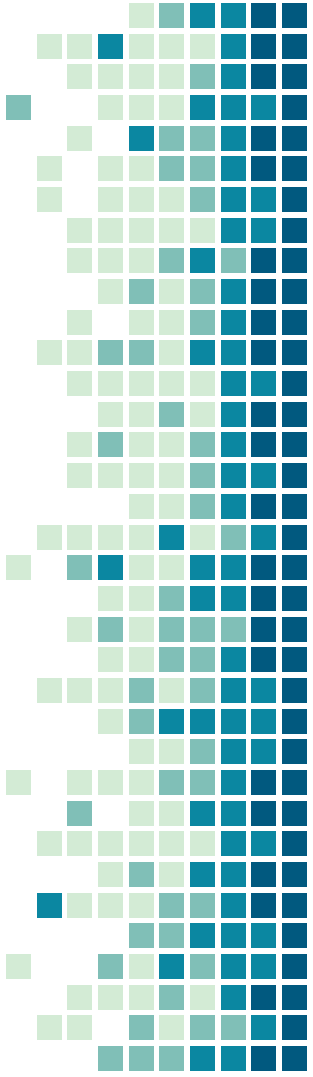
PROBLEM DEFINITION

The goal of this task is to predict if a loan will be paid back to the bank or not.

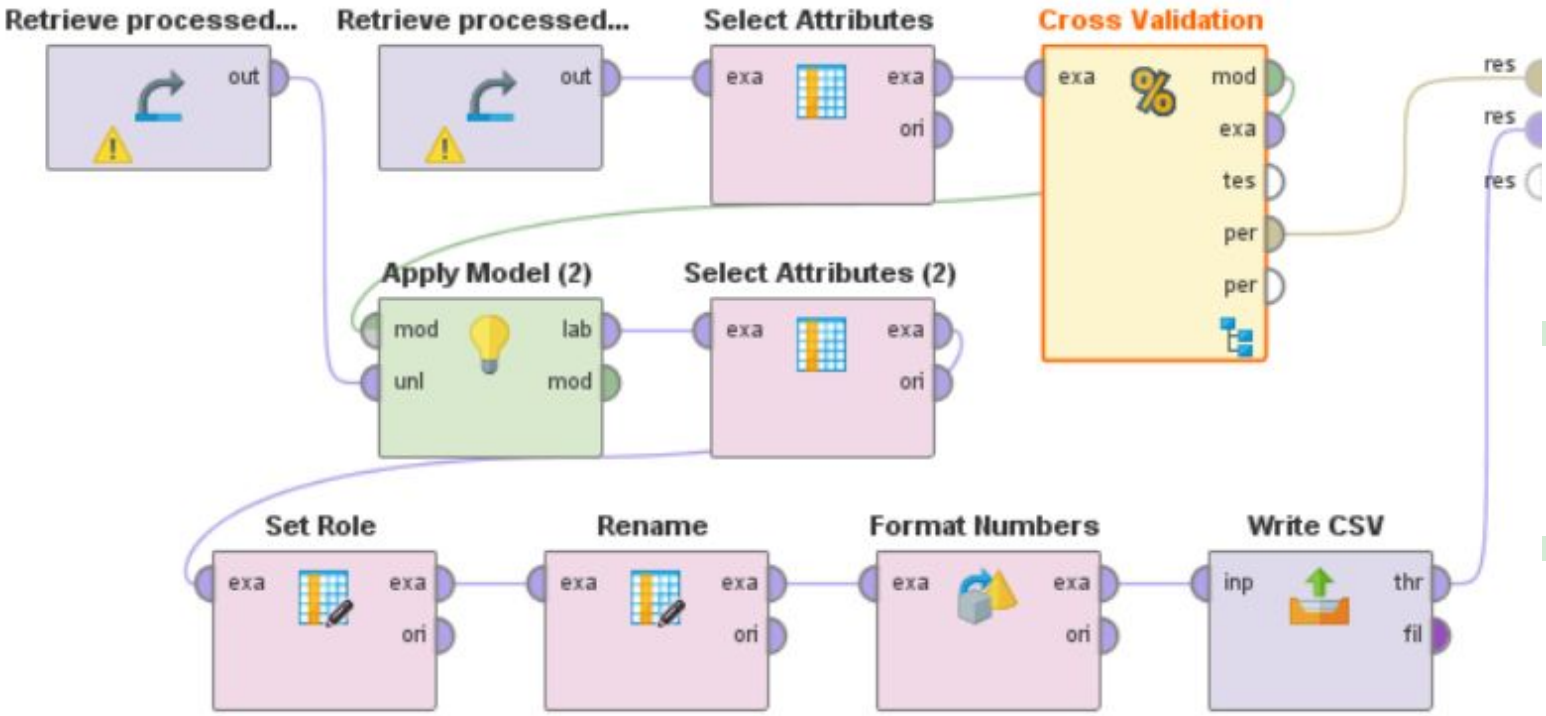
To achieve the goal, a model was trained using the data available. In order to get the maximum information from the data, some adaptations were made.

DATA PREPARATION

- Transform strings to integers
- Set missing values
- Get clients' gender through birth date, and normalize dates
- Filter to only show owners information
- Get clients' balance average and standard deviation
- Transform dates to date type



EXPERIMENTAL SETUP



RESULTS

- To test the performance of the model in RapidMiner, the Accuracy and AUC were chosen as metrics
- The best values the model got was: 89% Accuracy and 86% AUC
- The final score from Kaggle (before the end of the competition) was 85%

5. CONCLUSIONS, LIMITATIONS AND FUTURE WORK

CONCLUSIONS, LIMITATIONS AND FUTURE WORK

- Filtering the data has a huge impact in how well a model can be trained
- Some operators could be beneficial, such as optimize features or parameters
- Despite having an overall predictive score of 85%, we feel this could be improved by using other methods for training, which perhaps were not explored in the time given

ANNEXES

Deep Learning Parameters

- Activation: Rectifier
- Hidden layer sizes: 50 50
- Epochs: 20
- Train samples per iteration: -2
- Adaptive rate
- Epsilon: 1.0E-8
- Rho: 0.99
- Standardize
- (defaults)

THANKS!

Any questions?