Banking Data Mining Case Study

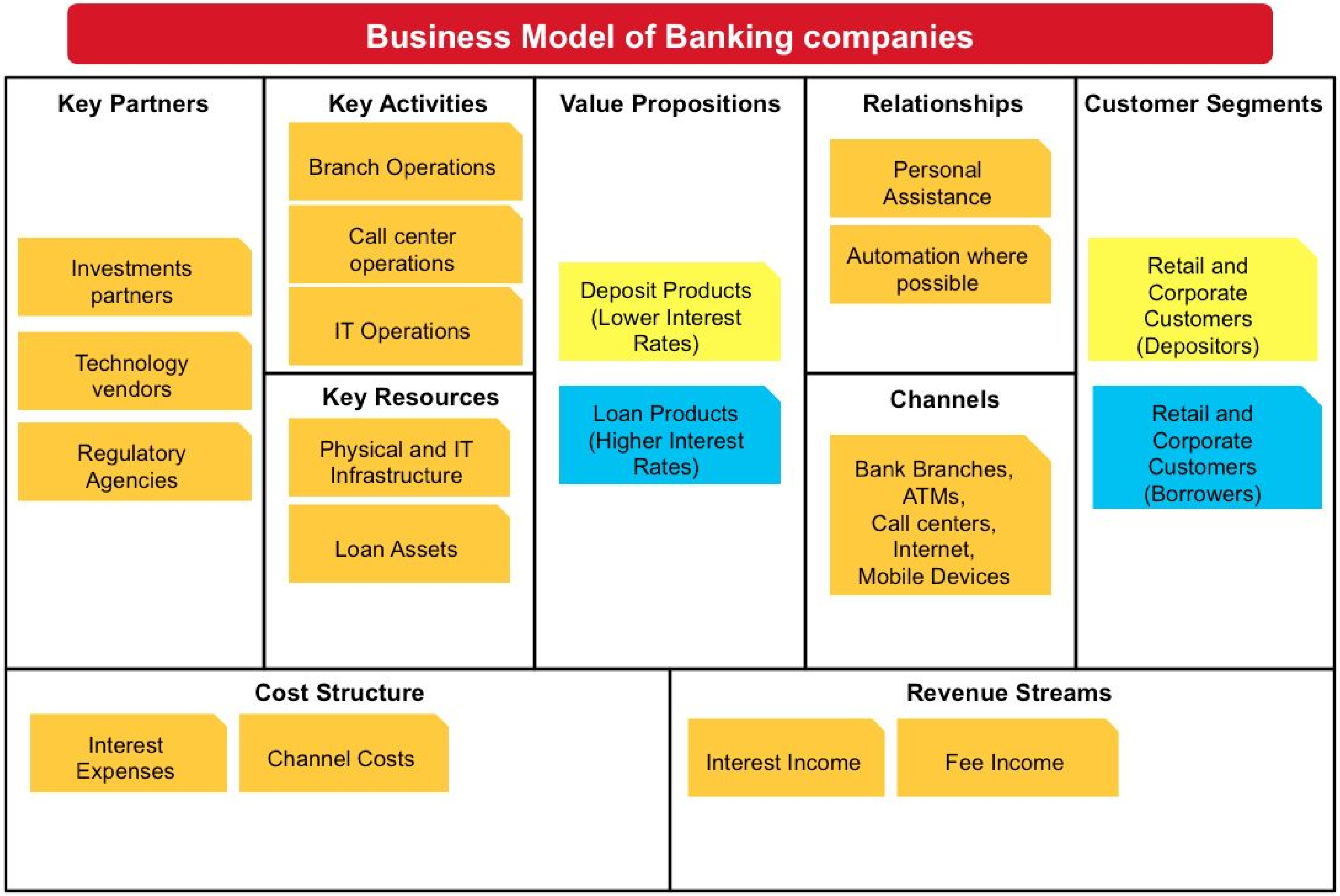
Knowledge Extraction and Machine Learning
2015/2016

Group 4
Ruben Fernando Pinto Cordeiro ei11097@fe.up.pt
Duarte Nuno Pereira Duarte ei11101@fe.up.pt

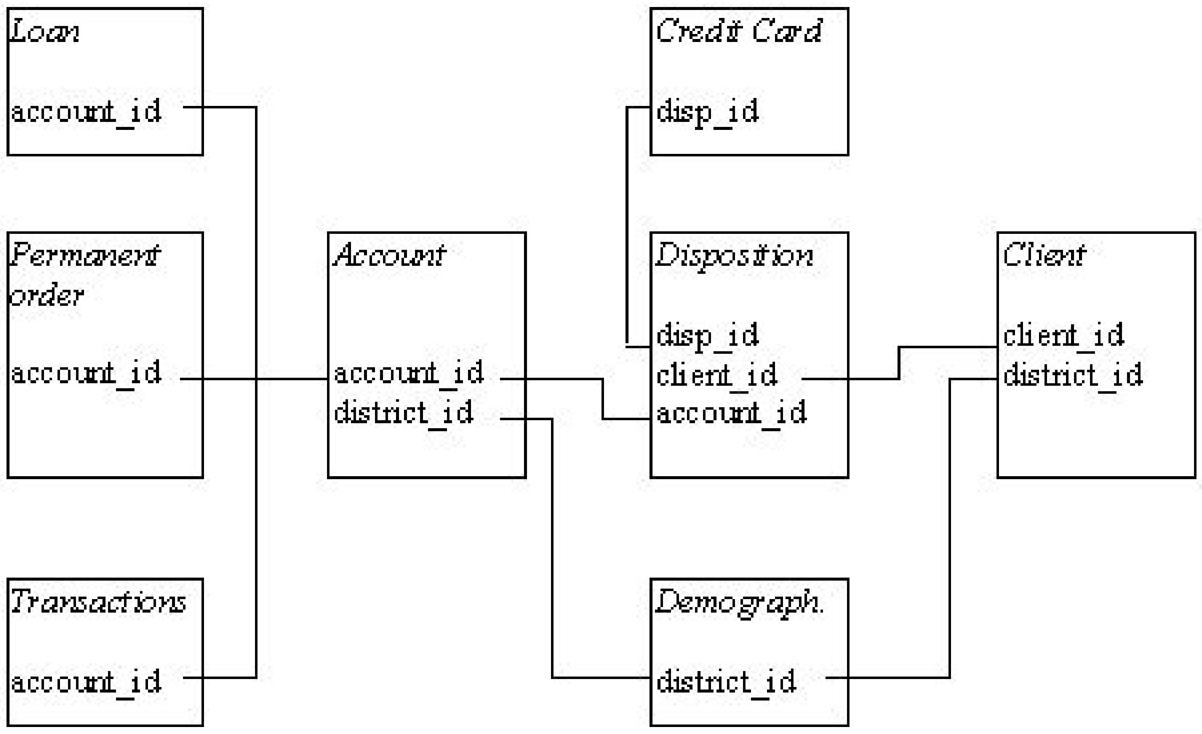
Outline

Methodology

Understanding the business



Understanding the data



Data preparation

- Data cleaning
  - Add labels where missing (district, ...);
  - Extract gender from birth date in the Client relation;
  - Format dates and numbers;
- Data transformation:
  - Calculate age in the Client relation;
  - Import raw data into RapidMiner;
  - Denormalize various relations for the descriptive and predictive analysis algorithms.

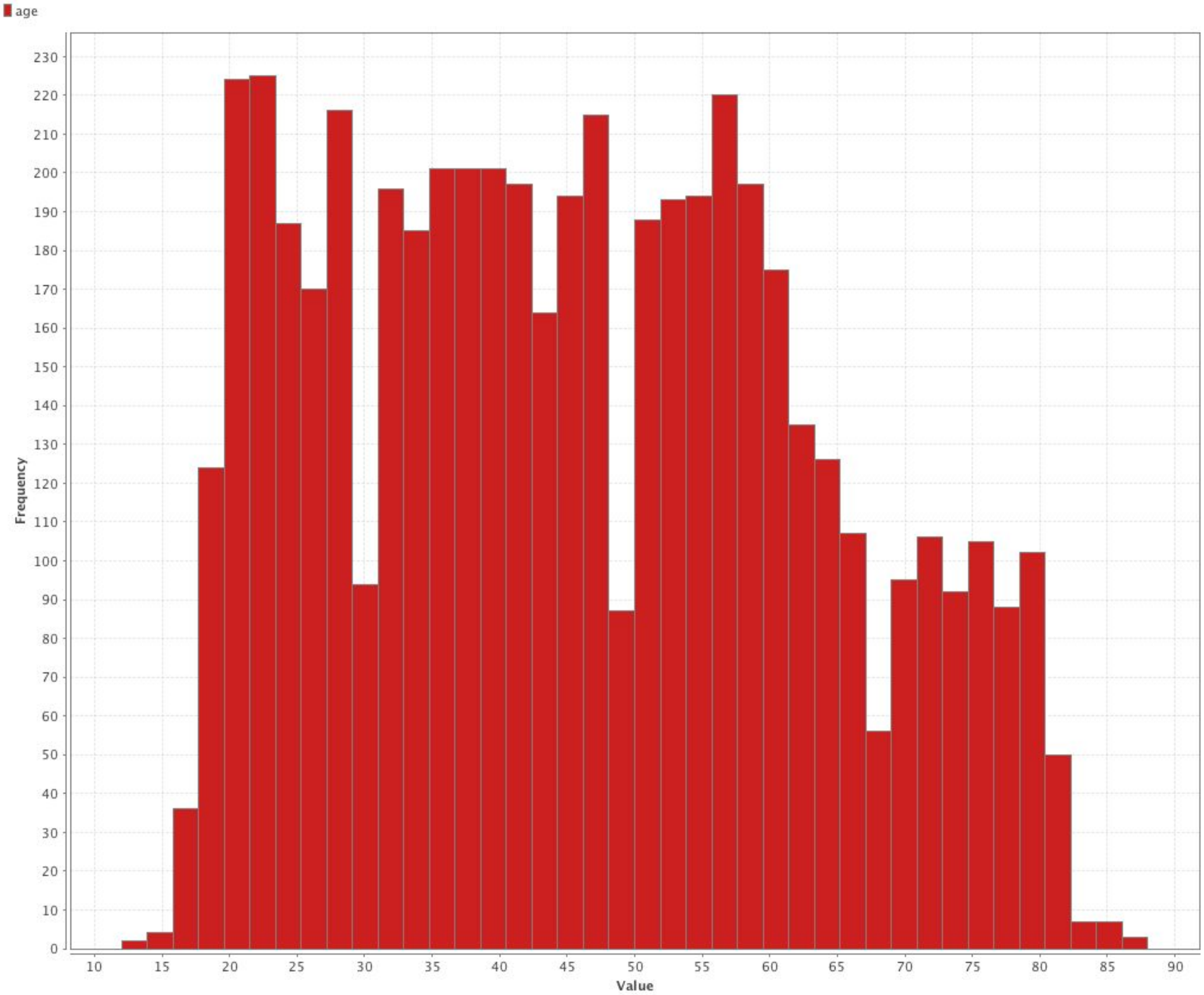
DESCRIPTIVE DATA ANALYSIS

Descriptive data analysis

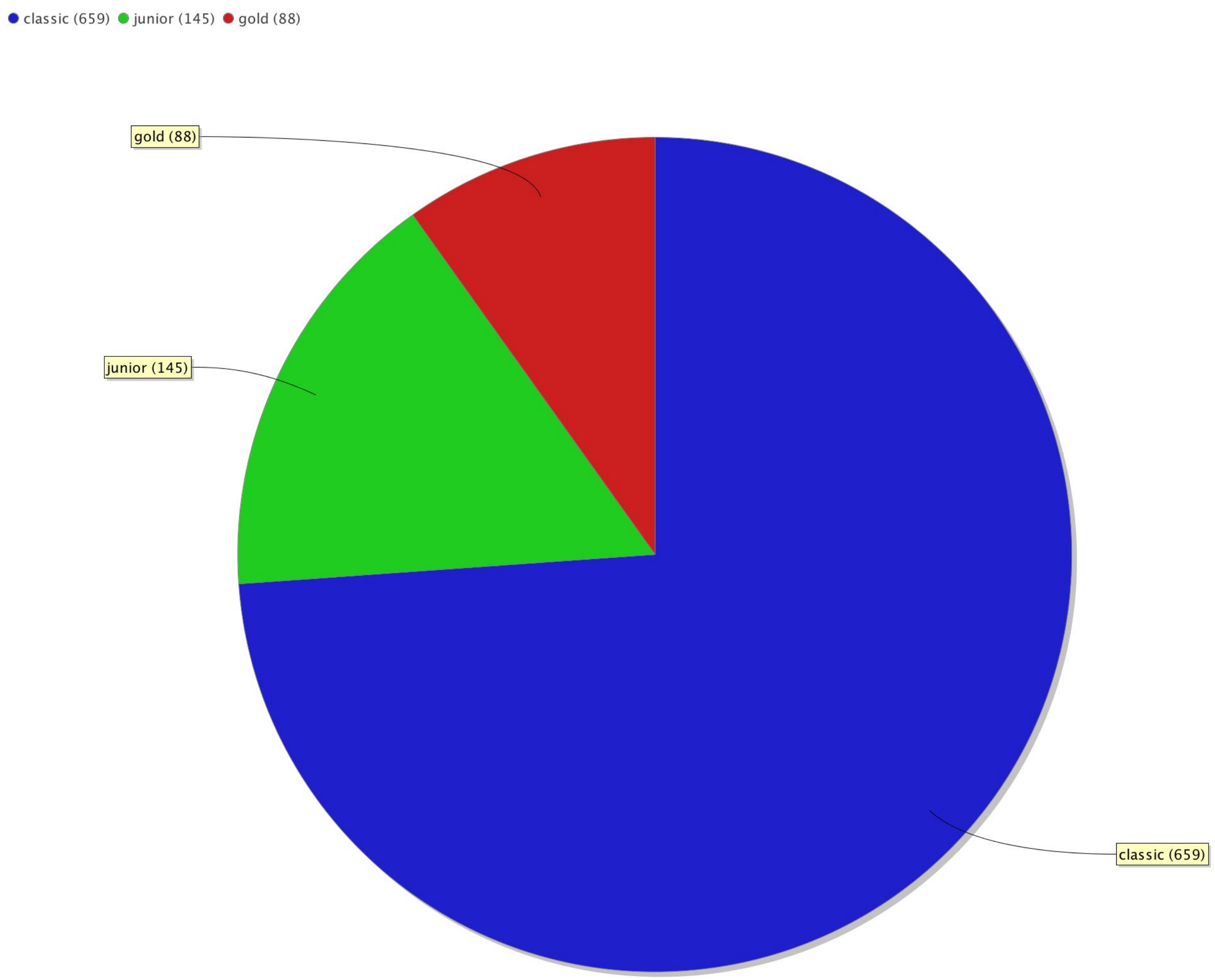
Segmentation

- Primary parameters:
  - Assets & Liabilities
  - Geo-demographic data
  - Profitability
- Secondary parameters:
  - Behavioristic segmentation
  - Life stage segmentation

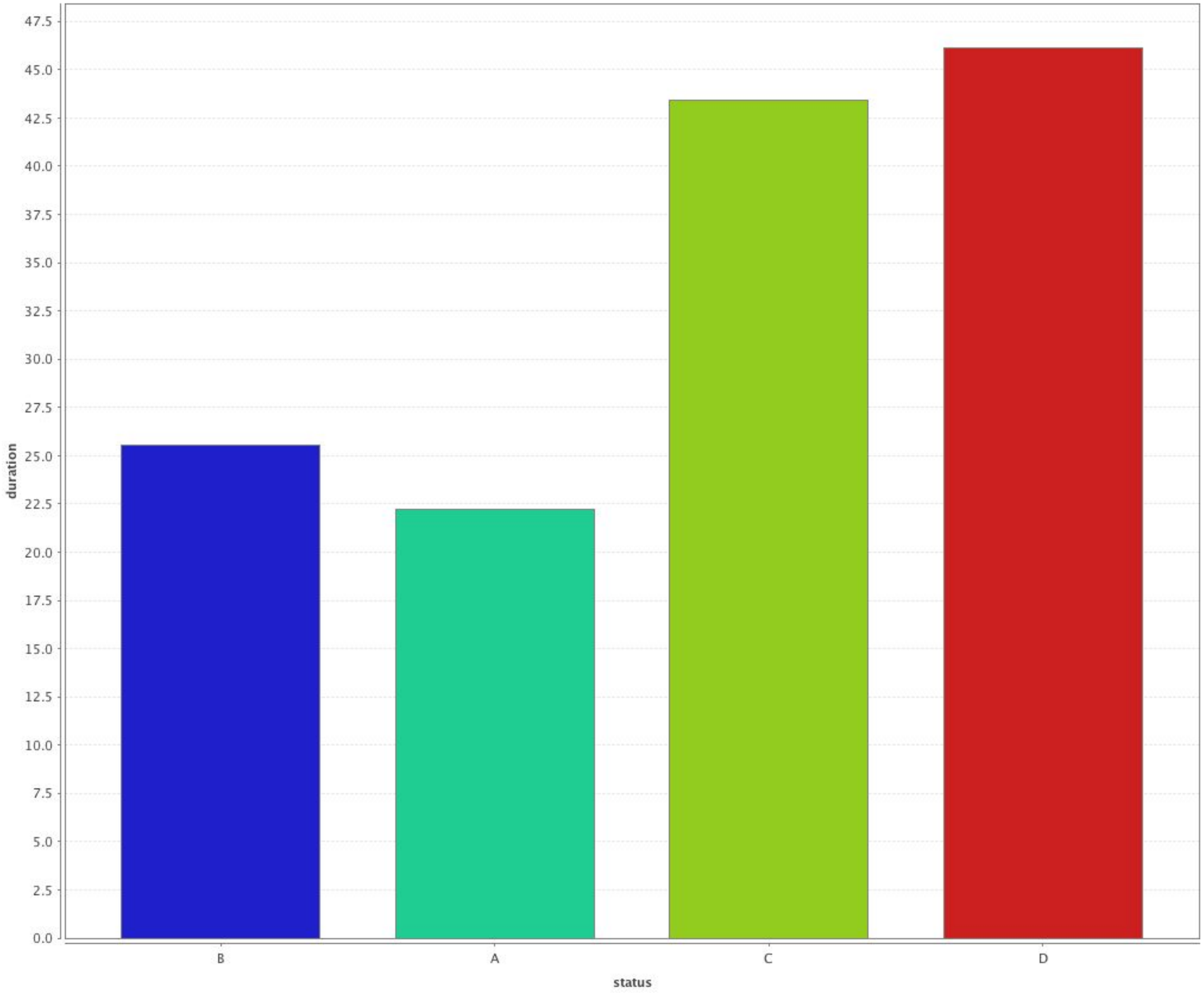
Age distribution



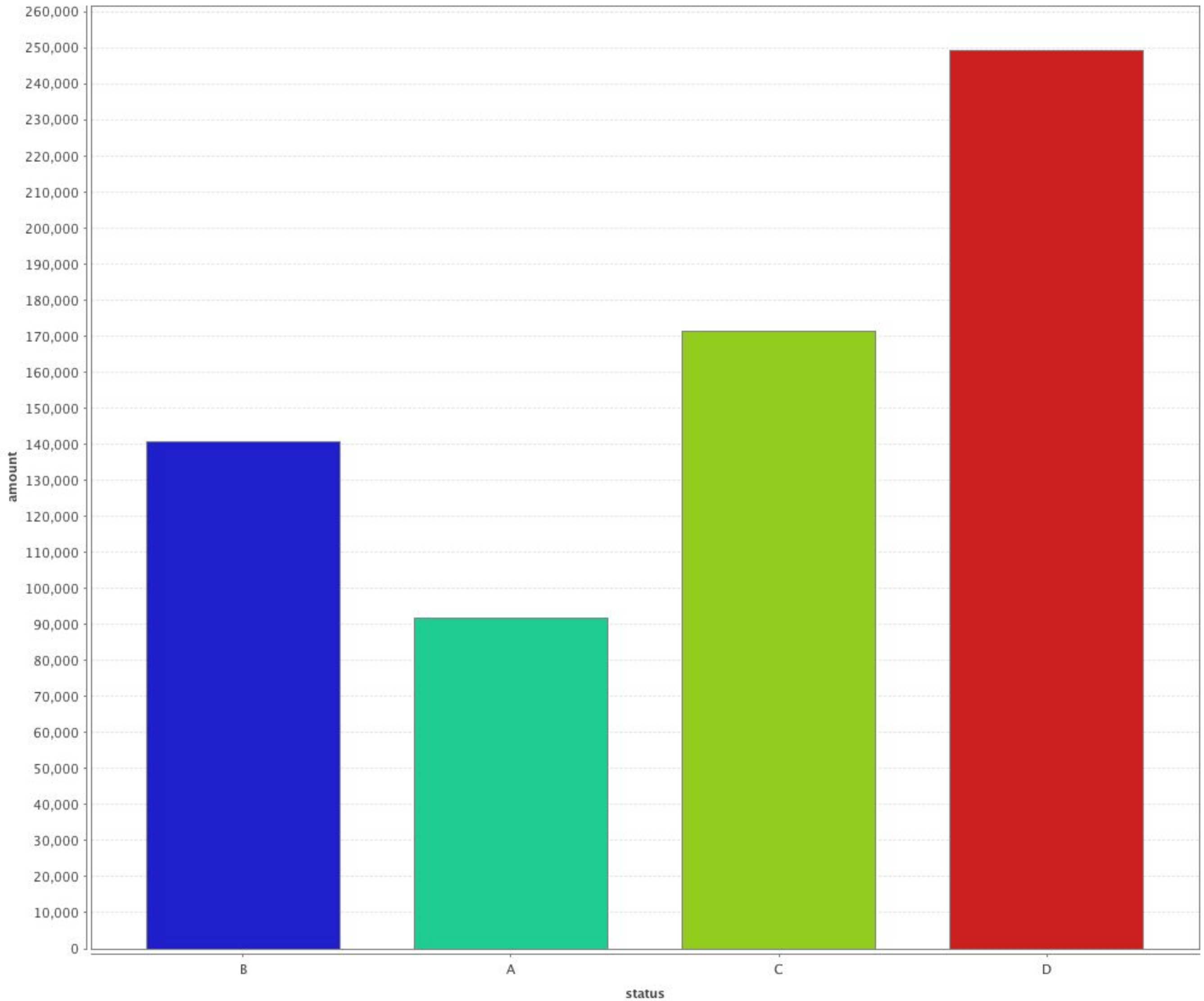
Credit card type distribution



Average loan duration distribution by status



Average loan amount distribution by status



Exploratory data analysis: conclusions

- Most of the issued credit cards are *classic* (649). There are more *junior* cards (145) than *gold* cards (88).
- *Good* loans (A, B) tend to be of a lower amount and a shorter duration when compared to *bad* loans (C, D).

Clustering: Case 1

- For the first clustering, the following attributes were pre-computed:
  - Amount: the total loan amount borrowed by each client;
  - Duration: the total loan duration by each client;
  - Payments: the monthly amount of loan payments made by each client;
  - Average Income: the monthly average income of each client;
  - Average Withdrawals: the monthly average withdrawals made by each client.

Clustering: Case 1

Attribute	Cluster 1	Cluster 2	Cluster 3
Amount	62396,58	152970,91	346722,23
Duration	26,26	36,29	53,54
Payments	3077,07	4251,54	6579,91
Average Income	26038,91	16010,17	33134,65
Average Widthrawals	23624,16	14759,80	29875,93

Clustering: Case 1

- With this clustering process, the goal was to segment clients according to their assets and liabilities.
- The clients in the Cluster 1 have less purchase power when compared to clients belonging to Cluster 2 and 3.

Clustering: Case 2

Attribute	Cluster 1	Cluster 2	Cluster 3
No. of entrepreneurs per 1000 inhabitants	160,10	112,05	115,43
No. of municipalities with inhabitants 2000-9999	0,00	7,39	5,64
No. of municipalities with inhabitants 500-1999	0,00	23,17	26,41
No. of municipalities with inhabitants < 499	0,00	21,52	74,67
No. of municipalities with inhabitants >10000	1,00	2,24	1,42
Ratio of urban inhabitants	100,00	70,42	55,75

Clustering: Case 2

- With this clustering process, the goal was to segment clients according to geo-demographic data.
- Clients in Cluster 1 belong to an urban environment with an higher ratio of entrepreneurs.
- Clients in Cluster 3 live in a district with the smallest ratio of urban inhabitants.

PREDICTIVE DATA ANALYSIS

- The loans dataset is relatively small: 682 loans in total.
- The status distribution of the loans is uneven: 606 loans are *good*, while only 76 loans are *bad*.
- These factors make the predictive task more difficult.

Decision Tree

- Given the small dataset, there is a great danger of overfitting the data.
- In order to mitigate the risk, the decision tree was pruned pre-pruned, thus reducing the complexity of the final classifier.

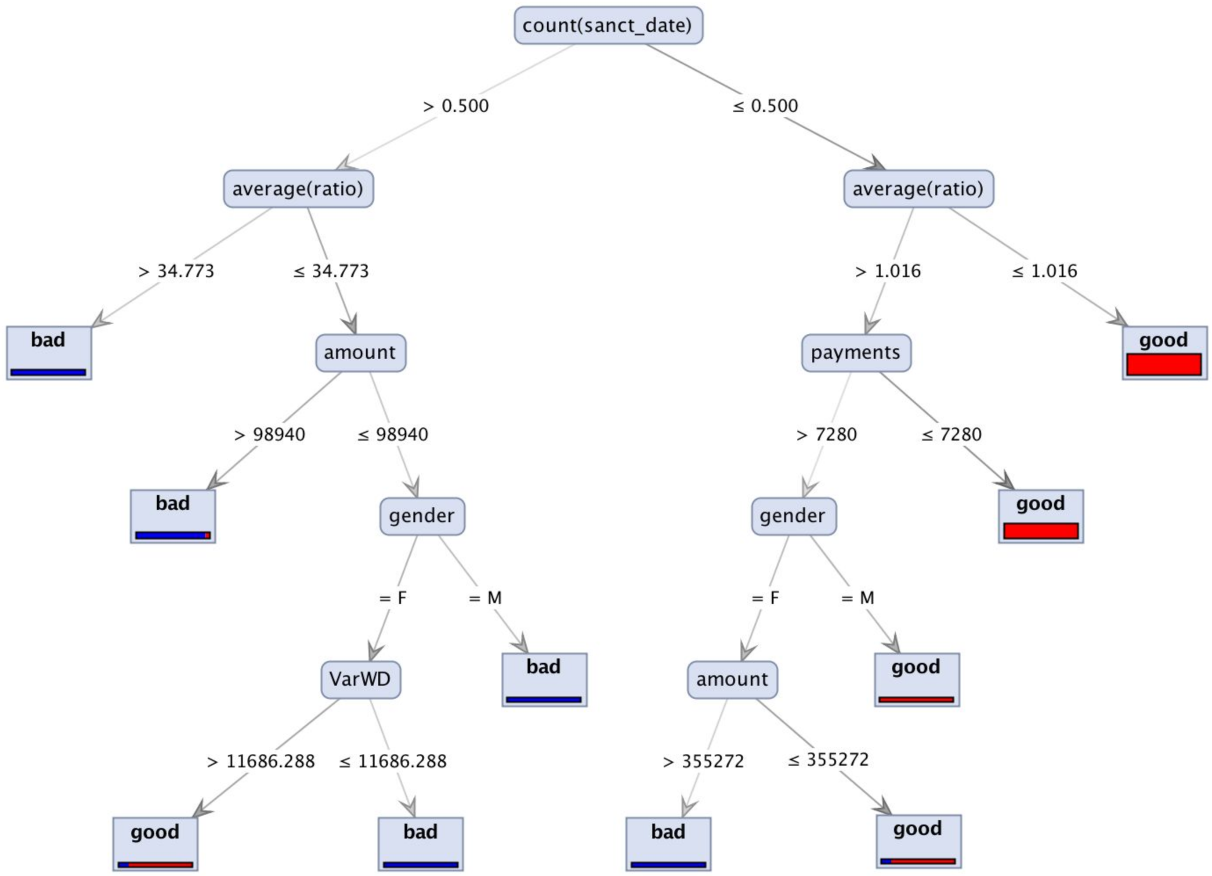
Decision Tree

- In RapidMiner:
  - The decision tree operator was used:
    - Criterion: *information_gain;*
    - Level of confidence set to 25%;
    - Applying pruning and pre-pruning;
    - Minimal gain set to 4%;
    - Maximal depth set to 6.

Decision Tree model

- The following relations were joined: Client, Account, Loan, Disposition and Transactions.
- For this classification, several attributes were pre-computed:
  - Average(ratio): Average debt-to-income ratio of the client.
  - Count(sanction_date): Number of sanctions that the client is responsible for. A sanction occurs when the client's bank account balance reaches negative values.
  - Amount: the total loan amount borrowed by the client.
  - Payments: the monthly amount of loan payments made by the client.
  - Gender: the client's gender.
  - VarWD: The monthly average variation of the withdrawals made by the client.

Decision Tree model



Decision Tree model: evaluation

- Using the Rapidminer ‘x-validation’ operator (cross validation):
  - Number of validations: 15.

accuracy: 97.21% +/- 1.71% (mikro: 97.21%)

	true bad	true good	class precision
pred. bad	65	8	89.04%
pred. good	11	598	98.19%
class recall	85.53%	98.68%	

K-Nearest Neighbor

- In Rapidminer:
  - The ‘K-NN’ operator was used:
    - K was set to 1;
    - Measure Types: MixedMeasures;
    - Mixed Measure: MixedEuclideanDistance.
- The following relations were joined: Client, Account, Loan, Disposition and Transactions.

K-Nearest Neighbor

- The model contains 682 examples with 11 dimensions of the following classes:
  - bad
  - good

K-Nearest Neighbor: evaluation

- Using the Rapidminer ‘x-validation’ operator (cross validation):
  - Number of validations: 15.

accuracy: 82.11% +/- 6.48% (mikro: 82.11%)

	true bad	true good	class precision
pred. bad	19	65	22.62%
pred. good	57	541	90.47%
class recall	25.00%	89.27%	

K-Nearest Neighbor: discussion

- This model performs worse than the Decision Tree model.
- The K-Nearest Neighbor generally performs badly with high-dimensional data (curse of dimensionality).

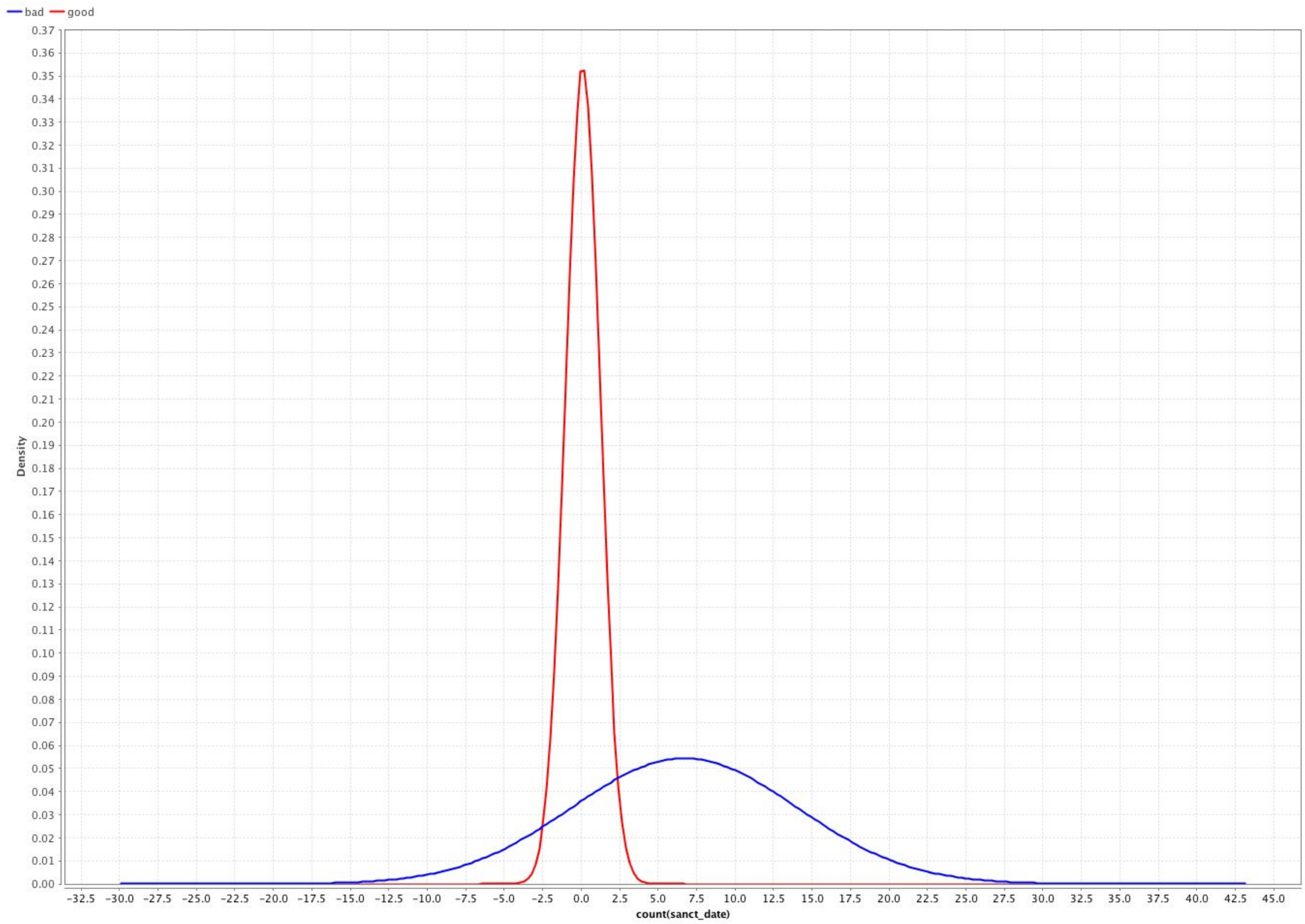
Naïve Bayes

- The following relations were joined: Client, Account, Loan, Disposition and Transactions.
- The same pre-computed attributes from the Decision Tree model were used.
- In RapidMiner:
  - The *Naïve Bayes* operator was used:
    - The laplace correction option was set to true.

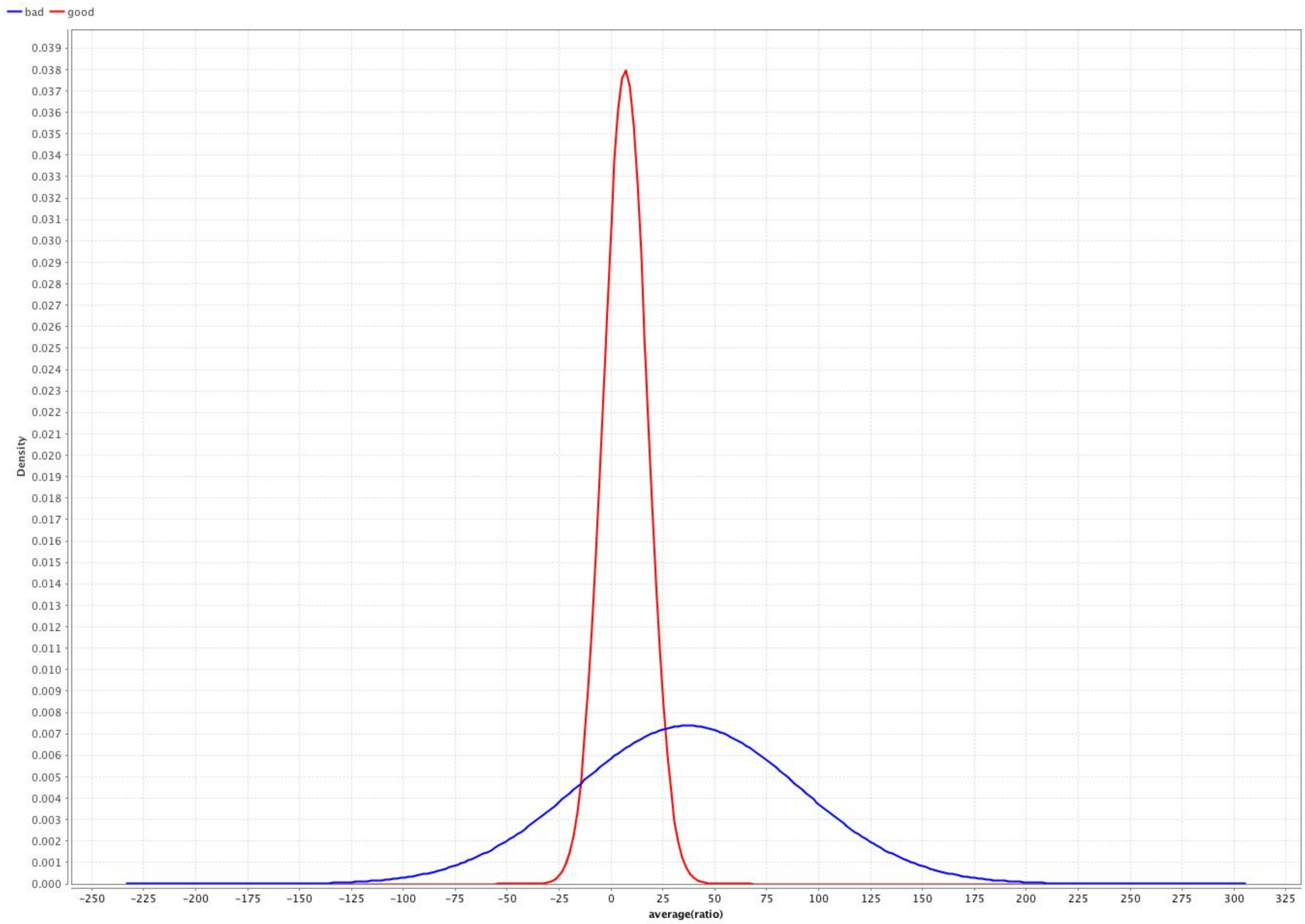
Naïve Bayes model

- Distribution model for label attribute status.
  - Class bad (0.111)
    13 distributions
  - Class good (0.889)
    13 distributions

Naïve Bayes model



Naïve Bayes model



Naïve Bayes model: evaluation

- Using the Rapidminer 'x-validation' operator (cross validation):
  - Number of validations: 15.

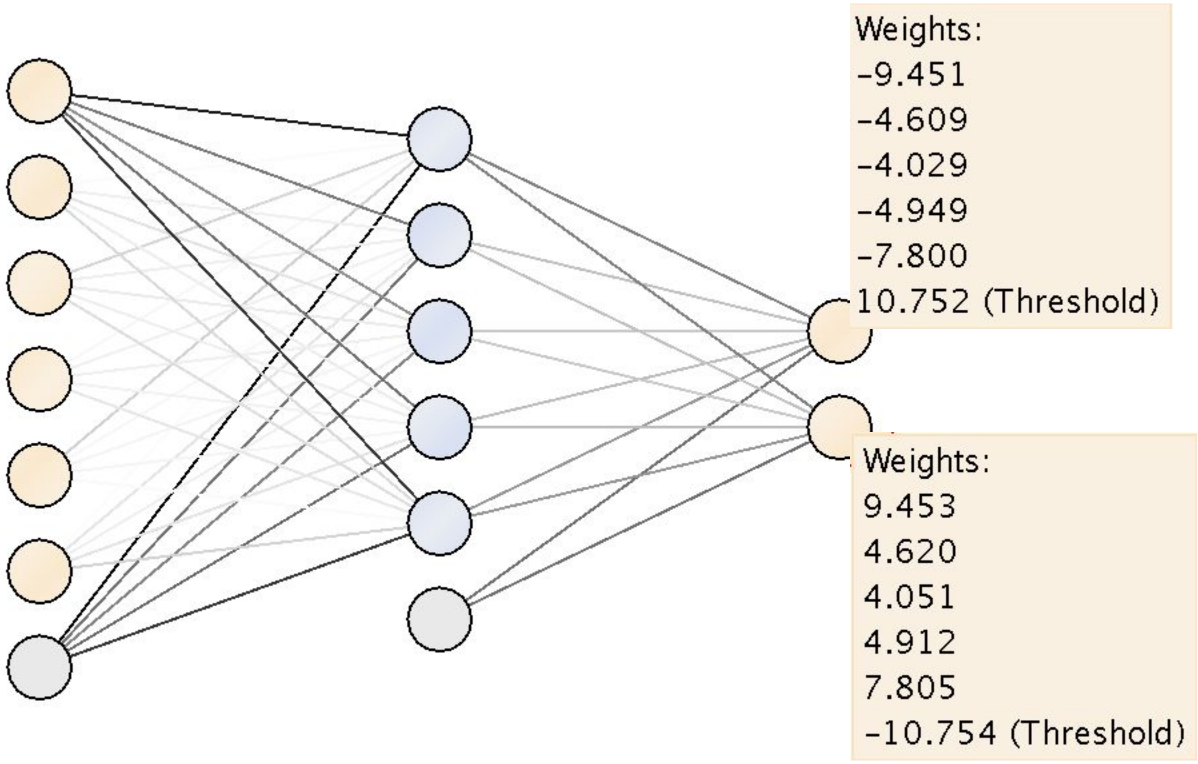
accuracy: 95.61% +/- 2.65% (mikro: 95.60%)

	true bad	true good	class precision
pred. bad	58	12	82.86%
pred. good	18	594	97.06%
class recall	76.32%	98.02%	

Neural Net model

- Using the Rapidminer ‘Neural Net’ operator:
  - Training cycles: 500;
  - Learning rate: 0.3;
  - Momentum: 0.3;
  - Shuffle: true;
  - Normalize: true.
- The following relations were joined: Client, Account, Loan, Disposition and Transactions.

Neural Net model



Neural Net model: Evaluation

- Using the Rapidminer ‘x-validation’ operator (cross validation):
  - Number of validations: 15.

accuracy: 97.07% +/- 2.07% (mikro: 97.07%)

	true bad	true good	class precision
pred. bad	66	10	86.84%
pred. good	10	596	98.35%
class recall	86.84%	98.35%	

classification_error: 2.93% +/- 2.07% (mikro: 2.93%)

Neural Net model: Discussion

- Better performance than all of the remaining models.
- However, there are some drawbacks:
  - Slow training process;
  - Difficult to interpret final weight values.

Conclusions and future work

- The data set example has a strong *class imbalance,* which can mislead some classification algorithms presented. There are two possible solutions:
  - Sampling of the input data;
  - Collection of new data.
- Apply more predictive data mining algorithms for the creation of new models:
  - Support Vector Machine;
  - Ensemble methods.

Conclusions and future work

- Empirical validation of the predictive models in a real world scenario is essential.