Banking
a data mining study case

Faculdade de Engenharia da Universidade do Porto

Extração de Conhecimento e Aprendizagem Computacional

MIEIC 2016/2017

Group 20

João Bordalo - ei12040
João Pereira - ei12023

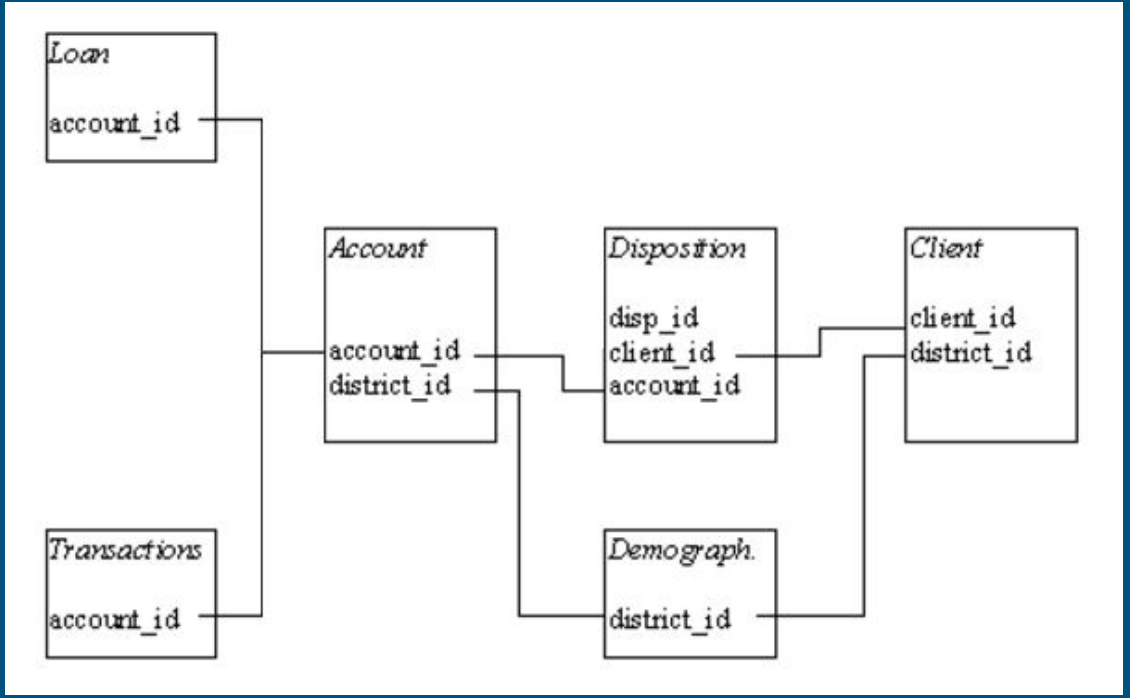
Problem and Context

The bank needs to improve their loans system and for that it’s needed to know who are the bad clients and the good ones.

Data Source

Used Datasets:

- account.csv;
- client.csv;
- disp.csv
- loan_test.csv;
- loan_train.csv;
- trans_test.csv;
- trans_train.csv.



Legend: Data relationships

Data Exploration

- Mean/Average values;
- Variance/Deviation;
- Mode;
- Correlation matrix;
- Charts visualization and analysis.

Data Cleaning

- Extract gender from *birth_number* in *Client*'s relation
- The conversion of the *date* in some datasets to a valid format
- Removing relations that the group has not considered relevants, i.e. the Cards dataset

Data transformation (semi-automatic)

The group has used the Pandas library for Python to calculate some features in an automatic way. By developing scripts to achieve the means and to see if an account has, at any time, negative balance.

This kind of operations were mostly used in the transactions dataset.

Sorting techniques to organize the data, using excel functions, and group by and aggregate with Python.

Data transformation (Manual)

The group has created scripts to calculate the majority of the values presented in our final dataset of features.

RapidMiner was used to join some simple attributes (non-calculated attributes) to the final dataset (train or test).

DM Descriptive: Problem

What might influence the success of a loan?
And how does it influence its success?

DM Descriptive: Algorithms

- **DBSCAN**
- **x-means**
- **k-medoids**
- **K-means**

DM Descriptive: Algorithms (DBSCAN)

- **Tuning Parameters:**
  - **Epsilon:** for the radius;
  - **Minimum points:** for the number of points allowed in epsilon range;
  - Comparing with other algorithms and analyze the clusters' properties

DM Descriptive: Algorithms (x-means)

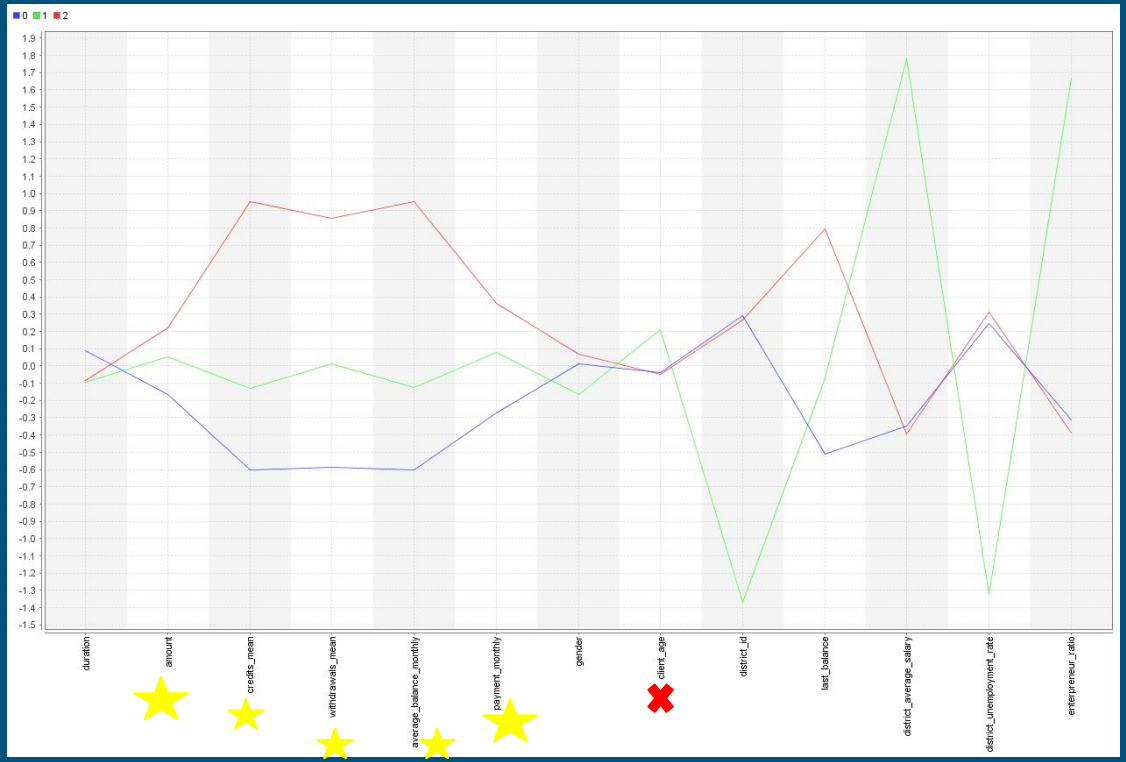
- **Tuning Parameters:**
  - **Minimum and maximum K value;**
  - **Maximum runs;**
  - **Maximum optimization steps.**
  - Comparing with other algorithms and analyze the clusters' properties

DM Descriptive: Algorithms (k-means and k-medoids)

- **Tuning Parameters:**
  - Drawing the line chart that shows the features' relation between clusters. This is to find the best K;
  - Internal indices: silhouette index and within groups of sum square index to find the best K
  - Comparing with x-means algorithm. This algorithm finds the best K for the given dataset;
  - Centroid variation to find the best K, related with the internal indices;
  - Variating the max number of runs of k-means with random initialization.
  - Comparing with other algorithms and analyze the clusters' properties

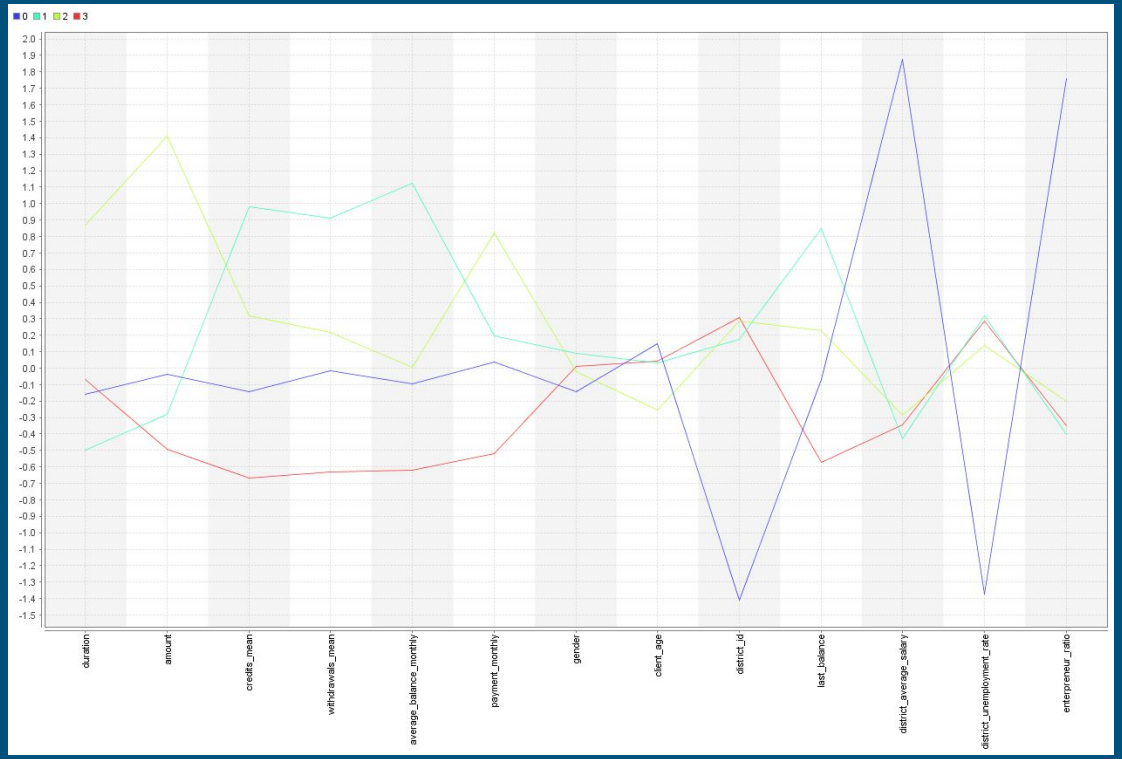
K-means algorithm was the one with the best results that we found within the selected features.

DM Descriptive: Algorithms (k-means)



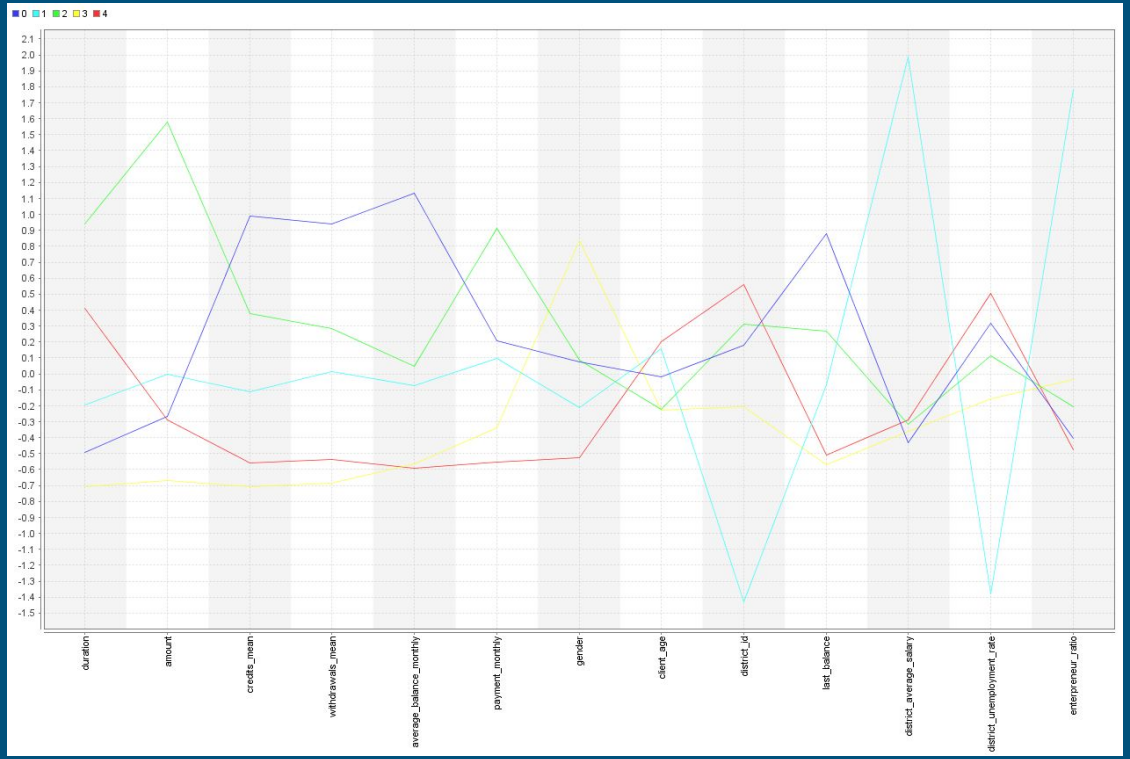
Legend: line chart with features' relation between clusters using k-means, where K = 3

DM Descriptive: Algorithms (k-means)



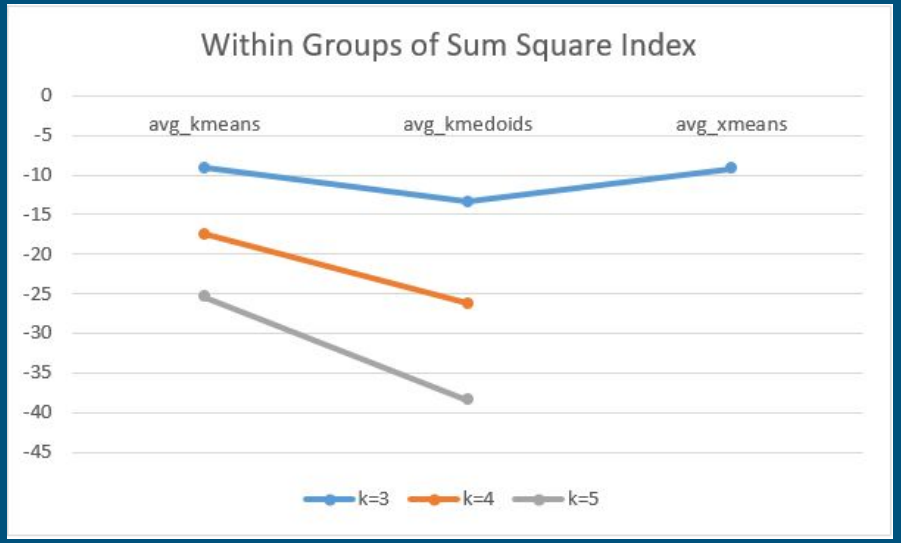
Legend: line chart with features' relation between clusters using k-means, where K = 4

DM Descriptive: Algorithms (k-means)

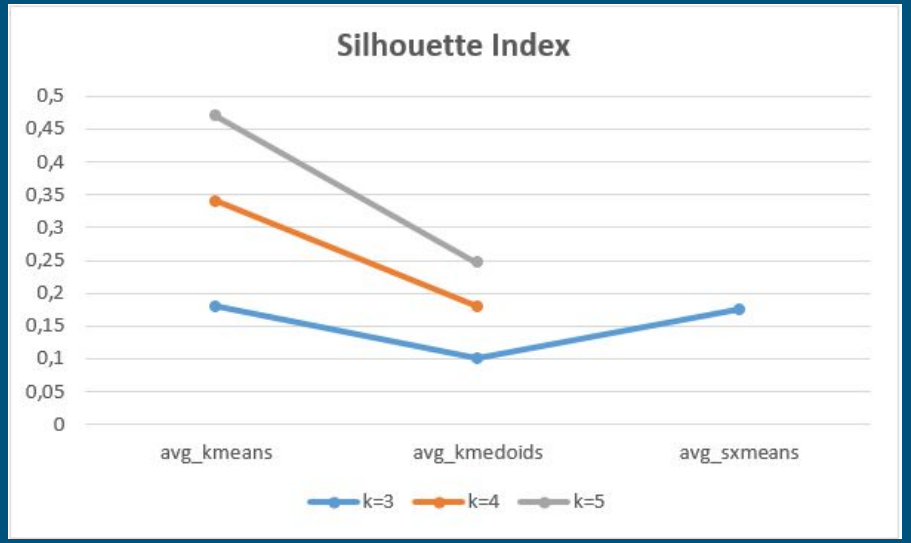


Legend: line chart with features' relation between clusters using k-means, where K = 5

DM Descriptive: Algorithms (Validation)



Legend: line chart with the average value of the within groups of sum square index for each algorithm and each k value



Legend: line chart with the average value of the silhouette index for each algorithm and each k value

DM Descriptive: Algorithms (Validation)

- The lowest the average value of t**he within groups of sum square index** the closer the cluster's elements are to each other (compactness);
- The average value of the **silhouette index** closer to 1 means that the cluster's elements are close to each other in the same cluster and far away from others clusters' elements (compactness and separation);
- With these metrics is possible to evaluate and validate a correct number of clusters, k, that helps understanding the dataset;
- Although, analysing the possible charts and other variables, we found that the algorithm K-means, with k = 4, gave us, in our opinion, a valid set of

DM Descriptive: experimental methodology

- Iterative Process
  - Increase of number of features in each iteration and evaluation
  - Selection of the best features in each algorithm used
- Features Normalization
- Clustering algorithm application
  - Evaluation of each cluster validation metrics (internal index, plot visualization, algorithm comparison)

DM Descriptive: experimental methodology

What might influence the success of a loan?

Loan

District

Account

DM Descriptive: experimental methodology

Loan

- Total Amount
- Average Amount
- Average Duration

District

- Average Salary
- Average Entrepreneur ratio
- Average Unemployment Rate

Account

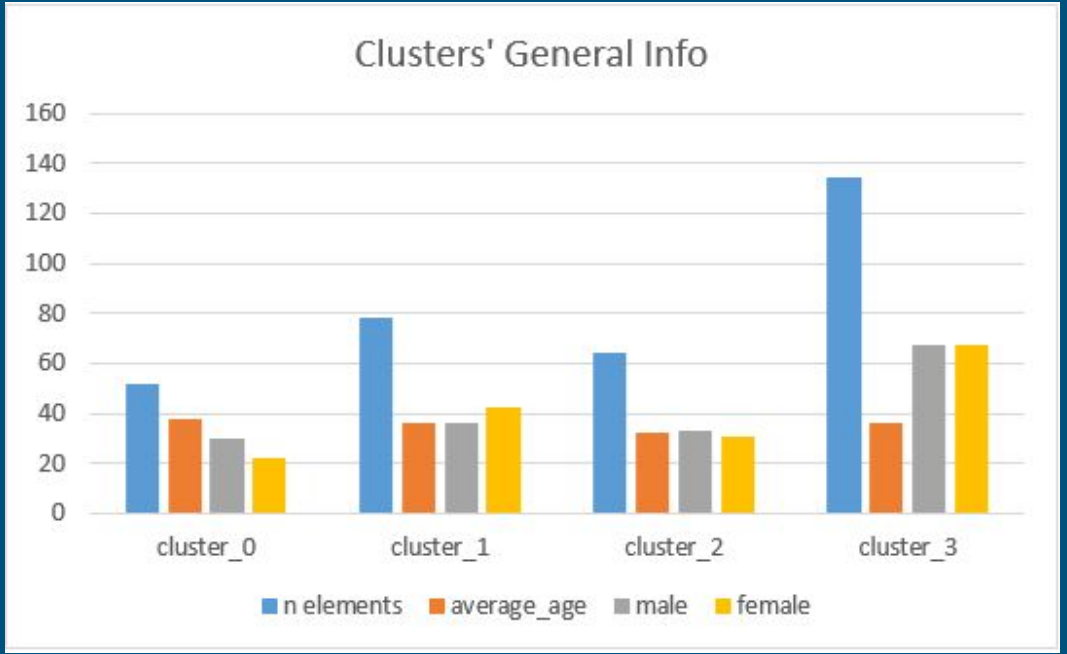
- Average Credit
- Average Withdrawals
- Average Monthly Balance

DM Descriptive: Results analysis (Overview)

	cluster_0	cluster_1	cluster_2	cluster_3
n elements	52	78	64	134
average_age	37,6	36,15	32,21	36,32
male	30	36	33	67
female	22	42	31	67

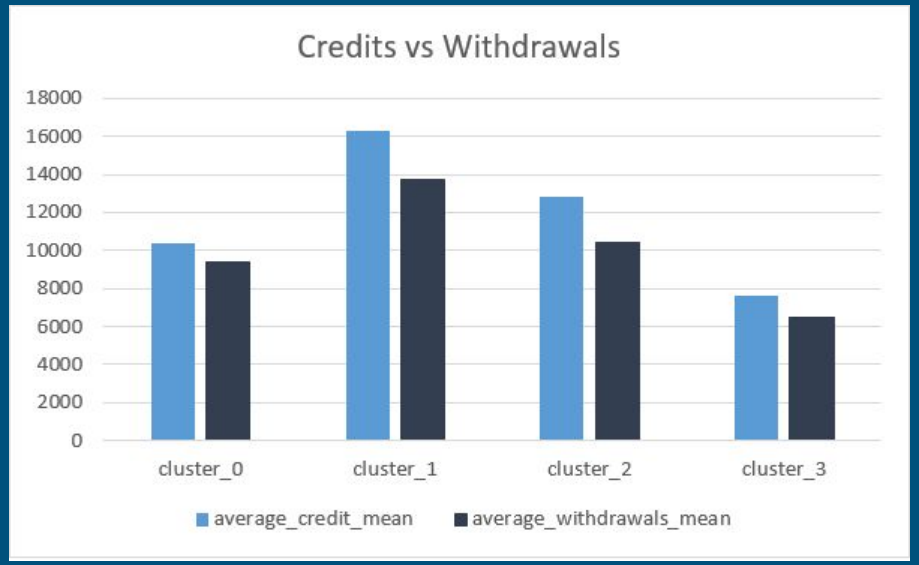
Legend: table with the clusters overview for each cluster using k-means, where K = 4

- cluster_3 has the highest number of elements;
- The average age in all clusters is about the same and they belong to active population;
- All clusters have the same gender proportion, although clusters 0 and 1 have a slight difference

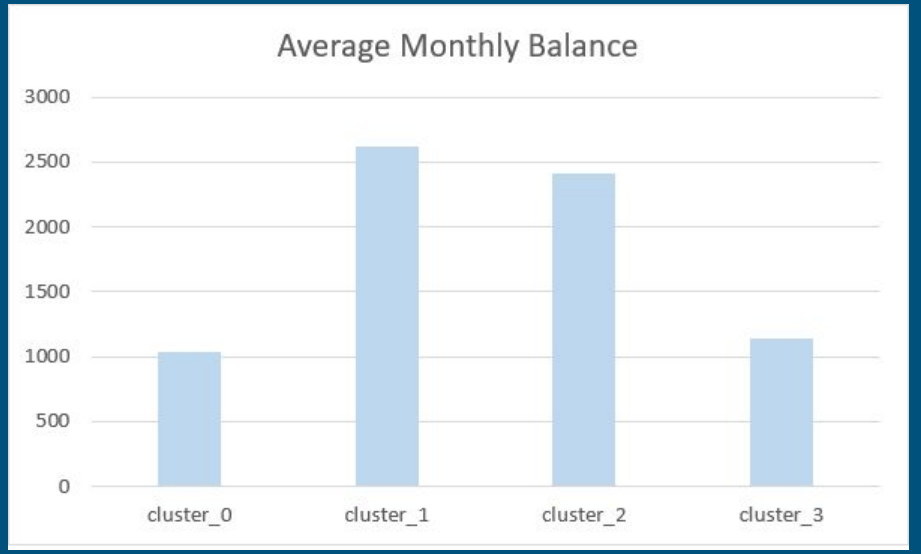


Legend: Bar chart with clusters' overview for each cluster using k-means, where K = 4

DM Descriptive: Results analysis (Account)



Legend: Bar chart with credits' and withdrawals' value for each month for each cluster using k-means, where K = 4

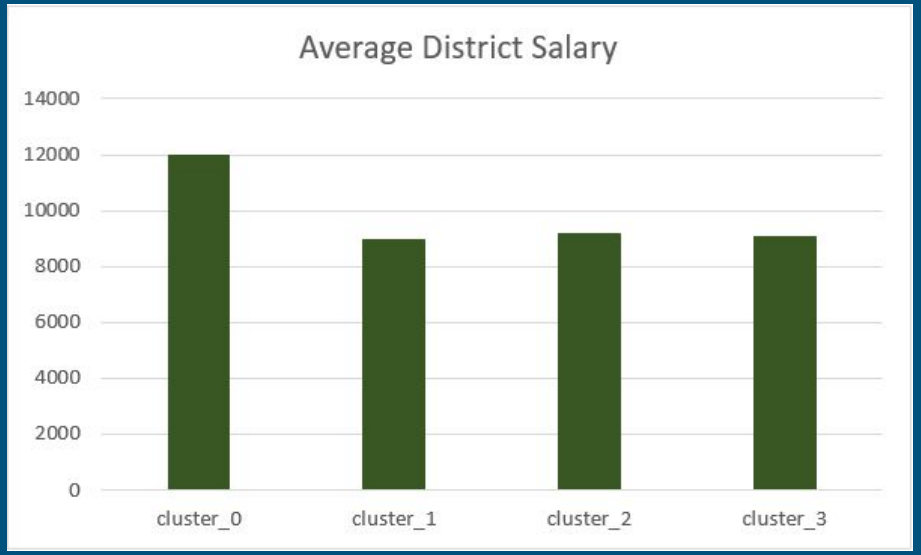


Legend: Bar chart with the average monthly balance for each cluster using k-means, where K = 4

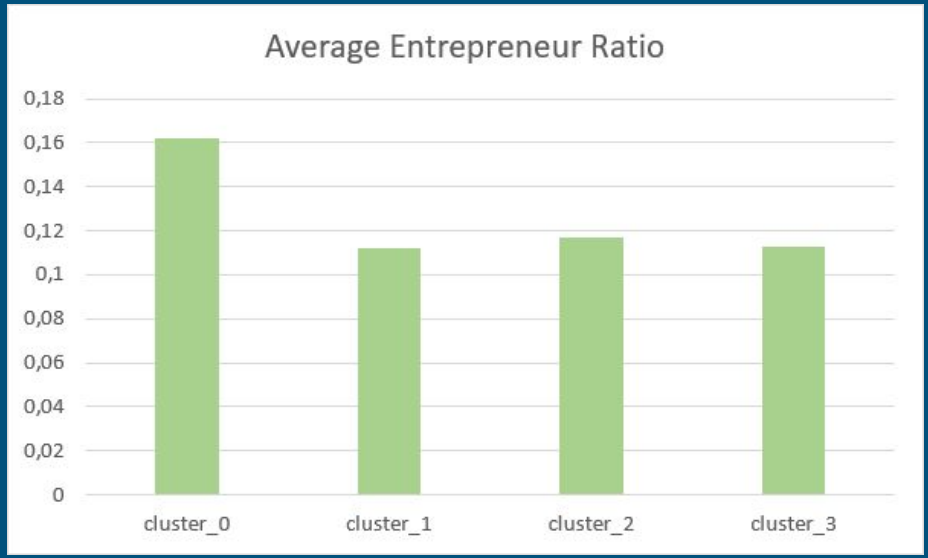
DM Descriptive: Results analysis (Account)

- Overall, all clusters have a positive balance, it means that there is more money entering than leaving the accounts;
- Although, cluster_1 is the one with the best balance. However, cluster_1 has the highest average withdrawals monthly value, which means that has the highest average credit monthly value.

DM Descriptive: Results analysis (District)

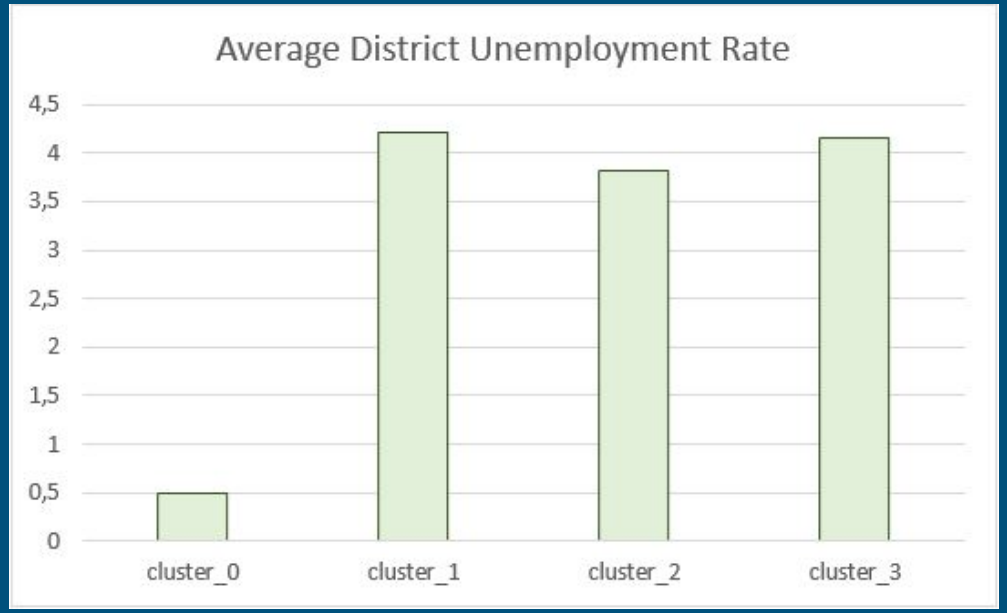


Legend: Bar chart with average district salary for each cluster using k-means, where K = 4



Legend: Bar chart with average district entrepreneur ratio for each cluster using k-means, where K = 4

DM Descriptive: Results analysis (District)

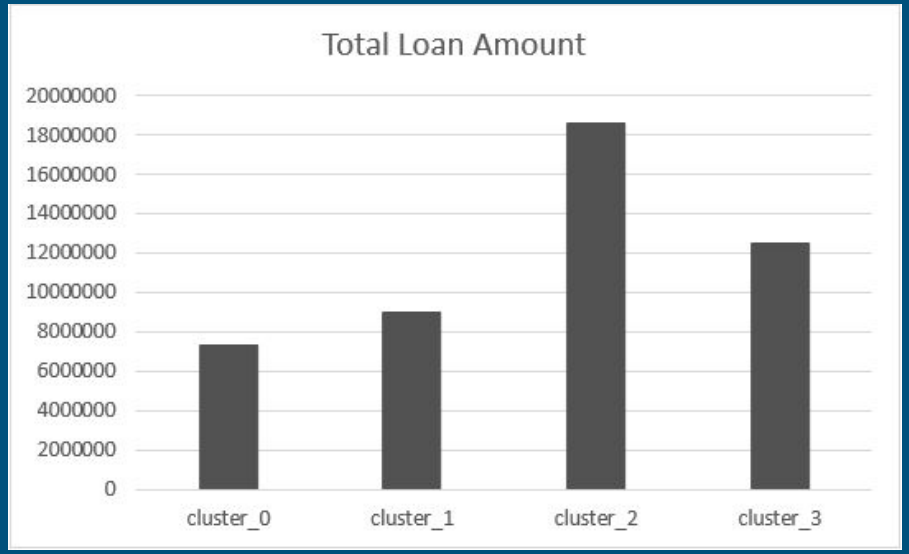


Legend: Bar chart with average district unemployment rate for each cluster using k-means, where K = 4

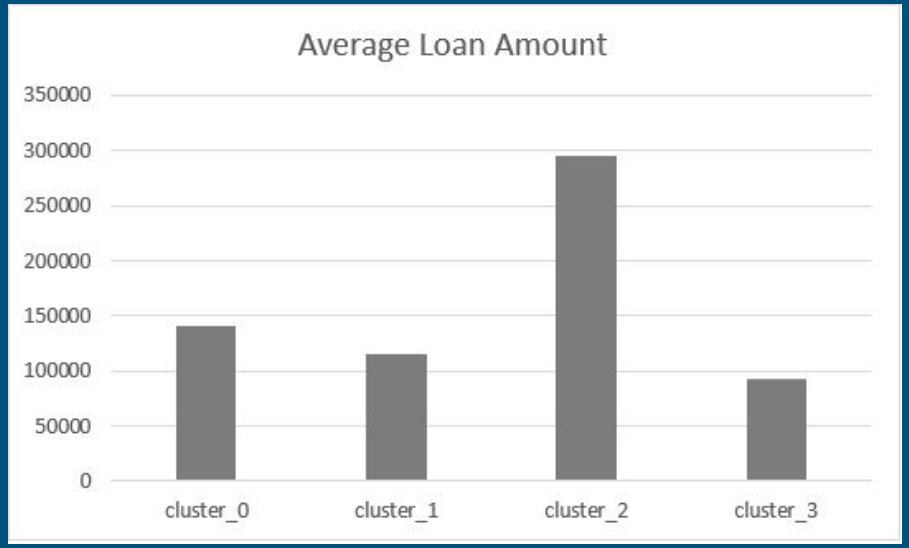
DM Descriptive: Results analysis (District)

- cluster_0 has the people with the best average district salary;
- It might be related with the lowest average district unemployment rate and the highest average district entrepreneur ratio;
- About the others clusters, all have about the same average district salary, average district entrepreneur ratio and average district unemployment rate;
- Although, cluster_1 stands out negatively because it has has lowest average district salary, lowest average district entrepreneur ratio and highest average district unemployment rate;

DM Descriptive: Results analysis (Loan)

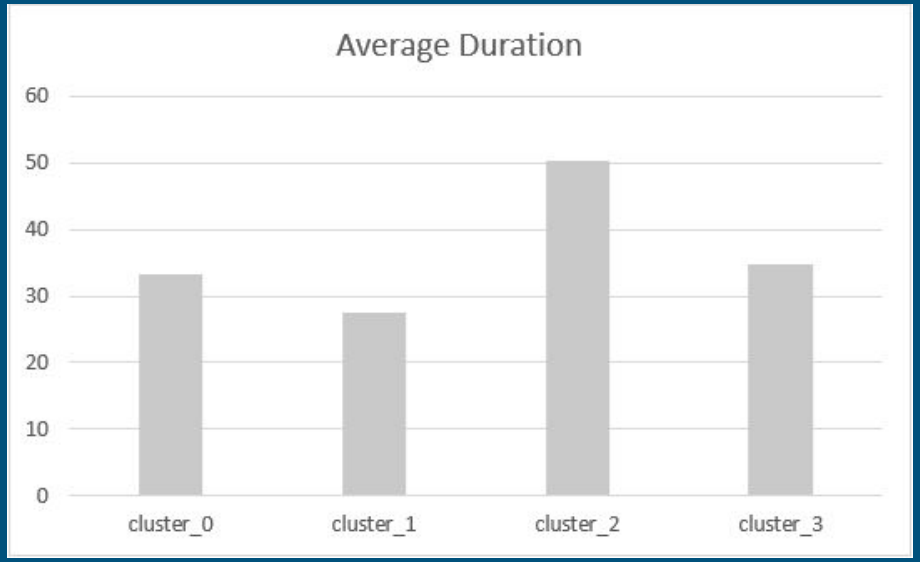


Legend: Bar chart with total loan amount for each cluster using k-means, where K = 4

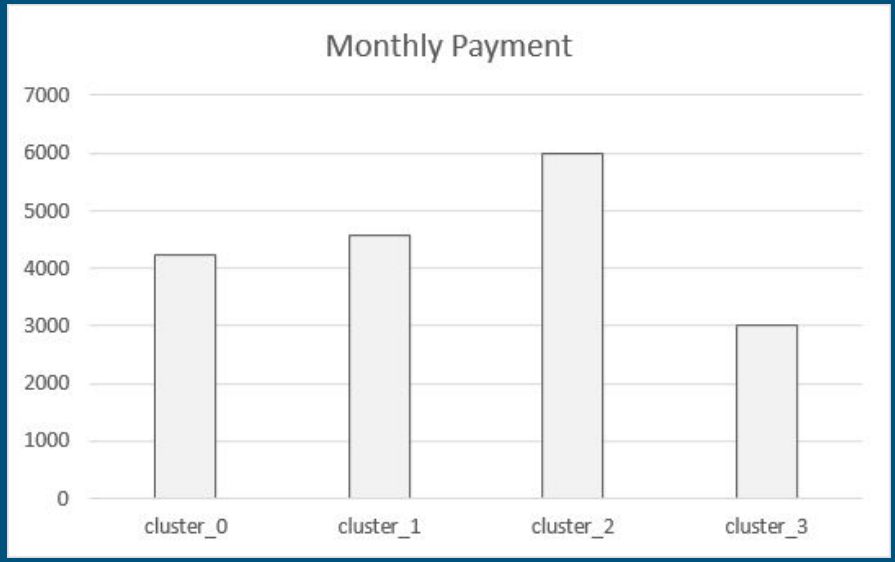


Legend: Bar chart with average loan amount for each cluster using k-means, where K = 4

DM Descriptive: Results analysis (Loan)



Legend: Bar chart with average loan duration for each cluster using k-means, where K = 4



Legend: Bar chart with monthly loan payment for each cluster using k-means, where K = 4

DM Descriptive: Results analysis (Loan)

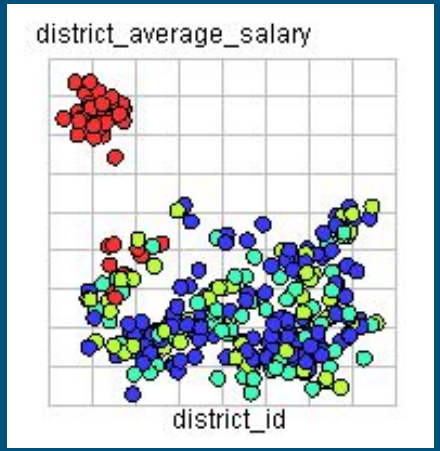
- Cluster_2 has both total loan amount and average loan amount highest values, which in turn corroborates with the fact that the cluster has the highest monthly payment value and loan duration;
- About the others clusters, there is no relevant analysis that can be concluded.

DM Descriptive: conclusions

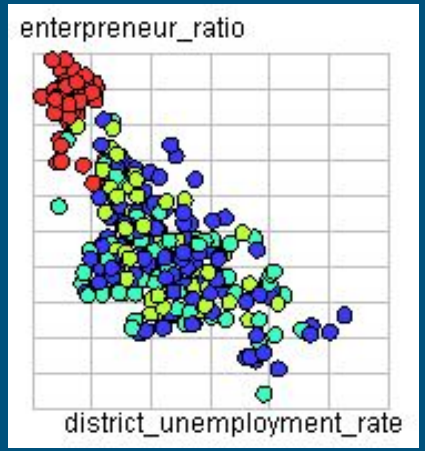
Legend: Bar chart with loan success ratio for each cluster using k-means, where K = 4

- Cluster_1 is the one with the best successful loan rate, this might be the fact that cluster_1 has the lowest loans' duration and a low average and total loan value.
- cluster_3 has the highest unsuccessful loan rate, due to its heterogeneous characteristic
- The other two clusters remaining have about the same loan success rate.

DM Descriptive: conclusions for cluster_0 (red)



Legend: plot chart with district average salary vs district id, using k-means, where K = 4



Legend: plot chart with district entrepreneur ratio vs district district unemployment rate, using k-means, where K = 4

- Homogeneous cluster;
- Rich cluster
- High purchasing power;
- Majority live in the same district (Hl.m. Praha);
- High probability to be an entrepreneur;
- Don’t need large amount of loan;
- Can easily pay their loan, reducing the loan duration;

DM Descriptive: conclusions for cluster_1 (greenish)

- From middle class;
- Can afford loans but with low duration and amount;
- This is plausible because they have a high average district unemployment rate, low average district salary and low average district entrepreneur ratio;
- With a positive and high balance makes them able to fulfil the loans' payment.

DM Descriptive: conclusions for cluster_2 (green)

- From middle class;
- But, probably, people starting their own business;
- Because they have the highest total loan amount, average loan amount highest values, and loan duration;
- Because they are starting their business, average district salary, average district entrepreneur ratio and average district unemployment rate aren't so significant;
- It is a gamble for them to pay their loan, since business could go well or not.

DM Descriptive: conclusions for cluster_3 (blue)

- From middle-low class;
- Heterogeneous cluster;
- The Poorest cluster;
- Impossible to know if they can afford loans or not;
- This might happen because the cluster is the biggest one and with a lots of noise;

DM Predictive

The loans training dataset is relatively small: ***328 loans*** in total.

There are ***282 successful*** loans and only ***46 unsuccessful***.

These factors make the predictive task more difficult because we are dealing with a unbalance classes problem.

DM Predictive: Algorithms

- **Support Vector Machine**
- **Decision Tree**

DM Predictive: Algorithms (SVM)

- **Tuning Parameters:**
  - **Kernel (Radial, Anova or Polynomial)**
  - **Balance cost;**
  - **Cost (C)**
- **Auxiliar Models/Operators:**
  - **Bagging**
    - **Number of iterations**
    - **Ratio Sample**
  - **Features normalization**

DM Predictive: Algorithms (Decision Tree)

- **Tuning Parameters:**
  - **Criterion (Information Gain, Gini Index)**
  - **Pruning**
    - **Confidence**
  - **Prepruning**
    - **Minimal Gain**
- **Auxiliar Models/Operators:**
  - **Bagging**
  - **Normalization**

DM Predictive: experimental methodology

- Model Validation
  - Non-exhaustive cross-validation
  - k-fold between 3 to 10 folds
  - Stratified sampling
- Iterative Process
  - Increase of number of features in each iteration and evaluation
  - Selection of the best features in each algorithm used
- Features Normalization
- Bagging
  - Reduce variance
  - Avoid overfitting

DM Predictive: experimental methodology

Most relevant features calculated/used:

1st Iteration:
Loan Amount, Credits mean, Withdrawals mean, Variation of Credits and Withdrawals, District unemployment rate, Loan duration, Entrepreneur Ratio, Last balance

2nd Iteration
Credits mean (another bank, cash, interest), Withdrawals mean (another bank, cash, credit card)

DM Predictive: experimental methodology

Most relevant features calculated/used:

3rd Iteration:
Account Sanctions (binary feature)

DM Predictive: models analysis

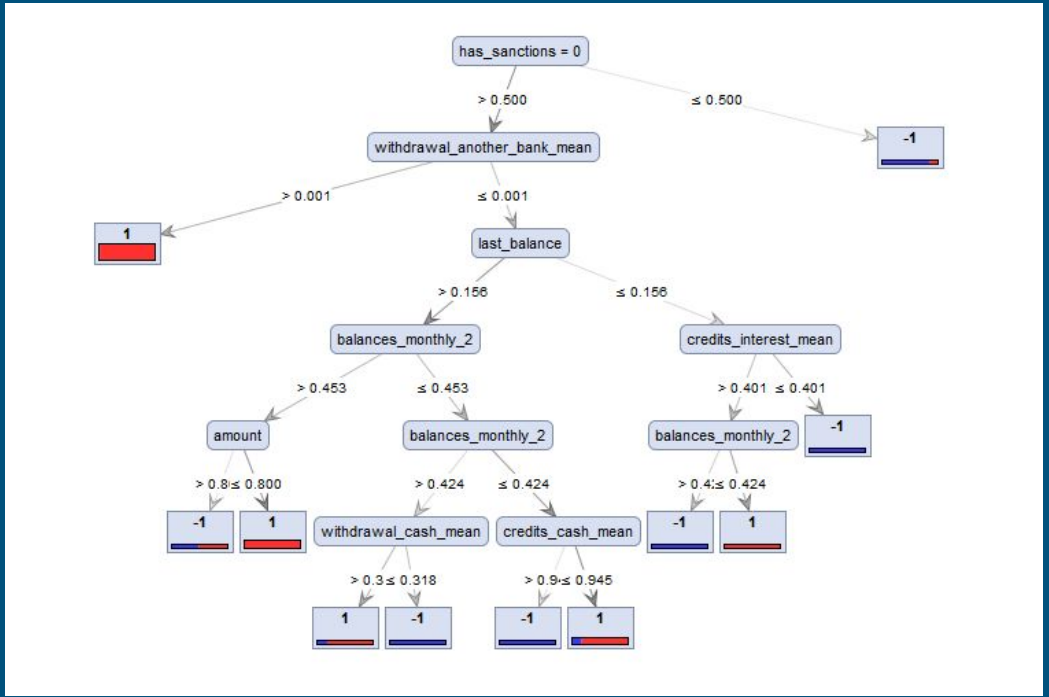
Decision Tree

Criterion: Gini_Index

Maximal Depth: 7

Confidence: 40%

Minimal Gain: 4%



Legend: Model of the decision tree algorithm.

DM Predictive: evaluation metrics

Decision Tree

AUC Treino
0.753 +/- 0.123

	true -1	true 1	class precision
pred. -1	20	15	57.14%
pred. 1	26	267	91.13%
class recall	43.48%	94.68%	

DM Predictive: models analysis

Support Vector Machine

Kernel: Anova

Kernel Gamma: 1.0

Kernel Degree: 2.0

Balance Cost: True

Kernel Model

```
Total number of Support Vectors: 328
Bias (offset): 1.691

w[has_sanctions = 0] = 0.010
w[has_sanctions = 1] = -0.227
w[amount] = -0.095
w[balances_monthly_2] = -0.008
w[last_balance] = 0.384
w[district_unemployment_rate] = -0.071
w[credits_cash_mean] = -0.117
w[credits_another_bank_mean] = -0.084
w[credits_interest_mean] = 0.134
w[withdrawal_another_bank_mean] = -0.073
w[withdrawal_cash_mean] = -0.258
w[withdrawal_credit_card_mean] = 0.003
```

Legend: Weights of the features used in the model.

DM Predictive: evaluation metrics

Support Vector Machine

AUC Training
0.793 +/- 0.128

AUC Public Test (Kaggle)
0.84193

Best obtained result

	true -1	true 1	class precision
pred. -1	19	25	43.18%
pred. 1	27	257	90.49%
class recall	41.30%	91.13%	

DM Predictive: Results analysis

- The results obtained with the algorithm SVM using kernel type Anova were the satisfactory ones. After several submissions, the decision tree algorithm results were not so good. The problem was, probably, the chosen features and the range of its values.
- Applying a discretization in the features of the decision tree the result will probably be better.
- The group has used the training AUC and the classes recall to conclude if it’ll be valuable to try, or not, a submission on Kaggle, but not always the best results were reflected. One conclusion we can take is the model was overfitted.

Tools

- RapidMiner
  - Models application and classification
- Excel
  - Visualize the data and take some relevant conclusions
- Python with Pandas
  - Data cleaning
  - Creation of relevant features for descriptive and predictive analysis

Conclusions and Future work

DM Descriptive:

- The best way to find the right number of clusters depends on the dataset behaviour, algorithm used, previous pre-processing, finding the best features, validating those features, finding the best algorithm's variables value and how to validate the quality and cluster's trustworthiness
- Although, the best way to find the number of cluster is to look into the data, instead relying only on numbers
- For future work, try to use more and/or better features
- Use a better validation on clusters to easily find better conclusions from the analysis, such as external measures (Entropy and Purity)

Conclusions and Future work

DM Predictive:

- The group should have take the problem considering the unbalanced classes, because we have more successful loans than unsuccessful
- The application of methodologies of features' selection, such as wrapper approaches, should have simplified the task of finding relevant features
- Training the models with others algorithms such as Deep Learning, or even Random Forest to evaluate if the testing results would be better