To Loan Or Not To Loan

T7G3

Afonso Caiado	up201806789@up.pt
Elias Lambrecht	up202102122@up.pt
José Miguel Maçães	up201806622@up.pt
Luís Miguel Afonso Pinto	up201806206@up.pt

Domain description

The human being has always learned by observing patterns, formulating hypothesis and testing them to discover new rules. Data mining does exactly that. It is the art and science of transforming raw data into useful information. Software is used to search for patterns and associations to determine connections between different variables, and even to create new ones.

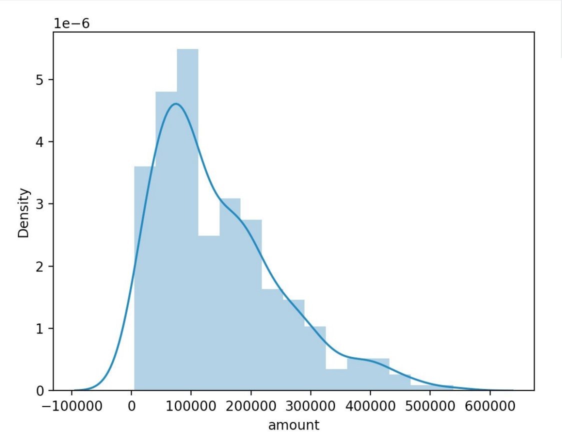
Data mining is used to better understand the clients, their patterns and motivations. By exploring large amounts of data, predictions for the future can be made.

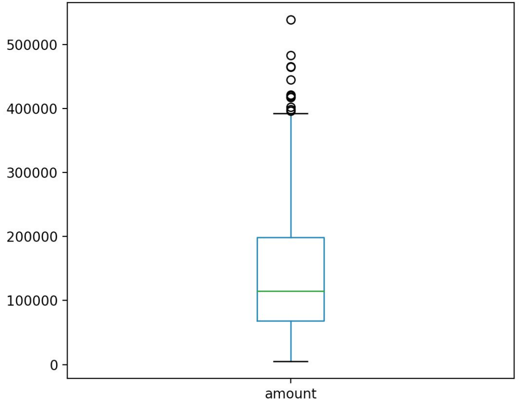
In our particular case, it is not a totally innovative idea, as we found 2 similar machine learning projects, concerning bank loans:

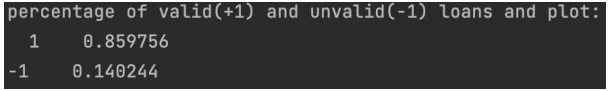
Machine learning: Predicting Bank Loan defaults and **Factors that affect loan giving decision of banks** (both links can be found in the annexes)

Exploratory data analysis: loan dataset

- Histograms to analyse each columns density
- Boxplots to analyse and detect possible outliers in each column
- Calculation of the percentage of valid vs invalid loans
- Average loan amount and average monthly loan amount by status analysis







```
percentage of valid(+1) and unvalid(-1) loans and plot:
  1    0.859756
-1    0.140244
```

Problem definition

Data about clients, accounts and loans are provided.

We need to predict whether a loan will end successfully, in order for the bank to be able to improve their understanding of customers and seek specific actions to improve services and prevent money losses with default loans.

To do so, we need to analyze the information we have, prepare the data accordingly, build our model and make our predictions.

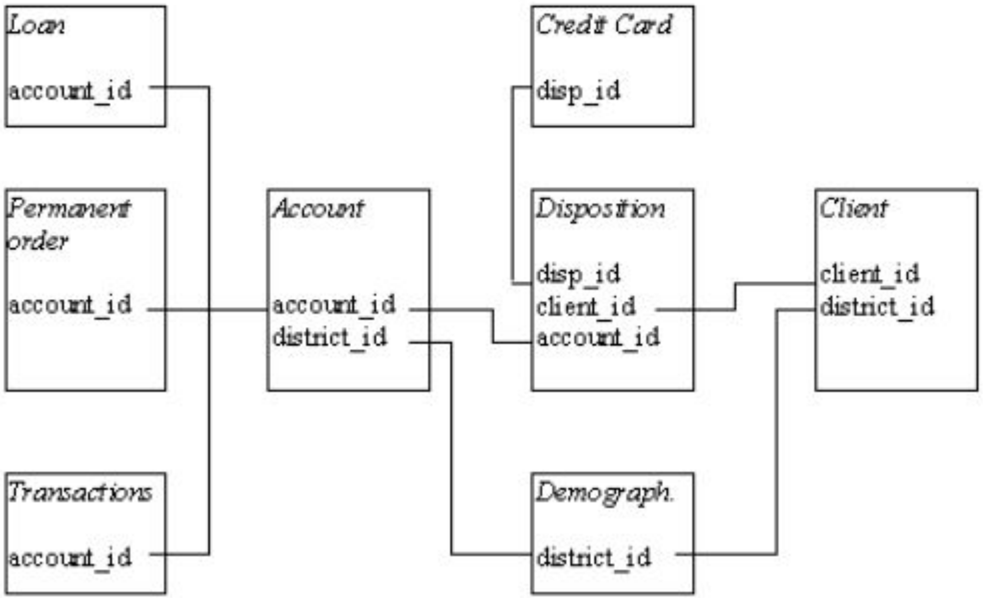
Data preparation

1. Join loan, account and district
   - Categorical to numeric
   - Missing values -> mean

2. Join transactions
   - 1 to N -> 1 to 1 relationship by aggregate of balance

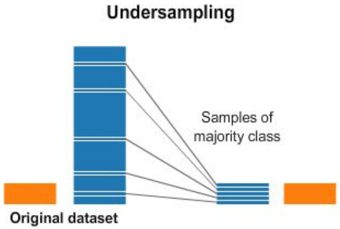
3. Join card
   - Left join
   - Categorical to numeric

4. Join client (date and gender)
   - Map account_id with client_id
   - Decipher birth_number

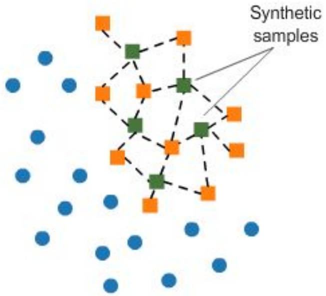


Data preparation

- Feature engineering
  - effort (=amount / average balance)
  - salary_effort (=amount/ average salary)
  - monthly effort (=payments / average salary)
- Imbalanced data
  - undersampling
  - SMOTE



SMOTE



Experimental setup

- Cross Validation: data splitting with a 70% / 30% ratio
- Data split taking the dates into account (train with pre-96 loans to predict post-96 loans)
- Another data splitting technique was experimented (0.73 for LR and 0.76 for RF), however it lead to overfitting and bad results on the competition
- AUC used for performance measure
- 4 different algorithms tested: Decision Tree, Random Forest, Gaussian Naive Bayes and Logistic Regression
- Python was used for the whole process

Results

- Gaussian Naive Bayes results:
  - 0.53 AUC -> considers every feature to be independent, has a simple approach to the problem
- Decision Tree Classifier results:
  - 0.52 AUC -> similar approach to Gaussian NB (tree was illogical for some parameters; gold card owners were given worse probabilities than junior card owners)
- Both algorithms are prone to overfitting, hence the bad results.

Results

- Random Forest Classifier:
  - 0.65 AUC -> the best parameters were found to be max_depth=5, n_estimators=41, random_state=5 (random search and grid search)
- Logistic Regression:
  - 0.70 AUC -> with a maximum number of iterations of 1000
- Empirical testing showed that our choice of criteria for data splitting had a better performance than random splitting. E.g. LR results with random splitting varied from 0.59 to 0.65 (100 experiments were made with different splitting each time)
- Both algorithms have a smaller chance of overfitting than the other presented

Conclusion

We addressed the data mining problem of a bank's loan prediction by preparing and analyzing the available data.

We found it difficult to take what we learned in the data analysis into our data preparation.

Many different tests and experiments have been left for the future due to lack of time (i.e. treatment of the detected outliers). We would have liked to have made a better data analysis and treatment. It could be interesting for example to have calculated a *balance_nearest_loan_date* variable, that would tell us the client's account balance nearest to the loan date.

Individual factors: We believe that the work was well divided between the members of our group, as everyone would have a individual factor of 1.

Annexes: Domain Analysis

- **Machine learning: Predicting Bank Loan defaults:** *https://towardsdatascience.com/machine-learning-predicting-bank-loan-defaults-d48bffb9aee2*
- **Factors that affect loan giving decision of banks:** *https://deepnote.com/@rhishab-mukherjee/Loan-Prediction-Project-TermPaper-VPSOpiywSu6FZeN2fK8fug*

Annexes: Exploratory data analysis (Distributions)

Not added yet