Data Mining Project

Work Summary

Autor

Machine Learning
Master's degree in Informatics and Computing Engineering

Porto, 2021

Resumo

Conteúdo

Capítulo 1

Business Understanding

- Analysis of requirements with the end user.
- Definition of business goals.
- Translation of business goals into data mining goals.

Capítulo 2

Data Understanding

- Diversity of statistical methods.
- Complexity of statistical methods.
- Interpretation of results of statistical methods.
- Knowledge extraction from results of statistical methods.
- Diversity of plots.
- Complexity of plots.
- Presentation.
- Interpretation of plots.
- Visual knowledge extraction.

2.1 Plots - Data Understanding

Looking at this plot, we can observe that most of the unsuccessful loans are usually located on the left part of the chart, which means that people with low balances on their accounts are prone to fail loan payments.

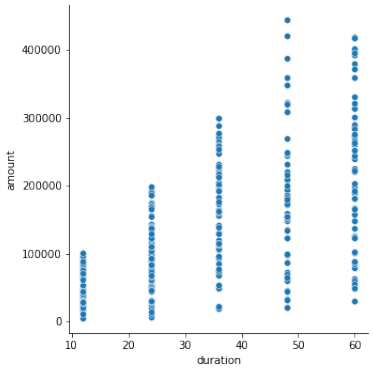

Figura 2.1: Plot

As we can see in this plot, higher duration loans result typically in higher amounts.

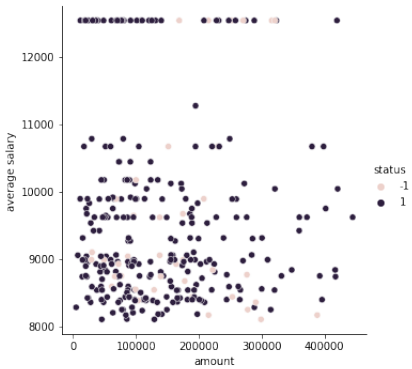

Figura 2.2: Plot

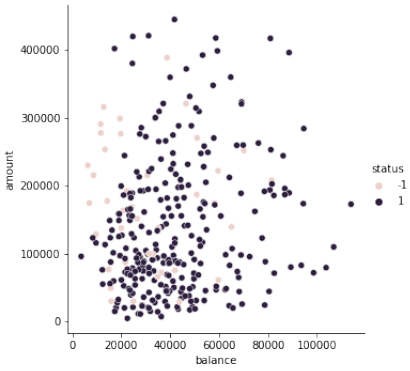

Figura 2.3: Plot

2.2 Plots - Processing

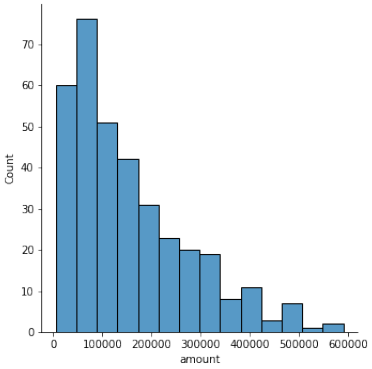

Figura 2.4: Plot.

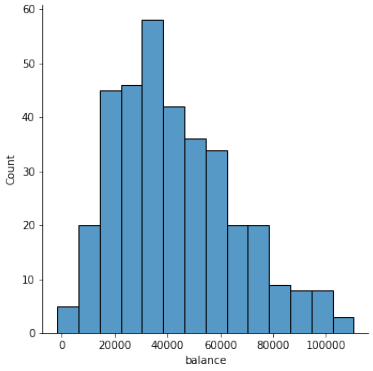

Figura 2.5: Plot.

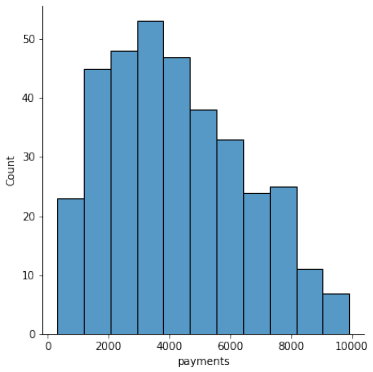

Figura 2.6: Plot.

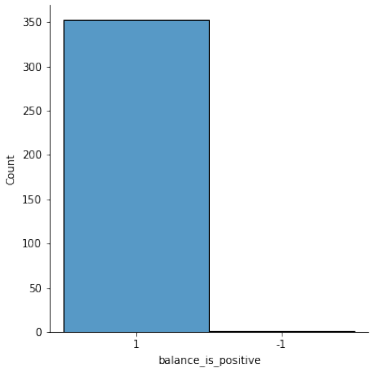

Figura 2.7: Plot.

Capítulo 3

Data Processing

- Data integration.
- Assessment of dimensions of data quality.
- Cleaning redundancy.
- Cleaning missing data.
- Cleaning outliers.
- Data transformation for compatibility with algorithms.
- Feature engineering from tabular data.
- Sampling for domain-specific purposes.
- Sampling for development.
- Imbalanced data.
- Feature selection.

Capítulo 4

Descriptive

- Diversity of algorithms.
- Parameter tunning.
- Understanding algorithm behaviour.
- Performance measure.
- Correct interpretation of performance measures.
- Comparative analysis of results.
- Model improvement.
- Analysis of resuls.
- Diversity of tasks.
- Diversity of algorithms.

Capítulo 5

Predictive

- Parameter tunning.
- Undertanding algorithm behavior.
- Performance estimation: training vs test.
- Performance estimation: other factors (time, ...).
- Performance estimation: performance measure.
- Performance estimation: correct interpretation of performance measures.
- Performance estimation: analysis of results.
- Model improvement.
- Feature importance.
- Analysis of "white-box"models.

Capítulo 6

Project

- Management methodology.
- Management plan.
- Project management tools.
- Collaboration tools.

We are using Github for version control and collaboration. We regularly do pair programming using Visual Studio Code live sharing feature.

Capítulo 7

Tools

- Analytics.
- Database.
- Other tools (data cleaning, visualization).

Capítulo 8

Presentation

- Quality of layout.
- Quality of content in slides.
- Delivery.
- Use of time.