SBD2 - Data-Drive Visualization for decision making (AS25)

Christoph Zangger, Branka Hadji Misheva and Lucia Gomez

Hackathon Handout

Starting Arguments

- Submission will close exactly at 12.15; please submit your solutions before that.
- You have 3 hours to complete this hackathon.
- This is a bring-your-own-device hackathon. Hence, you must come with your own device.
- You can use any programming language for the task (R, Python …) and you can use any notebook (R Markdown, Google Colab …) to submit the results.
- This is an open-resource exam. For the exam, you can use all academic resources (articles, book chapters, lectures, and lecture notes). Moreover, you allowed to use ChatGPT (or any other LLM or AI assistance).
- This is an **INDIVIDUAL EXAM**, group discussions or collaborations are not allowed.
- You should hand-in a codeNotebook with all your answers via Moodle.
    - In case you are submitting an R markdown, please submit the R markdown and the complied HTML file on Moodle
    - In case you are working in Google Colab, please submit a link to the google colab notebook. (Note: in case you are working only online, it is always a good practice to download the final output and store it locally.)
    - In other case, the notebook needs to include all code, results and your answers to the questions.
    - Your notebook needs to provide detailed text explanations covering the decision-making process that justifies the analytical pipeline executed and to discuss the results obtained and what they mean for the problem at hand (computational thinking and analytical interpretation).

Task Overview

Your challenge is to **analyze the Bondora dataset** (Bondora.csv) **and build a classification model** to predict whether a loan will result in **default** or **non-default**. The **target variable** in this challenge is the feature **default**, which takes the value **1 if the loan has defaulted** and **0 if the loan has not defaulted**. This is the variable you will try to predict using classification techniques. A complete description of all available features in the dataset can be found in the file **Description.csv**, which provides definitions and explanations of each variable to help guide your analysis.

Steps to follow:

1. **Data Analysis (10 points)**: get familiar with the data; look at the target distribution (default vs. non-default); Identify at least 3 interesting feature patterns.
2. **Data Transformation (10 points)**: discuss whether you need to run some data transformations. If yes, discuss in detail your choices & justify them.
3. **Modelling (10 points)**: propose at least **two possible classification approaches**. For each: explain briefly why it might work, and list **pros & cons.** Note: it is important to put this discussion into perspective of the actual data and use case. Generic arguments on the pros and cons linked to certain classification techniques will be awarded zero points.
4. **Implementation (20 points)**: **Choose one** of the approaches and build a classifier. Perform a train/test split.
5. **Evaluation (30 points)**: Use at least **two evaluation metrics** to discuss the result. You should not only **report the numbers**, but also **interpret what the results mean** in the context of loan default prediction. Consider which evaluation metric is most suited for your use case.
6. **Reflection (20 points)**: If you had more time/resources, what would you do next? Which additional datasets might improve your model? How do you interpret the current performance of your model? How do you interpret the parameters/features and their relevance in your model (if applicable)?

Evaluation Rubric

Completeness – Does the submission cover all required steps (EDA, data transformation, modeling approaches, implementation, evaluation, reflection) and respects required format (notebook including code, results and in text explanations)?
Correctness – Are the technical steps (analysis, transformations, modeling, evaluation) carried out correctly and logically?
Clarity – Are explanations, reasoning, and presentation clear, structured, and easy to follow?
Critical Thinking – Does the student go beyond surface-level answers, showing thoughtful consideration of data issues, model choices, and limitations?
Assessment – Are the evaluation metrics appropriate for the task, and is there meaningful discussion of model performance?