Homework 2: Credit Card Default Prediction

Prof. Alessio Martino & Lorenzo Ariemma & Mattia Cervellini

December 18, 2025

Scenario of Interest

Imagine you are working for a financial institution that has shared historical data on its credit card clients. As a data scientist, your task is to *analyze the dataset* and **develop a predictive model** to understand customers' credit risk profiles.

The ultimate objective is to predict whether a client is likely to *default on their next credit card payment* (1) or *not* (0). Accurate predictions can support the institution in improving credit risk management, optimizing lending strategies, and making more informed decisions regarding customer monitoring and intervention.

1 Goal #1: Data Exploration and Feature Engineering

Download the Credit Card Default Dataset from the MyLUISS course homepage and load it into a pandas DataFrame. Refer to Appendix A for a description of the variables.

1.1 Data Exploration

Perform exploratory data analysis to gain insights into the dataset. Generate summary statistics, visualizations, and any relevant plots to understand the distribution of features. At this stage be sure to identify any inconsistencies, missing values, or outliers in the data.

1.2 Feature Engineering

Handle missing values and outliers appropriately, make sure to motivate your decisions. Transform categorical variables using techniques we saw during the lab sessions (e.g., one-hot encoding, ordinal encoding, etc.).

1.3 Feature Augmentation

Create new features that could potentially improve your modelling capabilities by intelligently combining the original features in the dataset. Be creative but remember to justify your choices.

2 Goal #2: Build your classification model!

First of all, split your dataset into training and testing sets.

2.1 Baseline and Training

Start by selecting a suitable baseline approach for your classification task, it will be the core benchmark to understand whether your models are able to outperform a naive strategy. Then, choose three or more classification algorithms (e.g., logistic regression, decision tree, random forest, etc.) possibly belonging to different families of models (e.g., linear models, tree-based models, etc.) and train them on your data.

2.2 Model Evaluation

Evaluate the performance of each of your models using appropriate performance metrics and discuss the results as well as any insights gained from the evaluation. Motivate which metrics you took into consideration and why.

2.3 Model Comparison

Compare the results between the models chosen in Step 2.1. Discuss the differences and, if possible, try to intuitively understand why one performed better than the other ones. Otherwise, discuss on why the results are similar and comparable.

2.4 Fine-tuning

Experiment with hyperparameter tuning to optimize the model's performances, test at least two hyperparameter search strategies (e.g., Grid Search, Genetic Search, etc.). Make a comparison among the non-tuned versions of the models and their tuned counterparts.

Note: do not forget to use a validation set or cross-validation techniques to avoid overfitting and data leakage during the tuning phase.

2.5 Model Selection

Select the best-performing model based on your evaluations and justify your choice. Perform an in depth analysis of the selected model's performance, including any potential limitations or areas for improvement.

2.6 Dimensionality Reduction

Experiment with some of the dimensionality reduction techniques seen in class (e.g., PCA, K-PCA, etc.) to understand whether it is possible to reduce the existing dataset while keeping performances stable (or even increase them).

3 Goal #3: Summarize your findings

Produce a single Jupyter Notebook (.ipynb file) as a result of your work. The Notebook will work both as code and report submission (use "Markdown" cells). The Notebook should run without errors.

Reflect on the challenges faced during all the steps and propose any potential improvements or additional steps that could be taken. The final aim of the homework is not to get perfect classification performances but to simulate a (simplified) real-world data science project, do not worry if

performances are not outstanding, just make sure to carry out all steps properly and with appropriate motivations.

Appendix A: Data Description

Variable	Description
ID	Unique ID of each client
LIMIT_BAL	Amount of given credit (NT dollars)
SEX	Gender
EDUCATION	Education level
MARRIAGE	Marital status
AGE	Age in years
PAY_1	Number of months of payment delay as of September 2005
PAY_2	Number of months of payment delay as of August 2005
PAY_3	Number of months of payment delay as of July 2005
PAY_4	Number of months of payment delay as of June 2005
PAY_5	Number of months of payment delay as of May 2005
PAY_6	Number of months of payment delay as of April 2005
BILL_AMT1	Bill amount in September 2005 (NT dollars)
BILL_AMT2	Bill amount in August 2005 (NT dollars)
BILL_AMT3	Bill amount in July 2005 (NT dollars)
BILL_AMT4	Bill amount in June 2005 (NT dollars)
BILL_AMT5	Bill amount in May 2005 (NT dollars)
BILL_AMT6	Bill amount in April 2005 (NT dollars)
PAY_AMT1	Payment amount in September 2005 (NT dollars)
PAY_AMT2	Payment amount in August 2005 (NT dollars)
PAY_AMT3	Payment amount in July 2005 (NT dollars)
PAY_AMT4	Payment amount in June 2005 (NT dollars)
PAY_AMT5	Payment amount in May 2005 (NT dollars)
PAY_AMT6	Payment amount in April 2005 (NT dollars)
default.payment.next.month	Default payment status next month $\{1, 0\}$

Table 1: Columns description for the Credit Card Default Dataset.