Group Project LZSCC.460
Lent term of academic year 2024/25
Module leader: Dr Christoph Jansen

Group Report: Deadline 28.03.2025, 3pm

Group Presentations: 14.03.2025, 10am-1pm

This group work accounts for 50% of the overall assessment of the module *LZSCC.460 - Data Science Fundamentals* within the MSc Data Science. Each group will consist of three to four members.

This 50% is divided into two components as follows:

- **Group Report (35%):** An article written according to scientific standards, which documents the work and findings of the group in theory and practice. The article should not exceed 6 pages in the LaTeX template provided on Moodle (including figures, excluding references). Appendices containing *only* figures are permitted. At the end of your article it is *mandatory* to include a *contributions statement*, describing the specific contributions each of the group members has made (does not count towards the page limit).
- **Group Presentation (15%):** A presentation of the research strategy and the findings of the group work. The presentation should be self-contained and last 20 minutes. After the presentation, up to 10 minutes are reserved for questions and discussion.

For the report, please submit before the deadline the following two files:

- A pdf file with your typed report, and
- a R file containing the R code used to generate the results in the report.

The code contained in the R file should be commented in detail; this will be included in the assessment. The documents should be named group_x.pdf and group_x.r, respectively, where x is the number of your group. Submission must be done via Moodle (a submission point will be available on the course's Moodle page by then).

AI Category RED: The use of generative AI tools (such as ChatGPT or similar) is not permitted for this assessment.

MARKING CRITERIA

For the Report: (35 marks in total)

- Description of project and background (2 marks)
- Clear statement of research question(s) and objectives (3 marks)
- Research strategy and methods (5 marks)
- Data collection, pre-processing, and integration methods (5 marks)
- Presentation and interpretation of results (5 marks)
- Relation of results and findings to research question(s) (4 marks)
- Discussing and examining potential biases (3 marks)
- Reflections on the approach, and suggested improvements (3 marks)
- Comments in source files that clearly explain the source code (2 marks)
- Demonstration of executable code (3 marks)

For the Presentation: (15 marks in total)

- Articulation of the project's brief and motivation (2 marks)
- Explanation of the research question and objectives (2 marks)
- Research strategy and methods (1 marks)
- Explaining and justifying data processing and integration (2 marks)
- Explanation / rationale of the data analysis techniques used (2 marks)
- Analysis of the results (2 marks)
- Explaining the findings in relation to the research questions (2 marks)
- Appropriate presentation format and delivery (2 marks)

Note: All members of a group will be assigned the same number of marks for both the report and the presentation.

Project: Explanation and prediction of customer churn for a private bank

Background: The (fictitious) *Leipziger Privatbank (LPB)* has experienced strong customer churn in recent years. To investigate possible explanations for this, but also to be able to make predictions as to whether a potential new customer will leave the bank again in the future, LPB's Management Board had all three Leipzig branches compile a data set with data of selected of their current and former customers. These data sets can be found on Moodle. A description of the contained variables is given on the next page.

Objectives: Suppose the bank hires your group as external data scientists to analyze these data sets. The following analysis objectives are agreed on:

- *Pre-Processing, Cleaning, and Identifying Potential Biases:* The data sets should be checked for anomalies (missing data, outliers, errors, duplicates,...) and suitable strategies to deal with these anomalies (if any) should be implemented. Also it should be checked whether it is plausible to assume that the data sets are representative.

- *Exploratory Data Analysis and Visualization:* Appropriate descriptive analyses of the data sets should be carried out with the aim of discovering interesting information and patterns in the data. Results should be illustrated by using appropriate visualization techniques.

- *Statistical Modelling:* An appropriate statistical model should be identified to explain the relation between customer churn and (the) other (of the) variables contained in the data sets provided by the three branches. Based on this model, potential drivers of customers leaving LPB should be identified and (if appropriate) specific recommendations for the Management Board's action should be made.

- *Predictive Modelling:* At least three suitable machine learning models are to be considered and compared in order to implement a model for predicting the future churn of a potential new customer. Comparison of the models should be made by comparing predicitve accuracies of the models with tuned hyperparameters (if hyperparameters are involved).

The *report* should explain in detail how these four objectives have been adressed and what are the findings. The *presentation* should summarize the most important aspects suitable for the Management Board. The *code* should be well-commented and easily adaptable to comparable problems.

Description of the data sets: On the Moodle page for the course, you will find three data sets: `Branch1.csv`, `Branch2.csv` and `Branch3.csv`. Each of the data sets comes (as described on the previous page) from one of the three Leipzig branches of LPB.

Each of the data sets contains the following columns:

- `Customer_ID`: A unique identifier of a current or former customer.
- `Score`: Internal continuous credit score of the customer ranging from 350 to 850. Higher values indicate lower risk in lending.
- `Gender`: Gender of the customer.
- `Age`: Age of the customer (in full years).
- `Tenure`: Number of years the customer has been with the bank.
- `Salary`: Estimated annual salary of the customer in Euro.
- `Balance`: Account balance of the customer in Euro.
- `Products_in_Use`: Number of products (e.g., accounts, loans) the customer has with the bank.
- `Left`: The target variable indicating churn. Binary (1 = Yes, customer churned; 0 = No, customer retained).