Assignment 1: Churn prediction

A large Dutch energy supplier wants to predict which customers are most likely to churn (i.e., leave the company). For this they are asking a group of data scientists (i.e., you!) to develop the best model for churn prediction. To develop such a model, the firm has provided you with a dataset containing data on 20,000 customers. The dataset is balanced, meaning ~50% of the customers have churned within the observed time window. For each customer you have information on the following 14 variables:

1. **Customer_ID:** a unique customer identification number.
2. **Gender:** a dummy variable indicating if the customer who signed the contract is male (0) or female (1).
3. **Age:** the age of the customer in years.
4. **Income:** the monthly income of the customer's household in euros.
5. **Relation_length:** the amount of months the customer has been with the firm.
6. **Contract_length:** the amount of months the customer still has a contract with the firm. Zero means the customer has a flexible contract, i.e., (s)he can leave anytime without paying a fine. If the contract is more than zero months, the customer can still leave, but has to pay a fine when leaving.
7. **Start_channel:** indicating if the contract was filled out by the customer on the firm's website ("Online") or by calling up the firm ("Phone").
8. **Email_list:** indicating if the customer's email address is known by the firm (1=yes, 0=no).
9. **Home_age:** the age of the home of the customer in years.
10. **Home_label:** energy label of the home of the customer, ranging from A (good) to G (bad).
11. **Electricity_usage:** the yearly electricity usage in kWh.
12. **Gas_usage:** the yearly gas usage in cubic meters.
13. **Province:** the province where the customer is living.
14. **Churn:** a dummy variable indicating if the customer has churned (1) or not (0).

For the assignment, go through the model development steps, namely:

1. Defining the business problem
    - Can be very short here in the assignment
2. Design the research: formulate hypotheses, literature research, defining constructs
    - Based on literature, hypothesize about the most important variables and how they relate to the dependent variable (e.g., direction of the effect and shape of the relationship). This can be the basis for the baseline model! This part is maximum 2 pages long! (it is a very important step, but not the core of this course)
3. Data preparation: data cleaning and data transformation, including creating new variables based on the research design
    - Do check if everything is what you would expect and if it makes sense to do transformations and/or if there are problems in the dataset which need to be dealt with!
4. Explorative analysis: including correlations, statistical tests, histograms, etc.

- Provide descriptive information about the data, including summary statistics and graphs. Is there model free evidence of your hypotheses? Does this provide additional information about the structure of the data and what type of model specification is appropriate?

5. Modelling: Model specification (e.g., type and structure), Model estimation, Model validation, Model interpretation
    - Estimate a baseline model, based on your hypotheses. Estimate next to this at least 5 different models (e.g., (step-wise) logistic regression, trees, SVM). You are allowed to estimate different variations within each model (e.g., with different parameters and/or restrictions set). Explain your decision process and interpret the models. Validate the models by comparing the out-of-sample predictions based on the discussed validation criteria (i.e., hit-rate, top-decile lift, and Gini).

The written assignment is **maximum 15 pages long** (excl. title page, table of content, reference list, and AI-usage statement, incl. everything else (i.e., no appendices)!) and should consist out of:

- Title page (including title, group number, group members, student numbers)
- Table of content
- Managerial summary (1 page)
- Main paper, with the five steps discussed above
- Managerial conclusion (including recommendation for which model to implement)

For the assignment, use the following (standard) formatting:

- Font: Times New Roman or Calibri
- Font size: 12 points
- Spacing: Single
- Standard margins

Next to this, you have to hand in the R-script, with which all the results from the written assignment can be replicated!

Upload the written assignment (.pdf) and the R-script before the 5th of December 2024 at 17:00! Use the following file names: "assignment_1_group_x.pdf" and "assignment_1_group_x.R", where x is the sub-group number.

Rules regarding the use of Artificial Intelligence

- You are allowed to use AI for spelling, grammar, and rewriting.
- The base of the assignment should be written by you yourselves, but you are allowed to use AI (e.g., ChatGPT) for copy editing.
- You should write the R-code yourselves, but you are allowed to ask AI to improve the code, how to make (for instance) graphs, how to resolves issues/errors or want to better understand the output.
- You should, however, in the end understand all the code and the output yourselves!
- You are responsible for all content of the assignment.

You should add to the assignment a statement if you did or did not use AI, which AI (version) and for what purpose. This can be on the last page of the assignment and doesn't count towards the page limit.

Examples of such statements:

- We did not make use of AI to make this assignment.
- We used AI to copy edit the assignment, i.e., we did use it for checking spelling, grammer and rewriting. For this, we used ChatGPT 4o.
- We used AI to copy edit the assignment, i.e., we did use it for checking spelling, grammer and rewriting. We also used AI to improve our R-code and generate Figure 2. For all of this, we used ChatGPT 4o.

Below this, please put a signature of all group members!

Evaluation Form Assignment 1: Churn prediction

Group number:

Student name	Student number

	Remarks	**Score**
Research design (10%)		
Data preparation (10%)		
Explorative analysis (15%)		
Modeling (50%)		
Managerial conclusion, writing and reporting (15%)		
Overall grade		

Score: -- bad, - poor, +/- okay, + good, ++ very good