Business Analytics II first course assignment

The assignment should be done individually, the deadline is the 27th of October, 23:59. The assignment worth 30 points in total; in order to pass this assignment you need at least 15 points. You can find the required data files in Moodle and in the CSC platform, in the shared folder. You have to submit a report according to the specifications below, and a Jupyter notebook with the code that you used to solve the tasks that require working with the data (download the notebook from the CSC platform and upload it in Moodle).

Problem specification

In this assignment, you will apply the CRISP-DM framework to build a predictive model for customer churn in a subscription-based business. Predicting customer churn is a critical task for businesses that rely on recurring revenue, as losing customers can significantly impact profitability. Your goal is to explore how customer data, such as demographics and usage behavior, can be used to anticipate which customers are at risk of leaving. By following the steps of CRISP-DM, you will clean and prepare the data, build and compare predictive models, and discuss how these insights could guide business decisions to improve customer retention. You are required to prepare a comprehensive report by systematically following all the steps of the CRISP-DM framework. For each stage, address the specific questions provided.

In Moodle, you will find three academic articles that provide background information on the domain and will help you address the questions outlined below. You are free to identify and explore additional resources if needed, but please ensure that any extra references you use are appropriately cited in your report. The dataset and the specification of the columns can be found in the CSC platform, in the shared folder Assignment 1.

When writing the report, assume the role of a data scientist working within the company. Present the process of generating predictions and deriving actionable insights from the data. Structure your report as if you are explaining the entire analytical workflow—from understanding the business problem to modeling and evaluating predictions—to stakeholders within the organization.

Business Understanding

- What is customer churn, and why is predicting it important for subscription-based businesses?
- What is the potential impact of reducing churn on a company's revenue and customer retention strategy?
- What could be the specific business objectives? (e.g., Reduce churn by X%, increase retention in a particular customer segment, etc.)
- How can churn prediction affect marketing and customer service strategies?

Data Understanding

As part of this step to answer the questions, perform descriptive analysis in Python: calculate basic statistics (e.g., mean, standard deviation, correlation), create some basic visualizations (histogram, boxplots etc.) focusing on the outcome variable. Before this, you will have to perform one-hot encoding for categorical variables.

- What data is available to predict churn (e.g., customer demographics, subscription details, usage behavior, etc.)?
- What are the key features or variables that might affect customer churn?
- Are there any missing values or outliers in the dataset that need attention?
- What relationships can you observe between features and the target variable (churn)?
- Can you identify any early insights?

Data Preparation

As part of this step to answer the questions, perform data preparation in Python.

- How will you handle missing data or outliers?
- Do you need to normalize/standardize any of the features?
- What methods will you use to split the dataset into training and testing sets?
- Should any features be excluded based on relevance or redundancy?

Model Building

As part of this step to answer the questions, build machine learning models introduced in the course to predict customer churn.

- What measure(s) will you use to evaluate the performance of the models?
- Create a baseline model as a logistic regression without any parameter optimization.
- Create various tree-based models, also experiment with parameter optimisation: (i) decision trees, (ii) bagging, and (iii) random forest classifiers.

Evaluation

- How did the models perform overall? What were the strengths and weaknesses of each model?
- Which model performed best based on the performance evaluation metrics you chose, and why do you think it outperformed the others?
- Would the model(s) generalize well to new, unseen data?
- What are the four most important predictors according to the best decision tree model?

Deployment

- How would you implement this model in a real business environment?
- What steps would the company need to take to integrate this model into their customer management system?
- How could the model be used for targeted interventions (e.g., offering discounts to customers likely to churn)?
- What potential risks or limitations could arise during deployment (e.g., overfitting, ethical concerns, data drift)?