**Final Project**

**This is a group project. You will work with your team and deliver on CANVAS**

1. **one report on behalf of your group**
2. **one R-Studio file of your codes**

**German Credit Project**

*GermanCredit.csv* is the dataset for this case study.

**Background**

Money-lending has been around since the advent of money; it is perhaps the world’s second-oldest profession. The systematic evaluation of credit risk, though, is a relatively recent arrival, and lending was largely based on reputation and very incomplete data. Thomas Jefferson, the third President of the United States, was in debt throughout his life and unreliable in his debt payments, yet people continued to lend him money. It wasn’t until the beginning of the 20th century that the Retail Credit Company was founded to share information about credit. That company is now Equifax, one of the big three credit scoring agencies (the other two are Transunion and Experion).

Individual and local human judgment are now largely irrelevant to the credit reporting process. Credit agencies and other big financial institutions extending credit at the retail level collect huge amounts of data to predict whether defaults or other adverse events will occur, based on numerous customer and transaction information.

**Data**

This case deals with an early stage of the historical transition to predictive modeling, in which humans were employed to label records as either good or poor credit. The German Credit dataset has 30 variables and 1000 records, each record being a prior applicant for credit. Each applicant was rated as “good credit” (700 cases) or “bad credit” (300 cases). Table 1 shows the values of these variables for the first four records. All the variables are explained in Table 2. New applicants for credit can also be evaluated on these 30 predictor variables and classified as a good or a bad credit risk based on the predictor values.

The consequences of misclassification have been assessed as follows: The costs of a false positive (incorrectly saying that an applicant is a good credit risk) outweigh the benefits of a true positive (correctly saying that an applicant is a good credit risk) by a factor of 5. This is summarized in Table 3. The opportunity cost table was derived from the average net profit per loan as shown in Table 4. Because decision makers are used to thinking of their decision in terms of net profits, we use these tables in assessing the performance of the various models.

Table : First four records from German Credit dataset

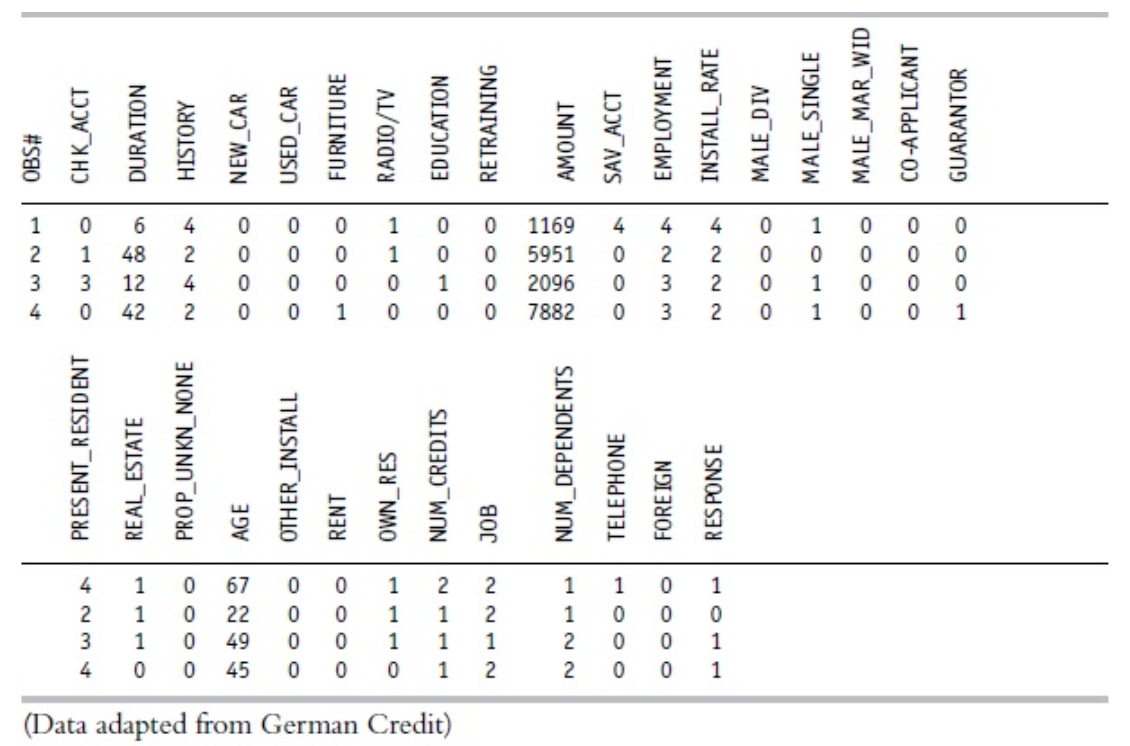


Table : Variables for the German Credit Dataset



The original dataset had a number of categorical variables, some of which were transformed into a series of binary variables and some ordered categorical variables were left as is, to be treated as numerical. (Data adapted from German Credit)

Table : Opportunity Cost Table (Deutsche Marks)

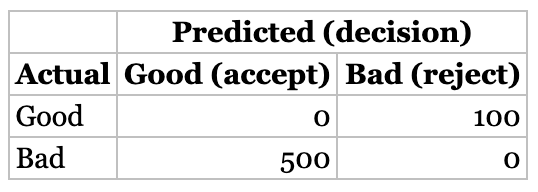
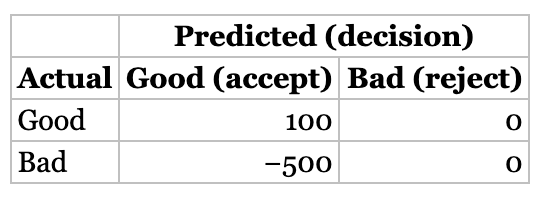


Table : Average Net Profit (Deutsche Marks)



### Assignment

1. Review the predictor variables (descriptive statistics) and guess what their role in a credit decision might be. Are there any surprises in the data?
2. Partition the dataset into 60% training and 40% validation (set the seed to 12345). Develop at least two classification models of your choice. Describe the two models that you chose, with sufficient details (method, parameters variables, etc.) so that it can be replicated.
3. Choose one model from each technique and report the confusion matrix and the cost/gain matrix for the validation data. Which technique has the highest net profit?