**BUDT733 - Homework 6  
Quantitative analysis of credit (contd.)**

In this Assignment we will finish our analysis of the Credit data. We use the same data, and the goal remains the same: to create a model that a bank or another financial institution can use to classify a new credit request as accept/not accept (CREDIT\_EXTENDED should be excluded from the analysis).

**Data Preparation**

1. Open the file credit3.xlsx. Create the outcome variable (PROFITABLE=1 if NPV>0, =0 otherwise), create factors for CHK\_ACCT, SAV\_ACCT, HISTORY, JOB and TYPE variables. Split the data using the sample function; 70% as training data and 30% as test data; setting the seed to 12345. (Do not use NPV as a predictor)

**Predicting profitable accounts with Classification Trees**

1. Run the Classification Tree algorithm using the data, with the PROFITABLE as the output variable. Set the seed to 123 and then use K-fold cross-validation (with K = 10) to prune back the tree. Attach the classification confusion matrix for the test data as **Exhibit 1** and a figure of the pruned tree as **Exhibit 2**.
2. How many decision nodes are in the full tree (using R default values)? \_\_\_\_\_\_\_\_\_\_\_\_\_\_

How many decision nodes are in the pruned tree? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Which model (the full or pruned tree) gives you better accuracy? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. How would the tree classify our student from the previous HW (the student is 27 years old, domestic, has $100 in her checking account but no savings account. The applicant has 1 existing credits, and a credit duration of 12 months, and the credit was paid back duly. The applicant has been renting her current place for less than 12 months, does not own any real estate, just started graduate school (the present employment variable is set to 1 and nature of job to 2). The applicant has no dependents and no guarantor. The applicant wants to buy a used car and has requested $4,500 in credit, and therefore the Installment rate is quite high or 2.5%, however the applicant does not have other installment plan credits. Finally, the applicant has a phone in her name)

Profitable / Not Profitable (pick one)

What is the predicted probability that the account is profitable?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Find the best pruned tree with only 4 terminal nodes. Describe the rule in words (in English)

**Predicting profit with Regression Trees**

1. Reset the seed to 123 and run a Regression Tree Algorithm to predict the NPV of each applicant. Use a pruned tree to score the data samples. Attach the pruned tree as **Exhibit 3**.
2. In the output for the test sample, the prediction for each node corresponds to the average NPV of all training records in that end-node. Therefore based on the training data we would extend credit to all requests with a positive predicted NPV.   
     
   Score the test data (i.e. compute predicted NPV), create a table that summarizes the number of records from the test data in each end node (each end-node has a distinct prediction value), and the total actual NPV of the test records in these nodes. Attached the table as **Exhibit 4**.

Based on your table and the predicted NPV values, how many customers in the test sample would you extend credit to?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

What would be the average profit per customer (that you extend credit to)?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

What is the overall profit for all customers you extend credit to in the test sample?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

How do these values compare with extending credit to everyone?

1. Compare the pruned classification tree to the pruned regression tree (Exhibits 2 and 3). These two trees are an indicator of what are some of the more important variables when classifying a profitable account and predicting the profit of an account. In what way are these trees similar/dissimilar? Briefly discuss.

**Selecting the “right” customers with multiple linear regression**

1. Build a linear multiple regression model to predict NPV. Attach **Exhibit 5** that neatly shows the variables and their corresponding coefficients. Use the training sample to select a cut-off value, which maximizes the overall profit (if the predicted NPV is above the cut-off the bank will extend credit, if the predicted NPV is below the cut-off value, the bank should decline the credit request).

HINT: Sort your training sample by the predicted value. Sum up the actual values for each possible cut-off value. Select the cut-off that results in the maximum sum.

What is your optimal cut-off value?

\_\_\_\_\_\_\_\_\_\_

1. Apply the cut-off value to the test sample. How many customers in the test sample would you extend credit to?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

What would be the average profit per customer (that you extend credit to)?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

What is the overall profit for all customers you extend credit to in the test sample?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Bagging, Random Forest and Boosting**

1. Use boosting, random forest and bagging to examine if performance of your classifier above can be improved. Discuss improvements (or not) both in terms of accuracy and profitability.