*BUDT758T - Homework 5***Quantitative analysis of credit**

*Please include a carefully annotated R script/markdown file with your submission.*

The Great Recession was due in a sizeable degree to the process of extending credit to people who defaulted on their loans (typically mortgages for their houses) as they were not able to repay them. Combined with decreasing real estate prices, many of the institutions that extended the loans ended up owning property that has decreased in value, and therefore lost significant amount of money.

In the spreadsheet credit3.xls under the tab “Data”, you will find Data pertaining to 1000 personal loan accounts at a bank. The tab “Data Dictionary” contains a description of what the various variables mean.

When a new applicant applies for credit, as a part of the application, the company collects information which is available in the form of Variables 2 to 21. The company then decides an amount to be credited (the variable CREDIT\_EXTENDED.) For these 1000 accounts, we also have information on how profitable did each account turn out to be (variable NPV). A negative value indicates a net loss and this typically happens when the debtor defaults on his/her payments.

The goal in this case is to investigate how one can use this data to better manage the bank's credit extension program. **Specifically, our goal is to develop a classification regression model to classify a new account as “profitable” or “not profitable”.**

**IMPORTANT: In this assignment it will be important for categorical predictors to be numerical dummy variables for k-NN (so that distances can be computed). For Naïve Bayes, you will need categorical variables to be factors (otherwise the algorithm will fit a Normal distribution to the data for the conditional probabilities).**

1. Carefully review the information in the Data Dictionary. Which variable if any cannot be used to classify an account performance?  
     
   Hint: If the model is built and deployed, it will be used to make credit decisions for new applicants. What information would be unavailable for such individuals?
2. The goal is to use classification methods to predict whether or not a new credit will result in a profitable account. Create a new categorical variable to use as the dependent variable in the model.
3. Starting with the original data set, create dummy variables for CHK\_ACCT, SAV\_ACCT, HISTORY, JOB and TYPE (*other categorical variables are ordered categorical and should be viewed as numerical*). Split the sample into training (70%) and validation (30%) samples with the seed set at 12345.   
     
   *Do not exclude the base category dummy variable from your analysis for k-NN and Naïve Bayes.*

**Predicting Profitability using k-nearest neighbors**

We will now conduct the analysis using the k-nearest neighbor algorithm.

1. Run the k-NN algorithm for classification, testing all values of *k* from 1 to 15, selecting to score the data on the best *k* (remember to normalize the data).
2. Using the output, plot the %Error on both the Training sample and the Validation sample. Include the plot as **Exhibit 4**.   
     
   What is the best value of *k*?

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

Briefly explain why the % Error is zero for the training sample when *k*=1, but not for the validation sample.

1. What is the error rate for the 2 classes on the validation sample?

Class 1:\_\_\_\_\_\_\_\_\_\_, class 0: \_\_\_\_\_\_\_\_\_\_

1. Repeat step 4 for seed values in the range 1:10 and determine your best *k*. What do you find?

**Predicting profitable accounts with Naïve Bayes**

1. Suppose the expected profit of a profitable account is around $500 and the expected profit of an unprofitable account is -$1500. If a random customer walked in and requested a loan, using only the information about the fraction of profitable accounts in the training data, would you extend credit to the customer (your only consideration is the expected profit)?

Yes / no (circle one)

Briefly explain:

1. Create the outcome variable (Y=1 if NPV>0, Y=0 otherwise) as before. Other than AGE, DURATION, and AMOUNT\_REQUESTED, convert all variables into factors. Split the data using 70% as training data and 30% as validation data, setting the seed = 12345.
2. Run a Naïve Bayes classification algorithm using the data. Attach the classification confusion matrix for the validation sample as **Exhibit 5** and The Conditional Probabilities as **Exhibit 6**.   
   (*Do not convert categorical variables to dummy variables*.)
3. Based on your model, how would you classify an applicant that is a 27 years old domestic student, 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, 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. Fit a Logistic Regression Model with all the relevant variables. Specifically, include: a constant term, AGE, CHK\_ACCT (factor), SAV\_ACCT (factor), NUM\_CREDITS, DURATION, HISTORY (factor), PRESENT\_RESIDENT, EMPLOYMENT, JOB (factor), NUM\_DEPENDENTS, RENT, INSTALL\_RATE, GUARANTOR, OTHER\_INSTALL, OWN\_RES, TELEPHONE, FOREIGN, REAL\_ESTATE, TYPE (factor), AMOUNT\_REQUESTED. [Note that variables that are *ordered* categorical variables are being treated as numerical.] Attach the regression output and confusion matrix of the validation data set as **Exhibit 7**.
2. You have now built three different classification models – the logistic regression model, the naïve Bayes model, and the best k-nearest neighbor model. Which model has the best prediction performance? Discuss using ROC curves for the three (on the same graph).