**Quantitative analysis of credit**

The recent slowdown of the US economy is 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. [1 point] 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. [no credit] 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 (e.g. a variable PROFITABLE which is 1 if NPV > 0 and 0 otherwise)
3. [no credit] 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.   
     
   You will need a distinct dummy variable for each possible value of these variables, e.g. create a new variable CHK\_ACCT0 = 1 if CHK\_ACCT = 0 and CHK\_ACCT0 = 0 otherwise.  
     
   *Do not exclude the base/reference 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. [4 point] 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).

What is the best value of *k*? \_\_\_\_\_\_\_\_\_\_\_\_\_\_

Attach the confusion matrices for both the training and the validation data sets as **Exhibit 1**.  
  
  
  
What is the validation error rate at the best value of *k*? \_\_\_\_\_\_\_\_\_\_\_\_\_\_

Using the output, plot the %Error on both the Training sample and the Validation sample.   
Include the plot as **Exhibit 2**.

1. [1 point] Briefly explain why the % Error is zero for the training sample when *k*=1, but not for the validation sample.
2. [2 point] What is the error rate for the 2 classes on the validation sample?

Class 1 (1-Sensitivity): \_\_\_\_\_\_\_\_\_\_, Class 0 (1-Specificity): \_\_\_\_\_\_\_\_\_\_

1. [1 point] Repeat step (4) above for data partitions corresponding to seed values in the range 1:10 and determine your best *k*. What do you find?