*Data Mining & Predictive Analytics*

**Homework 7**

This assignment is an extension of the exercise on analyzing credit. You will recall that we used the data to try and predict which applicants will have non-profitable loans. In this assignment we will use association rules and cluster analysis with the same data.

Recall that in the *Credit* spreadsheet under the “Data” tab, 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.

The bank has been making losses on this category of loans and the broader purpose of the study is to investigate how one can use this data to better manage the bank's credit extension program. Specifically, our goal is to use association rules to determine what features of applicants are associated with a *new* account being profitable or not profitable. We will also use cluster analysis to examine whether we can build a small number of profiles of applicants for loans. The latter would be of interest if the bank wanted to develop a better understanding of who its customers are, and what features lead to unprofitable loans.

For reasons discussed before, continue to exclude CREDIT\_EXTENDED from your analysis. Set the seed to 12345.

**Part 1: Cluster Analysis**

1. Create dummy variables for all categorical variables that are not ordered (CHK\_ACCT, SAV\_ACCT, HISTORY, JOB and TYPE).
2. Perform a K-Means clustering of the data with *K* = 5 using all of the data *except* NPV and the created PROFITABLE dummy variable. Include the table of cluster centers as **Exhibit 1.** Use 20 random starts to ensure that you have good clusters.
3. Comment on the profiles developed? Can they be clearly identified using meaningful labels? Try and describe each cluster in words.
4. The output allows you to identify, for each individual, the cluster they belong to. Combine this with the NPV column from the original data (making sure that the Row Id matches).
   1. Create a bar chart showing the percentage of people in each cluster.
   2. Create a table showing average of NPV split up by Cluster Id.
   3. Attach these as **Exhibit 2.**
5. Comment on your results above.

1. Experiment with two other values of *K* (4 and 6) and decide which set of clusters are most meaningful to you.

**Part 2: Association Rules**

In the second part of the assignment you are to look for good rules that can be used to better guide the extension of loans. I will leave this part of the study more open-ended (since, in real world settings, projects do not come in the form of well formulated exercises). Remember that for association rules all variables have to be binary, so (to save you some tedious work) you may restrict your analysis to the following list of variables: CHK\_ACCT, SAV\_ACCT, HISTORY, EMPLOYMENT, OWN\_RES, JOB and PROFITABLE. It is generally necessary to try out a few different values of support and confidence and to examine the resulting rules.

1. Conduct an association rule analysis of the data to identify attributes that are related to profitability. Attach any supporting output as **Exhibit 3.** [One approach would be to include the PROFITABLE dummy variable and focus on rules for which this variable is the consequent.]
2. List five good rules and describe their properties (in terms of criteria discussed in class). Describe the rules in words.

1. Suppose you were to use the best rule identified above as a guide whether to extend loans or not for the cases in your data set. What would the total profits be?

**Part 3: Combining Cluster Analysis and Association Rules**

1. Suppose you were to only use cluster membership (and PROFITABLE) to define association rules. Does this lead to good rules? How does the profitability of these rules compare with the profitability of rules that use the more detailed information?