Homework 1

(Due Wednesday, May 29, by 6:00pm in ICON Dropbox)

1. The file *SERU.csv* contains data on university students, some of whom filled out the Student Experience in the Research University (SERU) survey. The university would like to be able to predict which students are most likely to complete the survey, and the features that drive this prediction. The file *DataDictionarySERU.pdf* contains a description of the variables in *SERU.csv*.

   Partition the data set using the default 70/15/15 splits into training/validation/test sets. Each student should partition the data set using the last three digits of their university ID # (this will differentiate each student's training set). Please document your seed in your report.

   Using the Rattle methods discussed in class, lecture materials, and Chapters 3-5 of the DMRR text, as well as any additional R functions you find useful, perform a descriptive analysis of all the features in the training set.

   Submit a document that includes plots along with brief, precise textual descriptions of every conclusion you reach regarding the variables, or anything curious that you believe needs further examination. You should also consider any data transformations that you believe would be useful, although you are not required to perform them (yet). Are there columns which should be removed or combined? What else do you think should be measured to best predict the outcome of interest (whether a student completes survey)? You will be graded on the accuracy and completeness of your analysis as well as the clarity of the report.

2. The file *churn_imbalanced.csv* contains data on a collection of customers for a cell phone company, some of which have discontinued their service subscriptions and others who have remained customers. For this analysis, we are interested in understanding the characteristics of the customers who leave. Focus on these cases by first removing the negative cases (result = STAY). Partition the resulting data set using the default 70/15/15 splits into training/validation/test sets. Each student should partition the data set using the last three digits of their university ID # (this will differentiate each student's training set). Please document your seed in your report.

   Using the training set, cluster the data using *k*-means clustering. Do not use the class labels as a feature. You will need to choose an appropriate value for the parameter *k* (the number of clusters). You will need to justify this choice quantitatively. Document the approach you took for this choice and justify your conclusion, graphically if possible.

   You should next interpret your clusters by providing a short description of what makes each cluster "special," that is, what feature values make it different from the other clusters, or different from the general trends in the data.

Show a scatter plot of as many pairs of dimensions as you can (using the "Data" button). Which dimension pairs show a good separation of the clusters? Describe what you can learn from these plots. You will be graded on the accuracy and completeness of your analysis as well as the clarity of the report.