**Springboard Data Science Introduction in R**

Project Narrative

The goal of the capstone project was to predict the value of the column income50. The context of this column is a value the reflects incomes less than $50,000 or greater than or equal to that amount. Below is a list of columns in the data that were utilized to predict the variable:

* Age – continuous variable
* Workclass – categorial factor
* Fnlwwgt - continuous
* Education - categorial factor
* education\_num - continuous
* marital\_status - categorial factor
* occupation - categorial factor
* relationship - categorial factor
* race - categorial factor
* sex - categorial factor
* capital\_gain -
* capital\_loss -
* hours\_per\_week -
* native\_country - categorial factor
* income50 - categorial factor and predictor variable

The dataset adult\_data from the UCI Machine Learning Repository was used as input into a data frame.

Statics Recap

One of the challenges in preparing the data for modeling was to identify missing values. Utilizing ggplot2 bar charts for categorial variables, the columns of data from the data frame was identified has having missing values in the form of “ ?”. After replacing the “ ?” values with NA, k-nearest neighbor imputation function was implemented to select the best choice to replace the missing columns values.

After replacing missing values through imputation, a preprocessing model predictive variable model was implemented that:

* Centered the predictive data based on the mean
* Scaled the data based on the standard deviation
* Identified numeric predictor columns with a single value (nz)
* Excluded near zero values predictors

It should be noted that the capital gain and loss columns are sparsely populated.

Machine Learning Recap

A Support Vector Machine (SVM) model was developed to predict the variable income50. The svm model was tuned with a multitude of cost and gamma values in order to avoid overfitting and to correctly predict the data point income50. The following svm command was utilized for tuning along with optimized cost and gamma values:

* tune(svm, train.x = trainData[,-13], train.y=trainData$income50, kernel="radial", ranges=list(cost=exp(c(-1,0,3,5,7,9)), gamma=exp(c(-1,0,3,5,7,9))))
* cost = 1, gamma = 0.3678794 (i.e. recommended optimal values)

It should be noted that the capital gain and loss columns are sparsely populated. During the predictive modeling process, these two columns were removed by machine learning processes.