**RTI Exercise 1**

Edward Lim

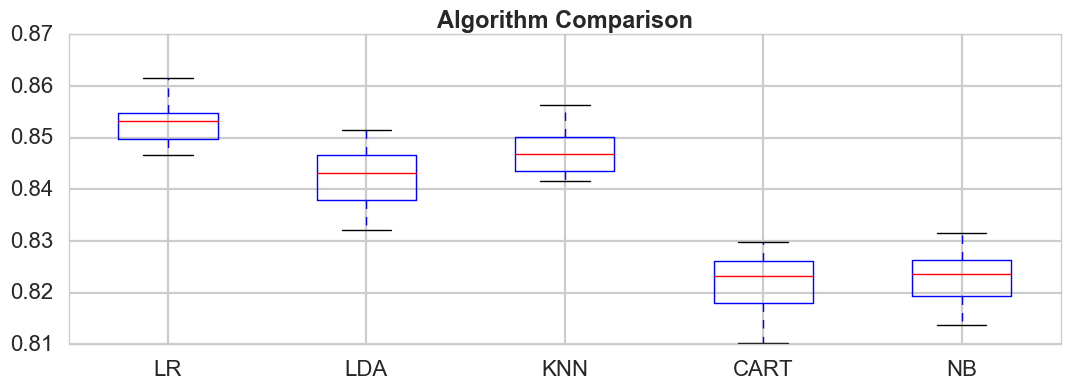
For all tasks from SQL, Exploratory Data Analysis, and Model Building, I used Python 2.7 and created two Jupyter notebooks that guide through all the code, analysis, visualizations, and explanations. They can be found on <https://github.com/doehun/data-scientist-exercise01>

**Part 1. Exploratory Analysis.ipynb**

Brief Summary: Variable distributions, interactions, and aggregations over different factors (counts, proportions, and averages) were explored in this section.

**Part 2. Model Building.ipynb**

Brief Summary: The variable *education level* was dropped because it had perfect correlation with *education number.* Using education number as a continuous feature was more computationally efficient and yielded higher accuracy in models as well.

The data was randomly shuffled and split 80/20 for training and testing. 10 fold cross validations were performed on training using Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbor, Decision Tree, and Naive Bayes. The accuracies for 10 fold cross validations were recorded and plotted. 

Logistic regression had the highest accuracy and the lowest variability of cross validation accuracy scores. On the test data set, the Logistic Regression model yielded about the same accuracy on the test set (~84%) and an AUC score of .902. Ensemble models did not yield significant improvement in prediction. Logistic Regression was chosen as the final model because it was the simplest and most interpretable model that could be explained to a nontechnical client.