Data Scientist Exercise – 01

Kasey Jones

A US Census dataset named Records containing 48,842 observations and 15 variables was analyzed and modeled to determine if a person earned over $50,000. The data was read in from an SQLite database using SQL and R, and all variables that contained ID numbers were transformed to categorical variables using the variable’s corresponding reference table. Education Number and Education Level were found to provide the same information and Education Level was dropped. There were concerns about the amount of perfect separation that occurred in the data. For example, all observations that had a Capital Gain of $15,024 made above $50,000 (513 records), while those with a Capital Gain of $5,013 made less than $50,000 (117 records). Variables that had clear problems with perfect separation had their values grouped together based on logical reasoning. This included Country, Education Number and Work Class, although more work needs to be done in the area.

The data was then split into training (60%), validation (20%), and testing (20%) datasets. The logistic regression model created on the training data correctly identified if a person made more than $50,000 84.7% of the time when tested on the testing dataset. Backwards selection was implemented, but it did not improve the model. Other modeling techniques such as Random Forests, and Neural Networks were attempted and even provided slightly better accuracy. However, the top model created was an xgboost model, and it had an accuracy of 87.3%. The parameters of the model were tuned using a large grid search.

All variables in the dataset improved the accuracy of the models, but some variables stood out. Consider the variable Relationship. If we only had this variable, we could achieve 93.5% accuracy for all observations not labeled “Husband” or “Wife” by guessing that the person did not make more than $50,000. This is 51% of the data. Below is a variable importance chart that resulted from running the Random Forest model. Values closer to 1 are more important, and values closer to 0 are less important.