VJ Davey

4/10/17

Big Data

Assignment 4 - Statement of Work

When I went about creating my models and prediction vectors for this assignment, I began by taking a raw look at the training data in the “insurance-training-set.csv” file. I found that the columns which contained missing values were AGE, YOJ, INCOME, HOME\_VAL, and CAR\_AGE. Following this, I considered the proportion and actual amount of data missing for each of these columns. All of these columns, with the exception of AGE, seemed to have about the same proportion and amount of data missing. Due to many values being repeated in AGE and the possibility of there being a “peaking” effect with driving ability and age, I decided it would be best to bin the values by 5 year gaps to turn it into a factor, and then create an extra level for the NA values.

Next, I went about doing research on data imputation in R. After some internet searching, I came across a paper from Columbia University on data imputation in addition to an R package called MICE, which offers data imputation functions. Using the MICE package, I further examined distribution statistics on the columns with missing data. I settled on using random forest based imputation in MICE due to the results of the function closely taking on the shape of the distribution of the non-missing values for all columns with missing data. After this, I saved the imputation prediction model to a variable so that I would be able to work with future datasets of the same structure missing the same columns.

Third, I did a bit of unsupervised learning to get an even closer look at the structure of the training data. I performed hierarchical clustering to get a fresh look at things, see which drivers clustered together, and see what traits they shared. I found that among clusters heavily populated with drivers who had crashed, that among drivers who crashed there was a tendency for traits including lower education levels, living in an urban area, having prior claims, and having omitted information in certain columns. I also found that among clusters which included large batches of drivers who did not crash, that there were still noticeable numbers of drivers who did crash, meaning there would probably be no absolute foolproof way to classify crashers vs non-crashers with the data available. So, while it would be unlikely that I develop a model that is incredibly accurate, I could at the very least be happy with a model that outperforms simply always choosing one class or another. Next, I did some automated preprocessing with the CARET package using the rfe() function to discover if certain columns in the data were correlated or possibly not needed for supervised learning model construction. I then broke my training data into training and validation subsets.

Lastly, I went about constructing the models. I ended up creating two logistic models to predict TARGET\_FLAG probability and class, and one random forest model to predict TARGET\_AMT, including TARGET\_FLAG as stated in the assignment pdf. The classification error rate for the logistic model predicting TARGET\_FLAG was approximately 20%. The mean absolute error for the random forest model predicting TARGET\_AMT was under $1800. I developed a confidence in these models by comparing their output to several other quick models I learned how to draft up in rattle recently in Murray’s class. These models seemed to perform at around the same level or better than competing methods including neural nets and support vector machines, so I chose to stick with them. I then used the models to make predictions on the data in the “insurance-test-set.csv” file. I compared the associated test predictor values along with my predictions to my observations from a hierarchical clustering, and found that the results seemed to be consistent with my earlier findings during the clustering stage.