M3: Data Assignment

Jason Merten

Data Exploration/Preparation:

The insurance data sets used in this analysis contains 25 columns of data across a training and a test data set. The training data set consists of 8161 rows of data while the test data set consists of 2141 rows of data. After adjusting the variable types, there are 6 columns that contain missing values: **AGE**, **YOJ**, **INCOME**, **HOME_VAL**, **OLDCLAIM**, and **CAR_AGE**. The majority of the variables within the data sets appear to follow some sort of statistical distribution (normal, beta, etc.) and a few have some outliers, though nothing extreme (less than 1%). All outliers were smoothed to the 99th and 1st quantile values for all numerical variables. While handling missing values, I decided to create 4 new variables in each data set (**HOME_OWNER**, **SQRT_TRAVTIME**, **SQRT_BLUEBOOK**, and **INCOME_bin**) to help with model accuracy.

Building Models:

When creating the models, I decided to try a varying array of methods from logistic regression to K-nearest neighbors. The first three models were developed using logistic regression, Model1 and Model1.1 were created using backwards selection and forwards selection including interaction terms. Both models produced fairly good results but took a long time to comb through all of the variables and remove/select the appropriate variables.

Model1.2 was created using a stepwise function from the base logistic regression function using forward and backward selection. While the stepwise function took a while to complete, it came up with a very good model for the data given.

Model2 was created using Linear Discriminant Analysis with the same variables and interaction terms from Model1.1. The difficulty with this model was how to document the results of the prediction versus the logistic regression model. I ended up using the posterior probabilities for predicting a 1 as the results.

Model3 was created using Quadratic Discriminant Analysis using a few different variables that I believed to be potentially useful for predicting the target variable. As with Model2, the results of the predict() function weren't the same as the logistic regression models, so I ended up using the posterior probabilities as the results.

Model4 to Model1.2 were created using K-nearest neighbors to see if I could get a better result from the previous models. I had to split the training data (80/20) into a new training/test data sets to be able to use KNN. I ultimately decided to not use them as part of the graded models because the results don't match what is required for the submission. I did end up getting an estimated 36% improvement in accuracy from the initial test data set.

Selecting a Model:

After finishing the models, I ran them through AIC, BIC, logLik, ROC, AUC, and ks_stat tests to determine the best model that fit the data. The only models that I could analyze with the AIC, BIC, and logLik tests were the logistic regression models (Model1-Model1.2). The other two models were primarily tested using the ROC, AUC, and ks_stat tests. I ended up selecting Model1.2 as the champion model because it had the lowest AIC, logLik, and AUC values and performed well overall compared to the other models.