M5: 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. I also renamed the blank **JOB** level name to 'None' to help reduce confusion.

Building Models:

I created three models for this assignment:

Model1 was created using the same method as the champion model from Module 3 (stepwise AIC selection in both directions) but using a 80/20 train/test split and a few interaction terms. Compared to the champion model, the AIC and BIC values are significantly better. The accuracy of this model against the test data set was 80.21%.

Model1.1 was created using the train function from the caret package with the glm function as the method and a 10-fold cross-validation against pre-processed data to center and scale the data around 0. The accuracy of this model against the test data set was 80.33%, slightly better than Model1.

Model2 was created using Linear Discriminant Analysis with the same variables and interaction terms from Model1. The accuracy of this model against the test data set was 80.7%, again slightly better than the two previous models.

Selecting a Model:

To analyze the performance of the models I created, I primarily relied on ROC curves and their AUC values. As an additional measure of performance, I looked at the confusion matrix for each model and noted their metrics. The AUC values for Model 1 and Model 1.1 were identical, with Model 2 closely behind. I decided to select Model 1.1 as the champion model for this module due to its high AUC value, high accuracy and sensitivity, and acceptably low specificity compared to the other models. A table with all the performance metrics gathered is below:

Model/Metric	AUC	Accuracy	Sensitivity	Specificity
Model 1	0.818	.7947	.9283	.4236
Model 1.1	0.818	.796	.9300	.4236
Model 2	0.813	.7929	.9292	.4144