Kaggle Assignment

Patrick Passafiume

November 24, 2017

Porto Segura Kaggle Competition

The purpose of the Porto Segura competition is to predict the probability that a driver will initiate an insurance claim. There are many applications for predicting whether a driver will initiate a claim. The company could use the predictions to tailor prices for customers based on the probability that the driver will file claim. This could allow good drivers to get a lower price while bad drivers are charged a premium based on that probability. It could also help the company be more selective about who they decide to cover in the hopes to minimize the amount they may have to pay out in future claims.

Data

The Porto Segura data was split into a training set and a test set. The training set is 595,212 observations with 59 features. The test set is 892,816 observations with 59 features as well. The data consists of categorical, ordinal, binary, and continuous variables. Omitted variables are intially coded as -1. The target (claim) variable is included in the training set with 0 and 1 to designate a claim or not.

Data Preparation

All of the missing data points (-1) were changed to NA's. A target variable was added to the test set to combine the training and test set to make sure any feature engineering and transformations done would affect both data sets. The binary and categorical data were transformed into factors. All of the created features were taken from the Kaggle example. Those include counts of NA's, sums of binary columns, sums of calulated columns, and "ind" binary column differences per row. The features that were intially chosen were based on the visuals from the Kaggle example and how they impacted the claims rate. Features were created by binning variables that seemed to have most importance based on the confidence intervals and they group together. Categorical variables and binned variables were all transformed by binning. This took care of the majority of the NA's in the data set. The rest of the NA's were imputed with the mean of the column. Different approaches were used throughout the process. In some models, the NA's were imputed by median and the initial binning was replaced with imputed all NA's with either mean or median. The binning with mean method was ulitmately what was used for the models chosen to submit.

Models

The first model used was a random forest. The features used initially were the ones based on the Kaggle example that had high importance and based on the new binned variables. The random forest model was not at predictive and predicted claims at a rate less than chance (3.6%). However, the random forest was helpful in optimizing the features chosen by looking at the feature importance. The next model used was a naive bayes model. This model was tried with all of the original features and with the random forest selected features. The naive model was more predictive than the random forest. It predicted 618 (7.5% of predicted) claims that were actually claims. The naive bayes model increased my Kaggle rank from the random forest model. The most predictive models were the logistic regression models. Different methods were used to experiment using different features. Using the random forest chose features produced the most predictive logistic regression model. All of the logistic regression models were evaluated using the AUC method. The highest AUC attained was 0.6334, which also produced the highest rank on Kaggle. KNN and a decision tress model were created as well, but were both terrible at predicting claims rates. Kaggle was used to evaluate models as well.

Evaluation and Deployment

The logisitic regression model does a decent job at predicting the claims rate without knowing what the variables mean explicitally. There are probably better ways to transform the data and bin the data to increase the predictive power compared to the model used. Logistic regression isn't the most powerful modeling technique to use, but for the purposes of this assignment, produced the best results on Kaggle for the models that were built.