Kaggle Competition: Safe Driver Prediction

Boh Young Suh, Morgan Wall, Ni Zhuang

* Who might care about this problem and why?

Predicting safe drivers and identifying whether or not they will file an insurance claim can be an important problem to analysis for many stakeholders. First of all, insurance companies will show high interest in these analyses because it directly affects their business. It can also be considered for auto insurance rates in the future to differ price based on the prediction and develop a reasonable insurance coverage for more drivers. Researchers in related fields would also care about this problem to analyze particular patterns related to safe driving that might be discovered through analysis as well as individuals who own a car and are looking for better auto insurance coverage.

* Why might this problem be challenging?

It is challenging to predict whether the driver will file for insurance or not because there are so many predictors in the data, the model built out of these predictors is hard to be perfect. What we end up with are all probabilities, which makes it hard to find the right split using the bagging and random forest methods. As for this dataset, the size of both the rows and column variables are quite large so it is difficult to run analysis with the whole data but using a subset of the full data may result in inaccuracy and biases. In the training data, most of the response variable has the value 0 with only a small fraction being 1, which makes it difficult to build the model for prediction. Also, factors in the train and test data have different levels so that prediction for the response variable “target” requires more data cleaning.

* What other problems resemble this problem?

There are many problems that may be similar to the Porto Seguro Safe Driver Competition. One of the defining features of the Porto Seguro competition is that our goal is to predict the probability of a binary outcome. In this scenario the probability is relevant because it characterizes the amount of risk that the insurance company is taking by insuring each driver. Some similar cases include predicting the outcome of a sporting event, particularly if it were an event that you were betting on, in which case you may want to know the probability of the outcome because of the risk of the bet. Another similar problem could be if you are trying to predict if a person will default on their loan. Similar to the car insurance scenario, in this situation understanding the probability that each individual has of defaulting can help you minimize your risk by deciding if this is a person to whom you would like to give a loan or what level of interest rate you may wish to charge.