ISYE 6740 Computational Data Analysis Final Report

Rahul Ghosh, Hao Wu, Wenhui Yang, Yiran Zhu

April 30, 2017

1 Introduction

Vehicle accidents affect millions of Americans and cause enormous economic loss every year. During 2015, roadway crashes caused 35,092 fatalities as well as large financial loss for the country (National Highway Transportation Safety Administration, 2015). As half of team members have four-year transportation engineering background, we selected 2015 Transportation Fatalities from National Highway Traffic Safety Administration as our dataset. The dataset contains 19 files including various information of each accident, such as crash types, locations, road environment, body type of vehicles, whether alcohol/drug involved, etc. The project aims to provide safety recommendations for drivers and reference for future machine learning analysis on traffic safety.

To utilize machine learning methods to detect statistical pattern between traffic accidents and factors, this project used Scikit-learn to investigate the problem. Scikit-learn is a free software machine learning library for the Python programming language. This library features various classification, regression and clustering algorithms, and is designed to inter-operate with the Python numerical and scientific libraries NumPy and SciPy. In this project, our approaches are divided into two branches: (1) predict whether the involved driver is drunk or sober using decision tree by three groups of features; (2) predict the severity level of injury by person and vehicle features. At last we test the accuracy of all the classifiers.

2 Demographics

2015 Transportation Fatalities dataset contains a large quantities of demographic characteristics. Initial demographic analysis help us to capture a better understanding of the dataset and the basic statistical influence of fatal accident features.

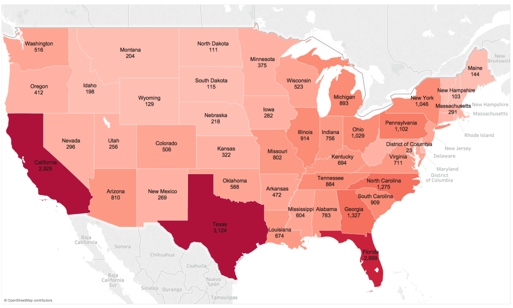

Figure 1: Fatal Accidents by States

Figure.1 shows the fatal accidents distribution by different states, which indicates the states with the largest amount of traffic fatalities are Texas, California, and Florida. In total they make up approximately 27 percent of the traffic fatalities in 2015. Breaking down the demographics of the accident victims, which is shown in Figure.2, we find that 70.9 percent of victims are males and 29.1 percent are females.

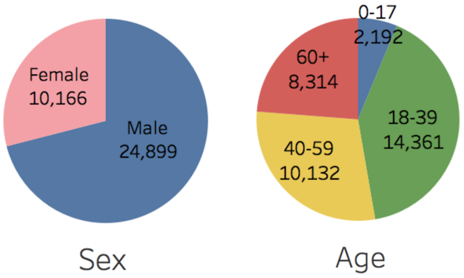

Figure 2: Victims in Fatal Accidents by Sex and Age

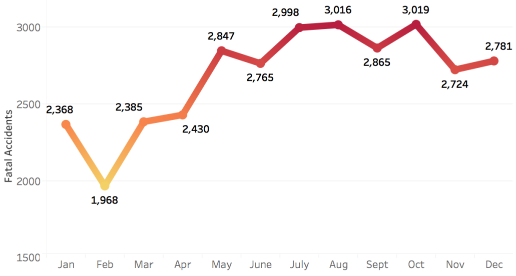

Figure 3: Fatal Accidents by Months

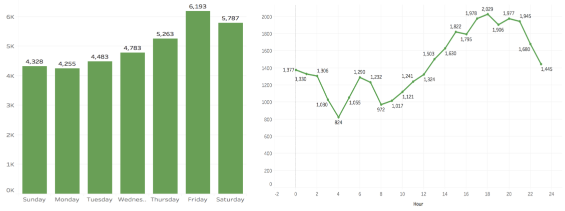

Figure 4: Fatal Accidents by Day and Hour

Figure.3 and Figure.4 show the time distribution of fatal accidents. According to the month distribution, there is an increase in the number of accidents between the months of February through October. According to the distribution of day of the week, most accidents tend to occur during weekends. What's more, the hour distribution states that most accidents occur between the hours of 3pm - 9pm of a day.

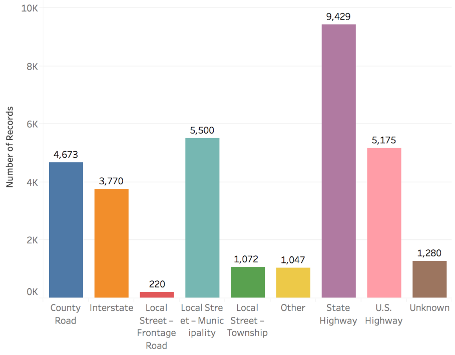

Figure 5: Fatal Accidents by Route Types

Figure.5 shows the fatal accidents distribution by road type. It indicates that state highway is the most risky road type for traffic fatalities, which is the scene of approximately 30 percent fatal accidents. The possible reason is that the vehicle speed is higher for state highway than for local road. The high speed will more likely result in severe accidents even cause fatalities. Besides, the state highway has the longest mileage among all the highway types, so it takes the largest number of accidents compared with other categories of highways.

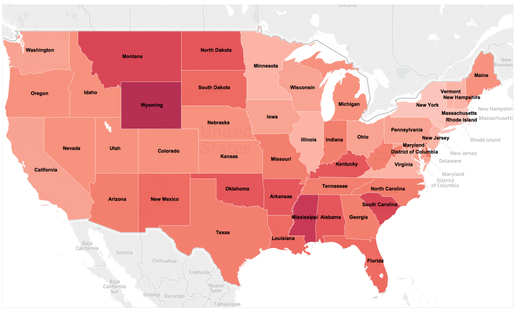

Figure 6: Victims Rate by Population

In addition to the above basic statistics, we calculate the rate of victims by the total population within the state. The rate distribution is displayed in Figure.6. Recalling the state distribution of total accidents, we can find that the states with largest number of accidents and the states with the highest victim rate are totally different. New York has the lowest victim rate per population (0.013 percent) across the country. It is because that the denser cities usually have lower probability of severe accidents for their low average speed. Meanwhile, in the state there exists a very dense city, New York City, which takes nearly the half of population in New York State. The low victim rate and large population of New York City account for the lowest rate of the state.

3 Paradigms

3.1 Predict Drunk or Sober

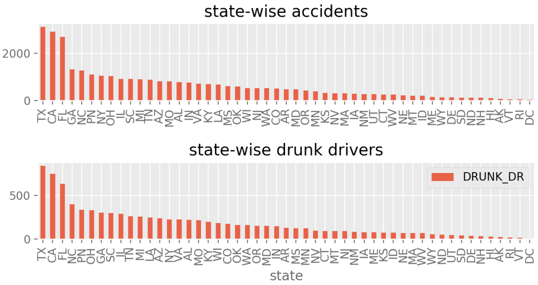

Figure 7: No. of Drunk Driving Related Fatalities

In order to dive deeper into analyzing the data, we will focus on a major factor related to traffic related fatalities, drunk driving. The first graph in Figure 7 shows the total number of traffic fatalities in each state in 2015, while the second graph depicts the total number of traffic fatalities involving one or more drunk drivers. From the graphs one can see that the three states with the greatest amount of fatalities are Texas, California, and Florida. This is not surprising due to the fact that these three states are the largest states in terms of population in the United States. So the number of traffic related fatalities in these states are proportional to the population. The same conclusion can be made of the number of drunk driving related fatalities in each state. The number of drunk driving related fatalities in each state is proportional to population of each state.

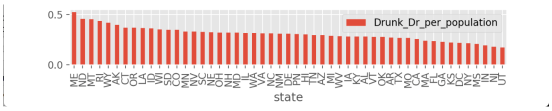

Figure 8: Percetage of Drunk Driving Fatalities

Even though the number of drunk driving related fatalities are proportional to the population of each states, it does not tell the full story of how the factor of drunk driving affects the total number of traffic fatalities in that respective state. Figure 8 depicts the percentage of traffic fatalities that are related to drunk driving in each state. One can see from this graph that even though Texas, California, and Florida had the highest total number of drunk driving related fatalities in 2015, the percentage of drunk driving accidents in each state is quite low. However, the three states with the highest percentage of drunk driving related fatalities are Maine, North Dakota, and Montana. The differences in perception based on how the data is viewed is a key reason why further exploration of the data is needed by using machine learning techniques.

3.1.1 Dataset: features and class labels

Within the accident data file, three different groups of columns are selected as features and whether the driver was drunk or sober is selected as binary label. Three groups are: Time and Environment, Accident Type, Weather. In the Time and Environment group, the month, day of week, day of month, hour, and how harmful the environment was are included; the Accident Type group contains common types such as angle, head-on, rear-end, sideswipe. etc.; in the Weather group, the weathers are divided into: clear, cloudy, rain, fog, snow, sleet, drizzle, blowing-snow, cross-wind, and blowing-sand.

3.1.2 Decision Tree Accuracy

For each of the above three feature groups, over one hundred times' random data trainings are performed and the average results showed the first group: Time and Environment has the closest relationship with whether the driver was drunk or sober. Even though the logic connection between these features and the drunk/sober label was not strong, the decision tree prediction accuracy was statistically acceptable.

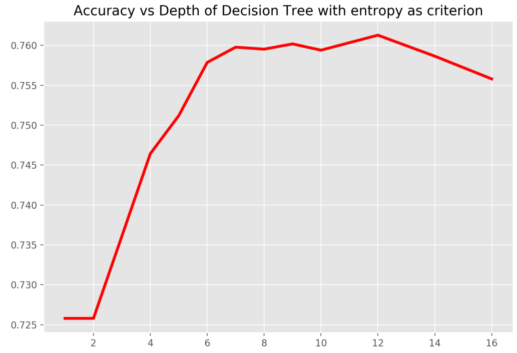

Figure 9: Accuracy V.S. Entropy as Leafing Criterion

Entropy and information gain were two leafing criterion chosen and both of them gave positive accuracy feedback to our decision tree: nearly 76.5 percent of the predictions are right after the tree depth exceeds six. The curve grows a bit faster in entropy figure but is more stable in information gain one. Both curves have a slight accuracy downhill because of the over-fitting.

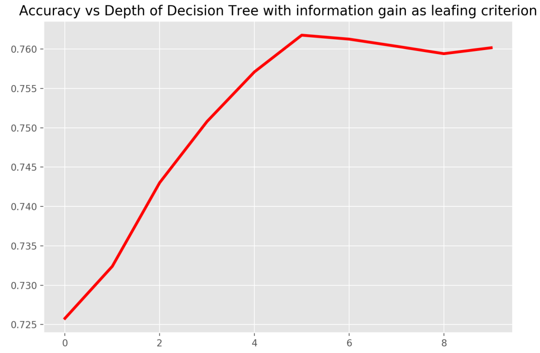

Figure 10: Accuracy V.S. Info Gain as Leafing Criterion

3.1.3 Multi-class Predictors

In addition to simple decision trees, various classifiers such as K-Neighbors, SVC, Random Forest, AdamBoost, Gradient-Boosting, Gaussian Naïve Bayes, Linear Discriminant Analysis were implemented for predictions.

The accuracy figure shows that the Linear Discriminant Analysis has the best prediction accuracy with nearly 80 percent, which is still, not a significant improvement compared to simple decision tree classifiers. The Gaussian Naïve Bayes has the worst accuracy of less than 45 percent.

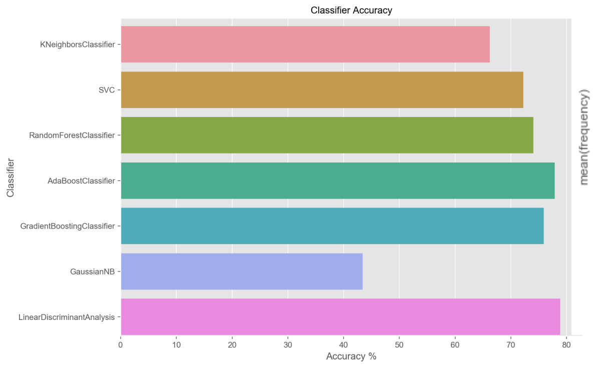

Figure 11: Classifier Accuracy

Going deep into the relationships within our feature groups, we found out that as the number of features we included increases, the "feature-independent" pre-assumption made by the Gaussian Naïve Bayes was violated. As Google developed the Scikit-Learn library, the GaussianNB classifier was not exempted from the assumption and further corrections was not made either, which caused the high log-loss, shown in Figure: Classifier Log Loss.

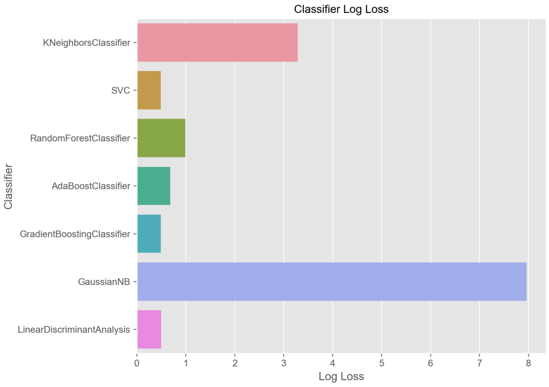

Figure 12: Classifier Log Loss

3.2 Predict the Injury Severity

3.2.1 Dataset: features and class labels

In this part, we construct several predictors for injury severity of victims in accidents by using following features: Vehicle Body Type, Age and Sex, Seat position, Railway units, Alcohol test, Drug test, Airbag condition and Restraint equipments usage. As provided by source data, class labels vary from 0 to 6 to represent: No injury, Possible injury, suspected minor injury, suspected serious injury, fatal injury, injury - severity unknown and death.

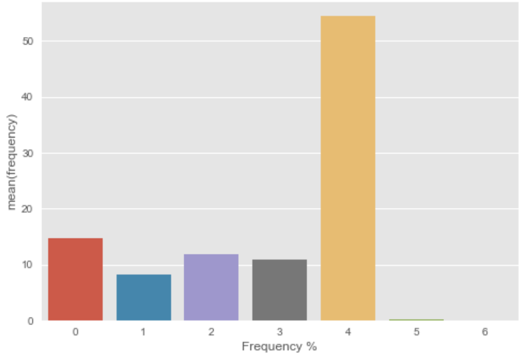

Figure 13: Distribution of classes

3.2.2 Multi-class predictors

Various methods are implemented for construct predictors such as SVM, Decision Tree, Decision Forest, AdaBoost, Neutral Network and Naive Bayes. Additionally, we combine decision tree, neutral network and SVM together to derive a majority voting classifier.

Some methods are originally designed for binary class, in our approach, we generate these methods to multi-class by deriving several one against rest binary predictors with same method.

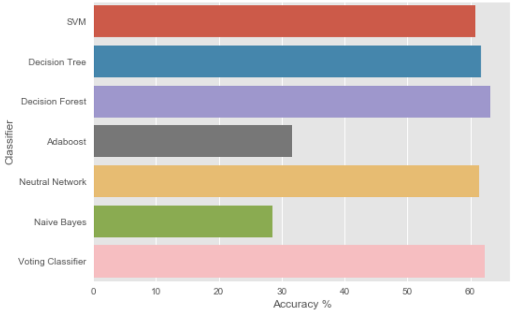

Figure 14: Classifier accuracy

The best accuracy is 63.2% given by Random Forest. Compared to accuracy 54.5% of the best trivial classifier which assigns all inputs to class 4, Random Forest's behavior is strong.

For understanding detailed performance of Random Tree and AdaBoost, we analyze how accuracy changes against parameters of these two methods.

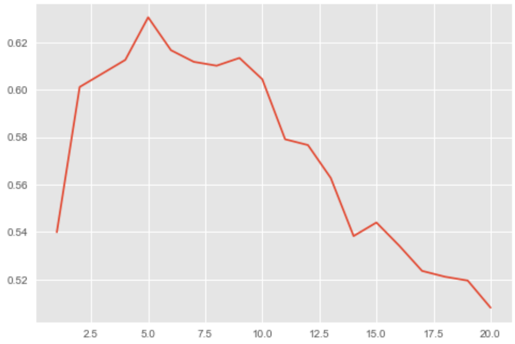

Figure 15: Accuracy vs depth of random tree

With growing depth of Random Tree, accuracy of the predictor increases at first and then decreases since overfitting occurs.

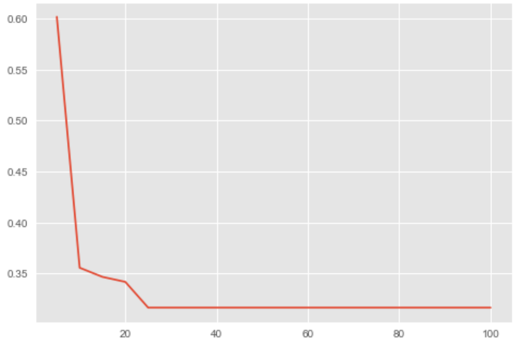

Figure 16: Accuracy vs number of weak classifiers

Increasing number of weak classifiers involved in AdaBoost leads to a worse result. Weak classifier used in the module is decision stump and we train a new classifier per iteration by bootstrapping. Although training data is sampled randomly, our experimental result shows overfitting is still serious.

3.2.3 Binary predictors

Since some classes in previous part overlaps, we split dataset into binary disjoint classes, namely injury and no injury, by combining all classes with label greater than zero. Features we use here are the same as before.

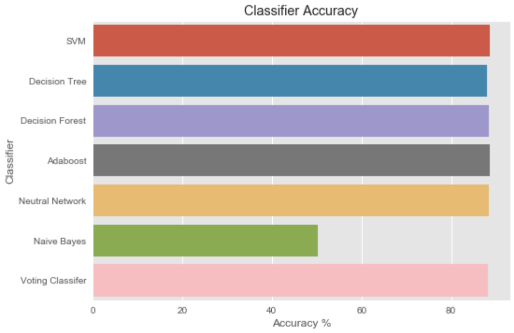

Figure 17: Classifier accuracy

In this case, the best accuracy is again given by Random Forest which is 88.2%. As following figure indicates, parameter analysis of depth in random tree shows the similar pattern to previous case.

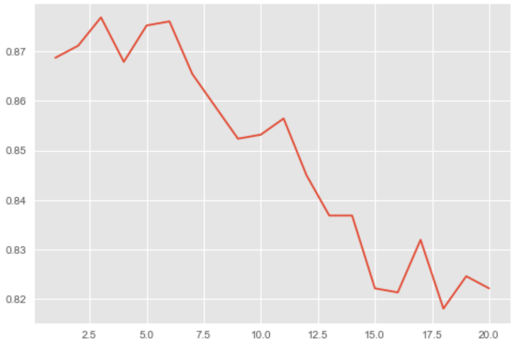

Figure 18: Accuracy vs depth of random tree

Though in this case, accuracy does not increases at the beginning. This phenomenon shows there is a strong connection between injury and alcohol test.

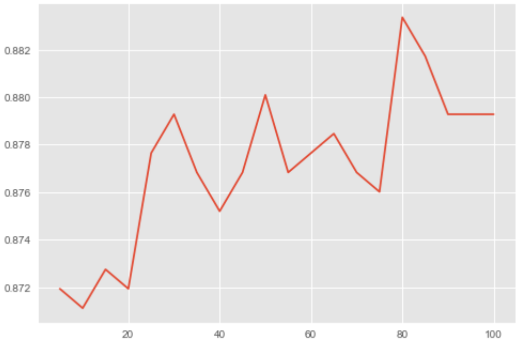

Figure 19: Accuracy vs number of weak classifiers

Different from previous case, in AdaBoost module, more weak classifiers means better accuracy.

4 Conclusion

According to our data analysis, the time distribution of traffic fatalities indicates inner rules significantly related to the traffic volume. Strong relationship between accident and whether drive was drunk or sober was shown. However, using the fact of accident and environment-related features to predict the driver's drunkenness is not accurate enough, no matter what type of classifier was chosen. Using the injury severity to predict the driver's drunkenness is relatively more reliable.

From the study of different predictors, we learned that data cleaning and pre-processing are very important. It is unwise to simply plug the data into various models without clarifying their assumptions and limitations. Also, we decided to choose features that are logically related to accidents but independent to each other. We haven't done thorough prediction study over all features, which could be the goal of our future effort.

5 Reference

1. National Highway Transportation Safety Administration, 2015. Traffic Safety Facts. Retrieved March 11, 2017 from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812318

2. Chong, Miao M., Ajith Abraham, and Marcin Paprzycki. "Traffic Accident Analysis Using Machine Learning Paradigms." Informatica (Slovenia) 29.1 (2005): 89-98.