Predicting Severity in US Traffic Accidents

—

Jingxuan Li * Jialiang Wei †

November 8, 2020

Abstract

This research employed gradient boosting techniques with a countrywide traffic accident dataset, which covers 49 states of the United States and is continually collected from February 2016 till now, to study the impacts of different factors, geographical locations, environmental stimuli and other relevant factors on the severity of accidents. The main goal of this research is to examine the significant factors that impact the severity of accidents, and to discuss how federal and local governments, as well as drivers can make changes to prevent or reduce the impact of accidents in different environments.

*Cornell University, *jl4267@cornell.edu*
†Cornell University, *jw2684@cornell.eduu*

1 Explanatory Data Analysis

1.1 Data Description

Our project employs the *US Accidents* from Kaggle.com. This is a countrywide traffic accident dataset, which covers 49 states of the United States. The data is continuously being collected from February 2016, currently there are about 3.5 million accident records. For every accident record we have data on the source of the accident report, description of the event, severity of the accident, start and end time, latitude and longitude information, zip code, time-stamp of weather observation record, temperature, wind chill, humidity, air pressure, visibility, wind direction and speed, presence of crossing, railway, traffic signal, junction, etc., and the period of the day. We will build learning models to study how different factors impact the severity of a car accident.

1.2 Data Visualization

1.2.1 Accidents Summary

Figure 1 shows the number of accidents in different states recorded in the dataset. Based on the bar chart, the state with index of 3 has much more accidents than the other states, which is referred to "CA" (California).

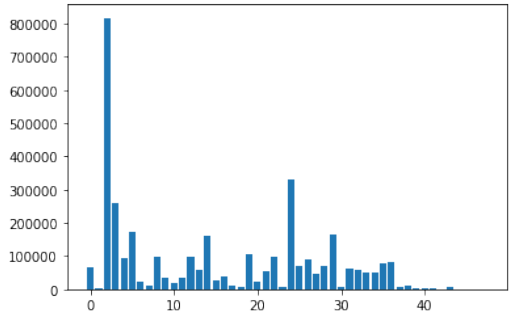

Figure 1: Number of Accidents

Figure 2 shows the normalized numbers of accidents over number of automobiles in different states (*US Department of Transportation*), as this ratio means on average how many car accidents per car in each state. Based on the figure 1, we want to know whether the number of accidents is correlated with states or other factors. It shows that except index 6 (South Carolina) has a higher ratio, all other states exhibit similar level as in figure 1. The correlation between the number of accidents and the total number of automobiles in each state is 0.91, which is consistent with what we expect that the more automobiles in a state, the higher possibility of traffic accidents.

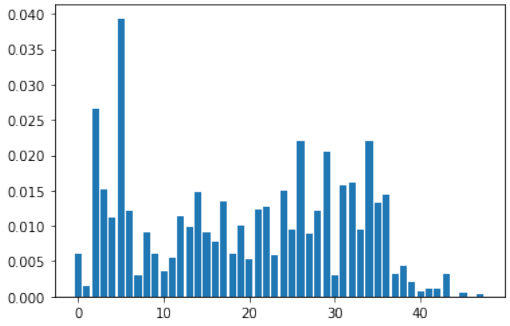

Figure 2: Traffic Accidents in Different States

1.3 Data Preprocessing

1.3.1 Missing Values

In the dataset, the missing values appears in Temperature (F), Humidity (%), Visibility (miles), Weather_Condition (rain, snow, thunderstorm, fog, etc.), and Sunrise_Sunset (day or night). Since the dataset is larger enough with 3 513 617 records in total, the records with missing values in any category are dropped. We still have 3 414 253 records with complete information which is large enough for us to do analyses.

1.3.2 Ordinal Values

The response variable in the model is the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic and 4 indicates a significant impact on traffic. As severity is a ordinal value, we will choose a model that can be adapted to fit a multi-classification response variable.

1.3.3 Nominal Values

The column "Start_Time" shows start time of the accident in local time zone. We classify them into four time windows (0 AM 6 AM, 6 AM 12 PM, 12 PM 6 PM, and 6 PM - 12 AM) and use one-hot encoding. The column "Street" shows the street name in address field. We believe that the accidents happened on interstate highways may be more severe than others, so we use one-hot encoding to classify the street with a name starting with "I-" which represents the interstate street. The columns "side" which contains nominal data and shows the relative side (right or left) in the accident on a street, "Weather_Condition", and "Sunrise_Sunset" contains nominal data and shows the period of day (day or night). We use one-hot encoding for each of them to make their values more expressive.

1.3.4 Additional Note

Since the data is recorded chronologically, we shuffle the data and split into training data and test data to avoid autocorrelation and overfitting. We tried with ten-fold cross-validation, however, the dataset is so large that our computational power is not capable.

2 Initial Modeling

2.1 Model Selection

Based on our parameter selection from data preprocessing and data encoding, we predict that there will be significant overfitting issues with the data. Indeed, when we run LASSO regression on our dataset, the lambda yields meaningless results (i.e. we get our optimal lambda = 0, which is the same as OLS). In addition, due to the large volume of data (1.98 Gigabytes, 481,409,673 data points), we can only consider models that are extremely efficient and more robust in preventing overfitting. Thus, we decide to use the gradient boosting method (R package: xgboost) to fit our data.

2.2 Gradient Boosting Algorithm

1: Initialize

$$\mu^{(0)}(x) = \arg\min_\theta \sum_{i=1}^{n} L(y_i, \gamma) \tag{1}$$

2: For m = 1 to M

a: For i = 1,2,...,n compute

$$r_i^{(m-1)} = -\frac{\partial L(y_i, \mu(x_i))}{\partial \mu(x_i)} \tag{2}$$

b: Fit a regression tree to the current residuals $r_i^{(m-1)}$ giving terminal regions R_{jm},j = 1,...,J compute

c: For j = 1,...,J compute

$$\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L(y_i, \mu^{m-1}(x_i) + \gamma) \tag{3}$$

d: Update $\mu^m(x) = \mu^{m-1}(x) + v \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm})$

3: Output $\hat{\mu}(x) = \mu^M(x)$

Due to page limit, for more information, please refer to *The Elements of Statistical Learning pp 359-361.*

Thus, we can see that to fit a gradient boosting algorithm, we need tuning process. Given limited computational power, we tuned the below factors: max_depth (the complexity of each fitting tree, positively correlated with J in above algorithm), eta (learning rate, v in the algorithm), subsample (subsample a portion of the data in our training data in each iteration), colsample_bytree (subsample a portion of input parameters in each iteration).

2.3 Tuning Parameter Selection

As our response variable (severity) is an ordinal value, we defined two error functions for multi-classification models that calculate the number of differences between our model prediction output and observed data. We define two parameters: percentage number of errors and percentage size of errors:

$$\text{percentage-number-of-error: } \frac{\sum_{i=1}^{N} 1_{\{y_i \neq \hat{y}_i\}}}{N} \quad (4)$$

$$\text{percentage-size-of-error: } \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i| 1_{\{y_i \neq \hat{y}_i\}}}{N} \quad (5)$$

We tuned a total of 9 models, with the tuning parameter as follows: (max_depth, eta, subsample, colsample_bytree) = (5, 0.3, 0.9, 1), (11, 0.3, 0.9, 1), (20, 0.3, 0.9, 1), (20, 0.3, 0.9, 0.8), (20, 0.3, 0.9, 0.5), (20, 0.3, 0.5, 0.8), (20, 0.5, 0.5, 0.8), (20, 0.5, 0.9, 0.8), (20, 0.7, 0.9, 0.8). Figure 3 shows the percentage number of errors and percentage size of errors. The fitted models are in the same sequence as the sequence from above.

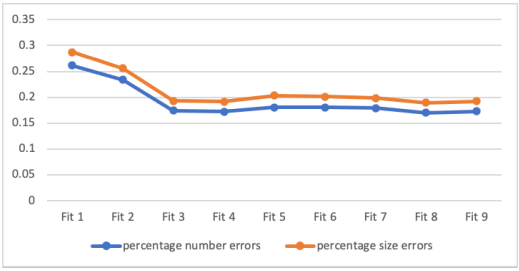

Figure 3: Errors of different Tuning Models

We can see that model 8 with tuning parameters (20, 0.5, 0.9, 0.8) has the smallest in both number of errors and size of errors. Thus, given our limited computational power, we chose this model to study the dataset's relative importance factors (to tune these 9 models, we ran for 15 hours with 3.8GHz CPU and 32G memory).

2.4 Relative Importance Factors

By fitting with our tuned gradient boosting model, we can see the relative importance of each factor. Figure 4 shows the percentage gains in loss function for each input parameters to the severity of car accidents.

We can see that geographical location (latitude and longitude), types of streets (Interstate highway or others), temperature, weather and humidity are some of

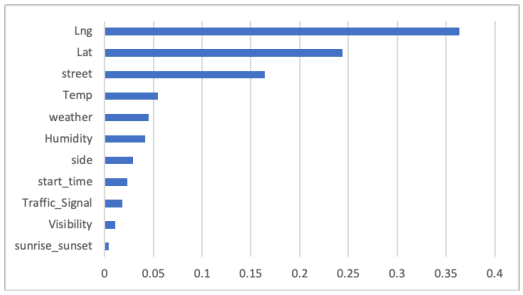

Figure 4: Relative Importance of Input Parameters

the most important factors to the severity of a car accident. However, some of other factors that we intuitively regard also important, such as sunrise_sunset (day/night), visibility and traffic signal are not as important. This shows that drivers already pays close attention when the road condition is complicated (in dark or foggy conditions), and drivers follow traffic signals closely. However, the high importance in location and weather shows authorities should invest more in certain regions for road maintenance such as snow removal. The high importance in types of streets also shows that drivers should pay close attention with high speed limit roads.

3 Potential Improvement

While our model shows impressive prediction results, there is still work to do to get a more accurate and holistic model. Firstly, we need to seek for additional computational power to continue researching this dataset, as it is beyond PC's capacity if we are going to try more complex models. Secondly, we will also explore other data encoding methods to include more input variables that we think is significant. Thirdly, we will try other models that can further improve overfitting and solve classification more effective, such as neural networks. Last but not least, we will examine balanced errors, as some parts of our data are not balanced, with most severities in level 2 and 3, and few of them are in level 1 and 4.