Analysis on the Car Accident Severity in US

Yi Li

Nov. 2020

1. Introduction

Traffic accidents have become a big threat to public safety and resulted in great amounts of economic loss around the world. A global status report on traffic safety indicated that the number of road traffic death continued to increase steadily, reaching to 1.35 million in 2016 [WHO, 2018]. Therefore, one of the important task for safety analysts and political makers, in order to mitigate the severity of the accidental consequence, is to make a comprehensive assessment of historical traffic accidents and then to increase the predictability of accidents.

Accident analysis and prediction has been discussed in many previous studies and covers a broad range of categories, including, for example, Environmental Stimuli Analysis, Accident Frequency Prediction, and Accident Risk Prediction [Moosavi, et. al., 2019a]. Environmental Stimuli Analysis assesses the environmental conditions (e.g. weather, and road conditions) that are correlated with the possibility or severity of traffic accidents. Accident Frequency Prediction is targeted on predicting the number of traffic accidents for a specific road-segment or geographical region. Accident Risk Prediction is similar to the previous one. However, instead of predicting the number of accidents, it is focusing on predict the possibility of road traffic accidents for real-time applications.

The analysis in this work belongs to the first category as I tried to seek the potential environmental stimuli in road traffic accidents. I used several exploratory data analysis (EDA) tools to investigate heterogeneity in the environmental factors and assessed the

impact of environmental stimuli on severity of the accidents in US using several different machine learning models. The results from my analysis may provide advice to political makers on whether new regulations are needed in specific roads or weather conditions to reduce the risk of traffic accidents. In addition, it may also give suggestions to car drivers to avoid certain road segments or to be vigilant on certain environmental conditions.

2. Data

The data set employed in this study is a countrywide traffic accident dataset (US-Accidents), which covers 49 states of the United States [Moosavi, et. al., 2019b]. The data were collected continuously from February 2016 to March 2019 and contains about 3.5 million accident records in total. This data set contains various attributes including time, location, severity and description of accidents, weather conditions, points of interest annotation (POI, e.g. whether there is a Stop sign in a nearby location). A summary table of all data attributes is shown in Table 1. Details of the attributes and data acquisition strategy can be found in Moosavi, et. al., 2019b. And the data set is available on Kaggle.com (https://www.kaggle.com/sobhanmoosavi/us-accidents).

Table. 1 Listing of attributes in US-Accidents data set

#	Attribute	Description	Nullable
1	ID	This is a unique identifier of the accident record.	No
2	Source	Indicates source of the accident report (i.e. the API which reported the accident.).	No
3	TMC	A traffic accident may have a Traffic Message Channel (TMC) code which provides more detailed description of the event.	Yes
4	Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).	No
5	Start_Time	Shows start time of the accident in local time zone.	No
6	End_Time	Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow was dismissed.	No

#	Attribute	Description	Nullable
7	Start_Lat	Shows latitude in GPS coordinate of the start point.	No
8	Start_Lng	Shows longitude in GPS coordinate of the start point.	No
9	End_Lat	Shows latitude in GPS coordinate of the end point.	Yes
10	End_Lng	Shows longitude in GPS coordinate of the end point.	Yes
11	Distance(mi)	The length of the road extent affected by the accident.	No
12	Description	Shows natural language description of the accident.	No
13	Number	Shows the street number in address field.	Yes
14	Street	Shows the street name in address field.	Yes
15	Side	Shows the relative side of the street (Right/Left) in address field.	Yes
16	City	Shows the city in address field.	Yes
17	County	Shows the county in address field.	Yes
18	State	Shows the state in address field.	Yes
19	Zipcode	Shows the zipcode in address field.	Yes
20	Country	Shows the country in address field.	Yes
21	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).	Yes
22	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.	Yes
23	Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).	Yes
24	Temperature(F)	Shows the temperature (in Fahrenheit).	Yes
25	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).	Yes
26	Humidity(%)	Shows the humidity (in percentage).	Yes
27	Pressure(in)	Shows the air pressure (in inches).	Yes
28	Visibility(mi)	Shows visibility (in miles).	Yes
29	Wind_Direction	Shows wind direction.	Yes
30	Wind_Speed(mph)	Shows wind speed (in miles per hour).	Yes
31	Precipitation(in)	Shows precipitation amount in inches, if there is any.	Yes
32	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)	Yes
33	Amenity	A POI annotation which indicates presence of amenity in a nearby location.	No

#	Attribute	Description	Nullable
34	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.	No
35	Crossing	A POI annotation which indicates presence of crossing in a nearby location.	No
36	Give_Way	A POI annotation which indicates presence of give_way in a nearby location.	No
37	Junction	A POI annotation which indicates presence of junction in a nearby location.	No
38	No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.	No
39	Railway	A POI annotation which indicates presence of railway in a nearby location.	No
40	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.	No
41	Station	A POI annotation which indicates presence of station in a nearby location.	No
42	Stop	A POI annotation which indicates presence of stop in a nearby location.	No
43	Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.	No
44	Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby location.	No
45	Turning_Loop	A POI annotation which indicates presence of turning_loop in a nearby location.	No
46	Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.	Yes
47	Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.	Yes
48	Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.	Yes
49	Astronomical_Twilight	Shows the period of day (i.e. day or night) based on astronomical twilight.	Yes

3. Methodology

3.1 Exploratory Data Analysis

Exploratory data analysis (EDA) was implemented on the US-Accidents data set in order to reveal the heterogeneity in the data attributes. The distribution in the time, location, weather conditions were analyzed to get a comprehensive understanding on the characteristics of the data set. The cross-corelation was also calculated for better investigating the relationship between the environmental stimuli and severity of

accidents, as well as for reducing the dimension of features for regression since only one of the highly correlated attributes is needed as the input feature of regression models.

3.2 Predictive Modeling for Severity Analysis

Due to the limitation of computational power, the severity analysis was only focused on data from one state in US. Multiple regression models, including multi-variate logistic Regression, Support Vector Machine, Random Forest and XGBoost, were employed to predict severity of the accidents. The performance of each model was evaluated and compared based on accuracy score of the prediction. The importance of features in predict accident severity were also discussed.

4. Exploratory Data Analysis

4.1 Distribution of Severity

Severity in US-Accidents data set (Figure 1) shows the severity of the accidents in US. It is represented as a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay). Figure 1. shows the imbalanced distribution of Severity in the data set. Over 95% accidents are labeled as medium severity, whereas only 0.8% events are labeled as level 1 severity. This should be considered when selecting severity as a target and making models to predict severity of the accidents.

4.2 Locations of the accidents

The locations of each accidents recorded in the data set were plotted in Figure 2. It is clear that most accidents happened in the west coast US and eastern US, whereas fewer accidents happened in the mid-US are recorded in the data set. We can also see that the traffic accidents in mid-US are mainly aggregated on the cross-country highways. Due to the inhomogeneous spacial distribution of accidents cross US, I further plotted the count of traffic accidents in each state in Figure 3. It is evident that

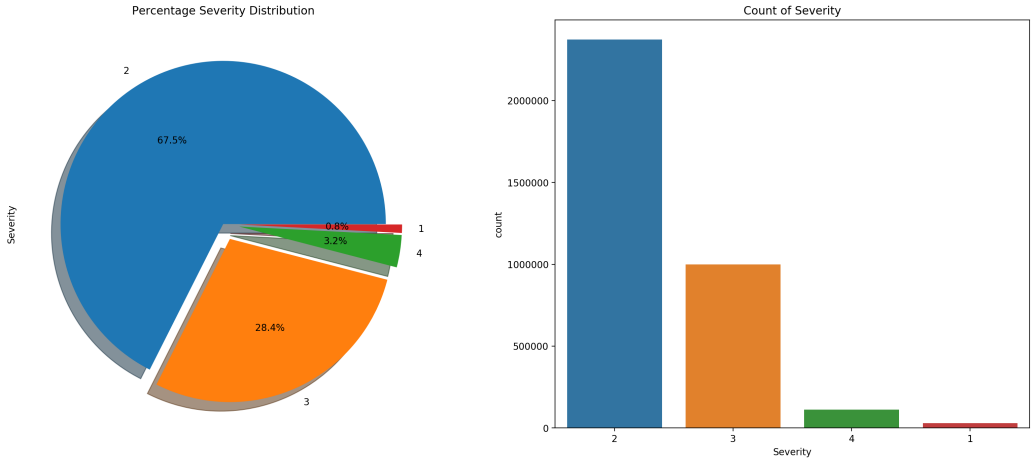

Figure 1. Percentage (left) and counts (right) of severity levels in US-Accidents data set

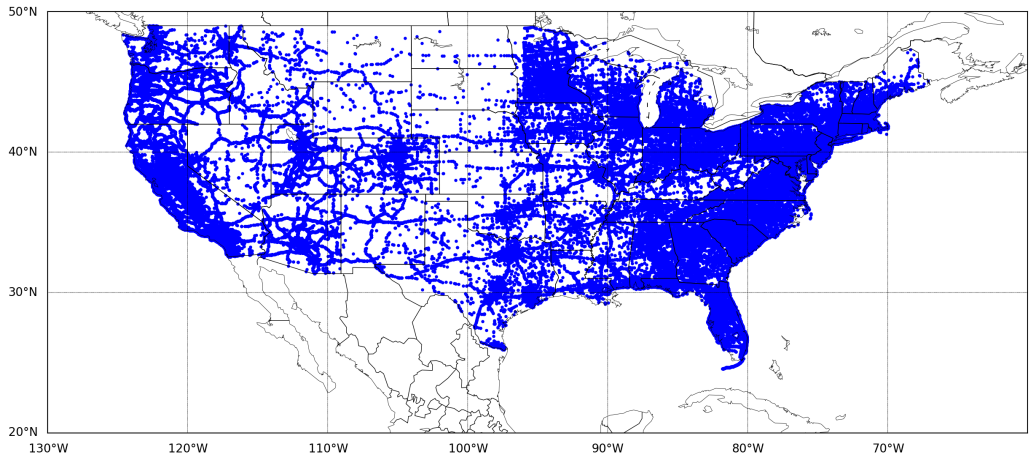

Figure 2. Locations of accidents in US-Accidents

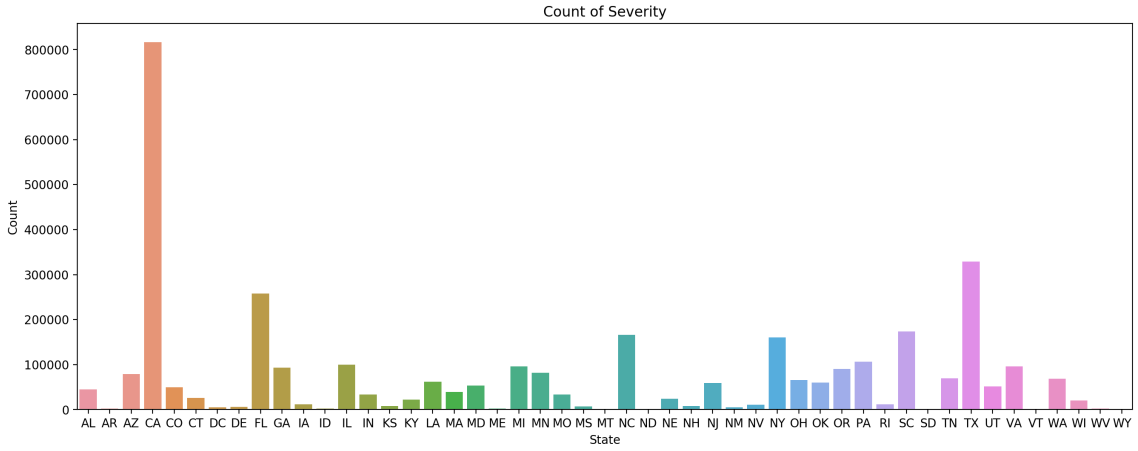

Figure 3. Counts of accidents in each state

state of California(CA), Texas(TX) and Florida(FL) are ranked as the top 3 states that have the most traffic accidents. These states have more than 780, 320 and 250 thousand cases of accidents respectively.

4.3 Time analysis

Characteristics of US-Accidents data set in terms of time analysis are shown in Figure 4-6. The weekday distribution in Figure 4 reveals that the number of accidents in weekdays are over two times more than that at weekends, though the counts of accidents in weekdays are quite even. The hour of day distribution of all accidents in weekday and weekends (Figure 5) have distinct patterns. During the weekdays, more accidents are happening at 7:00-8:00 and 16:00-17:00, which are synchronous with the peak hour on the roads. However more accidents are happening in the afternoon at weekends. The distribution and variance of duration time are shown in Figure 6. It is interesting that the median duration time for level 4 severity accidents is much longer than low-level severity events. This implies that the duration time might be an important feature to predict level 4 severity.

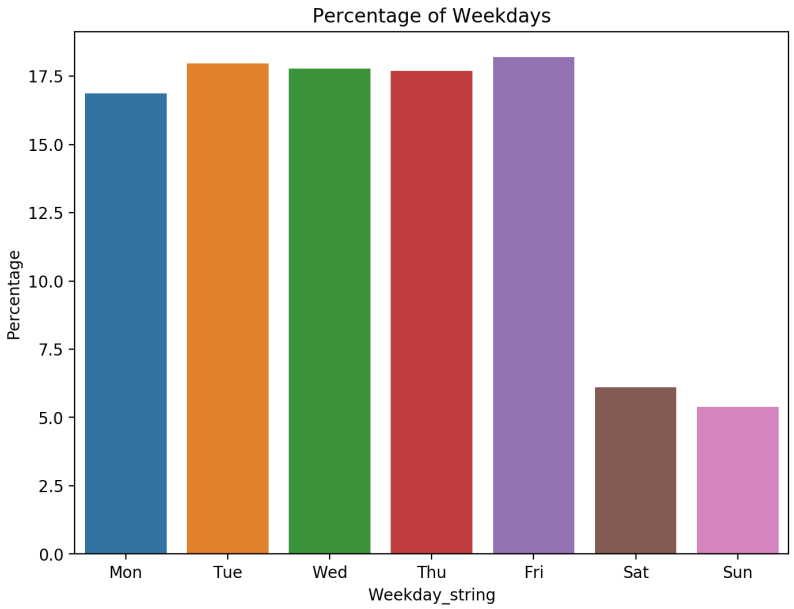

Figure 4. Weekday distribution of all accidents in data set

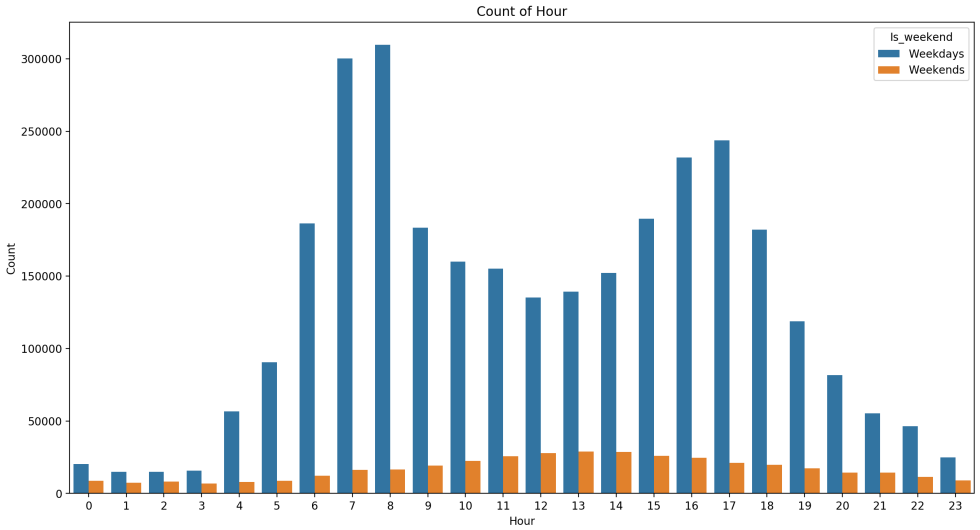

Figure 5. Hour of day distribution of all accidents in data set

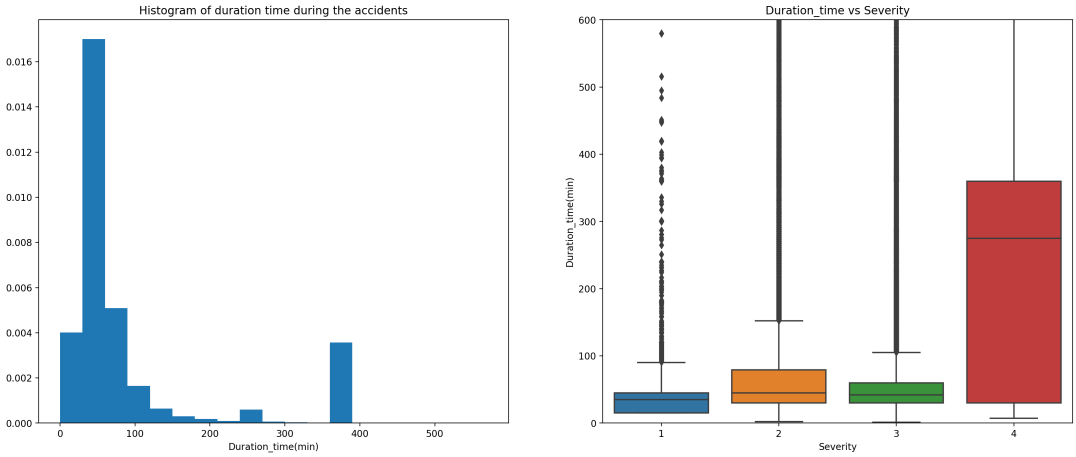

Figure 6. Duration time distribution of all accidents (left) and box-plot of duration time in 4 severity levels

4.4 Weather analysis

10 most weather conditions for all accidents are plotted in Figure 7. It is surprising to see that more accidents are happening in clear sky or fair condition. This might indicates that weather conditions may not be the determinative factor of traffic accidents. A closer look at the distribution of all weather conditions, I found that there

are no significant differences between different severity levels for most weather conditions, except for temperature, humidity, pressure and wind chill (not shown here). However there are a great amount of missing data in wind chill, so unfortunately this feature need to be dropped in further predictive modeling.

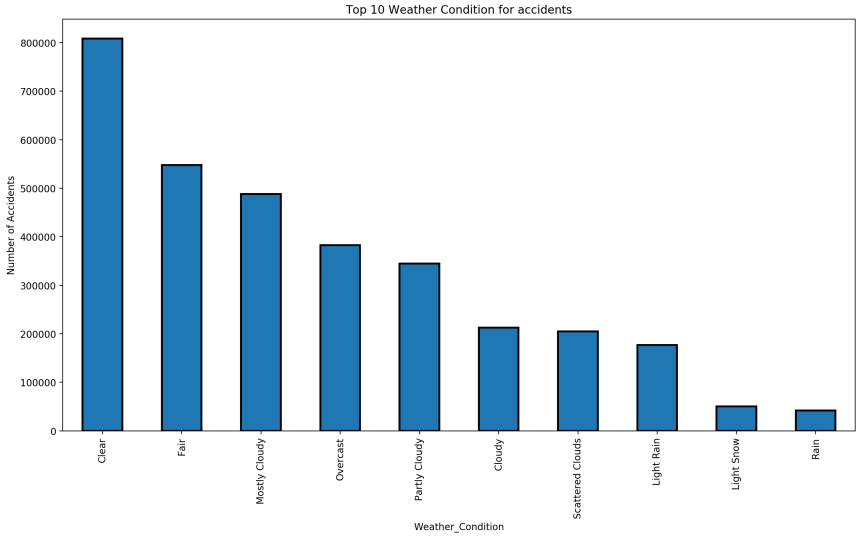

Figure 7. 10 most weather conditions for all accidents

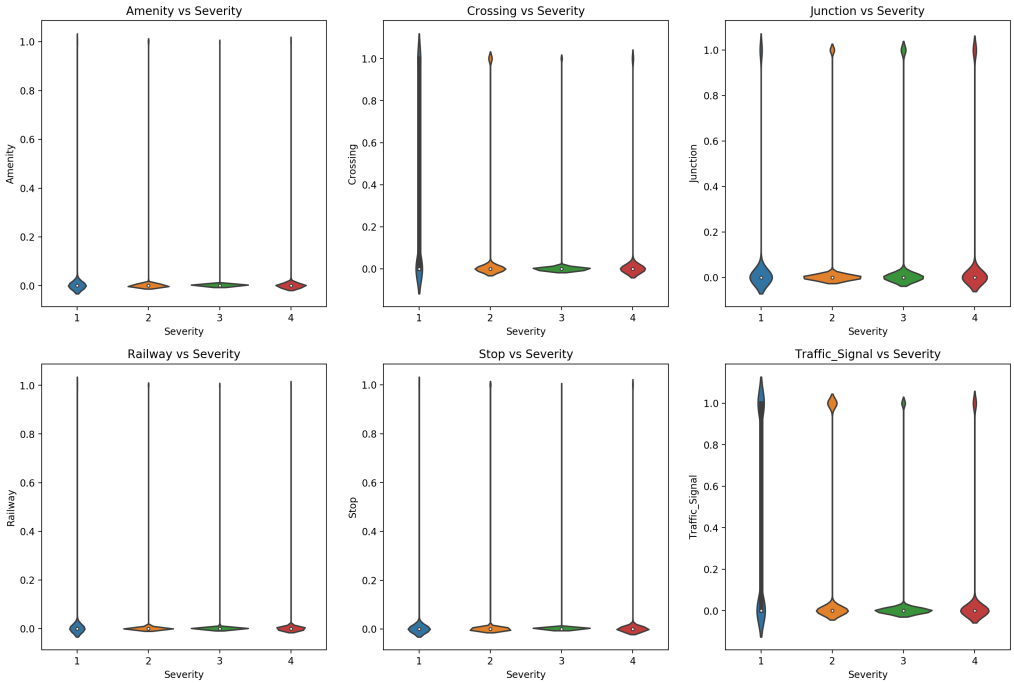

Figure 8. POI analysis for different severity levels

4.5 POI analysis

POI in the data set indicate whether there is a specific facility or traffic sign near the accident region. Figure 8 depict six POI distributions among all accidents at distinct severity levels. There are Crossing, Junction and Traffic signal existing in some of traffic accidents, whereas none or very a few Amenity, Railway and Stop appear near the region where accidents happened.

5. Feature Engineering

5.1 Correlation between features and severity level

The correlation between features and severity levels are depicted in Figure 9. Although all features have low correlation with Severity (<0.3), some of them show relatively higher correlation than others. So we may expect the Distance, Start_Lng, Stop, Junction, Year, Weekday, Traffic_Signal and Crossing to be important features to predict Severity.

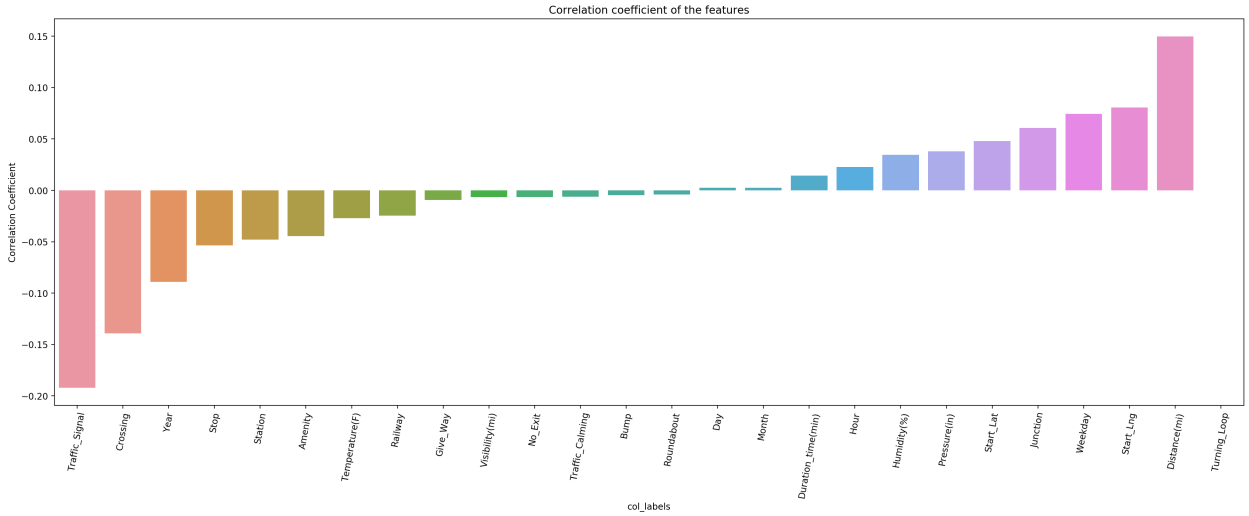

Figure 9. Correlation between features in US-Accidents and Severity

5.2 Cross-correlation of features

One way to reduce the dimension of features and improve the speed the modeling is to drop redundant features with high correlation. The cross-correlation map of all features are shown in Figure 10. The highest absolute value of cross-correlation is 0.4, indicating that we do not need to drop any feature as all features are quite independent.

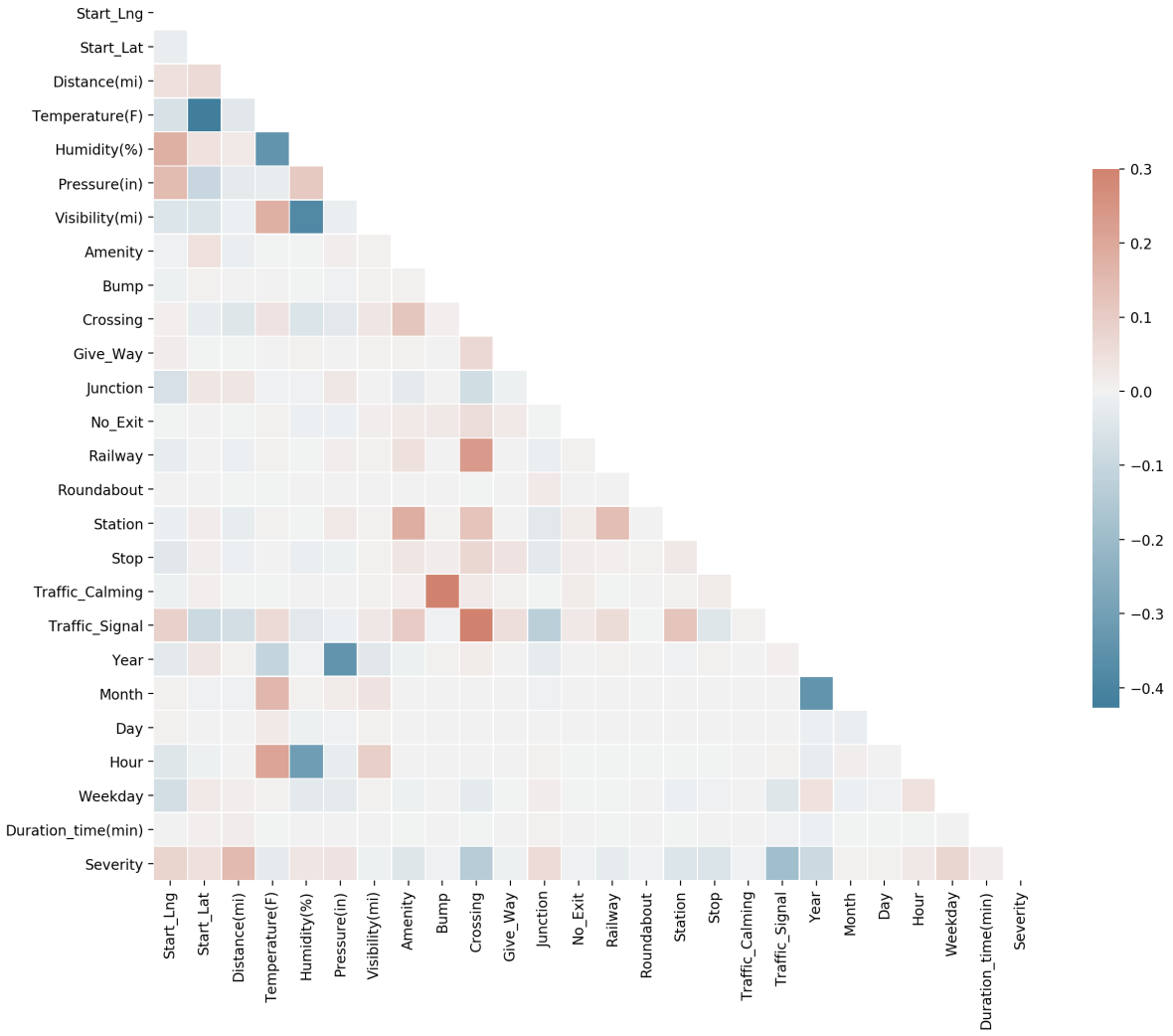

Figure 10. Cross-correlation map of all features

6. Predictive Modeling

This is a classic classification problem that we need to predict the severity level of traffic accidents based on the characteristics of the accidents. Due to the limitation of personal laptop, the whole US dataset with over 3.5 million samples is too big to handle. The following analysis will only focus on the accidents happened in the state of Georgia (GA). The reason to choose Georgia as the targeted state is because both the amount of traffic accidents and averaged severity levels in this state are in top 10.

6.1 Implementation of modeling

There are lots of information in US-Accidents data set. In order to simplify the problem, only the numerical and boolean variables are used in predictive modeling. The variables include location, time, weather condition and POIs are selected as predictive features. Noticed that weather conditions for each sample are converted to boolean variables using one-hot encoding and variables that over 30% missing values are dropped. Four classification models including Logistic Classifier, Support Vector Machine (SVM), Random Forest and XGBoost are implemented individually and evaluated for severity prediction. The GA data are split into 90% for training and 10% for validation.

6.2 Performance of different models

The performance of difference classification models are evaluated by the f1-score and accuracy of severity prediction are listed in Table 2. Overall, XGBoost outperformed among all models with both highest f1-score and accuracy. Comparing to Logistic Classifier and SVM, Random Forest and XGBoost significantly improved the predictability of level 4 severity, although all models has poor performance on predicting level 1 events. The poor predictability of level 1 events may be due to limitation of the highly imbalanced data set as level 1 events only account for 0.8% of the total data set.

Table 2. Performance of each classification models.

	F1-Score for different severity levels				Accuracy
	Level 1	Level 2	Level 3	Level 4	
Logistic Classifier	0.00	0.65	0.76	0.33	0.69
SVM	0.00	0.65	0.75	0.1	0.68
Random Forest	0.17	0.77	0.82	0.69	0.79
XGBoost	0.29	0.81	0.85	0.70	0.82

6.3 Importance of features

The importance of features in Random Forest and XGBoost are shown in Figure 11 and 12 respectively. Although the ranks of feature importance in two models are different, there are same features shown in the list of top 15 important features, including 'Traffic_Signal', 'Distance', 'Start_Lng', and ‘Duration_time’. This indicates that the location and duration time are key factors to predict severity as well as some POI information.

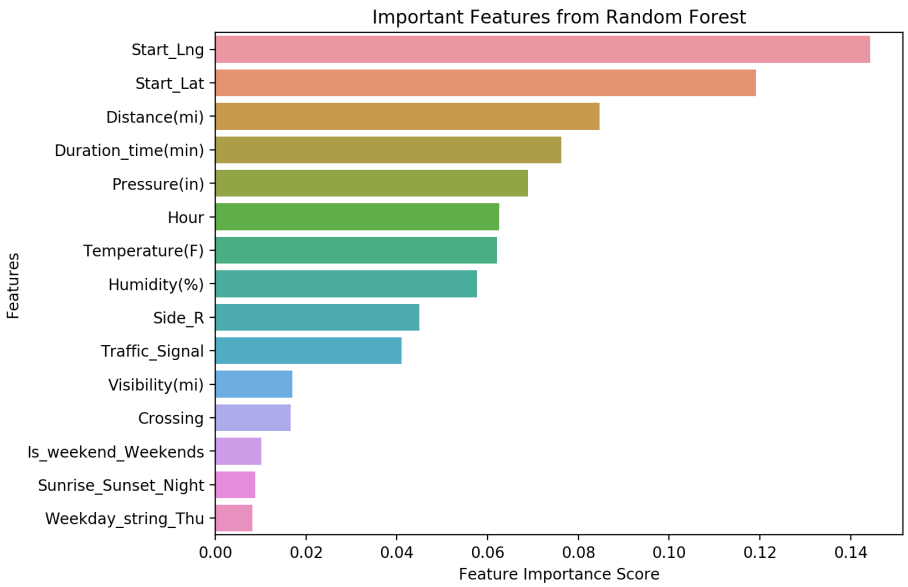

Figure 11. Importance of features in Random Forest model

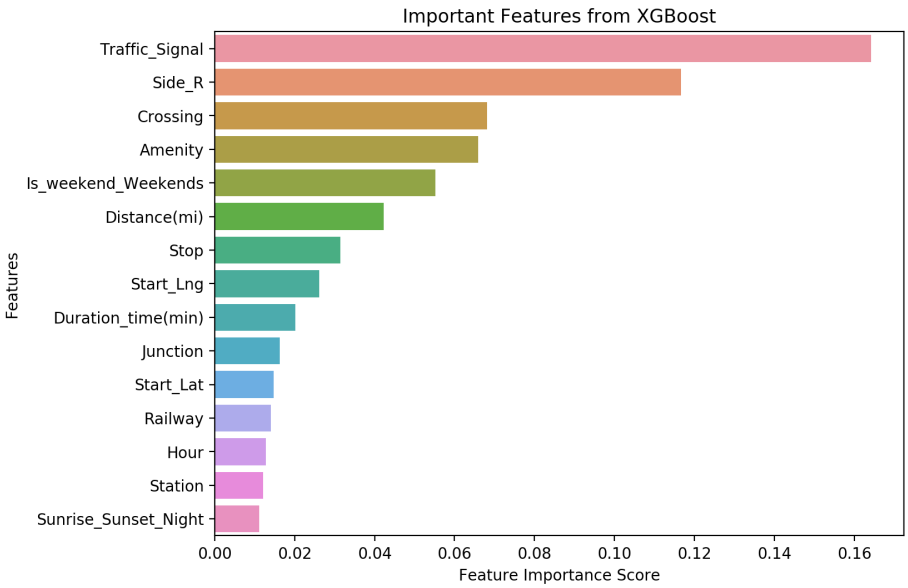

Figure 12. Importance of features in XGBoost model

7. Conclusions

In this project, I analyzed the accident data from February 2016 to June 2020 in US. There are over 3.4 million accidents are collected in this dataset. Over 95% accidents are labeled as medium severity (level 2 or 3), whereas only 0.8% events are labeled as loew severity (level 1). So the ability to distinguish between level 2 and 3 severity accidents would be the key factor that affect the accuracy of statistical models.
The locations and time of US traffic accidents show strong heterogeneity. Most accidents happened in the west coast US and eastern US, but fewer accidents in the mid-US are recorded in the data set, except for along the national high ways. And the top 3 states by the number of accidents are California(CA), Texas(TX) and Florida(FL). In terms of the time when accidents happened, I found that the number of accidents in weekdays are over two times more than that at weekends, though the distribution of accidents in weekdays are even. The accidents in weekdays and at weekends can also

happen at different time. There are more accidents happening at 7:00-8:00 and 16:00-17:00 in the weekdays, while more accidents are happening in the afternoon at weekends.

The correlation between severity and other variables were also assessed in this work. The variables, such as 'Start_Lng', 'Stop', 'Junction', 'Year', 'Weekday', 'Traffic_Signal' and ‘Crossing', are more correlated (positively or negatively) with severity prediction. Altough the 'Wind_Chill' may be a good classifier to distinguish level 1 severity from other levels, as the 'Wind_Chill' for level 1 severity accidents are slightly higher. Due to the high ratio of missing data, this variable had to be dropped for classification.

Machine learning models (Logistic Regression, SVM, Random Forest and XGBoost) were employed to predict the severity of accidents in Georgia (GA), where more accidents happened and state-averaged severity level is higher. Among four models, XGBoost has the highest accuracy (82%), whereas the performance of SVM is the worst. The key improvement of XGBoost is that the mismatches between level 2 and level 3 severity are less and the precision of level 1 severity prediction is higher. Both Random Forest and XGBoost implies that the varibles 'Traffic_Signal', 'Distance', 'Start_Lng', and 'Duration_time' are important features in severity prediction.

8. Discussions

In order to have a better evaluation on four models, it is suggested to use cross-validation to estimate the accuracy scores. Due to the limitation of the time, I will leave this for future work. It is also possible to achieve a higher accuracy of a single model through fine-tuning. According to Abellán, et. al., 2013, some variables, like age of drivers, cause of accidents, are also be important in severity analysis. These variables are either not included in the dataset or not used in my project. Gathering those data and including more preprocessing methods may contribute to a better severity classifier.

9. Reference

- Abellán, J., López, G., & De OñA, J. (2013). Analysis of traffic accident severity using decision rules via decision trees. Expert Systems with Applications, 40(15), 6047-6054.

- Moosavi, S., Samavatian, M. H., Parthasarathy, S., & Ramnath, R. (2019a). A Countrywide Traffic Accident Dataset. arXiv preprint arXiv:1906.05409.

- Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019b, November). Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights. In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 33-42).

- World Health Organization. (2018). Global status report on road safety 2018: Summary (No. WHO/NMH/NVI/18.20). World Health Organization.