Predicting the Car Accidents' Severity

1. Introduction

1.1 Background

According to a report on road safety from WTO in 2018, about 1.35 million people die every year on roads worldwide, and the traffic injuries are now becoming the leading killer of people aged 5-29 years. Facing such a high fetal rate, is there anything we can do to decrease the probability of car accident or at least lower the fetal rate? The answer might be yes. We can look at the historical car accident data and see what we can find out some clues from it. Therefore, this project aims to find out which characters influence the severity of a car accident based on historical car accident data using machine learning. The result might be helpful to give advance warnings on possible car accident and therefore decrease the probability and the fetal rate of the car accident.

1.2 Problem

This project uses traffic accidents records in the United States from February 2016 to June 2020 to find which features determine the severity of a car accident. The aim of this project is to forecast the severity of a car accident in the future and provide warnings in situations with high predicted probability of severity car accident.

1.3 Interest

The result of this project can offer some useful insights to the local transport authorities. For example, they can set addition signs or certain speed limitation in regions that are more likely to have severity car accident or during times with higher

chance of severity car accident. In addition, it can also be used to provide advance warnings or notifications to drivers exposed to an environment with high probability of severity car accident by tracking drivers' real-time location and the current weather, traffic or other conditions.

2. Data acquisition and cleaning

2.1 Data source

The dataset I use for this project can be found in Kaggle.[1] It has been collected in real-time, using multiple Traffic APIs, and it contains approximately 3.5 million accident records collected from February 2016 to June 2020 in the United States.

2.2 Data cleaning

The dataset includes 49 columns, including severity, the start and end date and time, the latitude and longitude of the place where the accident happens, the length of the road extent affected by the accident, the weather condition and so on.

First, I drop the variables that contain little information to forecast the severity of the car accident but has a large amount of missing value, including source, TMC[2], latitude and longitude of the end point, zip code, distance and street number[3]. In addition, sunrise-sunset, civil-twilight nautical-twilight and astronomical-twilight are all used to define day or night, and they give same value at most of the time. I only keep civil-twilight in the final dataset, because whether the natural light is enough for a driver to see clearly might be more related to the severity of a car accident. Second, I drop the observations with missing value for variables that might be highly related to the severity of car accident, such as visibility, weather condition and civil-twilight. Last, for variables that might have influence on the

[1] https://www.kaggle.com/sobhanmoosavi/us-accidents
[2] It identifies whether a car accident has a Traffic Message Channel (TMC) code or not. This code provides more detailed description of the event.
[3] The reason to delete this variable is that this dataset already contains street name and more than one third is missing values for street number.

severity of car accident but contain a lot of missing values (such as temperature, wind chill, humidity, pressure, wind speed and precipitation), I use scatter plot and correlation table to show whether there is a clear correlation between the severity of car accident and the numeric variables, and the correlations between the severity of car accident and the categorical variables are shown by linear regression in the next section. In addition, the variable 'wind_direction' contains several confusing values. For example, 'Calm' and 'CALM', which have very different average value for severity, indicating that they stand for different wind directions. Therefore, I drop the variable 'wind_direction' as there is not any further explanation for the meaning of each value.

2.3 Feature selection

After cleaning the data, the dataset has 3429500 observations and 37 columns (or features). Variables like visibility, weather condition, city, weekday and civil-twilight have straightforward connection with the probability and severity of car accident and therefore should be included in the model.

While, as mentioned in the last section, variables such as temperature, wind chill, humidity, pressure, wind speed and precipitation may or may not influence severity of car accident. Considering the severity of car accident is categorical variable, I use box plot to see which of these variables are closely correlated with the car accident severity and therefore should be included in the model.

2.3.1 temperature and the car accident severity

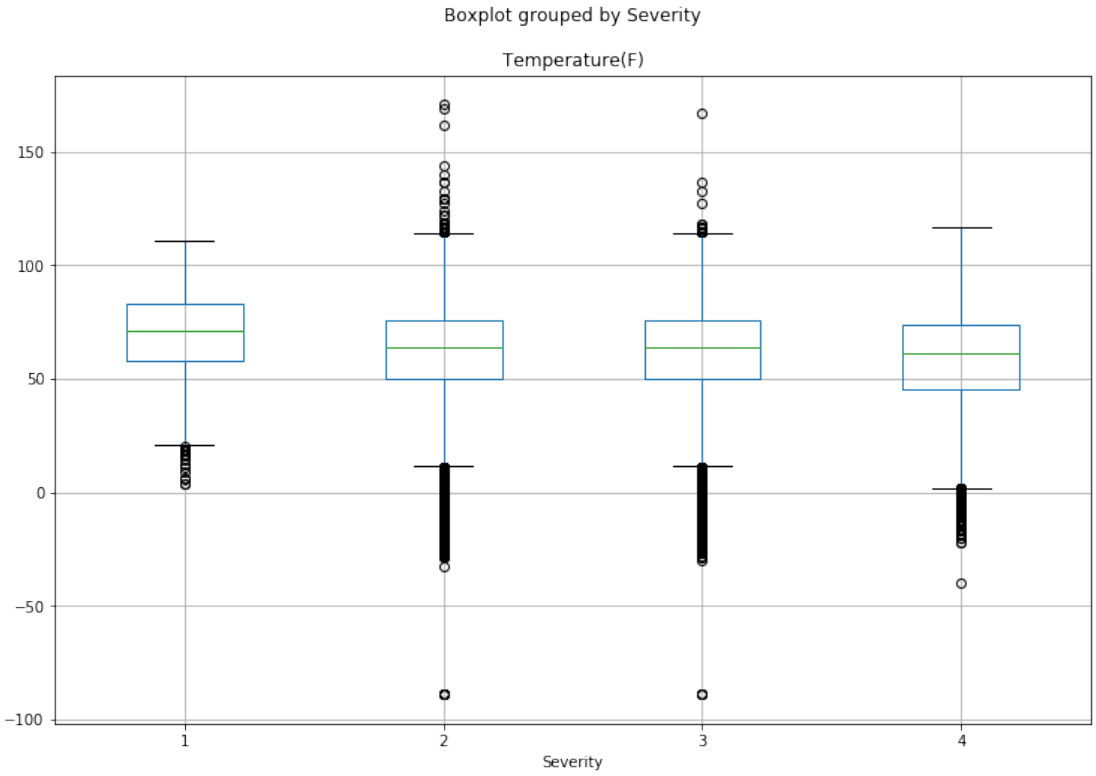


Figure 1. Box Plot of Temperature and Car Accident Severity

In figure 1, the median temperature for less severe car accident is slightly lower and the extreme temperature is associated with more severe car accident. The reason might be that drivers are more easily get distracted and pay less attention to the road when weather is not very pleasant. Therefore, I divide temperature into three categories, temperatures below 0 are grouped as cold, temperatures between 0 and 110 are categorized as temperate, and temperatures above 110 are categorized as hot. Table 1 shows the summary statistics for car accident severity of different temperature groups.

As shown in table 1, the cold weather group has a higher mean and min for car accident severity. Besides, as most of the days are marked as temperate, it is reasonable that the temperate group has a larger count of car accident. In conclusion, temperature should be included in the model to forecast the car accident severity.

	Severity			
	mean	count	min	max
temperature				
cold	2.408922	6456	2	4
temperate	2.338090	3411094	1	4
hot	2.242884	527	1	4

Table 1. Summary Statistics of Car Accident Severity for Different Weather Groups

2.3.2 wind chill and the car accident severity

As shown in figure 2, the large negetive value of wind chill is associated with more severity car accident, therefore wind chill should be included in the model.

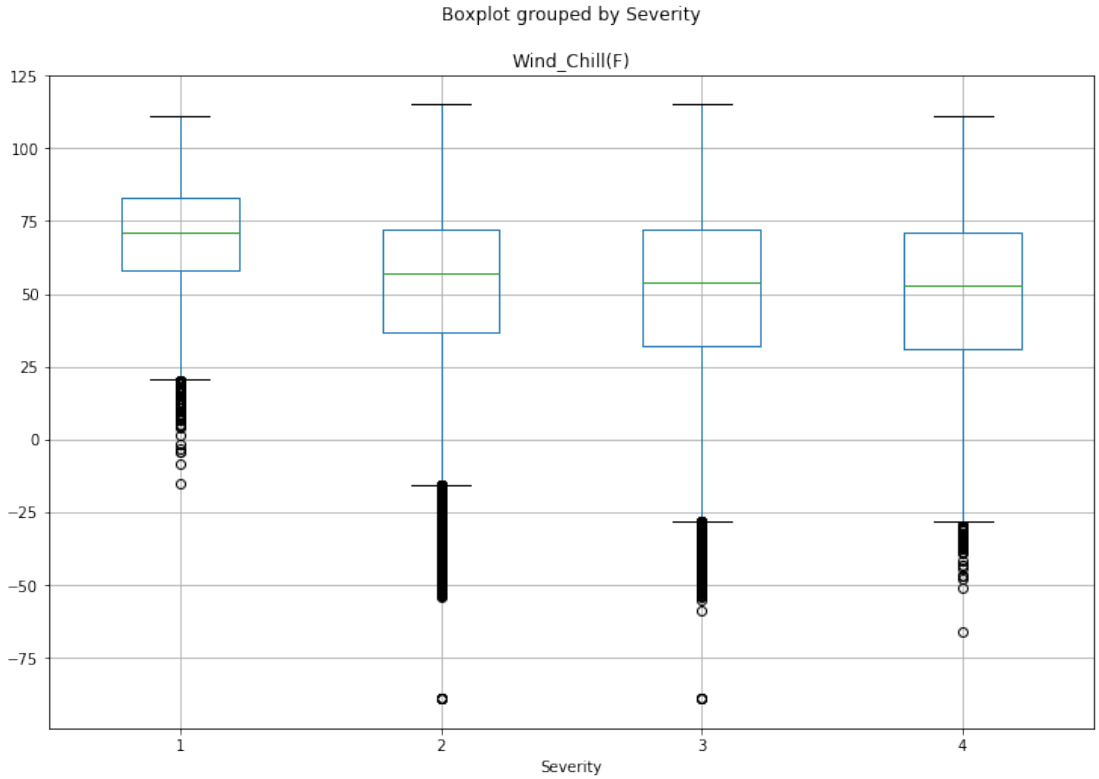


Figure 2. Box Plot of Wind Chill and Car Accident Severity

2.3.3 humidity and the car accident severity

In figure 3, the severe car accident (severity > 1) are associated with higher level of humidity. In other words, the humidity is related to the severity of car accident and therefore should be included in the model.

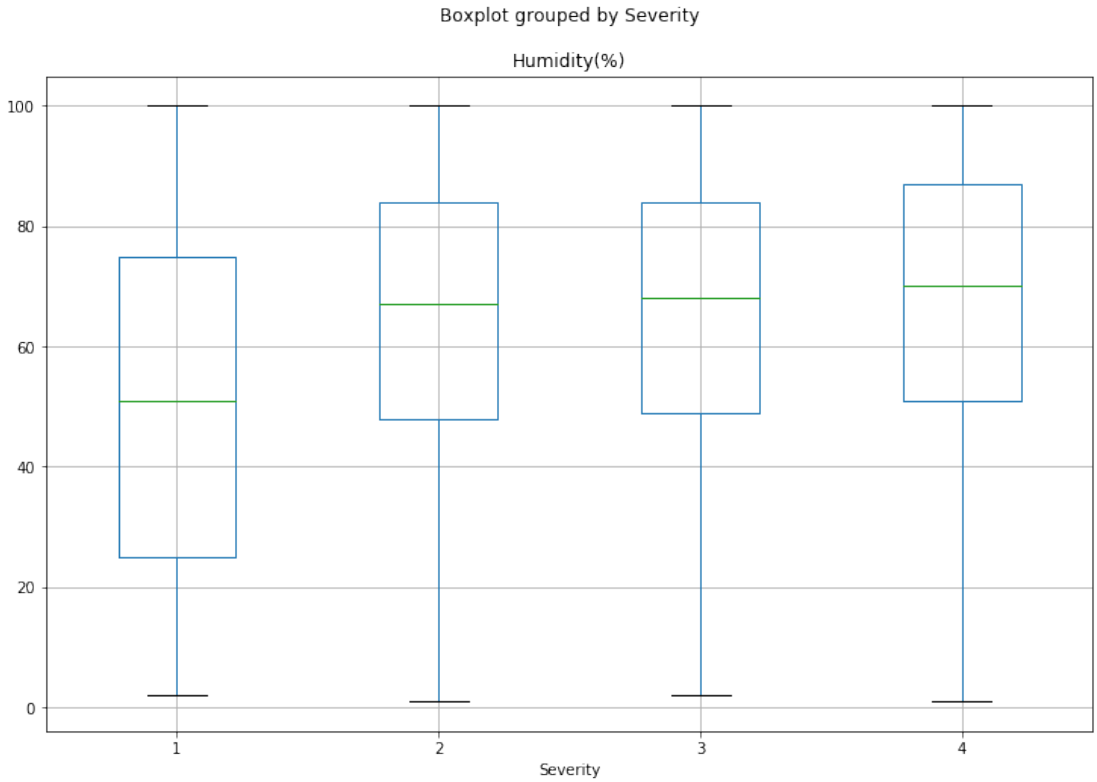


Figure 3. Box Plot of Humidity and Car Accident Severity

2.3.4 air pressure and the car accident severity

As shown in the graph below, the correlation between air pressure and car accident severity is not clear. The medians of air pressure for car accidents of different severity are basically around 30, so air pressure will not be included in the model.

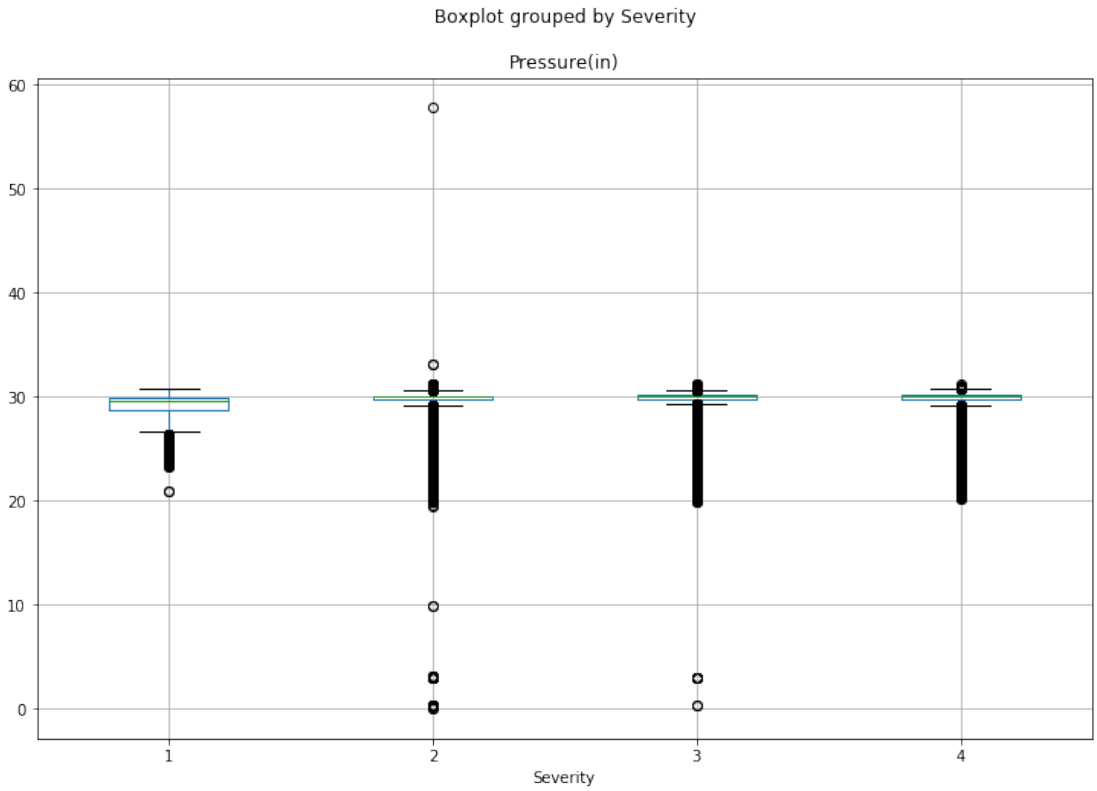


Figure 4. Box Plot of Air Pressure and Car Accident Severity

2.3.5 wind speed and the car accident severity

In figure 5, the medians of wind speed for car accidents that have different level of severity are generally around zero, while severe car accidents have more large value for wind speed, therefore I create a new dummy variable named ‘strong wind’ , which values 1 when wind speed larger than 50 miles per hour and zero otherwise, to indicate whether the wind speed is high.

Table 2 shows mean of the car accident severity grouped by the wind speed. The car accidents happen when the wind spend is higher than 50 miles per hour have a larger average regarding severity, indicates the car accident severity might be influenced by wind speed.

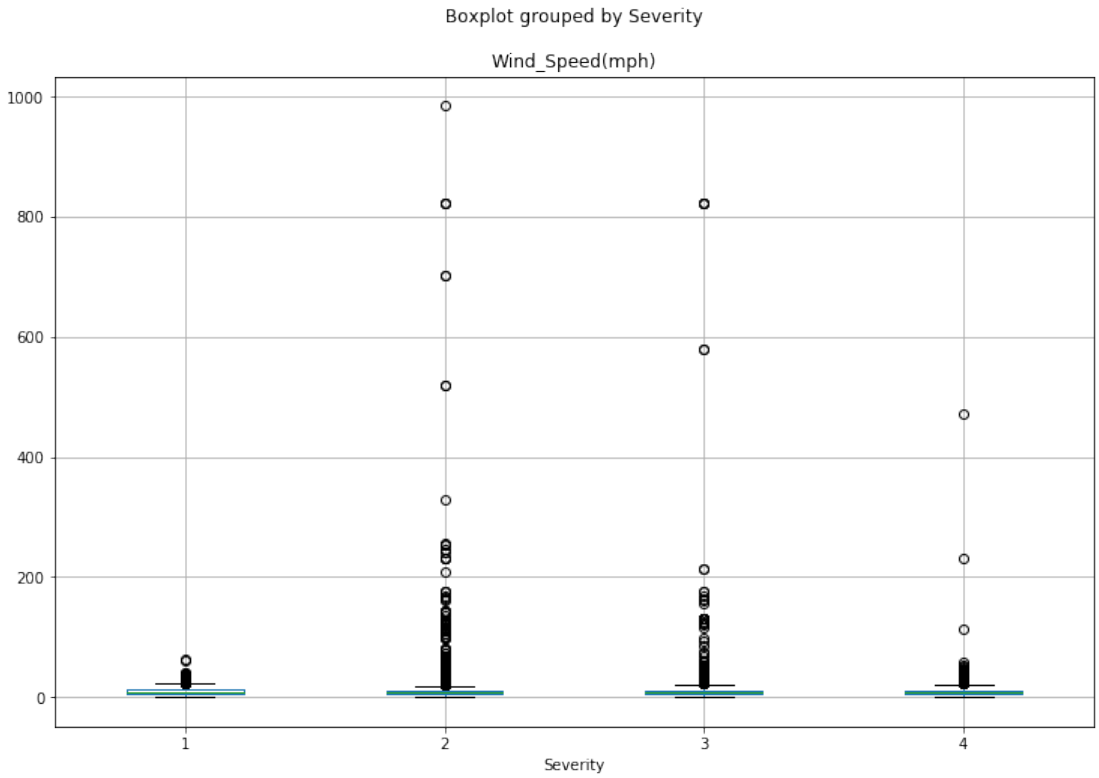


Figure 5. Box plot of wind speed and car accident severity

	Severity	
	mean	count
strong_wind		
False	2.338369	3429341
True	2.364780	159

Table 2. Mean Car Accident Severity Grouped by Wind Speed

2.3.6 precipitation and the car accident severity

Figure 6 shows the boxplot of precipitation and the car accident severity. Similar to wind speed, the car accident severity shows correlation with larger precipitation. Thus, I create a new dummy variable named 'large_Precipitation' indicates

precipitation larger than 2 inches. Table 3 shows the mean car accident severity for large precipitation.

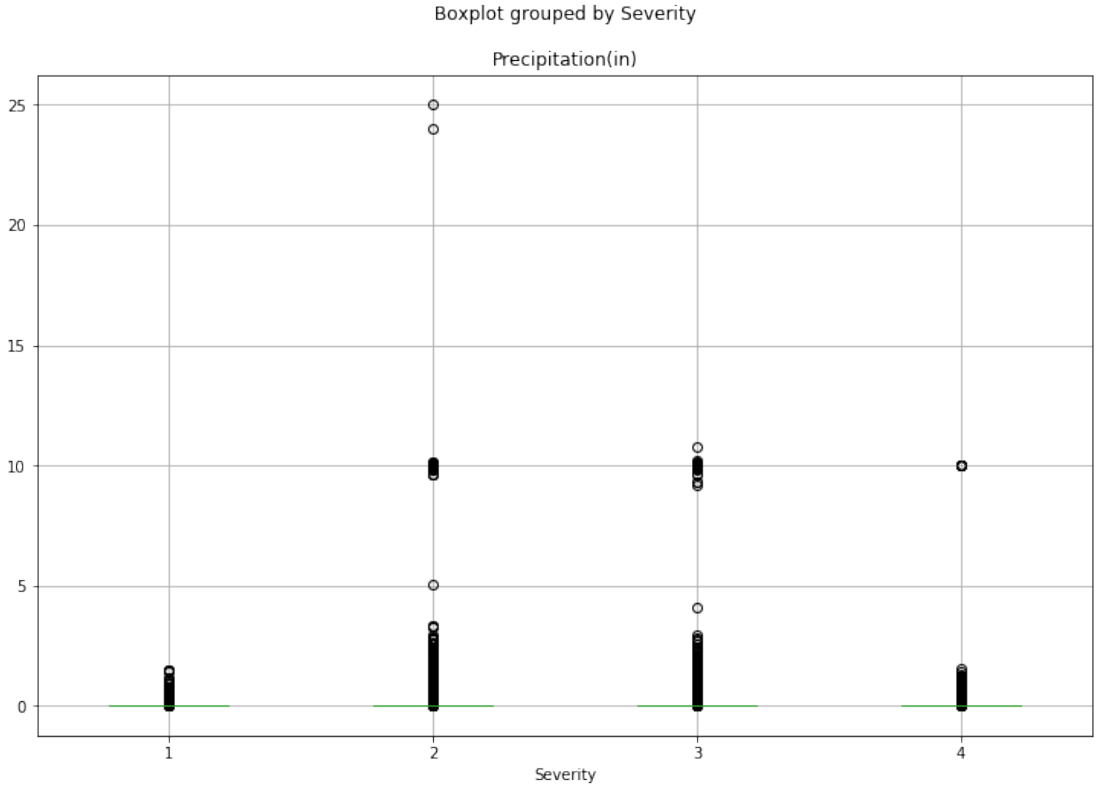


Figure 6. Box Plot of Precipitation and Car Accident Severity

	Severity	
	mean	count
large_Precipitation		
False	2.338357	3429106
True	2.449239	394

Table 2. Mean of Severity and the Number of Car Accident Grouped by Precipitation

The following table shows the variables name and description that will be included in the model to forecast the severity of car accident. For variables 'Humidity(%)' and 'Wind_Chill(F)', I drop the observations with missing value. For variables

'large_Precipitation', 'strong_wind' and ‘temperature’, I replace the missing value with the most frequent value (zero for 'large_Precipitation' and 'strong_wind, ‘temperate’ for ‘temperature’). The string variables in the following list will be used to draw choropleth map in which areas are shaded in proportion to the number of accidents.

Variable Name	Type	Description
temperature	categorical	Temperature lower than zero is marked as cold (-1), between 0 to 110 is marked as temperate(0) and above 110 is marked as hot(1).
Wind_Chill(F)	continuous	It shows the wind chill (in Fahrenheit).
Humidity(%)	continuous	It hows the humidity (in percentage).
strong_wind	categorical	Wind speed over 50 miles per hour is classed as strong wind and vaues 1.
large_Precipitation	categorical	Precipitation larger than 2 inchesis categorized as large precipitation and values 1.
dayofweek	categorical	It shows the weekday when the accident happens.
hour	categorical	It shows the accident happens at which hour
Civil_Twilight	categorical	It indicates dayshown as 1 or night shoqn as 0
Side	categorical	It shows the relative side of the street (Right shown as 1 /Left shown as 0) in address field.
Visibility(mi)	continuous	It shows visibility (in miles).
City	string	It shows the city in address field.
Street	string	It shows the street name in address field.
Start_Lat	string	latitude in GPS coordinate of the start point.
Start_Lng	string	longitude in GPS coordinate of the start point.

Table 3. Feature List

After finishing the selection of features, all categorical variables are assigned numeric values and then all the variables are normalized to avoid bias.

3. Exploratory data analysis

Panel A. Summary Statistics for wind chill and visibility

	Wind_Chill(F)	Visibility(mi)
count	1635505	1635505
mean	53.61466	8.965905
std	23.75038	2.990811
min	-59	0
25%	35.8	10
50%	57	10
75%	73	10
max	115	130

Panel B. Summary Statistics of other Categorical Variables

temperature	count	mean
cold	5589	2.41
temperate	1629725	2.30
hot	191	2.17

Civil_Twilight	count	mean
day	1259028	2.28
night	376477	2.36

strong_wind	count	mean
yes	73	2.40
no	1635432	2.30

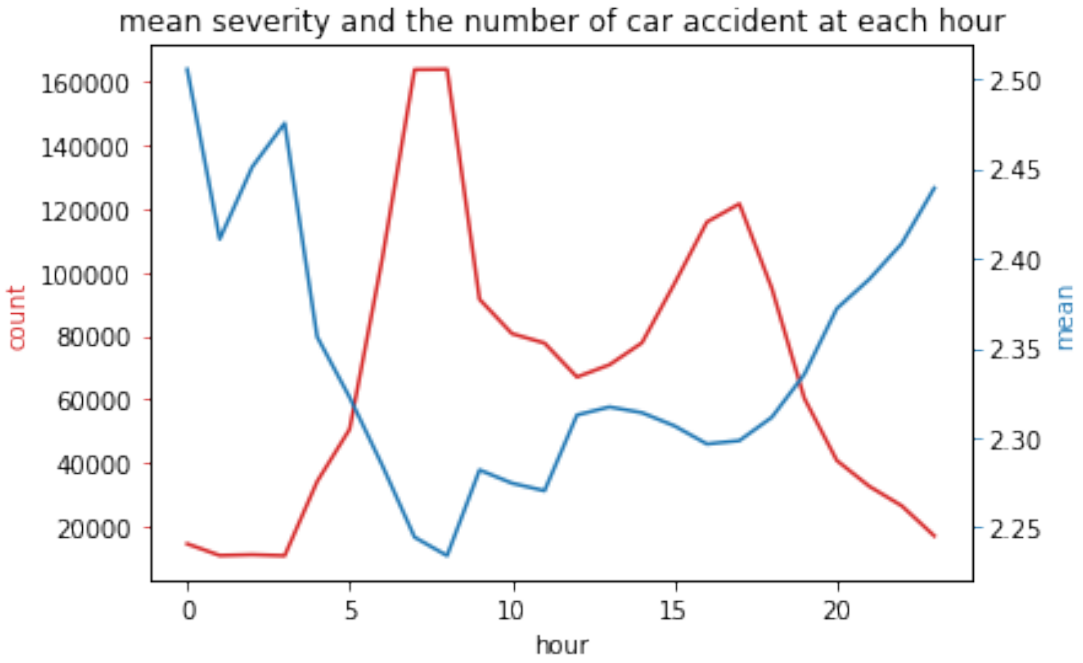
large_Precipitation	count	mean
yes	104	2.42
no	1635401	2.30

Side	count	mean
right	1331821	2.34
left	303684	2.13

Table 4. Summary Statistics

In Table 4, we can see the summary statistics for continuous variables and categorical variables. Panel A shows that the final sample contains 1635505 observations. Wind chill varies from -59 to 115, and visibility varies from 0 to 130. As shown in Panel B, the mean of severity is larger for cold weather, during the night, for strong wind, for accidents happened in the right side and when precipitation is large.

Panel A. Time of day and Severity



Panel B. Day of Week and Severity

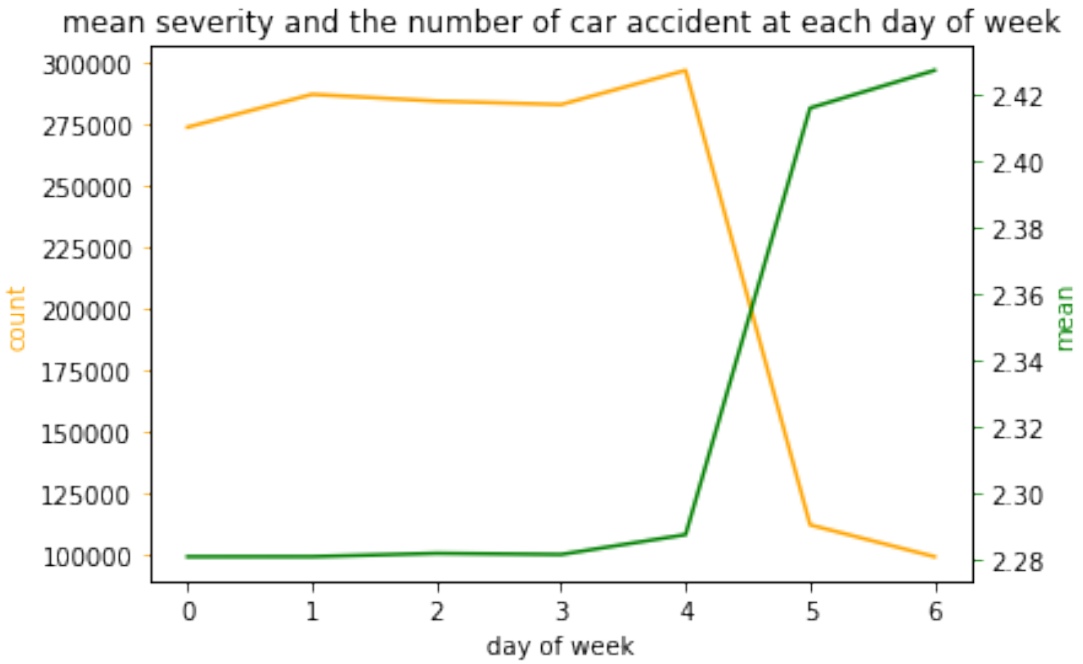


Figure 7. Box Plot of time variables and Car Accident Severity

As shown in figure 7, the severity and the number of car accident fluctuate a lot during a day and over a week. In Panel A, the number of car accident reach the peak at around 9 o'clock and then gradually decrease till about 18 o'clock in accordance with the traffic morning and evening peak. The mean value of severity is clearly higher during 20:00 in the evening to 3:00 next morning, which might because that drowsy driving is more likely to happen during the night. In Panel B， significant more car accidents happen during weekday in comparison with weekends, but the car accidents happen during weekends are more likely to be more serious, which might be explained that people tend to travel with family using one car during weekends, therefore more people get involved if car accident happens.

4. Predictive Modeling

There are two categories in supervised machine learning: linear regression and classification. But as the target variable (severity of the car accident) is not continuous, only classification model is available predict the severity. Specifically, K Nearest Neighbor (KNN), Decision Tree, Support Vector Machine and Logistic Regression will be used build the model. The dataset contains 1635505 observations after dropping the missing values, which takes too much time to run the codes using Jupyter, so I randomly choose 30000 observation as final dataset.

The final dataset is split into train set (80%) and test set (20%). The train dataset is used to train the model, and the test set is used to test the accuracy using Jaccard, F1-score and Log Loss.

The following table shows the frequency table of the true value of severity and the forecasted severity using different classification algorithms. As shown in table 5, all of the four methods misclassify the severity of some of the car accident as the most frequent category (in which severity equals to 2).

value	y_test	knn_yhat	DT_yhat	SVM_yhat	LR_yhat
1	105	0	0	0	0
2	4256	5904	6000	5995	5995
3	1456	96	0	5	5
4	183	0	0	0	0

Table 5. Frequency Table of Forecast Results

Table 6 shows model evaluation results. As most of the forecast values of the decision tree, support vector and logistic regression are the same, they have the same Jaccard index and F1-score. While the KNN performs better, as its F1-score is slightly higher than the other three.

	y_test	knn_yhat	DT_yhat	SVM_yhat	LR_yhat
Jaccard index	0.71	0.71	0.71	0.71	0.71
F1-score	0.6	0.59	0.59	0.59	0.59
LogLoss	NA	NA	NA	NA	0.74

Table 5. Model Evaluation

5. Conclusion

In this study, I use supervised machine learning to forecast the severity of car accident. As the original dataset takes too much time to process in Jupyter, I randomly select 30000 observations as the final dataset. The results show little difference for the four classification models I use, and the KNN model gives the best forecast result.

References

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "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, ACM, 2019.

World health organization. “Global status report on road safety.”, 2018.