Prediction of Accidents' Severity Based on Collisions

Introduction/Business Problem section

The automobile has become an important part of daily life. It influences how we live, work, and how we spend our time. It helps you to go to the workplace on time, sends your kids to school, and takes you and your family to a vacation on a beach. However, at the same time that it brings happiness, it also brings danger – car accidents. According to the statistics, there are 16,438 car crashes per day in the U.S. alone [1]. On a global scale, there is approximately 3287 death every day due to car accidents [1]. Generally, people would try not to drive at rush hour, bad weather, or dark place to avoid the chance to get into an accident. However, sometimes driving in a bad condition is unavoidable. For example, your kids are waiting for you to pick them back home after the piano class on a snowy night. This raises the question that how we maintain our daily needs and avoid accidents and related injuries at the same time. Thus, it is necessary to provide a model to predict the severity of an accident, thus one could estimate the dangers of different driving behaviors and stay safe.

Herein, we would build a machine learning model based on collision data to predict the accident severity. Data is provided by SPD and recorded by Traffic Records downloaded from the IBM course on Coursera. The provided dataset will be cleaned, processed, and used to build a classification model to predict whether an accident will lead to property damage or injury.

Data section

The data has a total of 37 attributes and 194673 instances[2]. According to the metadata form, we are going to first remove the data without any explanation. Thus, columns called OBJECTID, INCKEY, COLDETKEY, REPORTNO, EXCEPTRSNCODE, EXCEPTRSNDESC, STATUS, SEVERITYCODE.1. Some columns are duplicated. We only need to keep SDOT_COLDESC or SDOT_COLDCODE, ST_COLDESC or ST_COLDCODE, SEVERITYDESC or SEVERITYCODE. Because one column is the description of the other column. Because the purpose of the analysis is to predict the accident severity, we will remove parameters that describe the damage after the accidents. Thus, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, HITPARKEDCAR.

If we check the unique values in each column, PEDROWNOTGRNT, SPEEDING, INATTENTIONIND only has yes or nan (Figure 1). Thus, it would not be useful for prediction and thus removed.

```
PEDROWNOTGRNT :  [nan 'Y']
SPEEDING :  [nan 'Y']
INATTENTIONIND :  [nan 'Y']
```

Figure 1 Unique values in the column.

In addition, INCDATE and INCDTTM are reorganized to new columns called DOW, HOUR, DAY and MONTH to present the day of a week, hour of a day, day of a month and month of a year. Unknown values are treated as nan value. Because of the large amount of data we have, rows

with nan and empty values are removed. After the series of clean processes, SDOTCOLNUM is removed because there is a unique value for each instance. Columns with more than 10000 unique values will be reorganized to two unique values. The value is set to 1 when severity is likely to be 1 and value is set to 0 when severity is likely to be 2. In the process of model building, all categorical numbers will be adjusted to numerical numbers for further model training.

	SEVERITYCODE	COLLISIONTYPE	JUNCTIONTYPE	SDOT_COLCODE	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	ST_COLCODE	MONTH	DAY	DOW	HOUR	X_INDEX	Y_INDEX	LOC_INDEX	INTK_INDEX	SEG_INDEX	CROSS_INDEX
0	2	Angles	At Intersection (intersection related)	11	0	Overcast	Wet	Daylight	10	3	27	2	14	1	1	1	1	1	1
1	2	Angles	At Intersection (intersection related)	11	0	Raining	Wet	Daylight	10	1	28	2	8	1	1	1	1	1	1
2	1	Angles	At Intersection (intersection related)	11	0	Clear	Dry	Daylight	10	4	20	5	17	1	1	1	1	1	1
3	2	Cycles	At Intersection (intersection related)	51	0	Clear	Dry	Daylight	5	4	15	2	17	1	1	1	1	1	1
4	2	Angles	At Intersection (intersection related)	11	0	Clear	Dry	Daylight	10	3	20	0	15	1	1	1	1	1	1

Table 1 Example of cleaned data set.

Methodology

The data is explored in three main directions, location influence, condition influence, and timing influence. In location influence, accidents are grouped based on longitudes and latitudes. The statistic of accidents based on different areas will be plotted. Further, makers are plotted on the map of Seattle to present the number of cases for different types of severity. In addition, the correlation of accident type and junction type is studied by bar plots. Next, we consider the common factors that are expected to close correlated with accident severity, such as the influence of drug and alcohol, weather, road and light condition. Bar plot is used here to compare at different conditions whether there is a difference in the severity. Furthermore, time is considered. The records are reorganized into hour of a day, day of a week, day of a month and month of year. Similarly, bar plots are used to study the correlation.

To incorporate all the factors into a model to predict an accident severity, a machine learning model is built. Because the current data set only contains severity 1, property damage, and severity 0, injury, a classification model built on K-nearest neighbor, Decision tree, support vector machine, and Logistic Regression. Before training, data in each attribute is enumerated and normalized.

Results

1. location and accident type

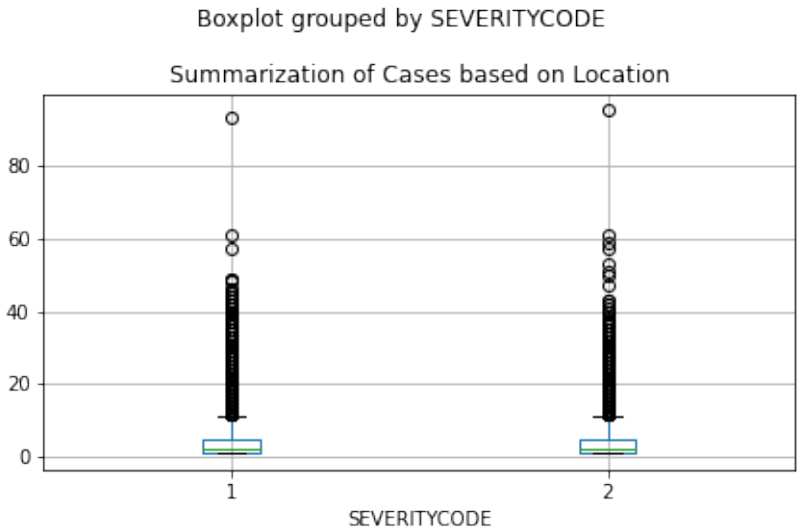

Figure 2 Summary of the number of cases based on location

From the summary of accidents based on various location (Figure 2), the distribution of accidents is inhomogeneous, which means some area is prone to have more accidents than other areas. Thus, we need to find the location with a high frequency of accidents. On the other hand, we can see from the statistic from the picture that there is no strong correlation between location and severity. It will be better if we can plot the represent the location, severity type and accumulated number of cases on the same figure.

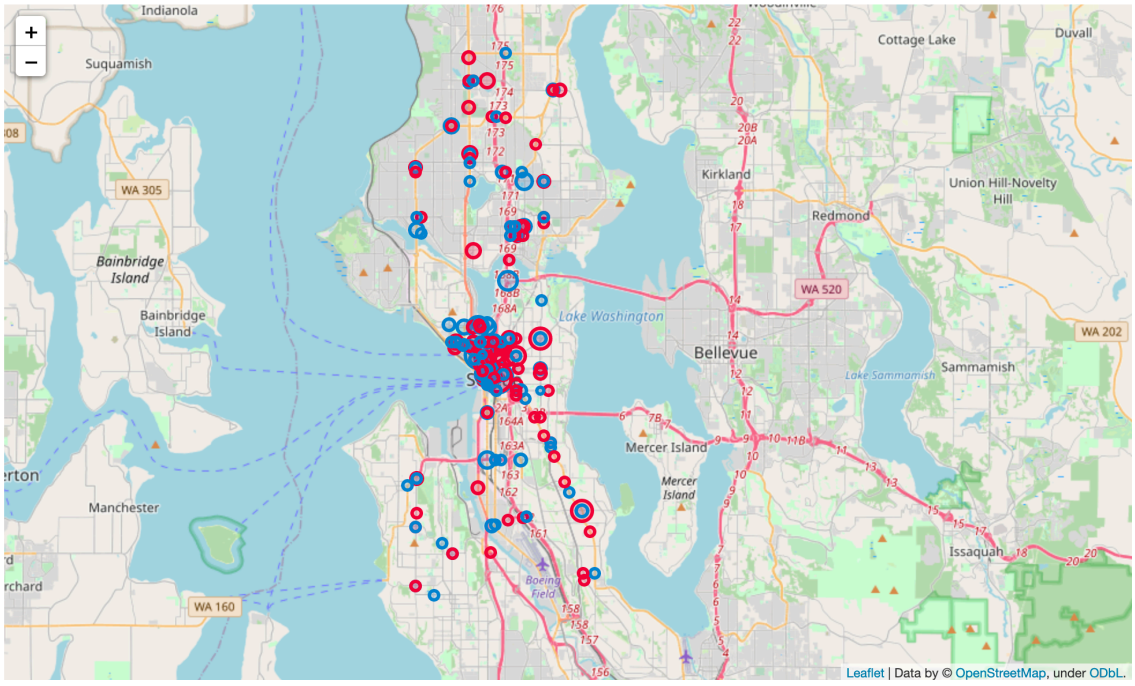

Figure 3 Top 100 locations with most cases of severity 1 and severity 2. The blue color indicates the severity is 1. The red color indicates the severity is 2. The size of the marker is proportional to the number of cases that happened during the research time.

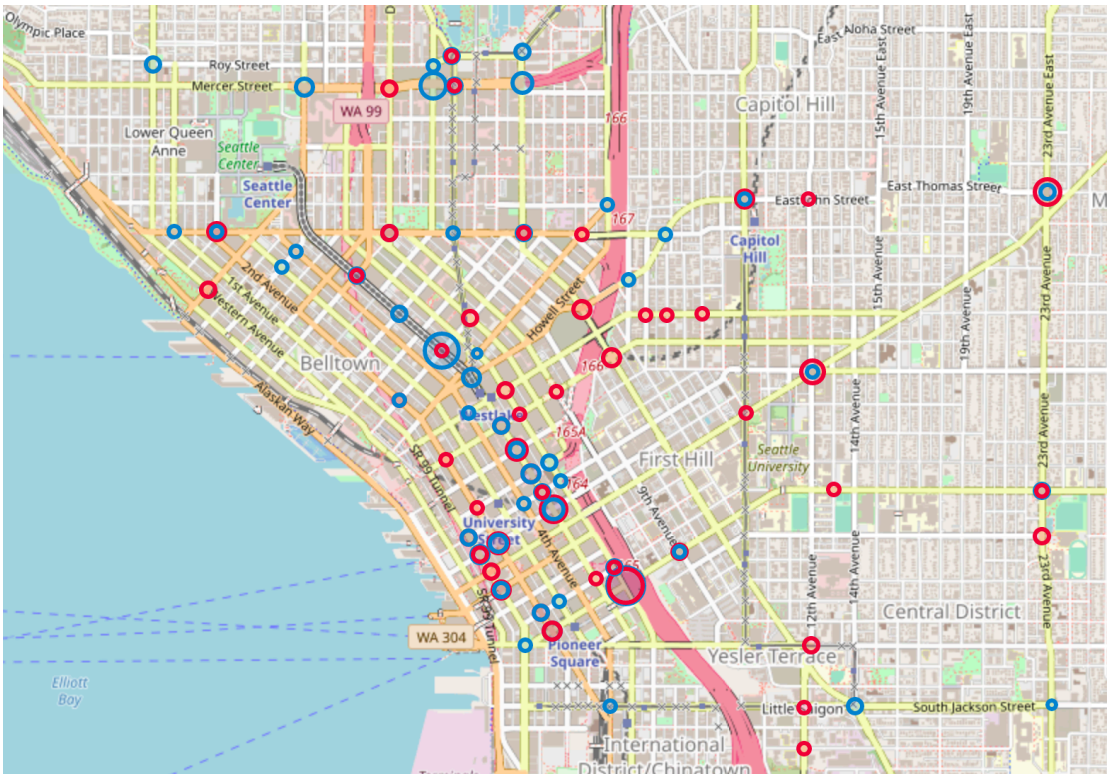

Figure 4 Location is downtown with a high risk for car accidents.

As we can observe from Figure 3, where the top 100 locations with severity 1 (blue) and top 100 locations with severity 2 are plotted (red), the majority of the cases locate at several main roads in the city, such as Rainer Avenue South, 1st Avenue, 4th Avenue, etc. In addition, downtown is the main area where most accidents happened. The data above suggests that it is better to avoid the downtown area when we travel. When we enter the area with a high chance of car accidents, it is better to choose the road without circles or small blue circles to avoid injury.

2. Condition

In the following section, we will explore the factor that might lead to different types of car accident severity. Before the interpretation of the statistics of the data, one thing we need to keep in mind is that the data only contains the cases with the severity of proper damage (0) and injury (1). It does not include serious injury (2b) and fatal accidents (3). Thus, factors, which seem to lead to a few cases in severity 1 and 2, might lead to a lot more cases in severity 2b and 3. On the other hand, we remove the rows with nan, empty and unknown. Therefore, if it is a common behavior to record nan, unknown, or empty value for severity 1 and severity 2, it will likely cause a systematic error. But the chance of this systematic error is very low.

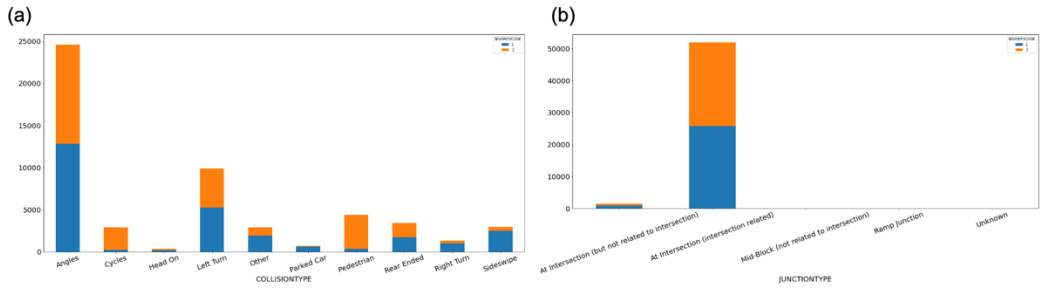

Figure 5 Correlation of collision type (a) and junction type (b) with a number of cases with different severity.

As shown in Figure 5a, angle collision, where one vehicle strikes another vehicle at an angle of approximately 90 degrees, is the main reason for collision with severity 0 and 1. Collision with cycles and pedestrian is likely to cause injury with severity 2. Most of the cases are at an intersection and related to the intersection (Figure 5b).

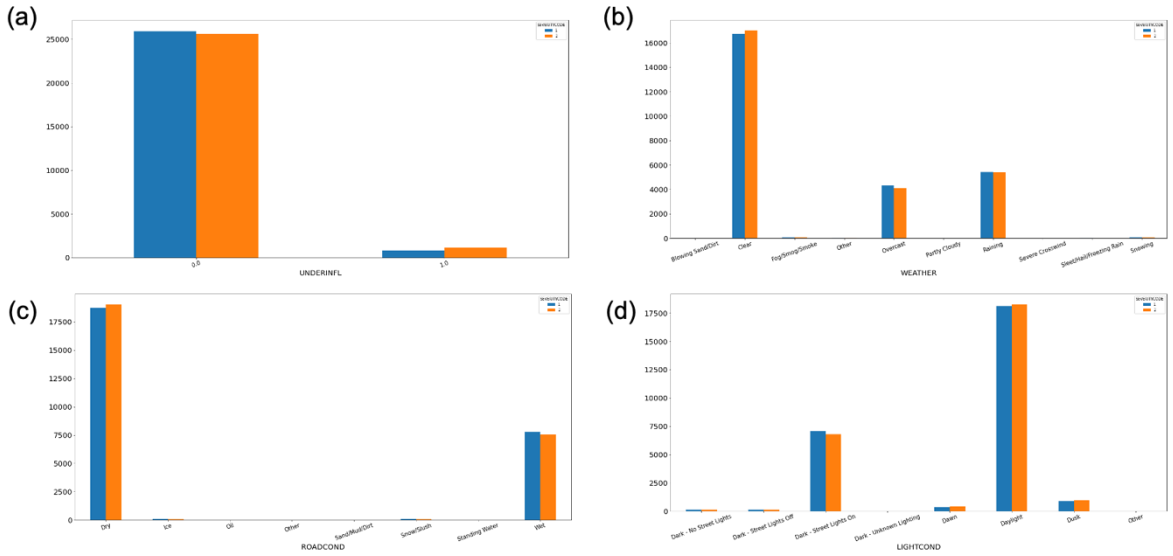

Figure 6 Correlation of influence of alcohol and drug (a), weather (b), road condition (c), light condition (d) with a number of cases with different severity.

There are commonly four factors one would think that will influence the risk of an accident, drug and alcohol influence, weather condition, road condition and light condition. The statistics are presented in Figure 6.

Against one's intuition, most of the cases which cause property damage and injury happened without drug and alcohol influence, but we need to keep in mind the data only shows the cases with severity of 0 and 1. Besides, the chance of injury is higher under drug and alcohol influence. Take a closer look at other factors, severity 0 and 1 does not have a strong correlation with any of the factors mentioned above. But it is clear, accidents with the severity 0 and 1 are likely to happen on a clear, raining, overcast day, when the road is either wet or dry, during daylight or when the streetlight on.

Timing is an important factor to take into consideration because human activity is expected to change at different times. For example, one would expect people to spend more time outdoors during the summer (walk, cycling, and drive), while stays at home in the cold winter. People are prone to drive in the morning and afternoon, but less unlikely to drive at late night. Thus, we summarize two types of accidents based on the hour of a day, day of a week, day of a month and month of a year.

As shown in Figure 7, accidents with the severity of 0 and 1 are relatively less in February and on weekends. During a day, the number of both types of accidents is proportional to the traffic volume. ~ 8 AM and ~5 PM are the time segments that accidents likely to happen. Interestingly, fewer accidents with severity 1 or 2 will happen at the beginning of the month, while there is no significant variance in each month or each week. More injuries prone to happen in the afternoon and from July to October.

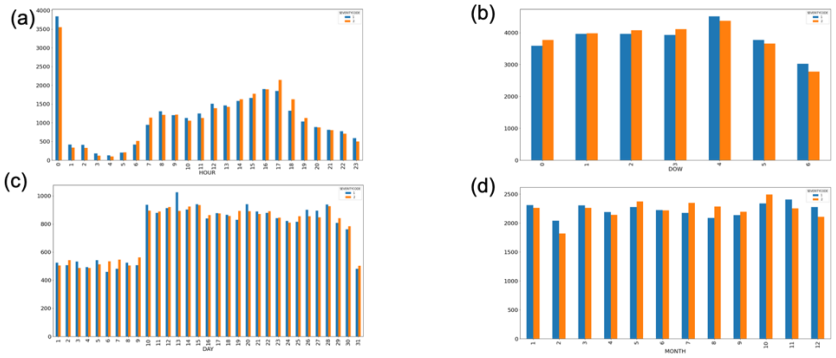

Figure 7 Correlation of hour of a day (a), day of a week (b), day of a month (c), month of a year (d) with number of cases with different severity.

3. Model

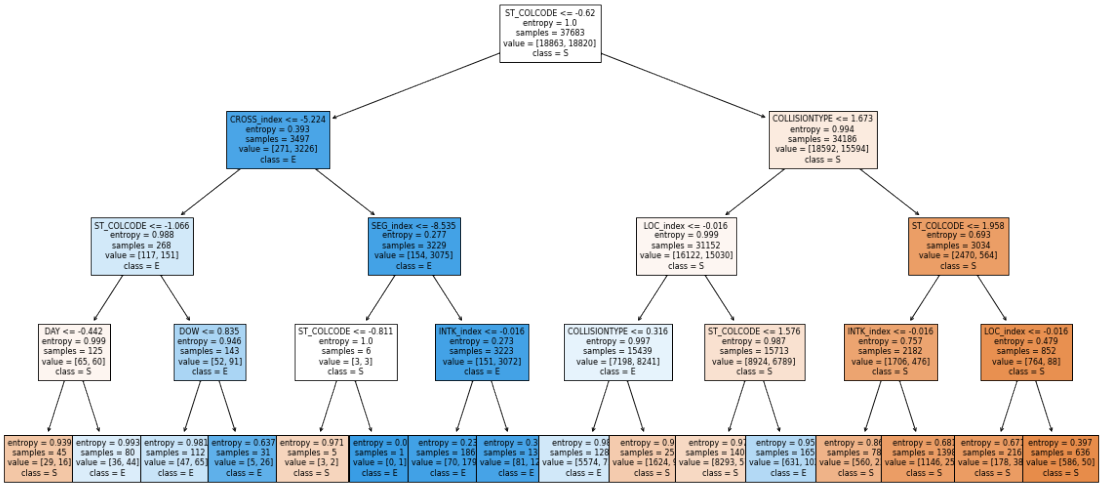

Figure 8 Decision tree built on the dataset with depth 4.

A classification model is trained to predict the severity of an accident based on collision type, junction type, whether under drug or alcohol influence, weather, road condition, light condition,

month, day, day of a week, hour, location, intersection, crosswalk and segment. The best model trained here is by decision tree with a depth of 4 with accuracy around 0.61 and an F1 score around 0.71.

Discussion

By analyzing the collision data in Seattle, we find the location where most accidents happened with severity 1 and 2. We exam the correlation of different factors and timing that are expected to influence the severity, but we find no difference for the current dataset. Moreover, a machine learning model is built to integrate all the factors discussed above to predict the severity of an accident. The model has the potential to be implemented in a navigation system. For example, with a similar amount of time to reach the destination, the navigation system would suggest a road with less probability to get into accidents or injury.

To improve the prediction of severity, a complete dataset includes severity 2b and 3 could be used. Especially, one would want to know whether an accident will cause fatal damage or not. Further, it would be great to discuss the dataset with the people who created it. They would provide valuable insights about which factor to keep and which factor to throw away. Besides, they can help to answer some questions about the dataset. For example, if we plot the light condition against hour of a day, we would notice daylight expands from 0-24. This is against my expectation because I expect daylight means bright conditions.

Conclusion

From the dataset, we successfully extract the information about the correlation of location, number of cases and severity. Unfortunately, we did not observe a strong correlation between the light condition, road condition, and other factors with severity. To further understand the impacts of those factors, it requires a complete dataset which includes severity 2b and 3. With a complete dataset, we can also ask the question about what factor contributes to the death in an accident, what would be the best time to travel, what would be the best weather to travel and etc. A classification machine learning model is built to integrate all factors available to predict the severity of an accident. The best prediction is made by a decision tree with a depth of 4.

Reference:

[1]50+ Car Accident Statistics in the U.S. & Worldwide (Source: https://www.thewanderingrv.com/car-accident-statistics/)
[2] Dataset proved by Coursera IBM course.