# Car Accident Severity Prediction

A. Introduction

What are the factors that will influence the severity of a car accident? The weather? The number of people in the car? Or the location and road condition? If we can find the factors that are likely to cause a severe car accident from past real reports, we can try to lower to severity in the future. If we can predict the severity of a car accident given some known condition, we can make alerts to drivers and passengers when needed. This model will greatly reduce the loss of lives and properties from car collisions.

B. Data Description

The dataset includes all types of collisions happened in Seattle from 2004 to present recorded by Traffic Records of Seattle Government. There are 37 attributes in total but not all attributes will be used to build the model. The severity of a collision is recorded using a code from 0 to 3. Other attributes include location, collision type, date, time, weather, road condition, etc. Most attributes are strings but some are numbers.
A detailed data description will be attached in another document.

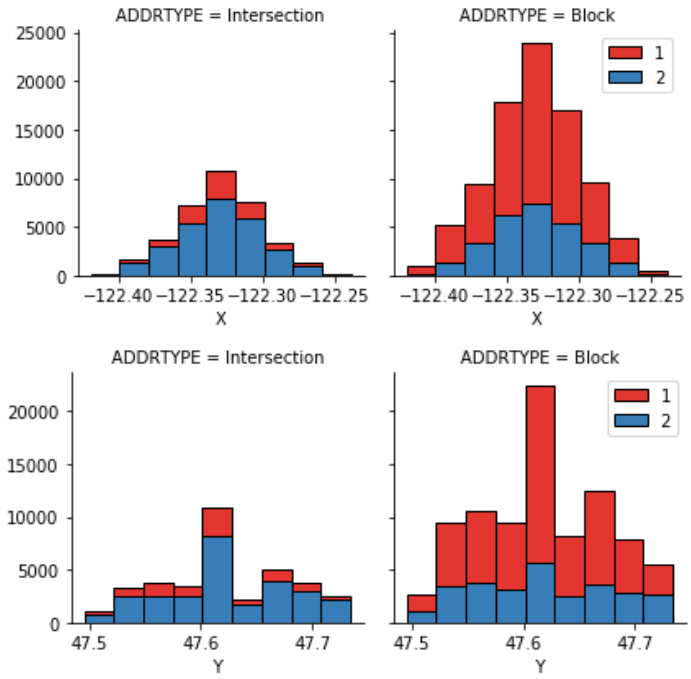
C. Methodology

The majority part of this project is the data engineering work before we build a model. After reviewing the whole dataset and description, I decided to choose 13 factors that may have effect on the severity of a collision. These factors are decided before an accident happen, that means, can be used for prediction. As for factors that are recorded after the accident are

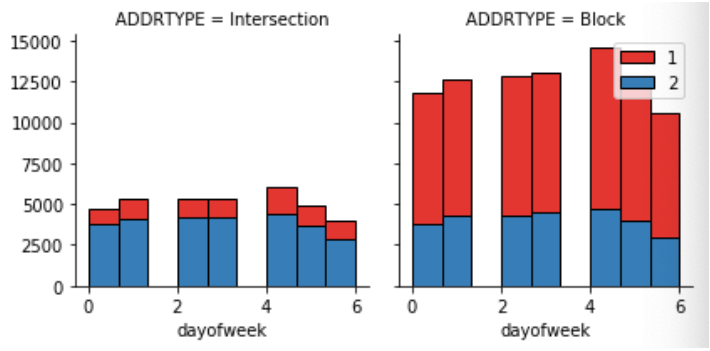
not considered here because we cannot know how many people will get injured in a collision or will there be a bicycle involved.
So the attributes I chooses include the address type, date and time, location, weather, road condition and some other factors. But we still need to do some engineering concerning the data.
For date and time, after exploring the histogram of different severity across day of the week and time of the day, I divide the date into weekend or weekday, and early morning or not. For weather and road condition and collision type, after group by these factors I find that every attribute will result in different proportion of severity. The solution is to transfer them into dummy variables. It makes my final training set a bit wide with 37 features in total and above 190,000 observations.
The machine learning method I choose to build a model is the K nearest neighbor.

D. Results

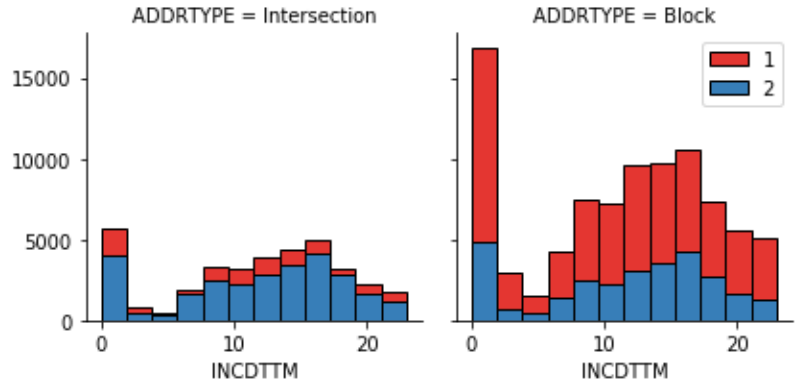
From the data analysis and graph analysis it can be easily found that date, time, weather, road condition and more factors have obvious effect on the severity of collisions.
In general, the location do have an effect on the severity. In some places, severe accident are more likely to happen while in some places things are in verse.



Whether it is on weekend is another factor while collisions happened on weekend are less severe.




As for the time in a day the disparity is more obvious and sensible because collisions in the early morning are less severe, due to less traffic.




For the analysis of other factors, whether drivers were speeding or whether pedestrians had obeyed the traffic rules will increase the probability of severe accident by a large amount. The difference of speeding is 8% and that of pedestrian not granting rule reaches 61% on average. And if the collision is hitting a parked car, the probability of severe accident (code=2) will decrease by approximately 24% precent.

E. Discussion

When I was cleaning the data I meet the first question: how should I deal with the “NaN” values? As we can see most of the PEDROWNOTGRNT and SPEEDING values are null, I want to drop the two features at first but it come to me that these could be important factors in a collision. So I assume these values as 0. For other features, as the NaN values are relatively small compare to the total size, I will drop these instances.

However, I am not sure what I am doing is proper. In real cases I think we need to talk with the people who take these records or maintain the database.

Another problem is the choose of dummy variables. I am wondering if there are too many dummies that make the running speed of my code becomes a little bit slow. But it is hard to divide these dummies into groups because there is not enough correlation between the proportion of severity and the dummies.


F. Conclusion

In this project I build a model to predict the severity of a car accident on some pre-decided conditions. It can be used in our transportation system for alert and reminder. Transportation department might be interested in it because they will know where and when to send warning messages to drivers. Other business that in transportation industries might want to utilize this model too. For example, it can be used in the auto-driving system and GPS guiding system.

The model reaches an accuracy of 0.7 but it can be improved from many other aspects. Future investigations could put into the pre-processing of data and exploration of other models.