Capstone Project - Car accident severity

Chris Han

Introduction/Business problem

Road accidents constitute a major problem in our societies around the world, also a significant source of death, injuries, property damage and a major cause of traffic jam. As a result, traffic safety has increasingly become a top issue today. As we all know a car accident cannot be happened single headed, there are varies of factor such as weather, driver's conditions and ext will cause a car accident, and depend on different situation, there are different level of severity, Therefore, this study is mainly aimed to evaluate the most effective factors in severity of these accidents based on the limited data that I found online.

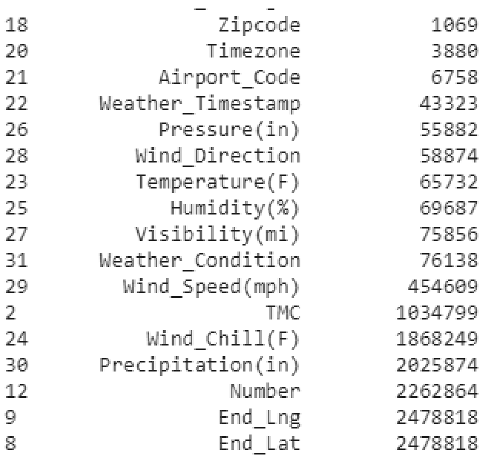
Data Description

- The dataset I chose for this study was US-Accidents: A Countrywide Traffic Accident Dataset which I found on Kaggle. This csv file covers around 3.5 million accident records from 49 states of the US, and this file has being collected continuously started from February 2016.

- This file has 49 different attributes for every single accident in record, such as temperature, humidity, and ext, some of the attributes are very useful for my study, so I believe this dataset will fit perfectly.

- For data cleaning, I printed out the missing value table, and clearly we can see that TMC, wind chill, precipitation, number, end Ing and lat has a relatively large number of missing values, so for accuracy, I decided to drop those columns form the table.

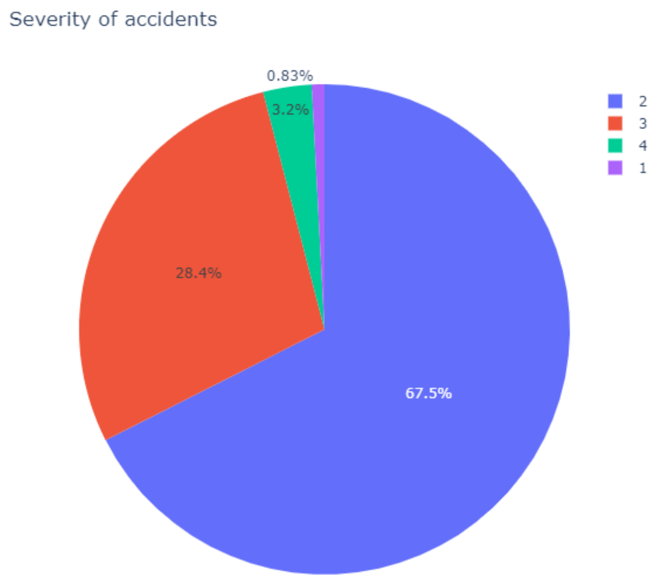
- I also use the example dataset on coursera for more reference. Since I only have to use one or two attributes of it, so I did not make any data cleaning for this table



```
18              Zipcode        1069
20             Timezone        3880
21         Airport_Code        6758
22    Weather_Timestamp       43323
26         Pressure(in)       55882
28       Wind_Direction       58874
23       Temperature(F)       65732
25          Humidity(%)       69687
27       Visibility(mi)       75856
31    Weather_Condition       76138
29      Wind_Speed(mph)      454609
2                   TMC     1034799
24        Wind_Chill(F)     1868249
30    Precipitation(in)     2025874
12               Number     2262864
9               End_Lng     2478818
8               End_Lat     2478818
```

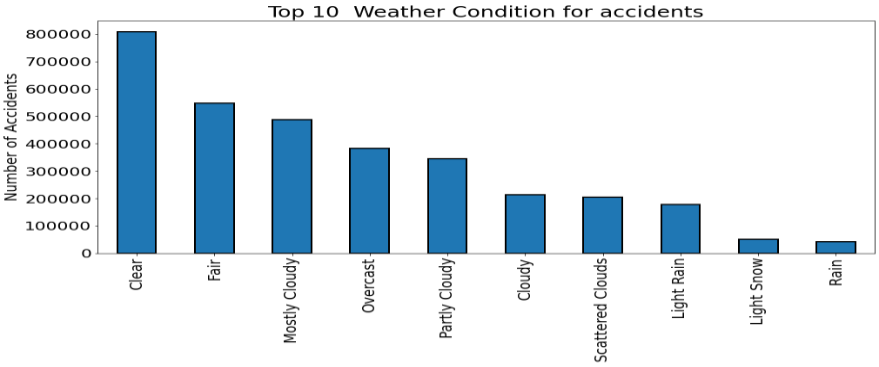
Methodology

I used colab on google for all of my coding and study. And for my dataset, I mainly forces on the relation between severity with other features. So first, I used value_counts for a distribution of different level severity, the number 1 to 4 represent 4 level of severity where 1 indicates the least impact on traffic, and 4 indicates the most serious one.
As we can see from the graph, majority of accidents was in severity 2 which are slightly more serious than a minor accidents.



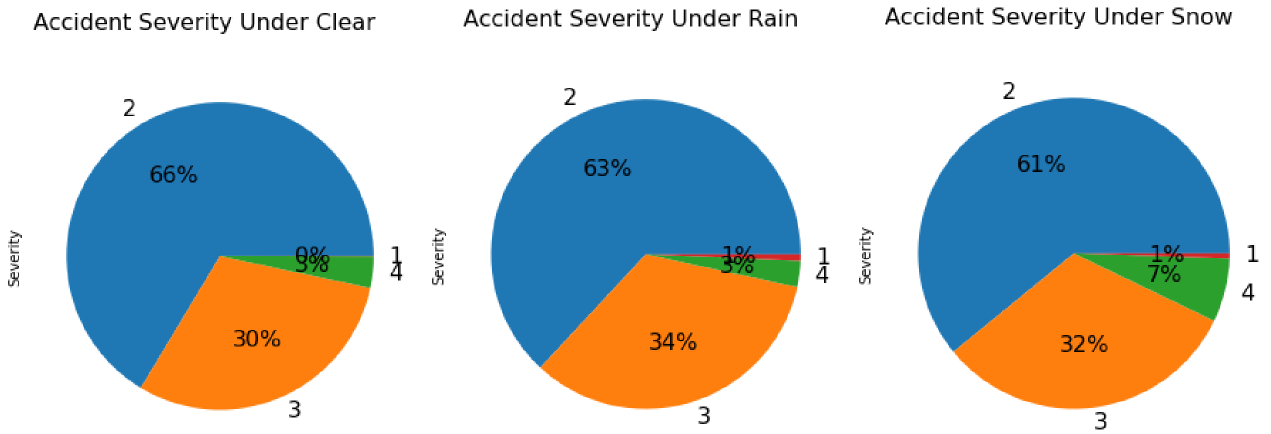
1. Weather

Severe weather conditions may have various impacts on traffic, involving the impacts on vehicle performance, road conditions and driver's conditions. These weather events can affect the transportation system both directly or indirectly, so in this section I mainly forced on the relationship between different weather condition and the severity level.

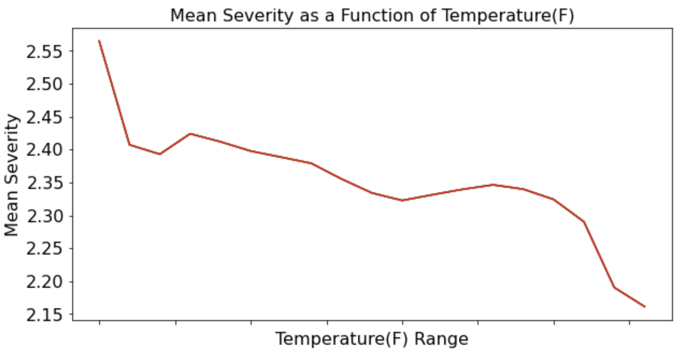


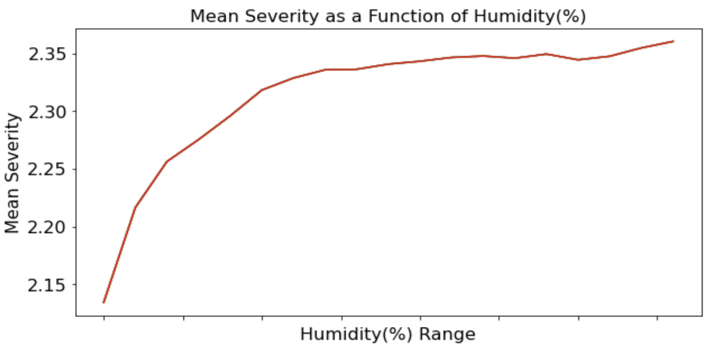
Here are top 10 condition for overall accidents.

- Compare clear, rain, snow relation with severity, it's clearly that condition under snow has the most proportion of level 3 and 4 accidents, possibly because of the slippery road.



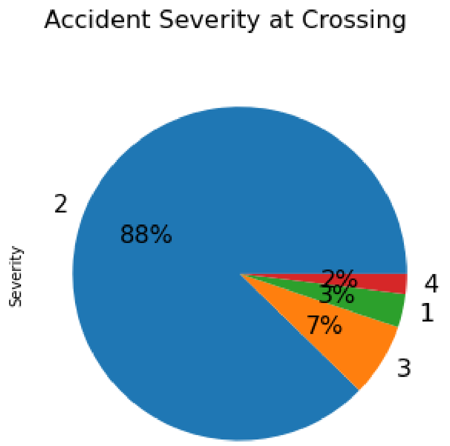
- Compare temperature and humidity, the graph shows the severity will goes up as the temperature decreasing, and the humidity increasing.

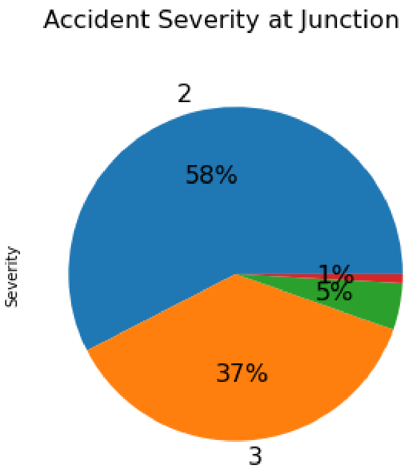


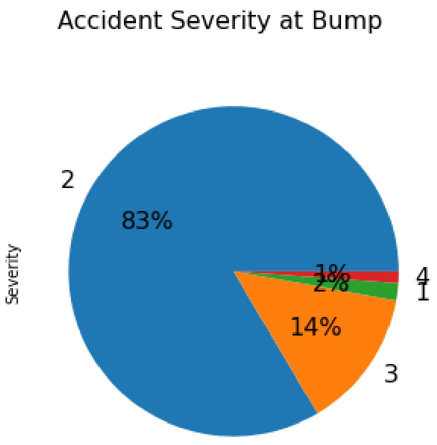


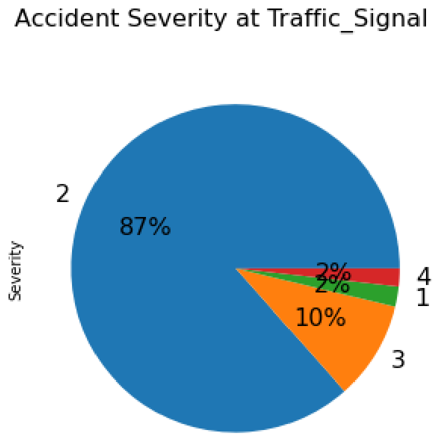
2. Place of accidents

- From those four graph we can tells that at junction it's more like will have a serious accidents.



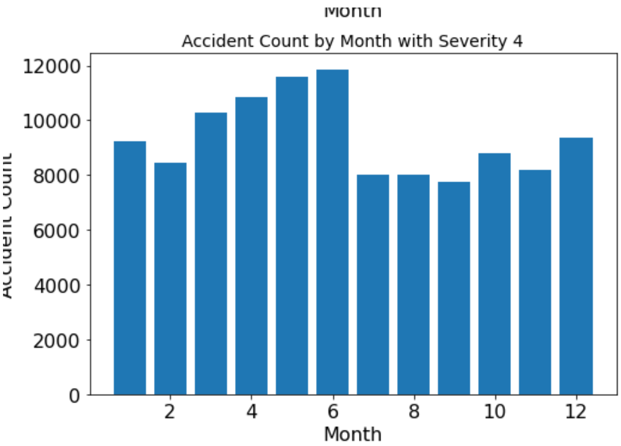






3. Time

Traffic data is required at every stage of project development in order to enable roads to be planned, projected, constructed, maintained and operated more rationally and scientifically. Expected number of vehicles on a planned road is the most important determination of the geometrical features of the road. The results obtained from studies that use this data affect the decisions and policies are formed on issues such as provision of road safety, efficient Murat Karacasu et al. / Procedia Social and Behavioral Sciences 20 (2011) 767–775 769 operation, whether land roads are feasible in economical aspect and in which period it will be feasible and how it will contribute in regional development (KGM, 2009)



From the graph we can tell that severity 4 has increased during march to June, especially on June, and decreases during July to September which is In summer months, schools are closed and people living there go to their cottages in cool locations, therefore, traffic flow decreases as well as the number of serious accidents.

Result / Discussion

This study's topic is the factors of accident severity, and the severity of an accident can be measured in terms of the number of fatal accidents or total economic loss. So there is such a difficulty in terms of clustering and classification, especially predicting. However, I still spat the data into a training set and testing set using the train test split model. First, I used the logistic regression model to perform the classification between minor and major accidents. The result was 77% of train accuracy and 76% of test accuracy, which is pretty decent but not so well. So I used the random forest for another try; this model turns out to have a 100% train accuracy and 97% of test accuracy, much better than the logistic regression model.

Conclusion

As I mentioned before, it is tough to predict the severity of accidents due to many factors. This study focuses on finding the factors which have the most impact on severity to bring benefits to the city and the people for preventing unexpected losses from accidents. The finding verified that city infrastructure might be the most influencing factor of the severity of crashes, so I suggest the city should put more effort into urban traffic planning; try to decrease the number of intersection may be a great solution.