Capstone- Predicting Car Accident Severity

Hetal Patel

Introduction

As the world progresses to a more and more advanced state the number of cars on the road are also increasing. Whether you are looking at first world countries or third world countries the number of vehicles year on year are increasing as households are able to afford more. With this in mind a very serious topic comes to light, according to an article by the WHO "Road traffic injuries are estimated to be the eighth leading cause of death globally for all age groups and the leading cause of death for children and young people 5–29 years of age. More people now die in road traffic crashes than from HIV/AIDS".[1]

The purpose of this project is to bring to light the factors that highly affect the chances of getting into a car accident and to encourage safe driving in the volatile conditions to reduce the risk factors. The data we have will help us better understand the severity of accidents in various conditions as well as the weather conditions that contribute to the accidents.

Data

The original data for this project is uploaded to the following Kaggle, it is provided by the IBM coursera and an untouched version of the data will be uploaded to the GitHub account. In another notebook I have begun preparing the data that I will be using so that we have the most relevant features for the prediction of traffic accident severity. There are 194,673 accident records and 38 variables in our data set which cover the years 2004 to 2020.

Objectives

- Review the data in order to understand patterns and trends
- Model the prediction using different algorithms and compare results.
- Conclude with which factors may have more impact on accidents

Missing Data

There is almost 40% of missing data in some of the features, with this in mind during the variable selection process such attributes have been removed.
Below are the variables we will be focusing on in order to classify the severity of the accidents.

- Weather: Weather conditions at the time of the collision/accident
- ROADCOND: Road conditions at the time of the collision/accident
- LIGHTCOND: Light conditions at the time of the collision/accident
- X: providing the longitude location of the accident
- Y: providing the latitude location of the accident

These variables also contain missing values however considering the total value of the dataset it only amounts to 3%.

Target Variable

The target variable in this data set is the SEVERITYCODE, it is split into two values as shown below.

- 1 – Property Damage
- 2 – Injury Collision

	Count of accidents	Accident Severity
SEVERITYCODE		
1	136485	Property Damage
2	58188	Injury Collision

Methodology

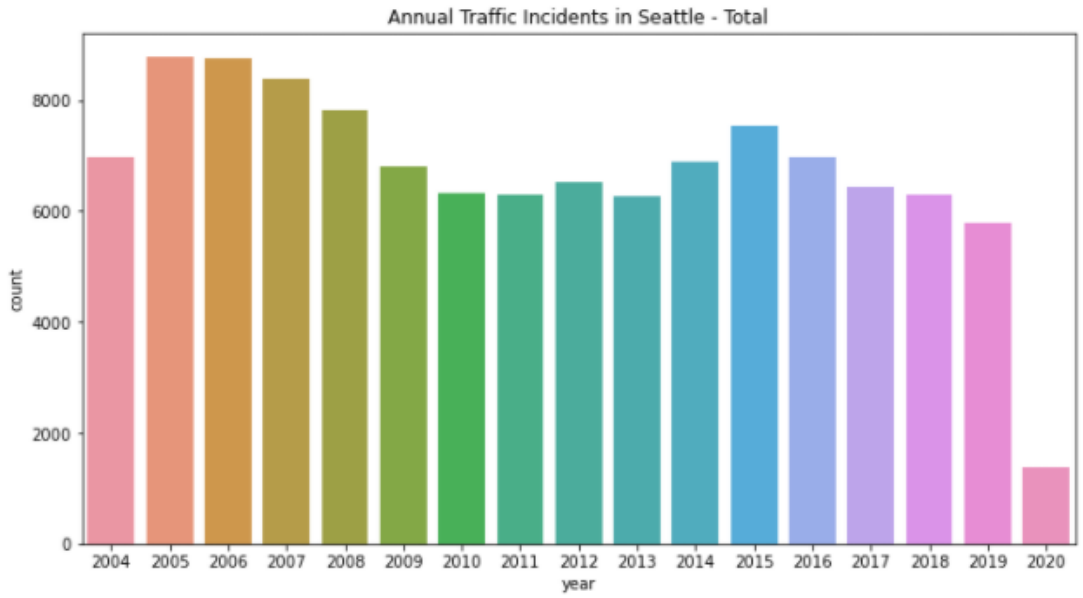
In order to classify our SEVERITYCODE we will use a limited number of features as shown below.

- WEATHER
- ROADCOND
- LIGHTCOND
- location (X)
- location (Y)

As we have already dropped the missing values in these features the next step is split the data for to training and test groups with 80% of samples for training and 20% for testing.
Finally in our analysis will be calculation and exploration of different models to find out the main problem for severity. We will use 4 classification models which are Logistic Regression, Decision Tree, KNN and SVM. After obtaining each model's predictions we will evaluate their accuracy, precision, f1-score, log-loss and compare and discuss the results.

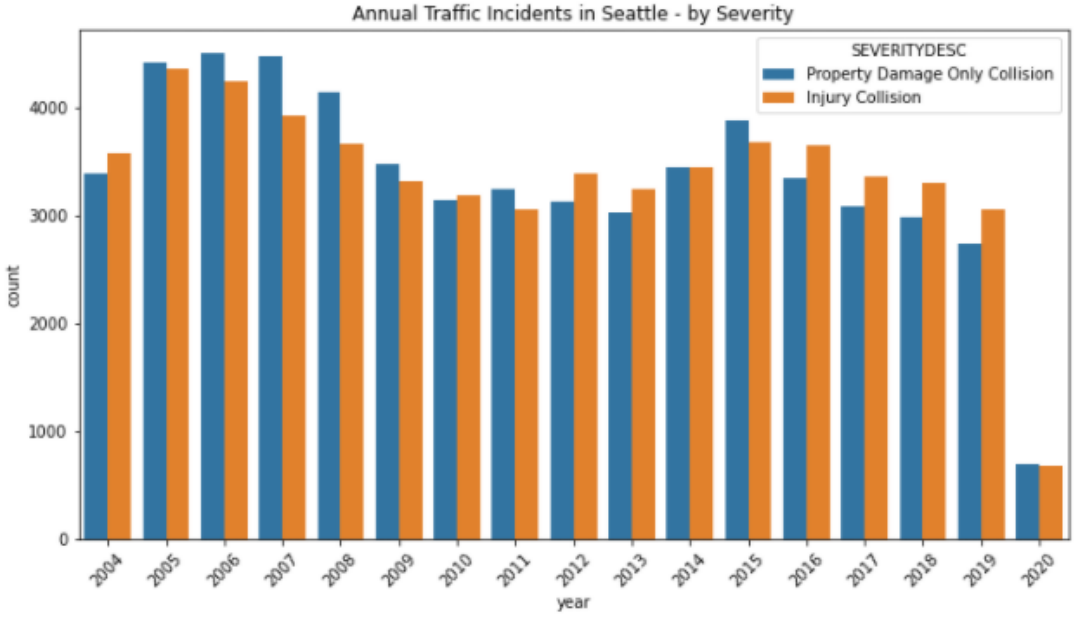
Annual Traffic Incidents

As the data provided spans over 16 years it is possible for us to compare year on year the number of traffic incidents. One thing to keep in mind is that the data for the year 2020 is limited (only up to May) and also we are experiencing a world-wide pandemic where the government and businesses have encouraged people to work from home, this means that the data we see is a direct result of this and it should not be compared to other years as we have not seen such circumstances before.



Another way of looking at the year on year breakdown is by studying the severity of the accidents year on year. One point to remember is that the data at this point has been balanced and so what we see below is a balanced representation of the severities.

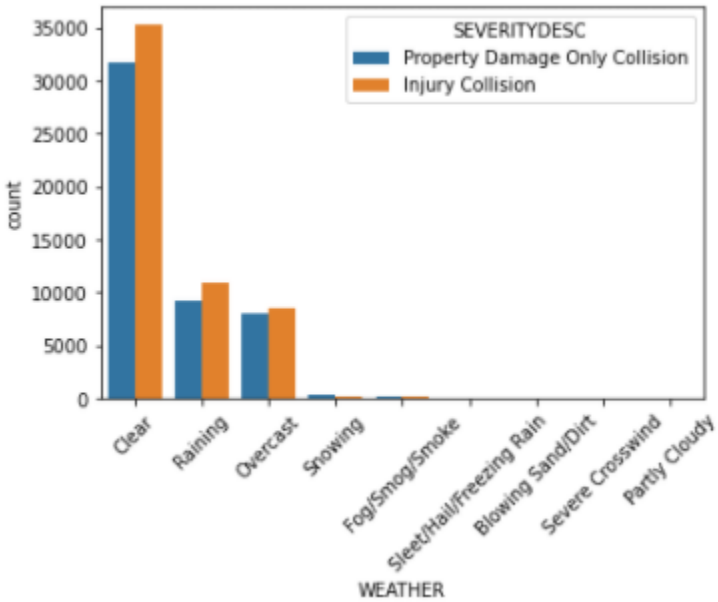
We cannot see a clear difference between the two severities, there are some years where 'Property Damage' is higher and other times where 'Injury by Collisions' is higher.



Weather Conditions

Weather conditions usually play the most pivotal role when looking at the cause of traffic accidents. In our data we are seeing a completely different outlook, by far we can see that the greatest number of accidents have taken place in dry conditions.

Weather Conditions	Count
Blowing Sand/Dirt	32
Clear	67017
Fog/Smog/Smoke	349
Overcast	16645
Partly Cloudy	5
Raining	20084
Severe Crosswind	17
Sleet/Hail/Freezing Rain	56
Snowing	486

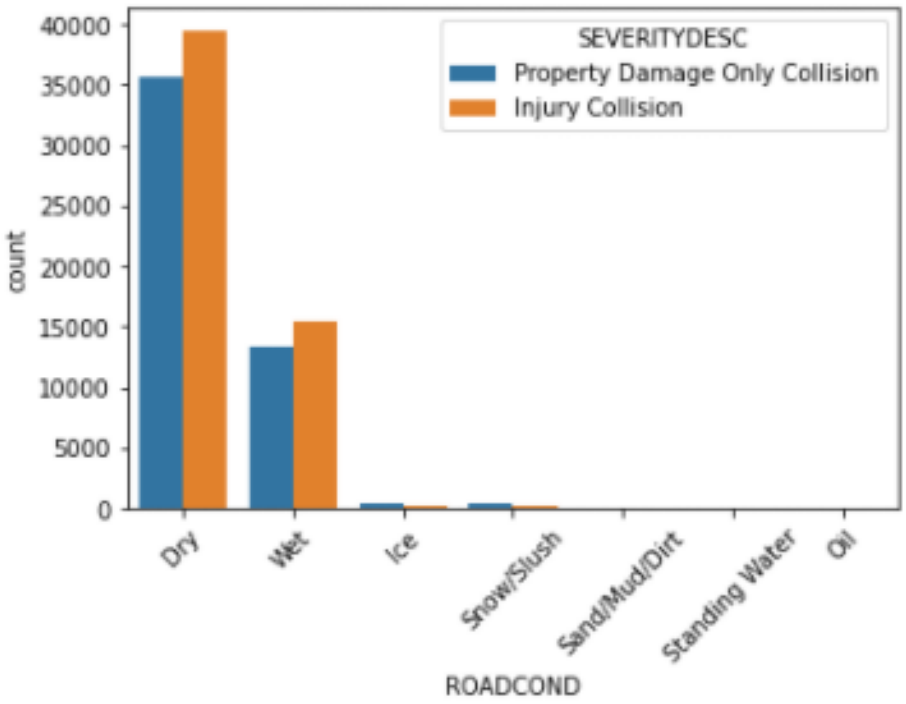


Originally, I had thought of using the 'INATTENTIONID' as one of the comparison factors however due to the amount of data that was missing or not confirmed I decided it would not be fair to this project.

Road Conditions

As most of us who drive cars understand that road conditions are usually something to look out for, our data shows us that the greatest number of accidents to occur have happen on dry conditions.

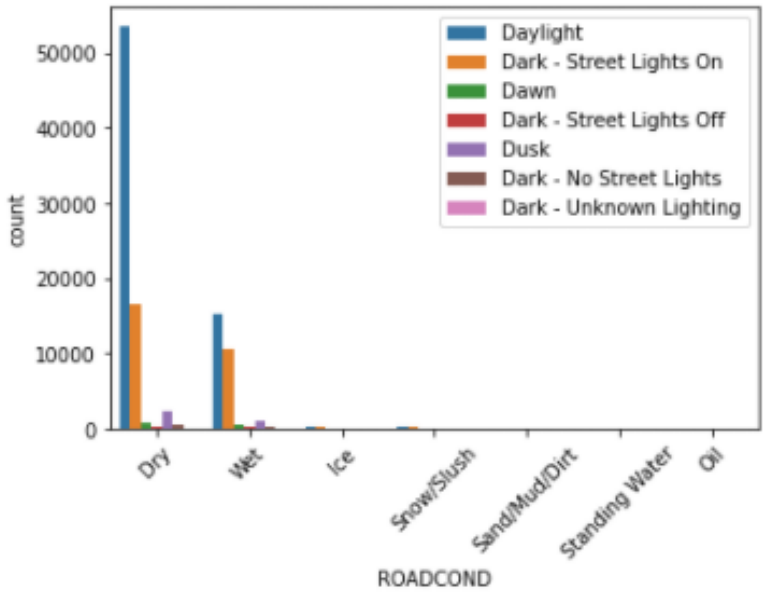
Road Conditions	Count
Dry	75123
Ice	661
Oil	33
Sand/Mud/Dirt	35
Snow/Slush	526
Standing Water	58
Wet	28687



With the help of the above graphs we are able to see the split in the severity description according to the road conditions, above the graph we can see that when there has been dry road conditions the most number of accidents have been recorded. Another factor to take into consideration would be the driving at various light conditions, as it is drivers would be a little less concerned when roads are dry. This leads to the question.

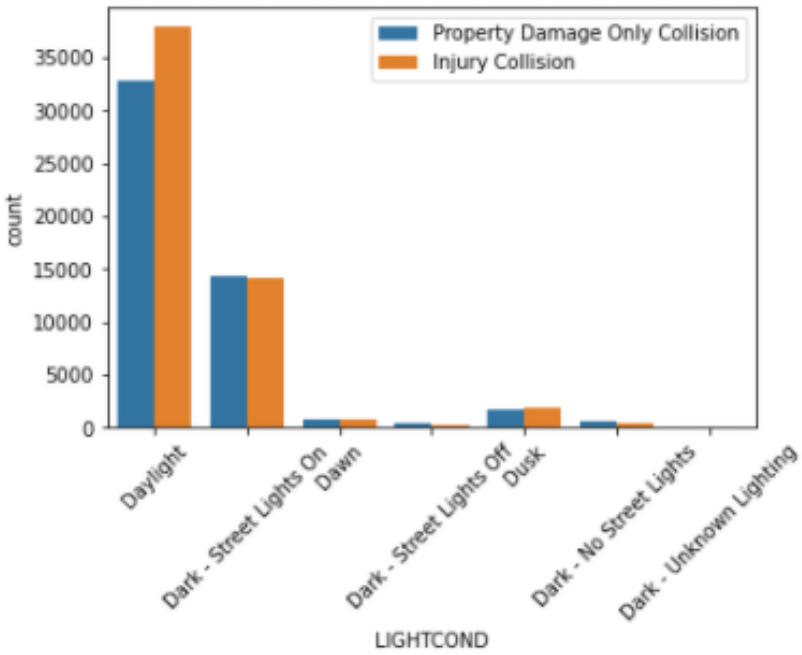
Do the road and light conditions lead to safer driving? Now normally you would assume that “yes if both the road and light conditions are good then there is less chance of an accident happening”

However, in our case it is the opposite, our data shows us that even when it is daylight and road conditions are dry, we have the greatest number of accidents.



Light Conditions

Our last comparison with road condition showed us that we have the largest number of accidents in the daylight.
To better understand this we bring in the SEVERITYCODE, which will help us understand the number of accidents that were responsible for injury as well as the number that were responsible for property damage.



Location

Location is a very important factor, it would normally make you be more wary if you know that accidents have happened in the neighbourhood. With our data as we have the ‘X’ and ‘Y’ coordinates we are able to plot all the data on a map to better understand the areas where the most accidents have happened.

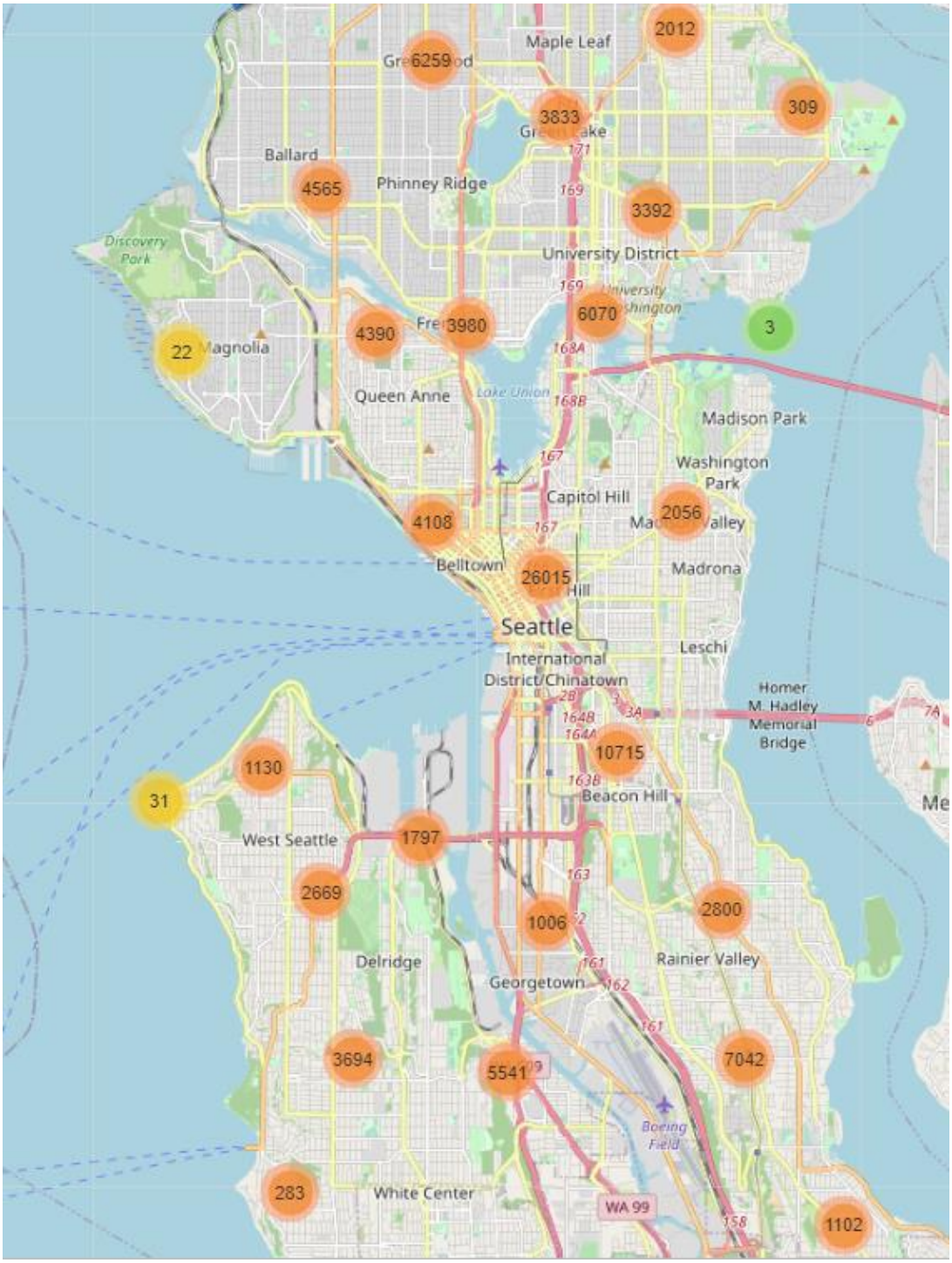


Image: Number of Traffic Accidents in Seattle, grouped by location.

Modelling

In order to build an accurate model, we will use the training set, then make use of the test set to gauge the accuracy of the individual models.

The following algorithms will be used

- K Nearest Neighbours
- Decision Tree
- Support Vector Machine
- Logistic Regression

We will try to ensure best model settings, to select best parameters for the model using Grid search, i.e. testing different parameters, like different number of K neigbors, and selecting one with highest estimate accuracy.

Results and Discussions

In this analysis we evaluated the performance of 4 machine learning algorithms on the Seattle Collision dataset to predict the severity of an accident by understanding the weather, road, light conditions and location. The three models performed very similary, but KNN stood out with a slightly higher F-1 score, but lower Jaccard index (lower accuacy). With KNN we were able to meet 54% accuracy and 50% accuracy with decision tree, SVM and Logistic Regression`.

Conclusions

The purpose of this project was exploring and understand the relationship between traffic accident severity and various characteristics that give insight to the situation. From the available 38 features we have gone on to select just 5 that we feel summarise the relationship as best as possible. An accuracy of 54% has been achieved and it is clear that the features that have been highlighted do have an impact on the severity of an accident. There is scope for improvement provided that additional variables are available, these can be a mix of variables we already have and other external variables. An example of this can be the last time of service, car safety features (for example air bags, anti-lock brakes, electronic stability control, traction control).