**FINAL REPORT**

**SEATTLE ROAD ACCIDENT SEVERITY PREDICTION**

***IBM Data Science Capstone***

**Bhavana Suresh**

September, 2020

1. **INTRODUCTION**

Let's imagine, that my family and I have made plans of going out in the evening and everyone is almost ready. But suddenly I see flash of lightening outside, hear the roaring thunder and the pitter-patter of rain drops. Everyone gets disappointed as the elders start advising that it is not safe to drive in such conditions. We cancel the plan. However, we cannot do the same if it is an urgent meeting you have to attend or catch a flight for an important business opportunity overseas. Because the elders are just making vague predictions based on their experience. It is correct to a certain extent but these appointments cannot be missed based on predictions made with no certain proof.

Thus, this project focuses on building a Machine Learning Model to predict the safety of the travelers (i.e.) the probability of an accident to occur based on various factors including weather, road conditions, traffic, etc. It helps make more accurate predictions based on the data collected from different accident locations in Seattle.

The information gathered by performing Data Analysis on the collected data will help the police forces and emergency medical evacuation team to warned and be prepared for being able to reach the location that are accident-prone areas. If found that the state of the roads are being a major cause for accidents, then the public works department can be notified about the amendments to be made for safe travel.

1. **ABOUT THE DATASET**

The details of various accidents and the severity of the accident occurring at multiple locations across Seattle are collected and recorded to analyze and find the cause of accident so as to warn the concerned to take appropriate measures to reduce the accident rates.

The data for analysis was retrieved from the *Road Accident Severity Data* from the *Seattle State Department of Transport* from [Data-Collisions](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv). It is a csv file that contains 194673 entries and 38 attributes. This file contains data about human factors that play a role in these accidents and the details of damage for each instance along with dates and other ‘Yes’ and ‘No’ questions regarding whether the driver was under influence of alcohol or not, or whether the accident involved speeding or not, or whether the pedestrian was granted the right of way to name some.

This dataset has an attribute SEVERITYCODE that classifies the accident into 5 different categories. A code that corresponds to the severity of the collision:

• 3- Fatality

• 2b- Serious Injury

• 2- Injury

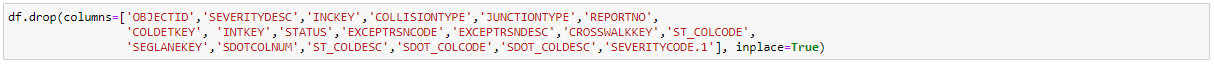
• 1- Property Damage

• 0- Unknown

We will extract data for address type, weather, road and light conditions, and time of day, location, inattentiveness, under influence, and speeding into a new data-frame which will be used for our model. We will clean this data, remove NaN values and categories that have a very small sample size in order to balance the data. Once data extraction and cleaning are completed we will move ahead with visualizations and modelling using KNN, Decision Tree or Logistic Regression.

1. **DATA EXTRACTION**

The required columns out of the 38 attributes are ADDRTYPE, ROADCOND, WEATHER, LIGHTCOND, X, Y, INCDTTM, INCDATE, LOCATION, INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SPEEDING and SEVERITYCODE. These will be the main attributes that we will be using throughout the project. So, we can drop the remaining unnecessary parameters in the dataset.

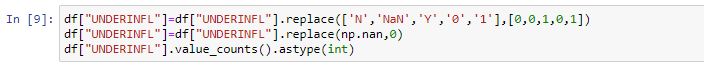


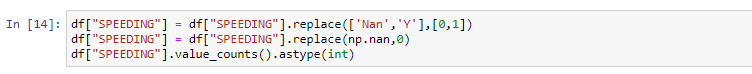
Dropping NaN values from primary attributes before moving to the next step. (We use the dropna function) Next step is data cleaning where modifications for missing values will be done and modifying contents of certain attributes will take place.

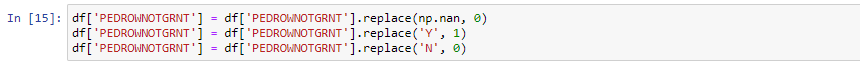


1. **DATA CLEANING**

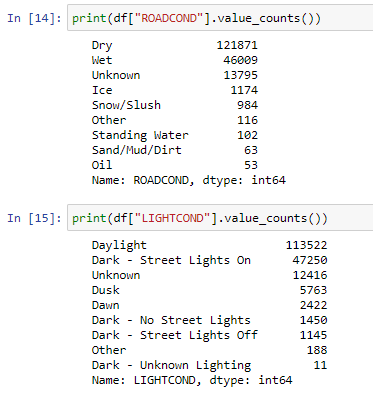
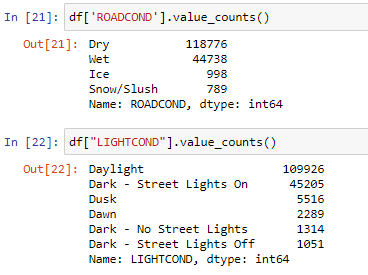
Starting with the extra attributes and homogenizing the data i.e. converting Yes/Y to 1, No/N to 0 and NaN values to 0 as most of the rows are 0 (average of entire column)

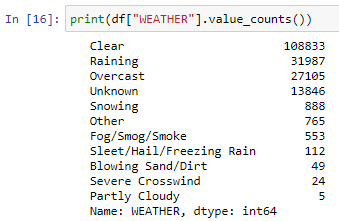
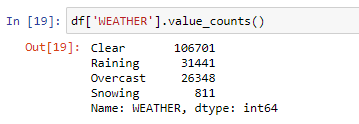






Next we need to clean the important attributes. Start by getting value counts for each attribute. It is observed that there are few categories that have a count very less compared to the categories with maximum instances. Also ‘unknown’ category within these attributes don’t help our model either. So we shall drop the entries that contain these values. (This cleaning is carried out on ROADCOND, WEATHER and LIGHTCOND)

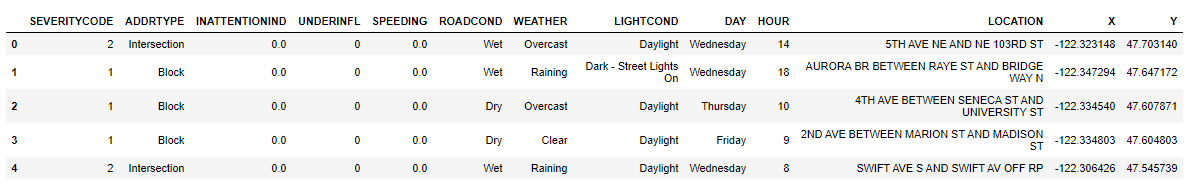
 

*Before cleaning After Cleaning*



Creating new columns named day and hour from INCDATE and INCDTMM respectively using dt.day\_name() and dt.hour functions. These attributes will be used for visualizations and modelling as this data makes more sense than YYYY-MM-DD HH-MM-SS-TTTT format.

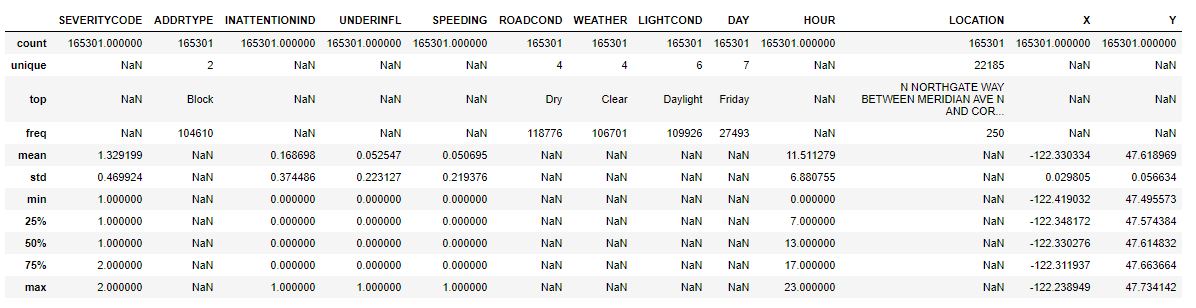
Once all the required data attributes are created and available, next step is to create a separate dataframe df\_new. The new dataframe has the following attributes: SEVERITYCODE, ADDRTYPE, INATTENTIONIND, UNDERINFL, SPEEDING, ROADCOND, WEATHER, LIGHTCOND, DAY, HOUR, LOCATION, X and Y.



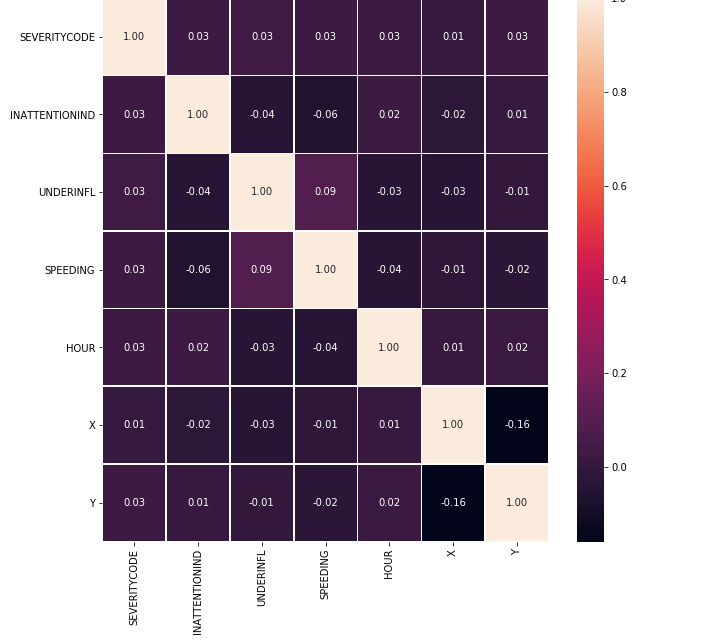
***Figure 1.*** *df\_new dataframe after dropping the unnecessary columns and removing the NaN values from all the primary attributes*

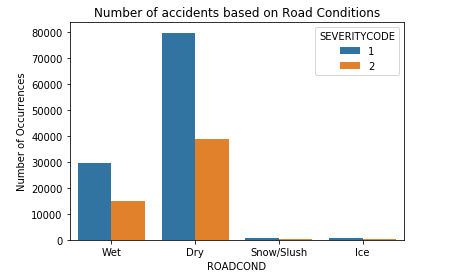
1. **EXPLORATORY DATA ANALYSIS**

Using the describe() method, we obtain the summary of the data. It calculates the basic statistics for all the numerical variables excluding NaN values



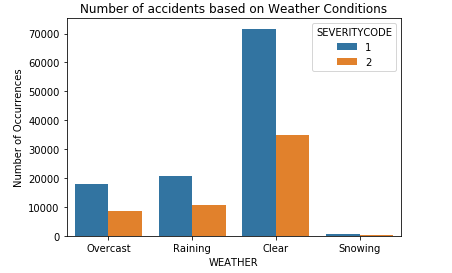
We can create a heat map to find the correlation among the features.





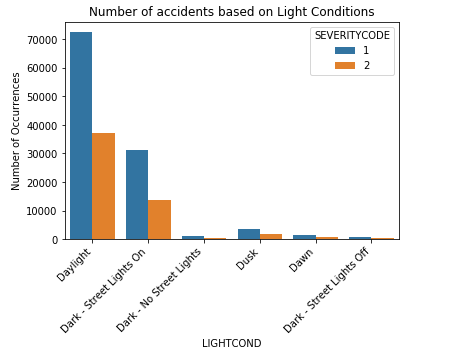
***Figure 3****. Number of accidents for various road conditions*

I used countplot() from the seaborn class to create the visualization of df\_new dataframe which shows the distribution of accidents along with the severity index classifications, based on different road conditions. Dry conditions causes the most number of accidents and the occurrence of injuries is greater than property damage.



***Figure 4.*** *Number of accidents for various weather conditions*

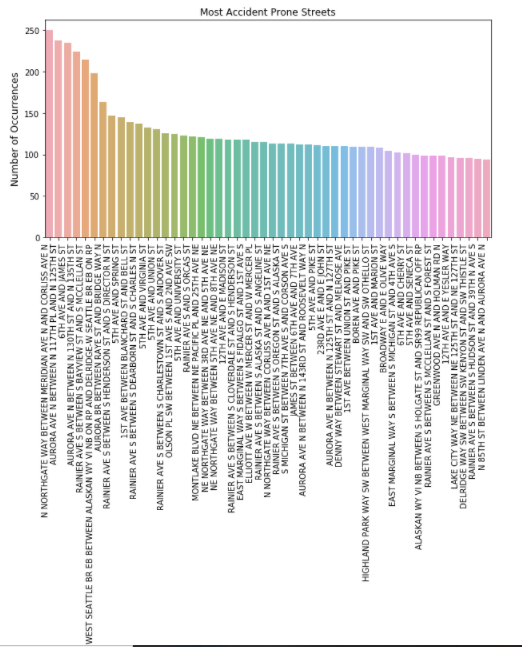
Plot based on different weather conditions. Clear conditions causes the most number of accidents and the occurrence of injuries is greater than property damage.



***Figure 5.*** *Number of accidents for various weather conditions*

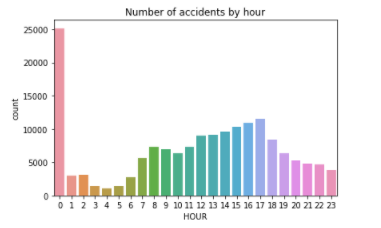
Plot based on different light conditions. Daylight causes the most number of accidents and the occurrence of injuries is greater than property damage.

The barplot() function is used to plot the top 50 most accident prone streets/blocks in Seattle.

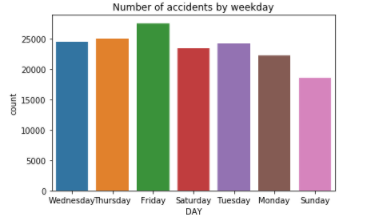


***Figure 8.*** *Number of accidents for top 50 locations*

The next two visualizations for accidents occurring on a particular weekday and at a particular hour of the day, which is carried out using seaborn countplot(). It can be seen that 0000 hours have the most number of accidents. Similarly Friday witnessed the maximum number of incidents. Although the distribution for weekday is not as dramatic as that for accident by hour.

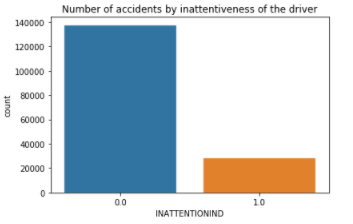


***Figure 7.*** *Number of accidents by hour*

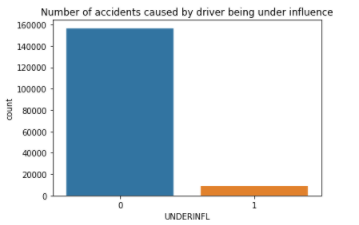


***Figure 6.*** *Number of accidents by weekdays*

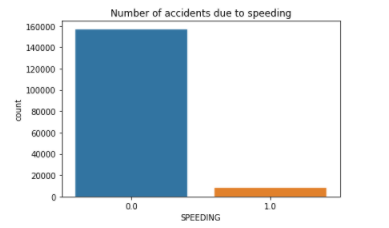
Furthermore, using countplot() I have visualized the number of accidents caused by inattentiveness of the driver, by driver being under influence or by speeding.



***Figure 6.*** *Number of accidents by inattentiveness of the driver*

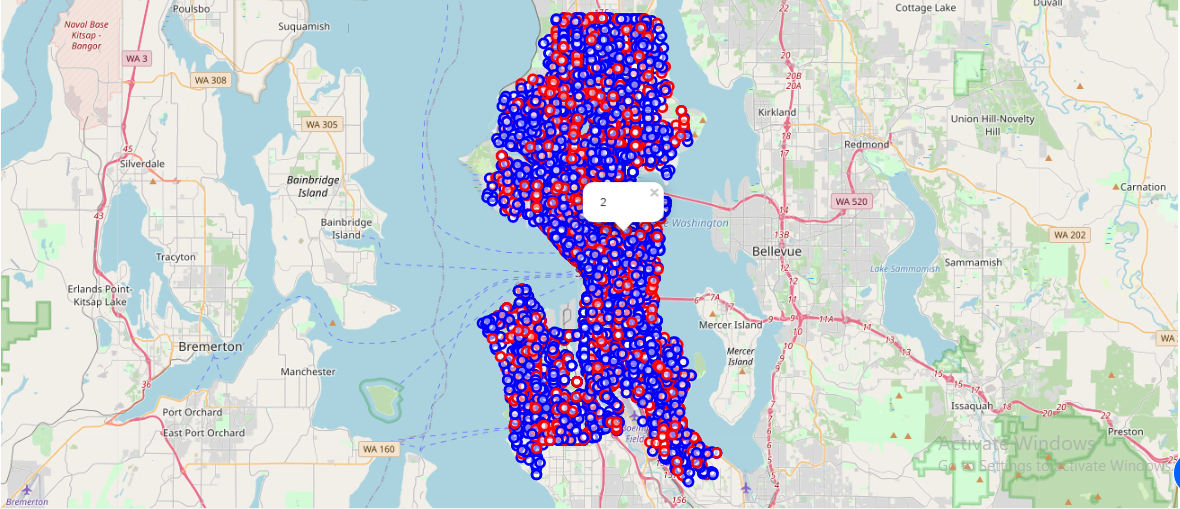


***Figure 6.*** *Number of accidents by driver being under influence*



***Figure 6.*** *Number of accidents by speeding*

Using Folium and Features, a map visualization is created to plot all the accidents in various colors and a pop up marker to indicate the severity of each incident. This plot shows the distribution of accidents and further modifications can be made to the parameters. For example, all the incidents in the past year can be plotted and a grouping can be done for accidents occurring in a particular region.

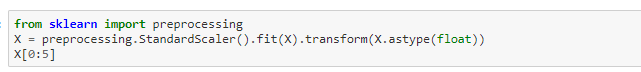


***Figure 9.*** *Plot of all the accidents on the map along with markers for severity index*

1. **PREDICTIVE MODELLING**

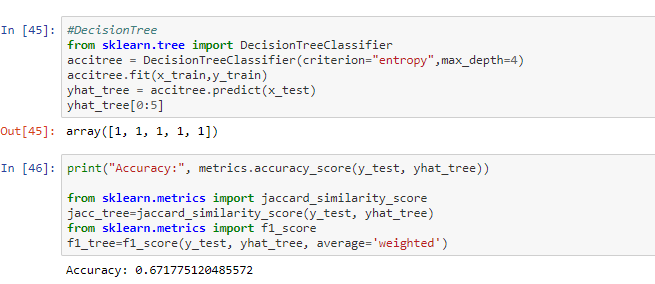
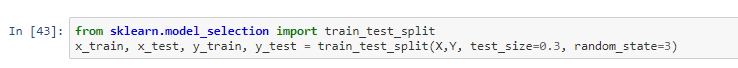
Moving ahead with the modelling, I have used K-Nearest Neighbor, Decision Tree and Logistic Regression as the three machine learning algorithms. For each method, the evaluation metrics used are Jaccard score and F1 score along with log loss for Logistic Regression. All the required libraries and packages are imported and after normalizing the data using StandardScalar(), the dataset is split into test and training set with 70% of the dataset being used to train the model.

Starting with the KNN first, using a loop and calculating accuracy for each k value from 1 to 10, I obtained maximum accuracy for the test set for a k value of 4.

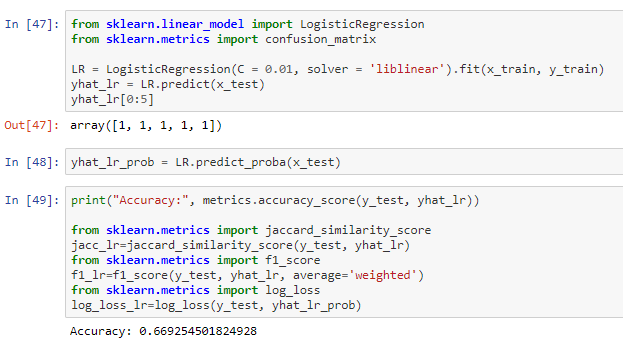




Next model is decision tree, which used a test size of 30% and random\_state of 3 to split the preprocessed data. The classifier is defined with a max depth 4, and achieved an accuracy of 67.18%. There is scope for refinement but I decided to go ahead with 4 as the max depth and fit the classifier using the training set and predicting target attribute using the test set. F1 score and Jaccard score for this model is also evaluated.

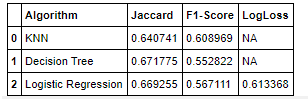


Lastly using Logistic Regression to create a predictor model, I tried implementing various solvers like newton-cg, liblinear, saga but all the solvers had similar accuracy for a C value of 0.01. The probability for the predicted values is also calculated along with the metric evaluation.



1. **RESULTS AND EVALUATION**

It seems like decision tree would be the best machine learning algorithm for this particular model as indicated by the highest Jaccard score of 0.671775. With further changes in max depth we could achieve a little more accuracy. But a score of 0.671775 shows a decent balance between recall and precision. It also indicates that road, weather and light conditions together with location and time of day with minor considerations of inattentiveness of driver and speeding can be used to predict severity of accident with an accuracy of 0.671775. Although, Logistic Regression can also be used as it would work well with binary classification which is the case with our dataset. (As it contains just two categories, 2- injury and 1- property damage).



1. **DISCUSSION**

Based on visualizations it looks like dry, clear skies and daylight conditions have caused the most accidents but these conditions are normal conditions and we could expect more driver negligence or other factors to be the main culprit. Few more lines of code that check the status for all cases where all the three conditions occur together can be carried out. Since it doesn’t help the current model. Analysis to understand the effect that each attribute has separately on the model and how does all of these attributes considered together affect the model can also be carried out.

1. **CONCLUSION**

Based on the dataset provided and the prediction model created using decision tree, we can conclude that particular weather, road and light conditions have some level of impact and responsible driving play a major role in deciding whether the travel could result in property damage or an injury. If implemented properly, it can result in saving many lives and huge sums of money on a yearly basis for the city of Seattle.