**Predicting the Severity of Accidents in Seattle**

Manohar Emani

August 30, 2020

1. **Introduction:**

Road accidents have always been an unsolved puzzle in the history of transportation. Road accidents have bad effects on various areas of life. A road accident can lead to physical damage to the victims and at times, it can lead to loss of lives. In the business perspective, road accidents result in damage of goods and delays in delivery of services. For the government, road accidents lead to huge traffic blocks. Public properties like the roads, pavements, etc., also get damaged due to the road accidents. To avoid more accidents, it will be helpful if the accident severity at various locations of the city of Seattle can be predicted based on the previous data so that the travelers can avoid the locations or at least, take care while they drive through the locations with high severity of accidents.

All the stake holders including the travelers, the government and the people who live near those locations of Seattle city show their interest to know the predicted severity of accidents so that they would cooperate with each other to resolve the issues in locations according to the predicted severity of accidents and make them safer for travel. Government can plan new infrastructure in those locations to avoid accidents.

1. **Data:**

The dataset provided in the Capstone course is used to build a good machine learning model that can predict the accidents’ severity of a location based on various attributes. The data is downloaded from [here](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv). The data has 37 attributes like the location, junction type, weather condition, light condition, accident severity, etc., and 194,674 records of various accidents happened at various locations in Seattle city. The data contains redundant columns, missing data, etc. All these anomalies are fixed and a few columns that are highly related to the accident severity are selected. Then, the processed data is used to build a good model that can predict the SEVERITYCODE (the column that contains accident severity codes based on the type of severity) based on the columns that are selected for the analysis.

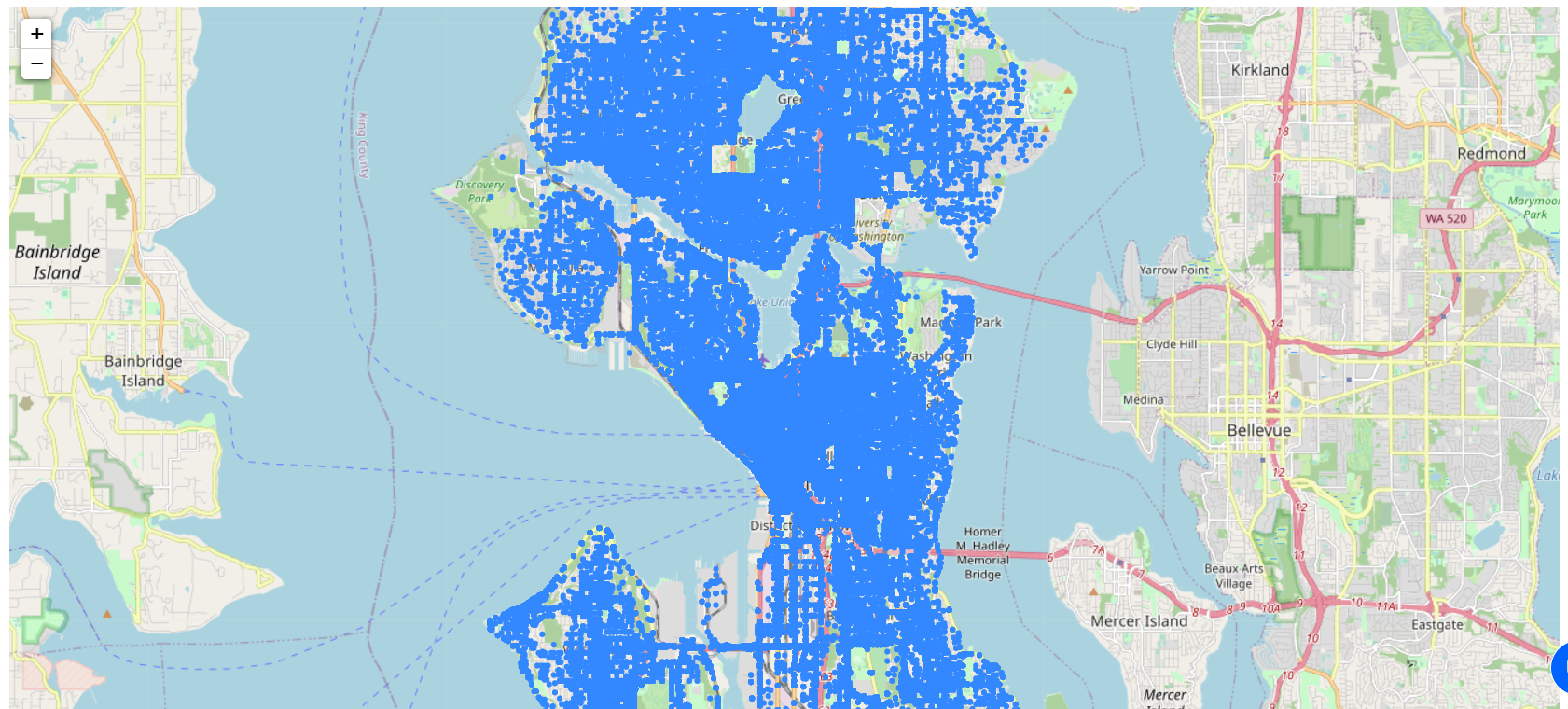
To build a machine learning model, I’ve chosen COLLISIONTYPE, PERSONCOUNT, VEHCOUNT, JUNCTIONTYPE, WEATHER, ROADCOND and LIGHTCOND as these are the features that affect the severity of an accident. The more people involved in the accident, the more severe is the accident. The case is same with the number of vehicles. The directions of the vehicles on the roads depend on the type of junction they pass through. Accident severity also depends on the type of collision. Moreover, bad weather, bad road conditions also lead to severe accidents. These features are used to build a model.

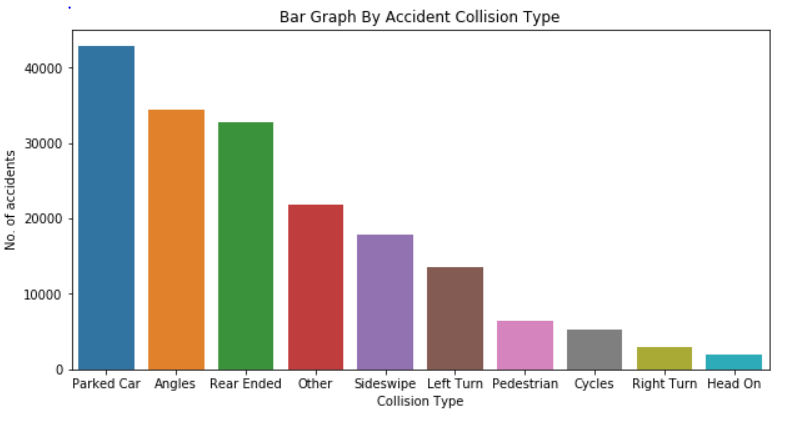
1. **Exploratory Data Analysis:**

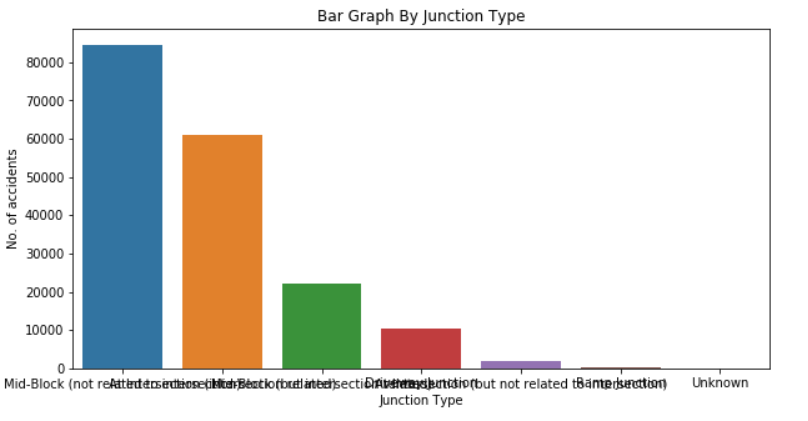
First, I visualized the data on a map using the given coordinates of the locations to observe the areas where the accidents happened. I analyzed the magnitude of accident severity of each feature that I used to build the model. Then, I converted the column values of the selected features into numbers so that it will be easier to apply statistical methods. Now, I need to build a model to predict the SEVERITYCODE. This is a classification problem as I need to predict whether the accident severity is of a particular type. Hence, I have applied various classification algorithms like K Nearest Neighbors, Decision Tree, Support Vector Machine and Logistic Regression. Before building the model, I split the data into training data and test data. Then, I applied the classification algorithms to the training data. After building each model, I fitted the data for training data and testing data. Then, I have calculated the accuracy of the models built using the f1 score, jaccard similarity score and logloss methods for the predictions on training data as well as test data.

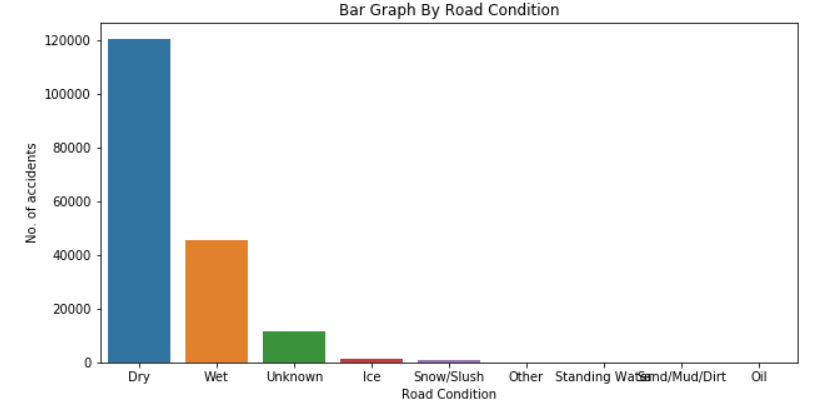
1. **Results:**

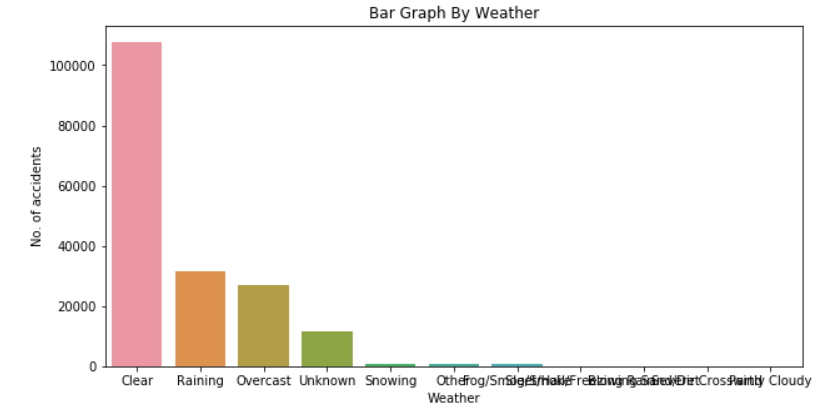
Below is the map of Seattle City. All the locations where these accidents took place are plotted in blue.

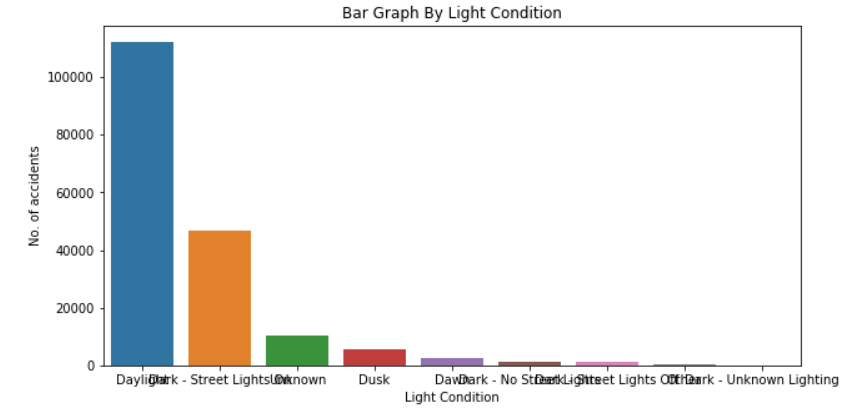












I have plotted the above graphs to observe various distributions of accident data. Accidents happened almost all over the city. Most of these accidents resulted in property damage, while the others lead to injuries. Interestingly, most of the accidents happened in Clear weather, Daylight and on Dry roads. A significant number of accidents happened on wet roads when it was raining, or it was an overcast sky. Most of the accidents happened at road intersections. The number of accidents varied more based on the type of collision. Hitting parked cars, angular hits and rear end hits were the most significant accident cases.

The goal is to predict the level of accident severity of a location. More accurate severity means most true positives. Under this criterion, the best fit model is Support Vector Machine with an average accuracy of ~75%. Here are the details of various models developed and their performance.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | K Nearest Neighbors (k=26) | Decision Tree | Support Vector Machine | Logistic Regression |
| F1 score | 0.7002 | 0.7064 | 0.6968 | 0.6463 |
| Jaccard Similarity Score | 0.7401 | 0.7391 | 0.7440 | 0.7078 |
| No. of True Positives | 23466 | 23039 | 23922 | 23563 |
| No. of True Negatives | 3191 | 3579 | 2876 | 1930 |
| No. of False Positives | 1320 | 1747 | 864 | 1223 |
| No. of False Negatives | 8037 | 7649 | 8352 | 9298 |

1. **Discussion:**

To discuss, there are few observations that I found during the analysis. First, the accident severity has many categories. But this data contains only two types of severity, property damage and injuries. Thus, the model developed on this data may not be helpful in predicting other types of accident severity unless the data related to those types of severity is available. Second, based on the analysis, the probability of accidents is maximum at road intersections. Drivers are needed to be careful at junctions to avoid accidents. One should also be careful with the parked cars and pedestrians while driving one’s car. Maintenance of street lights and the balance between street lights and the lights of the vehicles are also some important factors as accident probability is significant at locations in dark even when the street lights are turned on. There can be many other factors that can affect the accident severity at a place for which sufficient data is unavailable. Hence, there is still a scope for more analysis of the problem.

1. **Conclusion:**

It is to be concluded that the analysis performed for this problem resulted in the development of a machine learning model that can predict the accident severity. If there were sufficient data for other factors such as car speeding, which were dropped due to the unavailability of data, a better model would have been developed. This is not the end of this project as we know that,

Data Mining is a cyclic process and it is endless.

Thank you.