# Predicting the Severity of Car Collision

Tania Habib

September 18, 2020

## 1 Introduction

### 1.1 Background

Road safety is one of the most important aspect of modern infrastructure and urban design. As the governments are planning to move towards smart cities, they want to make decisions in light of facts which are driven directly from the data. Seattle Car Collision dataset is one such example by the local authorities to investigate the underlying conditions or patterns which make increases the chance of an accident. The *Collisions-All Years* is an open database which has been released by the Seattle Police Department that shows the records of all the collisions recorded on the roads since 2004.

### 1.2 Problem

The problem is to predict the severity of the collision which may happen at the intersection or mid-block of a segment. It is a classification problem where a suitable algorithm will be trained on the data set. The prediction will be made through a machine learning model trained on the *Collisions-All Years* dataset which has been recorded by the Traffic Records of the Seattle police department.

## 2 Data Acquisition and Cleaning

### 2.1 Data Source

The dataset used for this problem is provided by the Seattle Police Department traffic records which consists of the severity of collisions which may happen at the intersection or mid-block of a segment. Each case has been assigned a severity label ranging from 0 which means unknown, 1 which means proper damage, 2 which means injury, 2b means serious injury and 3 is for fatality. The data has 37 input attributes ranging from information about a specific accident, location of the accident, collision type, number of vehicles including bicycles, passengers and pedestrians, total injuries, serious injuries and number of fatalities. Furthermore, there is information about the date, time and the location of collision. It has been further supplemented with details about the road, weather and light condition during the collision. The other notable attributes are bout the conditions such as whether the driver involved was under any drug or alcohol influence. Whether speeding and inattention was a factor, whether a pedestrian right of way was granted, and whether the collision involved hitting a parked car. The miscellaneous attributes give details about collision code both at local and provincial level with description, lane segment and crosswalk keys.

## 2.2 Data Cleaning

After the table was converted into a table, various issues were observed in the dataset. There were empty fields for almost all the attributes. Another issue was that all the Severity Codes were not present in the dataset. Only cases with code 1 and code 2 were listed. The following bar chart shows the class distribution.

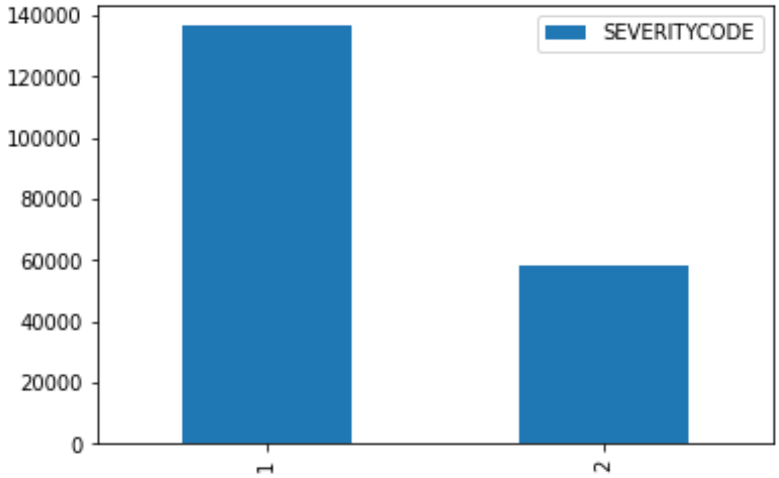

*Figure 1: Unbalanced Dataset.*

Out of a total of 194673 cases, 136485 were class 1, and only 58188 were class 2. Multiple techniques such as under-sampling and over-sampling will be explored to mitigate the effects of bias due to unbalanced dataset.

# 3 Methodology

## 3.1 Exploratory Data Analysis

Out of the 37 attributes listed in the data, not all of them can be deemed useful for predictions. Therefore, an exploratory data analysis exercise was done to see effects of most relevant features in the data set. The following bar char shows the weather conditions during the collision.

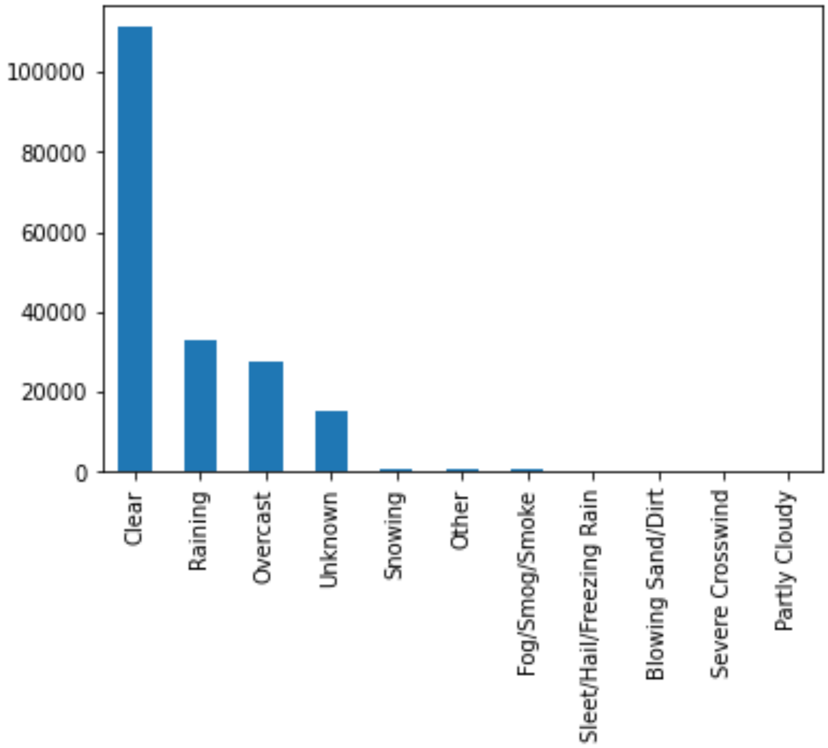

*Figure 2: Weather conditions during collision*

It is interesting to observe that majority of accidents happen in clear weather condition. The Road and light conditions were also investigated. The bar charts below show both the road and light conditions during collision.

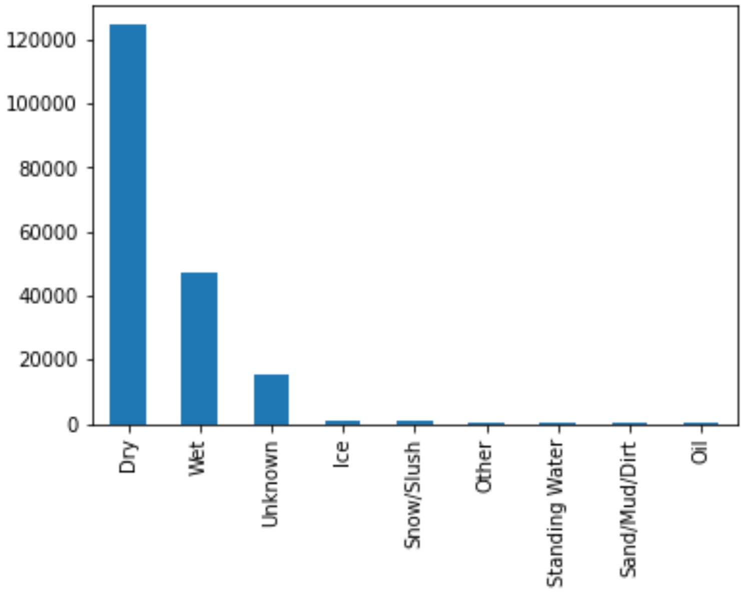

*Figure 3: Road conditions during collision*

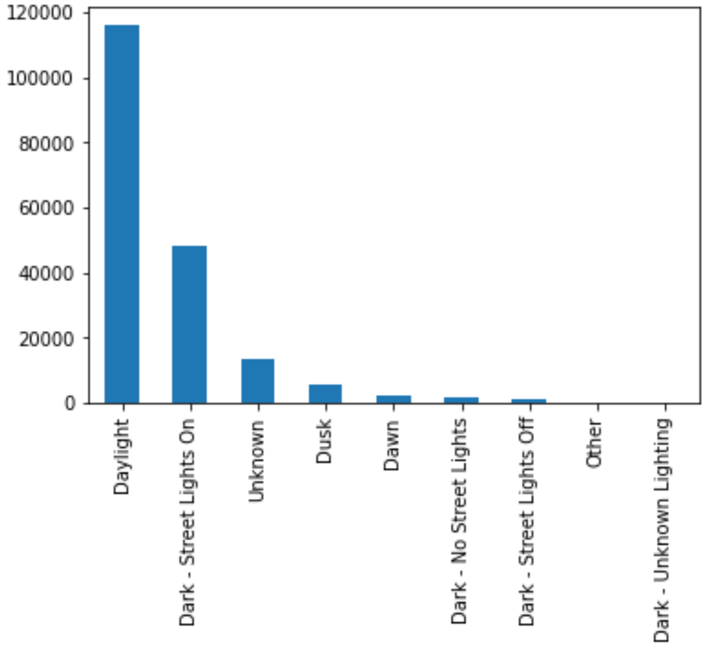

*Figure 4: Light conditions during collision.*

In both cases, road and light conditions were not hazardous but collisions still take place. Other than these cases, the driving under influence and speeding was also analysed, there were around 185340 entries which had no value for speeding attribute.

## 3.2 Feature Engineering

After careful analysis of the dataset and meta information, there were various attributes which did not show direct influence on the predicting the severity of the collision such as incident keys, intersection keys, object ID, location to name the few.

The next process was to find the missing values and see the frequency for each attribute, this highlighted few attributes which seem important for the problem, however there is very little information available in the dataset, for instance, speeding is a significant parameter to determine the severity of the collision but it was only available for 9333 cases in total. Another attribute is the pedestrian right of way granted which is available for only 4667 times in the data set. Another significant attribute is driver in-attention; however, it is available for only 29805 number of times.

Due to the above mentioned reasons, a subset of features was selected to train the model. The features included were as follows:

*Table 1: Description of the selected feature set*

| Feature Name | Type | Description |
|---|---|---|
| ADDRTYPE | Text | Collision address type: Alley, Block, Intersection |
| COLLISIONTYPE | Text | Collision Type |
| HITPARKEDCAR | Text | Whether or not the collision involved hitting a parked car. (Y/N) |
| PEDCYLCOUNT | Double | number of bicycles involved in the collision |
| VEHCOUNT | Double | number of vehicles involved in the collision |
| PERSONCOUNT | Double | total number of people involved in the collision |
| INATTENTIONIND | Text | Whether or not collision was due to inattention |
| UNDERINFL | Text | Whether or not a driver involved was under the influence of drugs or alcohol. |

## 3.3 Data Preparation

In order to prepare the data for machine learning algorithms, the rows for missing values were removed. Another issue relating to model bias which will happen due to un-balanced dataset, under-sampling technique was used. After that, the categorical values of attributes were replaced by dummy or indicator variables. The dataset was further split into 80-20 ratio so that the models trained on the 80% training data can be tested on the 20% testing set.

## 3.4 Model Selection

After the initial data analysis and data cleaning and preparation, various traditional machine learning methods were explored for the given problem. As there are only two labels in the data, Logistic Regression was used to train the model with a regularization to avoid over-fitting. SVM classifier with Radial Basis Function as a kernel was also used to train the model. As there are multiple categorical entries in the data set, decision tree was utilized to train the data. Lastly, K-Nearest Neighbour with N=2 was also explored. Precision, recall, F1-score, accuracy metrics were used to compare the results of all four algorithms. The results are shown in the table below:

*Table 2: Evaluation of Traditional Machine Learning Algorithms*

| | Precision | Recall | Accuracy | F1-score |
|---|---|---|---|---|
| Logistic Regression | 60% | 60% | 60% | 60% |
| SVM | 67% | 66% | 66% | 66% |
| Decision Tree | **68%** | **67%** | **67%** | **67%** |
| KNN | 62% | 62% | 62% | 62% |

Out of all the algorithms, Decision Tree performed the best, the second best was SVM followed by KNN. Logistic Regression was the least accurate method in this case, the regularization factor helps in over fitting but after trying a number of values, the best one was selected.

## 4 Results and Discussion

As this a traditional classification problem, hence evaluation metrics such as precision, recall, accuracy and F1-score were used to assess each classification method performance on the dataset. Even though many issues of the dataset were addressed and resolved and a handful of features were selected after careful data analysis. The performance of predicting the severity of collision is not quite high. Some attributes such as speeding, driver under influence and in-attention are significant, however a large of number of values were missing in the data set. Furthermore, initial findings related to weather, road and light effect on severity of collision was not significant.

## 5 Conclusion

The severity level of a car collision can be estimated with a good accuracy using the key features such as type of collision, the place of collision, driver’s attributes and other supplementary information for instance the number of vehicles, persons, pedestrians and bikers were present in each scenario.