**Predicting Accident Severity in the City of Seattle**

**Jitendra Kumar Tripathi**

Oct. 2020

**1. Introduction/Business Problem**

**1.1 Background**

With continuous increase in daily commutes and transportation, safety is the most important concern for people and the governments, and they are willing to spend lots of money on education, roads and vehicles safety. Unfortunately, accidents happen and an important factor in an accident is the severity of that.

It is very helpful if such severity can be predicted beforehand, based on some factors such as road, weather conditions and accident situation. This can have big advantages for public safety: The emergency system teams would know in advance and can be prepared for more sever accidents to have a prompt and proper response, and also the commuters can take into account such predictions and drive more cautiously or even reschedule their travel.

This can be done with the help of Data Science, which has always playing a crucial role in public health and safety in different areas.

**1.2 Problem**

In this project I will focus on how to predict accident severity based on different variables such as accident location, time, road and weather conditions and etc.

This will be a binary classification where outcomes are accident severity: 1-Property damage and 2-Injury.

**1.3 Interest**

Since this is a public safety issue everyone can be interested in this project, specifically municipalities and governments who are key responsible authorities for driving and road safety.

**2. Data acquisition and cleaning**

**2.1 Data sources**

The data has been provided in the course project. I also reviewed the same data set on Kaggle and found out they are very similar. Data includes accidents between 2004-2020 in the city of Seattle. Dataset provides several features, however not all of them are relevant to our work. Important ones are accident time, location, road, weather and light conditions, accident type and number of parties involved.

**2.2 Data Cleaning**

The data has 37 features. In the first glance it is clear that some of the features are not model predictors as they are duplicates or just some IDs. So, in the first step the following columns will be dropped from the data set.

['OBJECTID','INCKEY','COLDETKEY','REPORTNO','INTKEY','EXCEPTRSNCODE','EXCEPTRSNDESC','STATUS',

'SEVERITYCODE.1','SEVERITYDESC','SDOT\_COLDESC','SDOTCOLNUM','ST\_COLDESC','ST\_COLCODE']

Also the column 'LOCATION' can be dropped since X,Y coordinates are given, and that would be a duplicate.

The next step is to check the null values on the remaining columns, the following table shows only the columns that have null values and the corresponding action that was performed.

|  |  |  |
| --- | --- | --- |
| **Column** | **Number of missing entries** | **Action** |
| X | 5334 | Replace with average |
| Y | 5334 | Replace with average |
| ADDRTYPE | 1926 | Replace with frequency |
| COLLISIONTYPE | 4904 | Replace with frequency |
| JUNCTIONTYPE | 6329 | Replace with frequency |
| INATTENTIONIND | 164868 | Drop the column since most of the entries are missing |
| UNDERINFL | 4884 | Replace with frequency |
| WEATHER | 5081 | Replace with frequency |
| ROADCOND | 5012 | Replace with frequency |
| LIGHTCOND | 5170 | Replace with frequency |
| PEDROWNOTGRNT | 190006 | Drop the column since most of the entries are missing |
| SPEEDING | 185340 | Drop the column since most of the entries are missing |

**2.3 Feature selection**

One important feature in the accident severity can be the time of the accident, so from the two columns INCDATE and INCDTTM, I extracted ‘Hour’, ‘Day of week’, ‘month’ and ‘year’ of the accident and added these new 4 features to the dataset.

So, the final data used for modelling will have the following features:

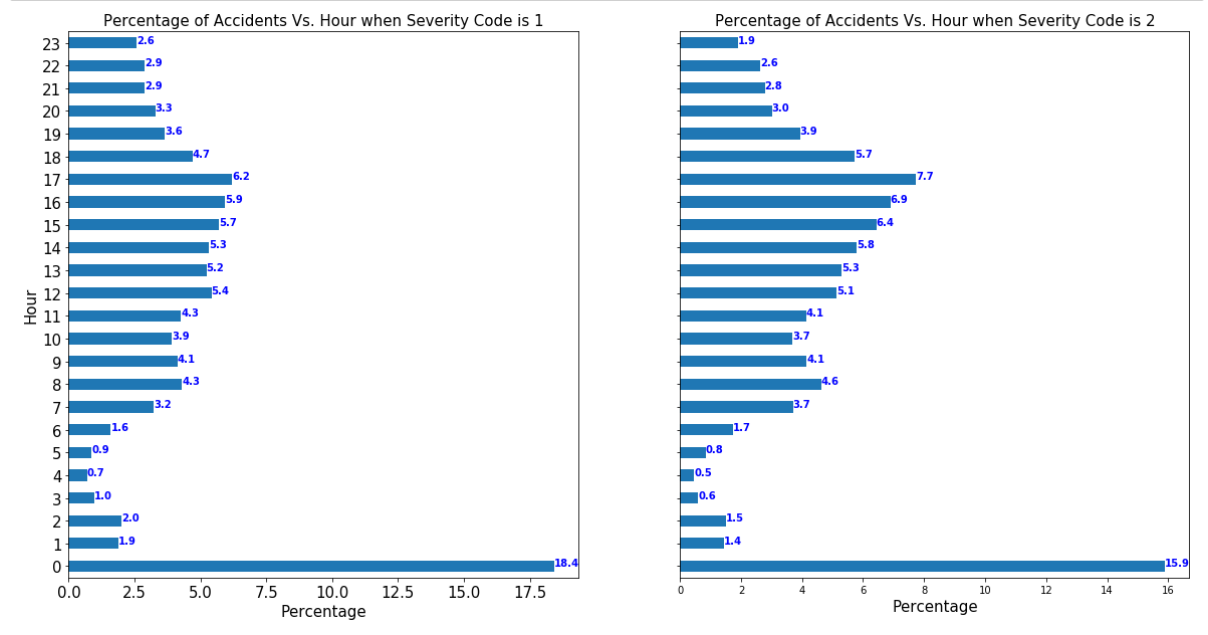
[ 'X', 'Y', 'ADDRTYPE', 'COLLISIONTYPE', 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'JUNCTIONTYPE', 'SDOT\_COLCODE', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR', **'year', 'month', 'dayofweek', 'Hour'**]

**2.4 Exploratory Data Analysis**

In this section I try to investigate the relationship between some important features and the target severity code.

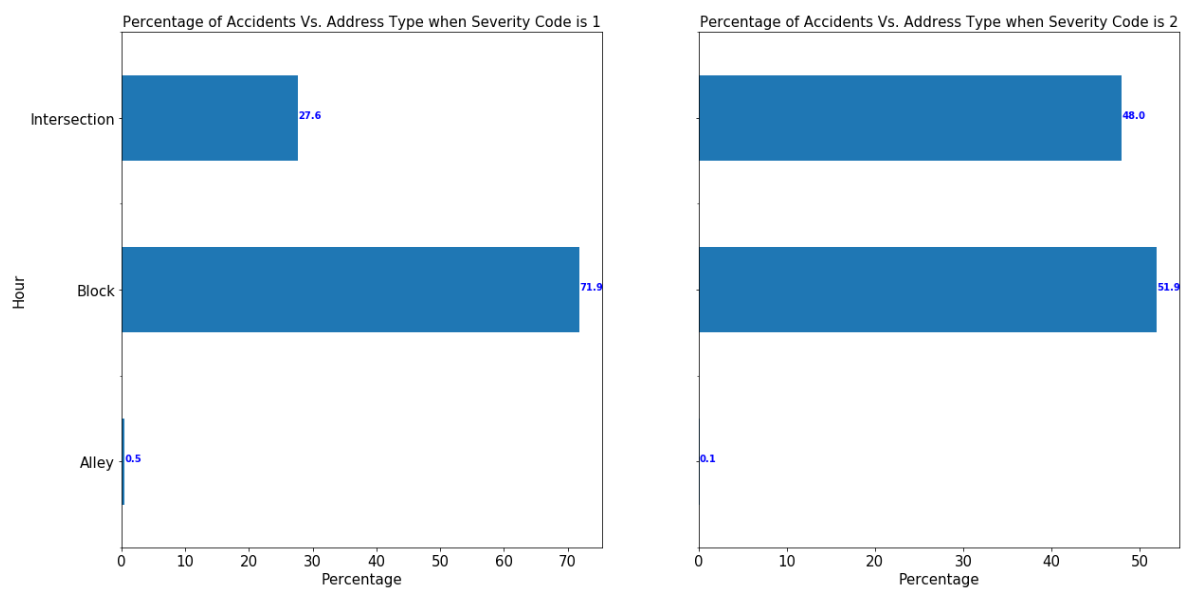
**2.4.1 Relationship between Hour of the accident and severity**

As mentioned earlier, the time of the accident is important in the accident severity, the following graph shows that more sever accidents (i.e. severity code=2) happen during the morning and afternoon rush hours. People should be more patient during rush hour times.



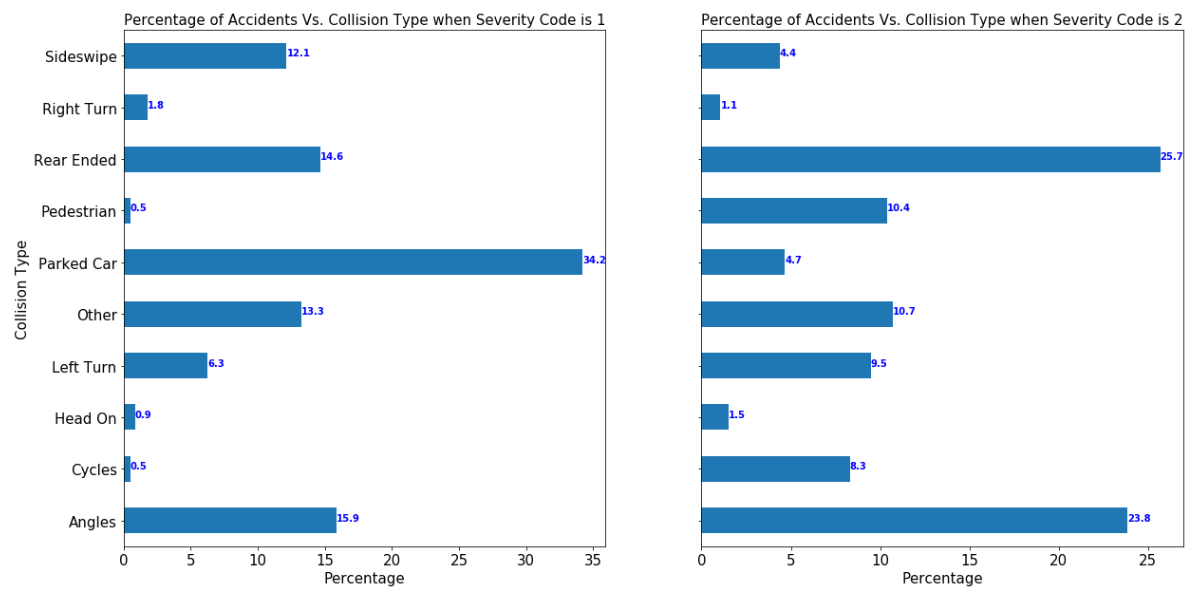
**2.4.2 Relationship between Address Type and accident severity**

According to the following graph the address type is also an important feature in predicting the accident severity: more sever accidents (i.e. severity code=2) happen at intersection, while no injury ones (i.e. severity code=1) mainly happen at blocks. Drivers, pedestrians and cyclists should be more careful when passing the intersections.



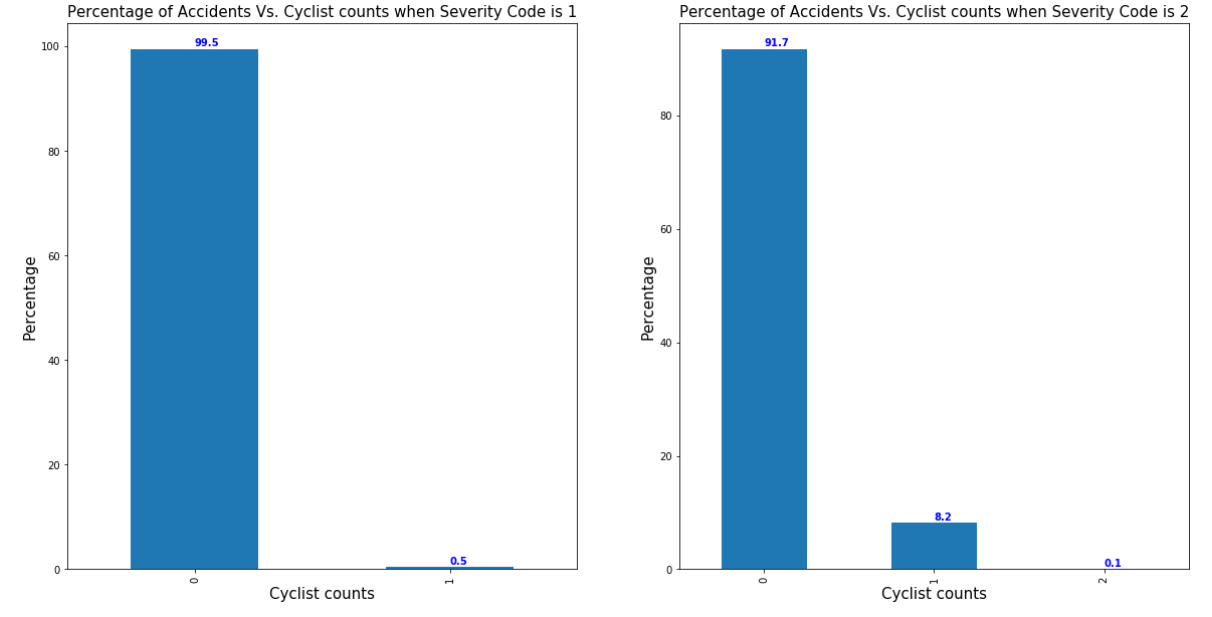
**2.4.3 Relationship between Collision Type and accident severity**

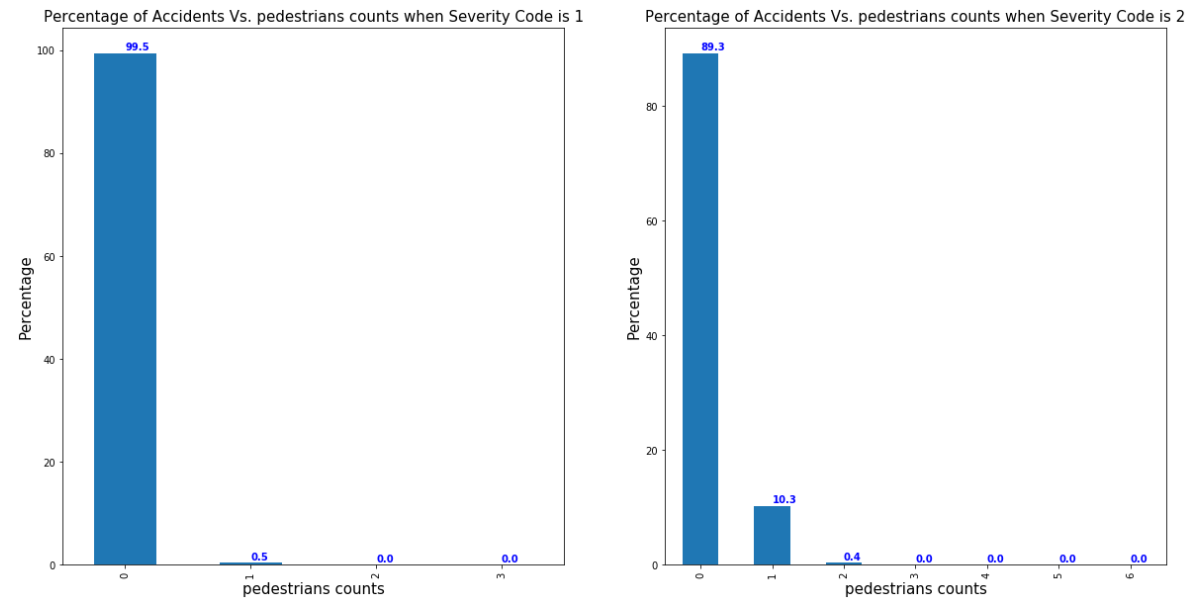
The following graphs show how the Collision Type can help predict the accident severity: for instance, injury accidents happen more if the accident is at Angles or Rear ended, while accidents involving Parked car are mainly cause property damage.



**2.4.4 Relationship between number of cyclists/pedestrians and accident severity**

It is clear when cyclists or pedestrians involved in an accident, injury can happen more frequently.





**3. Methodology**

With the cleaned data explained in section 2, in this part I conduct binary classification modeling. Four different models have been deployed and the results are compared, including KNN, Logistic Regression, Support Vector Machine (SVM) and Decision Tree.

The dataset was split to 80% train and 20% test data. Scaling was also performed on the features to make sure model coefficients are comparable.

**4. Results of Predictive Modelling**

The following Table summarizes the accuracy metrics of the four models used in this study:

|  |  |  |
| --- | --- | --- |
| Model used | F1 Score | Jaccard Similarity Score |
| Logistic Regression | 0.708 | 0.752 |
| KNN (k=6) | 0.715 | 0.743 |
| Decision Tree | 0.709 | 0.755 |
| SVM | 0.716 | 0.760 |

All the results are comparable, SVM slightly shows better performance however it is a more computationally expensive algorithm to implement, if this is not a concern then SVM would be the best choice

**5. Discussion**

In the original dataset some important features such as PEDROWNOTGRNT, SPEEDING and INATTENTIONIND have many missing values that I had no choice but to drop the feature. These features can be helpful in having a more accurate predictive model.

**6. Conclusion**

In this study, I tried to predict the accident severity based on several features. Binary classification methods have been deployed and the prediction metrics have been confirmed, although all of the predictive models showed similar behavior, SVM results are being slightly superior.