Predicting the severity of Road accident based on Traffic Incident data

Parag Sharma

Date- October 11, 2020

1. Introduction

1.1 Background

According to World Health Organization, every year the lives of approximately 1.35 million people are cut short as a result of a road traffic crash. More than half of all road traffic deaths are among vulnerable road users: pedestrians, cyclists, and motorcyclists. Seattle is a port city in West Coast of United States, The city has found itself "bursting at the seams", with over 45,000 households spending more than half their income on housing and at least 2,800 people homeless, and with the country's **sixth-worst rush hour traffic**. Seattle has the 8th worst traffic congestion of all American cities, and is **10th among all North American cities** according to Inrix. Seattle is also referred to informally as the "Gateway to Alaska" for being the nearest major city in the contiguous U.S. to Alaska, "Rain City" **for its frequent cloudy and rainy weather, and "Jet City". Seattle recorded the highest number of car accidents in the state that year (2018), at 14,508 in Washington.**

1.2 Problem Definition

We are trying to create a model to predict the accident severity by road and weather conditions so that the Drivers may act accordingly to the warnings provided. They may change routes, Drive to alternate direction or drive carefully. This will help in better traffic regulation and also provide warnings in case of Severe accident (Involving fatality is predicted).

Our goal is to create a machine learning model so that we can predict that under certain given weather or road conditions how severe is the accident which has the possibility of occurring. To do this we are using the data provide by the Seattle website.

2. Data acquisition and cleaning

2.1 Data Source

The source of the data is http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0/data

This data set is hosted by City of Seattle at an open data platform (Also available at other open data sources). The meta-data can be obtained form https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf.

2.2 A description of the data and how it will be used to solve the problem.

The data is downloaded from the website is form 2018, and it contains 221738 rows and 40 columns. The target Variable is the accident severity which contains four severity levels namely 1, 2, 2a, 3 on an increasing level of severity.

All the feature Names are-

'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS', 'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'SEVERITYCODE', 'SEVERITYDESC', 'COLLISIONTYPE', 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INJURIES', 'SERIOUSINJURIES', 'FATALITIES', 'INCDATE', 'INCDTTM', 'JUNCTIONTYPE', 'SDOT_COLCODE', 'SDOT_COLDESC', 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'PEDROWNOTGRNT', 'SDOTCOLNUM', 'SPEEDING', 'ST_COLCODE', 'ST_COLDESC', 'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR'.

Classification of the features

Locational Data	'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS', 'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC'
Severity Description	'SEVERITYCODE', 'SEVERITYDESC', 'COLLISIONTYPE'
Count of Entities Involved	'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT'
Injury Levels	'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT'
Date and Time data	'INCDATE', 'INCDTTM'
State Designated Code and There Description	'SDOT_COLCODE', 'SDOT_COLDESC'
Weather Road and Driver Conditions	'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'PEDROWNOTGRNT', 'SDOTCOLNUM', 'SPEEDING', 'HITPARKEDCAR'.
State Designated Codes for Crosswalks etc.	, 'ST_COLCODE', 'ST_COLDESC', 'SEGLANEKEY', 'CROSSWALKKEY',

2.3 Data Cleaning

The data we retrieved from the source was highly imbalance and the category of SEVERITYCODE '1' was in abundance. Also the categories '2a' and '3' representing 'serious injury' and 'fatality' were too less in count. So it was decided that the data will be clubbed in two categories namely '1' and '2' representing prop damage only and Human damage also respectively.

After categorizing ‘2a’ and ‘3’ the data needed to be balanced in order to gained an unbiased machine learning model. So we took the data randomly from category ‘1’ equal to the count of data in category ‘2’. The final data had both Categories with 50% count. We also dropped the Unknown severity code values since they were not helpful to us in prediction.

The features Fatality, Serious Injury, Severity Description were dropped because they were not to be used in our final model and they were also redundant since the SEVERITY CODE feature already contains all that information. Rest of the Variables are more of Post Incident features and can-not be used in prediction modelling.

Respectively we choose the following Features for our Feature set-

- Weather
- Road Conditions
- Light Conditions
- Junction Type
- Address Type
- Person Count
- Pedestrian Count
- Vehicle Count

We are Choosing Weather, Road Condition, Light Condition, Junction Type and address type as our variables

So we visualize and clean these features.

We are choosing these Features based on the logic that **only these are the features that will be known to us when our model is ultimately used to predict the Severity Code Based on data**.

We can create a model based on the weather and road conditions and used the Address type and Junction type to tell the driver that these conditions might originate based on where he is.

We are assuming that based on the traffic cams and general traffic pattern we can predict that what would be the **Count of people and Vehicles involved** so we are adding is to our feature set.

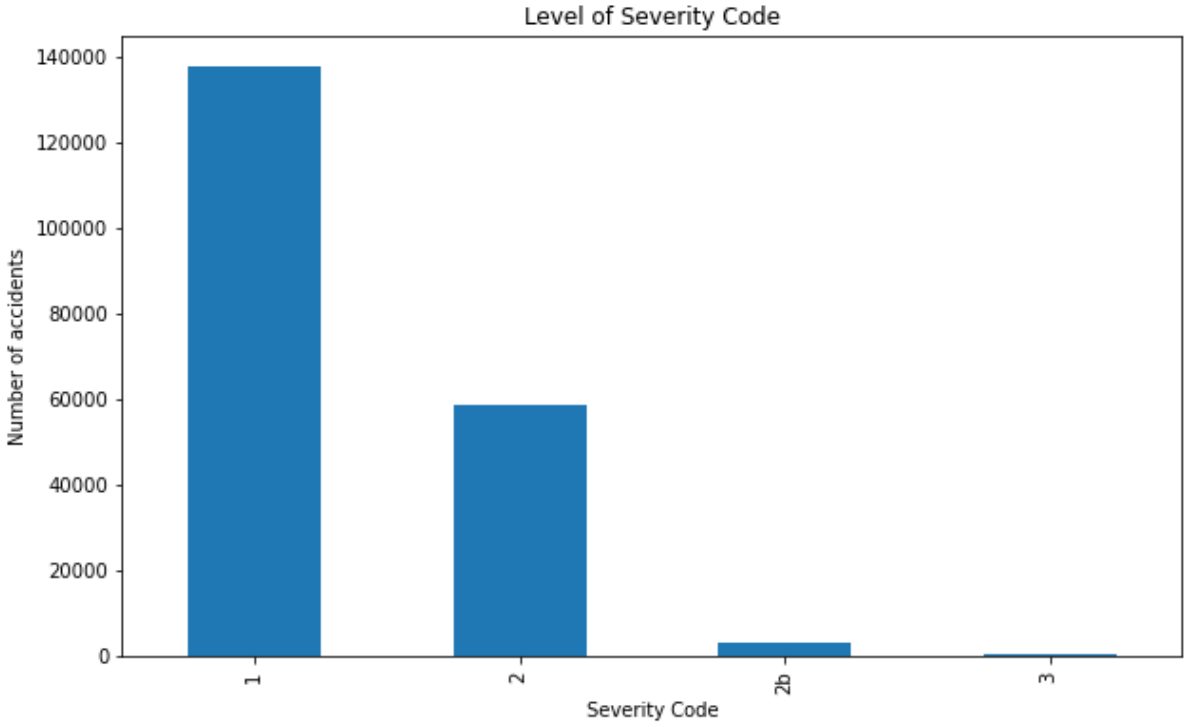
We dropped the ‘Unknown’ Feature type from each feature and also the feature types which had too less of a count were dropped from each feature individually.

One hot encoding of the data was done for the categorical variables using the pandas get dummies function and then the ‘NaN’ values were dropped using the pandas dropna() function.

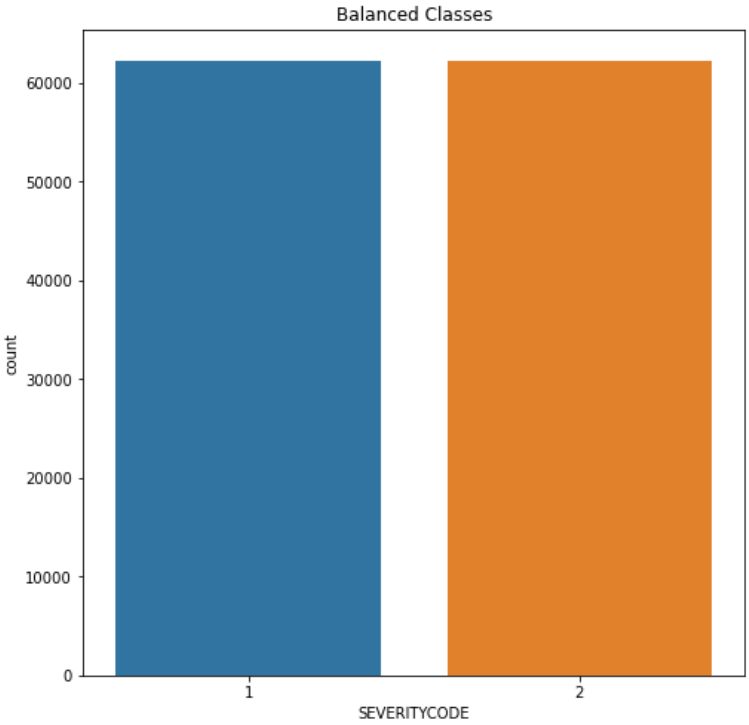
After cleaning the data we used **SciKit Learn’s** function Test Train Split to split our data into 30% test and 70% train. **Final Data Count-**

```
Train set: (77286, 30) (77286,)
Test set: (33123, 30) (33123,)
```

Initial Unbalanced Data

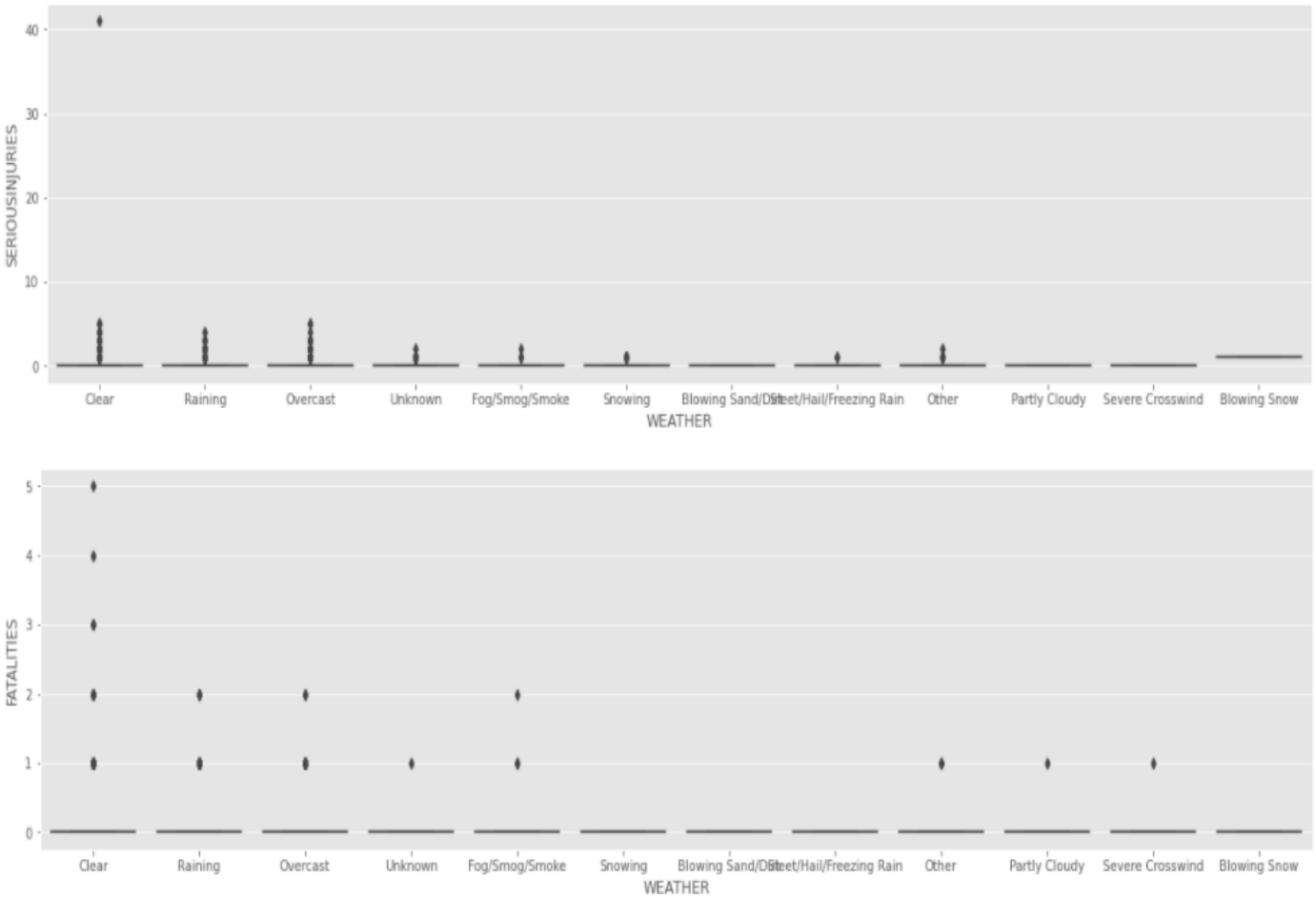


Balanced data



3. Exploratory Data Analysis

Serious Injuries and Fatalities in Different Weather



We can see that the serious injuries and fatalities are unevenly distributed among the weather and the Clear weather accounts mostly for both. Also we can see that there are some variables that are little to none in quantities so we will drop these variables.

In the **similar fashion we analyzed rest of the features** and based on the distribution we decided to drop the Feature types that were not much beneficial to our prediction model. Hence we created a clean data set for our categorical variables.

4. Predictive Modelling

We used classification algorithms since our end result was the classification of a probable incident among two classes i.e. Severity Code ‘1’ and Severity Code ‘2’. We used four algorithms for our classification problem-

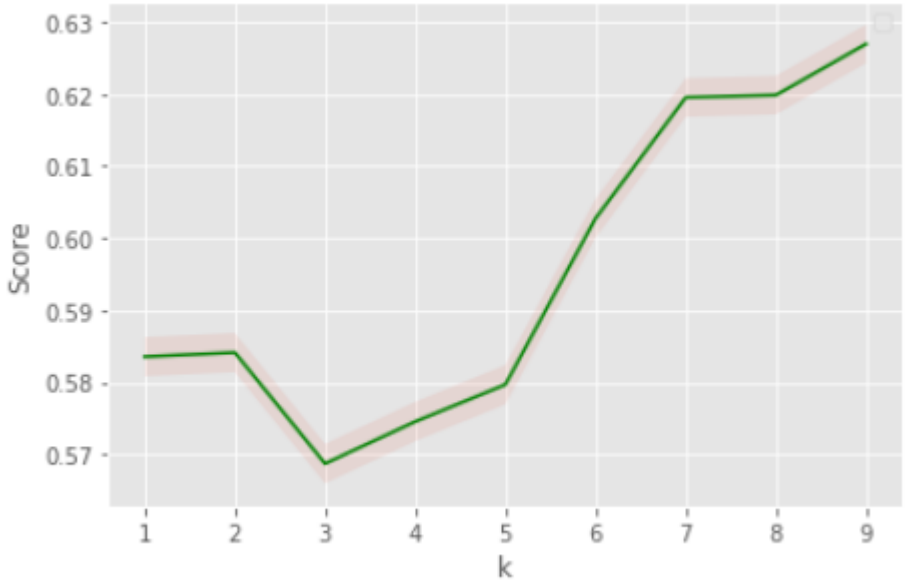
- KNN
- Support Vector Machine

- Decision Tree Classification
- Logistic Regression

KNN Classification Model

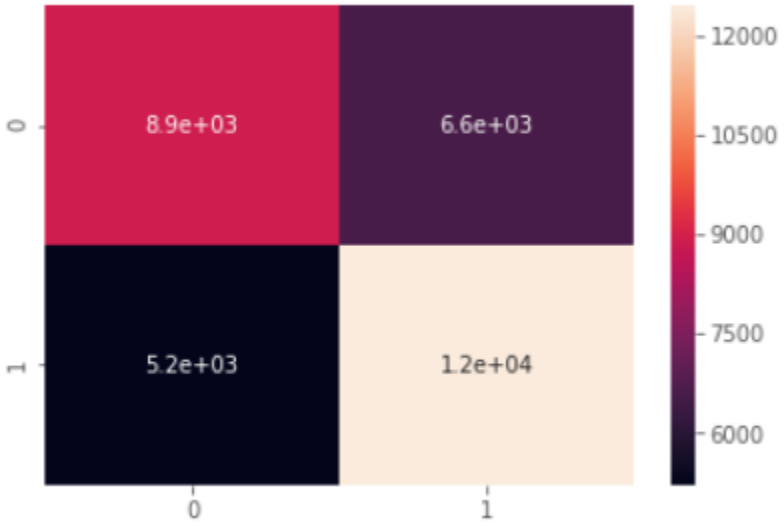
We ran the algorithm from cluster values of 0 to 10 and based on the best result we formulated our model.

K=9 provided us with the best accuracy score for our train data.



Support Vector Machine

The following Confusion Matrix was obtained

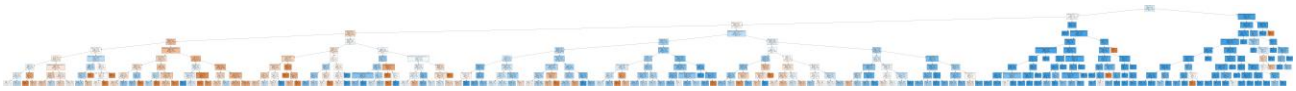


Decision Tree Classifier

We ran the decision tree model for varying depth of 1 to 10 and based on the best accuracy score we chose the depth for our model.

```
Accuracy Score for Tree depth 1 : 0.5331944570238203
Accuracy Score for Tree depth 2 : 0.5698457265344322
Accuracy Score for Tree depth 3 : 0.641819883464662
Accuracy Score for Tree depth 4 : 0.646046553754189
Accuracy Score for Tree depth 5 : 0.6462880777707334
Accuracy Score for Tree depth 6 : 0.6468616973100263
Accuracy Score for Tree depth 7 : 0.6481296983968844
Accuracy Score for Tree depth 8 : 0.6493071279775383
Accuracy Score for Tree depth 9 : 0.6512997011140296
Accuracy Score for Tree depth 10 : 0.6500920810313076
```

The Tree obtained-



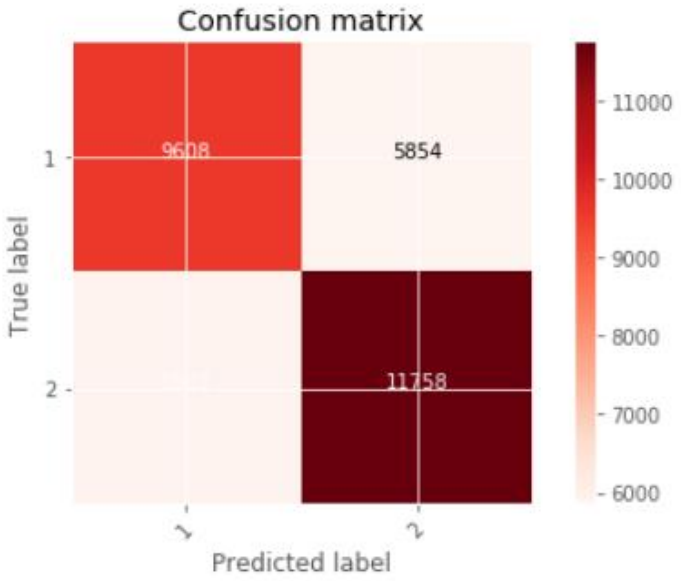
It isn't visible so I am providing link to the image-

https://github.com/Parag-Sharma/Projects/blob/6f3f69e4bcab4788a1dcfed36130b552a65a73f1/IBM_capstone_Project/tree.png

Logistic Regression

We chose Liblinear for our modelling and ran the test for different values of C to obtain the best results for the model.

The following confusion matrix was obtained-



```
                precision    recall  f1-score   support

             1       0.62      0.62      0.62     15462
             2       0.67      0.67      0.67     17661

   micro avg       0.65      0.65      0.65     33123
   macro avg       0.64      0.64      0.64     33123
weighted avg       0.65      0.65      0.65     33123

Confusion matrix, without normalization
[[ 9608  5854]
 [ 5903 11758]]
```

4.1 Choosing the best model based on scores

Algorithm	Jaccard	F1-score	LogLoss
Logistic Regression	0.644416	0.644726	0.608068
Decison Tree	0.651300	0.649031	NA
KNN	0.626936	0.626915	NA
SVM	0.644114	0.642526	NA

Based on the following matrices we choose the Decision Tree as the best classifier for our model.

5. Conclusion and Future Direction

The project was focused on finding the best model for predicting the severity code of the Seattle traffic data and the built model was able to predict the test data with 65% accuracy. Although the variables available were not enough to increase the accuracy of model (Provided we chose variables only which can be available to us prior to any incident). The created model's efficiency can be increased by concatenating other features to our feature set. The current models relies on the road condition and Weather condition along with the Persons and Vehicles involved hence it is limited to the conditions that have previously originated.