Car Accident Severity

Introduction

Car accidents are considered to be one of the most common form of accidents in the world. Accidents can take place due to various reasons. It can be due to driver negligence, bad weather, bad road quality, bad traffic and varius other man-made or environmental reasons. It is important to understand the severity of the car accidents, the cause for these accidents and how to reduce the number of accidents occurring every year.

Data

The collision data chosen can be found from the link: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

The dataset provides several attributes such as the weather during the time of accident (WEATHER), road condition (ROADCOND), visbility of the area (LIGHTCOND) and type of road junction (JUNCTIONTYPE). It also shows a severity type and the collision type of the accident which provides a detailed description of the type of accident and the conditions it occurred in.

The solution will be to compare a predictor with the SEVERITYCODE as the target variable. It will be able to measure and proedict the severity of an accident based on a scale of 0-5. The attributes used will be 'WEATHER', 'ROADCOND' and 'LIGHTCOND'. The scale is described as below:

a) 0: Little to No Probability (Clear Weather Conditions)

b) 1: Very Low Probabibility (Chance of Property Damage)

c) 2: Low Probability (Chance of Injury)

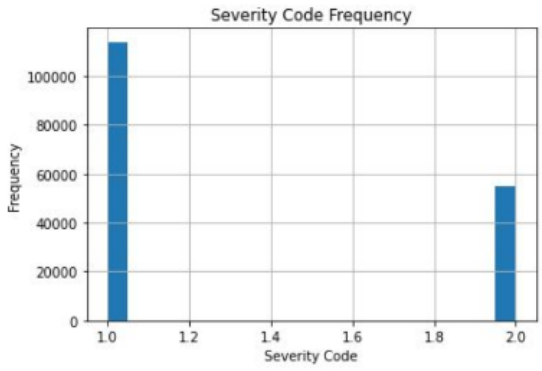
d) 4: High Probability (Chance of Fatality)

The dataset had some missing data which would be required for the algorithm. These values were filled with “N/A” or “Unknown” in the number of vehicles or persons injured. As a result, these rows had to be ignored as it would give an inaccurate result with the algorithms used.

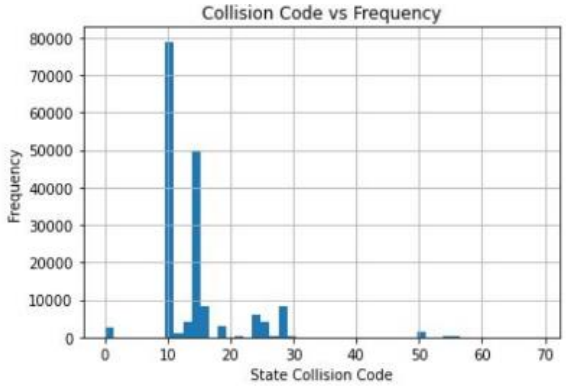
Methodology

First, to compare the data with predictors, histogram plots were made with several predictors against the severity codes.

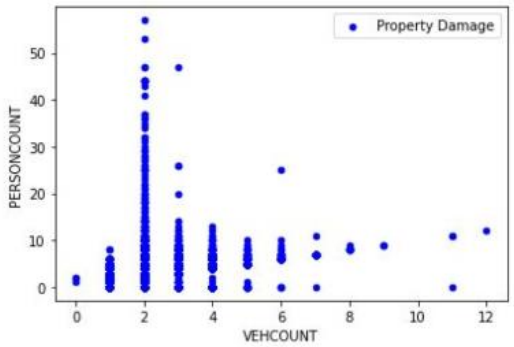
The plot of Frequency against Severity Code showed that the severity code is 1 for the maximum of cases. A severity code of 2 was also generated for frequency of around 5800 cases.



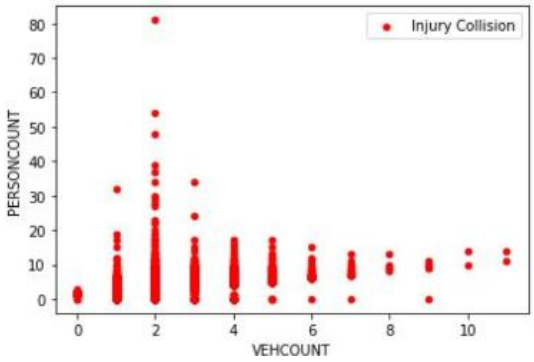
The collision type was also plotted against frequency. This showed a better specifications than the severity code. It showed a code of 10 or 11, representing “entering at an angle” or “both going straight, both moving, sideswipe” respectively was the highest amongst the other type of collisions.



Next graph plotted was property damaged during collisions.



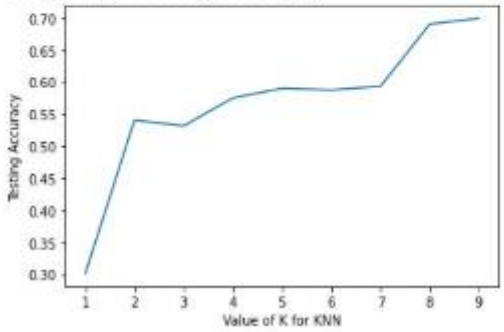
The next graph is similar to the property damage graph except that it takes into consideration the collisions resulting into injury.



A Decision Tree and K – Nearest Neighbour (KNN) was then implemented to predict the accident severity. However, the Decision Tree was results was not too precise as shown below.

The KNN implementation however resulted into a much better prediction for accuracy as shown below.

```
Test set Accuracy at k= 1 :  0.30121733966745845
Test set Accuracy at k= 2 :  0.5406360424028268
Test set Accuracy at k= 3 :  0.5316246382802811
Test set Accuracy at k= 4 :  0.5752497371188223
Test set Accuracy at k= 5 :  0.5901874914465581
Test set Accuracy at k= 6 :  0.5876661840200079
Test set Accuracy at k= 7 :  0.5934058393659725
Test set Accuracy at k= 8 :  0.69045830202855
Test set Accuracy at k= 9 :  0.6993428081193953
Text(0, 0.5, 'Testing Accuracy')
```



Results

The K value of 9 showed the highest accuracy at 0.699. A predicted value of y_hat was then put in and the model predicted 12 correct values out of 20 trials. Using this model, the accuracy for predicting the car accident severity was found to be 68.5% accurate.

```
X= df[["VEHCOUNT", "PERSONCOUNT", "SDOT_COLCODE", "SEGLANEKEY"]].values
y = df["SEVERITYCODE"].values
print("Actual values of the test cases: " + str(y[0:20]))
```

```
Actual values of the test cases: [2 1 1 1 2 1 1 2 1 2 1 1 1 2 2 2 2 1 2]
```

```
k = 9
KNN = KNeighborsClassifier(n_neighbors = k).fit(X_train, y_train)
y_hat = KNN.predict(X)
print("Predicted values using k = 9: " + str(y_hat[0:20]))
```

```
Predicted values using k = 9: [1 1 2 2 1 1 1 2 1 1 1 1 1 1 1 2 1 2 1 1]
```

```
print("KNN F1-Score: " + str(f1_score(y, y_hat, average = "weighted")))
print("KNN Jaccard Score: " + str(jaccard_score(y, y_hat)))
```

```
KNN F1-Score: 0.6846631437653313
KNN Jaccard Score: 0.7011003992599084
```

Discussion

The model was hard to work with as a lot of the values were inconsistent and missing. It was difficult to test for over sampling because of highly imbalanced classes. Most of the algorithms used showed that it is biased towards the most frequent class. Efficient pre-processing and corresponding imbalanced data techniques should give optimal results. However, it showed that KNN model provided better accuracy in determining the accuracy of the model.

Conclusion

During this study several relationship factors and their link with severity code for collisions were compared and analysed. Two classification models namely Decision Tree and KNN model was

used to predict the severity of a car accident. The accuracy of the developed model was found to be better calculated using the KNN model.