Car Accident Severity

Monika Singhal

1. Introduction

1.1 Background

In spite of significant advances in road safety, a lot of crashes in high severities still occur. Investigation of influential factors on crashes enables engineers to carry out calculations in order to reduce crash severity. According to the World Health Organization there were approximately 1.35 million people killed on roadways around the world. Averagely there are almost 3,700 people killed globally in road traffic crashes. The causes of accident ranges from weather condition, driver condition to road condition. Prediction of traffic accident severity is one of the crucial steps in the traffic accident management process. Timely recognition of key influences may play an important role for improving traffic safety. This study intends to contribute for a better understanding of the factors that affect the occurrence of accidents and those that affect its severity.

1.2 Problem

Accidents in traffic lead to associated fatalities and economic losses every year and thus is an area of primary concern to society from loss prevention point of view. Modeling accident severity prediction and improving the model are critical to the effective performance of road traffic systems for improved safety. Predicting the probability and severity of vehicular accidents based on weather and other characteristics, can help in improving the traffic accident management process. Accident severity prediction can provide crucial information for emergency responders to evaluate the severity level of accidents, estimate the potential impacts, and implement efficient accident management procedure

1.3 Solution

In this report, I present the classification models for road traffic accidents that predict the severity of injuries received by the accidents. First, I extracted the feature information on traffic accidents and store them into an accidental feature database. Accidental feature data are split into two disjoint sets: training and test dataset. Next, I generate classification models through machine learning techniques using the training data set, and then verify the performance of the model produced using the test data set. I have verified each model, and the experimental results show that our classification models achieve the considerable classification accuracy. The finding of this study can be used to the traffic accident management system for improving the prediction accuracy of traffic accident severity.

1.4 Interest

This analysis will interest the people who are driving as they can consider various factors before driving to avoid the accidents. It could also be used by the Department of Transportation to find ways on how to improve the traffics within UK and provide safeness to its people.

2. Data Description

2.1 Data Source

This dataset provides detailed road safety data about the circumstances of personal injury road accidents in 2018 in UK, the types of vehicles involved and the consequential casualties. The statistics relate only to personal injury accidents on public roads that are reported to the police, and subsequently recorded, using the STATS19 accident reporting form. It contains 157342 rows and 47 columns. It has been derived from https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

There are two files: *Road_Safety_Data-Accident 2018* which has 122635 rows and 32 columns. File *Road_Safety_Data-Casualties 2018* has 160597 rows and 16 columns. Both the files have been joined on a common feature Accident_Index.

3. Methodology

3.1 Data Pre-processing

Upon examining the features, it was found that there were a few features that does not contribute to the prediction of accident severity. For example, Location_Easting_OSGR

, Location_Northing_OSGR, Casualty_IMD_Decile, 2nd_Road_Class. These few features were drop instantly from data modeling consideration.

	Accident_Index	Location_Easting_OSGR	Location_Northing_OSGR	Longitude	Latitude	Police_Force	Accident_Severity	Number_of_Vehicles	Number_of_Casualties	Date	...	Age_Band_of_Casualty
0	2018010080971	529150.0	182270.0	-0.139737	51.524587	1	3	2	2	01/01/2018	...	NaN
1	2018010080973	542020.0	184290.0	0.046471	51.539651	1	3	1	1	01/01/2018	...	NaN
2	2018010080974	531720.0	182910.0	-0.102474	51.529746	1	3	2	1	01/01/2018	...	NaN
3	2018010080981	541450.0	183220.0	0.037828	51.530179	1	2	2	1	01/01/2018	...	NaN
4	2018010080982	543580.0	176500.0	0.065781	51.469258	1	2	2	2	01/01/2018	...	NaN

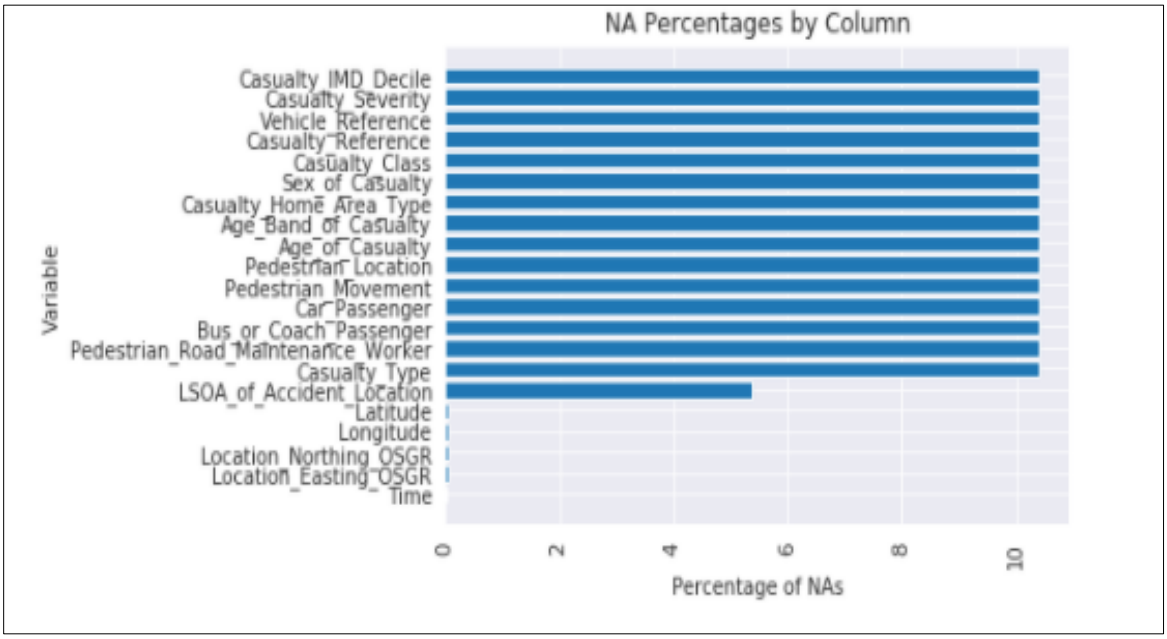
After dropping:

```
df.head()
```

	Accident_Index	Longitude	Latitude	Police_Force	Accident_Severity	Number_of_Vehicles	Number_of_Casualties	Date	Day_of_Week	Time	...	Age_Band_of_Casualty	Casualty_Severity	Pedestr
16384	2018010130027	-0.215455	51.556878	1	3	2	1	01/09/2018	7	12:48	...	6.0	3	
16385	2018010130028	-0.299739	51.552014	1	3	1	2	01/09/2018	7	14:00	...	6.0	3	
16386	2018010130028	-0.299739	51.552014	1	3	1	2	01/09/2018	7	14:00	...	1.0	3	
16387	2018010130030	0.057318	51.524353	1	3	2	2	01/09/2018	7	11:40	...	7.0	3	
16388	2018010130030	0.057318	51.524353	1	3	2	2	01/09/2018	7	11:40	...	7.0	3	

5 rows × 45 columns

Also, it was examined that almost 10% of the data was missing from many of the labels. It was decided to drop the null values as assuming the values would have introduced error.



It was assumed that Accident_Severity and Casualty_Severity were highly correlated with correlation factor of 0.86. So, the missing values of casualty_Severity were imputed on the basis of Accident_Severity.

```
#checking the correlation between accident severity and casualty severity
df[['Accident_Severity', 'Casualty_Severity']].corr()
```

35]:

	Accident_Severity	Casualty_Severity
Accident_Severity	1.000000	0.851583
Casualty_Severity	0.851583	1.000000

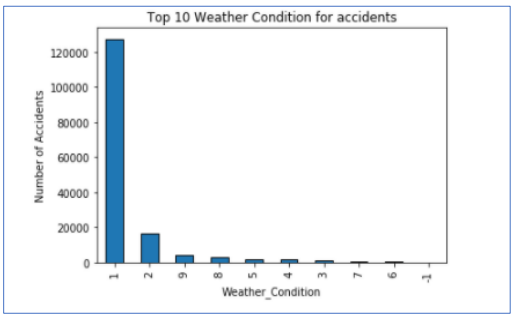
The Accident Severity was pre-divided into three categories:

```
3    102837
2     27213
1      2450
```

With 1 being Fatal, 2 being Severe and 3 being Slight. So, after looking at the counts we merged two close categories (Serious and Fatal) into one and creating classification dummy for Severity of accident (Serious = 1, Slight = 0)

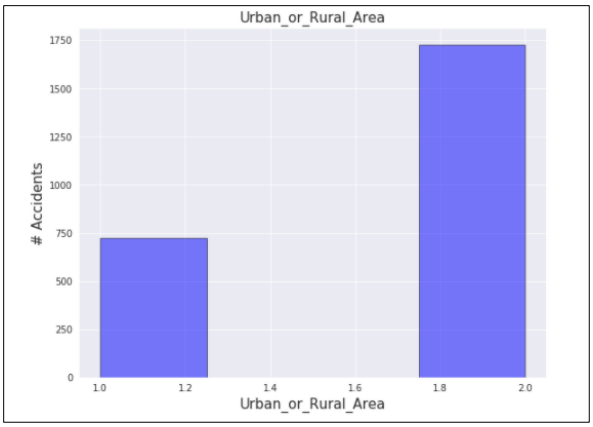
```
df['Severity_dummy'].value_counts()

0    102837
1     29663
```

More accidents occur in Fine weather with no high wind speed

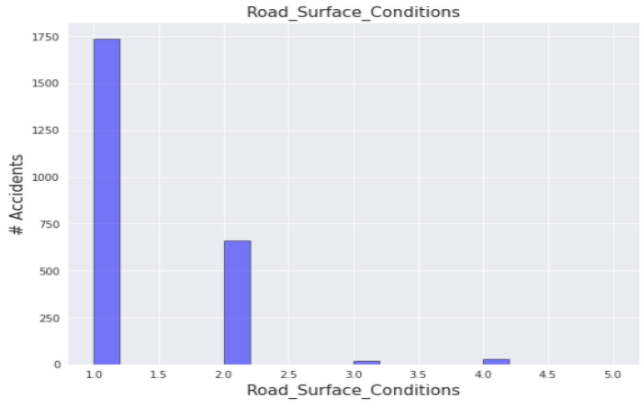


code	label
1	Fine no high winds
2	Raining no high winds
3	Snowing no high winds
4	Fine + high winds
5	Raining + high winds
6	Snowing + high winds
7	Fog or mist
8	Other
9	Unknown
-1	Data missing or out of range

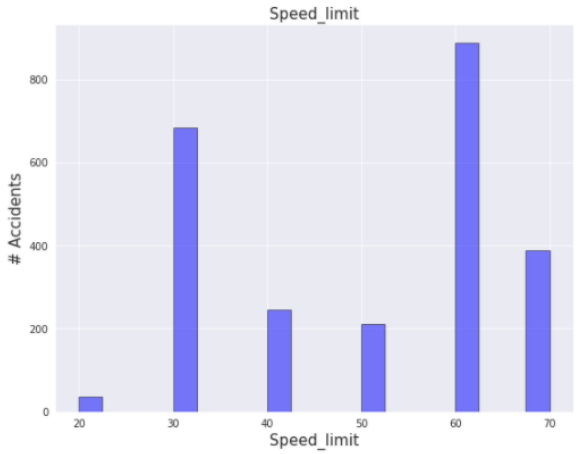
More accidents occur in Rural area than Urban area



More accidents occur on dry road surface



code	label
1	Dry
2	Wet or damp
3	Snow
4	Frost or ice
5	Flood over 3cm. deep
6	Oil or diesel
7	Mud
-1	Data missing or out of range



Data Analysis

There were a few classification methods that was considered for the modeling. KNN, Decision Tree, Logistic Regression and Random Forest. The dataset was then split into training and testing data with size of 80/20. The data was split into train and test data as to not have overfitting in the results which leads to inaccurate prediction for general datasets. By running the Logistic Regression, it was calculated that accuracy is 77.9%. By running the model and using Decision Tree, it was calculated that the accuracy is 78%. By having a Decision Tree of depth 6, the F1 score for the test set was calculated to be 84.93%. By running the KNN, it was calculated that the train accuracy was 80.9% and test accuracy was 76.08%. We predicted using the K value of 6. However, the best accuracy of 77.2% can be achieved by the k value of 2. The average F1 score is 0.75.

Results

We have verified each model, and the experimental results show that our classification models achieve the considerable classification accuracy. The finding of this study can be used to the traffic accident management system for improving the prediction accuracy of traffic accident severity. In this study, the logistic regression model and decision tree model were used to examine the influence of a number of factors on the injury severity faced by motor-vehicle occupants involved in road accidents. The model estimation results suggest that some types of road accidents, namely the rollover-type, run-off-road, collisions against fixed objects and head-on collisions, appear to be the major contributors for the most severe injury level. Also, those who travel in a light vehicle, at a two-way road and on dry road surface tend to suffer more severe injuries than those who travel in a heavy-vehicle, at a one-way road, and on a wet road surface. In contrast, the driver´s seat is clearly the safest seating position, and urban areas, although presenting the highest accident occurrence frequency, are linked to decreased severity level. Also, women tend to be more likely to suffer serious or fatal injuries than men.

Conclusion

In this study, the relation between several data sets and the severity of accidents were found. The data used has a direct correlation on the severity of the accidents. This model could be very useful for the UK Department of Transportation as they could use it to determine red zones for accident and could deploy more staffs on maintaining the safeness of road users based on certain conditions. The department responsible could also improve or maintain the infrastructure of the road such as the light conditions or the road conditions.

Road users could also determine the risk of driving when route planning as to avoid certain conditions that could result in an accident.