Severity classification of traffic accidents in the UK

Introduction

Traffic accidents may be caused by a long list of factors including weather, mechanical problems, road conditions, driver state of mind, road design.

A traffic accident occurs when a vehicle collides with another vehicle, pedestrian, animal, road debris, or other stationary obstruction, such as a tree, pole or building. Traffic collisions often result in injury, disability, death, and property damage as well as financial costs to both society and the individuals involved. In 2013, traffic accidents resulted in 1.4 million fatalities and 54 million injuries. [Wikipedia]

Proper reporting, recording and analysis of traffic accidents data will be helpful for urban planners, car designers and drivers to paint a more detailed picture on where, when and how accidents happen, leading to reductions in the total number of traffic accidents and hopefully elimination of fatal accidents all together. This report is intended to present the results of traffic accidents data analysis and ML modelling with the aim of predicting the severity of accidents based on a number of factors.

Data

This analysis is based on data made available by the government of UK, the dataset lists over 2 million records of accidents in the time frame between 2005 and 2017 within the geographical boundries of UK. a total of 34 attributes details each of the accidents records, as follows;

1 - Accident_Index

2 - 1st_Road_Class

3 - 1st_Road_Number

4 - 2nd_Road_Class

5 - 2nd_Road_Number

6 - Accident_Severity

7 - Carriageway_Hazards

8 - Date

9 - Day_of_Week

10 - Did_Police_Officer_Attend_Scene_of_Accident

11 - Junction_Control

12 - Junction_Detail

13 - Latitude

14 - Light_Conditions

15 - Local_Authority_(District)

16 - Local_Authority_(Highway)

17 - Location_Easting_OSGR

18 - Location_Northing_OSGR

19 - Longitude

20 - LSOA_of_Accident_Location

21 - Number_of_Casualties

22 - Number_of_Vehicles

23 - Pedestrian_Crossing-Human_Control

24 - Pedestrian_Crossing-Physical_Facilities

25 - Police_Force

26 - Road_Surface_Conditions

27 - Road_Type

28 - Special_Conditions_at_Site

29 - Speed_limit

30 - Time

31 - Urban_or_Rural_Area

32 - Weather_Conditions

33 - Year

34 - InScotland

Following are general observations based on data explorations;

- The number of missing data ("unknown" or "data missing or out of range") is minor and can be removed without impacting the integrity of the dataset.
- The data set is evidently imbalanced when it comes to Accident Severity, the three categories (Slight, Serious, Fatal) are distributed at 84.7%, 14% and 1.3% respectively.
- It was noticed that "Hazards" are in some cases indicated as "none" while special conditions indicate present hazards, such as "oil/diesel spell" for example. Hence the decision to merge both attributes in one.
- The number of accidents vary depending of weekdays, the highest number of accidents, 335183, took place on Fridays and the lowest being 225327 on Sundays.
- The data shows that around 40% of accidents happen with no junction/crossing within 20 meters, however, 30% took place at T-junctions.
- Surprisingly, 80% of accidents took place on days with fine weather.
- Almost 70% of accidents took place on dry roads and approximately 28% on wet/damp roads
- Almost 75% of accidents took place on single carriageway (undivided roads)

In the methodology section we will discuss how attributes were selected, cleaned and modelled using different machine learning algorithms to predict the severity of a traffic accidents.

Data wrangling and cleaning

The dataset dimensions were 2,047,256 x 34, which proved heavy and consumed a lot of time and processing resources to wrangle and model. Therefore, a random sample of 5% was taken for the purpose of this project, with dimensions of (102,363 x 34).

Out of 34 attribute, the following 11 attributes were selected to prepare a feature set that will eventually be used for modelling;

- Accident_Severity

- Date

- Day_of_Week

- Junction_Control

- Junction_Detail

- Light_Conditions

- Road_Surface_Conditions

- Road_Type

- Speed_limit

- Urban_or_Rural_Area

- Weather_Conditions

Next, all occurrences of '*Data missing or out of* range', *'Unallocated'* or *'Unknown'* were replaced with NaN and removed along with originally existing NaN or missing values.

Date column was reformatted to Date/Time data type, then month and year were extracted into separate columns, eventually dropping the original Date column.

As the dataset was heavily imbalanced, a sampling compensation was performed. Out of the 102K rows; 54643 were categorized "*Slight*", 8022 "*Serious*", 559 "*Fatal*". The number of Slight records was under-sampled and the Fatal records were over-sampled to match the 8022 number of records of the "Serious" category. This will not impact the model accuracy since the objective is to predict severity of an occurrence rather than predict probabilities of occurrences. The end result was a dataframe with 24066 evenly distributed between all 3 Severities.

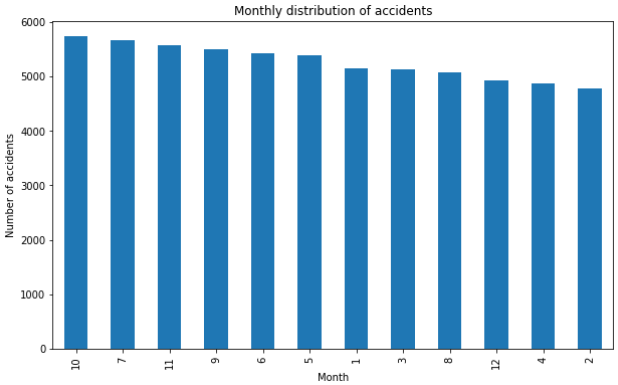
Values under Lighting Conditions we re-categorized into *Daylight* and *Dark/Night*.

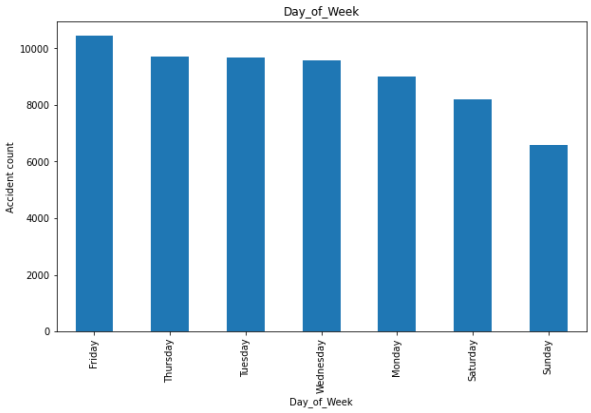
Methodology

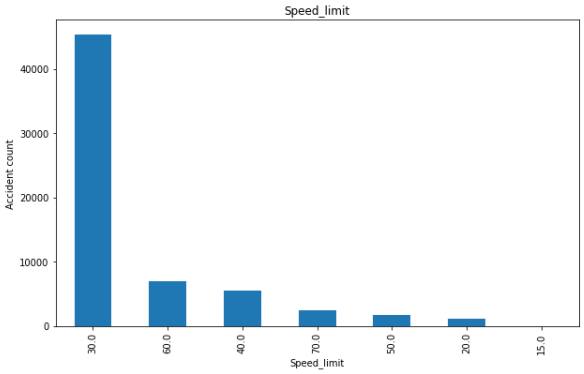
Feature set visualization

Below are some of the attributes visualized

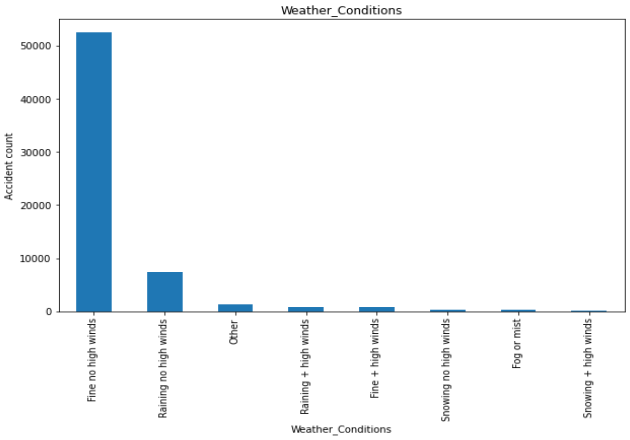
- Highest number of accidents occur in October and the least is in February
- Most accidents happen on Fridays with the least on Sunday

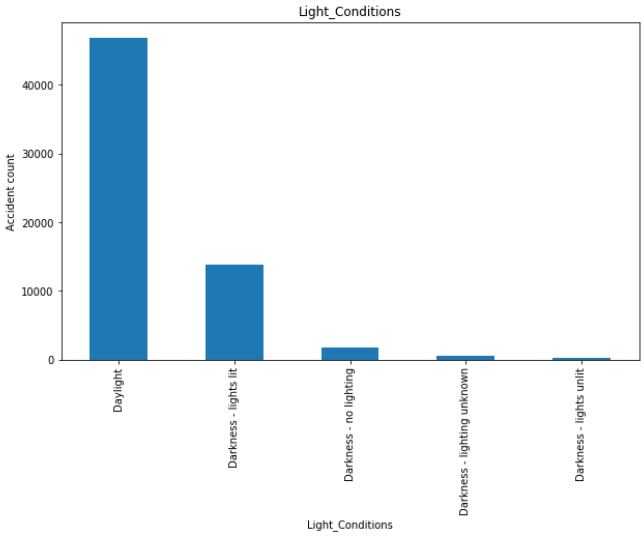




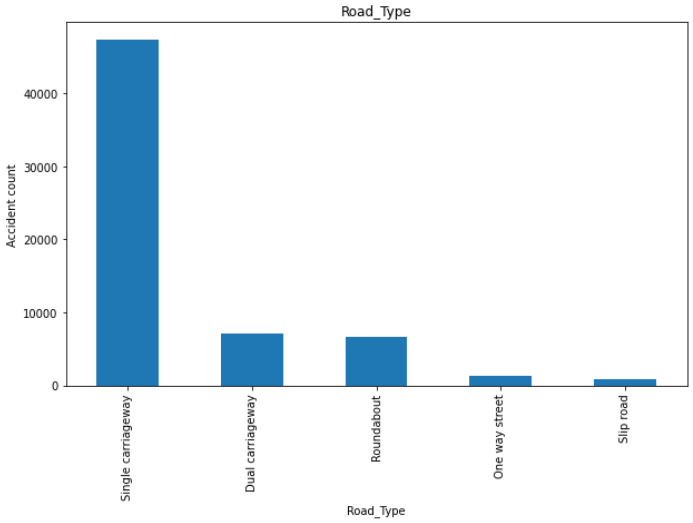


- Most accidents took place on roads with 30 mph speed limit. (left)
- Most accidents took place in good weather. (below)

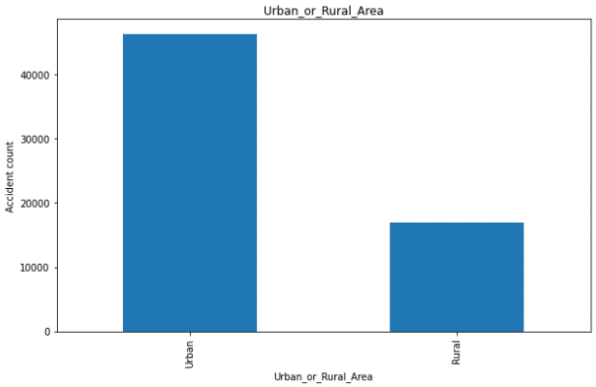


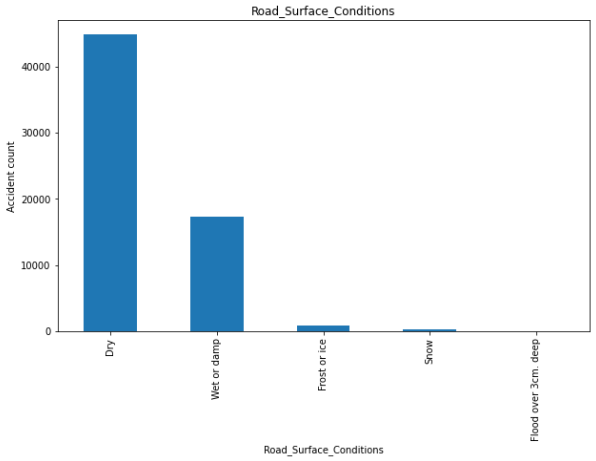


- Most accidents took place on broad daylight. (left)
- Most accidents took place on single carriageways. (below)



- Most accidents took place in Urban areas
- Most accidents took place on roads with dry surfaces





First 5 rows of the feature set

Accident_Severity	Day_of_Week	Junction_Control	Junction_Detail	Light_Conditions	Road_Surface_Conditions	Road_Type	Speed_limit	Urban_or_Rural_Area	Weather_Conditions	year	month
Fatal	Friday	Give way or uncontrolled	Other junction	Daylight	Wet or damp	Single carriageway	30.0	Urban	Fine no high winds	2008	10
Fatal	Tuesday	Give way or uncontrolled	Roundabout	Daylight	Dry	Dual carriageway	60.0	Rural	Fine no high winds	2005	6
Fatal	Friday	Give way or uncontrolled	Crossroads	Daylight	Wet or damp	Single carriageway	60.0	Rural	Raining no high winds	2012	9
Fatal	Friday	Give way or uncontrolled	T or staggered junction	Dark/Night	Dry	Single carriageway	30.0	Urban	Fine no high winds	2017	10
Fatal	Friday	Give way or uncontrolled	T or staggered junction	Dark/Night	Wet or damp	Single carriageway	30.0	Urban	Raining + high winds	2005	9

Modelling

One hot encoding was used to convert the feature set into a binary matrix, which was split into a training set of 16,846 records and a test set of 7,220.

A classification model is required to predict a given accident into one of 3 classes (namely Fatal, Serious and Slight). The training set was used to build 4 classification models;

K Nearest Neighbor(KNN)

A non-parametric classification and regression algorithm.

Decision Tree

A predictive modelling algorithm based on decision tree flowchart structure to move from observations to conclusions on target values.

Support Vector Machine

A machine learning algorithms that aims to building separation lines or hyperplanes to classify a scatter of data points into classification groups by finding that maximum distance between hyperplanes and the nearest data point.

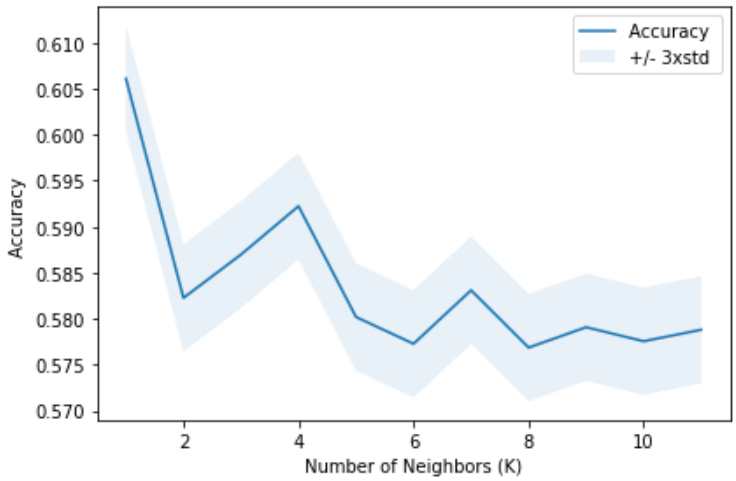
Logistic Regression

A multinomial version of the binary logistic regression can be helpful in cases of multi-class classifications.

Results

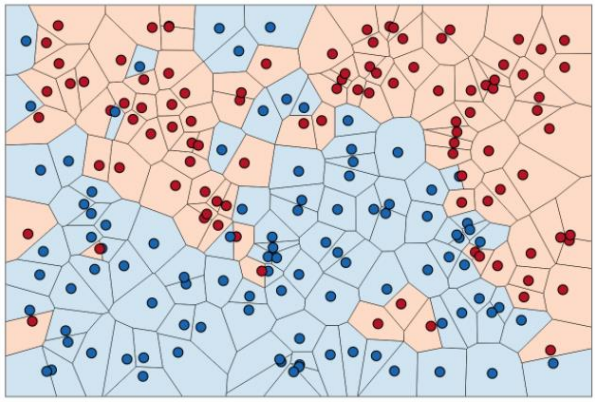
K Nearest Neighbor(KNN)

Best K value was identified to be 1 with an accuracy score of 0.6




```
The best accuracy was with 0.6060941828254848 with k= 1
```

A K value of 1 can be visualized as shown in the figure below



Source: https://stats.stackexchange.com/questions/367010/training-error-in-knn-classifier-when-k-1/367015

Sample comparison between actual values and values predicted by KNN model

```
# Fitting the model using best K
neigh = KNeighborsClassifier(n_neighbors=mean_acc.argmax()+1).fit(X_train, y_train)
yhat = neigh.predict(X_test)
print('sample of predicted values',yhat[0:10])
print('corresponding actual values',y_test[0:10])
```

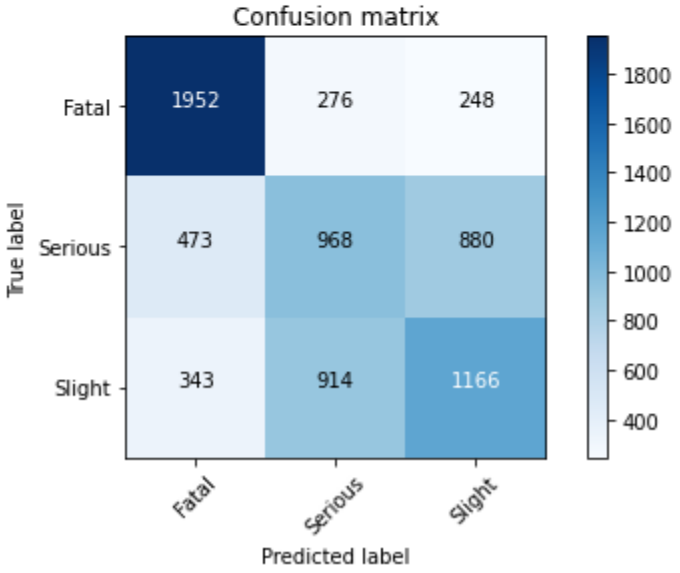
```
sample of predicted values ['Fatal' 'Slight' 'Serious' 'Serious' 'Slight' 'Serious' 'Fatal' 'Slight'
 'Serious' 'Fatal']
corresponding actual values ['Fatal' 'Serious' 'Serious' 'Slight' 'Slight' 'Serious' 'Fatal' 'Fatal'
 'Slight' 'Fatal']
```

Support Vector Machine

A confusion matrix was built to evaluate the model output'

	precision	recall	f1-score	support
Fatal	0.71	0.79	0.74	2476
Serious	0.45	0.42	0.43	2321
Slight	0.51	0.48	0.49	2423
accuracy			0.57	7220
macro avg	0.55	0.56	0.56	7220
weighted avg	0.56	0.57	0.56	7220

```
Confusion matrix, without normalization
[[1952  276  248]
 [ 473  968  880]
 [ 343  914 1166]]
```



The confusion matrix above show that the model was able to correctly predict 1952 Fatal severity cases out of 2476 with accuracy of 0.74. However, accuracy for predicting Slight and Serious severity cases were lower with accuracy of 0.49 and 0.43 respectively. The average accuracy of the model was at 0.57.

Decision Tree

Decision tree model achieved around 0.64 accuracy.

Sample comparison between actual values and values predicted by decision tree model

```
sample of predicted values ['Fatal' 'Slight' 'Serious' 'Serious' 'Slight' 'Serious' 'Fatal' 'Serious'
 'Serious' 'Fatal']
corresponding actual values ['Fatal' 'Serious' 'Serious' 'Slight' 'Slight' 'Serious' 'Fatal' 'Fatal'
 'Slight' 'Fatal']
```

Logistic Regression

Decision tree model achieved around 0.47 accuracy.

Sample comparison between actual values and values predicted by logistic regression model

```
sample of predicted values ['Fatal' 'Serious' 'Fatal' 'Slight' 'Slight' 'Fatal' 'Serious' 'Serious'
 'Slight' 'Slight']
corresponding actual values ['Fatal' 'Serious' 'Serious' 'Slight' 'Slight' 'Serious' 'Fatal' 'Fatal'
 'Slight' 'Fatal']
```

Discussion

Model comparison

The below table shows accurecy score of 4 models used for Accident Severity prediction

Algorithm	Jaccard	F1-score	Log Loss
KNN	0.610942	0.601854	NA
Decision Tree	0.649169	0.639322	NA
SVM	0.565928	0.560169	NA
Logistic Regression	0.473546	0.468204	1.01672

The above table shows that the decisoin tree model has achieved the best accurecy, closely followed by KNN model.

Observations:

Some of the observations noted during the course of this study;

- Fatal accidents are near evenly distributed through out the weekdays, with Saturdays being the highest at 17% and Tuesday being the lowest at 11%. Distributions of Serious and Slight accidents across the week are very similar to that of Fatal accidents.
- Give way junctions (uncontrolled) are home to almost 68% of Fatal accidents, 78% of Serious accidents and 76% of Slight accidents.
- T-junctions are where almost 50% of all accidents take place.

- Daytime is when 60% of Fatal accidents, 70% of Serious accidents and 75% of Slight accidents occur.
- 77% of Fatal accidents take place on single carriageways.
- 42% of Fatal accidents occur on road with speed limits of 30 mph and 31% on roads with 60 mph.

Conclusion

During this study; traffic accident dataset provided by the government of the UK was studied and analyzed, and 4 different classification models were built in efforts to predict severity of traffic accidents. The feature set included weekdays, lighting conditions, weather conditions, road conditions, accident location, among other attributes. The Decision Tree model was able to predict accident severity with an average accuracy of 64% and was able to predict Fatal accidents with accuracy of 74%.

The study has shown that most fatal accidents take place during daytime, fine weather, good road conditions and on low speed limit roads, this suggests that reckless driving could be the cause behind most accidents, specifically, the fatal ones.