Classifying Car Accident Severity in the UK

Kwok Chin Hung, Michael

25 August 2020

1. Introduction

Background

Car Accidents are caused by a lot of different factors, such as weather conditions, time of the day, light conditions or road conditions. Once car accident happened, police and ambulance have to arrive as soon as possible to handle the case. Without any prediction, police and ambulanceman might not be able to be prepared for the case. Therefore, I would like to create a map which fetches the real-time conditions of the traffic and weather and shows the predicted severity of car accidents that might happen. Police and Ambulance might make use of this map to be prepared for the potential accidents.

Problem

Data that might contribute to estimating the accident severity might include day of the week, time of the day, weather, his position, light and road conditions, and location. This project aims to provide an estimated severity of certain location for the police force and ambulance based on these data.

Interest

Police and ambulance might need to prepare enough human resources for accidents. They would be very interested in accurate estimation of the severity in order to be well prepared for the accidents. Others, such as the government, can also investigate into the data and draft policies that can reduce the severity of car accidents in the United Kingdom.

2. Data acquisition and cleaning

Data sources

Most of the historical data can be found in Kaggle datasets, such as here and here. This two datasets includes the information of accidents, such as the Accident Severity, day of the week, road conditions, weather conditions, light conditions, etc. The dataset, “UK Accidents 10 years history with many variables”, is chosen. This is due to the completeness of the accident information.

Data cleaning

The dataset downloaded from Kaggle consists of three tables, namely Accidents0514.csv, Casaulties0514.csv, and Vehicles0514.csv. Accidents0514.csv and Vehicles0514.csv are mainly used. There are 52 variables in total.

However, the dataset includes too many data that might not be relevant to the cause of the accidents, such as Road number, junction control, LSOA, etc. Therefore specific features are selected in order to predict the severity of the potential accidents in that area.

Target Variable

In this project, the target variable is the accident severity. There are three type of accident severity in the dataset, namely slight, serious and fatal.

code	label
1	Fatal
2	Serious
3	Slight

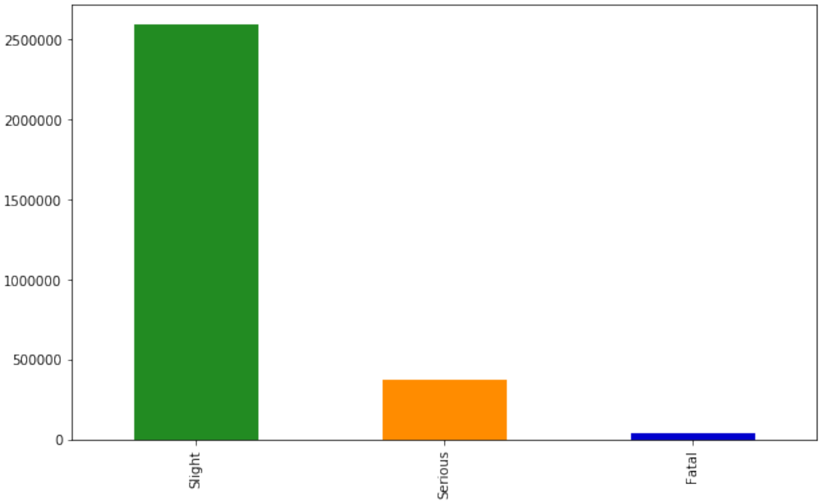

Figure 1 - Accident Severity Distribution

Feature selection

In order to predict the severity of the potential accidents, data has to be uncontrolled. Day of week, time of the day, light conditions, weather conditions, Road surface conditions are selected. In order to predict the severity in each area, local authority and police force are also included. This reduces the feature sets from 52 to 9.

1. **Day of week**
2. **Time (hour)**
3. **Light Condition**

There are five different types of light condition in the dataset. The corresponding label of each code is as follow:

code	label
1	Daylight
4	Darkness - lights lit
5	Darkness - lights unlit
6	Darkness - no lighting
7	Darkness - lighting unknown
-1	Data missing or out of range

4. **Weather Conditions**

There are nine different types of weather condition in the dataset. The corresponding label of each code is as follow:

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

5. **Road conditions**

There are seven different types of light condition in the dataset. The corresponding label of each code is as follow:

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

6. **Number of Vehicles**
7. **Sex of Driver**
8. **Age band of Driver**

There are seven different types of light condition in the dataset. The corresponding label of each code is as follow:

code	label
1	0 - 5
2	6 - 10
3	11 - 15
4	16 - 20
5	21 - 25
6	26 - 35
7	36 - 45
8	46 - 55
9	56 - 65
10	66 - 75
11	Over 75
-1	Data missing or out of range

9. **1st point of impact**

There are four different types of light condition in the dataset. The corresponding label of each code is as follow:

code	label
0	Did not impact
1	Front
2	Back
3	Offside
4	Nearside
-1	Data missing or out of range

10. Road Type

There are seven different types of light condition in the dataset. The corresponding label of each code is as follow:

code	label
1	Roundabout
2	One way street
3	Dual carriageway
6	Single carriageway
7	Slip road
9	Unknown
12	One way street/Slip road
-1	Data missing or out of range

11. Hit Object in Carriageway

There are twelve different types of light condition in the dataset. The corresponding label of each code is as follow:

code	label
0	None
1	Previous accident
2	Road works
4	Parked vehicle
5	Bridge (roof)
6	Bridge (side)
7	Bollard or refuge
8	Open door of vehicle
9	Central island of roundabout
10	Kerb
11	Other object
12	Any animal (except ridden horse)
-1	Data missing or out of range

12. Hit Object off Carriageway

There are twelve different types of light condition in the dataset. The corresponding label of each code is as follow:

code	label
0	None
1	Road sign or traffic signal
2	Lamp post
3	Telegraph or electricity pole
4	Tree
5	Bus stop or bus shelter
6	Central crash barrier
7	Near/Offside crash barrier
8	Submerged in water
9	Entered ditch
10	Other permanent object
11	Wall or fence
-1	Data missing or out of range

13. Vehicle Leaving Carriageway

There are nine different types of light condition in the dataset. The corresponding label of each code is as follow:

code	label
0	Did not leave carriageway
1	Nearside
2	Nearside and rebounded
3	Straight ahead at junction
4	Offside on to central reservation
5	Offside on to centrl res + rebounded
6	Offside - crossed central reservation
7	Offside
8	Offside and rebounded
-1	Data missing or out of range

14. Special Conditions at Site

There are eight different types of light condition in the dataset. The corresponding label of each code is as follow:

code	label
0	None
1	Auto traffic signal - out
2	Auto signal part defective
3	Road sign or marking defective or obscured
4	Roadworks
5	Road surface defective
6	Oil or diesel
7	Mud
-1	Data missing or out of range

15. Skidding and Overturning

There are six different types of light condition in the dataset. The corresponding label of each code is as follow:

code	label
0	None
1	Skidded
2	Skidded and overturned
3	Jackknifed
4	Jackknifed and overturned
5	Overturned
-1	Data missing or out of range

3. Exploratory Data Analysis

Relationship between variables

1. Accident Severity and Time of the day

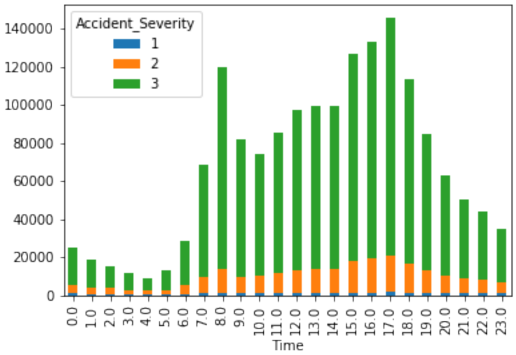

Figure 2 - Accident Severity and Time of the Day

From Figure 1, throughout the day, more accidents happen during peak hours, that is 8:00am and 5:00pm.

2. Accident Severity and Light conditions

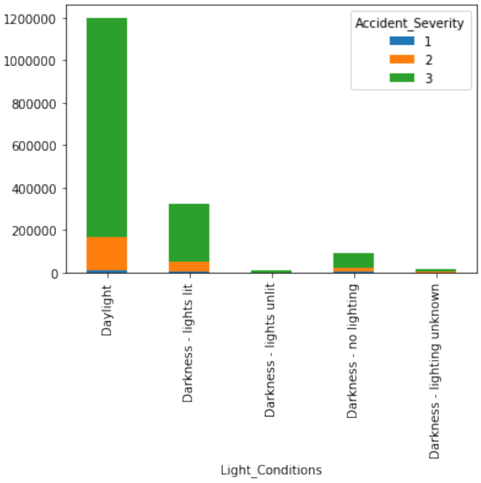

Figure 3 – Accident Severity and Light Conditions

According to Figure 2, most of the accidents happened during daylight time.

3. Accident Severity and Weather conditions

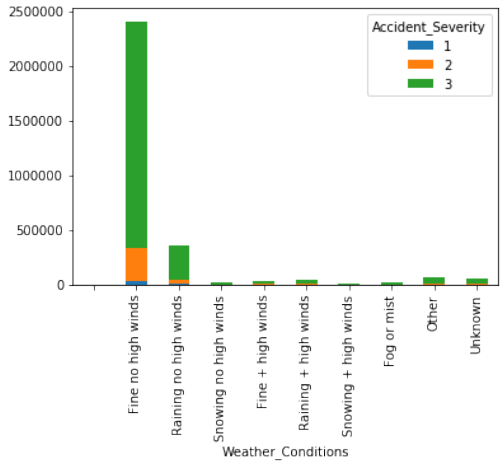

Figure 4 - Accident Severity and Weather Conditions

For Weather conditions, most of the accident happened when there were no high winds.

4. Accident Severity and Sex of driver

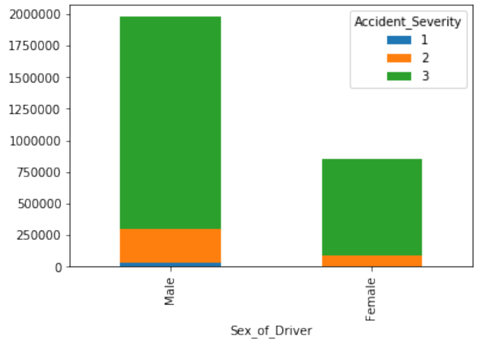

Figure 5 - Accident Severity and Sex of the Driver

Most of the accidents happened when the driver was a male.

5. Accident Severity and 1st point of impact

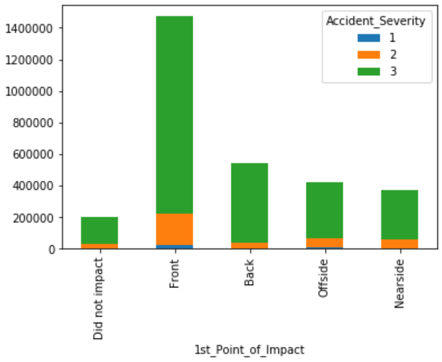

Figure 6 - Accident Severity and 1st point of impact

6. Accident Severity and Age band of driver

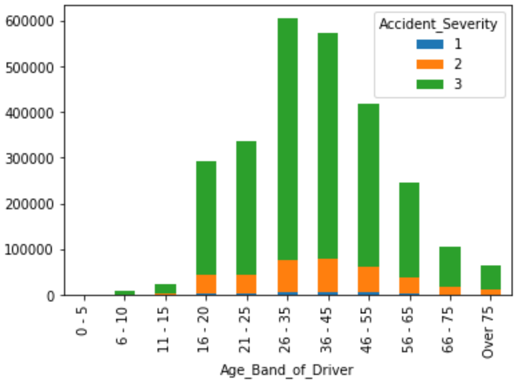

Figure 7 - Accident Severity and Age band of driver

From Figure 6, we can see that the majority of accidents happened when the driver is aged between 25-35. Despite being the major age band for the car accident, it is not the age band group with the highest serious accident severity.

7. Heatmap of the car accidents across the UK

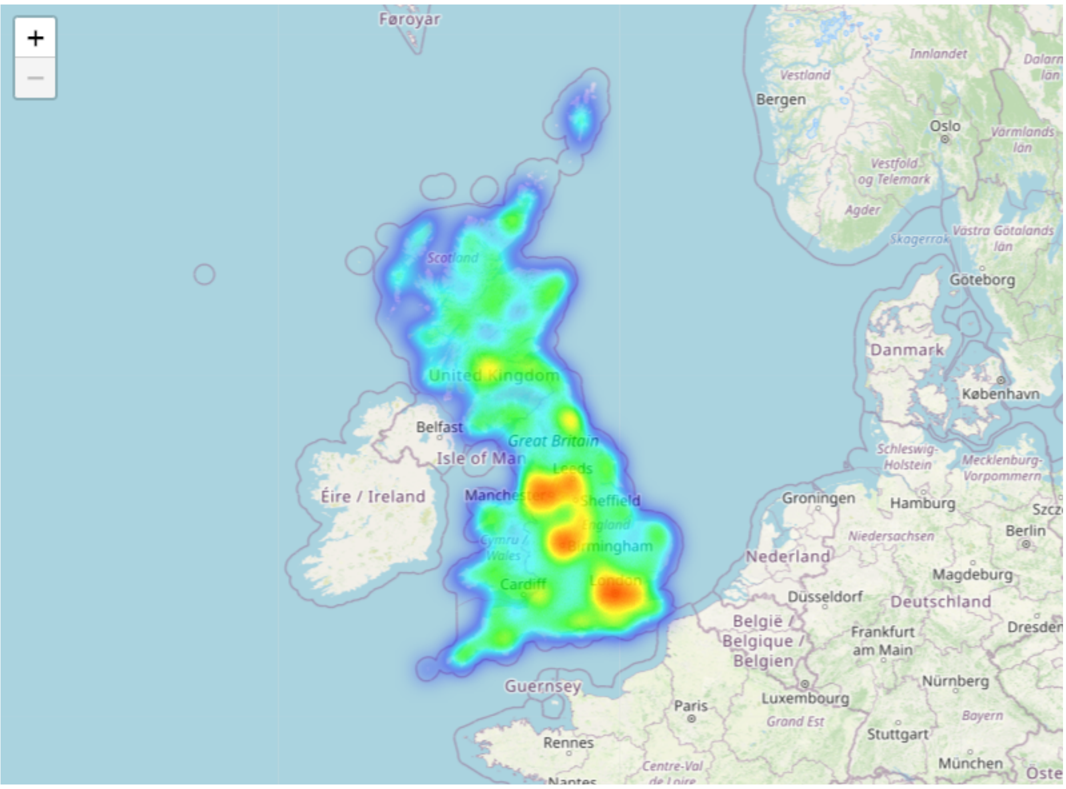

Figure 8 - Heat Map

From Figure, we can see that the majority of the car accidents happened in big cities. Therefore, polices in those area has to be prepared for car accidents.

16. Classification Modelling

K-Nearest Neighbours (KNN) model

a. Applying standard algorithms and their problem

KNN classification model classify the target variable based on its k nearest neighbours'' classification. It finds out the k nearest neighbour and their classification, and then define the classification of the target variable.

This method highly depends on the value k. If k-value is too small, it will lead to overfitting. It means that a larger k value is preferred. However, if k-value is too large, it makes the classification model too general.

Therefore, we have to find the best k-value in order to train the best model.

b. Solution to the problem

To remedy this situation, 10 k-values were used to find out the best model. Values 1 to 10 were used. Throughout the 10 k-values, the 9th k-value has the best performance.

Decision Tree Model

a. Information gain

Decision Tree Model finds out the decisions needed to classify the target variable based on the feature set. To find out the best decisions, it utilises the information gain to decide which feature to depend on.

b. Performance of the model

After training the model with the training set, the test set is used to test the accuracy of the model.

In order to test the performance of the model, the metric 'accuracy score' is used. For the Decision Tree Model here, the accuracy score is 0.858071.

Logistic Regression Model

a. Applying standard algorithms

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. In this case, since there are three outcomes for the target variables, we use multinomial logistic regression for this project.

b. Performance of the model

Three evaluation metrics are used to test the performance of this model, namely F1 score, Jaccard Similarity Score, and Log loss.

Jaccard Similarity Score	0.858061

F1 Score	0.792538
Log Loss	0.443054

17. Conclusions

After testing the above classification models, both Decision Model and Logistic Regression model stood out

From the above table, it is shown that the Decision Tree Model has the highest Jaccard Similarity Score while the Logistic Regression Model has the highest F1-score.

18. Future directions

Car Accident Severity is an important topic. Therefore, it is not enough to classify the severity based on given features. For future study, the predictive model can be built in order to predict the number of car accidents and its severity based on a set of features, so drivers can stay alert while police and ambulance can be well-prepared to face the accidents based on the prediction results. This can help handle car accidents efficiently and effectively.