Data Science - IBM Coursera Capstone Project

Author Name: Hue Dinh - Data Science Student

Car Accident Severity Analysis and Prediction with Logistic Regression

1. Introduction

Background

Cars accident is a crucial problem for all countries in the world. It annually makes over 26.000 people die or seriously injured in Great Britain alone (reported in 2018). It indeed brings along numerous social and personal consequences.

Car accidents casualties are various, controllable and uncontrollable. Therefore, an accurate analysis prediction model is necessary to help drivers and lawmakers to have awareness in particular situations so that severity will be minimised.

This project will focus on analysing accident data - The record of car accidents in the UK from 2005 to 2015, which include contributing factors to severity.

Problem

The data might includes bias information which is a part of natural society development. This project will focus on predicting the severity of car accidents in particular conditions, which can be aware of.

Interest

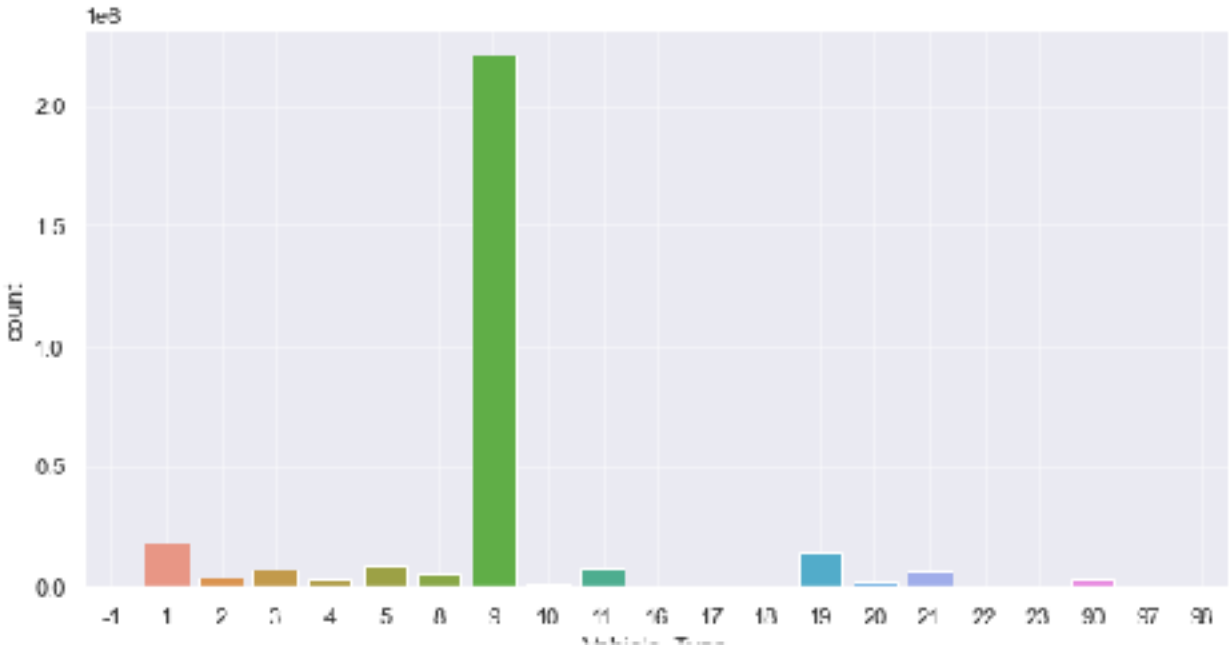
Local governments, especially police force will work much more effective if they know how severe a day can be. Appropriate actions will be taken, workforce will be more productive. Moreover, the application can be

developed to send driver appropriate alerts on a day which is dangerous for driving so that they can consider to take public transport instead.

2. Data acquisition

2.1 About the Dataset¶

Data set will be used in this analysis are Accidents and Vehicle from Road Safety Data, provided by UK government. They can be found **HERE** and fully explained **HERE** . They recored accidents and vehicles in those accidents, including acidental scenes, conditions at site, vehicle details, etc…



Vehicle Type distribution

Vehicle_type has 20 different attributes, however car (number 9), has the significant frequency. Since other vehicles have different set of contributing factors to the severity, and to make the model to be more more accurate, we will focus on only car accident.

The original dataset has 3 millions observations, after filter Vehicle type to car only, the new dataset has 2,2 millions observations, relatively large.

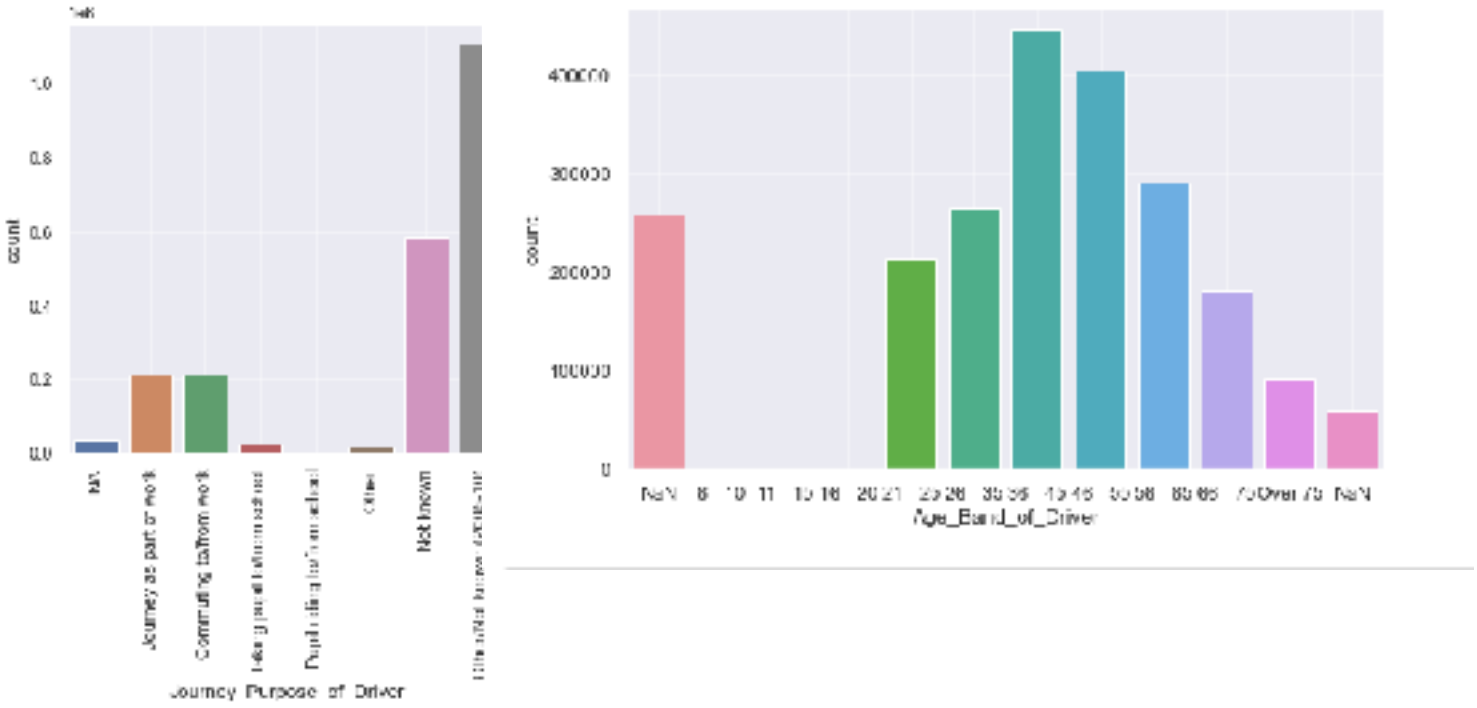
Each observation in the dataset is the record of one vehicle in an accident. Each accident may contain more than one vehicle, so the picture from the dataset, how each factor attribute to the severity is more clear.

The dataset is represent by Label-Encoding, which represent each category in each variable as a number. All missing values are represented by -1. As this is a large dataset, so it allows us to drop all missing values for analysis.

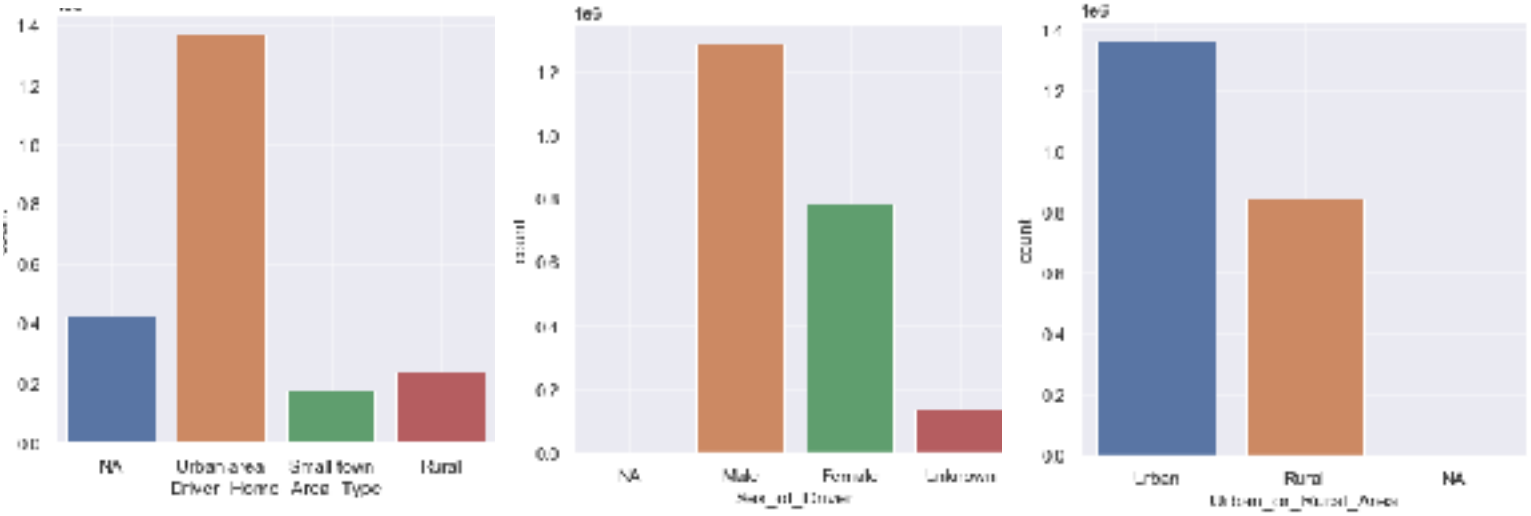
This dataset is contributed mostly by categorical variables, which is easy to be understood, as the record of accidents are based on check-list from the police.

2.2 Highly bias variables

Below are the frequency charts of some factors that normally get our first impression



Journey purpose of driver and Age band of driver distribution



Distribution of Driver Home Area Type, Sex of Driver, and Urban or Rural Area

These factors seem to have considerable participation in causing accidents. However, they are simply the natural result of social development. Such as people at the age of 30s and 40s are more likely to own and to use a car than the rest, urban areas are more density than rural areas, so more people drive, more accidents occurs. Therefore, we do not use these variables.

2.3 Choosing Variable for prediction.

As mentioned in the introduction section, and the application of the analysis, we will focus on factors that authorities - our user - can be aware of beforehand, so that they can take proper actions to limit the damages.
We will run the logistic regression model later so that choosing a small number of good variables will also avoid overfitting the model.

Dropping Features	Reason for dropping
`Towing_and_Articulation`, `Skidding_and_Overturning`, `Hit_Object_in_Carriageway`, `Vehicle_Leaving_Carriageway`, `Hit_Object_off_Carriageway`, `1st_Point_of_Impact`, `Number_of_Vehicles` , `Number_of_Casualties`	Consequences, only can be collected once an accident happens.

Dropping Features	Reason for dropping
`Was_Vehicle_Left_Hand_Drive?, Journey_Purpose_of_Driver, Sex_of_Driver, Age_of_Driver, Age_Band_of_Driver , Engine_Capacity_(CC) , Propulsion_Code, Age_of_Vehicle, Driver_IMD_Decile , Local_Authority_(District)`	Bias information, and the process of controlling these variables will be extremely complicated and bring along many consequences.
`Driver_Home_Area_Type, Location_Easting_OSGR , Location variables`	These variables lead to one factor: The more density an area is, the more car accidents happen, so we will not focusing on these factor as we can make a hypothesis that authorities of all area in the UK understand their location and society well to have an appropriate awareness on car accident severity.

Therefore, we will be going to do further analysis with these variables.

Vehicle_Manoeuvre, Junction_Location, Accident_Severity, Day_of_Week ,
Junction_Detail, Junction_Control,
Light_Conditions,
Weather_Conditions,
Road_Surface_Conditions, Special_Conditions_at_Site

3. Exploretory Analysis

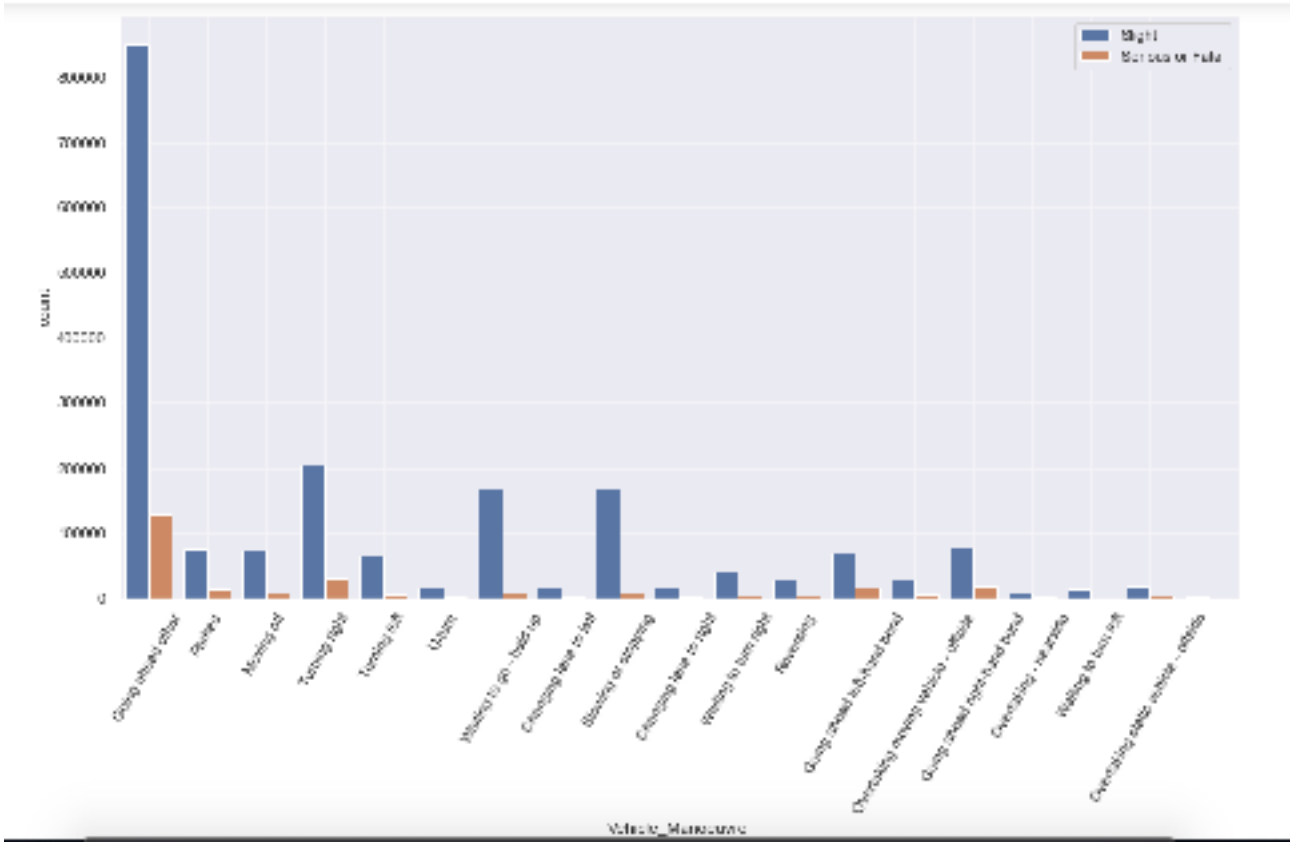
This section focuses on clarifying relationship between accident frequency and selected features.

3.1 Vehicle_Manoeuvre

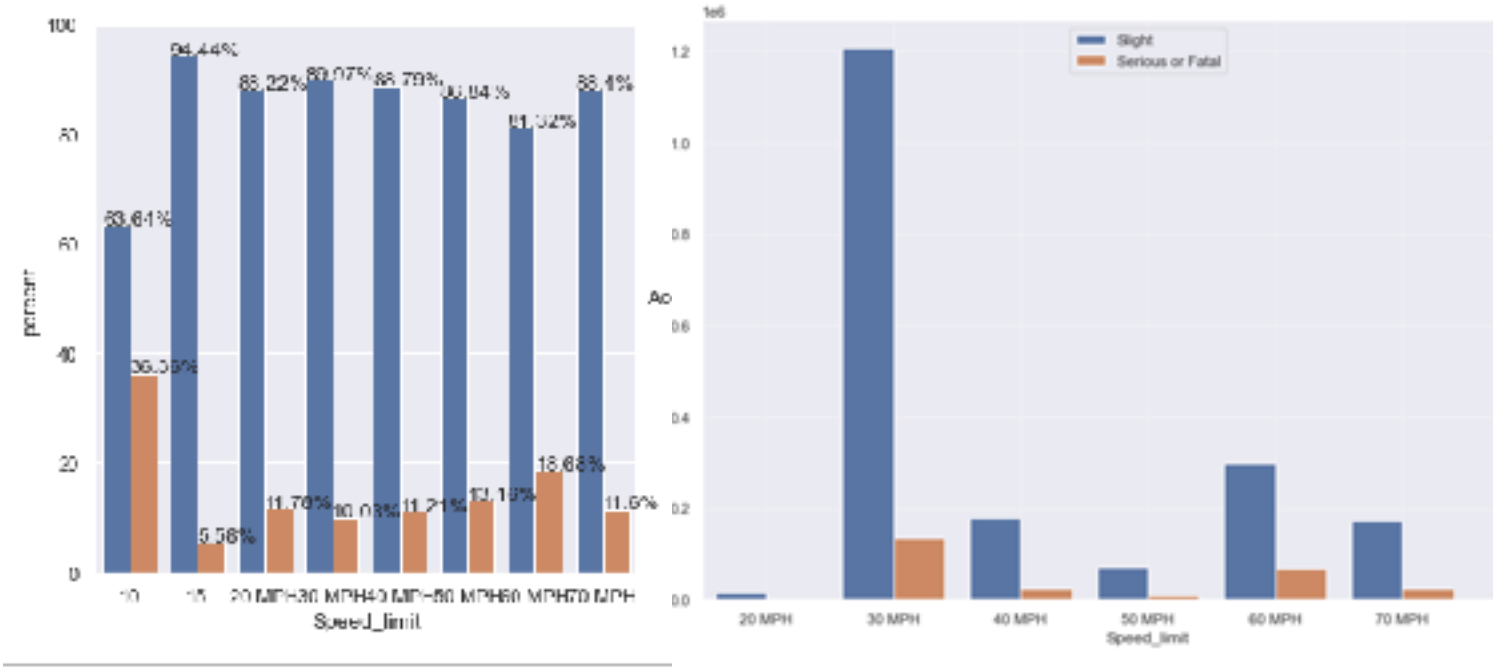
The dataset shows that manoeuvres have the most frequency of occurring an accident are going ahead of others and changing their directions. This factor is indeed difficult to be managed, however, if camera system was installed in intersections, watching the traffic from above will give the police officer a better view to predicting which vehicle is in a sever movement.

3.2 Speed limit

At each threshold of speed limit gives us a different understanding of this relationship.
Recall the causations of an accident: either vehicles crash to the other, or crash to an object. The data set shows us, more vehicles involve in accidents in the lower speed limit road because the traffic is more density in these. At higher speed limit ones, accidents seem to be more severe.

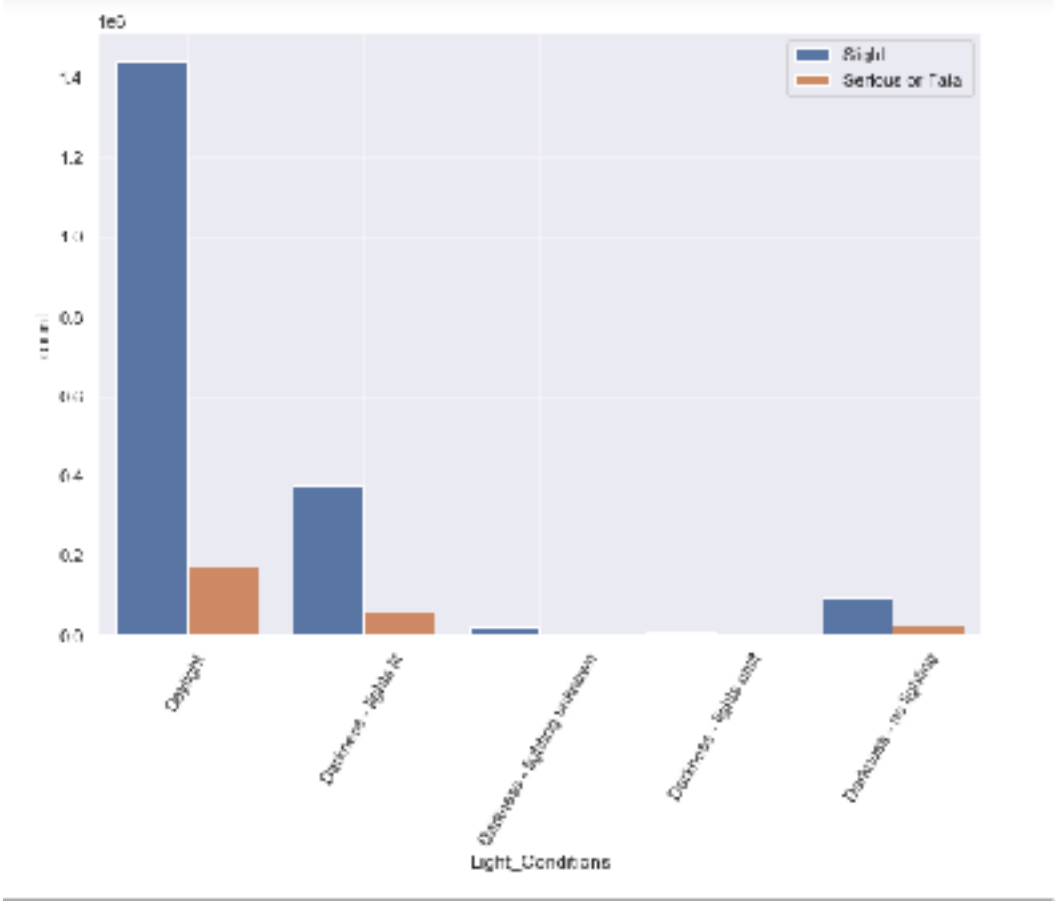


Vehicle_Manoeuvre



Proportion of accident at each speed threshold and counting observations

3.3 Light conditions

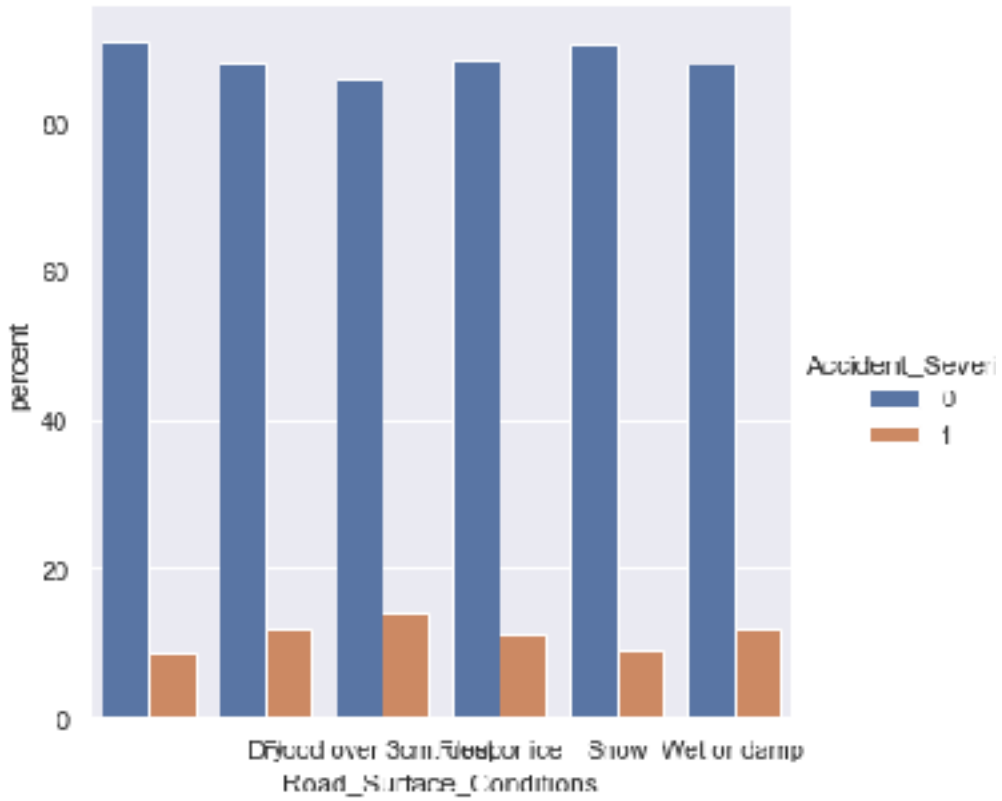


Accident counts at light conditions

Daylight has the most accidents involve due to the normal demands of joining traffic in the day time. However, in the darkness conditions, even the density of traffic is low, the frequency of accidents is considerable. Which suggests authorities to check the lighting system in their local areas.

3.4. Road surface condition.

Rain is a norm in the UK, that makes the road surface is often to be wet. However, an accident happens on the wet surface road tends to be more severe than when it is dry. Therefore, appropriate alert for the were weather condition is necessary.



Proportion of accidents at road conditions.

4. Predictive Modelling

The chosen model for this application is pretty straight forward as our outcome are boolean, classification model.

The dataset recorded vehicles already encountered accidents. That means, for the default of this application, we should always aware of accidents can be happened at any time. This model is to predict if outside factors could make consequences more dangerous for vehicles so that authorities will have appropriate action to minimise the severity.
Our chosen factors for analysing are all categorical, so I choose the logistic regression model for prediction.

The reason behind our problems is drivers are not aware enough about outside conditions, notice boards are not visible enough or weather and traffic conditions change from region to region…

Notice board and police force have been working hard to minimise car accident severity. However, it would be better if they have a system to tell them if a day, in a particular location, needs more attention from them, or not that much. The whole system will work much more effective.

Logistic Regression is a variation of Linear Regression, useful when the observed dependent variable, y, is categorical. It produces a formula that predicts the probability of the class label as a function of the independent variables.

After running the model on our data set, we get the report:

```
print (classification_report(y_test, preds))
```

	precision	recall	f1-score	support
0	0.89	0.99	0.94	741951
1	0.30	0.04	0.07	90318
accuracy			0.89	832269
macro avg	0.60	0.52	0.51	832269
weighted avg	0.83	0.89	0.84	832269

The F1 score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0. It is a good way to show that a classifier has a good value for both recall and precision.
And finally, we can tell the average accuracy for this classifier is the average of the F1-score for both labels, which is 0.84 in our case.

5. Discussion

As mentioned earlier, car accidents have many factors of causation, many of them starting from subjective factors of drivers, which is out of the authorities control, and even drivers control sometimes. However, from the authorities perspective, they want to try the best as they can to limit the consequences.

Even the model is highly accurate in prediction, we still need to work on the operation parts to arrange workforce and requirements.

This model heavily based on the hypothesis that the local authorities know well about their road accident severity problem, and each of them has a certain level of resources which fits along with the local social situation.

When developing the model, I found out the optimal threshold is 0.62 (0.5 is the default of regression model), when I moved the threshold of the model to 0.62, the model had better accuracy.

6. Conclusion

With the high accuracy of this model, it can be applied in real-life situations at local authorities in controlling traffic. In particular, they can base on the weather forecast, which is highly accurate nowadays, and local traffic conditions to:

- Upgrade traffic Infrastructure if it's allowed
- Distribute workforce to 'hotspot' areas, day and time.
- Have appropriate signs to raise awareness for drivers in high-risk conditions
- And many more

This analysis has focused on factors which can implement easily in daily basic, of any local authority. It also avoids bias information which is the nature of social development.
Not only authorities but also people who join the traffic can implement this model to reduce the severity caused by car accidents.