How severe could a Car-Crash be in the UK ?

Introduction

Background -
Road accidents happen everyday all over the world, according to the WHO, road traffic injuries caused an estimated 1.35 million deaths worldwide in the year 2016 alone. It can seem like they are an old topic and yet with the advancement capabilities of cars and their technology, especially self-driving technology, it is even more necessary to have tools and measures to prevent them.

Problem -
Predicting the severity of a car crash is no easy task. And even when possible, precision levels will vary significantly depending on, among many factors, the data available and how well the problem has been modeled. But here using the data available to public from the UK government website and supervised machine learning methods, we try to predict an accident severity, given date, time, weather, light, road conditions, ezc.

Stakeholders
It would definitely help the Development Authority to make any necessary changes to prevent road accidents. Local Police force and first-responders could also be assisted with this data in advance. Even civilians themselves could be armed with this knowledge to make travel-plans in a much informed manner.

Data

Data Source -
The data come from the **Open Data** website of the UK government, where they have been published by the Department of Transport.

The dataset comprises of two csv files:

- Accident_Information.csv: every line in the file represents a unique traffic accident (identified by the Accident_Index column), featuring various properties related to the accident as columns. Date range: 2005-2017
- Vehicle_Information.csv: every line in the file represents the involvement of a unique vehicle in a unique traffic accident, featuring various vehicle and passenger properties as columns. Date range: 2004-2016

The two above-mentioned files/datasets can be linked through the unique traffic accident identifier (Accident_Index column).

Data PreProcessing -
This dataset covers a wider date range of events. Most of the coded data variables have been transformed to textual strings using relevant lookup tables, enabling more efficient and "human-readable" analysis. It features detailed information about the vehicles involved in

the accidents. While the first dataset contains 34 columns/attributes, the second dataset contains 24 of them. But before merging the datasets a number of observations about missing data was made and much of necessary data was formatted.

Based on initial intuition and an educated guess, many columns in the accidents-dataset like ‘InScotland’, which do not seems to contribute to good prediction, were dropped. Then about 0.5% of the records which had missing valued were dropped too. It seems important that time-of-day seemed to play an import role and so the ‘time’ attribute for was divided to 5 time-periods and numerical values were assigned to each. After formatting the ‘date’ attribute , vehicle-dataset, with chosen attributes only, was merged with formatted dataset, with the Accident_Index as the guiding column, to give final dataset, where features to needed to be formatted – categorical to numercal.

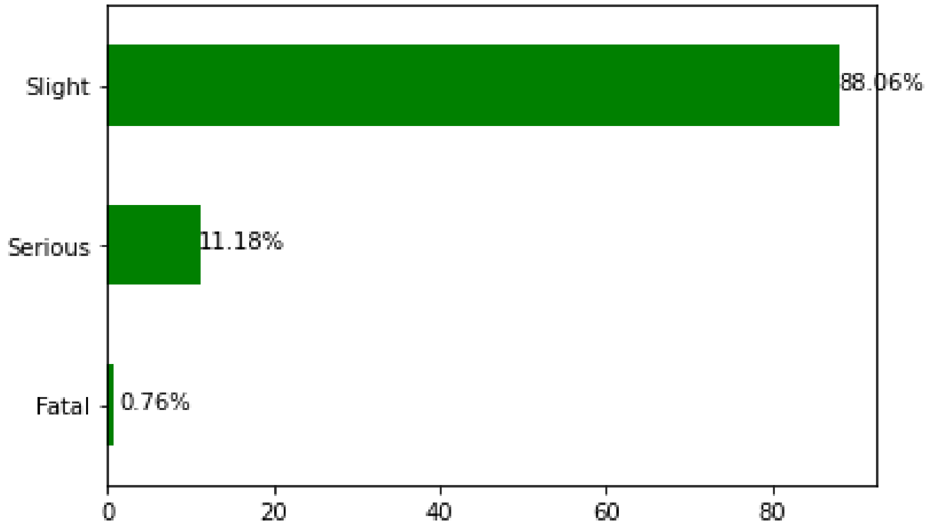
Features selected -
The features for model building were selected carefully based on data exploration and analysis:

(From Accidents_Information dataset)Time_of_Day, Road_Surface_Condition, Urban_or_Rural_Area, Light_Condition, Speed_Limit, Weather_Conditions

(From Vehicle_Information dataset) Age_of_Vehicle, Age_Band_of_Driver, Vehicle_Manoeuvre.

Exploratory Data Analysis

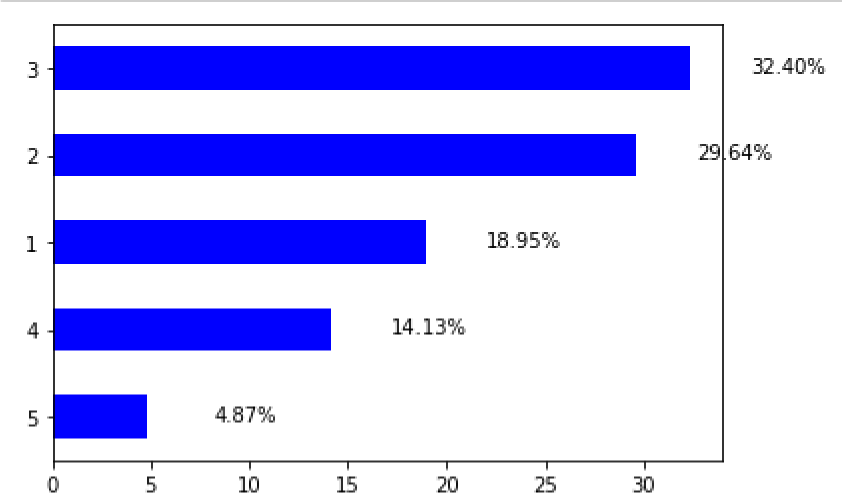
We finally have features-set with 968614 records with 60 columns/features. After understanding what each feature represents, we need to explore their respective values to understand their distribution and if there are any inconsistencies from their definition.
In any case, the most important thing that we have learned from this step is that there is a huge imbalance of classes for *our* target variable:



Given that ‘Slight’-severe crashes are the most common ones and that as the severity of the crash increases, so does the importance of making a good prediction; we need to address this issue if we hope to produce a good operational model. To that end, we will resort to under and oversampling techniques, which are described in the Predictive Model section.

Relationship between number of accidents and time of day -

It can be intuitively said that the time of day definitely decides the probability of an accident. Looking at the relation :



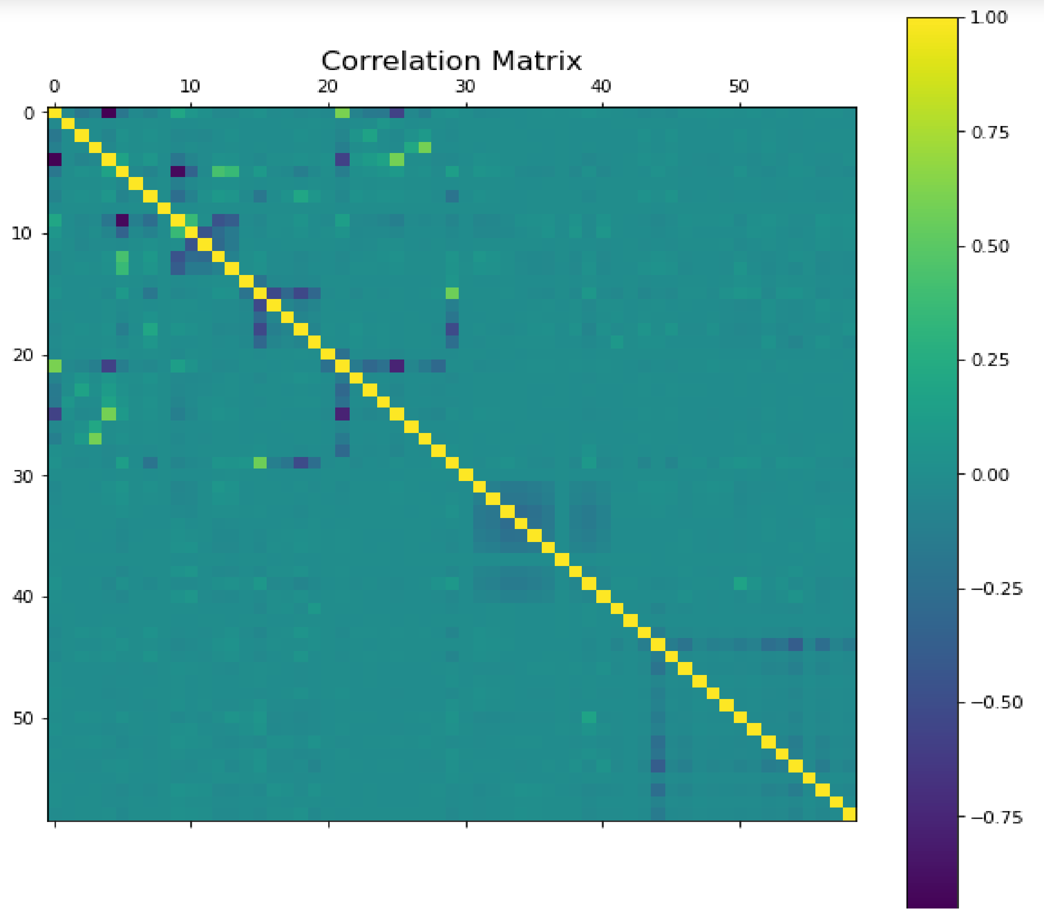
The graph

seems to match one's intuition, with highest number of accident cases occurring during the Afternoon-Rush hour between 15:00 and 19:00. Also it appears that 'Daytime' clearly would be a very good feature to predict using classification models.

Features correlation-

We need to check for possible correlations between a pair of features. We take the p-value for the test of each pair of features and consider a 0.05 threshold on the Null hypothesis for independence.

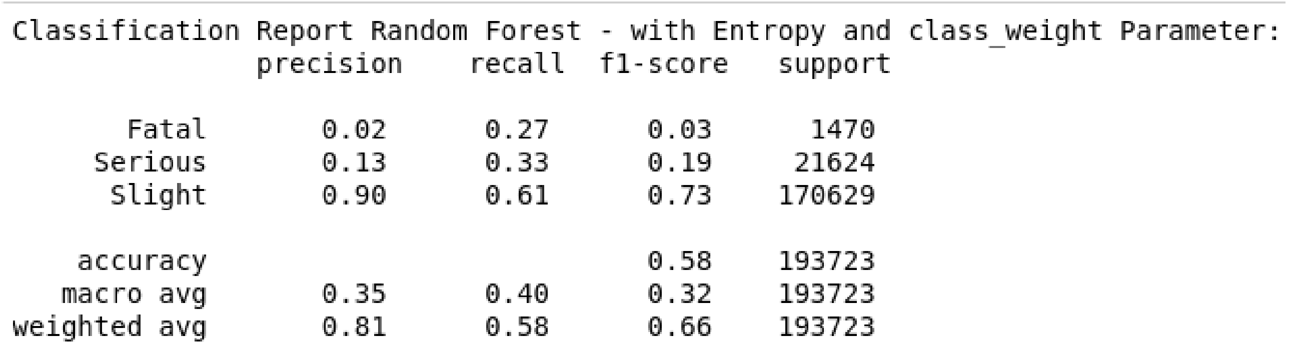
The results are summarized in the picture below:



We can see that every feature is dependent on almost every other feature. Making it virtually impossible to remove any of them due to correlation, since they all carry valuable information. As a result, we decide to keep all the remaining features.

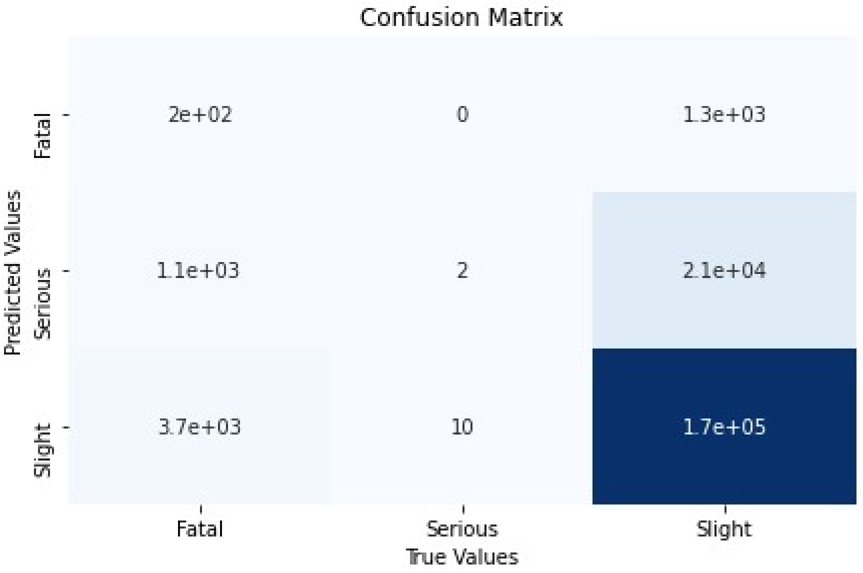
Predictive Model

Since all of the features in the dataset are categorical and also, it is not clear which ones are the most meaningful or relevant to the classification task. Therefore, we will focus first on decision tress as a and we will train and optimize a **Random Forest** classifier. The Random Forest Classifier we will use now is a popular classification algorithm and includes a class_weight parameter, which allows us to have the algorithm adjust for imbalanced classes.

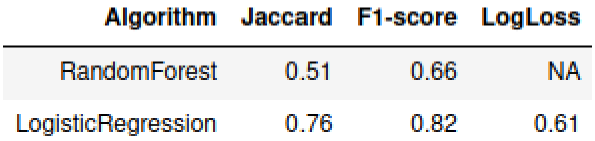


```
Classification Report Random Forest - with Entropy and class_weight Parameter:
               precision    recall  f1-score   support

       Fatal       0.02      0.27      0.03      1470
     Serious       0.13      0.33      0.19     21624
      Slight       0.90      0.61      0.73    170629

    accuracy                           0.58    193723
   macro avg       0.35      0.40      0.32    193723
weighted avg       0.81      0.58      0.66    193723
```

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.
Based on the predicted values -



Accuracy Report-



Algorithm	Jaccard	F1-score	LogLoss
RandomForest	0.51	0.66	NA
LogisticRegression	0.76	0.82	0.61

Results and Conclusion

The Logistics Regression seems perform well in comparison to the Random Forest classifier, which was unexpected initially. The RF is able to produce a mere 0.51accuracy and 0.66 F1 score. While the F1 score is balanced across classes –meaning that the F1 score for each class has a similar value to the averaged one. This makes our model relevant, yet barely usable.

Nevertheless, there is much room for improvements. For example, a few other algorithms that could prove useful considering the shape and form of the dataset are SVM, XGBoost, and LGMB. These could also be combined –including the Random Forest.

Possible Improvements-

We could also train a binary classifier for each class to have a more customized one-vs-all approach. And also try other evaluation methods like precision/recall and ROC curves. In this case, we could implement a voter for the final classifier. And we could also leverage a deeper analysis of predicted probabilities for each class. Helping understand better what are the most representative characteristics of each class.

Also, for the imbalanced nature of the dataset, some cost sensitivity algorithm could be very helpful.

Recommendations-

Even though our model may not have been very good, yet after assessing the data and the output of the models, a few recommendations can be made for the stakeholders. The Pubic Development Authority of UK could assess how much of these accidents have occurred in a place where road or light conditions were not ideal for that specific area and could launch development projects for those areas where most severe accidents take place in order to minimize the effects of these two factors. Whereas, the car drivers could also use this data to assess and make travel-plans carefully on the road under the given circumstances of light, road-surface and weather-conditions, in order to avoid a severe accident.

-o O-o-