PREDICTING THE SEVERITY OF ROAD CRASHES BASED ON ENVIRONMENTAL, ROAD AND OTHER CONDITIONS

by

Musab Faiyazuddin

for the Capstone project in IBM's Data Science Professional Certificate

Table of Contents

Introduction

Traffic accidents are a major cause of death globally. More than 38K people die every year in crashes in US roads. The fatality rate is 12.4 deaths per 100,00 inhabitants[1]. If your home is metropolis city chances are that you have heard, witnessed, or even involved in one. If we can predict the traffic accidents or their intensities, it can potentially save many lives. Even though some of the accidents happen because of the careless nature of the people involved, which would be impossible to predict. But most of the remaining accidents are influenced by many quantifiable factors like weather conditions, car types, driving speeds, road structures and many others.

Imagine the possibilities where the predicted high risk for accidents environment can be used. For instance, the areas where under a certain condition is expected to have high risk of accidents the emergency services can be in close range to that to reduce the response time and which may be potentially save lives. Similarly, the highway patrols can impart some mechanism during certain conditions to reduce the speed of the passing vehicles which can very likely reduce the risk of the accident to happen in the first place. Another situation where this knowledge could be used is the department of roads construction. They can study more into road nature and how improving the way its build so that it can adapt to weather change with no added risk factor for the people commuting on it.

In light of all the mentioned instances and many more possibilities where this predictive analysis can be very crucial understanding what factors are having the most influence on the severity of the accident and how we could possibly save lives eventually.

Given a dataset of all those quantifiable features and the intensity of the accidents can potentially help in building a predictive model. Fortunately, several such public datasets are available, like the one being used here.

[1] Based on analysis of data from US Department of Transportation in 2019

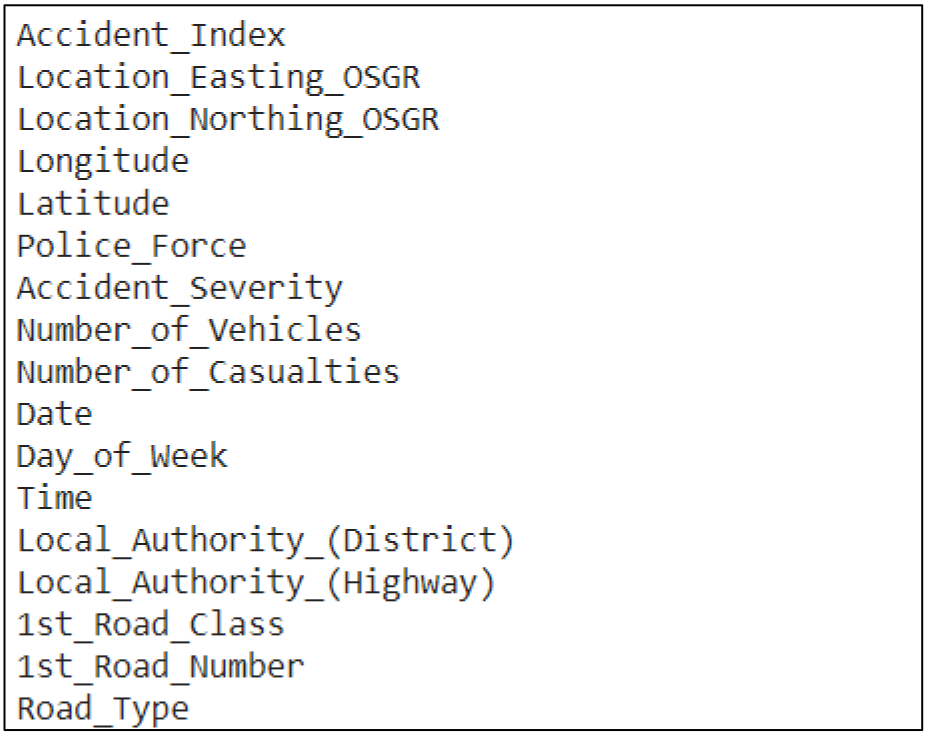
Data Description

The Open dataset that I will be using for this exercise is from UK Department of Transport from Kaggle. I have chosen this because it is an amassed traffic data for a total of six years with many factoring columns which is always good. The other weighing factor was also because the same dataset has been used by Data scientists before for different activities, as it has all necessary properties enabling us to model using it.

The shape of the dataset at the start was 1504150 rows and 33 columns, that is big enough for our needs. As we will be working on predictive analysis of the severity of the accidents, we need to know if the dataset is balanced between all categories. It turns out its not, but we can apply some methodologies to overcome this setback[2].

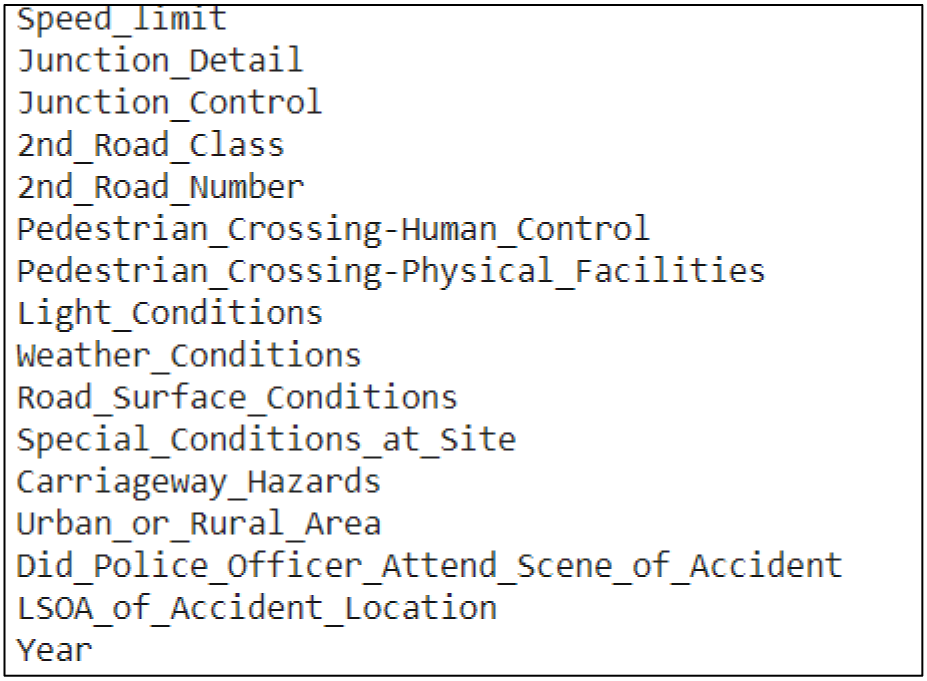
There are 33 columns to start with, but we would not be using all of those as they might not be significant enough. We will be deciding that on the basis on some statistical and distributive aspects of the data which will be discussed in detail in later sections of the report.

Also, it is observed that the data has some Null/Unknown values which will be treated or removed based on what is more fitting for us. The following is the list of columns we have initially in our dataset,



```
Accident_Index
Location_Easting_OSGR
Location_Northing_OSGR
Longitude
Latitude
Police_Force
Accident_Severity
Number_of_Vehicles
Number_of_Casualties
Date
Day_of_Week
Time
Local_Authority_(District)
Local_Authority_(Highway)
1st_Road_Class
1st_Road_Number
Road_Type
```

[2] To be discussed in the methodology section



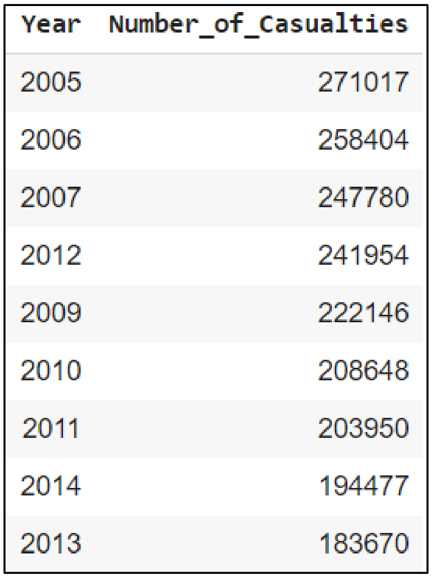
```
Speed_limit
Junction_Detail
Junction_Control
2nd_Road_Class
2nd_Road_Number
Pedestrian_Crossing-Human_Control
Pedestrian_Crossing-Physical_Facilities
Light_Conditions
Weather_Conditions
Road_Surface_Conditions
Special_Conditions_at_Site
Carriageway_Hazards
Urban_or_Rural_Area
Did_Police_Officer_Attend_Scene_of_Accident
LSOA_of_Accident_Location
Year
```

Based on the columns and the overall size of the dataset, it looks like a great dictionary for our use-case needs right now. The naming convention of the columns are all self-explanatory which makes it easier to work with it and not having to refer to the definition file of the dataset.

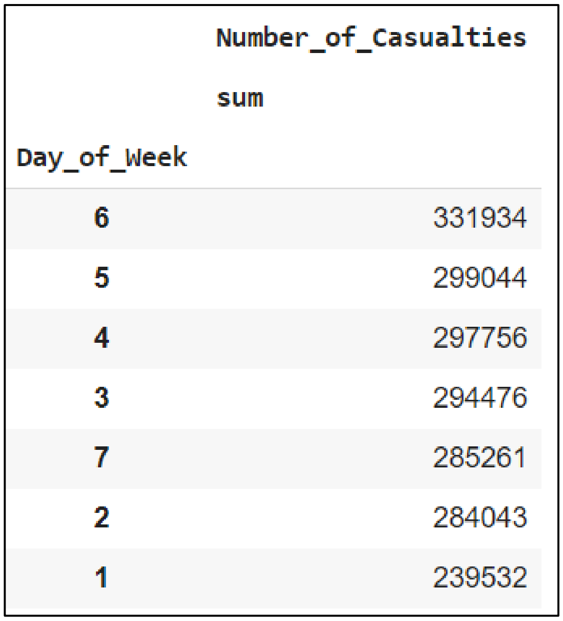
Methodology

After examining the distribution of the target variable, the first assumption was to adapt a Machine Learning model that would go well with an unbalanced data, like XGBoost Classifier. During the Exploratory Data Analysis (EDA some very interesting discoveries about the data were made. Seeing the distribution of accidents when grouped by Year, it was seen that the number of casualties occurring each year had been declining every year. Not significantly but there has been decrease, indicating there may have been some improvements made in the departments that had been the main cause for accidents to happen, like speed limits at some places may be. Learning from the historic data and making changes to the future is the core idea behind Analytics and Machine Learning. So there seems to be a good sign that we have been continuously trying to learn and use the recorded data to improve our futures.

The other observation that was made was to project the same distribution on day of the week instead of the year, something obvious yet very notable was to see there were significantly more fatalities towards the end of the week in comparison to the start of it. The day of the week will be highly correlated with the amount of traffic flow as well.



Year	Number_of_Casualties
2005	271017
2006	258404
2007	247780
2012	241954
2009	222146
2010	208648
2011	203950
2014	194477
2013	183670

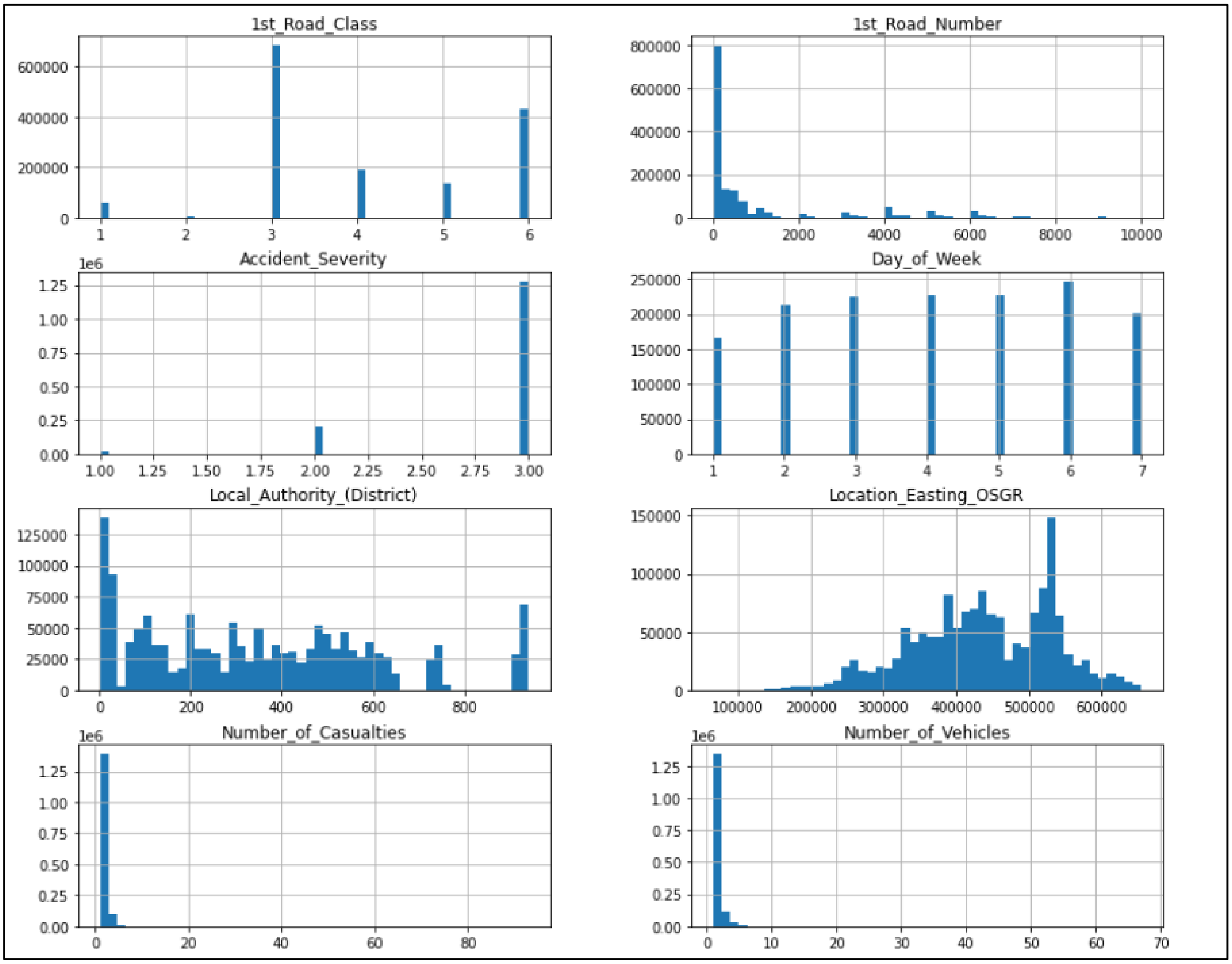


	Number_of_Casualties
	sum
Day_of_Week	
6	331934
5	299044
4	297756
3	294476
7	285261
2	284043
1	239532

After seeing such distributive insights, the task was to identify the useful columns among the 33 and if more features could be engineered using any of those features. What was used to complete this task is to first see the distribution of variables, so that if any of those are too skewed and mostly identical could be dropped as that would not have any impact in the model. The second step to that was see the correlation between each of those and the target variable (i.e, Accident Severity) and any variables with very small values that are not having effect will also be removed.

The latitude and Longitude value columns were scatter plotted using an overlay of UK map to see if there were regions which accumulated more of fatal accidents. It was observed as the data is of 6 years the spread was all over the map except a few spots towards North. But, as the correlation value of Latitude/Longitude with the target was not very poor, the assumption that some points on the map must be more high-risk zones than others could stand true.

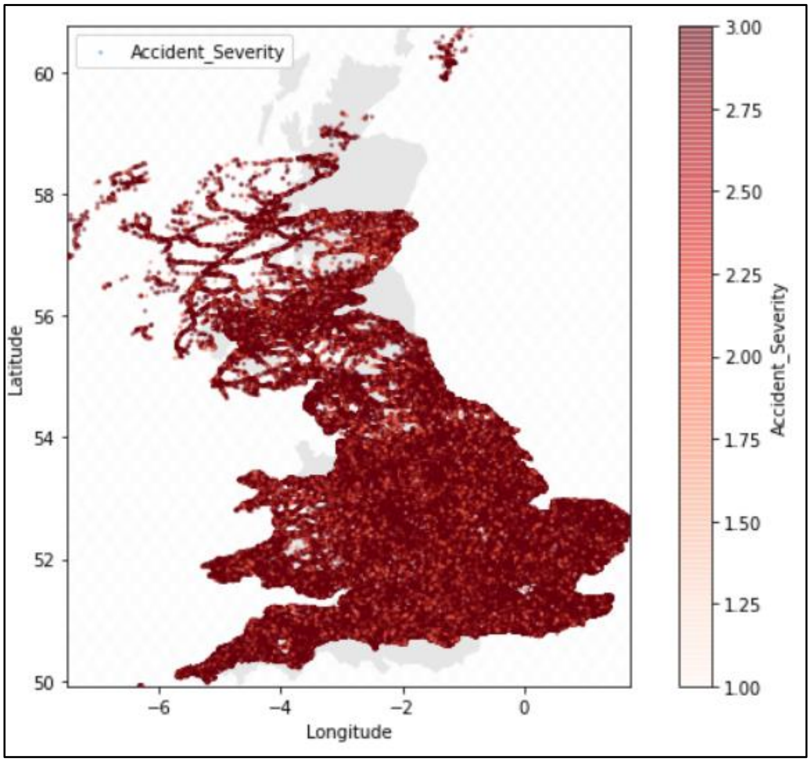
During the Data preparation stages, the null/missing values within the data and the uniformity of format of the data was corrected that is acceptable by the Machine Learning model. As the total size of the data was quite big, correspondingly the number of missing data was less, dropping all such rows was preferred instead of imputing them to maintain originality of the dataset. As for the data type format, for all the categorical string type columns where encoded using LabelEncoder as ML models do not accept string inputs. Similarly, the date column was reduced to a number format. At the end of this step we were left with all columns in a numeric format (Float/Int).



Choosing a right Machine Learning model was challenging. In the first attempt what I considered the most is to use a model that can accurately classify without creating the bias of unbalanced data. For that needs, the XGB Classifier seemed like the best option for classification.

After the data was split into test and train with a ratio 3:7 respectively the model was trained and tested. An accuracy of about 88% was achieved, but to our disappointment the confusion matrix proved that the model was in fact biased towards one class.

To fix the unbalanced class data, a random_sampler was used, and a technique called Over Sampling was adapted to get a balanced dataset. When the same model was trained and tested on a balanced data the accuracy of XGB was not satisfying enough then the hunt for a better classifier continued again. After some research and try outs turned out Random Forest Classifier was the best performing Ensemble model providing us with a good overall accuracy.

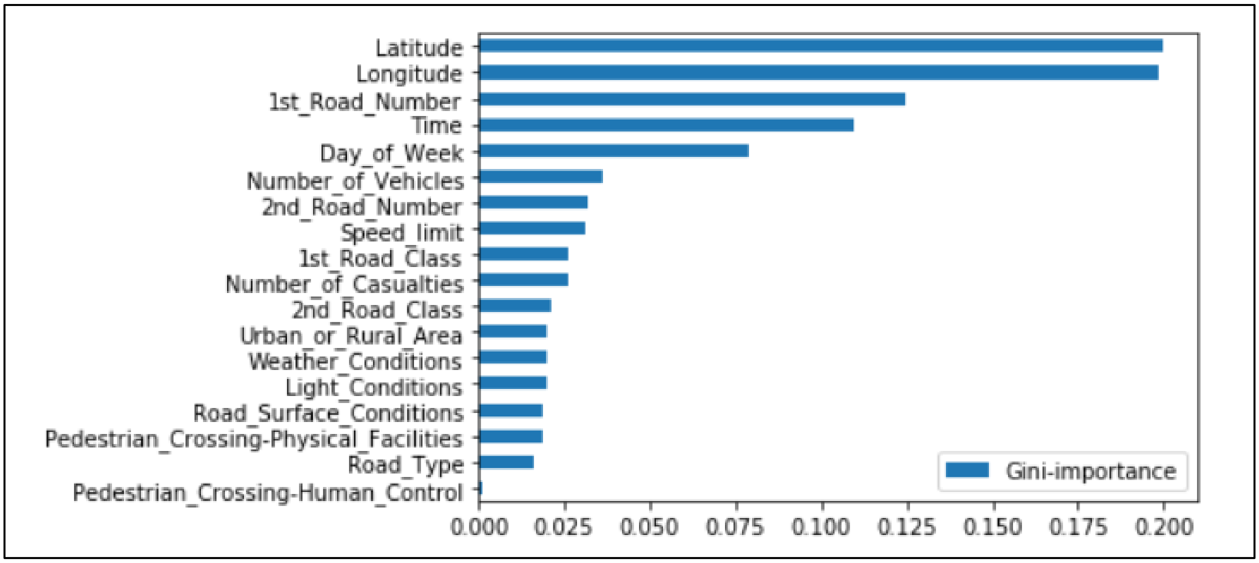


To further enhance the results of the classifier some hyper-parameter tuning techniques were also used, like GridSearch. Which provided with the best parameters for the model. Using those new hyper-parameters, the model was again trained and tested improving its overall accuracy.

Results

The final RandomForest classifier was classifying the accident severity (3 classes) with an accuracy of 84%. The confusion matrix also showed that though there has been misclassification overall, but it was not biased towards one class.

Observing the accuracy, the proportion of misclassification and the low feature importance it will be sound if more data engineering is performed on the data and repeat the modelling again. That should potentially increase the accuracy and decrease misclassifications.



Discussion Section

The idea behind adopting Random Forest as a classifier was also because it does not require much preprocessing of data. So, even if the data were to lack the quality the Random Forest classifier would adapt to that.

After we have a model now that can predict how a severe an accident based on some conditions, or how severe could an accident happen with a certain condition. This can now be used in Applications and as an input while constructing and designing roads.

Conclusion Section

Even though the size of features was less here the accuracy achieved in the end was pretty good. This exercise here can be a steppingstone to perform more research on the factors leading to an accident and then perform data gathering activities. Machine learning models can be retrained anytime a new attribute is discovered that improves the predicting ability.
There has been continuous research work being done in the department of Transport as a result we have observed in the exploratory data analysis that there has been decline in the number of yearly fatalities in road accidents. This gives us an encouragement that Machine Learning can solve real-life problems and there is scope for improvement always.