- Data Link: https://www.kaggle.com/tsiaras/uk-road-safety-accidents-and-vehicles

2 Data

Traffic accidents are extremely common, and it is more frequent in the sprawling metropolis. Because of their frequency, traffic accidents are a major cause of death globally, cutting short millions of lives per year. By 2017, the total number of cars in the UK is 31,200,182, indicating that per 1000 people have 471 motor vehicles, rank 35th in 160 countries. However, road accidents happen every day. In 2018, there were a total of 160,597 casualties of all severities in road traffic crashes.

We will use various analysis techniques and build models to predict accident severity, including Logistic Regression, Random Forest, Linear SVC, KNN, Decision Tree. By using predictive analysis and comparing the models, we can have a better understanding of different variables like what kind of road condition, vehicle condition and what factors are involved that contribute to road accidents.

The dataset on road traffic accident analysis has about 2 million rows and 63 columns. The predictor variables in the dataset had information geographical locations, weather conditions, type of vehicle, number of casualties and vehicle manoeuvres. The dataset had missing values which were dropped.

During the exploratory data analysis part, I found some compelling details like the age group between 26 years to 35 years was more frequent with accidents than any other age group. We also explored that most of the accidents happened on Friday. We were surprised to see that weather conditions did not play much role at the time of accident. Most of the accidents happened during the afternoon rush that is between 3.00 pm to 7.00 pm. This data exploration gave us clear insights that road conditions, rush hours, weekdays, age group, age of vehicle, and junction details played a major role in road accidents.

In order to increase the practicability of the dataset, extensive feature engineering for all the predictor variables was performed. All the categorical variables were recorded and interaction terms between two predictor variables was also performed i.e. age of predictor variable and speed limit. Following this, Z -score standardization for all the numeric variables was applied so that the scale of any numeric variable does not influence the predictions during the modeling.

I classified our multiclass target variable into 2 classes i.e. slight as ’0’ and Fatal or severe as ‘1’. The original dataset had only 13% rows for severe or Fatal accident severity, this was a highly unbalanced dataset. In order to balance it for the modeling, I used a sampling technique to balance out the Slight and Severe or Fatal accident severity. After under-sampling technique, I had a ratio of 1:1 for slight and severe or Fatal accident severity which was used for modeling and training the data.

Model Description

The final modeling dataset of 381K rows and 62 columns was split into training data with 70% and test data with 30%. Spot-checking technique to determine the machine learning model was implemented which was best suited to predict the accident severity. Machine learning models which I used are Logistic Regression, Random Forest Classifier, K Neighbors Classifier, Gaussian NB, Perceptron, SGD Classifier, Decision Tree Classifier, Gradient Boosting Classifier, Linear Discriminant Analysis, Extra Trees Classifier, and

Bagging Classifier. Based on the accuracy, from 11 different models the dataset was narrowed down to 6 best models whose accuracy is more than 60%. The dataset was too huge to do the hyperparameter tuning and gridsearchcv, so we subset the data with the accidents that happened in 2016. Hyperparameter was used for tuning to modify parameters in all the 6 models which are Logistic Regression, Random Forest classifier, Gradient Boosting Classifier, Linear Discriminant analysis, Extra Trees classifier and Bagging Classifier to improve the performance of the models.

Further permutation testing was implemented for feature importance for the best model i.e. Gradient Boosting. Here it was observed that the features like Engine capacity, Number of vehicles, speed limit, age group of drivers are some of the features which were believed to be important in making predictions. While it was also seen that features like pedestrian crossing human control, vehicle location restricted lane, road surface condition code, weather condition code, were comparatively less important features.