Road Accident in UK (2018)- Accident Severity Prediction

1. Introduction

Millions of people die due to accident every year. About 1.35 million people died in road accident in the year 2016. Furthermore, there are millions of people who have suffered serious injury due to road accident causing long term health issue. It is currently eighth highest cause of death and WHO has predicted it to be seventh highest cause of death by the year 2030.

This project is performed to predict the road accident and its severity of the United Kingdom in the year 2018. Analyzing different features such as location, weather condition, Light condition, age, sex, etc., we can predict the severity of the accident. These information thus can be used to alert ourselves from the accident prone areas, help the authorities to take necessary actions to decrease the severity of the accident.

2. Data

The data for this project is collect from https://data.gov.uk . Here, all the information related to road accident in the United Kingdom is found. The data is based on the police report, so, it will only contain the officially recorded road accidents. The data of the year 2018 is taken for the analysis. Initially, data of accidents, vehicles and causalities were taken and merged for the analysis. After merging there we 69 different features of the accidents. Some of the major features were location, weather conditions, light conditions, age, sex, accident, severity, etc. some categorical datasets were given numbers like weather conditions had numbers from 1 to 9 which each number giving different conditions. Similarly severity 1 meant fata, 2 meant serious and 3 meant slight. This information can be found on http://data.dft.gov.uk/road-accidents-safety-data/variable%20lookup.xls.

The merged data was cleaned for further analysis. Any missing value on the data was dropped. The Location_Easting_OSGR, Location_Northing_OSGR, Longitude and Latitude had 120 missing values, for which the rows of missing values were dropped. LSOA_of_Accident_Location had 15762 missing values, for which entire row was dropped for the analysis.

3. Methodology

3.1 Exploratory Data Analysis

At first the analysis was performed to check the accident based on hour, month and day of a week. Bar chart was plotted for each case.

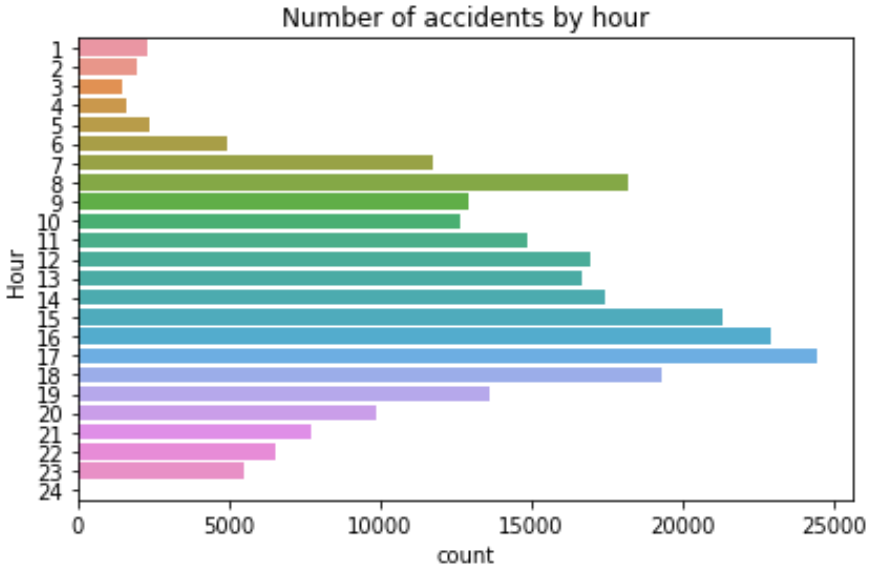

Figure 1 number of accident by hour

The graph shows that the accidents was low during morning and night time while it was most during 1500 to1800 hours.

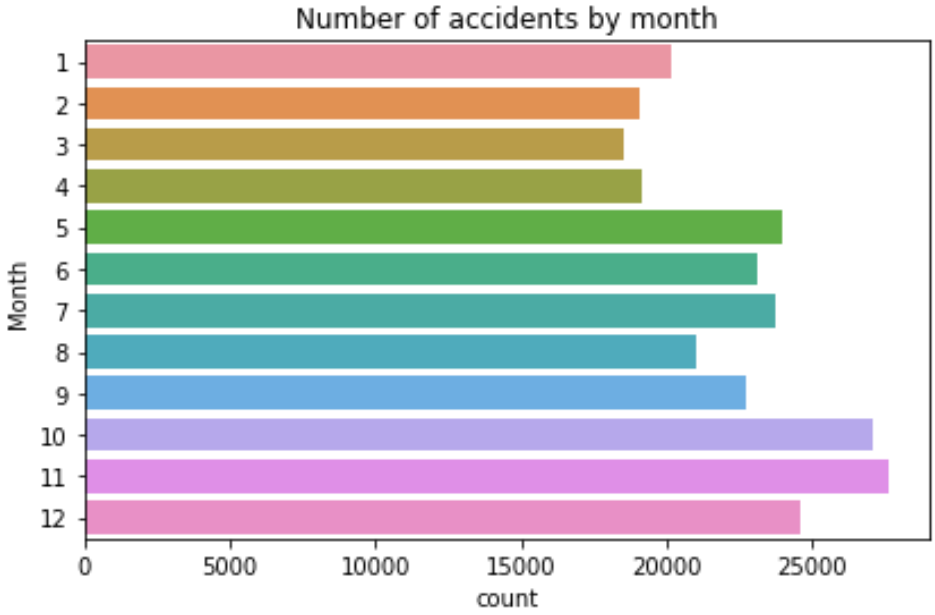

Figure 2 number of accidents by month

The month data showed nearly similar count of accidents throughout the month with 10th and 11th month being slightly higher.

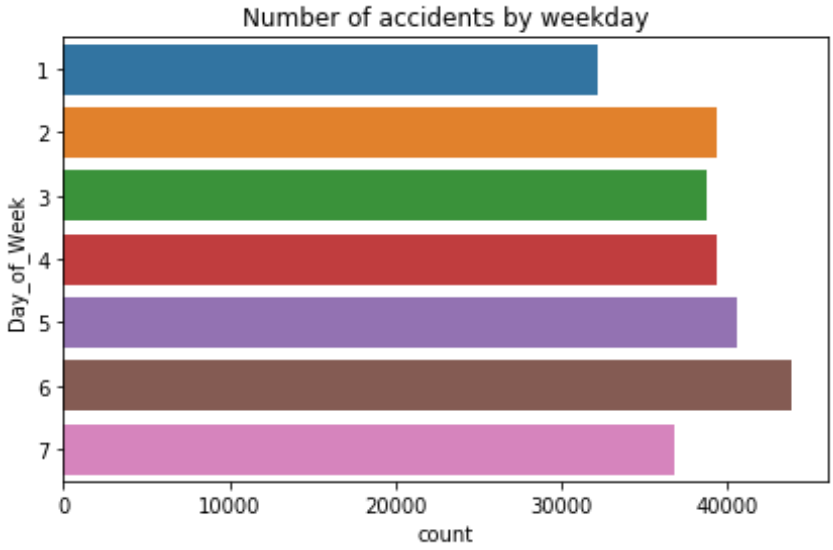

Figure 3 number of accidents by weekday

The 6th day of the week had slightly higher accidents than other days whereas 1st and 7th day had slightly lower accident than other days.

The analysis was also done to check the accident severity in a month. The result is as below:

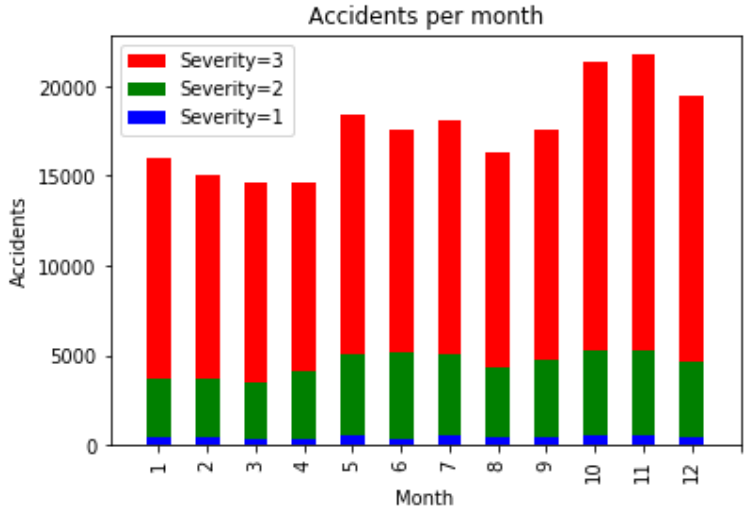

Figure 4 severity and accident per month

Then, a heat map was prepared to visualize the correlation between accident severity, weather condition, light condition and hour.

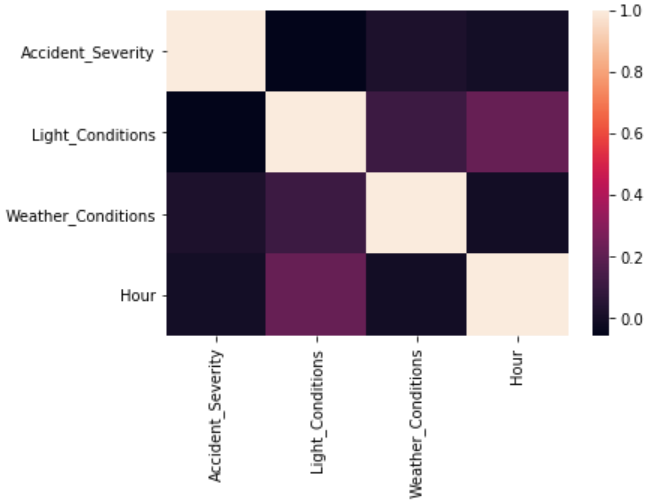

Figure 5 heat map

The next analysis was done to check the hourly accident severity caused due to different weather condition.

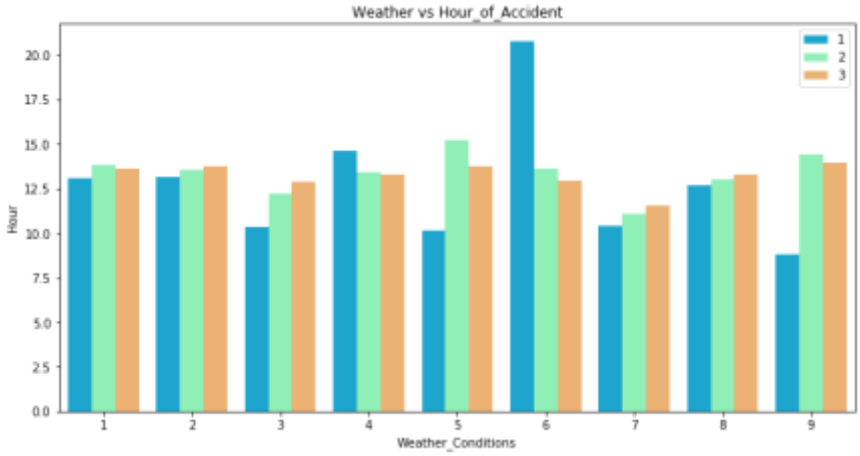

Figure 6 weather vs hour of accident

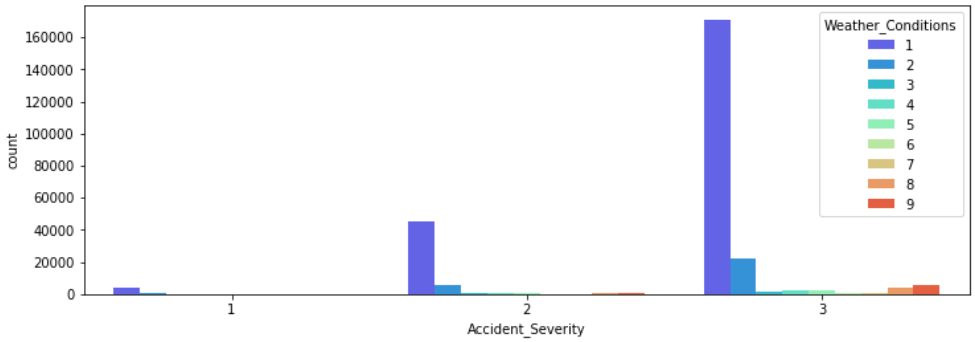

Figure 7 severity and weather condition

The figure shows that accident usually occurred during afternoon due to weather condition. The accidents with slight severity occurred the most. Accidents usually took place when the Weather

conditions were fine and also there weren't any high winds: meaning which the weather conditions didn't effectively contribute to occurrences of accidents.

The relation between journey purpose of the driver and their age and sex is given in the following graph

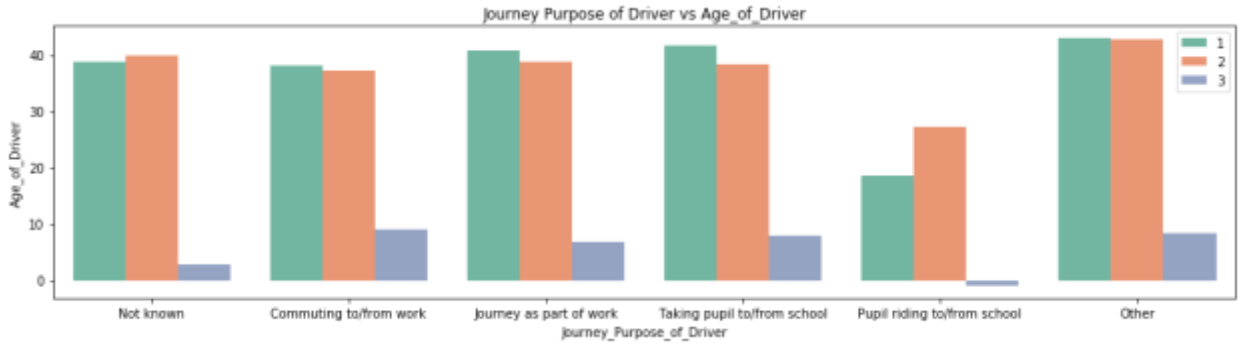

Figure 8 journey purpose of driver vs age

- Here 1 means male, 2 is female and 3 is unknown

3.2 Predictive Modelling

Firstly the rows were split into 80-20 ratio between the training and testing sets. Furthermore, the training samples were again split into 80-20 creating the validation set for the development of models.

Different classification models were used to build the classification model to predict the accident severity. These algorithm provided a supervised learning predicting with certain accuracy and computational time. These two properties have been compared in order to determine best suited algorithm.

Decision tree (Random Forest), Logistic Regression and K-Nearest Neighbour were used for predicting the severity. SVM was not done as the computational time was very high in this case. For the prediction, 9 features were used to predict the severity. Similarly, random forest was done with 100 decision tree, logistic regression with c =0.001, solver as libilinear and for KNN, the value of k was chosen as 19.

4. Results and Discussion

The metrics used to compare the accuracy of the model were jaccard score, f1- score, precision and recall.

Their weighted average were taken.

Algorithm	Jaccard score	F1-score	precision	recall	time
Random Forest	0.89	0.87	0.90	0.88	41.34757
Logistic Regression	0.78	0.68	0.61	0.78	3.51829
KNN	0.78	0.71	0.71	0.78	3.63040

From the table, we can say that Random Forest is the best model to predict the severity of the accident. The logistic regression and KNN model showed similar accuracy.

5. Conclusion

In this study, I analyzed the relationship between severity of accident and different characteristics of the accident. I took 9 different characteristics to predict the severity of the accident. Three different classifications models were built to predict the severity of accidents and random forest classifier was the best model for prediction.

6. Reference

- https://github.com/cptmacp/Accident-Prediction/blob/master/Accidents.Prediction.Project.ipynb
- https://www.kaggle.com/sohamsave/accident-data-analysis
- https://github.com/Josep-at-work/Coursera_Capstone/blob/master/Predicting_Traffic_Accident_Severity_.pdf
- https://data.gov.uk