Analysis of UK Traffic Accidents

1. Introduction

1.1 Background

Road accident causing enormous losses it did not affect only those who are victims. But it also causes impact to the economy and society. Because of the deaths and injuries both physically and mentally from the accident causing the victims and their families to lose their productivity affect the overall productivity of the country

Accidents also generate other costs such as legal costs. Costs from impact on traffic conditions, etc. The economic and societal impact of traffic accidents cost British citizen million British pounds every year.

1.2 Problem

It would be great if we can find what are the most common causes, to decrease or prevent the accidents. we might be able to make well-informed actions and better divide financial and human resources.

1.3 Interest

The data comes from government website www.data.gov.uk. UK police forces collect the accidents data using the form called Stats19. This project aim to analysis on U.K accidents data from year 2015 to predicting the accident severity. Government who want to decrease prevent cost. Insurance company would be very interested in accurate prediction of the accident severity, for Insurance premium and risk management. And a driver who want to increase they chances of staying safe on the road.

2. Data acquisition and cleaning

2.1 Data sources

Road accidents and safety data comes from government website www.data.gov.uk. This data is consist of 3 files that are Accident Circumstances, Vehicle and Casualty. Every column of the dataset is in numerical format. With a variable lookup to explain each numerical category in accidents dataset was provided on the www.data.gov.uk website as well.

This dataset can answer some questions like

- what the most accident occour on the days of a week?
- what time that had the highly accident?
- what the most age of driver that involved in accident?

we can find this answer by using data visualization

2.2 Data cleaning

Data downloaded and 2 files Accident Circumstances and Vehicle were joined into one table. Because I think that was the internal and external factors to cause accident. There is one problem with the datasets. This data had 2 types of missing values '-1' and 'Nan'. I did not imputing any mean or median value as the dataset are big enough to execute analysis.

2.3 Feature selection

After data cleaning, there were 70,383 samples and 53 features in the data. When I inspected the correlation of independent variables, most of them not have strong correlations between each other feature. I also consider some features to predictors for machine learning algorithm. After all, 11

features were selected as day of week, speed limit, light conditions, weather conditions, road surface conditions, did police officer attend scene of accident, vehicle type, sex of driver, age of driver, engine capacity (CC), age of vehicle.

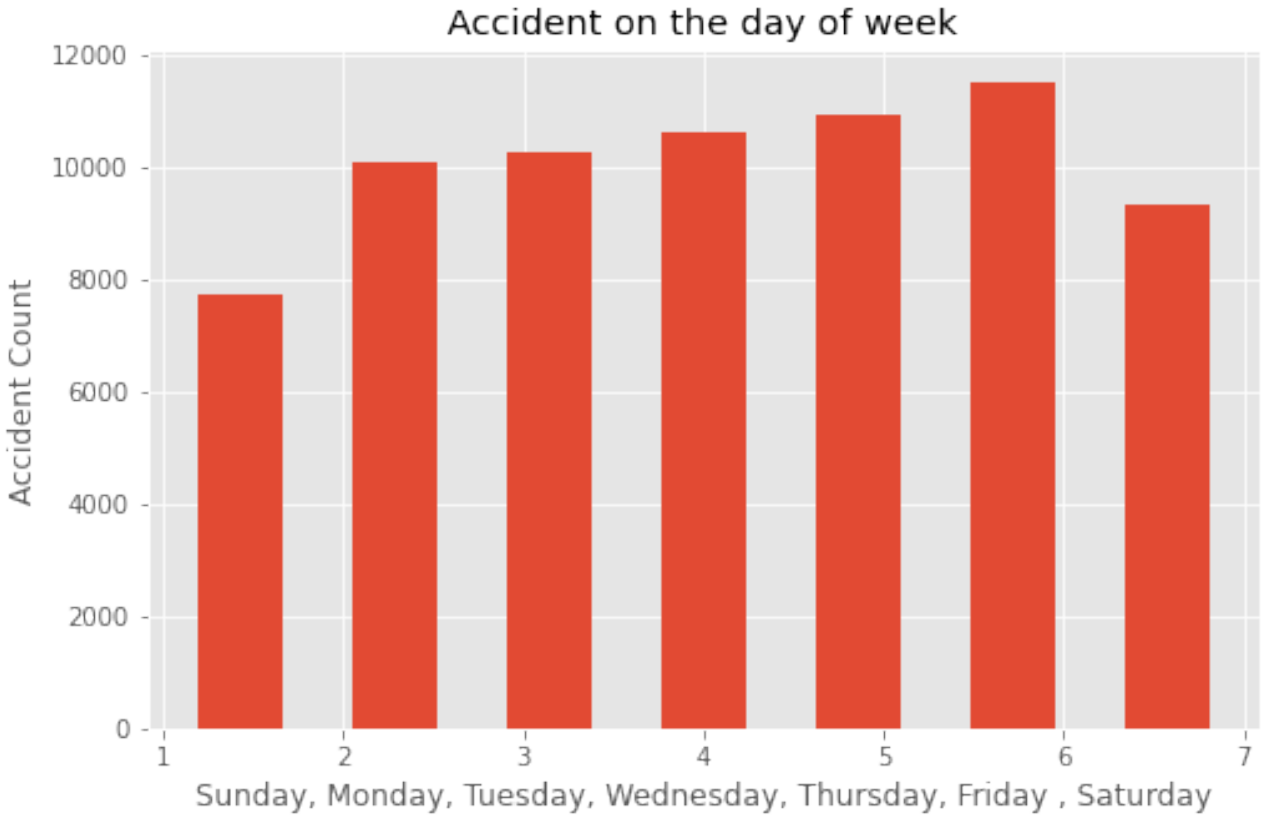
3. Exploratory Data Analysis

3.1 Data Visualization

The important thing about analysis uk road accident to understand the factors involved and their impact. Accidents that happen often consist of many variables like time, age of driver, weather condition etc. In this section data visualization will find out about what time of accidents to get intution and some driver's age who are involved in the accident.

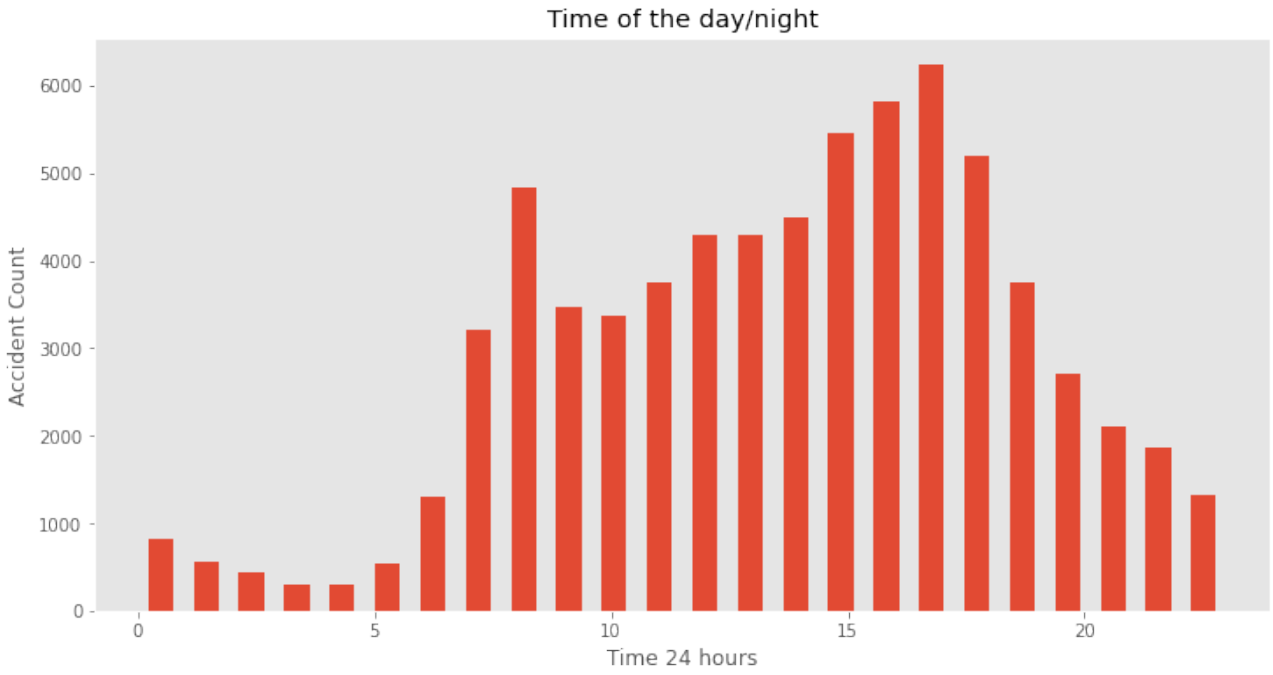
3.2 What the number of accidents on the days of a week?

This simple graph visualizes the number of accidents and distribution on the day of week. As we can see that friday has the highest volume of accidents in this dataset. We have to keep in mind that number of accidents could be rely on traffic amount on certain day.



3.2 What time had the most accident?

In this case, we can see most of accident are occur in the morning and evening rush-hours (07:00 - 10:00 and 16:00 - 19:00). We can assume that this time of the day has the most traffic moving such as people go to and get off work.

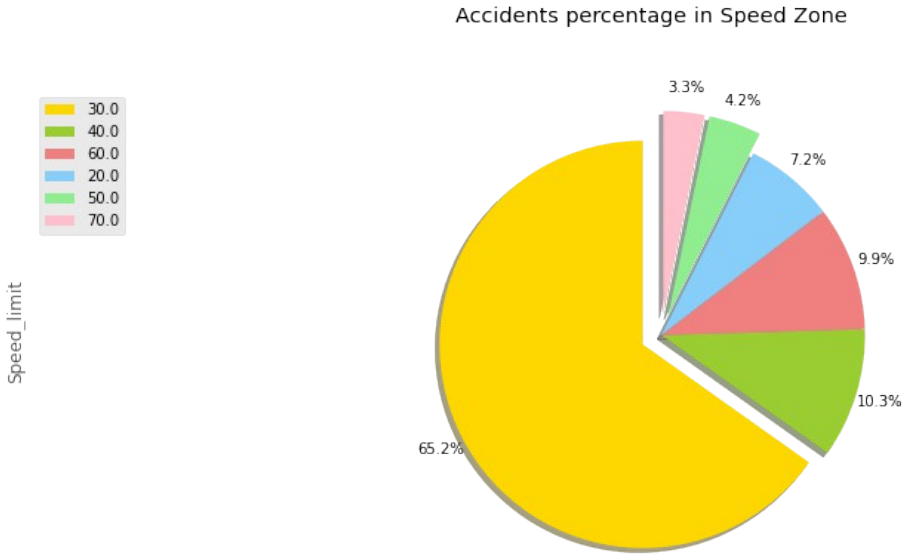


3.3 What are the age group are most likely to be involved in accidents?

This is very interesting fact about this dataset. the number of accident increase by age. Most of a driver who age around 26 - 35 are involved in the accident. When age over 36 number of accident are decrease Significantly. For me they are 2 ways to think, First people who age around 26 - 35 having risky driving habits Than any other age. Second Age distribution in the United Kingdom age 25 to 35 are more in the number of drivers with different age.

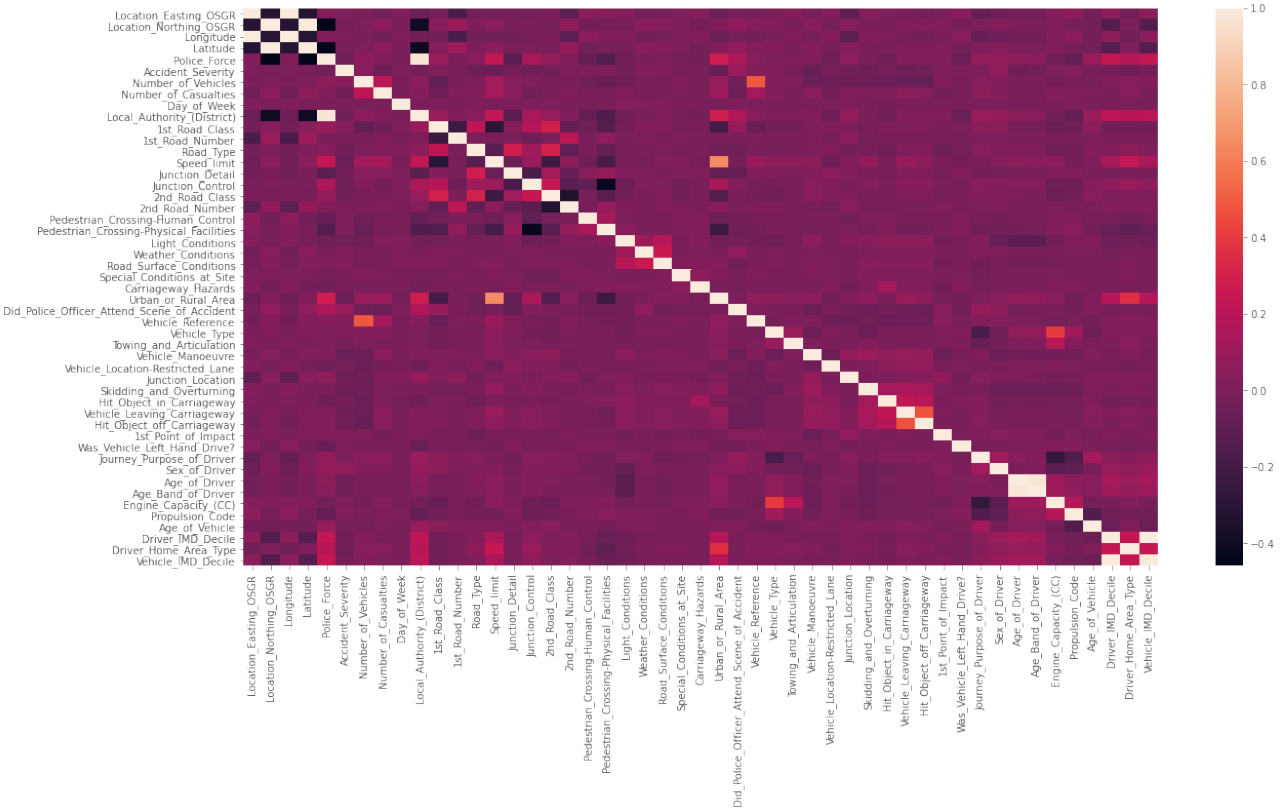
3.4 What the Speed Zone had the most accident?

It really surprise me Most of the accidents occurred on the road where the speed limit is 30. I think the more speed limit the more dangerous. Maybe people can increase their risk when they feel safe, so if road users realize they are at greater risk of accidents, Will be more careful with driving behavior.



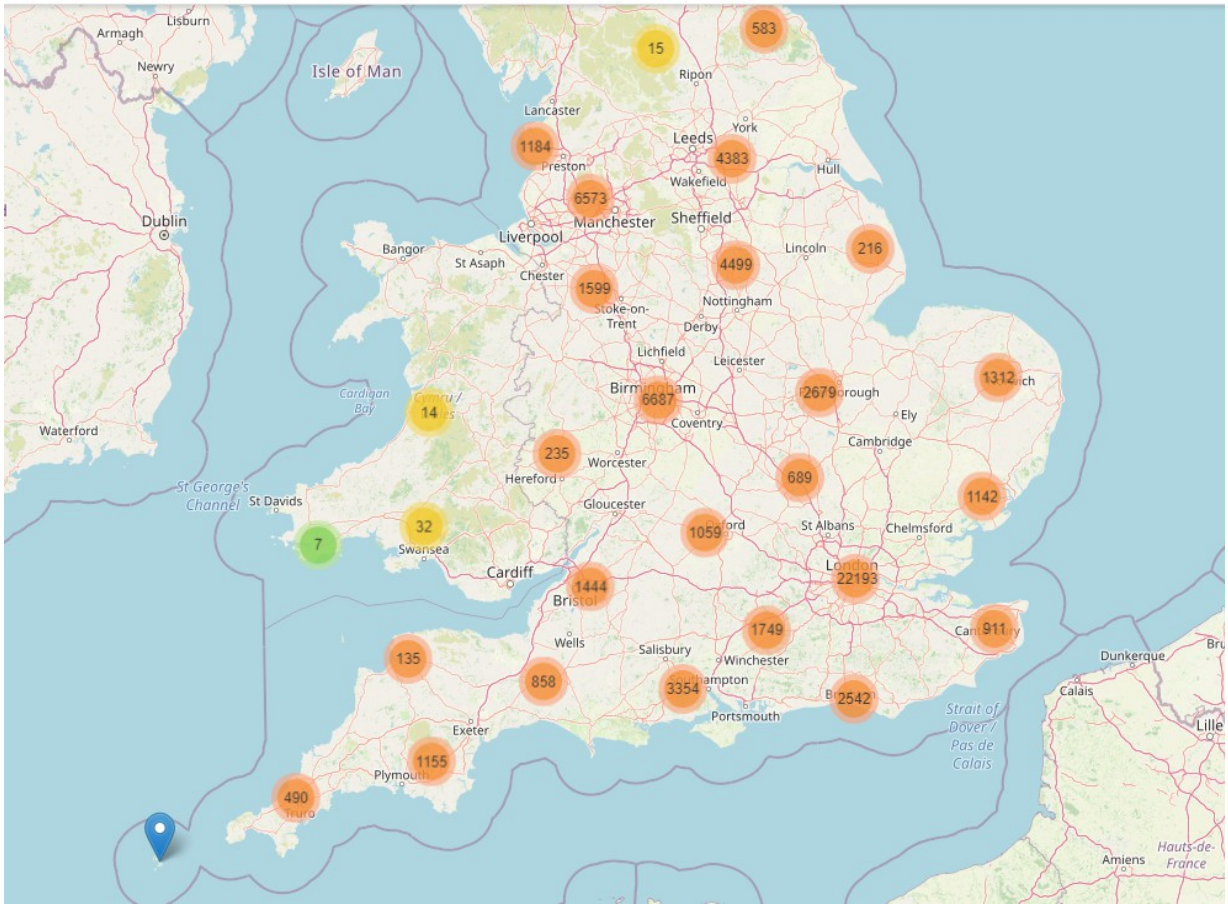
3.5 Correlation between categorical variables

This data had many variables but there is not so much strong correlations between any variables. I can spot few interesting correlations for example: speed limit and Urban or Rural Area.



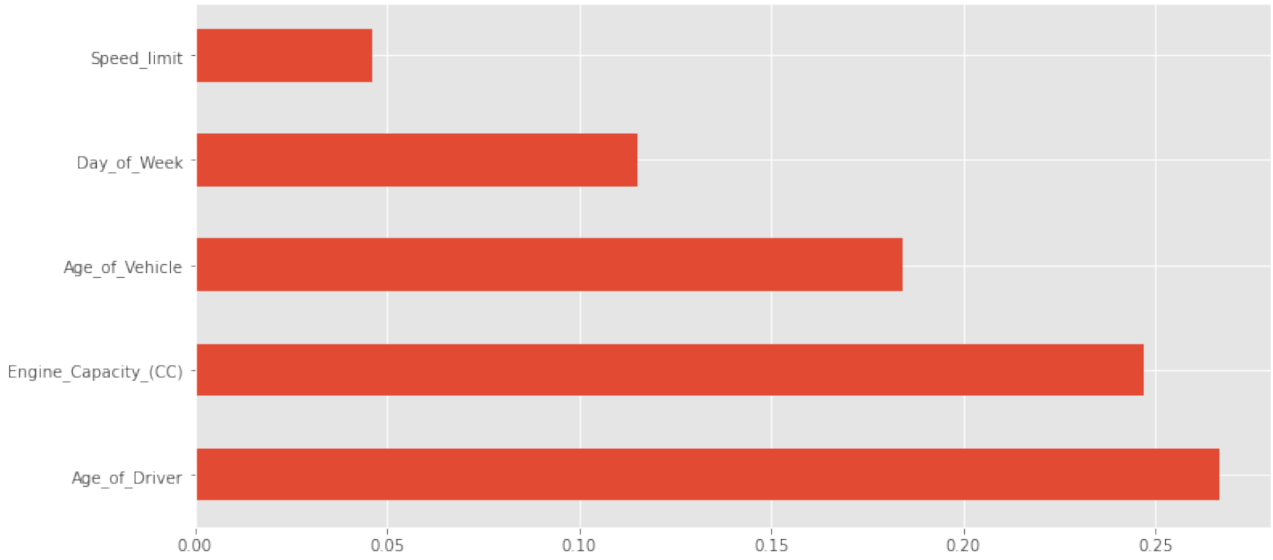
3.6 Plotting accidents Location on Google Maps

I will be using UK maps to plot the accidents. This data It includes the location, date, and time for every traffic accident, London has the most accidents. However, it also depends on the amount of traffic of each area.



3.7 Important Feature

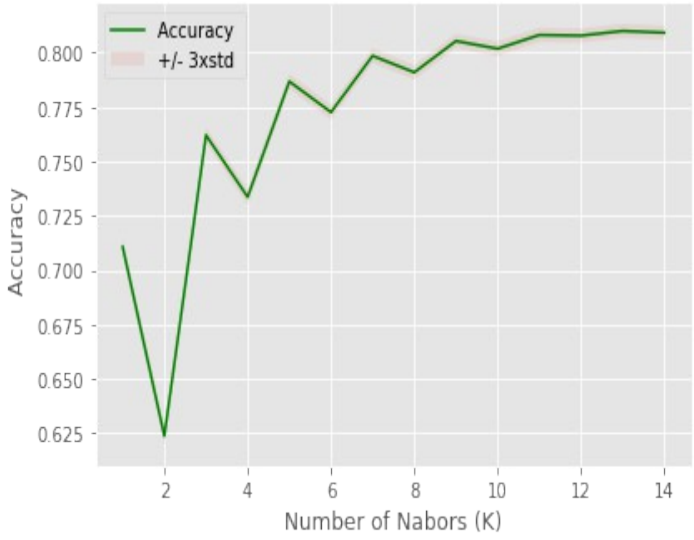
When we use Sklearn's random forest library to check out what the most important features and visualize in ascending order it will let's us know what most important features to cause the traffic accident.



4. Predictive Modeling

The Accident Severity that we want to predict is a categorical variable with discrete values. So classification model can be used and focus on the probabilities a traffic accident might happen. I don't use Support Vector Machine because it is not very efficient computationally when dataset is very big, this data has more than 1,000 rows. So I just use K Nearest Neighbor(KNN) and Decision Tree algorithm to predict accident.

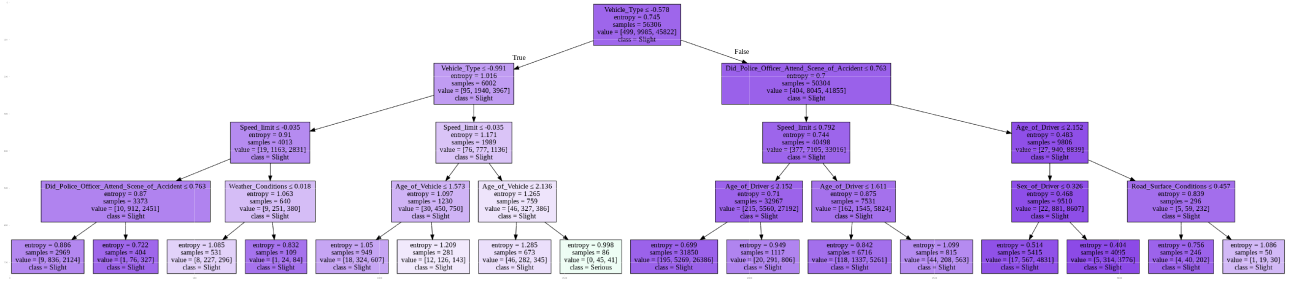
4.1 K Nearest Neighbor(KNN)




```
[33]  1 print( "The best accuracy was with", mean_acc.max(), "with k=", mean_acc.argmax()+1)
```

```
The best accuracy was with 0.809973715990623 with k= 13
```

4.2 Decision Tree



```
[36]  1 from sklearn import metrics
      2 yhat = Tree.predict(X_test)
      3 print("DecisionTrees's Accuracy: ", metrics.accuracy_score(y_test, yhat))
```

```
DecisionTrees's Accuracy:  0.8148753285501172
```

Both algorithms provide quite similar accuracy values where the Decision Tree perform slightly better.

5. Conclusions

In this study, I analysis of UK traffic accidents and predicting the accident severity. I identified day of week, speed limit, light conditions, weather conditions, road surface conditions, did police officer attend scene of accident, vehicle type, sex of driver, age of driver, engine capacity (CC), age of vehicle among the most important features that cause the traffic accident. I built classification models to predict accident severity. These models can be very useful in helping decrease traffic accident we can suggestion to government for road safety campaigns and communication or law enforcement to road users.

6. Future directions

I was able to achieve 81% accuracy in the classification problem. It performance really good because the ability of predicting accident severity is very important since drivers' fatality has the highest cost to society economically and socially. However, I think this data still don't has 2 very important factors causing different accident severity. First the actual speed that the vehicle was going when the accident happened. Second alcohol usage it is well-known that the alcohol is one of the leading causes of accidents. If the actual speed and alcohol usage was available, it is extremely likely that it could have helped to improve the performance of models studied in this paper.