Predicting Car accidents severity in UK

1. Introduction

i. Background:

One of the greatest problems that our communities are facing today is traffic accidents. Nowadays, there are many traffic accidents that happen daily. Worldwide, averages of seventeen people are killed in these traffic accidents every minute. There are another twenty to thirty people that end up handicap after being involved in these tragic, traffic accidents. There are a few main causes of traffic accidents.

ii. Problem:

The kind of data that we will use and study will help us determine or predict if there is a chance for a car accident based on different features for example light, weather, and road conditions. So this project's aim is to predict car accidents in the UK based on specific data that we have in our hands.

iii. Interest:

The interest in this project will be dragged mostly by government facilities that are responsible for road safety such as Traffic police. Also, it may help road travelers to decide whether the road is safe to use or not.

2. Data

i. Data Source:

UK police forces collect data on every vehicle collision in the ukon a form called Stats19. Data from this form ends up at the DfT and is published at here and the data csv files found in kaggle datasets found here. This Database collected between (2005 to 2015).

ii. Data Cleaning:

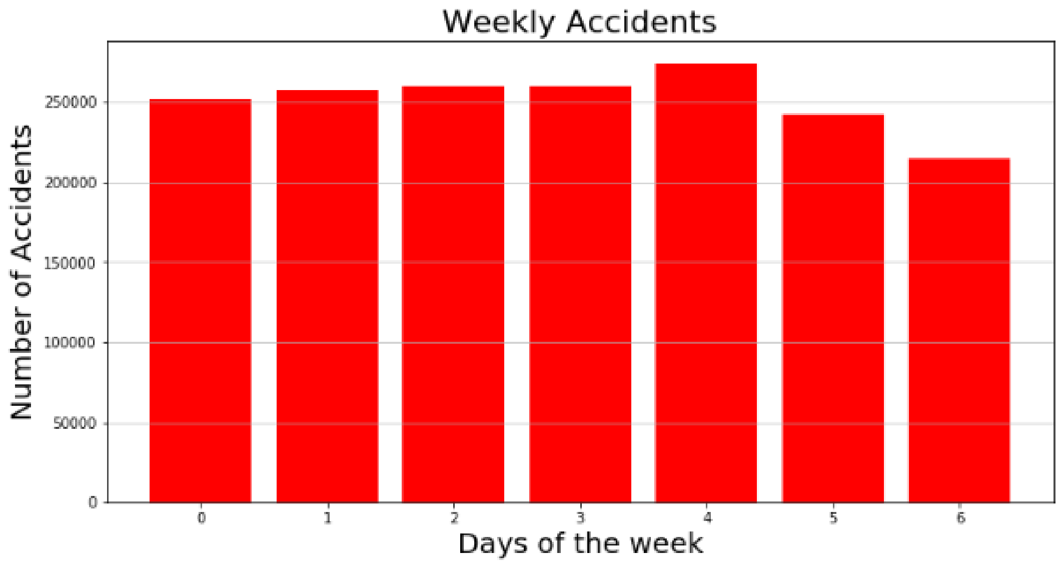
This data consists of three files that are accidents, causalities, and vehicles. However, we have one more file which is general information about the traffic count for the year 2000 to 2015. We can use general traffic information data for the machine learning part. There are two types of missing values '-1' and 'Nan'. So, We will solve missing data problem by investigating each column with total missing values. We will not be imputing any mean or median value since the dataset is big enough to perform analysis.

iii. Data Visualization:

I will try here to describe the data that I have using visual figures to walk you through understanding the data and giving better view on what we are dealing with.

The first thing we can do is to find out about accidents time to get intuition and some driver's age who are involved in the accident.

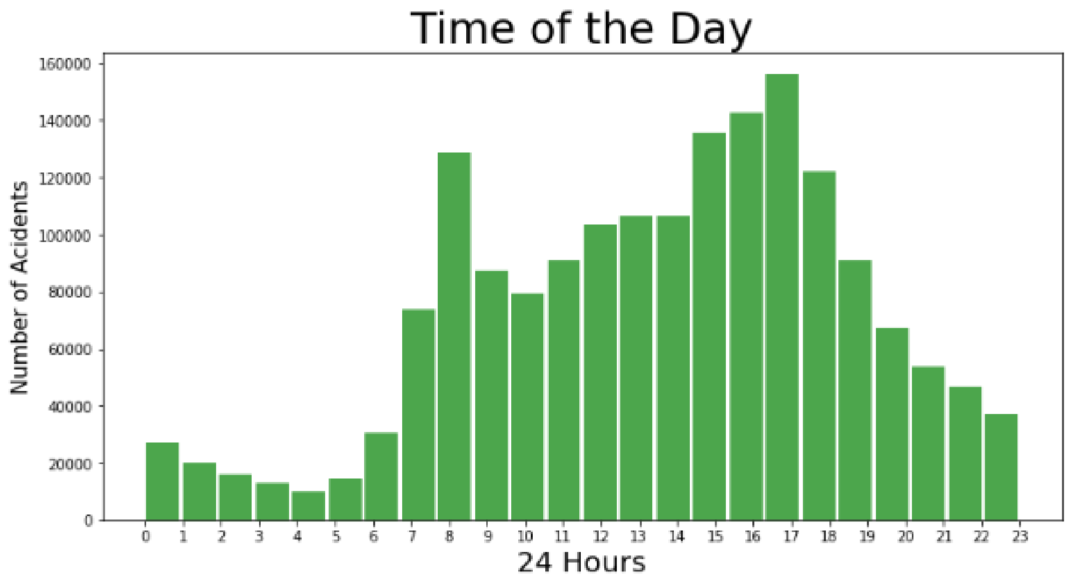
1- We can find out the number of accidents on the days of a week.



0-> Monday , 1-> Tuesday , 2-> Wednesday , 3-> Thursday , 4-> Friday , 5-> Saturday , 6->Sunday

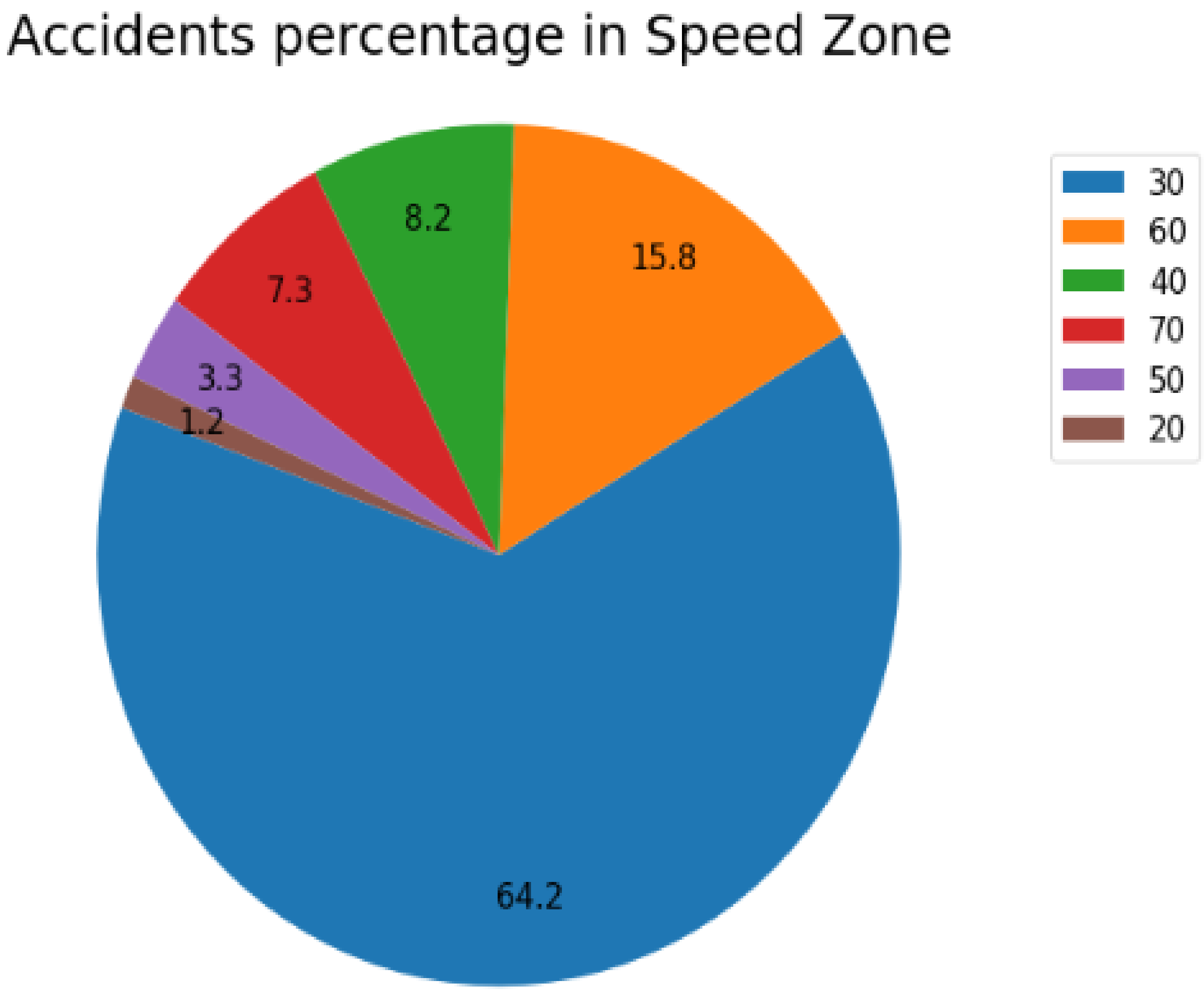
We can see that the most day accidents happened in UK is on Friday. Also, we have to keep in mind that accidents numbers could be depending on traffic amount on particular day.

<u>2- We can find out about the accidents number using hours of the day.</u>



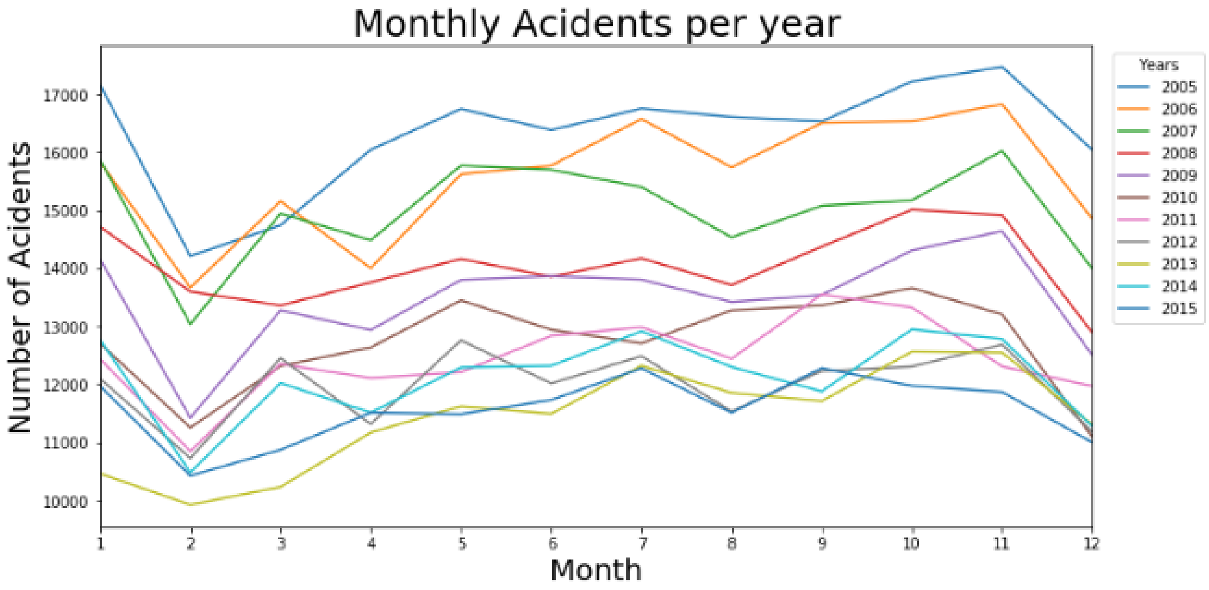
We can see that the most of accidents happened around after noon. We can assume that this time of the day has the most traffic moving such as people leaving from work.

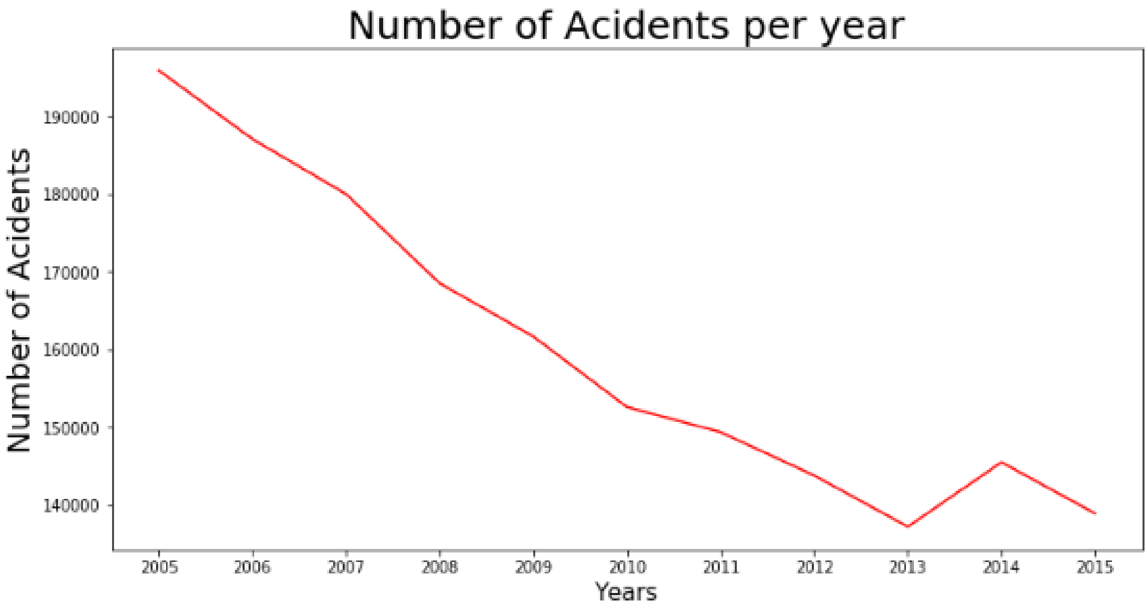
3- We will study the speed limits of the roads



Most of the accidents occurred on the road where the speed limit is 30. I was expecting more accidents on highway or major roadways. Some of the accidents could be cause of stop sign, changing lanes or turning into parking lot etc.

4- Is the number of accident increasing by the time or decreasing.



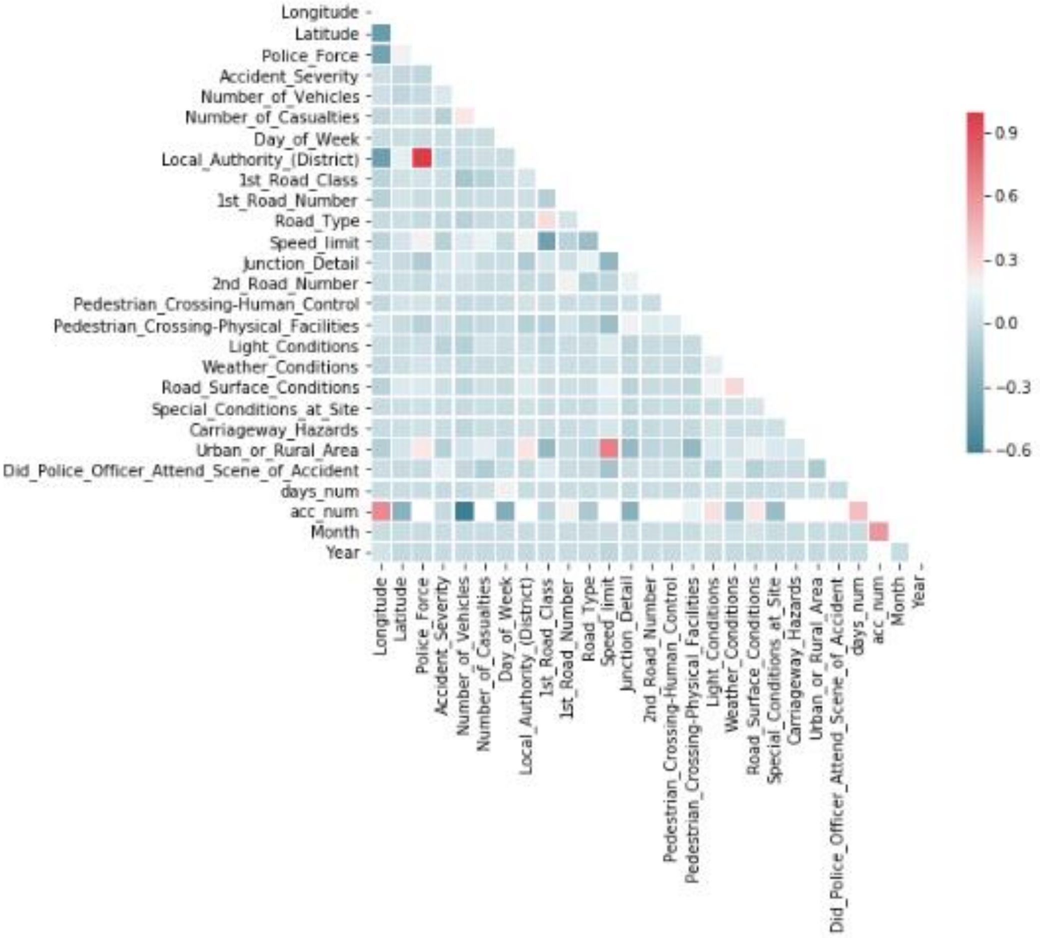


We can see that most of the accidents happened between (October – November) throughout the years.it may refer to the time of the year where Workers and Students finish their holiday and return to their work.

Also, The Second Line graph illustrates how the number of accidents decreases dramatically during the year (2005 – 2015).it may because the government set new traffic rules that brought more safety to the roads and applied careful driving on drivers.

After all, we are trying to understand the data and how each feature effect on it. We will now try to find any relation between the features to use it to find a solution to predict accidents severity using machine learning models.

Correlation between all features:



As shown in the figure that represents a heat map for the correlation between the features of the database. There are not many correlations that could help us during the process. We can see that there is a strong positive correlation between Local_Authority_ (District) and Police_force but it will not help us in our study. There is another strong positive correlation between Urban_or_Rural_Areas and Speed_limits we can use it.

3. Problem solving:

We can use different supervised machine learning algorithms to predict car accidents severity such as Logistic Regression and Decision tree.

Machine Learning Algorithms:

Before we start using ML algorithms on our data, we need to do some normalizing on it

Normalizing Data:

It is the first step we take before we apply ML models to our data. The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. There are a few columns that we will standardize, so it would not affect negatively our machine learning algorithms. The age of the driver is from 18 to 88 in the dataset and we can normalize it. Also, the age of the vehicle is also from 0 to 100 and it can skew the performance of your machine learning algorithm and we will normalize this predictor too.

1- Logistic Regression:

We will take (Did_Police_Officer_Attend_Scene_of_Accident' , 'Age_of_Driver' ,'Vehicle_Type', 'Age_of_Vehicle','Engine_Capacity_(CC)','Day_of_Week' , 'Weather_Conditions' , 'Road_Surface_Conditions', , 'Light_Conditions', 'Sex_of_Driver' ,'Speed_limit') columns to use it as an input for our logistic regression model. In addition, I used (Accident_Severity) column as the output or prediction for the model.

We will train the model using 80% of the whole data that we have and the other reminded 20% we will use it for testing.

	precision	recall	f1-score	support
1	0.00	0.00	0.00	1507
2	0.00	0.00	0.00	23102
3	0.88	1.00	0.94	181109
Avg/total	0.78	0.88	0.82	205718

After we trained the model we will have the ability to predict the accidents severities but first we should test the accuracy for our model and I did that using two methods (jaccard index – Log Loss) and we got

Methods	Accuracy by percentage
jaccard index	88.0%
Log Loss	38.0%

So we can see that we could use logistic regression to predict car accidents severity but we will test another model to see if it perform better.

2- Decision Tree:

We will use the same input and output for our model that we used in logistic regression model.

Accuracy by percentage
88.0%

Conclusion:

As we saw, we could use both algorithms to predict car accidents severity in the UK based on the data that we have from 2005 – 2015.

Some Recommendations:

After we reviewed the data and analyzed it, there are a lot of factors that could help the authorities to reduce the number of car accidents such as the age of the driver, the day of the week, hours of the day, road condition and specific areas or locations that have the most accidents rate.