**Predicting the Severity of Road Accident.**

**1. Introduction**

**1.1 Background**

Car crashes and road accidents could be considered an old topic. Yet, with the progress of cars and the capabilities of the technology they carry, it is even more important to have tools and means available to mitigate their occurrences, as well as the implications and consequences they have for the people involved. But predicting the severity of a car crash is no easy task. And even when possible, precision levels will vary significantly depending on, among many factors, the data available and how well the problem has been modeled.

**1.2 Problem**

More than 3000 people die in road crashes every day worldwide, including approximately 1000 people in the UK. Such accidents cost $518 billion globally and $230.6 billion per year in the United Kingdom. In particular, many young people are affected by traffic accidents, as most traffic accidents are caused by the 25–45 age group (Association for Safe International Road Travel Data). Hence, substantial efforts have been made to address the main factors that cause traffic accidents, such as speeding, distracted driving, driver fatigue, road conditions, use of mobile phones, and poor weather conditions.

**2. Data acquisition and cleaning**

**2.1 Data sources**

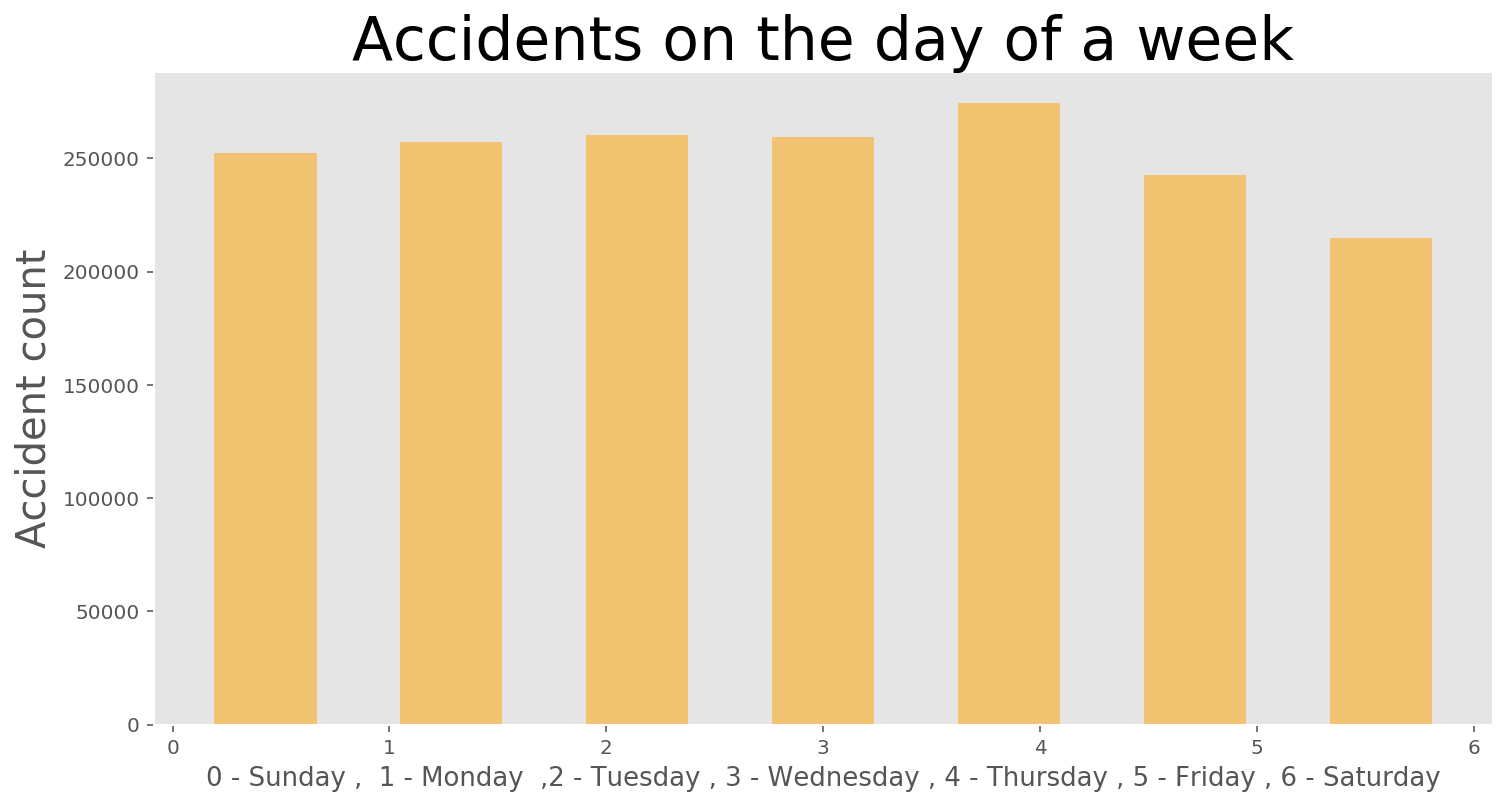
While searching for a meaningful dataset for my capstone project, I came across [data.govt.uk](http://data.govt.nz/) ; a repository of open datasets related to all kind of activities throughout United Kingdom and published by the central government. UK police forces collect the accidents data using the form called Stats19. The data consists of all kind of vehicle collisions from 2005 to 2015. Every column of the dataset is in numerical format. A supporting document to understand each numerical category in accidents dataset is provided on the [www.data.gov.uk](http://www.data.gov.uk/) website as well.

**2.2 Data cleaning**

Data downloaded or scraped from multiple sources were combined into one table. There were a some missing values in the dataset. In this particular dataset, there are two types of missing values '-1' and 'Nan'. We will investigate 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.

**3. Exploratory Data Analysis**

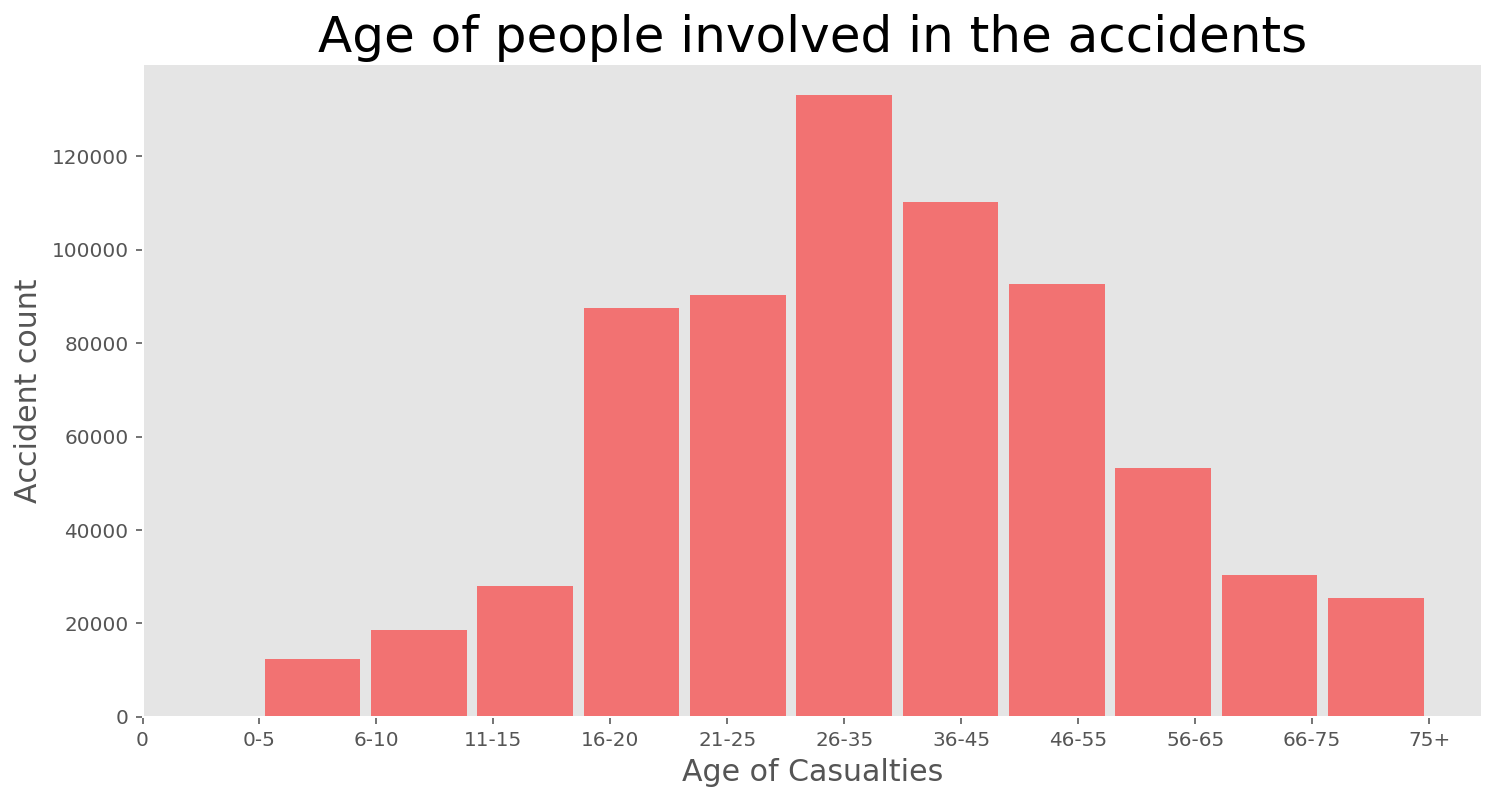
**3.1 Relationship between day of a week and accident**



As we can see that thursday has the highest amount of accidents in this dataset from 2005 to 2015. We have to keep in mind that accidents numbers could be depending on traffic amount on particular day.

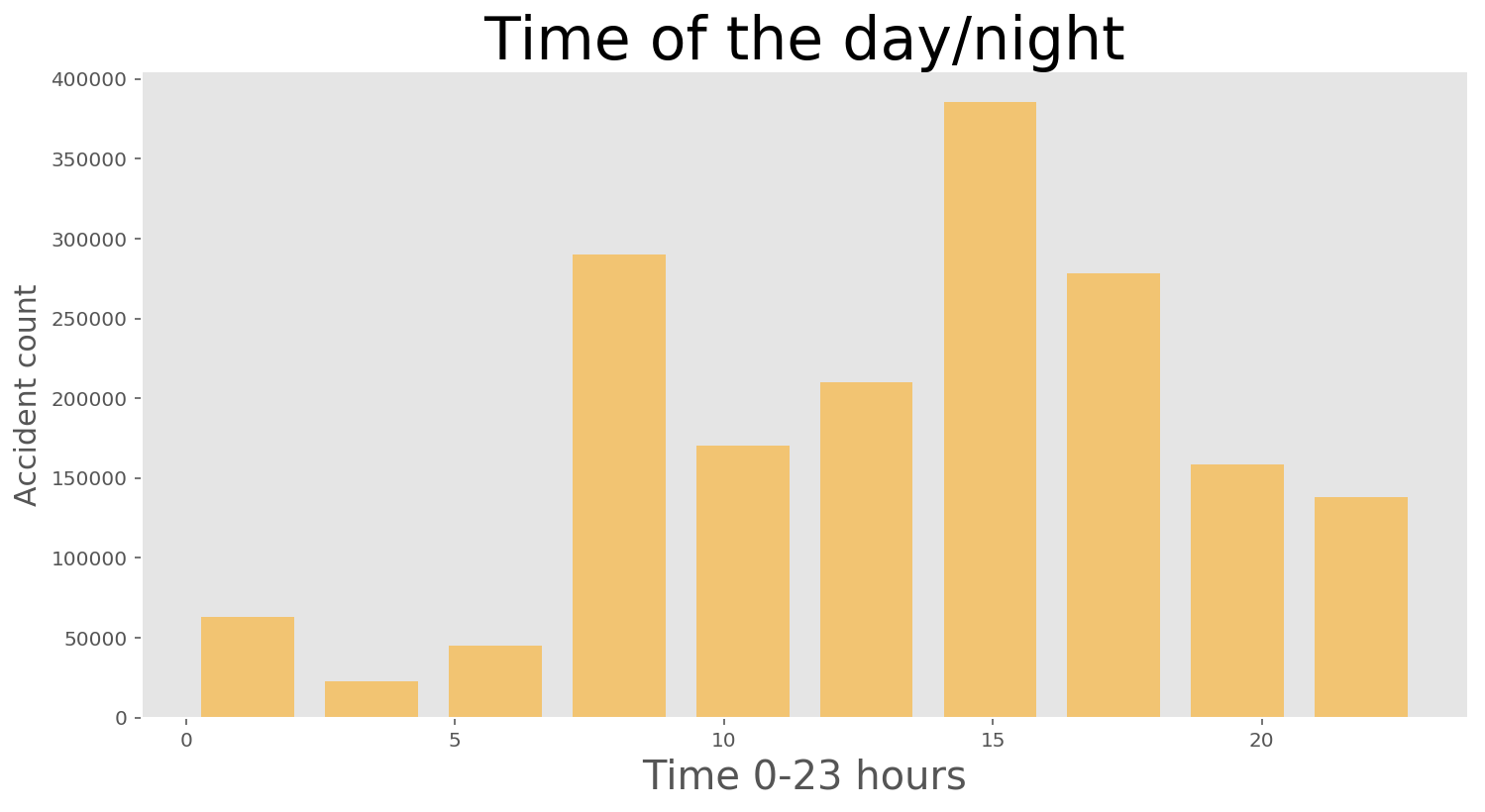
**3.2 Relationship between accident and age**

It is widely accepted that younger generation are more likely to prone to accident, and it was indeed supported by our data.



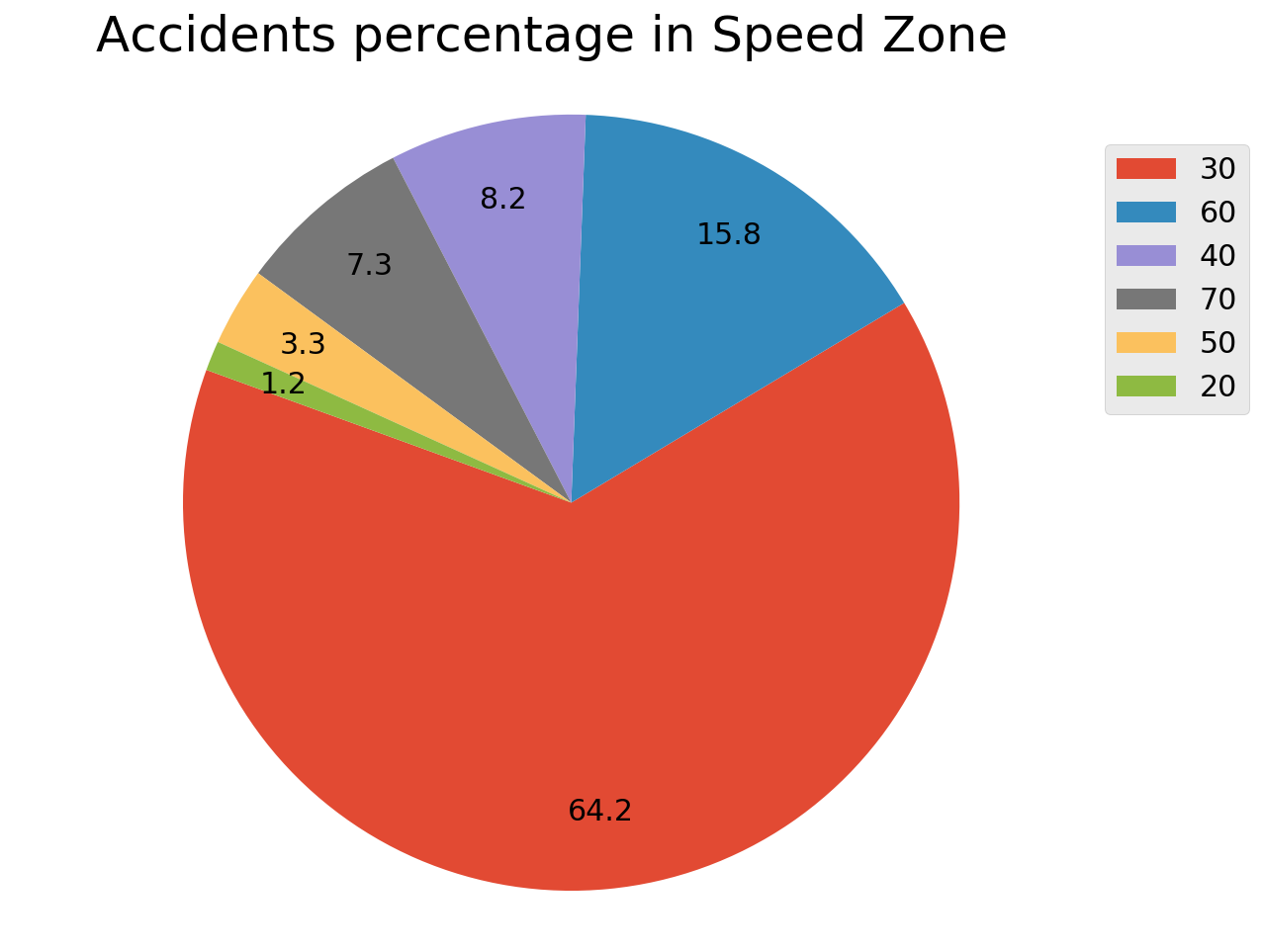
This is very interesting fact about this dataset. Most of the drivers’ age is around 25 to 35 who are involved in the accident. However, we do not know the number of drivers with age 25 to 35 on the road compare to other ages. Intuitively, I would assume that the driver with age 25 to 35 are more in the number of drivers with different age.

**3.2 Relationship between accident and time of the day/night**



We found out 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.

**3.2 Relationship between accident and speed limit**



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. Predictive Modeling**

**4.1 Logistic Regression**

We got 84.96% accuracy using this model.

precision recall f1-score support

1 0.217391 0.001100 0.002188 4547

2 0.348315 0.000641 0.001279 48395

3 0.849757 0.999816 0.918699 299166

micro avg 0.849589 0.849589 0.849589 352108

macro avg 0.471821 0.333852 0.307389 352108

weighted avg 0.772671 0.849589 0.780770 352108

| **Predicted** | **1** | **2** | **3** | **All** |
| --- | --- | --- | --- | --- |
| **Actual** |  |  |  |  |
| **1** | 5 | 10 | 4532 | 4547 |
| **2** | 11 | 31 | 48353 | 48395 |
| **3** | 7 | 48 | 299111 | 299166 |
| **All** | 23 | 89 | 351996 | 352108 |
|  |  |  |  |  |

**4.2 Decision Tree**

We got 75.73% accuracy using this model.

precision recall f1-score support

1 0.049082 0.054102 0.051470 4547

2 0.184749 0.182374 0.183554 48395

3 0.860492 0.860943 0.860718 299166

micro avg 0.757259 0.757259 0.757259 352108

macro avg 0.364774 0.365806 0.365247 352108

weighted avg 0.757137 0.757259 0.757195 352108

| **Predicted** | **1** | **2** | **3** | **All** |
| --- | --- | --- | --- | --- |
| **Actual** |  |  |  |  |
| **1** | 246 | 1037 | 3264 | 4547 |
| **2** | 1075 | 8826 | 38494 | 48395 |
| **3** | 3691 | 37910 | 257565 | 299166 |
| **All** | 5012 | 47773 | 299323 | 352108 |

**5.Conclusion**

As we have implemented the Logistic Regression, Decision Tree and K Nearest Neighbor algorithms to predict the accident severity. There are two things that we can conclude from this learning.

**Machine Learning Conclusion**

As we have tried two different algorithms to predict the accident severity. It was clear that Decision tree performed much better in terms of predicting all the classes of accident severity. Logistic regression has better accuracy but it does not mean it did better than other algorithm.

**Recommendation for Public or Law Enforcement**

There are few things that were clear from this project. First, the most of the accidents occurred in an area where speed limit is 30. Secondly, most of the casualties are in age band of 25 to 35 years. We already know that the car insurances are expensive for the young age people and this is one of the reasons. We found out 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.