Study of Car Accidents Severity in the UK

By Eric Yu

20 September 2020

Car Accident Severity

## ****Introduction****

* 1. **Background**

**Hundreds of millions of people commute in the road every day for work, schools and other social activities. According to World Health Organization1, each year, 1.35 million people are killed on roadways around the world, every day, almost 3,700 people are killed globally in road traffic crashes involving cars, buses, motorcycles, bicycles, trucks, or pedestrians and more than half of those killed are pedestrians, motorcyclists, and cyclists. Road traffic injuries are estimated to be the eighth leading cause of death globally for all age groups and the leading cause of death for children and young people 5–29 years of age. More people now die in road traffic crashes than from HIV/AIDS.**

**This makes traffic safety not only an engineering problem, but also a political, economic and public health issue of major significance. In order to find appropriate solutions to this problem, it is necessary to understand the causes of crashes and the factors that affect crash severity.**

* 1. **Business Understanding/Problem Description**

Since I’m looking to move to the United Kingdom, I was drawn to analyze this problem in this country.

Car crashes and road accidents could be considered an old topic. Yet, with the progress of the technology included in in cars and their new capabilities, it is ever more important to have tools and means available to mitigate their occurrences, as well as their implications and consequences for the people involved. With the data obtained, I will try to identify the variables that influence the severity of car accidents and the following questions:

1. What is the trend of road accidents in UK?
2. What are the accident occurrence rate in UK in different time?
3. What conditions (taking all conditions into account i.e. road, weather and light conditions) caused the most road accidents in UK?
4. How many numbers of casualties occur per accidents in UK and what is their distribution in terms of the total amount of road accident that occurred?
   1. **Target Audience**

This project intends to serve two groups of audience:

1. Policies makers who want to understand the factors influencing road accident severity to design safer traffic rules.
2. Transportation companies which could make significant impact to make road safer to use.

## ****Data****

A dataset of UK traffic data from 2005 and 2015 from Kaggle was used. It is amassed by the UK Government recording over 1.7 million accidents.

The original dataset on Kaggle has the following columns:

Accident\_Index

Location\_Easting\_OSGR

Location\_Northing\_OSGR

Longitude

Latitude

Police\_Force

Accident\_Severity

Number\_of\_Vehicles

Number\_of\_Casualties

Date

Day\_of\_Week

Time

Local\_Authority\_(District)

Local\_Authority\_(Highway)

1st\_Road\_Class

1st\_Road\_Number

Road\_Type

Speed\_limit

Junction\_Detail

Junction\_Control

2nd\_Road\_Class

2nd\_Road\_Number

Pedestrian\_Crossing-Human\_Control

Pedestrian\_Crossing-Physical\_Facilities

Light\_Conditions

Weather\_Conditions

Road\_Surface\_Conditions

Special\_Conditions\_at\_Site

Carriageway\_Hazards

Urban\_or\_Rural\_Area

Did\_Police\_Officer\_Attend\_Scene\_of\_Accident

LSOA\_of\_Accident\_Location

The following features were dropped due to significant portion of missing data of irrelevant to the analysis.

Location\_Easting\_OSGR

Location\_Northing\_OSGR

LSOA\_of\_Accident\_Location

Junction\_Control

2nd\_Road\_Class

The dataset has an initial total of 1,760,539 samples, each with [x] different features. The feature we want to predict is Accident\_Severity and it is a categorical feature that takes the values 1, 2, 3 to indicate that the crash was of class Fatal, Serious, or Slight, respectively. The remaining features are both numerical and categorical. Additionally, there’s also four geo spatial features used to indicate longitude and latitude of the crash site.

## ****Methodology****

* 1. Data Collection

There are a lot of car accident data available. I chose one in Kaggle.com :

https://www.kaggle.com/silicon99/dft-accident-data

The data set contains accident information of the UK from 2005 to 2015. It contains three files which are “Accidents”, “Causalities”, and “Vehicles”. This data set is selected because the data is continuous, every accident can be uniquely identified with a specific ID.

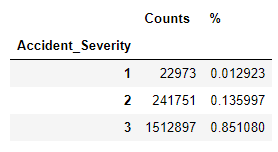
* 1. **Analytic Approach:**

As the dataset has many features, I will first perform exploratory data analysis using a few selected features to discover properties of data

The target we want to predict here (“Accident\_Severity”) is of categorical feature with three classes (‘fatal’, ‘severe’ and ‘slight’) which are mutually exclusive to each other. Therefore, we will use a logistic regression model to predict severity of accidents. We will use an F1 score to evaluate the performance of the model that we will train.

## **Analysis**

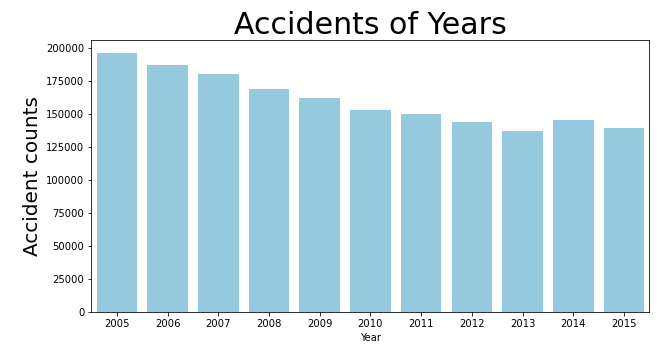
The aim of this study is to predict ‘Accident\_Severity’, let’s see how the severity is distributed:



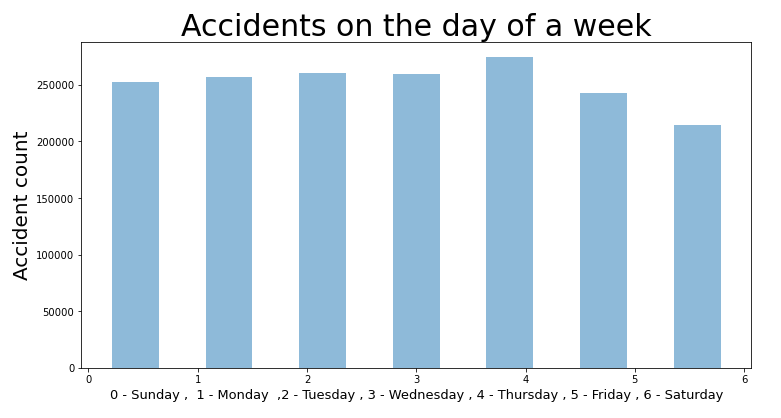
Over 85% of the data is actually ‘3’ which is slight.

* 1. **Exploratory Data Analysis**

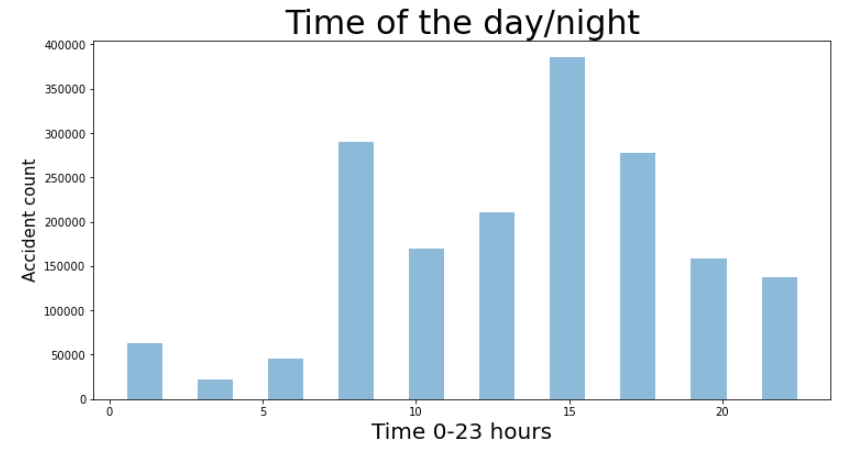
Let’s start with the trend of the total number of accidents over the years:



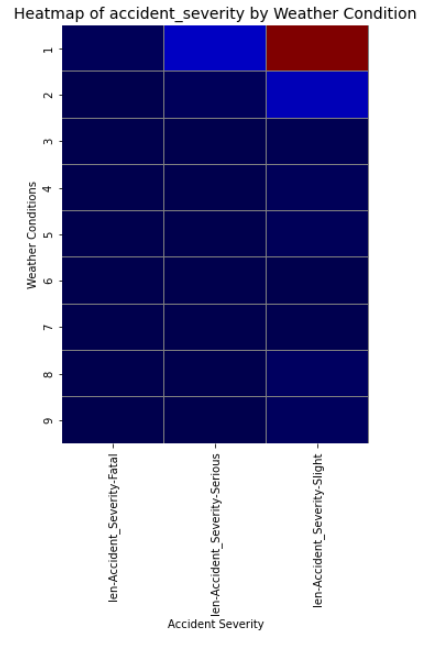
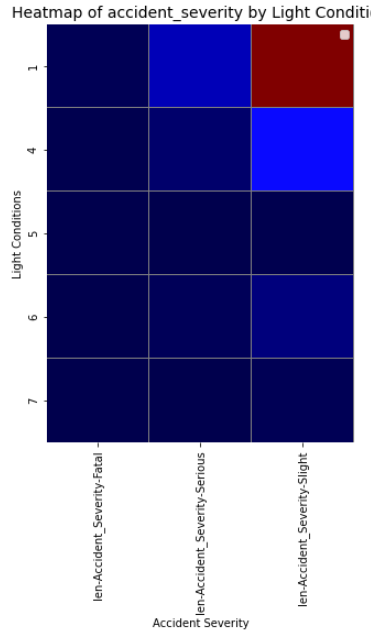
The total number of accidents decreased from 2005 to 2015.



As we can see that Thursday has the highest amount of accidents in this dataset from 2005 to 2015 and Saturday has the lowest amount. But the variation is not large.



We found out that the most of accidents happened around after noon.



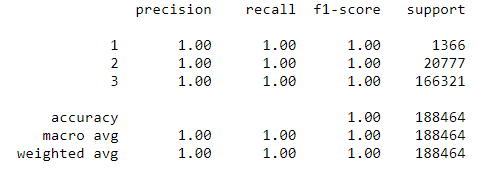
Weather conditions and Light conditions do not seem to have impact on Accident Severity.

* 1. Machine Learning

We will use logistic regression to train the data. In order to include the features in ‘Vehicles’ dataset into the model we first outer join ‘Accidents’ and ‘Vehicles’. The data was not joined in performing exploratory analysis because the ’Vehicles’ data contains multiple entries for some accidents (i.e. some accidents involved more than one vehicles), including them would make duplicated entries for the dataset we analyze. The joined data is cleaned up again eliminating Nan and unknown values for training, then object features were dropped and ‘Vehicle\_Type' was one hot encoded. The dataframe is then standardized and fit in to the model using 80/20 split. We will use f1-score to evaluate the model.

## **Results**

The results was astonishingly accurate. The accuracy was 1 and f1-score is also 1. The model can accurately predict all three classes of ‘Fatal’, ‘Severe’, and ‘Slight’.



## **Discussion**

The selected dataset can accurately predict the severity of accidents in the UK during 2005-2015. However, which of the used features are more important in predicting the severity? This question could be further studied with the probability of each features. Furthermore, the ultimate study question would be – how accidents could be prevented? This study would need a lot of data of all kind of traffic information.

## **Conclusion**

We explored and analyzed various aspects of traffic accidents the UK with data science. We used an existing dataset and performed EDA and machine learning models in our pursuit of solutions. We were able to find satisfactory answers to the questions we posed before the study.

Although this study cannot give a solution to lower traffic accidents, it certainly shed light on what affected the severity of a traffic accident.

Reference:

1: World Health Organization (WHO). Global Status Report on Road Safety 2018. December 2018. [cited 2019 April 8]. Available from URL: https://www.who.int/violence\_injury\_prevention/road\_safety\_status/2018/en/external icon