

# Traffic Accident Severity Prediction.

Abhijit Kumar Dutta

September 2020

## 1 Introduction

### 1.1 Background

Every year car accidents cause hundreds of thousands of deaths worldwide. According to a research conducted by the World Health Organization (WHO) there were 1.35 million road traffic deaths globally in 2016, with millions more sustaining serious injuries and living with long-term adverse health consequences. Globally, traffic crashes are a leading cause of death among young people, and the main cause of death among those aged 15-29 years. Road traffic injuries are currently estimated to be the eighth leading cause of death across all age groups globally.

Leveraging the tools and all the information nowadays available, an extensive analysis to predict traffic accidents and its severity would make a difference to the death toll. Analyzing a large range of factors, including weather conditions, locality, type of road and lighting among others, an accurate prediction of the severity of the accidents can be performed. Thus, trends that commonly lead to severe incidents can help identifying the highly severe accidents. This kind of information could be used by emergency services, to send the exact required equipment to the place of the accident, leaving more resources available for accidents occurring simultaneously. Moreover, this severe accident situation can be warned to nearby hospitals which can have all the equipment ready for a severe intervention in advance.

Consequently, road safety should be a prior interest for governments, local authorities and private companies investing in technologies that can help reduce accidents and improve overall driver safety.

## 1.2 Problem

Data that might contribute to determining the likeliness of a potential accident occurring might include information on previous accidents such as road conditions, weather conditions, exact time and place of the accident, type of vehicles involved in the accident, information on the users involved in the accident and the severity of the accident. This project aims to forecast the severity of accidents with previous information that could be given by a witness informing the emergency services.

## 1.3 Interest

Governments should be highly interested in accurate predictions of the severity of an accident, in order to reduce the response time of the emergency medical services and to make a more efficient use of the resources, and thus save a significant amount of lives. Others interested could be private companies investing in technologies aiming to improve road safety procedures.

# 2 Data

## 2.1 Data source

The data can be found in the following Kaggle data set [click here](#).

## 2.2 Feature Selection

The data is divided in 5 different data sets, consisting of all the recorded accidents in France from 2005 to 2016. The characteristics data set contains information on the time, place, and type of collision, weather and lighting conditions and type of intersection where it occurred. The places data set has the road specifics such as the gradient, shape and category of the road, the traffic regime, surface conditions and infrastructure. On the user data set it can be found the place occupied by the users of the vehicle, information on the users involved in the accident, reason of traveling, severity of the accident, the use of safety equipment and information on the pedestrians. The vehicle data set contains the type of vehicle, and the holiday one labels the accidents occurring in a holiday. All the data sets share the accident identification number.

An initial analysis of the data was performed for the selection of the most relevant features for this specific problem, reducing the size of the dataset and avoiding redundancy. With this process the number of features was reduced from 54 to 28.

[Link](#) to the feature selection notebook.

## 2.3 Description

The dataset that resulted from the feature selection consisted in 839,985 samples, each one describing an accident and 29 different features.

These features were the following:

From the characteristics dataset: lighting, localization, type of intersection, atmospheric conditions, type of collisions, department, time and the coordinates which are described in the Kaggle dataset [here](#). In addition, two new features were crafted, date to perform a seasonality analysis of the accident severity and weekend indicating if the accident occurred during the weekend or not.

Regarding the places dataset, the selected features were: road category, traffic regime, number of lanes, road problem, road shape, surface condition, situation, school nearby and infrastructure.

The users dataset was used to craft some new features:

- Number of users: total number of people involved in the accident.
- Pedestrians: whether there were pedestrians involved (1) or not (0).
- Critical age: whether there were users between 17 or 31 year old involved in the accident.
- Severity: Severity of the accident. Unscathed or light injury (0), hospitalized wounded or death (1)

The holiday dataset was used to add a last feature, labeling the accidents which occurred in a holiday.

## 2.4 Data Cleaning

The data cleaning is the process of giving a proper format to the data for its further analysis. The first step was to deal with missing values and outliers. Initially the latitude, longitude and road number were dropped from the data frame as more than a 50% of its values were NaN or 0 which is an outlier in this case.

Then keeping with replacing the missing values, the analysis was divided in two groups of features. The first group had in all features a label which described other cases, for instance the feature describing the atmospheric conditions had a value of 9 for any other atmospheric condition not labeled with the other 8 values. Therefore, the missing values and outliers were replaced with the other cases label for the features of atmospheric conditions, type of collision, road category and the surface conditions. For the second group of features instead, the distribution of their values was analyzed. Then two features were dropped, the infrastructures and reserved lanes, as the outliers represented more than 75% of its data. Finally with the rest of the features with missing values, the traffic regime, the number of lanes, the road problem and shape and the situation at the time of the accident, the NaN and outliers were replaced with the feature's most popular value.

Last format changes were performed to the school and department values. The school feature had all samples divided either in the 0 or the 100 values, thus all the 100 values were replaced with a 1. Similarly the department feature had an extra 0 added at the units position, so all values were divided by 10.

Regarding the type of the data, all features had a coherent data type except for the date feature which was needed with the string type. I used the `to_date` function of pandas to denote the date feature with the datetime type. After all, 24 features existed for the model building.

### 3 Exploratory Data Analysis

First, the distribution of the target's values was visualized. The plot confirmed that it is a balanced labeled dataset as the samples are divided 56-54 with more cases of lower severity. Then a seasonality analysis was performed, visualizing the global trend of daily accidents as well as the amount of accidents grouped by years, month of the year, and day of the week.

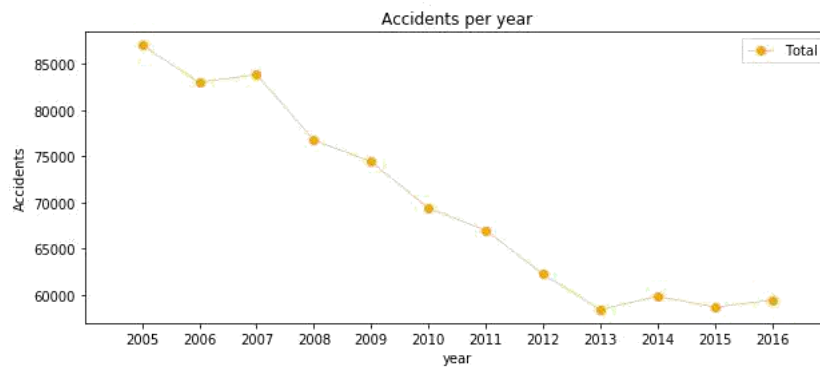


Figure 1: Lineplot of total amount of accidents per year.

The previous image show that the number of traffic accidents decreased over the years from 2005 to 2013, after which the trend became stable. Analyzing the yearly trend there is a seasonal pattern where the number of accidents increase around March and then again in September. This pattern can be seen in following two figures.

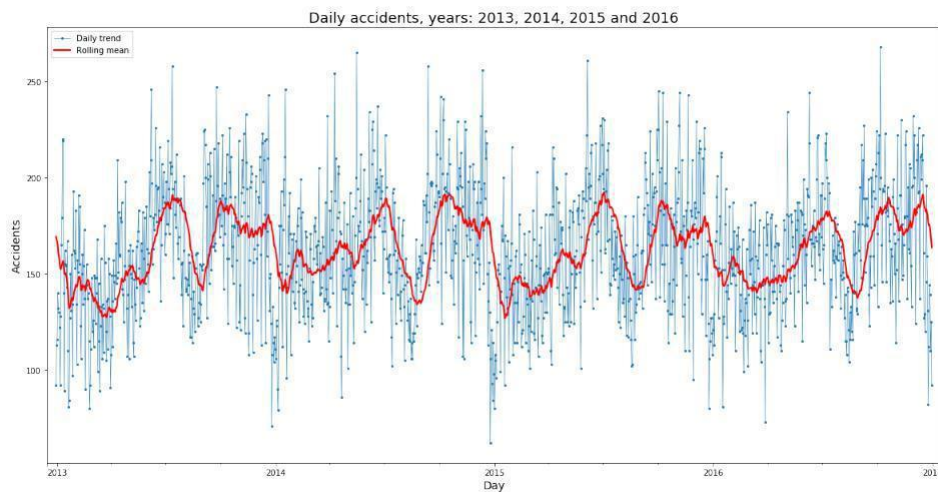


Figure 2: Lineplot of the amount of accident per day during the 2013, 2014, 2015 and 2016. The plot includes the rolling mean, with a window size of 30 days.

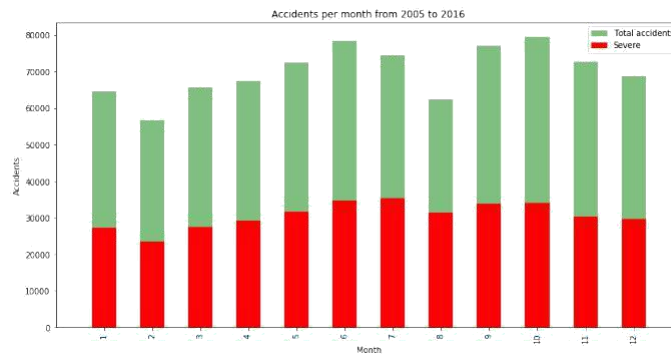


Figure 3: Barplot: Amount of accident per month from 2005 to 2016.

Regarding the day of the week there is not a significant difference between them, Figure4. There is a steady trend during the week with more accidents on Friday, and Sunday is the day with less recorded accident of all.

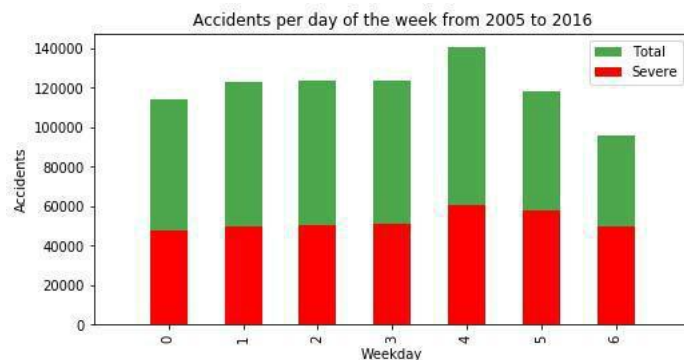


Figure 4: Barplot: Amount of accident per day of the week from 2005 to 2016.

Lastly analyzing the accidents per hour, there are clearly two spikes, one at 8 am, the time people go to work and another one between 5 p.m. and 6 p.m., time when people return home. The number of accidents decreases between these two spikes, nothing unusual but it proves there is a pattern here.

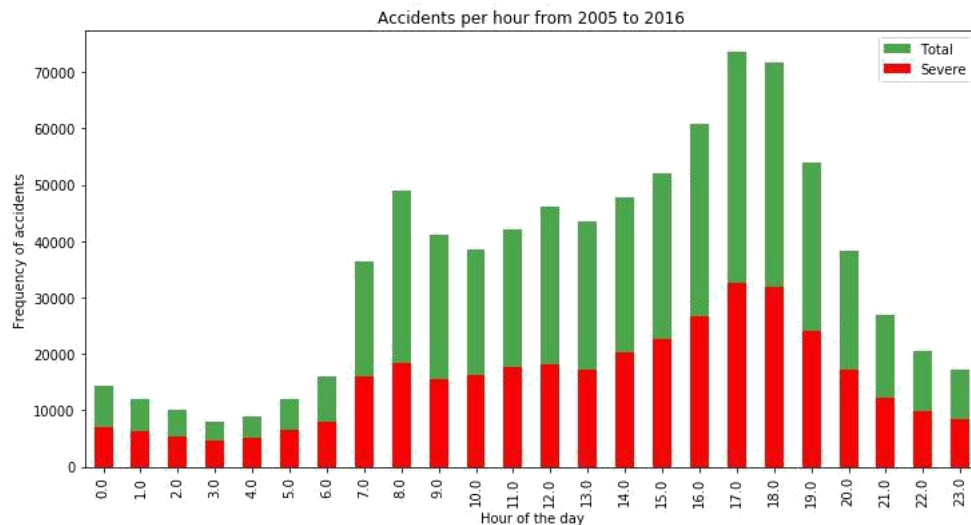


Figure 5: Lineplot of total amount of accidents per year.

The trend of highly severe accidents is proportional to the global trend, for both the accidents divided per month of the year and per day of the week. Same thing happens with the amount of highly severe accidents by hour of the day as we can see on Figure:5. One aspect to highlight from the hourly trend is that the proportion of severe accidents from noon to morning is higher, to be precise, the percentage of severe accidents from 9pm to 6am is 50.67% of the total amount of accidents occurring between these hours, while from 7am to 8pm is 42.41%. Due to the results of the former analysis, to features were added; month and day as the day of the month.

The next statistical analysis was the correlation of the features with the severity of an accident. The Pearson correlation showed weak or null correlation with all features. Further visualizations were performed for a better understanding of the data. Some conclusions of this analysis were for instance that accidents involving people above 84 years old tend to have a high severity.

## References

- [1] Alexander Popov. Road Traffic Injuries. WHO, Global Health Observation Data, 2016. [https://www.who.int/health-topics/road-safety#tab=tab\\_1](https://www.who.int/health-topics/road-safety#tab=tab_1)

-