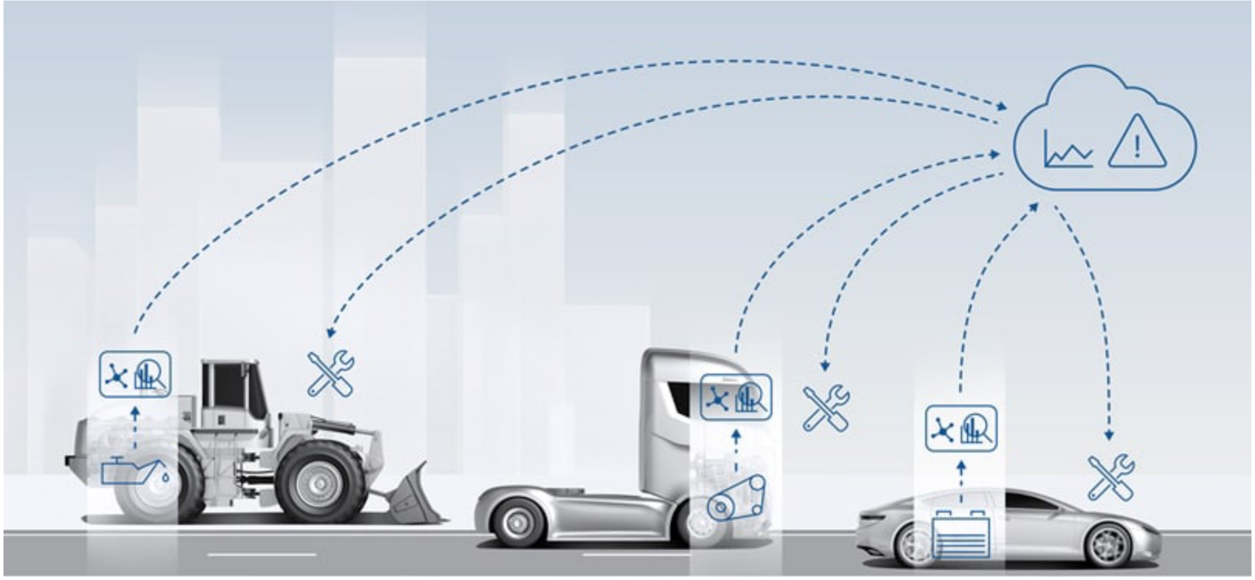
Applied Data science Captsone

# Car Accident Severity Prediction

## Data Science Professional Certificate



## Table of content

1. Introduction
2. Data
3. Data Cleaning
4. EDA
5. Data Preparation
6. Model Development
7. Results

# 1. Introduction

### 1.1 Background

Every year car accidents cause hundreds of deaths world wide. Road traffic injuries are currently estimated to be the eighth leading cause of death across all age groups globally, and are predicted to become the seventh leading cause of death by 2030.

Road traffic crashes result in the deaths of approximately 1.35 million people around the world each year and leave between 20 and 50 million people with non-fatal injuries. More than half of all road traffic deaths and injuries involve vulnerable road users, such as pedestrians, cyclists and motorcyclists and their passengers. The young are particularly vulnerable on the world's roads and road traffic injuries are the leading cause of death for children and young adults aged 5-29. Young males under 25 years are more likely to be involved in road traffic crashes than females, with 73% of all road traffic deaths occurring among young males in that age. Developing economies record higher rates of road traffic injuries, with 93% of fatalities coming from low- and middle- income countries.

In addition to the human suffering caused by road traffic injuries, they also incur a heavy economic burden on victims and their families, both through treatment costs for the injured and through loss of productivity of those killed or disabled. More broadly, road traffic injuries have a serious impact on national economies, costing countries 3% of their annual gross domestic product. Measures proven to reduce the risk of road traffic injuries and deaths exist and the 2030 Agenda for Sustainable Development has set ambitious targets for reducing road traffic injuries.

Leveraging the tolls and all the information nowadays available, an extensive analysis to predict traffic accidents and its severity would make a difference to the death toll. Analysing a significant range of factors including weather conditions, locality, type of road and lighting among others, an accurate prediction of the severity of the accident can be performed. Thus, trends that commonly lead to severe traffic incidents can help identify the highly severe accidents.

## 1.2 Audience

This kind of information could be used by emergency services to send the exact required staff and equipment to the place of the accident., leaving other resources for accidents occurring simultaneously.

Hospitals can be prepared if such accident information is forecasted to ensure the appropriate needed equipment and resources can be made available.

In addition, this knowledge of a severe accident situation can be warned to drivers so that they would drive more carefully or even change their route if it is possible or to hospital which could have set everything ready for a severe intervention in advance.

Governments should be highly interested in accurate predictions of the severity of an accident, in order to reduce the time of arrival and thus save a significant amount of people each year. Others interested could be private companies investing in technologies aiming to improve road safeness.

### 1.3 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 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.

### 1.4 Interest

Government should be highly interested in accurate predictions of the severity of an accident, in order to reduce the line of arrival and to make a more efficient use of the resources, and thus save a significant amount of people each year. Others interested could be private companies investing in technologies aiming to improve road safety.

## 2. Data

### 2.1 Data source

The data has been picked from Kaggle data set available through thelink

### 2.2 Selecting the features

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 consists 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.
In the user data set it can be found the place occupied by the users of the vehicle, information on the user involved in the accident, purpose of travel, severity of the accident, the use of safety equipment and information on the pedestrians. The vehicle data set contains the flow and type of vehicle and the holiday once labels the accidents occuring in a holiday. All 5 data sets share the accident identification number.
An initial analysis of the data was performed for the selection of the most relevant features for the specific problem, reducing the size of the dataset and avoiding redundancy. The number of features was reduced from 54 to 28 in this process.

## 2.2 Selecting the features

The original data for this project comes from the Kaggle data set. In a previous notebook, Selecting Features, I performed a selection of the most relevant features for the prediction of traffic accident severity.

The features of the dataset resulting are the following:

From the characteristics dataset, I will keep the features: lighting type of intersection, atmospheric conditions, type of collisions, department, address, time and the coordinates. 2 new features were added to this original dataset: "date" and "weekend" indicating if the accident occurred during the weekend or not.

In the places dataset, the selected features were: road categories, traffic regime, number of traffic lanes, road profile, road shape, surface condition, situation, school nearby and infrastructure.

The users dataset, following features were crafted::

- Number of users: : total number of users involved in the accident.
- Pedestrians: Whether there are pedestrians involved or not.
- critic_age: If there is any user in between 17 and 31 y.o.
- severity : maximum gravity suffered by any user involved in the accident:
    - 0 = Unscathered or Light injury

- 1 = Hospitalized wounded or Death

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

## 2.3 Data cleaning

The data cleaning is the process of giving a proper format to the data for its further analysis. The 1st step was to deal with missing data and outliers. Initially the latitude and road number were dropped from the data frame as more than 50% of its value were 0.
Then keeping with the replacing the missing values, the analysis was divided in 2 groups of features.

1. The 1st group had in all features a label which described other cases, example the feature describing the atmospheric conditions had a value of 9 for any other atmospheric conditions not labeled with the other 8 values. Therefore the outlier and missing information were replaced with the other cases label for the feature of atmospheric condition, type of collisions, road category and the surface conditions.
2. The 2nd group of features instead, the distribution of their values was analysed. Then 2 features were dropped, as the outlier represented more than 75% of the data. Finally with the rest of the missing values, the traffic regime, # of lanes, road profile and the shape and the situation of the accident, the 0 and outliers were replaced with feature's most popular value.
3. Last format changes were performed to the school and department values. The school feature had all the samples divided either in the 0 or 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 the values were divided by 10.

Regarding the type of data, all features had a coherent data type except for the date feature which was defined with the string type.

# 3. Exploratory Data Analysis (EDA)

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. Thena seasonality analysis was performed, visualizing the global trend of daily accidents as well as the amount of accidents grouped by year, month and day of the week.

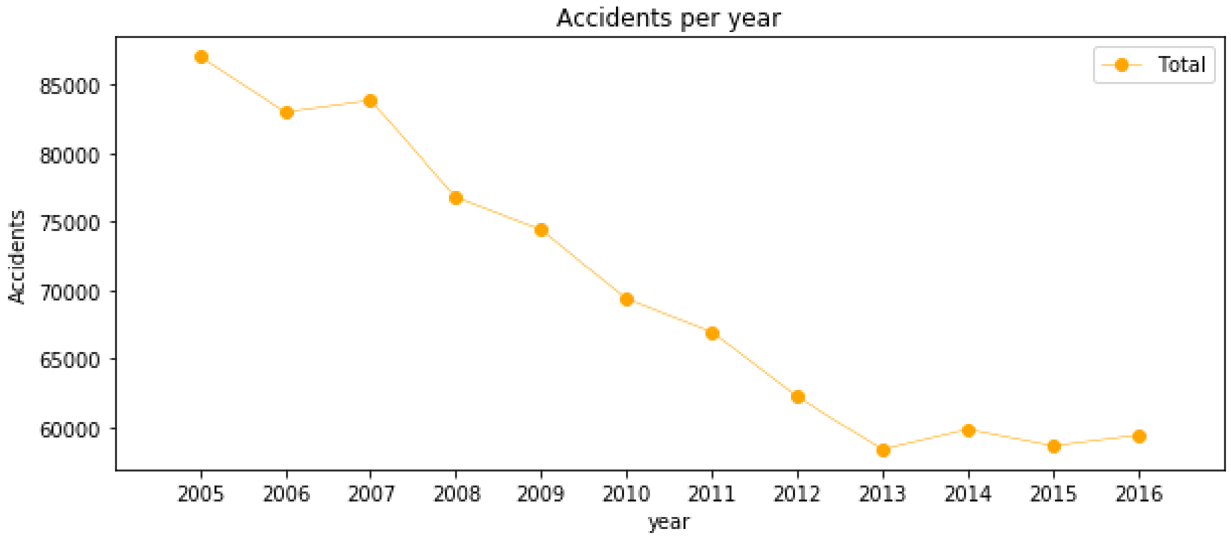


Figure 1. Lineplot of total accidents per year

The previous image shows 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 around September, This pattern can be seen in the next 2 images:

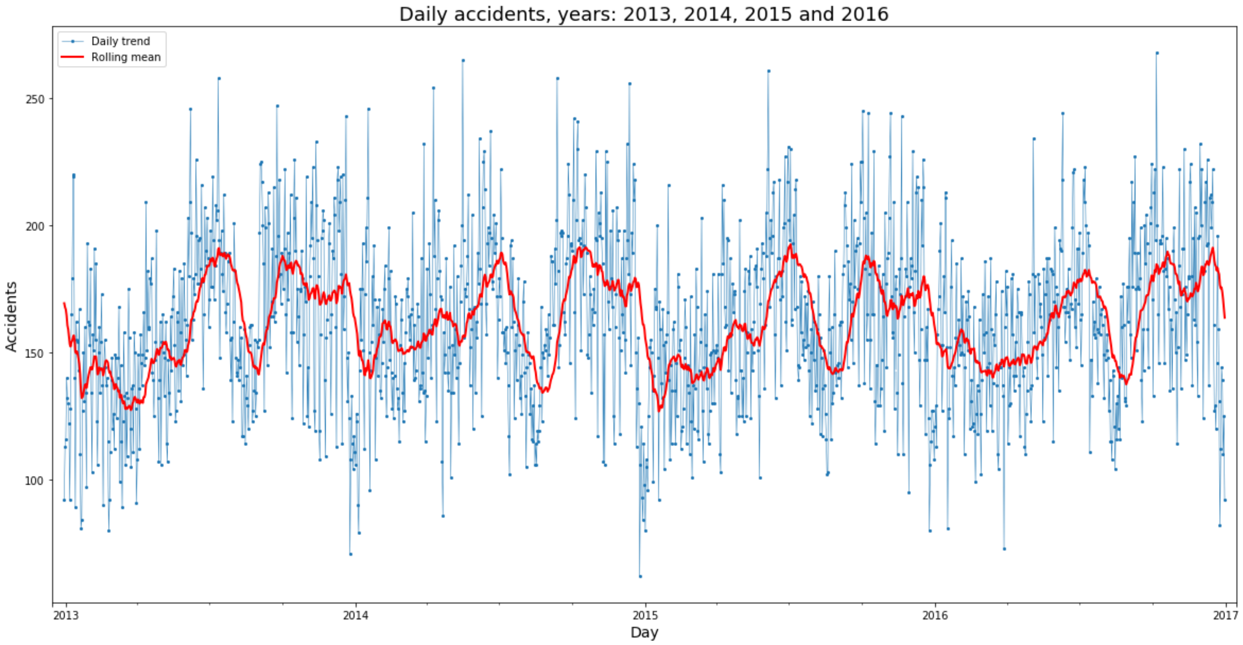


Figure 2: Lineplot of accidents per day during 2013 -16.

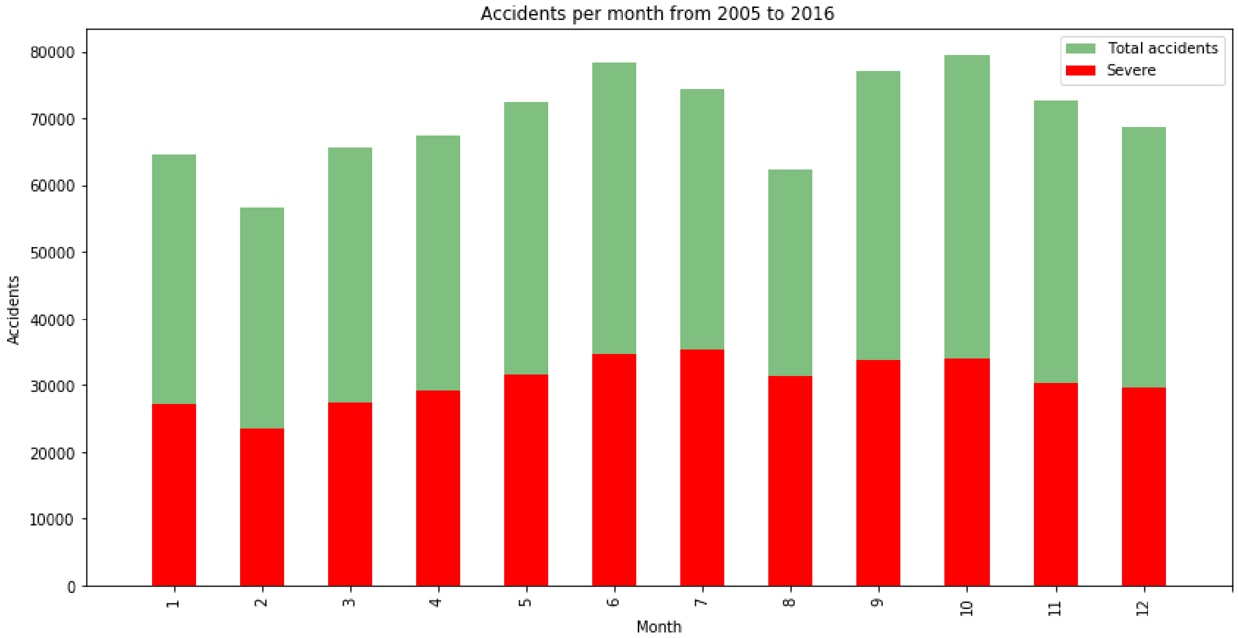


Figure 3: Barplot: Accidents per month from 2005 to 2016

Regarding the day of the week there is not a significant difference between them as seen in Figure 4. There is a steady trend during the week and more on Friday and Saturday with SUnday having the least number of recorded accidents.

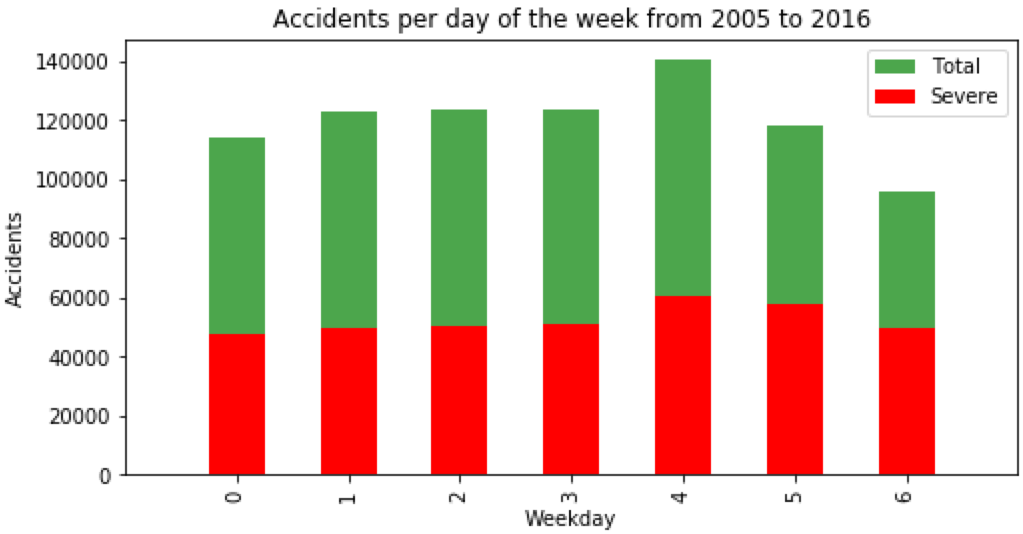


Figure 4: Barplot: Accidents per day of the week from 2005-16

Lastly analyzing the accidents per hour, there are clearly 2 spikes, one at 8AM and between 5-6PM which office commute hours. The accidents decrease between these spikes, thus showing a pattern.

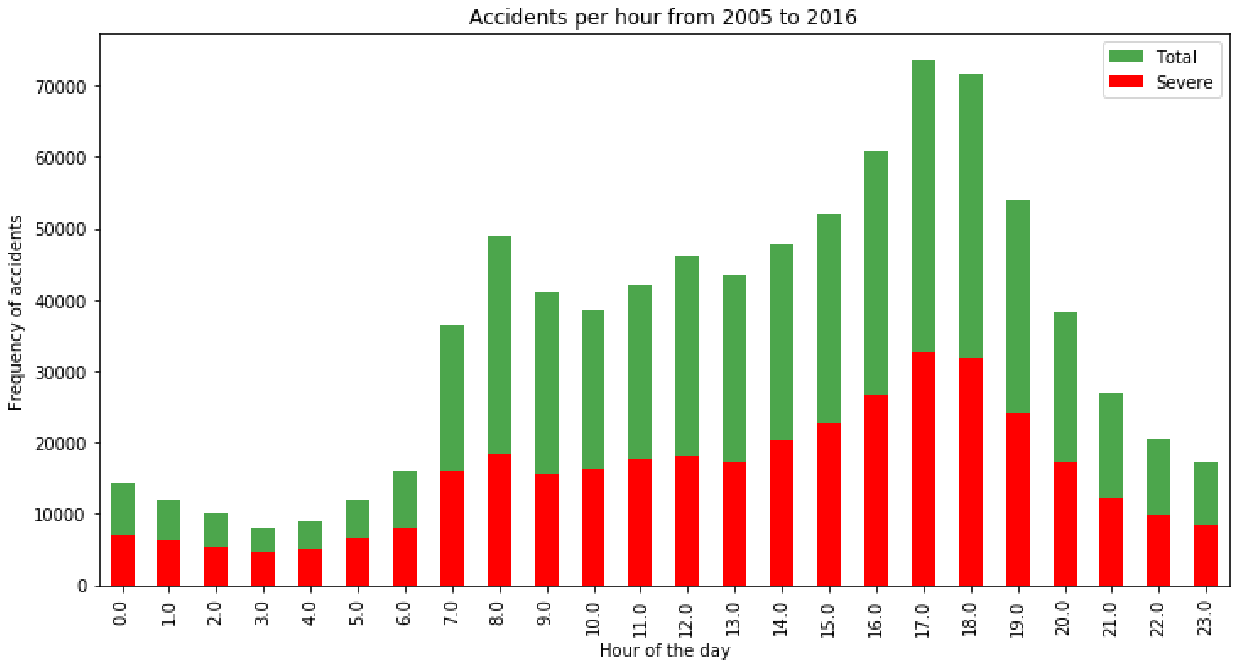


Figure 5: Lineplot of total accidents per year

The trend of highly severe accidents is proportional to the global trend for both accidents divided per month of the year and per day of the week. The same is noticed for severe accidents by hour of the day as in FIgure 5. One aspect to highlight from the hourly trend is that the proportion of severe accidents from noon to morning is higher - the %age of severe accidents from 9PM to 6AM is 50.67% of the total accidents occurring between these hours while from 7AM to 8PM is 42.41%. Due to the results, 2 features were added - month and day of the month.

The next statistical analysis was the correlation of features with severity of an accident. The Pearson correlation showed weak or null correlation  with all features, MOre visualizations were done and some conclusions of this analysis were that the accidents involving people above 84 years tend to be of high severity.

# 4. Predictive Modeling

Different classification algorithms have been tuned and built for the prediction of the level of accident severity. These algorithms provided a supervised learning approach predicting with certain accuracy and computational time. These 2 properties have been compared in order to determine the best suited algorithm for this specific problem.

The 83k rows were split 80/20 between training and test sets, afterwards 80/20 split was performed among the training samples creating the validation set for the development of the model. Then the data was standardized giving zero mean and unit variance to all features.

3 different approaches were used:

1. Decision tree - Random forest
2. Logistics REgression
3. K nearest neighbour
4. Supervised vector machine

Same process was followed for each algorithm. With the train and validation sets the best hyperparameters were selected and using the  test set the accuracy and time for the development of the models were calculated.

The decision tree model was upgraded to Random forest. With default random forest, features were sorted by impurity based importance in the prediction of the severity. Thus the 10 least important features were dropped to improve performance. After evaluating the parameters for each algorithm, these were the models.

- Random forest: 10 Decision tree, max depth 12 features and MAx 8 features compared for the split.
- LR: c = 0.001
- KNN: k = 16

# 5. Results:

The metrics used to compare the accuracy of the models are the Jaccard score, f1 score, Precision and Recall. The table reports the results:

| Algorithm | Jaccard | F1 score | Precision | Recall | Time |
|---|---|---|---|---|---|
| RF | 0.73 | 0.72 | 0.72 | 0.72 | 18.58 |
| LR | 0.56 | 0.43 | 0.53 | 0.56 | 4.48 |
| KNN | 0.75 | 0.75 | 0.76 | 0.76 | 20 |

The recall is more important than the precision as the high recall will favor that the required resources will be equipped up to the severity of the accident. The LR, KNN models have similar accuracy, but the time for LT is the best.

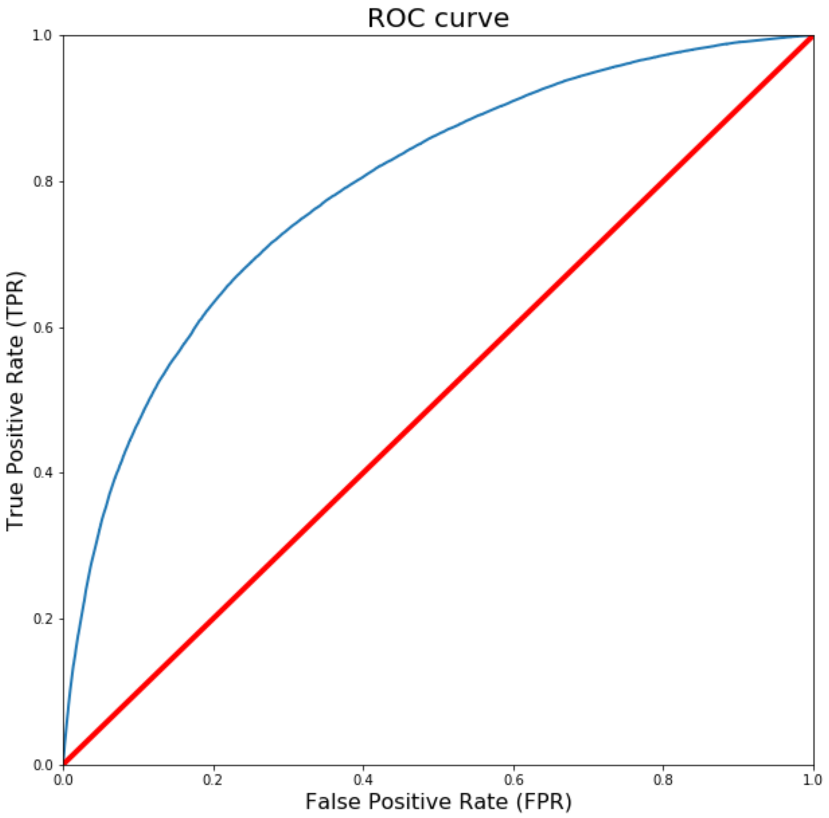


Figure 6: ROC  curve from RF model

False positive rates are less important than higher true positive rates. It is more important to properly predict the high severity accidents properly, so as to avoid them.

## 6. Conclusion

I analysed the relationship between severity of an accident and some characteristics which describe the situation that involved the accident. Earlier, it was felt factors such as atmospheric conditions, lighting or being a holiday would influence more but it was found that department, day of the time, the road category and type of collision among the most important features.

I compared 3 different classification models to predict if an accident would have a high or low severity. These models can have multiple applications in real life. Example: if emergency services have a application with some features like date, time, department and then with the information from witness calling to report an accident, they could predict the severity before actually going there and be equipped with the right resources.

## 7. Observations

About 65% accuracy was achieved and there is significant variance that was not predicted by the models used. Some factors that can also impact  are: travel time, speed, etc. which could have increased the accuracy.

The next action could be to add an accident prediction model could be to predict locations most vulnerable to accidents.

## 8. References

Kaggle: Data sourced :
https://www.kaggle.com/ahmedlahlou/accidents-in-france-from-2005-to-2016