

Ravi Singh

October 2020

1 Introduction

1.1 Background

Car accidents keep increasing year after year causing a large number of deaths worldwide. According to a study conducted by the (WHO) there were a total of 1.35 million road traffic deaths worldwide in 2016, with a big number sustaining serious injuries and living with long-term injuries. Worldwide, traffic crashes are the dominant cause of death among people, and also the main cause of death among aged groups 15-29 years. Road traffic injuries are estimated to be the 8th leading cause of death across varying age groups and are predicted to become the 7th highest cause of death by 2030[1].

Utilizing the tools and all the technology available, extensive dataset analysis is possible in order to predict traffic accidents and its severity would make a huge difference to the death toll. Analyzing a significant range of factors, including locality, type, weather conditions of road and lighting among others, an accurate estimate of the severity of the accidents is performed. This, trends lead to severe traffic incidents that can help to identify the highly severe accidents. The information can be used by public/non-profit services to allocate and send the required resources and equipment to the location of the accident, leaving more resources available for accidents occurring simultaneously. Moreover, this scenerio can be communicated to nearby hospitals that can have all equipment ready for a severe intervention in advance.

Consequently, road safety should be a priority of every government, and non-government organization investing in technologies which can help reduce accidents and improve overall driver safety.

1.2 Problem

Dataset which will contribute in determining the potential of a life threatening accident might be including 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 o course 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 must be highly interested to accurately predict any of the severity of an accident, so as to reduce the time of response to the accident and to make a more client 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 safeness.

2 Data

2.1 Data source

The data can be found in the following Kaggle data set given in the notebook.

2.2 Feature Selection

The dataset 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 for traveling, the severity of the accident, the use of safety equipment and information on the pedestrians. The vehicle

data set contains the ow and type of vehicle, and the holiday one labels the accidents occurring in a holiday. All ve data sets share the accident identifications 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, [click here](#). With this process, the number of features was reduced from 54 to 28.

2.3 Description

The original data for this project comes from the following Kaggle data set. In a previous notebook, Feature Selection, 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:

In the characteristics dataset, I will keep the features: "lighting", "localisation"(agg), "type of intersection", "atmospheric conditions", "type of collisions", "department", "adress", "time" and the coordinates. Two new features were added from this original dataset: "date" and "weekend" indicating if the accident occurred during the weekend or not.

In the places dataset, I will keep only the features: "road categorie", "traffic regime", "number of traffic lanes", "road profile", "road shape", "surface condition", "situation", "school nearby" and "infrastructure".

Using the users dataset, I have created the following features:

num_us: total number of users involved in the accident.

ped: Wether there are pedestrians involved or not.

critic_age: If there is any user in between 17 and 31 y.o.

sev : maximum gravity suffered by any user involved in the accident:

0 = Unscathered or Light injury

1 = Hospitalized wonded or Death

I used the holiday dataset to craft a new feature indicating the accident occurred during a holiday.

2.4 Data Cleaning

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

Then keeping with replacing the missing values, the analysis is divided in two groups of features. The first group that had described all features of a label that 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 are 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 and shape and the situation at the time of the accident, the NaN and outliers were replaced with the feature's most popular value.

Final format changes is 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 unit's 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 defined with the string type. I used the `to_datetime` function of pandas to define the date feature with the DateTime type. After all, 24 features remained.

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.

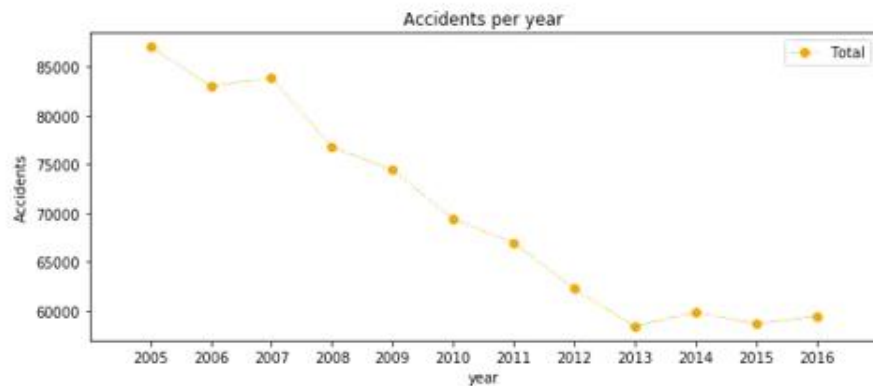


Figure 1: Lineplot of total amount of accidents per year.

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 increases around March and then again in September. This pattern can be seen in the following two figures.

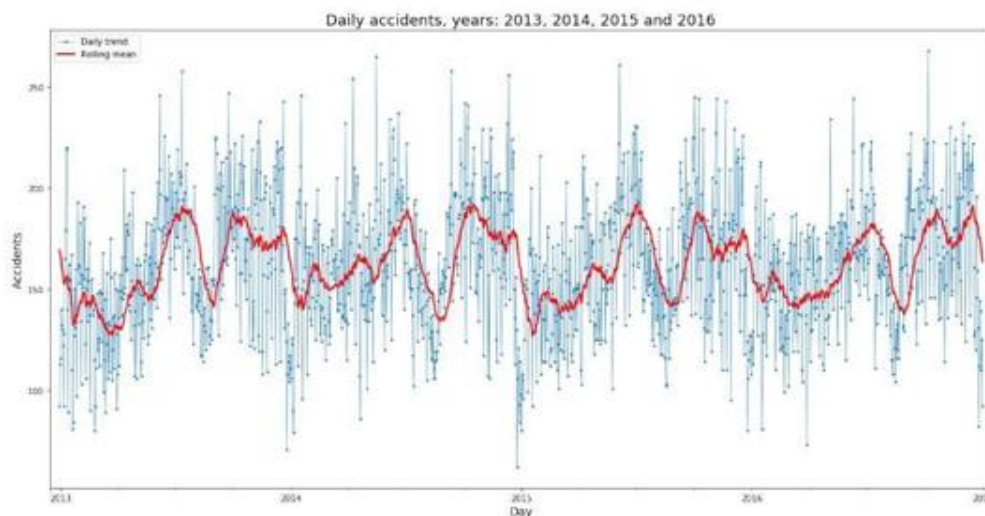


Figure 2: Lineplot of the amount of accident per day during the 2013, 2014, 2015 and 2016. The plot includes the rolling mean, with a window size of 30 days.

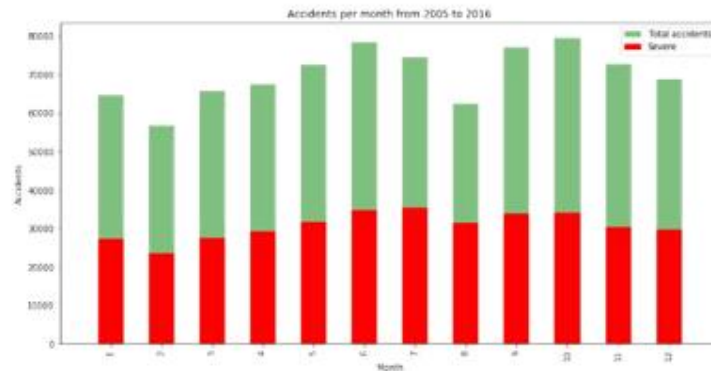


Figure 3: Barplot: Amount of accident per month from 2005 to 2016.

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 the less recorded accident of all.

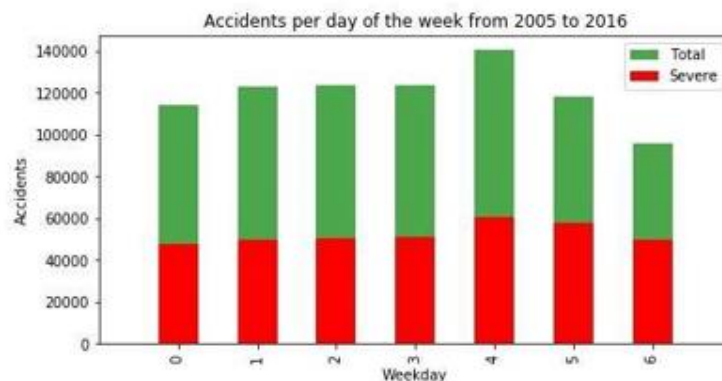


Figure 4: Barplot: Amount of accident per day of the week from 2005 to 2016.

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

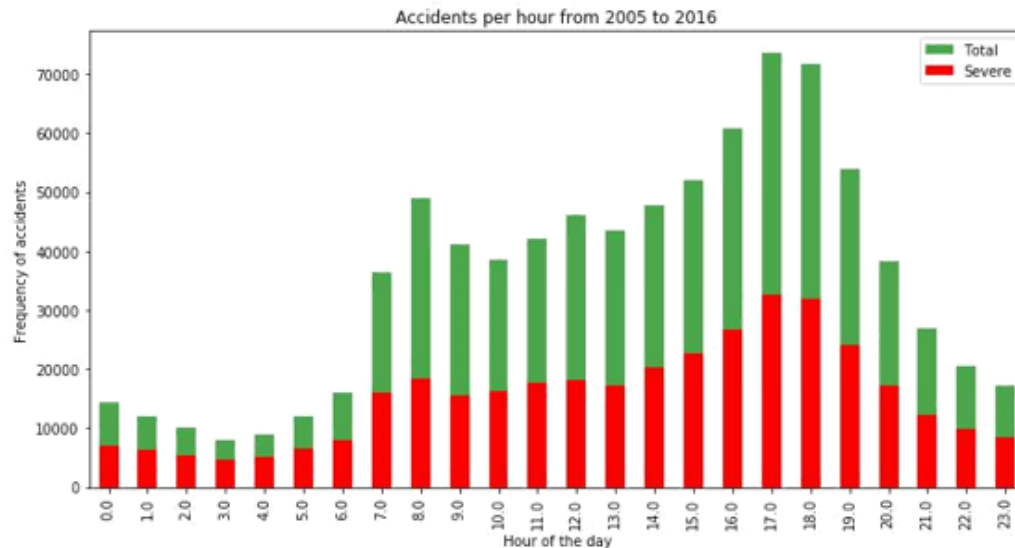


Figure 5: Lineplot of total amount of accidents peryear.

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 9:00 pm to 6:00 am 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.

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 two properties have been compared in order to determine the best suited algorithm for this specific problem.

Firstly, the 839,985 rows were split 80/20 between the training and test sets; afterwards an additional 80/20 split was performed among the training samples creating the validation set for the development of the models. Then the data was standardized giving zero mean and unit variance to all features.

Four different approaches were used:

Decision Tree, Random Forest Logistic Regression

K-Nearest Neighbour Supervised Vector Machine

The same modus operandi was performed with each algorithm. With the train and validation sets the best hyperparameters were selected and using the test set the accuracy and computational time for the development of the models were calculated.

The decision tree model was upgraded to the random forest. With the default random forest the features were sorted by impurity-based importance in the prediction of the severity. Thus, the 10 least important features were dropped to decrease the computation complexity for the KNN and SVM models. Keeping with 13 features the accuracy stayed the same and the computational time decreased significantly. After evaluating the parameters for each algorithm, these were the models.

Random Forest: 10 decision trees, maximum depth of 12 features and maximum of 8 features compared for the split.

Logistic Regression: $c=0.001$. KNN: $k=16$

SVM: size of the training set= 75,000 samples.

The following visualizations show how the parameters for KNN and SVM models were selected. The SVM model is computationally inefficient with huge sample sets. Therefore, an equilibrium between accuracy and computational time was found evaluating different training sizes. The training set was reduced from 537,590 to 75,000 rows. On Figure:7, the accuracy is increasing as the training size does, however Figure:9 shows how this comes with an important increasing of the computational time.

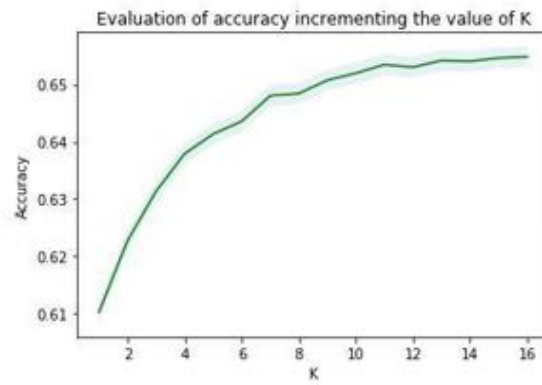


Figure 6: Accuracy of KNN models increasing the value of K.

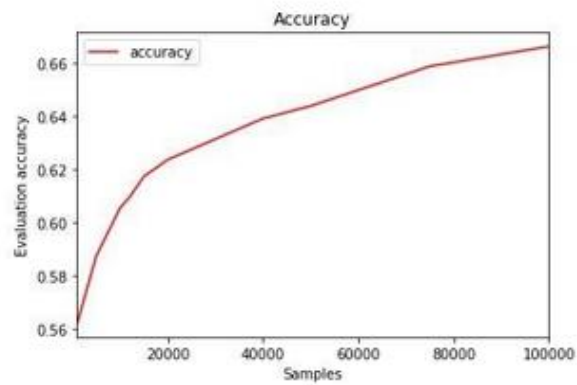


Figure 7: Accuracy of SVM increasing the training sample's size.

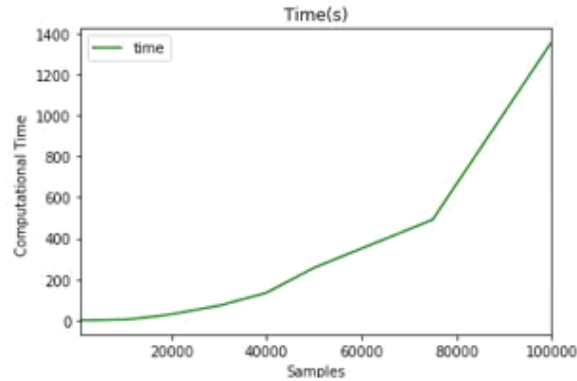


Figure 8: Computational time of SVM increasing the training sample's size.

5 Results

The metrics used to compare the accuracy of the models are the Jaccard Score, f1-score, Precision¹ and Recall². This table reports the results of the evaluation of each model.

Algorithm	Jaccard	f1-score	Precision	Recall	Time(s)
Random Forest	0.722	0.72	0.724	0.591	6.588
Logistic Regression	0.661	0.65	0.667	0.456	6.530
KNN	0.664	0.66	0.652	0.506	200.58
SVM	0.659	0.65	0.630	0.528	403.92

In this case, the recall is more important than the precision as a high recall will favor that all required resources will be equipped up to the severity of the accident. The logistic regression, KNN, and SVM models have similar accuracy, however the computational time from the regression is far better than the other two models. With no doubt the Random Forest is the best model, in the same time as the log. res. it improves the accuracy from 0.66 to 0.72 and the recall from 0.45 to 0.59.

¹Proportion of predicted severe accidents that were truly severe

²Proportion of truly severe accidents that were properly predicted

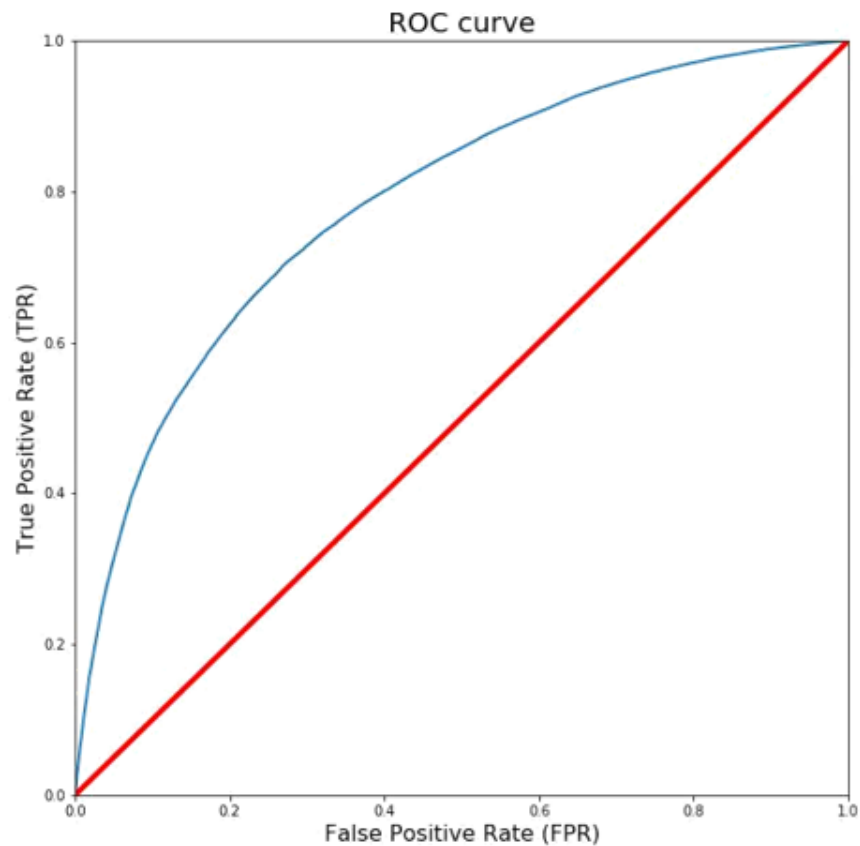


Figure 9: Representation of the ROC curve from the results of the Random Forest model.

I also evaluated the best model using their ROC curves. In this particular problem, lower false positive rate is less important than a higher true positive rate. In other words, it is more important to properly predict the high-severity accidents properly, if there is room for doubt it is better to prevent.

6 Conclusion

In this study, I analyzed the relationship between severity of an accident and some characteristics which describe the situation that involved the accident. Initially I thought that features such as atmospheric conditions, the lighting or being a holiday would be the most relevant ones, yet I identified the department, the day and time of the accident, the road category and type of

collision among

the most important features that affect to the gravity of the accident. I built and compared 4 different classification models to predict whether an accident would have a high or low severity. These models can have multiple application in real life. For instance, imagine that emergency services have a application with some default features such as date, time and department/municipality and then with the information given by the witness calling to inform on the accident they could predict the severity of the accident before getting there and so alert nearby hospitals and prepare with the necessary equipment and staff. Also by identifying the features that favor the most the gravity of an accident, these could be tackled by improving road conditions or increasing the awareness of the population.

6 Observation

I was able to achieve 68% accuracy in the. However, there was still significant variance that could not be predicted by the models in this study. I think other features like speed or uninterrupted time of traveling could be used to predict a more accurate classification. These are characteristics that may be impossible of knowing right now, but at the incredible pace that technology is evolving nowadays, soon cars will be able to track them so that the emergency services could use them.

One problem I think these features had is that the target of this classification problem was simplified to two different classes, low and high severity. Labeling severity with a range of punctuation from 0 to 100, for instance, could allow the possibility of developing regression model.

The next step on this problem could be to add a accident prediction model able to not just predict the accuracy but also the critical time and spots where potential accidents can occur in advance.

References

[1] Alexander Popov. Road Traffic Injuries. WHO, Global Health Observation Data, 2016.
<https://www.who.int/health-topics/road-safety#tab=tab1>