PREDICTING TRAFFIC ACCIDENT SEVERITY

SWATHI R,

OCTOBER , 2020

1 INTRODUCTION

1.1 BACKGROUND

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 crashes are becoming a global health crisis and, as such, require comprehensive measures to prevent them, including a better understanding of the social impacts of road-related deaths and injuries.

Analyzing a significant range of factors, including weather conditions, special events, road works, traffic jams among others, an accurate prediction of the severity of the accidents can be performed.

These insights, could allow law enforcement bodies to allocate their resources more effectively in advance of potential accidents, preventing when and where a severe accidents can occur as well as saving both, time and money. 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.

This project consists of several parts divided in two different notebooks.

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 o course the severity of the accident. This projects aims to forecast the severity of accidents with previous information that could be given by a witness informing the emergency services.

2 Data

2.1 DATA SOURCE

The data can be found in the following Kaggle data set

2.2 DATA 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 species such as the gradient, shape and category of the road, the 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 follow and type of vehicle, and the holiday one labels the accidents occurring in a holiday. All 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 OF DATA

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

These features where 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 where: road category, traffic regime, number of traffic lanes, road profile, 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 years old. Involved in the accident.

Severity: maximum gravity by any user involved in 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.

3 EXPLORATORY DATA ANALYSIS

First, the distribution of the target's values was visualized. The plot corned 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.

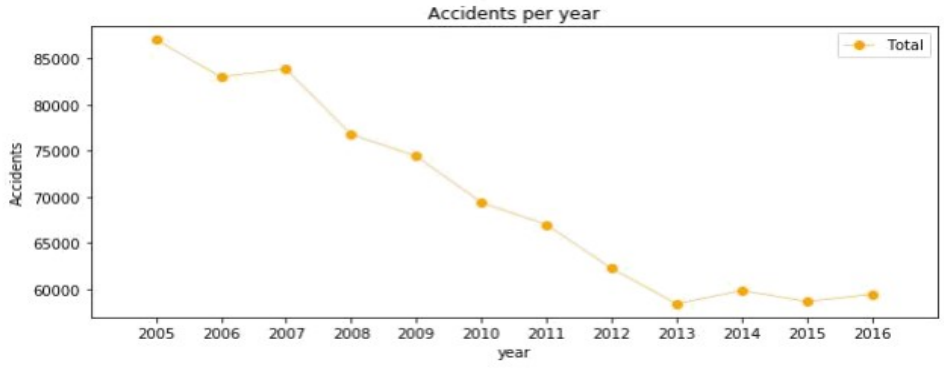

Figure 1: Lineplot of total amount of accidents per year.

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.

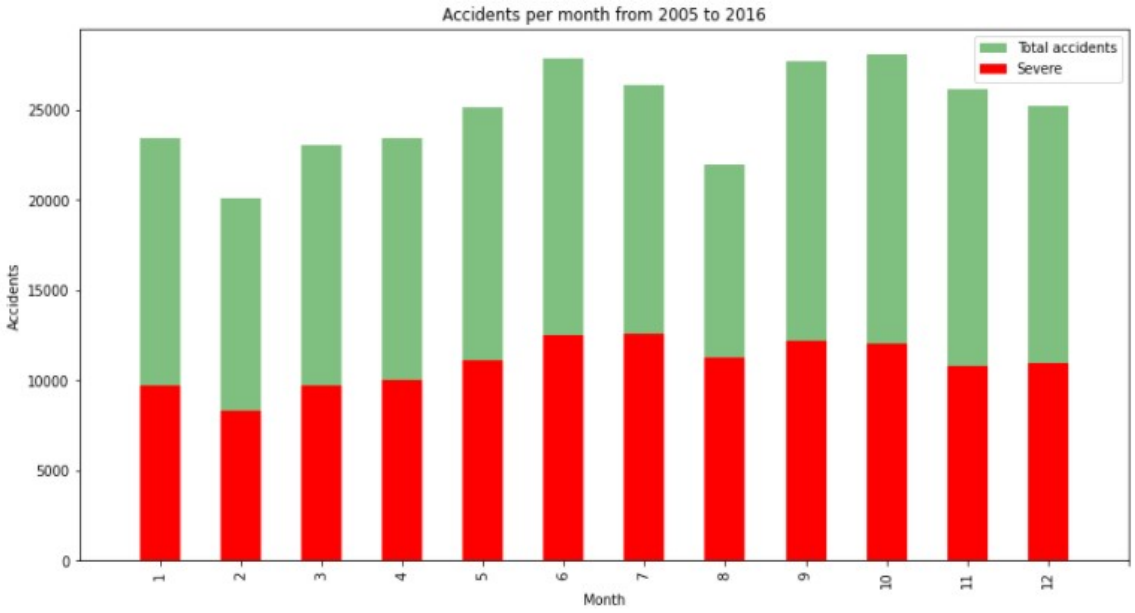

Figure 2 : Bar plot of Amount of accident per month from 2005 to 2016.

Regarding the month of the year there is not a significant difference between them, Figure2. There is a steady trend during the year with more accidents on July, September and October is the month with more recorded accident of all and February and August is the month with less recorded accident.

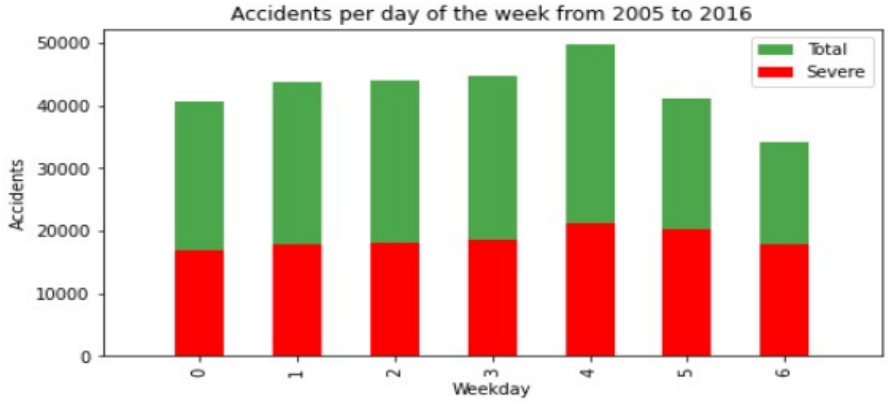

Figure 3 : Bar plot of Amount of accident per week from 2005 to 2016.

Regarding the day of the week there is not a significant difference between them, Figure3. There is a steady trend during the week with more accidents on Thursday, and Friday .Sunday is the day with less recorded accident of all.

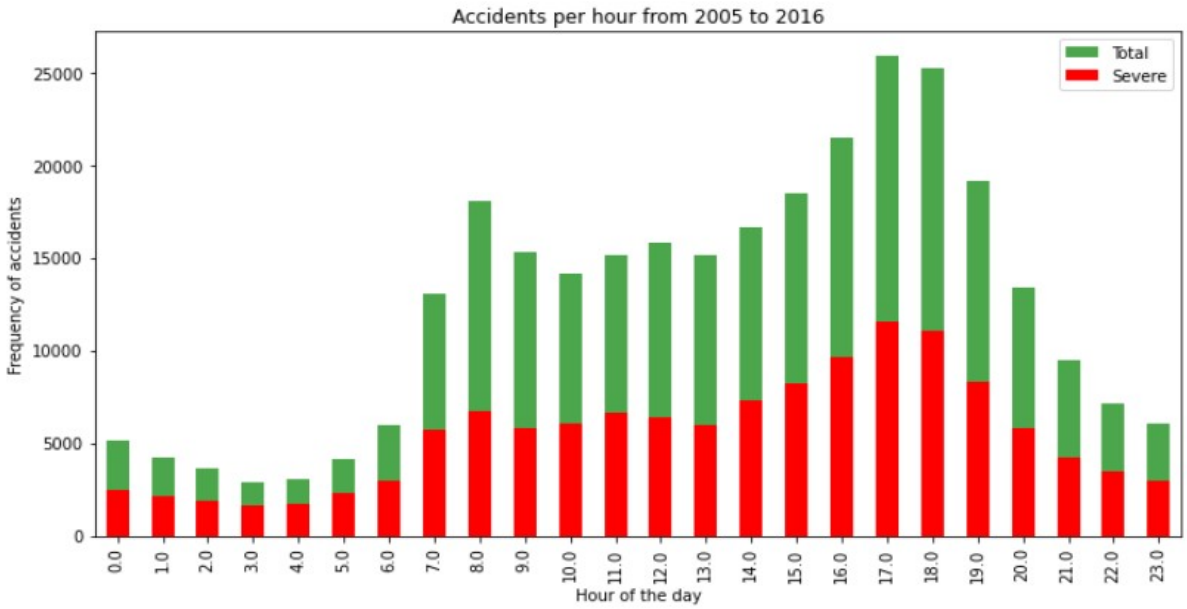

Figure 4 : Bar plot of Amount of accident per day from 2005 to 2016.

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 49.87% of the total amount of accidents occurring between this hours, while the percentage of deathly accidents from 7am to 8pm is 42.61%.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.

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 7 approach predicting with certain accuracy and computational time. These two properties have been compared in order to determine the best suited algorithm for his specific problem.

Firstly, the 839.985 rows where 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 Neighbor
- Supervised Vector Machine

The same modus was performed with each algorithm. With the train and validation sets the best hyper parameters 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 incident with huge sample sets. Therefore, equilibrium between accuracy and computational time was a found evaluating different training sizes. The training set was reduced from 8 537,590 to 75,000 rows.

On Figure: 5, the accuracy is increasing as the training size does, however Figure: 6 shows how this comes with an important increasing of the computational time.

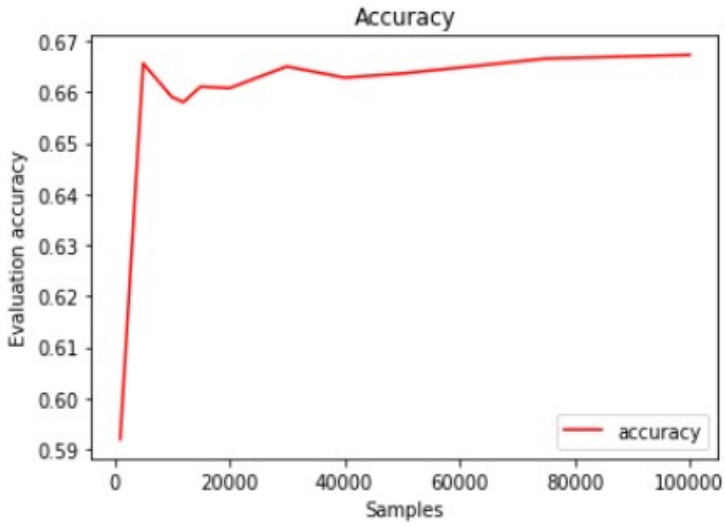

Figure 5: Evaluation Accuracy

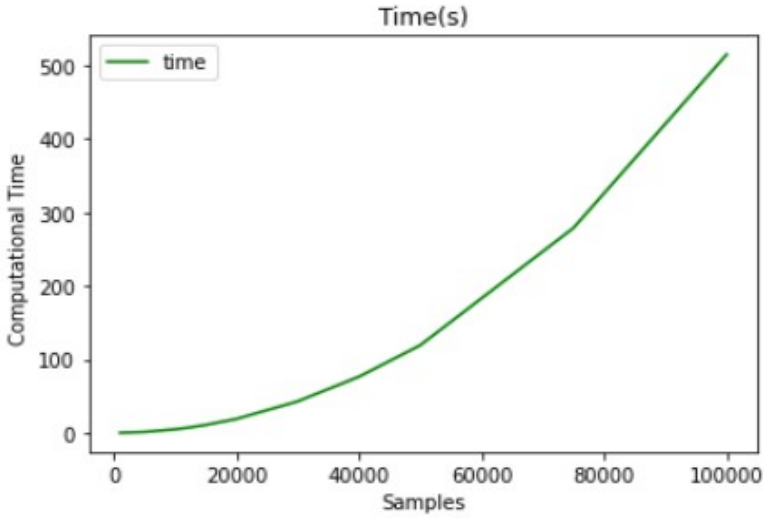

Figure 6: Computational Time

5 RESULTS

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

ALGORITHM	JACCARD	F1-SCORE	PRECISION	RECALL	TIME(S)
RANDOM FOREST	0.7150	0.77	0.71	0.84	2.6406
LOGISTIC REGRESSION	0.660	0.73	0.66	0.82	1.37
KNN	0.66	0.733	0.67	0.79	22.95
SVM	0.660	0.73	0.6	0.82	273.70

For this specific problem precision means the % of predicted severe accidents that were truly severe. The recall instead, is the % of truly severe accidents that were properly predicted. For this specific problem, 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.

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 0.71 and the recall from 0.79 to 0.84.

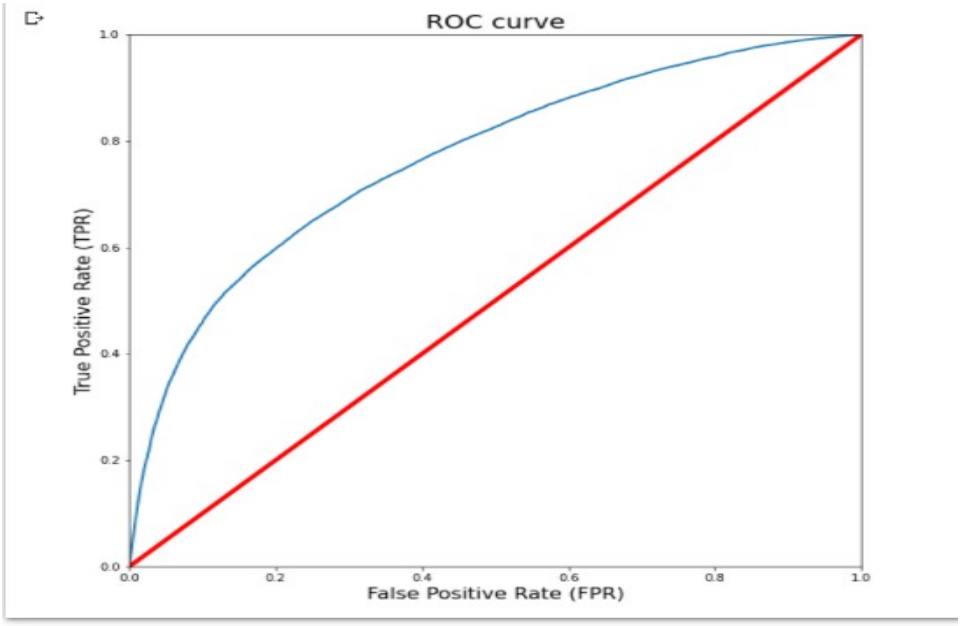

Figure 9: Representation of the ROC curve from the results of the Random Forest model.

I also evaluated the best model using their ROC curves. In this particular problem, lower false positive rate is less important than 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 11 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 applications 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. 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.