Coursera Capstone Project:

Predicting Severity and Causes of Road Accidents.

1 Introduction

1.1 Background

Consistently fender benders cause a huge number of passings around the world. As indicated by research led by the World Health Organization (WHO), there were 1.35 million street traffic passings universally in 2016, with millions additionally supporting genuine wounds and living with long haul unfavorable wellbeing results. Around the world, street car accidents are a main source of death among youngsters and the primary driver of death among those matured 15–29 years. Street traffic wounds are at present assessed to be the eighth driving reason for death over all age bunches worldwide and are anticipated to turn into the seventh driving reason for death by 2030[1].

Utilizing the apparatuses and all the data these days accessible, a broad examination to foresee car crashes and its seriousness would have any kind of effect on the loss of life. Investigating a critical scope of elements, including climate conditions, territory, kind of street, and lighting among others, a precise expectation of the seriousness of the mishaps can be performed. Accordingly, patterns that generally lead to extreme traffic occurrences can help recognize exceptionally serious mishaps. This sort of data could be utilized by crisis administrations, to send the specific required staff and hardware to the spot of the mishap, leaving more assets accessible for mishaps happening at the same time. Besides, this serious mishap circumstance can be cautioned to close by clinics which can have all the gear prepared for an extreme mediation ahead of time.

Thus, street wellbeing ought to be an earlier enthusiasm for governments, neighborhood specialists, and privately owned businesses putting resources into innovations that can help decrease mishaps and improve in general driver security.

1.2 Problem

Information that may add to deciding the likeliness of a potential mishap happening may remember data for past mishaps, for example, street conditions, climate conditions, specific time and spot of the mishap, kind of vehicles associated with the mishap, data on the clients engaged with the mishap and obviously the seriousness of the mishap. This venture expects to gauge the seriousness of mishaps with past data that could be given by an observer illuminating the crisis administrations.

1.3 Interest

Governments ought to be exceptionally intrigued by, exact expectations of the seriousness of a mishap, so as to lessen the hour of appearance and to utilize the assets, and accordingly spare a lot of individuals every year. Others intrigued could be privately owned businesses putting resources into advancements intending to improve street safeness.

2 Data

2.1 Data source

The data can be found in the following Kaggle data set click here.

2.2 Feature Selection

The information is separated into 5 diverse informational indexes, comprising the apparent multitude of recorded mishaps in France from 2005 to 2016. The attributes informational collection contains data on the time, spot, and kind of crash, climate and lighting conditions, and sort of convergence where it happened. The spots informational collection has street particulars, for example, the inclination, shape, and class of the street, the traffic system, surface conditions, and foundation. On the client informational collection, it tends to be discovered the spot involved by the clients of the vehicle, data on the clients engaged with the mishap, the explanation behind voyaging, the seriousness of the mishap, the utilization of security hardware, and data on the walkers. The vehicle informational collection contains the stream and kind of vehicle, and the occasion one names the mishaps happening on a vacation. Each of the five datasets share the mishap distinguishing proof number.

An underlying investigation of the information was performed for the choice of the most important highlights for this particular issue, decreasing the size of the dataset and dodging repetition, click here. With this cycle, the quantity of highlights was diminished from 54 to 28.

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 where the following:

From the *characteristics* dataset: lighting, localisation, 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, traf- fic 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 y.o. involved in the accident.
- **severity** : maximum gravity suffered 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.

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 profile and shape and the situation at the time of the accident, the NaN and outliers were replaced with the future'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 defined with the string type. I used the to data 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.

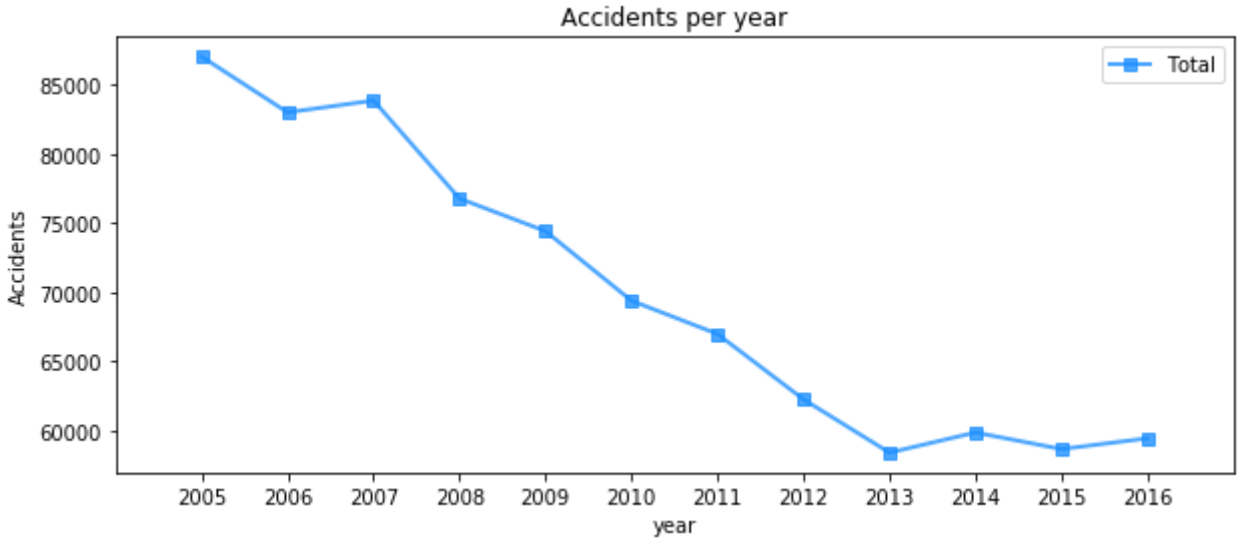

Figure 1: Line Plot of total amount of accidents per year.

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 again in September. This pattern can be seen in the following two figures.

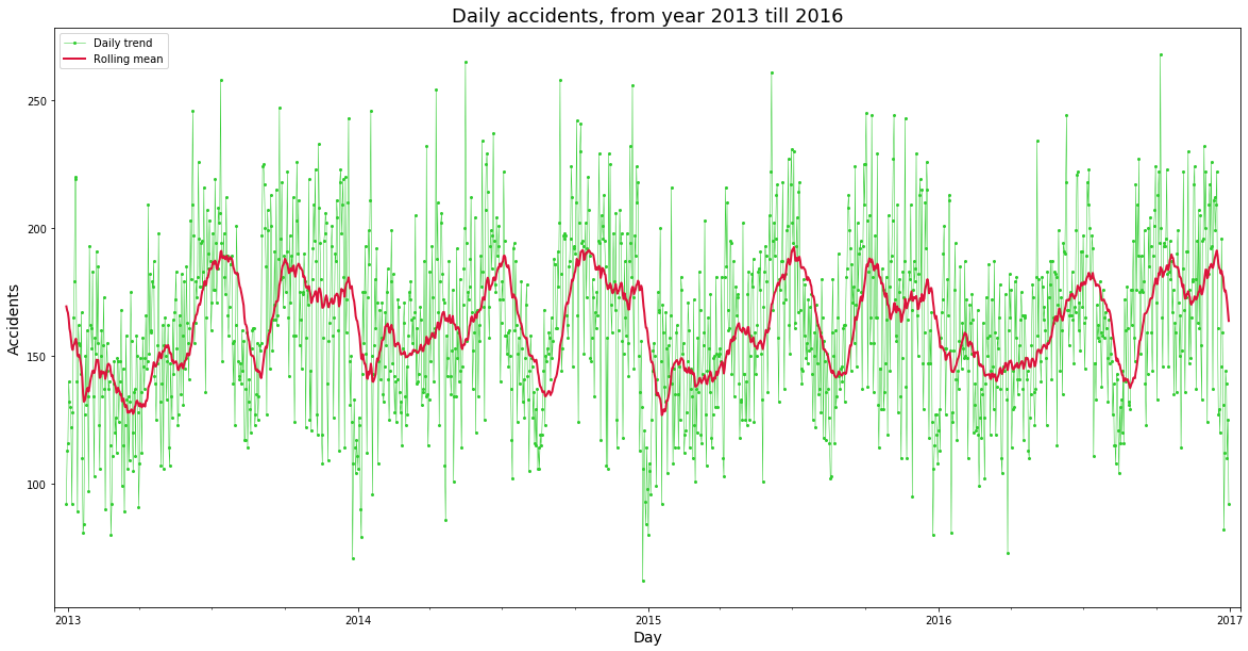

Figure 2: Line Plot of the amount of accidents per day during the year 2013 to 2016. The plot includes the rolling mean, with a window size of 30 days.

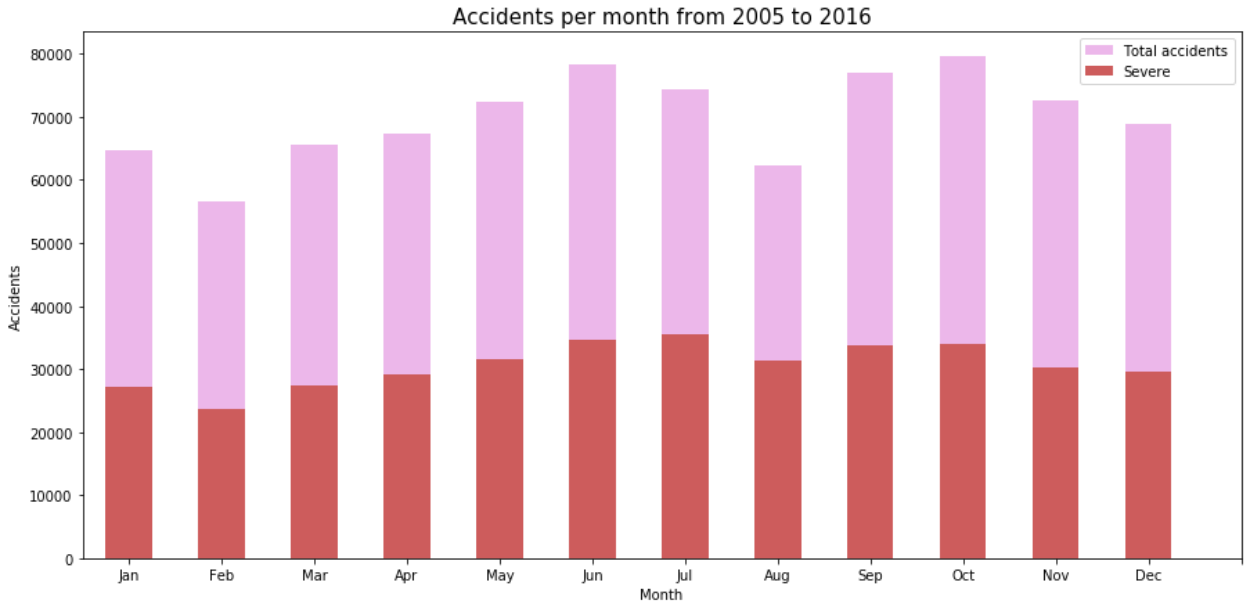

Figure 3: Bar graph: Amount of accidents per month from 2005 to 2016.

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 accidents of all.

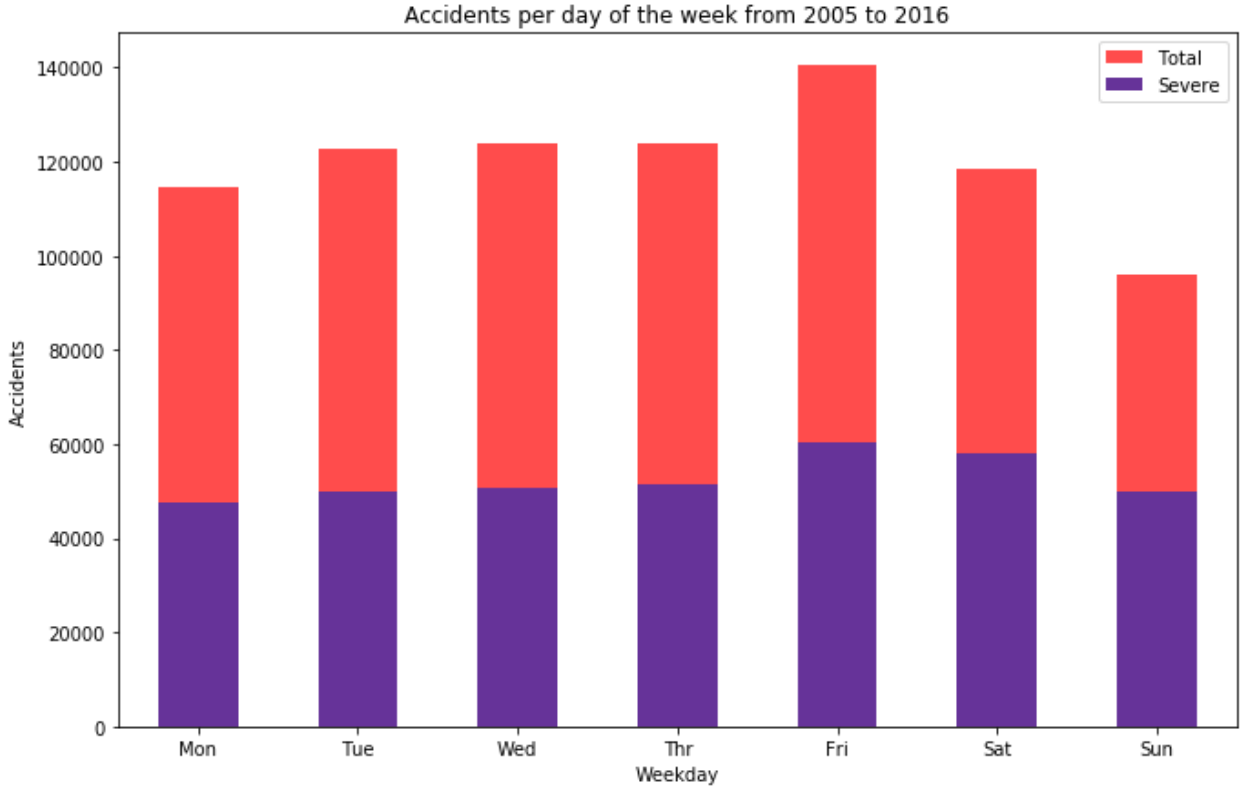

Figure 4: Bar graph: Amount of accidents per day of the week from 2005 to 2016.

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 6pm, time when people return home. The number of accidents decreases between these two spikes, nothing unusual but it proves there is a pattern here.

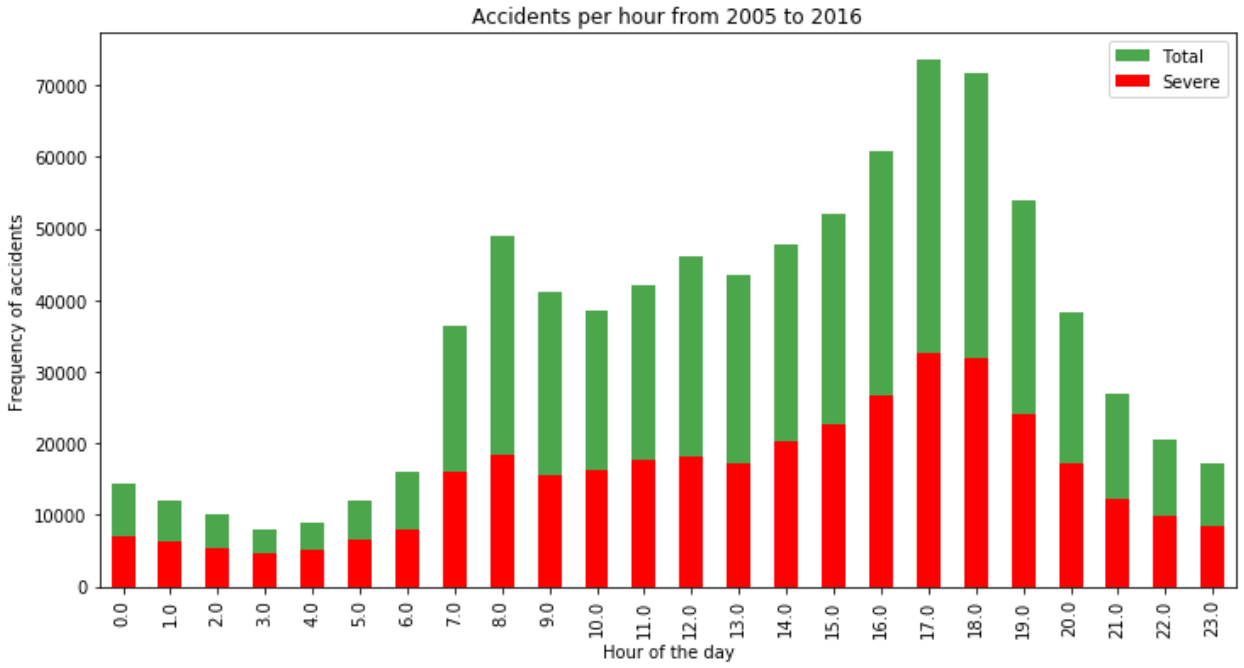

Figure 5: Bar graph of total amount of accidents per year.

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, two 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

Distinctive grouping calculations have been tuned and worked for the expectation of the degree of mishap seriousness. These calculations gave a regulated learning

approach foreseeing with certain exactness and computational time. These two properties have been contrasted all together with deciding the most appropriate calculation for his particular issue.

Right off the bat, the 839.985 lines were part 80/20 between the preparation and test sets, a while later, an extra 80/20 split was performed among the preparation tests making the approval set for the improvement of the models. At that point the information was normalized giving zero mean and unit fluctuation to all highlights.

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 increase of the computational time.

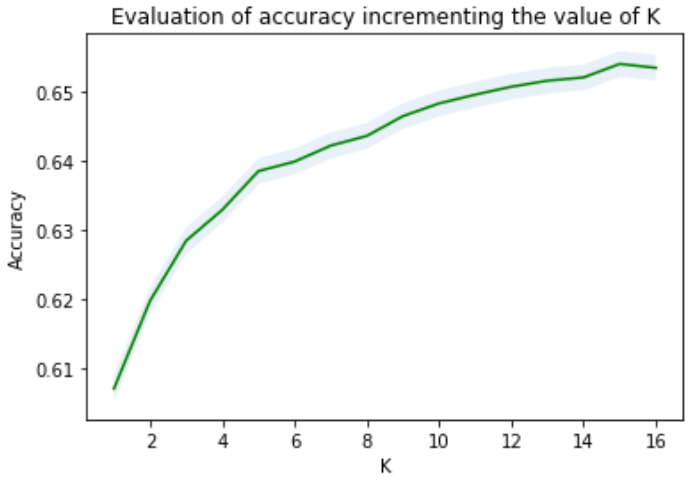

Figure 6: Accuracy of KNN models increasing the value of K.

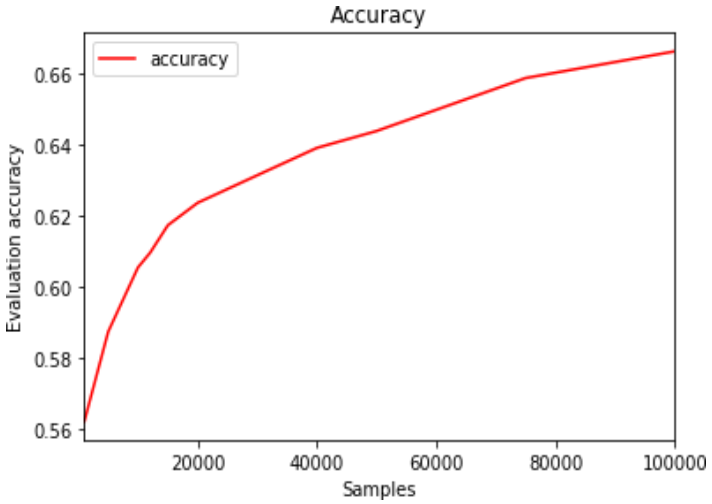

Figure 7: Accuracy of SVM increasing the training sample's size.

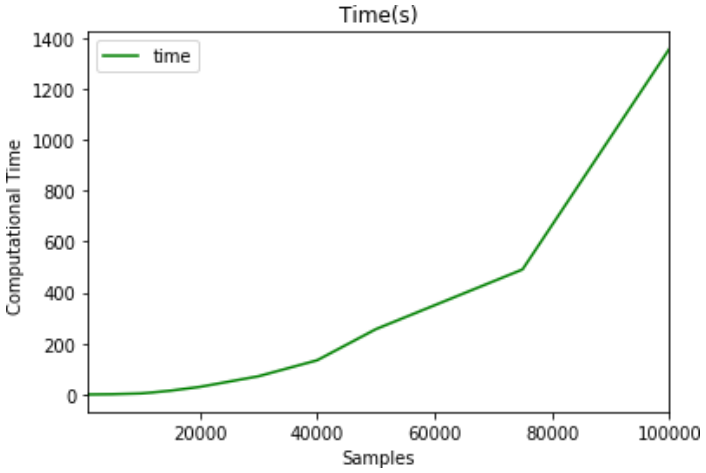

Figure 8: Computational time of SVM increasing the training sample's size.

5 Results

Here I 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	0.722	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

Figure 9: Final results

For this situation, the review is a higher priority than the accuracy as a high review will support that all necessary assets will be prepared up to the seriousness of the mishap. The strategic relapse, KNN, and SVM models have comparable exactness, notwithstanding, the computational time from the relapse is much better than the other two models. Without no uncertainty, the Random Forest is the best model, simultaneously as the log. res. it improves the precision from 0.66 to 0.72 and the review from 0.45 to 0.59.

1. Proportion of predicted severe accidents that were truly severe
2. Proportion of truly severe accidents that were properly predicted

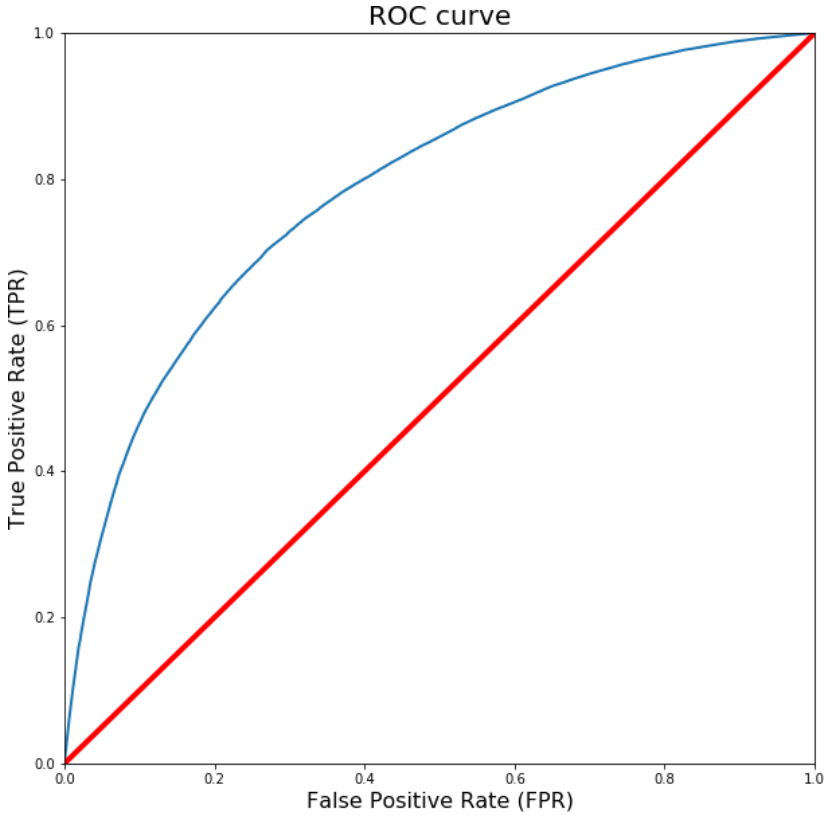

Figure 10: Representation of the ROC curve from the results of the Random Forest model.

I also evaluated the best model using their ROC curves. In this particular problem, a 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 assessment, I dismembered the association between the earnestness of an incident and a couple of characteristics that depict the situation that incorporated the setback. From the start, I envisioned that features, for instance, cools, the lighting, or being an event would be the most appropriate ones, yet I recognized the workplace, the day and period of the incident, the road grouping, and kind of crash among the most huge features that impact the gravity of the setback. I gathered and differentiated 4 unmistakable game plan models with predict whether a disaster would have a high or low reality. These models can have various applications, in reality. For instance, imagine that emergency organizations have an application with some default features, for instance, date, time, and division/area, and thereafter with the information given by the onlooker calling to instruct on the disaster they could envision the reality of the setback before showing up in this way ready near to clinical centers and plan with the principal apparatus and staff. Also, by perceiving the features that favor the most the gravity of an incident, these could be taken care of by improving road conditions or extending the experience with the general population.

7 Observation

I had the option to accomplish 68% exactness. Notwithstanding, there was as yet a critical change that couldn't be anticipated by the models in this investigation. I think different highlights like speed or continuous season of making a trip could be utilized to anticipate a more precise grouping. These are qualities that might be difficult to know at this moment, yet at the extraordinary harmony that innovation is developing these days, soon vehicles will have the option to follow them so the crisis administrations could utilize them.

One issue I think these highlights had is that the objective of this characterization issue was streamlined to two distinct classes, low and high seriousness. Marking seriousness with a scope of accentuation from 0 to 100, for example, could permit the chance of building up a relapse model.

The subsequent stage on this issue could be to include a mishap expectation model ready to foresee the precision as well as the crucial time and spots where potential mishaps can happen ahead of time.

References

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