Predicting Traffic Accident Severity

IBM/Coursera Data Science Professional Certificate

Capstone Project

('Predicting Traffic Accident Severity')

Table of contents

1. Introduction

1.1 Background

Car accidents claim lives of thousands of people every year. According to the CARE (Community database on Accidents on the Roads in Europe)

database, road traffic accidents in the member states of the European Union claimed about 25,600 lives and left more than 1.4 million people injured in 2016. Research conducted by the World Health Organization (WHO) in 2016 found that there were 1.35 million road traffic deaths, with millions sustaining serious injuries and living with long-term adverse health consequences. According to the same report, road traffic accidents is a leading cause of death among young people.

Leveraging the tools and all information available plus an extensive analysis to predict road accidents and severity can make a difference in helping to decrease death toll. I believe that analysing multiple factors such as weather conditions, type of road and lighting an accurate prediction of road accidents would be possible. Armed with this analysis about factors and trends which lead to multiple accidents we can identify severe accidents. Such an analysis would benefit emergency services who can plan resources much better. Hospital facilities can also be on standby in case of certain severe conditions worsening on a particular day. Road safety is a major concern for both governments and private organizations investing in technologies to reduce road accidents.

1.2 Problem

Past data on accidents can help us to predict the potential for a particular accidents occurring in the future. When an accident takes place, certain information is usually collected such as weather conditions, time of accident, and place of accident, users and road conditions. My project is an attempt to predict the severity of an accident based on accounts of witnesses to emergency services.

1.3 Interest

Every government has an interest to reduce road carnage and save human lives. It would be highly helpful if one can predict the severity of accident which could lead to a better planning of emergency resources. Private companies involved in development of technologies for road safety might also be interested in information about the severity of road accidents.

2. Data description

2.1 Data source

The dataset can be found from the Kaggle link below:

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

2.2 Feature selection

The data consists of five data sets containing all accidents recorded in France between 2005 and 2016. First the characteristics data set contains information on time, place, type of collision, weather, lighting and the type of intersection where it occurred. Second the places data set specific information about the road such as conditions, gradient, shape and category of road surface conditions and infrastructure. Third the user data contains information about the position of the car occupants, reason for traveling, accident severity, the use of safety equipment and the state of pedestrians. Fourth the vehicle data set contains information about the type of the vehicle. Last but not least the holidays data labels accidents occurring on a holiday. All data sets contain a common unique identifier number of each accident. I performed an initial analysis of all five data sets which led to selection of the most relevant features for my study. Through this process the number of features was reduced from the original 54 to 28. This helps me to avoid redundancy in the data set.

2.3 Description

The data set which resulted from the feature selection exercise consisted of 839,985 samples with each one describing an accident and 28 different features.

First the following features we selected from the characteristics data set: lighting, localization, type of intersection, atmospheric conditions, type of collision, department, time and coordinates. Additionally two new features were developed i.e. date to perform seasonality analysis of the accident and weekend to indicate whether the accident occurred during a weekend or not.

Second the following features were selected from the places set: road category, traffic regime, number of traffic lanes, road shape, surface condition, situation, school nearby, road profile and infrastructure.

Third these features were selected from the users data set: number of users, pedestrians, critical age meaning whether there were users between 17 and 31 years old in the accident, maximum gravity suffered by the accident victims, unscathed or light injury (0), hospitalized wounded or death (1).

Fourth the holiday dataset was used to add a last feature through labelling of accidents which occurred on a holiday.

2.4 Data cleaning

Data cleaning is the process of giving a proper format to the before further analysis. First, I dealt with large values and outliers. I dropped longitude, latitude and road number from the data set because more than 50% of the values were 0 or NaN. Then keeping with replacing the missing values, analysis was split into two groups of features. The first group had all features with a label which described other cases. For example the feature describing the atmospheric condition had a value of 9 or any other atmospheric condition not labelled has a value of 8. Thus the missing values and outliers were replaced with other cases labels for features of atmospheric conditions, type of collision, road category and surface conditions. For the second group of features, the distribution of their values was analysed instead. Then two features were dropped, infrastructure and reserved lanes as the outliers represented more than 75% of the 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 accident, the NaN and outliers were replaced with the features most popular value.

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

On the data type, all features had an appropriate data type except for date feature which was defined with the string type. I used the to-data function of pandas to define the date feature with datetime type. At the end I had 24 features.

3. Methodology (Exploratory Data Analysis)

First the distribution of the target's value was visualized. The plot confirmed that it is a balanced labelled dataset as the samples were divided 56-54 with more cases of lower severity. I then performed a seasonality analysis visualizing the global trend of accidents as well as the number of accidents grouped by years, month and day of the week.

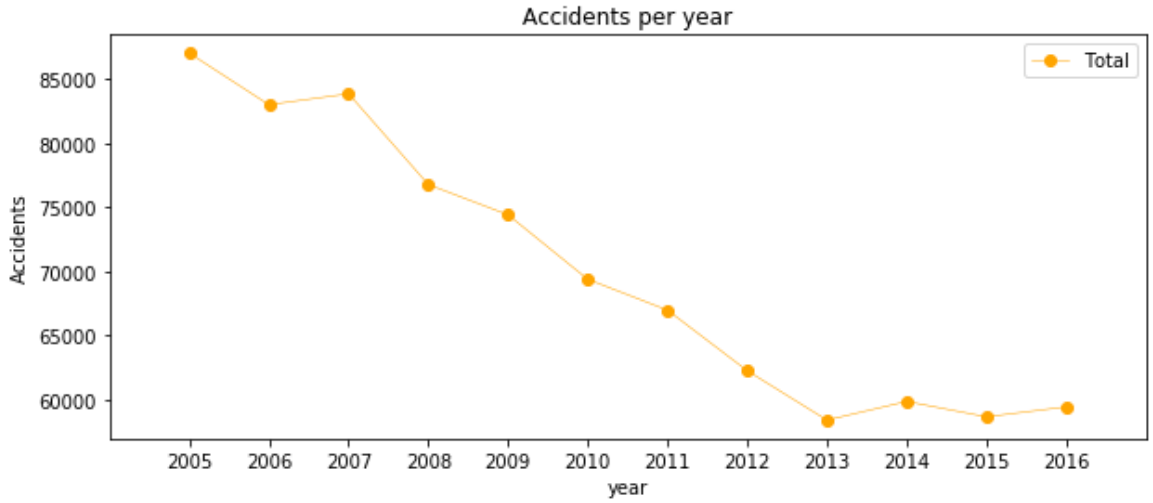

Fig. 1: Line plot of decreasing total number of accidents per year

Fig. 1 shows that there was a continuous decrease of total accidents from 2005 to 2013 before the trend became stable. A further analysis of the yearly trend reveals that accidents tend to peak in the months of March and September.

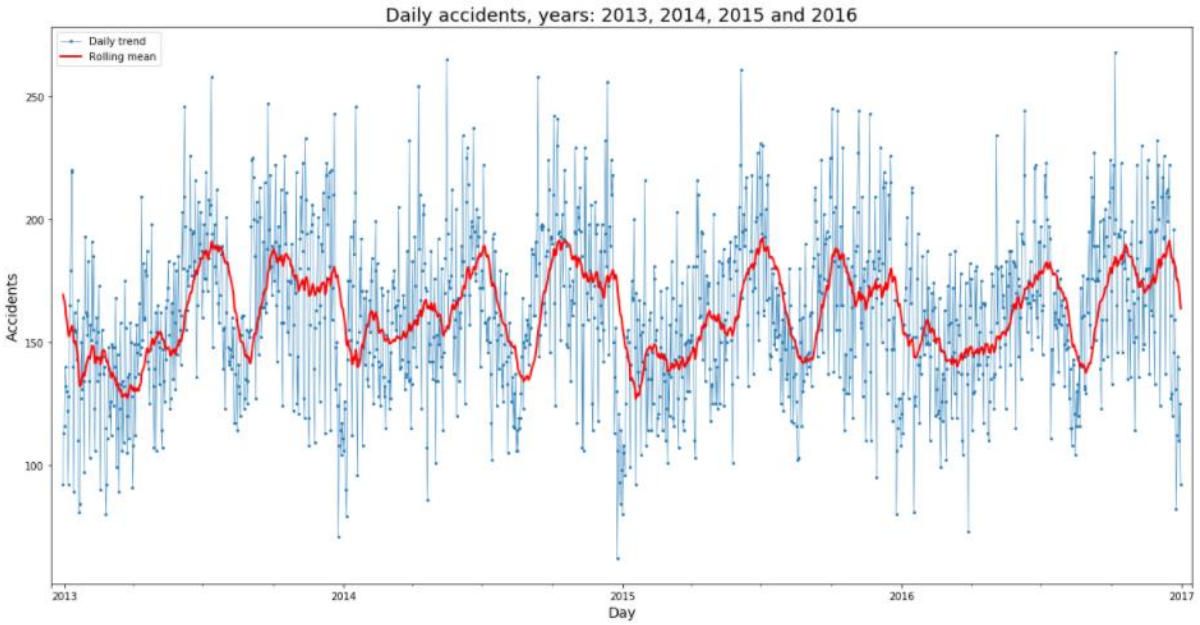

Fig. 2: Line plot of the amount of accidents per day from 2013 to 2016. The plot includes the rolling mean, with a window size of 30 days.

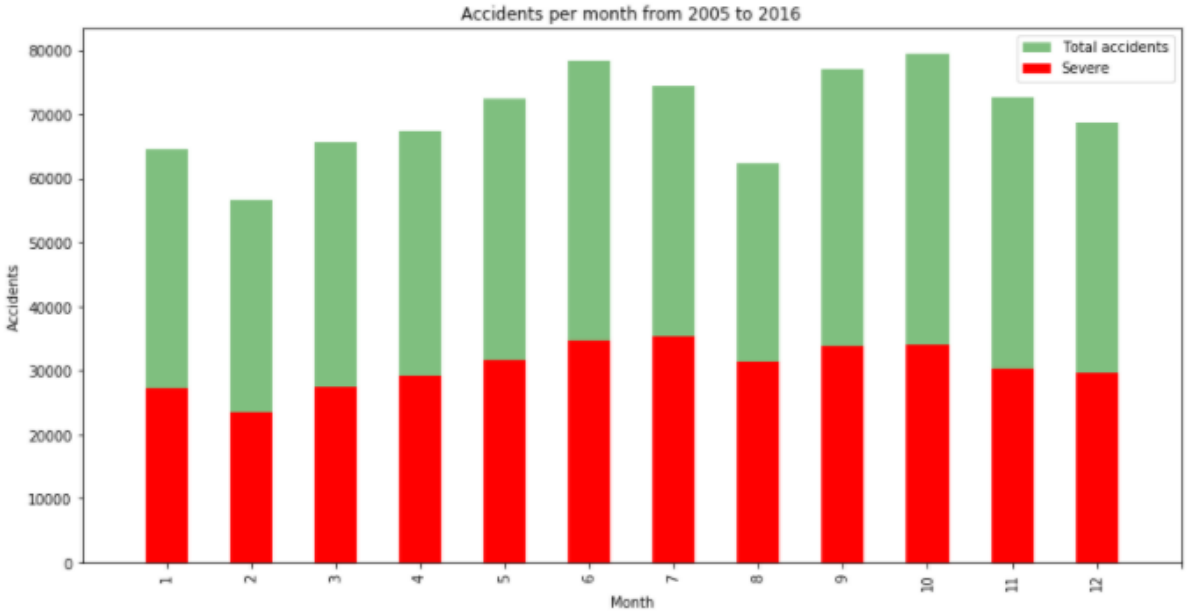

Fig. 3: Bar plot of cumulative total of accidents per month from 2005 to 2016

In terms of days of the week, there is not significant difference in terms of total number of accidents on a given day. The trend is that Friday recorded the highest number of accidents while Sunday the least.

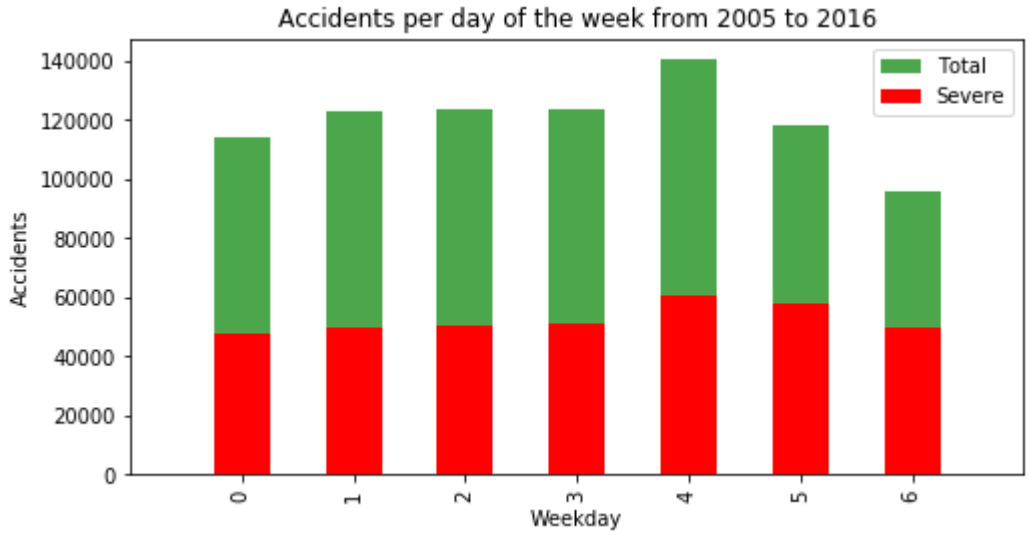

Fig. 4: Bar plot showing the number of accidents per week day from 2005 to 2016

Analyzing the accidents per hour, there are clearly two spikes of accidents one at 8 am, perhaps when people are rushing to work and another one between 5 pm and 6 pm when people return home. The number of patterns tends to decrease between 8 am and 6pm indicating that there is a pattern.

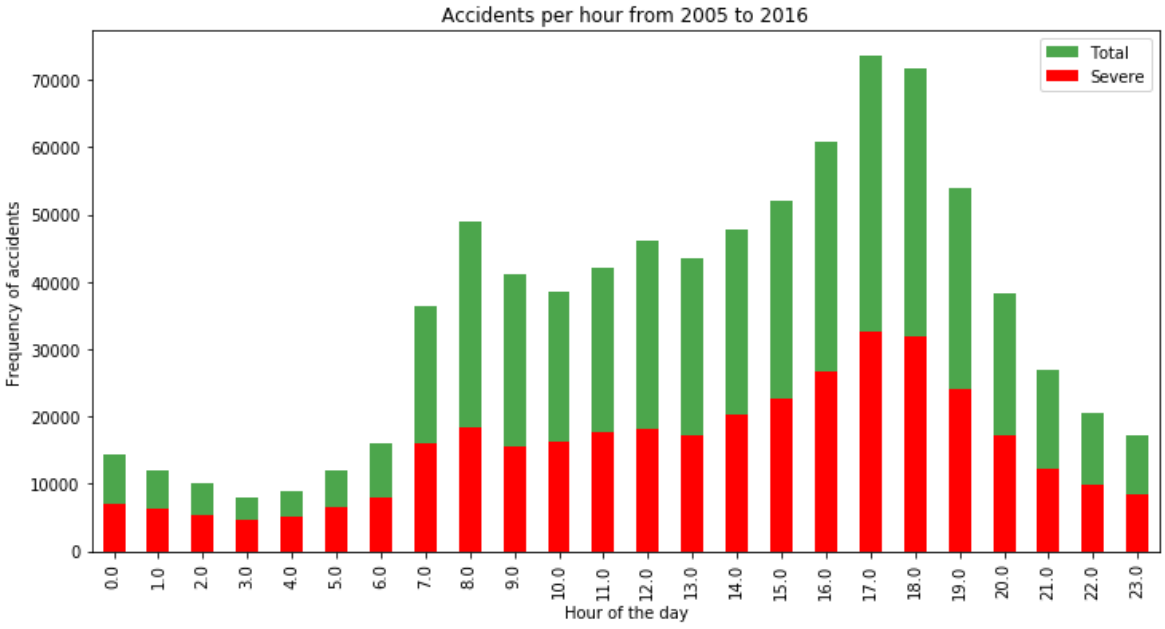

Fig. 5: Bar plot of total accidents per hour from years 2005 to 2016

Zooming on the number of severe accidents on a given hour, we can see that most of them occur between noon and evening.

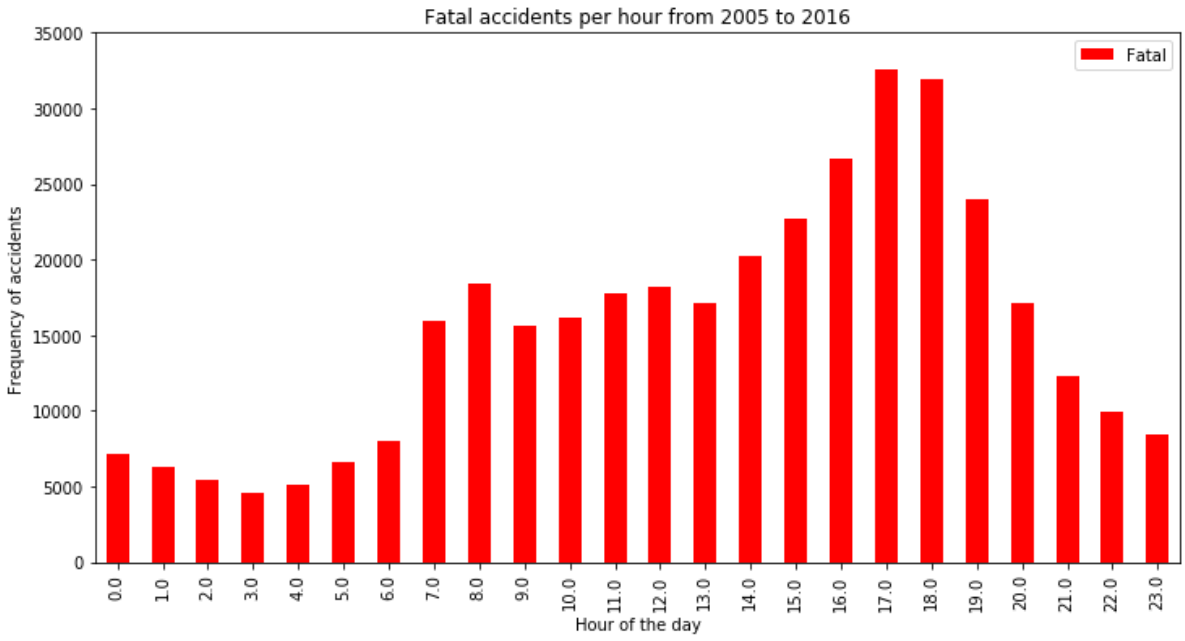

Fig. 5: Bar plot of severe accidents per hour from 2005 to 2016

A further analysis revealed that accidents involving above 84 years of age tend to be more severe.

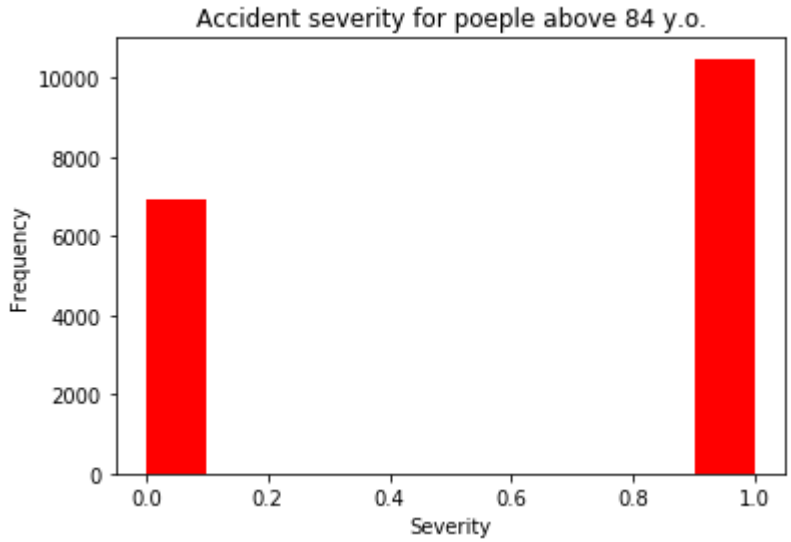

Fig. 6: Histogram showing high severity of accidents for people above 84 years of age

4. Predictive modelling

Several classification algorithms are suitable for prediction of accident severity. These algorithms provided a supervised learning approach to predict accuracy and computational time for the severity of an accident.

These two properties have been compared in order to determine the most suitable algorithm for this problem.

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

I used four different algorithm approaches:

- Decision Tree (Random Forest)
- Logistic Regression
- K-Nearest Neighbour
- Supervised Vector Machine

The same procedure was performed using each algorithm. For each procedure, the accuracy and computational time for development of each model was calculated.

The decision tree model was modified into a random forest. With the default random forest the features were sorted by impurity based on importance and ability to predict the severity of an accident. Hence, the 10 least important features were dropped to decrease the computational complexity for supervised vector machine and k-nearest neighbour models. I kept 13 features which saw the accuracy stay the same and computational time decrease significantly. I evaluated the parameters for each algorithm and came up with the following models:

- Random forest: 10 decision trees, maximum depth of 13 features and maximum of 8 features compared for the split
- Logistic Regression: c=0.001
- K-Nearest Neighbour: k=16
- Supervised Vector Machine: size of training set = 75,000 samples

The visualizations below show how the KNN and SVM models were selected. The SVM model is computationally inefficient with large sample set. The training set was reduced from 537,590 to 75,000 rows. On Fig. 7 the accuracy is increasing as the training size increases. Fig. 9 shows that this increasing accuracy comes with an increase in computational time.

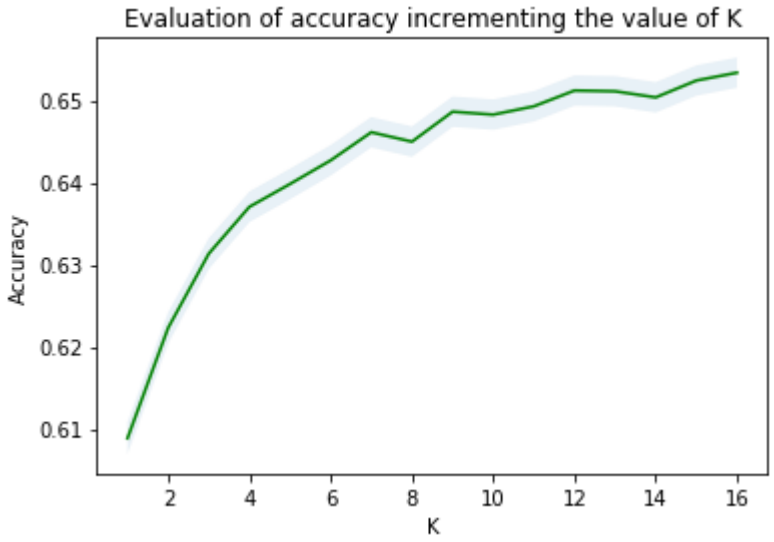

Fig. 7: Line graph showing accuracy of KNN models when increasing the value of K

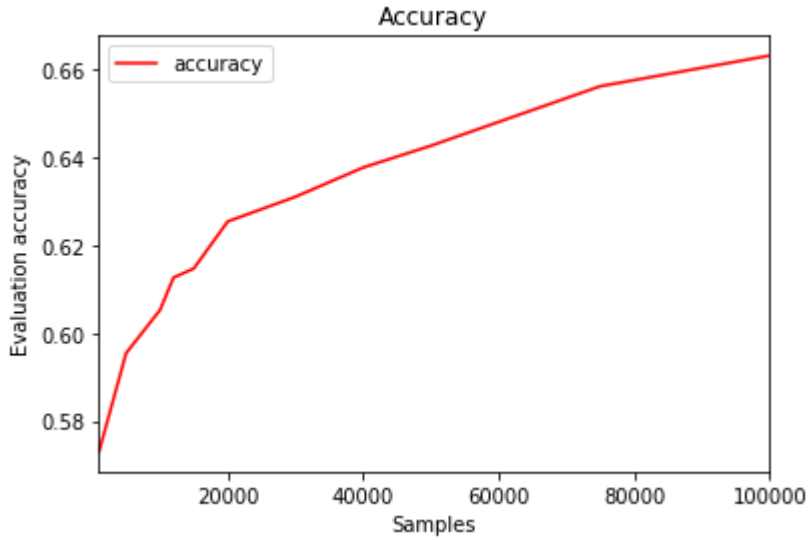

Fig. 8: Line graph showing accuracy of SVM model when increasing sample size

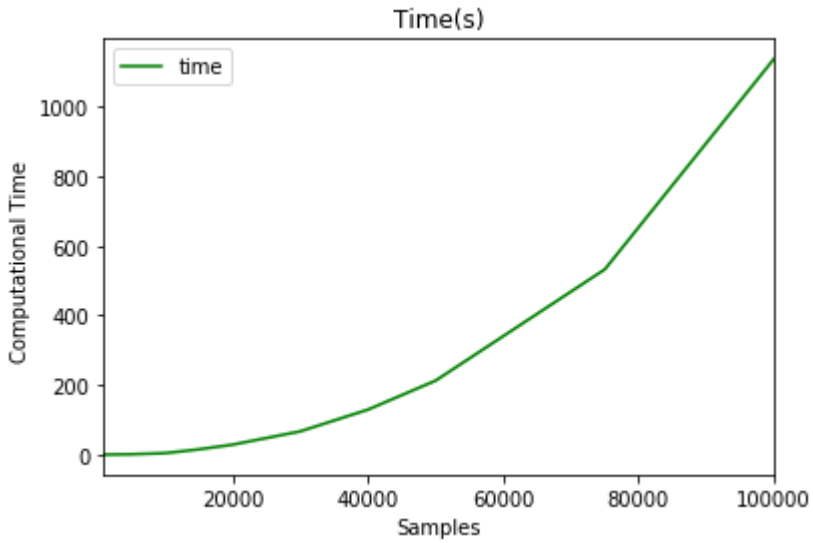

Fig. 9: Line graph showing computational time of SVM increasing the training sample size

5. Results

The metrics used to compare the accuracy of the models are the Jaccard[i] Score, F-Score[ii], Precision[iii] and Recall[iv]. Fig. 10 below reports 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

Fig. 10: A table of metrics used to compare the accuracy of models

In the evaluation above, recall is more important than precision because a high recall will favour that all the required resources will be equipped up to the severity of an accident. The logistic, SVM and KNN models have similar accuracy. However, SVM and KNN have a higher computational time than the logistic regression. Random forest has proven to be the best model while posting the same time as the logistic regression. It improves the accuracy from 0.66 to 0.72 and the recall from 0.45 to 0.59.

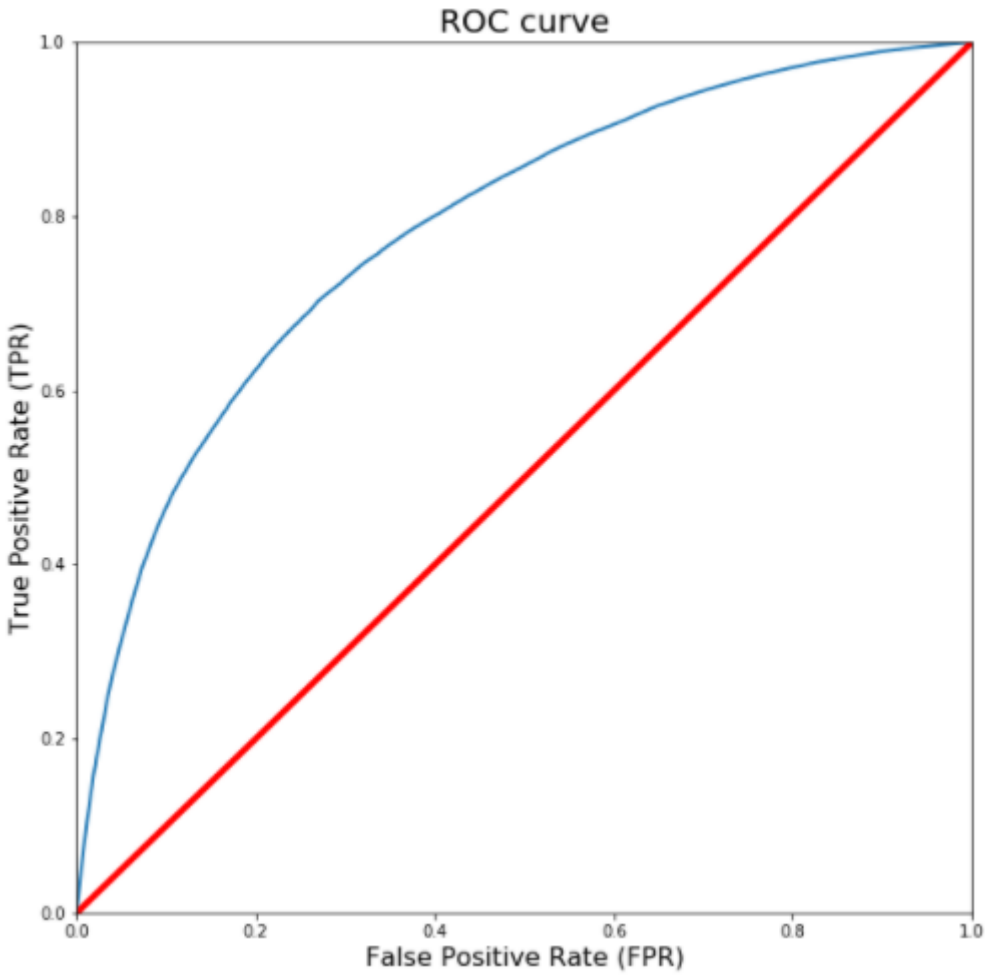

Fig. 11: A Receiver Operating Curve[v] (ROC) from the results of Random forest model

I have also evaluated the best model (Random Forest) using the ROC curves. In this example, lower false positive rate (FPR) is less important than higher true positive rate (TPR). It is more important to properly predict the higher severity accidents properly.

6. Conclusion & discussion

In this study, I analysed the relationship between the severity of an accident and several characteristics which describe the situation of a particular accident. My initial assumption was that features such as weather conditions, lighting or whether it is a holiday would be the most relevant. However, I identified day, time of the day, road category and type of collision as some of the features which affect the severity of an accident. I built and compared 4 different classification models to determine whether an accident has a high or low severity. There are

multiple application of these models in real life. The emergency department could have an application that has as default features such as date, time, municipality and then the information given by the witness calling to inform about the accident to predict the severity of an accident. Predicting the severity of an accident, would allow the nearby hospitals to prepare in advance while emergency services send enough staff to the accident scene. The road works authority could also study the factors responsible for the most severe accidents by improving road conditions and increasing awareness of the population.

7. Observation

With the random forest model, I was able to achieve 72% accuracy. The four models showed different levels of accuracy which doesn't explain all the variance. The model lacked other factors responsible for an accident such as speed which would have provided more accuracy when determining the severity of an accident. The model could be improved by including the speeds of cars involved in a traffic accident. Another improvement of the model might be in the form of a prediction model for telling which spots on the road accidents are likely to occur in order to forewarn drivers.

8. References

Annual Accident Report 2018. European Union (2018). https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/statistics/dacota/asr2018.pdf <Retrieved on 13.10.2020>

Adeloye, D. et. al. Bulletin of the World Health Organization (2016). The burden of road traffic crashes, injuries and deaths in a Africa: A systematic review and meta-analysis. https://www.who.int/bulletin/volumes/94/7/15-163121/en/<Retrieved on 14.10.2020>

[i] Jaccard index is a statistic used for gauging the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the intersection divided by the size of the union of sample sets.

[ii] F-Score is a measure of a tests' accuracy. It is the harmonic mean of precision and recall. It is calculated from the precision and recall of a test.

[iii] Precision is the number of correctly identified positive results divided the number of all positive results including those not identified correctly.

[iv] Recall is the number of correctly identified positive results divided by the number of all samples that should have been identified as positive.

[v] ROC is a graphical plot which illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.