PREDICITNG TRAFFIC ACCIDENT SEVERITY

1. INTRODUCTION

1.1 Background

Hundreds of thousands of deaths worldwide are caused by traffic crashes every year. In 2016, there were 1.35 million road traffic accidents worldwide, with millions more suffering severe injuries and dealing with long-term adverse health effects, according to a survey undertaken by the International Health Organization (WHO). Road accidents are a major cause of death for young people globally, and the leading cause of death for 15-29 year-olds. Currently, road traffic deaths are estimated to be the eighth leading cause of death among all age classes worldwide and are expected to become the seventh leading cause of death by 2030.

A comprehensive analysis to forecast traffic incidents and their magnitude will make a difference to the death toll by using the equipment and all the details now available. A reliable estimation of the severity of incidents can be carried out by evaluating a significant number of variables, including environmental patterns, location, type of road and lighting, among others. Trends that normally lead to serious traffic injuries will also help to classify the exceptionally severe accidents. Emergency services could use this type of information to send the exact status and equipment required to the accident location, leaving more resources available for accidents occurring simultaneously. In addition, it is possible to alert nearby hospitals of this serious emergency scenario, and would get all the resources available for a serious operation in advance.

Road safety should therefore be a priority for governments, local authorities and private companies to invest in technologies that can help reduce accidents and improve the overall safety of drivers.

1.2 Problem

Information on past incidents, such as road conditions , environmental patterns, the precise time and position of the crash, the type of vehicles involved in the accident, information about the users involved in the accident and, of course, the seriousness of the accident, may lead to the estimation of the probability of a possible accident happening. The purpose of these projects is to predict the magnitude of injuries using advance knowledge that could be provided by an emergency witness.

1.3 Interest

In order to minimize the time of arrival and to allow more effective use of resources and thereby save a large number of lives per year, policymakers should be particularly interested in reliable forecasts of the seriousness of an accident. Private businesses involved in technology that seek to improve road safety may also be involved.

2. DATA

2.1 Data source

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

2.2 Feature Selection

The data were split into 5 separate sets of records, comprising of all the incidents reported in France between 2005 and 2016. The data set of characteristics provides detail about the location, position, and manner of collision, the circumstances of the weather and lighting, and the manner of intersection where it occurred. Road parameters such as the gradient, form and category of the road, surface conditions and facilities are set for the position data set. The user data collection includes information on the area occupied by the occupants of the car, information on the occupants involved in the crash, the cause for driving, the seriousness of the crash, the use of protective devices and pedestrian information. The vehicle data set contains the vehicle type, and the holiday one marks the accidents that occur on a holiday. The crash identification number is spread across all vehicle data sets. To determine the most important features for this particular challenge, an initial review of the data was carried out, reducing the size of the dataset and preventing duplication. With this process, the number of characteristics was decreased from 54 to 28.

2.3 Data Cleaning

The method of providing the data a suitable format for further study is data cleaning. Dealing with missed values and outliers was the first step. The latitude, longitude and road number were originally dropped from the data frame when more than 50% of its values were NaN or 0, which in this case is an outlier. Then, the study was split into two classes of features to replace the missing values. The 1st category had a symbol representing other situations in all characteristics, such as the characteristic defining the atmospheric conditions had a rating of 9 on every other atmospheric state not branded like the other 8 labels. Therefore, for the characteristics of atmospheric conditions, crash form, path category and surface conditions, the missing values and outliers were replaced with the other case marks. The distribution of their values was evaluated for the second category of features instead. Then two features, the facilities and reserved lanes, were dropped, as the outliers accounted for more than 75% of their data. Finally, with the remaining features with missed values, the number of lanes, the structure and the condition at the time of the crash, the most common value of the feature was replaced with the NaN and outliers.

The last changes in format to the principles of the school and department were made. All samples were split into either 0 or 100 values in the school function, so all 100 values were replaced with 1. Similarly, at the location of the machine, the department feature had an additional 0 attached, so all values were divided by 10. Regarding the type of the data, except for the date feature that was denied with the string type, all features had a coherent data type.

3. EXPLORATORY DATA ANALYSIS

First, it visualised the distribution of the values of the target. The plot verified that as the samples are split 56-54 with more cases with lower intensity, it is a healthy labelled dataset. A seasonality analysis was then carried out, visualising the global frequency of everyday incidents as well as the number of injuries clustered by the years, month of the year, and day of the week.

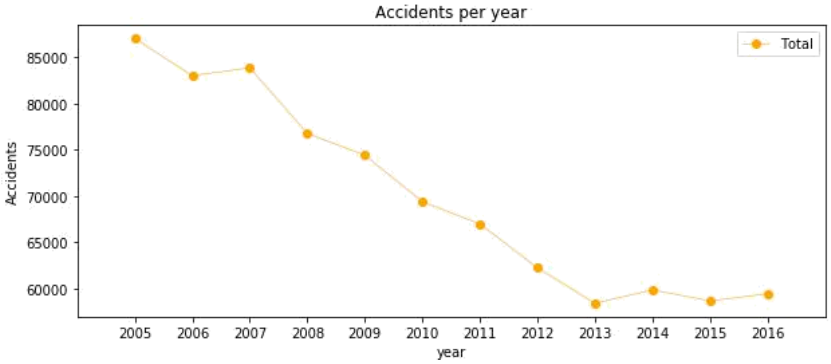

Figure 1: Lineplot of total amount of accidents per year

The previous picture indicates that, from 2005 to 2013, the number of traffic incidents declined, after which the pattern became steady. There is a temporal phenomenon that analyses the annual average, where the number of incidents increases in March and then again in September. This pattern can be seen in the two figures below.

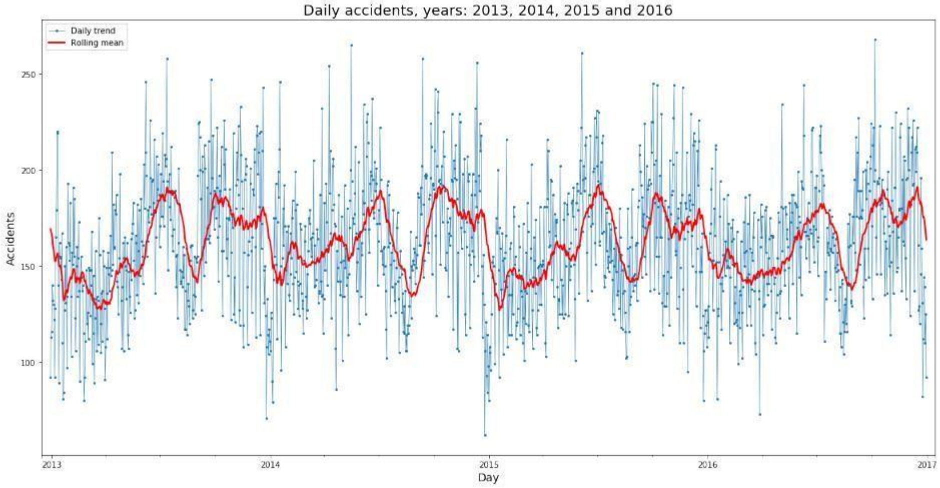

Figure 2: Lineplot of the amount of accident per day during the 2013, 2014, 2015 and 2016. The plot includes the rolling mean, with a window size of 30 days.

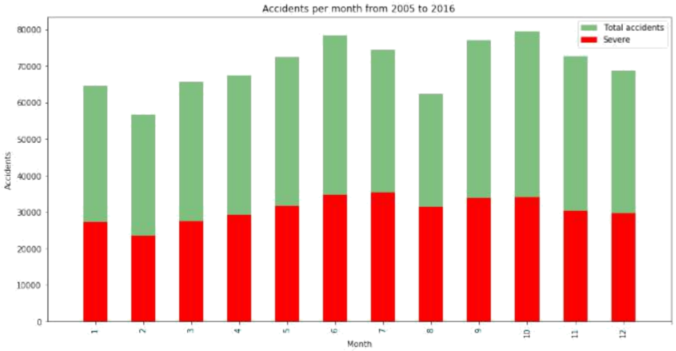

Figure 3: Barplot: Amount of accident per month from 2005 to 2016.

There is no major distinction between them as respect to the day of the week, Figure 4. There is a consistent trend of more crashes on Friday throughout the week, and Sunday is the day with the least reported crash of all.

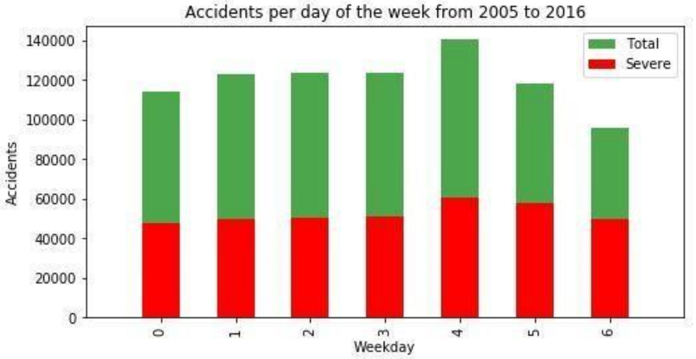

Figure 4: Barplot: Amount of accident per day of the week from 2005 to 2016.

Finally, there are obviously two spikes, one at 8 a.m., when people go to work, and another between 5 and 6 p.m., when people come home, evaluating the injuries every hour. Between these two spikes, the number of injuries drops, nothing odd, but it proves there is a trend here.

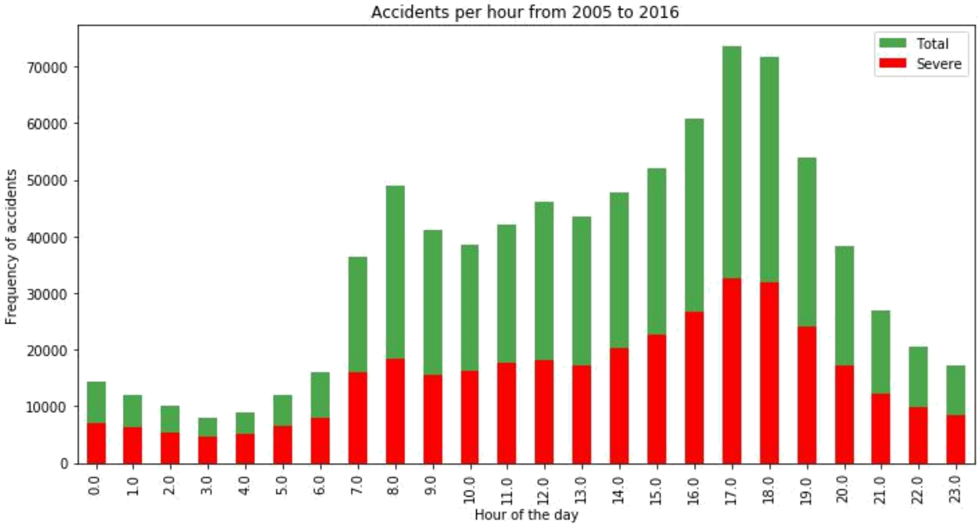

Figure 5: Lineplot of total amount of accidents per year.

For all incidents, split per month of the year and per day of the week, the pattern of extremely serious incidents is relative to the global trend. As we can see in Figure:5, the same phenomenon exists in the number of exceptionally bad injuries every hour of the day. One point to note from the hourly average is that the proportion of serious incidents is higher from noon to midnight, to be exact, the ratio of serious incidents from 9 pm to 6 am is 50.67 percent of the overall number of accidents between these hours, while 42.41 percent is from 7 am to 8 pm. Features were added according to the effects of the previous analysis; month and day as the day of the month.

The association of the characteristics with the seriousness of an event was the next mathematical study. The Pearson correlation for all features revealed a poor or negative correlation. For a better understanding of the results, more visualisations were done. For example, some of the results of this study is that incidents affecting individuals over 84 years of age appear to be of high seriousness.

4. PREDICTIVE MODELING

For the calculation of the degree of accident seriousness, separate classification algorithms have been refined and designed. These algorithms offered a supervised approach to learning that estimated computational time and accuracy with some precision. In order to decide the best suited algorithm for his particular problem, these two properties were compared.

First of all, the 839,985 rows were divided 80/20 between the training and test sets, followed by an additional 80/20 divided between the trial samples that generated the validation set for model creation. The data was then optimised to include all characteristics with zero mean and unit variance.

Four different approaches were used: Decision Tree, Random Forest Logistic Regression, K-Nearest Neighbor, and Supervised Vector Machine

For each algorithm, the same modus operandi was conducted. The best hyperparameters were chosen with train and validation sets and the accuracy and computational time for the creation of the models were measured using the test set.

The architecture of the decision tree was upgraded into a random woodland. With the default random tree, the features in the prediction of severity were sorted by impurity-based value. Therefore, to reduce the computational complexity for the KNN and SVM versions, the 10 least essential features were discarded. The precision remained the same for 13 traits, and the calculation time decreased substantially. This were the models after the parameters for each algorithm were evaluated.

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 visualisations demonstrate how parameters have been chosen for KNN and SVM models. With large sample collections, the SVM model is inefficient in computational terms. Therefore, the measurement of multiple training sizes was found to be a compromise between precision and computational time. From 537,590 to 75,000 line, the training set was reduced. The accuracy improves as the training size grows in Figure: 7, but Figure: 9 illustrates how this comes with a large increase in processing time.

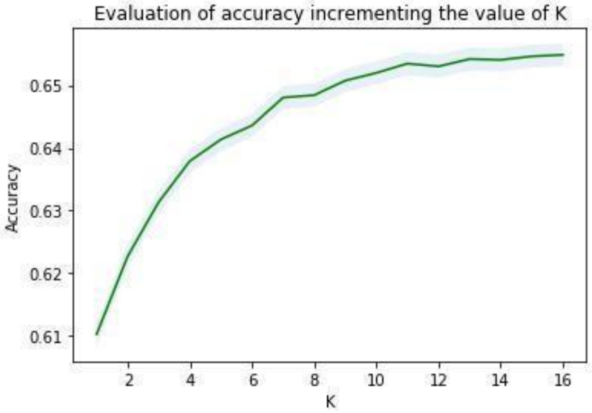

Figure 6: Accuracy of KNN models increasing the value of K.

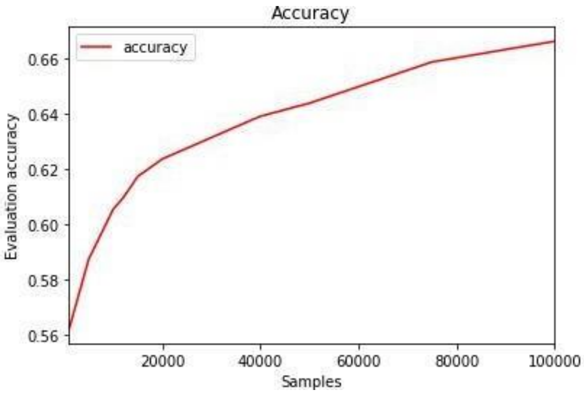

Figure 7: Accuracy of SVM increasing the training sample's size.

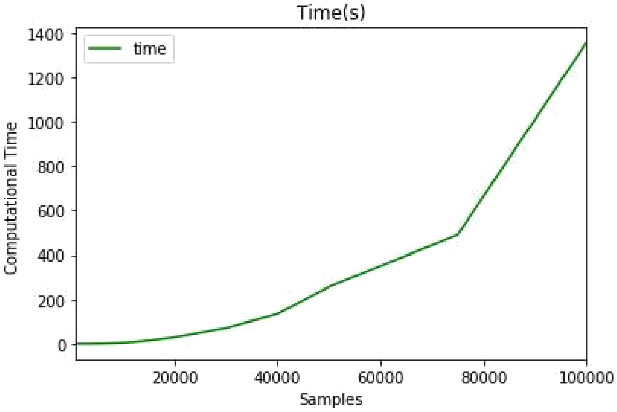

Figure 8: Computational time of SVM increasing the training sample's size

5. RESULT

The Jaccard Score, f1-score, Precision1 and Recall2 are the parameters used to compare the models' precision. The results of testing each model are recorded in this table.

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 scenario, the recall is more important than the accuracy as a high recall would favour the machinery in all appropriate resources up to the seriousness of the accident. Related precision is present in the logistic regression, KNN, and SVM models. The computational time from the regression, however, is significantly better than the other two models. Without a doubt, at the same time as logistic regression, the Random Forest is the best model. This boosts the precision from 0.66 to 0.72 and the recall from 0.45 to 0.592.

6. CONCLUSION

In this analysis, the relationship between the seriousness of an accident and certain attributes explaining the condition involved in the accident was studied. Initially, it was assumed that the most important factors such as weather conditions, illumination or having a holiday will be the most significant ones, but the service, the day and time of the crash, the category of road and the type of collision were listed as the most important features that influence the crash's gravity. To predict when an accident will have a high or low impact, I constructed and compared 4 different classification models. In real life, these models may have many implementations. Imagine, for example, that ambulance responders had an application of some default characteristics such as date , time and department / municipality and then using the details supplied by the eyewitness calling to advise about the accident, they may determine the severity of the accident before getting there and then warn surrounding hospitals and plan for the equipment and state available. This may also be resolved by upgrading road quality or increasing the knowledge of the public by defining the factors that benefit the gravity of a crash the most.

7. OBSERVATION

I was able to achieve 68% accuracy. There was also, though, a substantial uncertainty that the models in this analysis did not forecast. I assume that other attributes, such as speed or continuous travel time, might be used to estimate a more precise classification. There are features that may be difficult to realize right now, but with the amazing progress that technology is developing now, vehicles will eventually be able to detect them so that they can be used by emergency responders.
One flaw I believe these features had was that two separate types, low and high severity, were generalized to the aim of this classification problem. For example, marking gravity with a punctuation range from 0 to 100 might encourage the possibility of creating a regression model.

The next move on this topic may be to incorporate a model of accident prediction capable of estimating not only the accuracy, but also the crucial time and locations where likely injuries will occur in advance