

Predicting Accident Severity.

Sudeep Suresh

Coursera IBM Capstone

Project

1 Introduction

1.1 Background

Every year car accidents cause hundreds of thousands of deaths worldwide. According to a research conducted by the World Health Organization (WHO) there were about 1.35 million road accident deaths globally in 2016, with millions more sustaining serious injuries and living with long-term adverse health consequences. Globally, road crashes are a leading cause of death among younger people and the main cause of death among those aged between 15 -29 years. Road accident 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[1].

Leveraging the tools and all the information is currently available, an extensive analysis to predict road accidents and its severity would make a difference to the death toll. Analyzing a wide range of factors, including weather conditions, locality, type of road and lighting among others would be able to create an accurate prediction of the severity of the accidents. Thus, trends that commonly lead to severe injuries can help identify the type of accidents that causes it.

1.2 Problem

Data that might contribute to determining the likeliness of a potential accidents 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 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 should be highly interested in accurate predictions of the severity of an accident, in order to reduce the time of arrival and to make a more 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 safety.

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 is divided in 5 different data sets, consisting of all the recorded accidents in France between 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 specs such as the gradient, shape and category of the road, the 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 of traveling, severity of the accident, the use of safety equipment and information on the pedestrians. The vehicle data set contains the type of vehicle and other details as well. All the data sets share the accident identification 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 dataset that resulted from the feature selection consisted in 839,985 samples, each one describing an accident and 29 different features.

These features were 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 place's dataset, the selected features were: road category, traffic regime, number of traffic lanes, road profile, road shape, surface condition, situation, school nearby and infrastructure.

The user's 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 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 next 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 where 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 rest 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 feature'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.

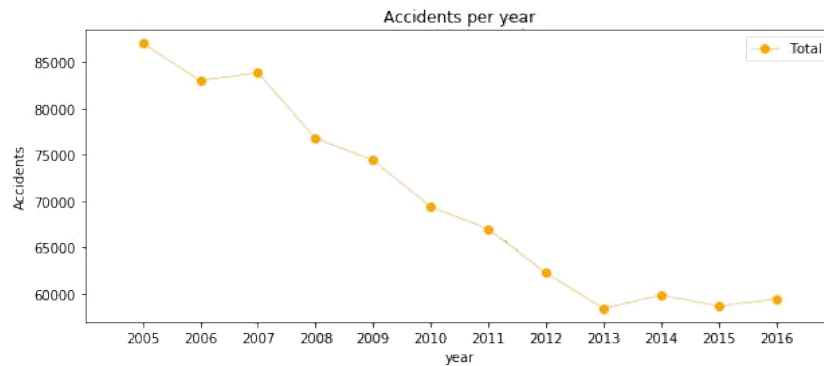


Figure 1: Lineplot of total amount of accidents per year.

The previous image show that the number of tra c 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 following two figures.

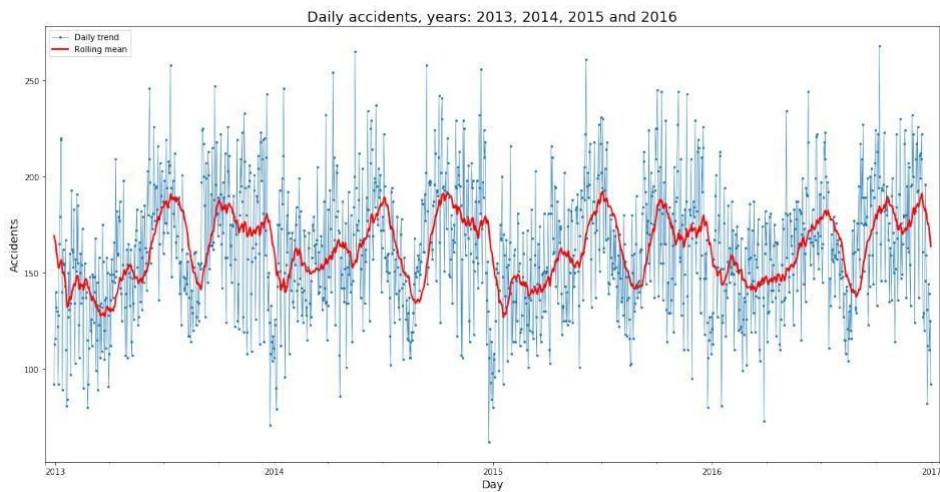


Figure 2: Lineplot of the amount of accident per day during the 2013, 2014, 2015 and 2016. The plot includes the rolling mean, with a window size of 30 days.

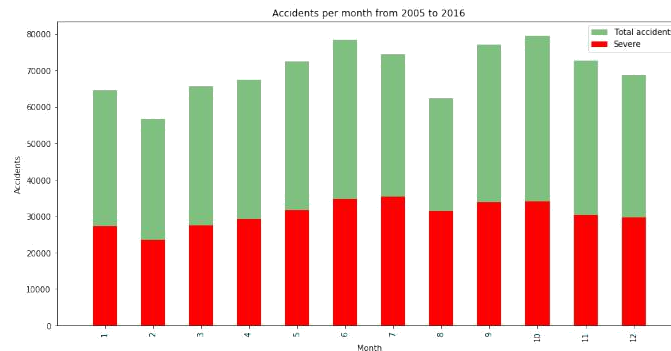


Figure 3: Barplot: Amount of accident per month from 2005 to 2016.

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 less recorded accident of all.

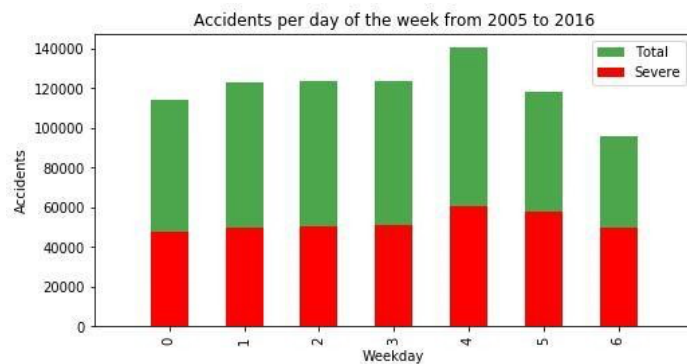


Figure 4: Barplot: Amount of accident per day of the week from 2005 to 2016.

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 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.

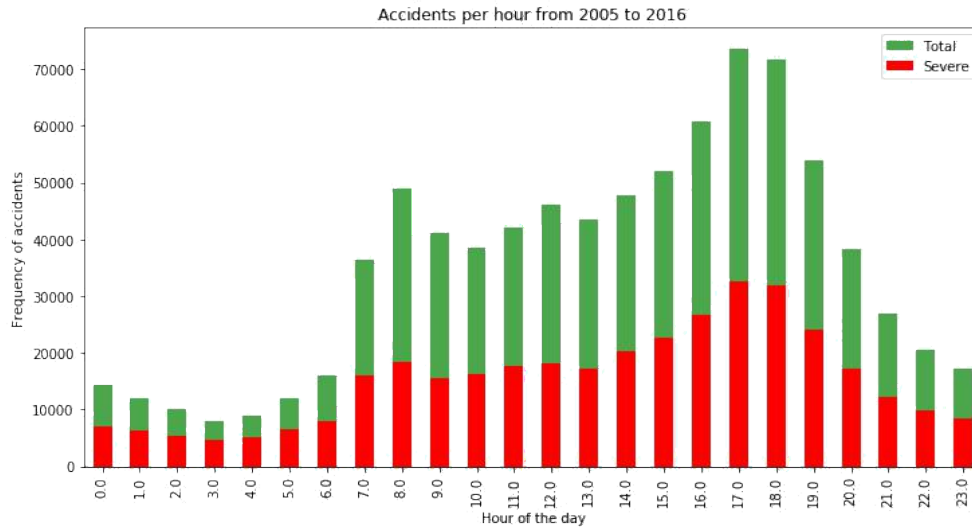


Figure 5: Lineplot of total amount of accidents per year.

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, 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 his 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 Neighbors

Supervised Vector Machine

The same modulus 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.

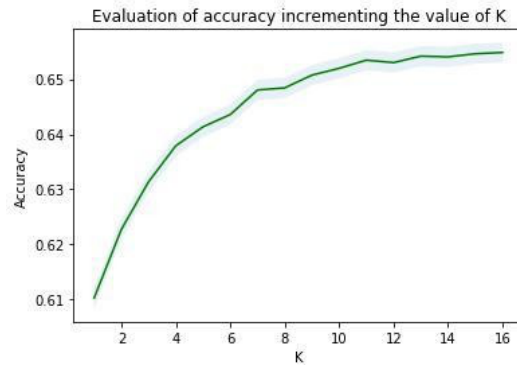


Figure 6: Accuracy of KNN models increasing the value of K.

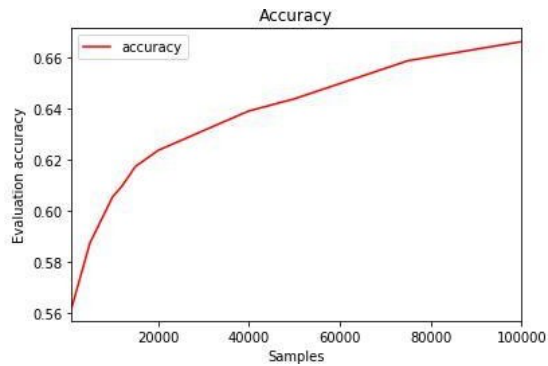


Figure 7: Accuracy of SVM increasing the training sample's size.

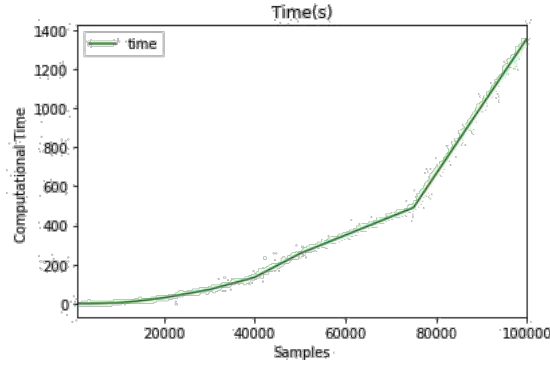


Figure 8: Computational time of SVM increasing the training sample's size.

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

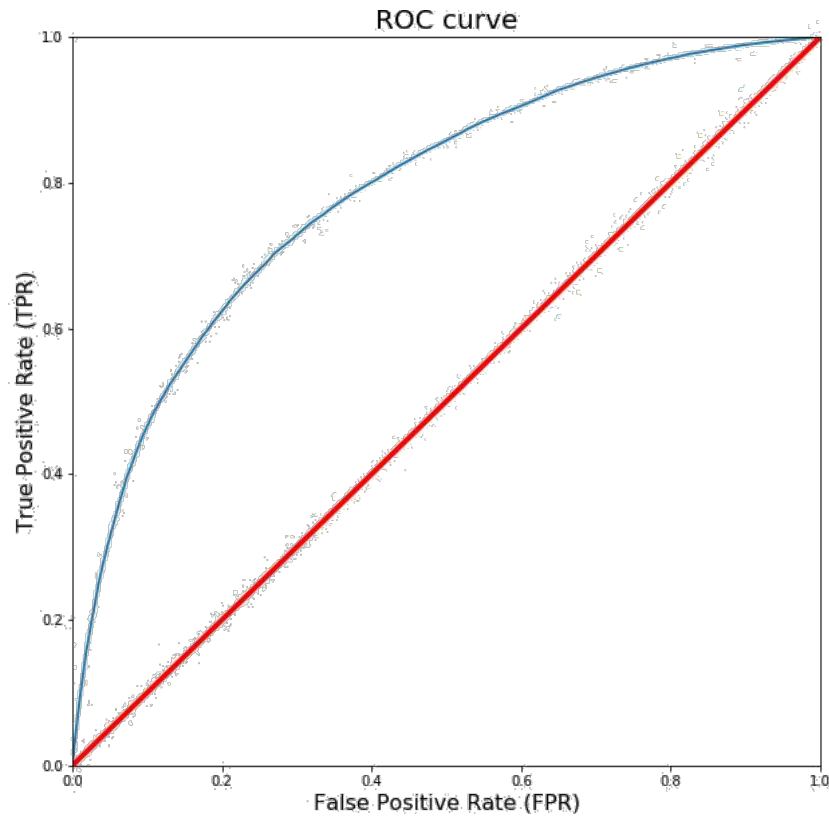


Figure 9: Representation of the ROC curve from the results of the Random Forest model.

I also evaluated the best model using their ROC curves. In this particular problem, lower false positive rate is less important than 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 managed identified the department, the day and time of the accident, the road category and type of collision among

the most important features that effect 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.

7 Observation

I was able to achieve 68% accuracy. 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 peace that technology is evolving nowadays, soon cars will be able of 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 an 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 Tra c Injuries. WHO, Global Health Observation Data, 2016. <https://www.who.int/health-topics/road-safety#tab=tab 1>

-