Predicting Traffic Accident Severity.
Geeta Akshata
October 2020

1.INTRODUCTION

1.1 Background

In 2013, 54 million people worldwide sustained injuries from traffic collisions. This resulted in 1.4 million deaths in 2013, up from 1.1 million deaths in 1990. About 68,000 of these occurred in children less than five years old. Almost all high-income countries have decreasing death rates, while the majority of low-income countries have increasing death rates due to traffic collisions. Middle-income countries have the highest rate with 20 deaths per 100,000 inhabitants, accounting for 80% of all road fatalities with 52% of all vehicles. While the death rate in Africa is the highest - 24.1 per 100,000 inhabitants, the lowest rate is to be found in Europe -10.3 per 100,000 inhabitants.
A 1985 study by K. Rumar, using British and American crash reports as data, suggested 57% of crashes were due solely to driver factors, 27% to combined roadway and driver factors, 6% to combined vehicle and driver factors, 3% solely to roadway factors, 3% to combined roadway, driver, and vehicle factors, 2% solely to vehicle factors, and 1% to combined roadway and vehicle factors. Reducing the severity of injury in crashes is more important than reducing incidence and ranking incidence by broad categories of causes is misleading regarding severe injury reduction. Vehicle and road modifications are generally more effective than behavioral change efforts with the exception of certain laws such as required use of seat belts, motorcycle helmets, and graduated licensing of teenagers.

1.2 Problem overview

The aim is to forecast the severity of accidents with previous information that could be given by a witness informing the emergency services.
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 of course the severity of the accident.

2.DATA

The data can be found in the following Kaggle data

- ❏ The data is divided in 5 different data sets, consisting of all the recorded accidents 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 trac 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 law and type of vehicle.
❏ The holiday one labels the accidents occurring in a holiday.
All the datasets share the accident identification number.

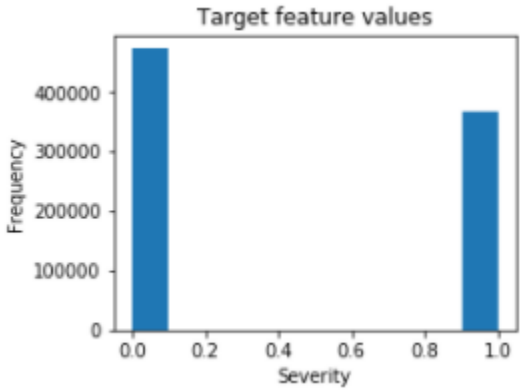
2.1 Description

The dataset that resulted from the feature selection consisted in 839,985 samples, each one describing an accident and 29 different features.
The users dataset was used to craft some new features:
1. number of users: total number of people involved in the accident.
2. pedestrians: whether there were pedestrians involved (1) or not (0).
3. critical age: whether there were users between 17 or 31 y.o. involved in the accident.
4. severity : maximum gravity suffered by any user involved in the accident.

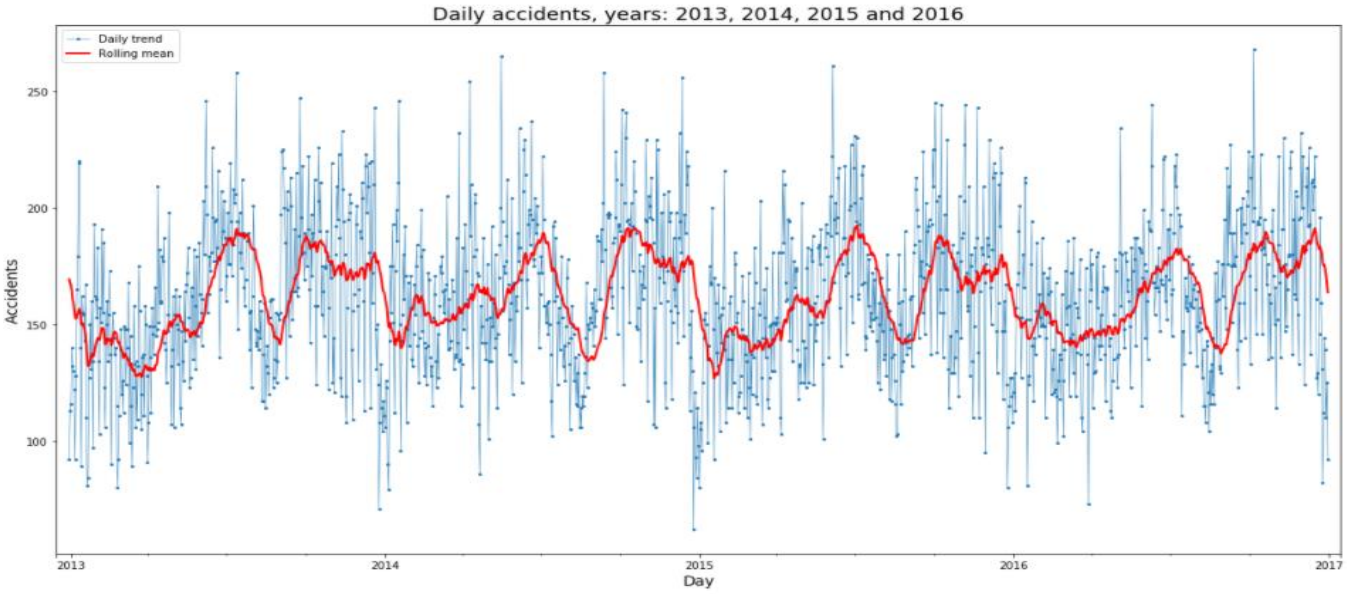
Unscathed or light injury (0), hospitalized wounded or death (1)
The holiday dataset was used to add a last feature, labeling the accidents which occurred on a holiday.

2.2 Analysis

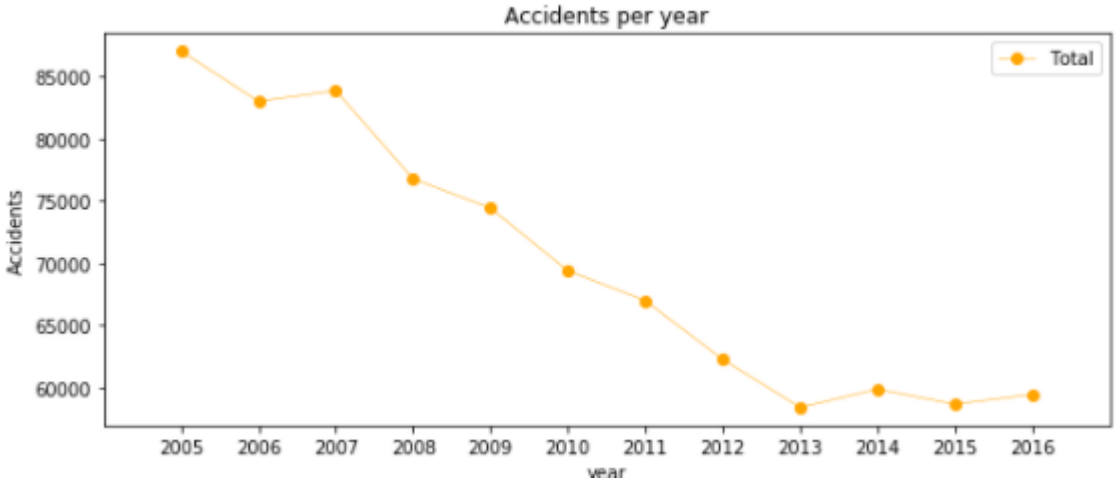
Severity of the accident:



Seasonality of accident:



Accidents by year:



Fatal accidents by the hour of day:

3.METHODOLOGY

Four different approaches were used:
1.Decision Tree-Random Forest
2.Logistic Regression
3.K-Nearest Neighbor
4.Supervised Vector Machine

Random Forest: 10 decision trees, maximum depth of 12 features and maximum of 8 features compared for the split.

Jaccard : 0.7221974201920273

	precision	recall	f1-score	support
0	0.72	0.82	0.77	94297
1	0.72	0.59	0.65	73700
accuracy			0.72	167997
macro avg	0.72	0.71	0.71	167997
weighted avg	0.72	0.72	0.72	167997

Logistic Regression: c=0.001.

Jaccard : 0.6617499122008131

	precision	recall	f1-score	support
0	0.66	0.82	0.73	94297
1	0.67	0.46	0.54	73700
accuracy			0.66	167997
macro avg	0.66	0.64	0.64	167997
weighted avg	0.66	0.66	0.65	167997

KNN: k=16

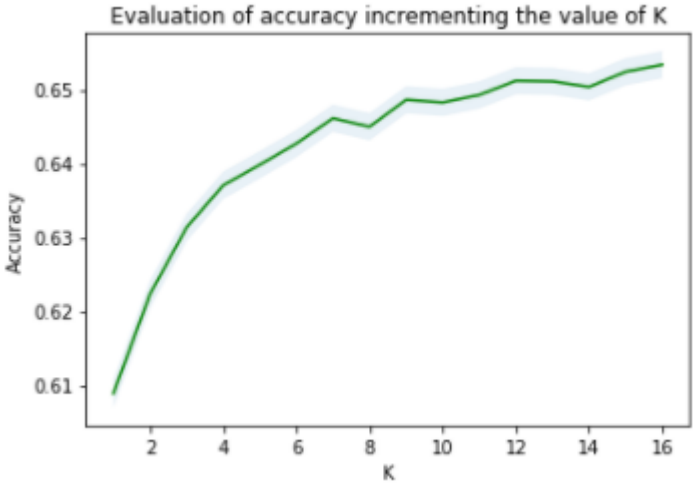


Fig:Accuracy of KNN models increasing the value of K.

SVM: size of the training set= 75,000 samples.

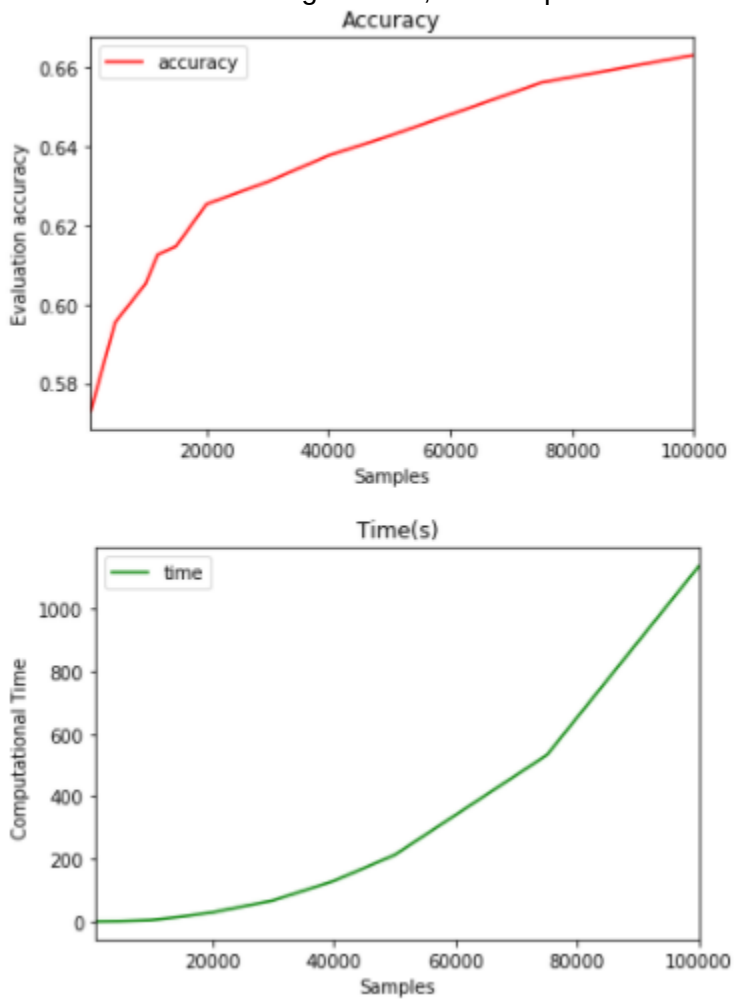


Fig: Accuracy of SVM increasing the training sample's size and computational time.

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.

4.RESULTS

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

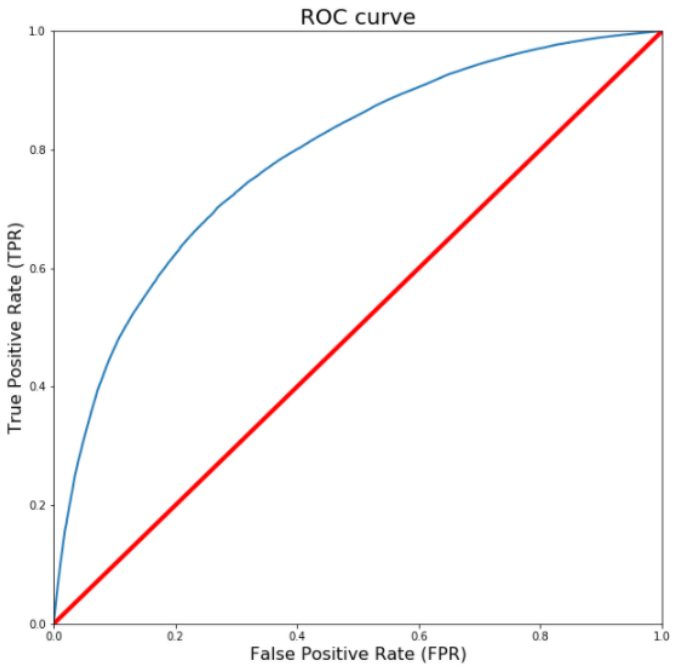
In this case, the recall is more important than the precision as a high recall will favor that all required resources will be equipped up to the severity of the accident. The logistic regression, KNN, and SVM models have similar accuracy, however the computational time from the regression is far better than the other two models. With no doubt the Random Forest is the best model, at 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.

5.DISCUSSIONS

The best performing model was able to achieve 68% accuracy in the. 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 to know right now, but at the incredible peace that technology is evolving nowadays, soon cars will be able to track them so that the emergency services could use them.
One problem is features had is that the target of this classification 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 a regression model. The next step on this problem could be to add a accident prediction model able to not just predict the accuracy but also the critical time and spots where potential accidents can occur in advance.

6.CONCLUSION

In this study, 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 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 and state. 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.


ROC curve
True Positive Rate (TPR)
False Positive Rate (FPR)
0.0
0.2
0.4
0.6
0.8
1.0