Predicting Car Accident Severity

Author:- Prabal Ganguly
Date:- September 2020

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 1.35 million road accidents deaths globally in 2016, with millions more sus-taining serious injuries and living with long-term adverse health consequences. Globally, road car crashes are a leading cause of death among young people, and the main cause of death among those aged 15429 years. Road car accidents caused 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 nowadays available, an extensive analysis to predict car accidents and its severity would make a difference to the death toll. Analysing a significant range of factors, including weather conditions, locality, type of road and lighting among others, an accurate prediction of the severity of the accidents can be performed. Thus, trends that commonly lead to severe car accidents can help identify the highly severe accidents. This kind of information could be used by emergency services, to send the exact required services and equipment to the place of the accident, leaving more resources available for accidents occurring simultaneously. Moreover, this severe accident situation can be warned to nearby hospitals which can have all the equipment ready for a severe intervention in advance.

Consequently, road safety should be a prior interest for governments, local authorities and private companies investing in technologies that can help reduce accidents and improve overall driver safety.

1.2 Problem

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. 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 efficient 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 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 data is divided in 5 different data sets, consisting of all the recorded accidents in France 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 specifics such as the gradient, shape and category of the road, the traffic 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 the holiday one labels the accidents occurring in a holiday. All the data sets share the accident identifications number.

An initial analysis of the data was performed for the selection of the most relevant features for this specfic 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 of 839,985 samples, each one describing an accident and 29 different features.

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, traffic regime, number of lanes, road problems, road shape, surface condition, situation, school nearby and infrastructure. 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 years. involved in the accident. 4. Severity: Maximum damage suffered by any user involved in the accident. Unscathed or light injury (O), hospitalized wounded or death (1)
5. 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 dataframe as more than a 50% of its values were NaN or O 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 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 problems 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 O 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.