Predicting Traffic Accident Severity.

Muhammad Samiullah

August 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 traffic deaths globally in 2016, with millions more sustaining serious injuries and living with long-term adverse health consequences. Globally, road traffic crashes are a leading cause of death among young people, and the main cause of death among those aged 15–29 years. Road traffic 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 traffic 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 traffic incidents can help indentifying the highly severe accidents. This kind of information could be used by emergency services, to send the exact required staff 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 acci-dent 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 off course the severity of the accident. This projects aims to forecast the severity of accidents with previous information that could be given by a witness informing the emergency services.

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 flow and type of vehicle, and the *holiday* one labels the accidents occurring in a holiday. All five 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 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 where the following:

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 where: road category, traffic regime, number of traffic lanes, road profile, road shape, surface condition, situation, school nearby and infrastructure.

The users 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 y.o. 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.