CAR ACCIDENT SEVERITY

Pavithra Kollipara

Introduction : Business Problem

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.

Analysing a significant range of factors, including weather conditions, special events, roadworks, traffic jams among others, an accurate prediction of the severity of the accidents can be performed.

These insights could allow law enforcement bodies to allocate their resources more effectively in advance of potential accidents, preventing when and where severe accidents can occur as well as saving both time and money. In addition, this knowledge of a severe accident situation can be warned to drivers so that they would drive more carefully or even change their route if it is possible or to hospital which could have set everything ready for a severe intervention in advance.

Governments should be highly interested in accurate predictions of the severity of an accident, in order to reduce the time of arrival and thus save a significant amount of people each year. Others interested could be private companies investing in technologies aiming to improve road safeness.

This project consists of several parts divided in two different notebooks.

Data Description :

Data source:

The data can be found in Kaggle.

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 tra c 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, and the holiday one labels the accidents occurring in a holiday. All datasets share the accident identification number.

Description:

The dataset that resulted from the feature selection consisted of 839,985 samples, each one describing an accident and 29 different features. These features were 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 were : road category, traffic regime, number of lanes, road shape, surface condition, situation, school nearby and infrastructure.