Car Accidents Severity Analysis

Capstone using Machine Learning Models

Coursera IBM Data Science Certification

Josue Irizarry

Introduction

Background

Based on the https://www.driverknowledge.com/car-accident-statistics/ information, the facts are along 6 millions of car accidents on average every year in the U.S., 90 people die every day in car accidents. In addition 3 millions of people are injured every year in car accidents, where 2 millions of them experience permanent injuries every year. Is wondering that 1 of 7 people only use the seatbelt, using the seatbelts the deather are reduced by 45% and cutting the risk of injury by 50%.

Do you know that 6% of crashes result in fatalities, 27% on non-fatality injury and in overall 72% result in property damage?

Roughly the experts understand the causes of severity of accidents are due alcohol, speeding and reckless driving. Another reason is distracted driving, where each day, more than 9 people are killed due to distracted driving. Distracted driving means people using their cell phone, texting.

Did you know Montana from 2011–2016 the motor vehicle fatality rate is 19 per 100,000 people compared to the national wide rate that is 11 per 100,000.

Regarding World wide, the number of road fatalities per 100,000 inhabitants per year is 18.2, resulting in 1,300,000 total fatalities in the data obtained from WHO in 2016. In the United States alone the rate is 12.4 of road fatalities per 100,000 inhabitants per year, resulting in 39,888 from the WHO report available from 2018.

Problems

Based on the statistics described above, everyone that uses the car for transportation is exposed to the risk of an accident and a stress of uncertain things that could happen in case of a car accident. As well the car accidents can have a financial impact on the individual, families and insurance.

The CDC defines the Road Traffic Injuries and Deaths, as a Global Problem. Resulting in road traffic accidents leading to the cause of the death in the United States in people aged 1 to 54 and the main reason of death for U.S. citizens or travelling abroad.

Worldwide, it is not uncommon to see the roads are shared in addition to cars, by buses, trucks, motorcycles, pedestrians and animals, in less or more degree in each country.

Image the scenario that you are planning to drive for a business trip or vacations to a place that you may not ever visit before or a place with a lot of people driving. In this scenario adding the complexity of the weather, road conditions and other factors. What about in this day that the data is the new petroleum use the existing (historical) public datasets and available web information to develop a machine learning model that you have an added value and know the factors results in car accidents and in the base case avoid it, doing predictive or prescriptive data analysis.

In this project the goal is to develop a machine model, perform features selection, and what is the best performing supervised learning model to perform the data analysis.

Stakeholders

The audience of this project is wided open to individuals or organizations interested in the reduction of car accidents severity drived by through a machine learning model.

Data

This project relied on the public car accident dataset in CSV format from the Seattle Police Department, with more than 190,000 collected records in a period of year 2004 to 2020. The selected data set is composed of 38 columns, between category and numerical features.

Data Preparation:

This following metadata available from https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf will be used to support the data preparation phase. Briefing reviewing the data is understood the data is far to be perfect, required to remove missing values or replaced by arbitrary value, transformation and balance the data. The dependent variable is the severity code column. The result of the data preparation is to construct the final data to be the modeling tool.