Accident Severity Prediction in USA

IBM CAPSTONE PROJECT

Nandan Rajeev | September 2020

Introduction

With the continuously growing number of automobiles on the road, the road safety issues have been on the rise. Globally, approximately 1.35 million people die in road crashes each year, on average 3,700 people lose their lives every day on the roads. In the USA alone, more than 38,000 people die every year in crashes on roadways. The U.S. traffic fatality rate is 12.4 deaths per 100,000 inhabitants. Road crashes are the leading cause of death in the U.S. for people aged 1-54.

What can be done to reduce the risk of being in a road crash? Unfortunately, we can't control the weather and get rid of poor driving conditions. Can we autonomously control all the cars on the road so that the driver mistakes can be minimized? Maybe in the future we can but right now, it's not a feasible solution. Predicting accidents is difficult, but not impossible. Various factors come into play which can decide the severity of an accident. Such as, car speed, weather, driver experience, location of incident, road type and so on. Can we make use of such features to predict the possibility of accidents and their severity? Yes! By studying and analyzing previous recorded data of accidents, we can construct predictor models which can help us predict the severity of accidents which will enable us to take adequate measures in reducing the risk or severity of the accident.

In this project, I will be using machine learning models to predict the severity of accidents by taking the various features as input. We will also come across various insights into the data such as – Which state has the highest number of accidents? Which factor contributes the largest to the severity of the accident?

Such models can be utilized to predict accidents in real-time which can lead to a considerable decrease in the dangers of road accidents.

Data

The data used in this project is a dataset consisting of road traffic accidents that occurred in USA between February 2016 and June 2020, covering all 49 states. The dataset consists of around 3.5 million entries. This data has been collected by Lyft, for analyzing the delays caused by accidents.

Our target variable here, will be the severity rating. Higher the severity of an accident, greater the traffic delay as the roads are often partially closed while the medical and road services are at work. By relating the traffic delay to the severity of the accident, the accidents are assigned a severity rating from 1 to 4 with 4 being the most severe.

The independent variables which will be used as the features here consist of entries such as – Visibility, Presence of speed bumps, Presence of roundabouts, and so on. By using these as the features, the model will be trained and tested to accurately predict the outcome. Different ML Algorithms will be applied and the best will be chosen based on their accuracies on an out-of-sample set.

The dataset will be split into 3 sets – Training Set, Test Set and Validation Set. We require the validation set to get better out of sample accuracy as certain ML methods will utilize the test set in the model building process (for example, K-Nearest Neighbors Method will utilize the test set for choosing the best k value).