Coursera Capstone - Predicting Accident Severity

Arabinda Behera

August 31, 2020

1 Introduction

1.1 Background

The World Health Organization describes the road traffic system as the most complex and the most dangerous system with which people have to deal every day. Millions of people around the world die in road accidents each year. The affected people and their family go through mental stress and face financial problems. Road accidents also puts pressure on police and like insurance agencies. They have to be prepared and be efficient in dealing with unexpected cases. The study of road accidents is very crucial and some models for predicting accident severity would be very beneficial for police, insurance and other agencies to deal with accidents efficiently. It would also be helpful for the awareness of general public on the important causes and patterns and may help prevent accidents to some level.

1.2 Problem Statement

The numbers and the severity of the accidents can depend on several factors like location, time, traffic conditions, weather conditions, state of the vehicle, driver's physical and mental state, etc. Predicting the probability and severity of accidents using simple statistical models is difficult. A huge amount of open access data is available on road accidents from different governments around the world. The important question is can this vast data be used to predict accident severity?

This project aims on developing machine learning models using available and labeled data to predict severity of road accidents.

2 Data

2.1 Data Source

The data for this project is taken from Kaggle. The link is given below :
https://www.kaggle.com/tsiaras/uk-road-safety-accidents-and-vehicles

The data originally comes from the Open Data website of UK government published by the Department of Transport. The dataset contains detailed information on the road accidents from different places in the country. The dataset contains two csv files :

- Accident_Information.csv: each row represents a unique traffic accident (identified by the Accident_Index column), featuring various properties related to the accident as columns. Date range: 2005-2017

- Vehicle_Information.csv: each row represents the involvement of a unique vehicle in a unique traffic accident, featuring various vehicle and passenger properties as columns. Date range: 2004-2016

2.2 Data Cleaning

2.3 Feature Engineering

3 Modeling

4 Evaluation

5 Results

6 Summary and Discussion