Car Accident Severity Prediction

Introduction

As per a report from USA's CDC, 1.35 million people die each year due to road accidents, apart from the damage to property. If we can understand why road accidents happen, if we can identify the important factors behind accidents, we can create a curative plan to reduce the number of road accidents. This report aims to identify the factors behind accidents. This analysis can be used in a plethora of industries and scenarios. For e.g. Home deliveries have been increasing in the last 10 years, and has recently shot up due to COVID-19. All the delivery agent traversing the last mile (from Pick up point to a customer's address) can benefit from this analysis. If our model states that a given stretch of road has a high risk of accident, routing algorithms (like Google maps) can direct the riders towards a different road.

This report will identify the important factors behind road accidents and predict the severity of road accidents in Seattle Area. This will be useful for police departments as they can identify the accident hotspots for better manpower allocation. This can also be utilized by last mile delivery agents.

Data

I am using the data collected by Seattle City's Police Department from 2004 to Feb 2020. It has around 194k+ observations regarding the reported road accidents during the past 15+ years. Each observation can have 38 attributes. Some of the attributes are SEVERITYCODE, LOCATION, STATUS, VEHCOUNT, etc.

For e.g. SEVERITYCODE corresponds to the severity of the collision. '3' represents a fatal accident, whereas '1' represents property damage. Similarly, ADDRTYPE lists out the type of address (Alley, lock or intersection) where the accident has occurred. A detailed description of attributes can be found at link.

Most of the attributes are categorical in nature and will have to be encoded for analysis. For e.g. JUNCTIONTYPE explains the category of junctions where the accident has taken place. It has the following distinct values - At Intersection (intersection related)'; 'Mid-Block (not related to intersection)'; 'Driveway Junction'; 'Mid-Block (but intersection related)'; 'At Intersection (but not related to intersection)'; NAN; 'Unknown' and 'Ramp Junction'. Post encoding it was changed to values 0, 1, 2, 3, 4, 5 and 6.

I used the 'ADDRTYPE', 'COLLISIONTYPE', 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'SDOT_COLCODE', 'UNDERINFL', 'ROADCOND', 'LIGHTCOND' and 'HITPARKEDCAR' columns as independent variables to build the model. 'SEVERITYCODE' is the dependent variable for the model.

Atul Anand