Car Accident Severity: Seattle, Washington

(Applied Data Science Capstone – IBM)

This project aims at understanding what factors play a vital role in the severity of car accidents in Seattle, Washington using Data Science Toolkit and predicting them prior to take necessary measures to avoid them using Machine Learning techniques.

Name: Vedant Mane

Project: Car Accident Severity Prediction

Course: Applied Data Science Capstone

Specialization: IBM Data Science Professional Certificate

Understanding Data:

The dataset requires pre-processing so that the machine learning model classifies the prediction to a categorical variable. The dataset has a total of 37 features and 194673 observations with variation in number of observations for every feature. The total dataset has high variation in the lengths of almost every column in the dataset due to the missing data values. These values could have been beneficial for our prediction algorithm had the data been present. The Metadata information for the dataset can be found at the link here.

Pre-processing Data:

The model aims to predict the severity of the car accident, considering that, the variable "SEVERITYCODE" of Severity Code that was in the form of 1 (Property Damage only) and 2 (Injury Collision) were encoded as 0 (Property Damage only) and 1 (Injury Collision). The attribute "INATTENTIONIND" for Driver's Attentiveness at the time of accident was considered 1 were value was 'Y' and 0 where data appeared missing. Similarly, the attribute "UNDERINFL" was encoded as 0 for the values 'N', '0', 'Nan' while 1 for 'Y', '1'. In the case of "SPEEDING", the 'Y' value was considered as 1, while missing values were considered to have the value 0. The Light conditions were at the time of accident were categorized into 4 groups based on the type of light and time of day as Daylight, Dark with Street Lights, Dark without Street Lights & Dark with unknown lighting conditions. Similarly, Weather Conditions were categorized as Clear, Cloudy or Low Visibility, Windy, Pouring for Raining or Snowing or Sleet/Hail/Freezing Rain and Unknown for missing or unknown values. The condition of the roads was classified as Dry, Wet, Blocked or Unknown conditions. The "JUNCTIONTYPE" attribute was classified into 2 types based on whether the accident took place at the intersection or elsewhere. Then, dummy variables were created for all the categorical columns (ROADCOND, LIGHTCOND, WEATHER).

Further we will normalize this data using the StandardScaler function and fit this data for our dataset and transform the same for model building.

Feature Selection:

Feature	Description
Location (Address Type)	Description of the general location of the Collision.
Weather Condition	A description of the weather conditions during the time of the collision.
Car Speeding	Whether or not speeding was a factor in the collision.
Light Conditions	The light conditions during the collision.
Road Condition	The condition of the road during the collision.
Junction Type	Category of junction at which collision took

	place
Number of People involved	The total number of people involved in the Collision.
Number of Vehicles involved	The number of vehicles involved in the collision.
Driver's Attentiveness	Whether or not the Driver was attentive.
Driver under influence	Whether or not the Driver was under influence.
Severity Code	A code that corresponds to the severity of the collision: • 1 — Property Damage • 2 — Injury Collision