Week 2 – Capstone Project
Severity of Car Accidents in Seattle
Problem and Data Understanding

1. Description of the problem and a discussion of the background

1.1 Background and Problem Understanding

Car accidents are one of the most common hazards we all face in daily life. From minor fender-benders to catastrophic, multi-car pileups, getting into an accident in a motor vehicle can end a life or change it forever. The Seattle Department of Transportation's annual traffic report illustrates the constant challenge to the city posed by car accidents.

Different tropic conditions, locations, weather conditions, road conditions, light conditions, day of the week, junction type, speed range and other types of factors are major attributes causing the car accidents.

In this project, we focus on the subject of predicting the severity of an accident in the city of Seattle. Some attributes will be evaluated like weather and road conditions which contribute in the severity of the car accidents.

1.2 Target Audience

This project will be most beneficial for:

- People who travel on a regular bases by car.
- Truck drivers.
- Police officers who want to reduce the accident rate and severity.

2. A description of the data and how it will be used to solve the problem.

2.1 Data resource

The data was retrieved from **https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv**

This data for the capstone project was already provided by Applied data Science Capstone course by Coursera.

2.2 Data Description

The Collisions dataset includes records of collisions that happened on road from 2004 to Present. The dataset contains 194673 rows and 38 columns, each row is a record of the accident, and each column is an attribute. The first column "SEVERITYCODE" is the labeled data, which describes the fatality of an accident. The remaining 37 columns have different types of attributes. Some or all can be used to train the model.

There are some problems that need to be fixed in this dataset:

- There are missing values which needs to be removed
- Some columns need to be removed as they are not helpful in building our model
- The data type of some attributes is not correct. For e.g some data needs to be changed from object to integer
- The data has unbalanced labels which needs to be balanced

2.3 Data usage for the solution of the problem

2.3.1 Catplots and Distplots

Catplots and Distplots have been created for the severity of the accidents. The target variable “SEVERITYDESC” i.e a detailed description of the severity of the collision. Independent variables that have been chosen are “ADDRTYPE”, “COLLISIONTYPE”, “JUNCTIONTYPE”, “SDOT_COLCODE”, “ST_COLCODE”, “UNDERINFL”, “ROADCOND”, “LIGHTCOND”, “WEATHER”, and “PERSONCOUNT”.

1. **ADDRTYPE** - Collision address type:
   • Alley
   • Block

- Intersection

2. **COLLISIONTYPE** – Collision type
3. **JUNCTIONTYPE** - Category of junction at which collision took place
4. **SDOT_COLCODE** - A code given to the collision by SDOT.
5. **ST_COLCODE** - A code provided by the state that describes the collision.
6. **UNDERINFL** - Whether or not a driver involved was under the influence of drugs or alcohol.
7. **ROADCOND** - The condition of the road during the collision.
8. **LIGHTCOND** - The light conditions during the collision.
9. **WEATHER** - A description of the weather conditions during the time of the collision.
10. **PERSONCOUNT** - The total number of people involved in the Collision
11. **SEVERITYDESC** - A detailed description of the severity of the Collision

2.3.2 Seattle map for the Severity of Car Accidents

Seattle map for the occurrence of car accidents was built.

2.3.3 K-Nearest Neighbors, Decision Tree, Logistic Regression and ROC Curves

K-Nearest neighbors, decision tree and Logistic Regression algorithms were used for the classification problem to predict the severity of accidents. “SEVERITYCODE” was used as the target variable

SEVERITYCODE - A code that corresponds to the severity of the collision:

- **3**—fatality
- **2b**—serious injury
- **2**—injury
- **1**—prop damage
- **0**—unknown

ROC curves were then build to interpret the above models.