1. Business Problem

1.1. Background

Every year, car accidents cause more injuries and deaths than any other type of personal injury. There are many factors that go into whether a person gets into a car accident that the severity of the event. With the progress of technology included in cars and their new capabilities, it would be important to have the tools and means available to provide drivers with a warning, given the weather and road conditions about the possibility of getting into a car accident and how severe it would be. Therefore, the driver would drive more carefully or even change his/her driving route if he/she is able to do so.

1.2. Problem

Data that might contribute to determining the severity of a car accident may include weather, road conditions, lighting conditions, and other metrics that would help describe the surrounding environment of the driver. This project aims to predict how severe an accident would be based on these data

1.3. Interest

Obviously, it would be of great interest for drivers to have to avoid an accident from knowing how severe it may be by driving more carefully. This will also enable transport, security, and emergency agencies all around the world to have different predictive models for quickly analyzing crashes when they happen and dispatch an appropriate response swiftly.

2. Data

2.1 Data Sources

The data utilized for this report came from collision data provided by SPD and recorded by Traffic Records. This dataset includes all types of collisions from 2004 to present, where collisions will display at the intersection or mid-block of a segment.

2.2 Data Cleaning

The data was downloaded from the course organizers at IBM. I first dropped all rows with missing values, as they would not contribute to predicting the severity of an accident. Similarly, rows containing ambiguous values were also dropped. For example, under the WEATHER column, the rows that have “Other” or “Unknown” have been dropped.

2.3 Feature Selection

The columns that were decided to be kept from the original dataset to predict our severity score, SEVERITYCODE, is as follows:

- ADDRTYPE: Collision address type
- PERSONCOUNT: The total number of people involved in the collision
- VEHCOUNT: The number of vehicles involved in the collision
- WEATHER: A description of weather conditions during the time of the collision
- ROADCOND: The condition of the road during the collision
- LIGHTCOND: The light conditions during the collision

The other columns in the dataset are either redundant to the features chosen above or do not have any significance in determining accident severity.