WEEK1 – Report(Car Accident severity)

Introduction

Background

Seattle is the largest city in the state of Washington that has around 745000 people. The rate of people growth is 2.1% where is the second-fastest growth in US cities. Also, the number of tourism people is increased. Also, the number of workers is increased. So, all these factors could affect transportation where the number of cars is increased which lead to an increase in the number of accidents. Some of the accidents are dangerous and some not. It is important to know the severity of accidents and what are factors that affect them by predicting the severity of car accidents in future by using data of previous accidents.

Problem

Data that might contribute to determining the severity of accidents might include weather, speed, road condition, the light of road, address type, and the pedestrian right of way was granted or not. The aim of the project is to predict whether of car severity of accidents.

Interest

Seattle department of transportation would be very interested in accurate prediction of SEVERITY of accidents of cars. Also, this prediction could help to solve some factors that affect car accidents.

Data acquisition and cleaning

Data sources

Transportation SeattleCityGIS has shared the dataset in its website. Dataset contents data of report number, location, rain, weather, date, severity code, road condition, light condition, speeding, State Collision Code and so on.

Data Cleaning

There were a lot of missing values that could affect the prediction. I decided to add the results of some missing values of data that can add them or change. For example, the column of SPEEDING has yes, or No value and some records have a null value. So, I change the null value to No. This process does it with INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SPEEDING, and JUNCTIONTYPE columns.

Another clean data method is removed all rows that has null values and cannot edit the value of columns such as the column of State Collision Code.

Also, I have used column of date to get the year of the accident as well as the day of the accidents are weekends or weekdays.

After that, I will change the columns name to be clear and I can use them for exploring data.

Feature Selection

When I cleaned data, I got 187504 sample and 38 features. There are some redundant features that could not use for predication. For example, duplicate columns such as severity code and date timestamp. Also, I dropped some columns that could not use as features such as report number, longitude, latitude, unique key for the incident, Secondary key for the incident, Key

that corresponds to the intersection associated with a collision, and so on. After this process, I got around 19 columns. It could be changed when I start experiment.