Car Accident Severity

IBM Capstone Project

Prepared by: Yevgeniy Dobrynin

13-Oct-2020

Introduction

Statistically the most dangerous transport is car. Approximately 1.35 million people die each year as a result of road traffic crashes. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury. Seattle city has an intensive traffic and as the result frequent collision accidents with fatalities and heavy injuries. From 2004 to 2020 it was recorded 78,596 injuries and 328 fatalities as the result of road accidents in Seattle city. Analysis of such data could reveal dangerous areas and other factors leading to high probability of accidents to occur.

Stakeholders who may benefit from the analysis are: drivers, insurance companies, hospitals and police.

Data Preparation

The information was provided by Seattle Police Department from 2004 to 2020 containing 221,737 records.
Database was obtained from the following link:
https://data-seattlecitygis.opendata.arcgis.com/datasets/collisions
Dataset contains 39 columns of accident records.

Using the Severity Code assigned to each accident we could identify accident severity

- 0: Unknown/no data
- 1: Property damage only
- 2: Minor injury collision
- 2b: Major injury collision
- 3: Fatality collision

Dataset in their original form is not suitable for analysis, thus the following preprocess were applied:

1. Drop columns that is not going to be used in analysis. 26 columns were dropped in total including unnecessary and redundant columns.
2. Data incompleteness. Around 15% of the data were missing one or more key features required for analysis. As the result of data cleaning and preprocessing such data were dropped from the dataset including NaN and Unknown values.
3. Convert categorical data to numerical. In order to apply Machine Learning models it requires to convert all categorical data to numeric. As an example, “ROADCOND” column contains strings that can’t be directly processed by Machine Learning Engines. Picture below demonstrates conversion from categorical to numerical string.
4. The same transformation were also applied to “SEVERITYCODE”, “WEATHER”, “LIGHTCOND” and “SPEEDING”

Example of categorical to numerical conversion

Categorical data		Numerical data	
Dry	123439	0	123439
Wet	46329	7	46329
Ice	1093	1	1093
Snow/Slush	842	5	842
Other	102	3	102
Standing Water	99	6	99
Sand/Mud/Dirt	59	4	59
Oil	50	2	50

5. Data balancing and standardization. The target variable for this study is SEVERITYCODE which represents severity of accidents contains the following values

    1: Property damage only - 113156
    2: Minor injury collision - 55584
    2b: Major injury collision - 2945
    3: Fatality collision - 328

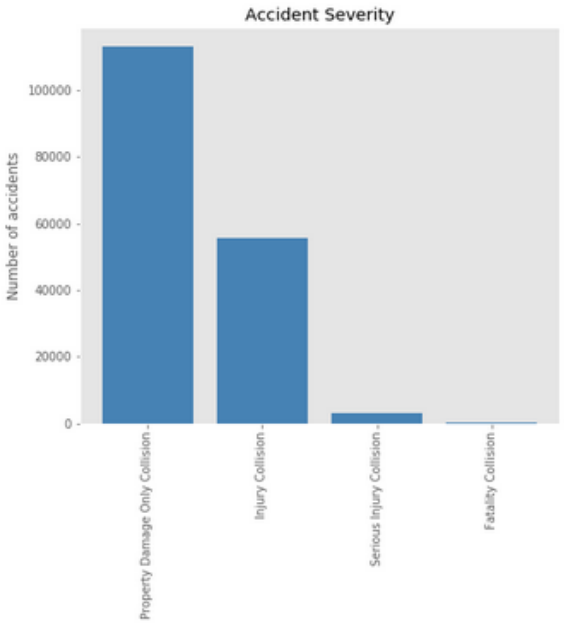
    This real-life representation of accidents outcome may bias the model, thus the model has balanced and binarized to the following 2 categories.

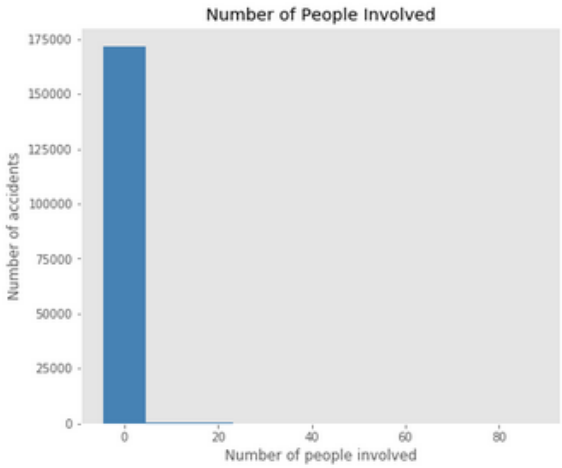
    0 – Minor accidents that will include property damage only
    1 – Major accidents that will include all type of injuries: minor, major injuries and fatality collisions

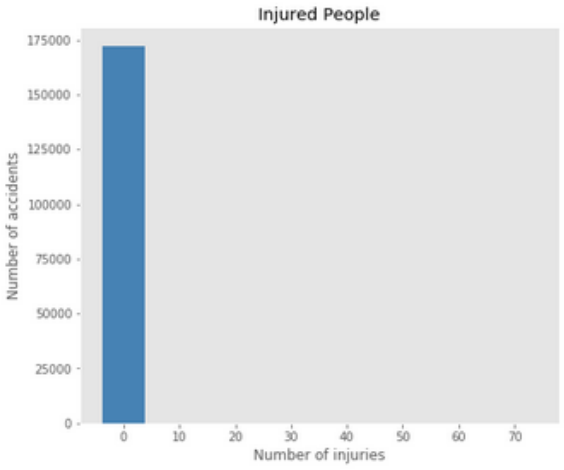
    Binarized and balanced result

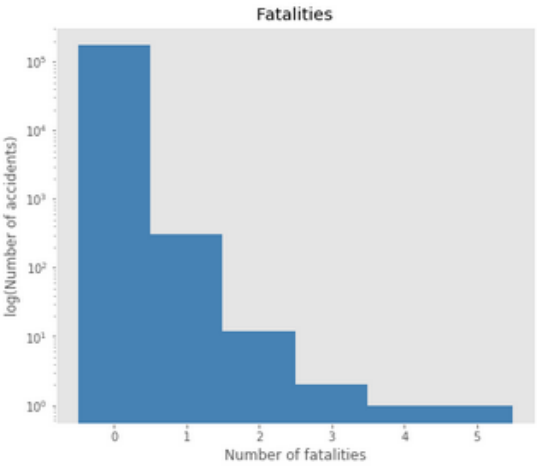
    0 – 58587
    1 – 58587

Plotting the data to see values distribution
Visualize quantities of accidents and number of people involved

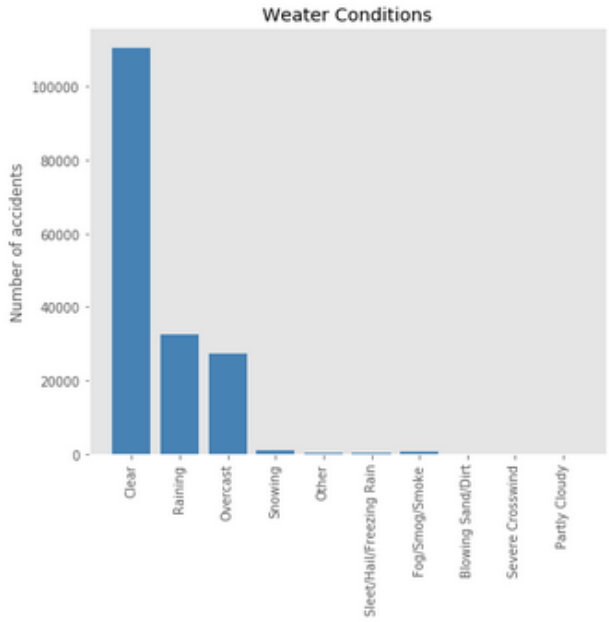


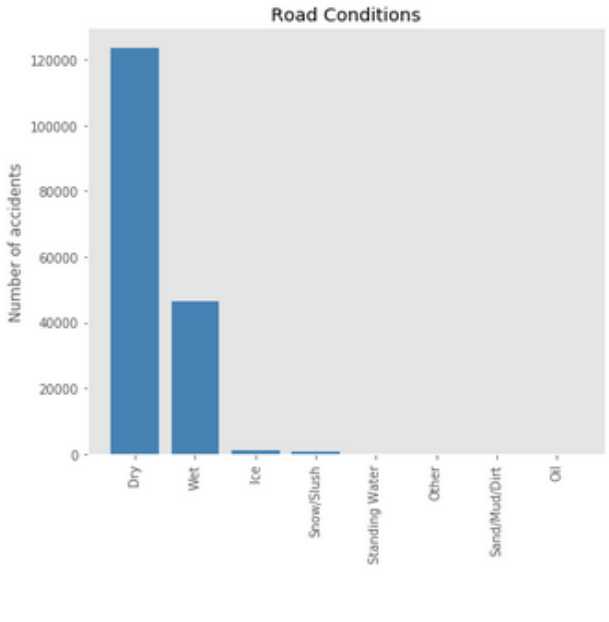


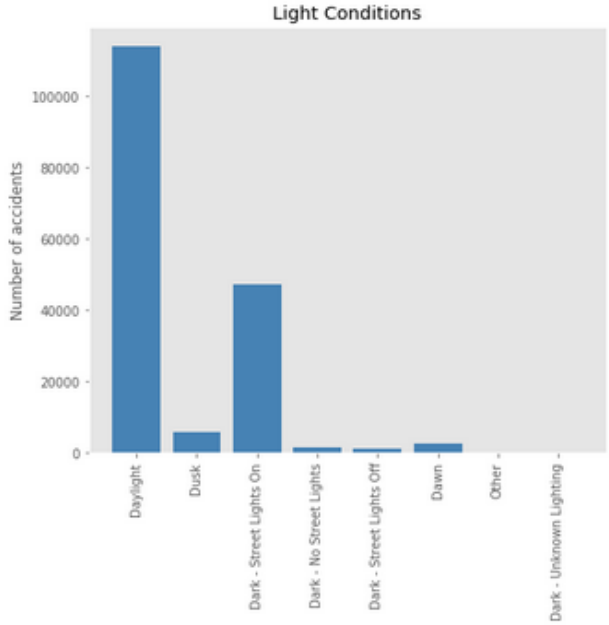


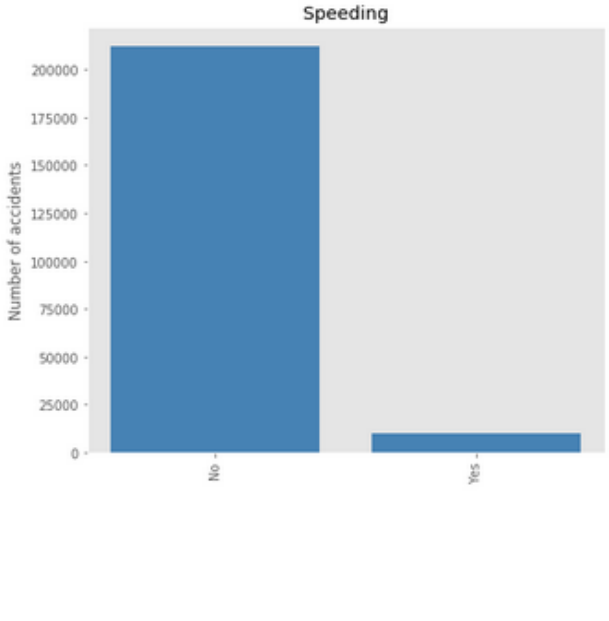


Visualize parameters that affect accident severity









Build correlation matrix to discover which columns need to be included for model building