Introduction

PROBLEM DESCRIPTION

This project is about analyzing the severity of collisions in Seattle area in the USA. As part of this capstone project we will be developing a predictions model to predict the collisions severity in Seattle area in the United States of America. The main problem: What is the severity magnitude for a collision to occur in Seattle?

Audience consists of people who lives in Seattle or travel through it and goal is to provide them information about current situation on a roads and possible dangers.

Audience should be interested in this problem because knowing the relationships between conditions and likelihood of collision can save their money and life.

DATA DESCRIPTION

This dataset contains about 6.5 mln car collisions records. The data include all types of collisions. Collisions will display at the intersection or mid-block of a segment. The data are collected from a timeframe of 2004 to Present. The data is updated on a weekly basis and is provided by SPD and recorded by Traffic Records.

Data set has 194673 rows and 38 columns. After checking the data there is a need to clean them and select main features used in the further analysis.

The analysis focuses on the following columns because they are relevant to the analysis:

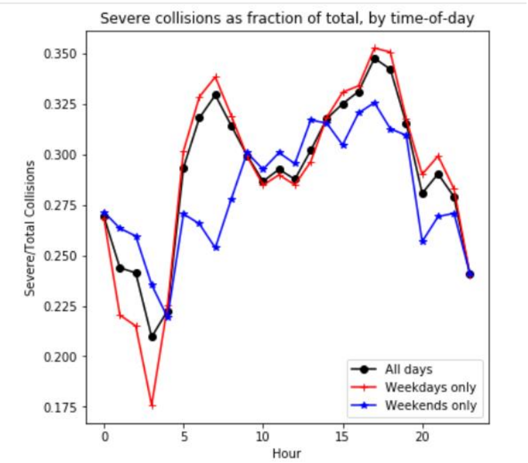
- COLLISIONTYPE: A keyword describing the collision, eg 'head-on', 'angled', 'cycles', etc.
- PERSONCOUNT: Total number of people involved in the collision.
- PEDCOUNT: The number of pedestrians involved in the collision.
- PEDCYLCOUNT: The number of bicycles involved.
- VEHCOUNT: The number of vehicles involved in the collision.
- INCDATE/INCDTTM: Date and time recordings for the incident records.
- INATTENTIONIND: If collision was due to inattention.
- UNDERINFL: If driver was under influence of drugs/alchohol.

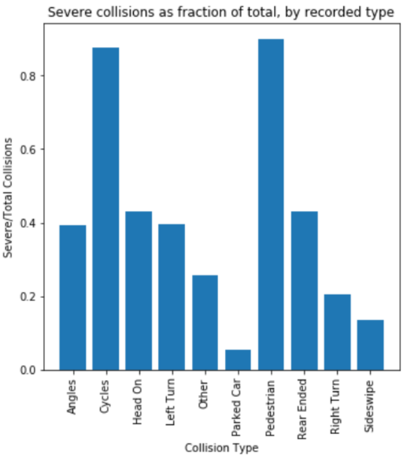
- WEATHER: Weather at the time of incident.
- ROADCOND: Condition of road at the time of incident.
- LIGHTCOND: Light conditions at the time of incident.
- SPEEDING: If speeding was a factor in the collision.

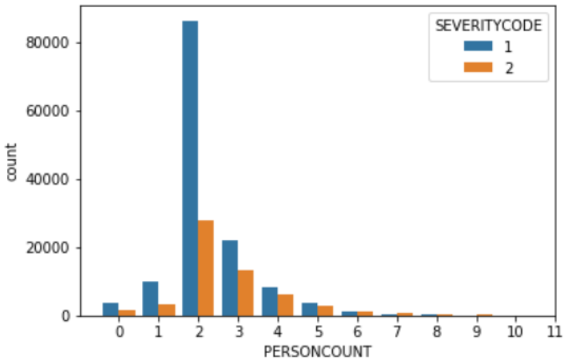
By utilising independent values in machine learning models I determined the SEVERITYCODE.

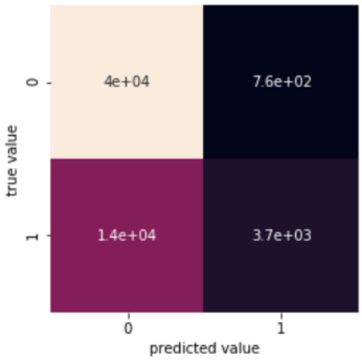
METHODOLOGY

I've done several analyses, mainly consisting of:
- check data types with df.dtypes
- check value counts of each column in order to find NaN/Unknown values/corrupted values
and accordingly adjusted them
- use of plots such as bar charts for visualising counts of data distribution
- use graphs to analyse trends
- Correlation and regression
- among many others









DISCUSSION

According to data provided we can predict with models what value will severity code have based on the model. It will be valuable to have complete dataset and minimise NaN values and also in some cases there is only one binary classifying value and it's impossible to deduct what the missing values are. These features will be useful for analysis of Severity of collision.

Logistics regression has a score of 0.75.

CONCLUSION

During the analysis we have observed several interesting relationships in data and this kind of classification problem is good to predict with a given accuracy because of minimal differences in Severity code 1 and Severity code 2 in reality. Generally, there is higher probability of SEVERITYCODE 1 collision, and they are mainly happening during daylight on a dry road, followed by dark - street lights on a wet road. The proportion of collisions resulting in injury is 1.28 times higher when speeding is involved compared to when it's not involved.