Predicting Accident Severity with Different Factors

Chi-Fat Wong

August 31, 2020

1. Introduction

1.1 Background

Traffic accidents have different level of severity and cause different levels of traffic jam and delayed in the transportation. The severity of traffic accidents depends on different factors such as the location of the accident occurred, the number of vehicles or people involved. The objective of this project is to develop a classification model to predict the severity of traffic accidents so that based on the traffic accidents attribute the model can alert travelers on the severity and the potential delays in the traffic.

1.2 Problem

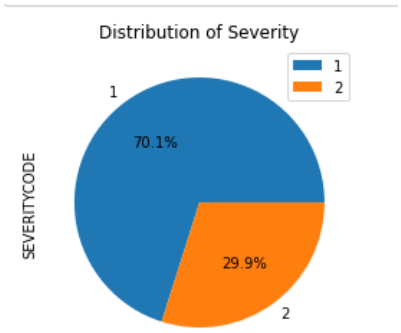
The problem is to determine the severity of traffic accidents using the traffic accidents data in Seattle city in United States.

2. Data acquisition and cleansing

2.1 Data sources

The traffic accidents data is obtained for the Seattle city in this link. The metadata description of the columns of the data is available in this link. The data period is from January 1, 2004 to May 20, 2020.

The column 'SEVERITYCODE' is target label that the model would predict. It is notated as 1 and 2 which represent prop damage and injury respectively. The distribution of the severity level of the data is presented as follow. About 70.1% of accidents were with level 1 and 29.9% were with level 2. Note that the distribution between level 1 and level 2 are not even and additional adjustments to the data might be required later to avoid biased classification model.



2.2 Data cleansing

There are in total 194,674 rows of data. Most of the data columns are categorical and some of them are numerical. The following columns are removed as they are the index of map objects or the reporting number of the accident which has no value for the classification problem.

- OBJECTID
- INCKEY
- COLDETKEY
- REPORTNO
- SEGANKEY
- CROSSWALKKEY

The column INCDATE is converted to date object in the data set in order to identify the date of the data.

2.3 Features selection

The following 16 features were selected for the classification problem as they are potentially related to the accident severity, including:

Categorical:

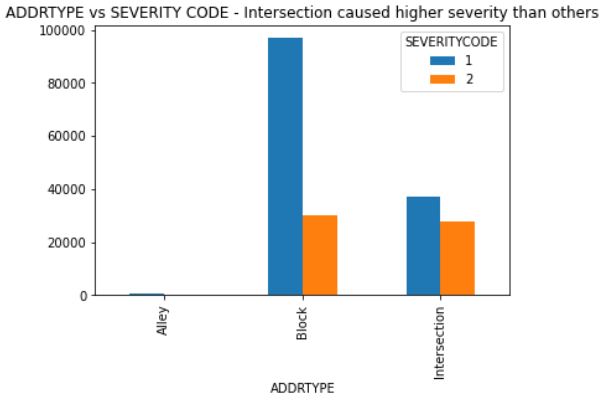
- ‘ADDRTYPE’
- 'EXCEPTRSNDESC',
- 'COLLISIONTYPE',
- 'JUNCTIONTYPE',
- 'SDOT_COLDESC',
- 'INATTENTIONIND',
- 'UNDERINFL',
- 'WEATHER',
- 'ROADCOND',

- 'LIGHTCOND',
- 'SPEEDING',
- 'ST_COLDESC',
- 'HITPARKEDCAR',

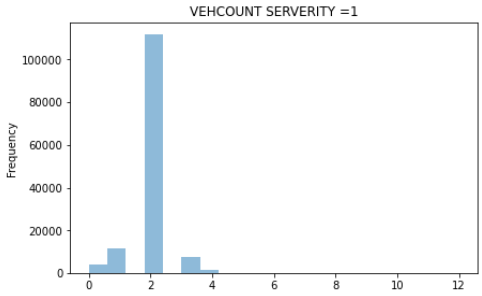
Numerical:

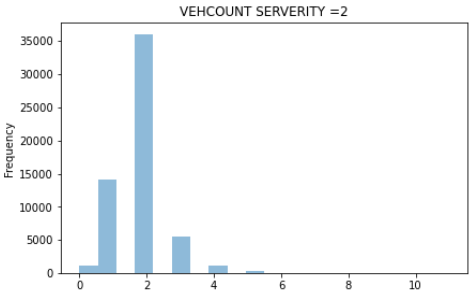
- 'PERSONCOUNT',
- 'PEDCOUNT',
- 'VEHCOUNT',
- 'PEDCYLCOUNT'

For categorical feature, we try to analyze whether the category has any relation to the severity. For example, the below is the ADDRTYPE category and the number of accidents with different severity. Obviously, ADDRTYPE of Intersection has a high chance of level 2 accident and the other two categories. We repeatedly plotting the relation of each categorical feature versus the severity to try to identify if any potential relationship exists.



For numerical feature, the histogram is plotted for the frequency of accidents' severity level for each feature. For example, the below shows the number of vehicle count involved in the accident and the frequency of the severity. However, there is no obvious relationship found.


VEHCOUNT SERVERITY =1
Frequency
100000
80000
60000
40000
20000
0
0
2
4
6
8
10
12


VEHCOUNT SERVERITY =2
Frequency
35000
30000
25000
20000
15000
10000
5000
0
0
2
4
6
8
10