Predicting the Severity of a Car Accident

1. Introduction

1.1 Background

A car accident can happen to everyone at every time in everywhere. There are several levels of an accident, starting from the light one, which involved no injured people to the level where it can kill a number of people. These level of an accident is known further as The Severity Level. When we talk about accidents, we cannot ignore the factors that contribute to the existence of the accidents, some of them are weather, road condition, light condition, driver condition, etc. These factors can collectively influence the severity level of an accident.

1.2 Problem

Our main problem in this project is that we want to try to predict a severity level of a car accident based on the features/attributes that we are going to choose from the given dataset.

1.3 Interest

This project might be useful for wide range of people, especially for them who travel a lot as a protective information so that they can safely do their driving. This project would also benefit the road management institution in work for increasing the safety level of the road.

2. Data Acquisition and Cleaning

2.1 Data Source

The data that we are going to use is a car accident data from Seattle Police Department and Accident Traffic Record Department from 2004 to present. The data consists of 37 columns and almost 200.000 rows of observation. This dataset is dominated by categorical and numerical data. The target data, severity code, is filled with different code with the type of integer that indicate different level of severity.

2.2 Data Cleaning

Real world data is a raw data that have not been processed, where it consists of many missing value, wrong format, unnecessary columns, duplicates, wrong value, etc. So, in order to maximize our data analysis process, we have to clean this raw data first. The most important thing in data cleaning process is handling missing value. There are two major ways for doing this, those are dropping or filling. Dropping means we drop all the missing values, while filling means we fill all the missing value with certain value. In our case, because the number of missing value is not too big, we then chose to drop it all.

We also change some value in our data, for example in column “UNDERINFL” we simply change all value 1 and 0 to Y and N.

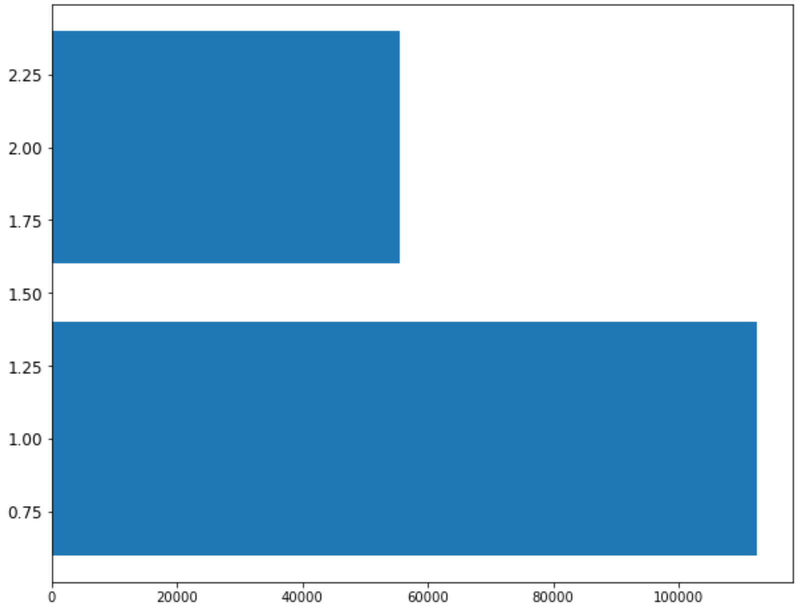
2.3 Features Selection

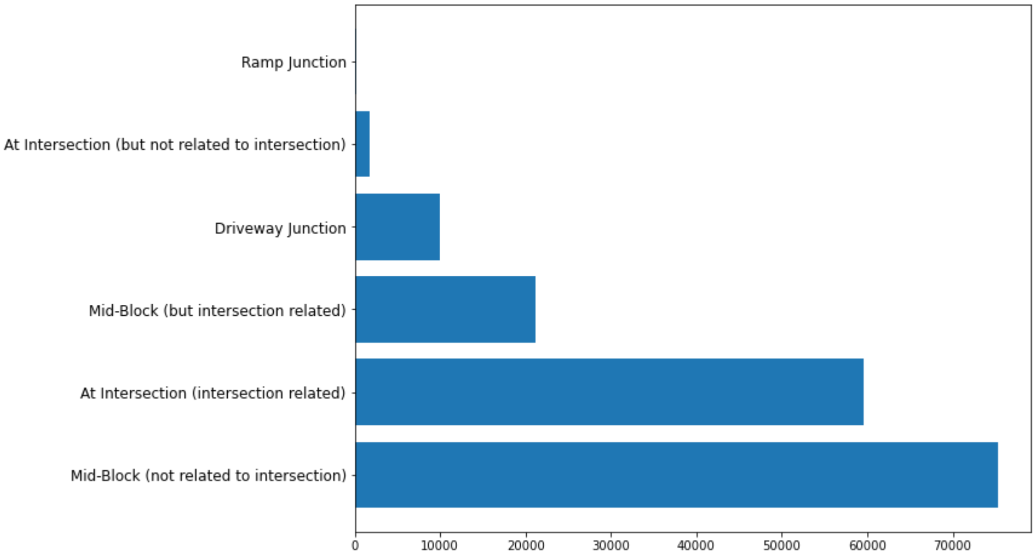
Not all features in our data can be useful for our model. Thus, we need to select some features that we think can make our model perform well. In order to do that, we can select it by simply using natural or common understanding. For instance, the level of severity of an accident should be logically influenced by weather, road condition, light condition, junction type, driver condition, and speed of the vehicle. So, we put this factors as our new feature along with severity code as the target.

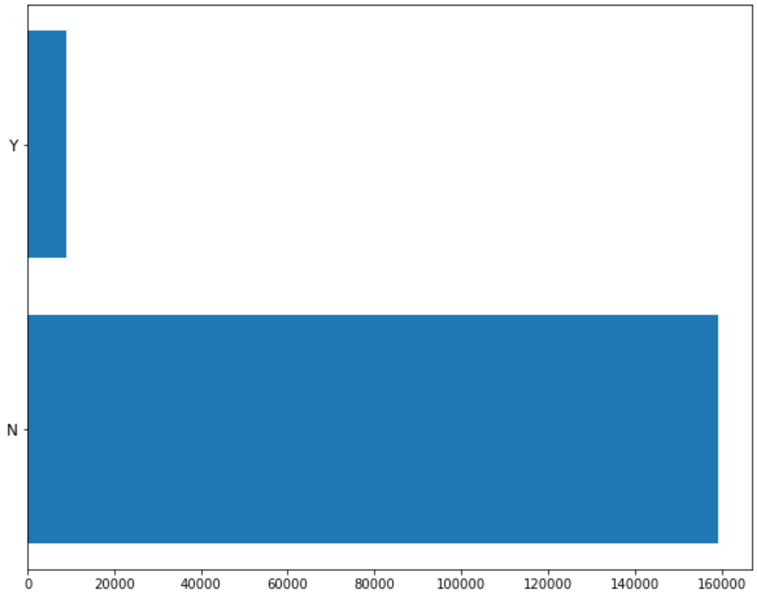
3. Exploratory Data Analysis

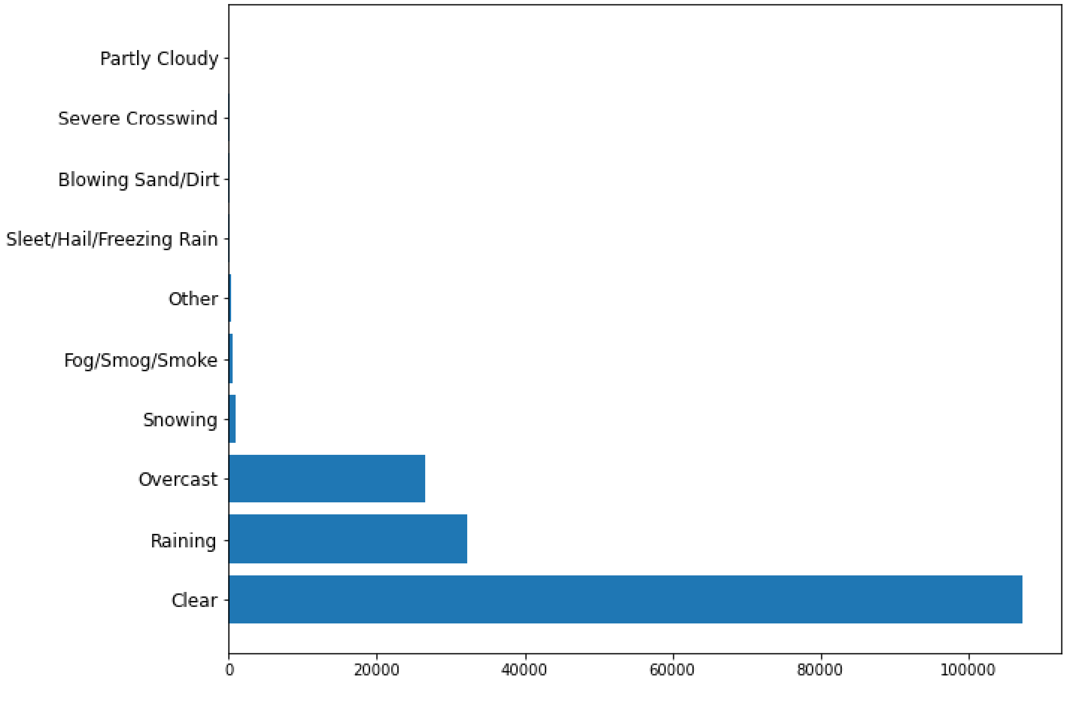
3.1 Univariate Analysis

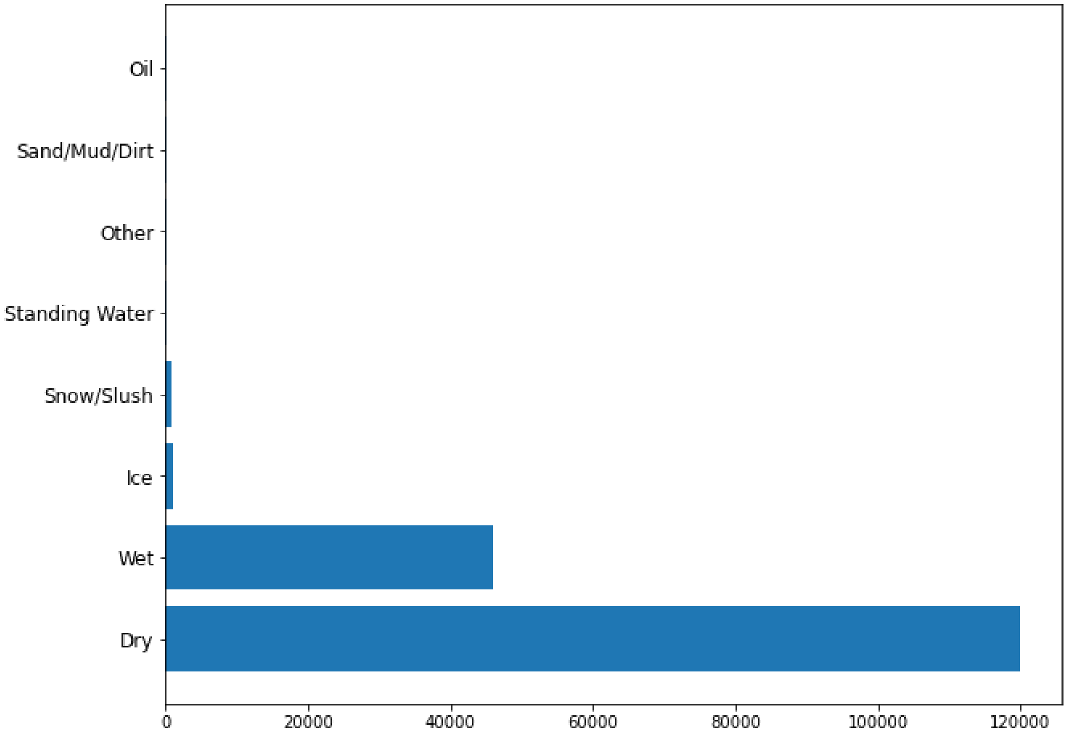
Univariate analysis is a type of analytics process that involve one variable only. In other words, we analysis every column in our datasets. This following figures is the result of our univariate analysis.

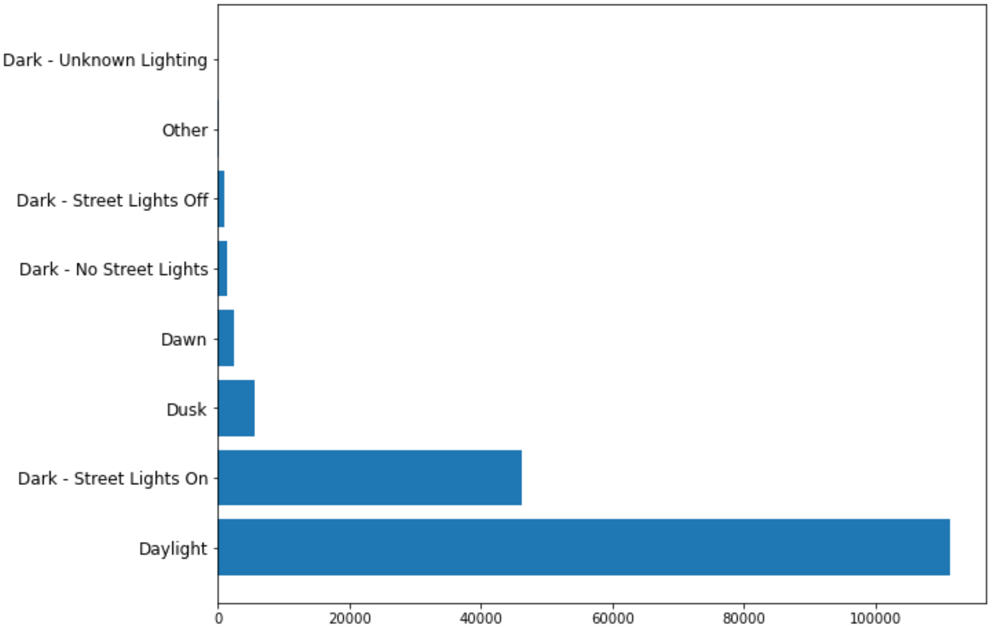


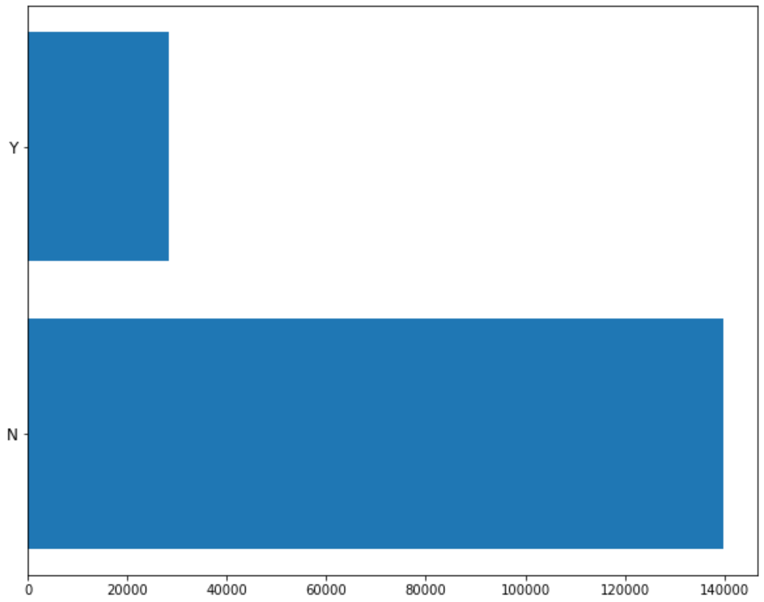


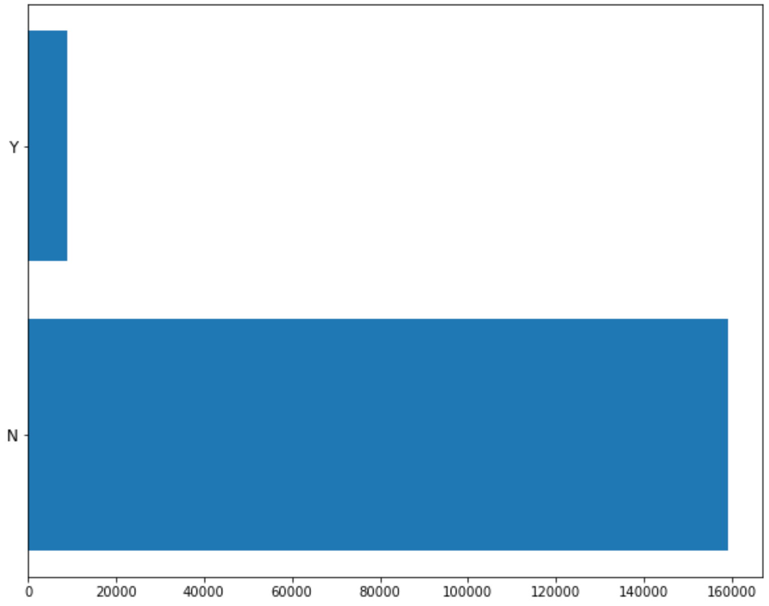

Y
N
0
20000
40000
60000
80000
100000
120000
140000
160000


Partly Cloudy
Severe Crosswind
Blowing Sand/Dirt
Sleet/Hail/Freezing Rain
Other
Fog/Smog/Smoke
Snowing
Overcast
Raining
Clear
0
20000
40000
60000
80000
100000


Oil
Sand/Mud/Dirt
Other
Standing Water
Snow/Slush
Ice
Wet
Dry
0
20000
40000
60000
80000
100000
120000


Dark - Unknown Lighting
Other
Dark - Street Lights Off
Dark - No Street Lights
Dawn
Dusk
Dark - Street Lights On
Daylight
0
20000
40000
60000
80000
100000

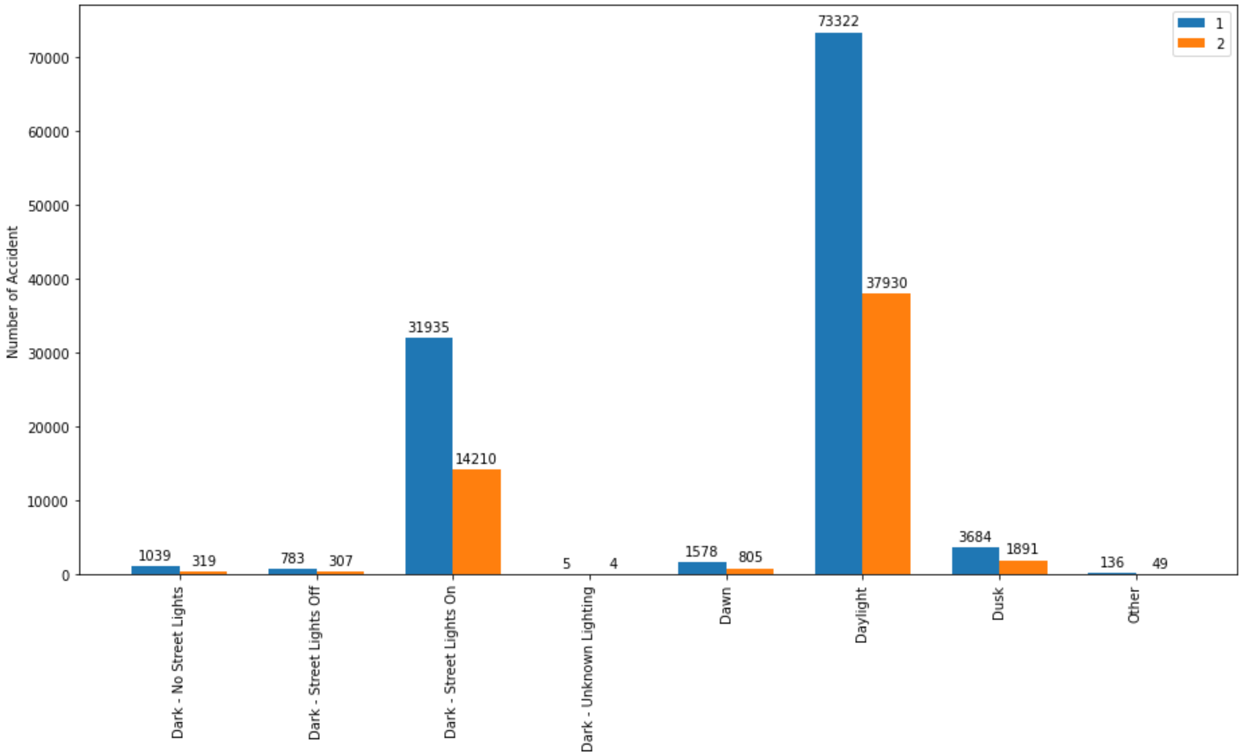


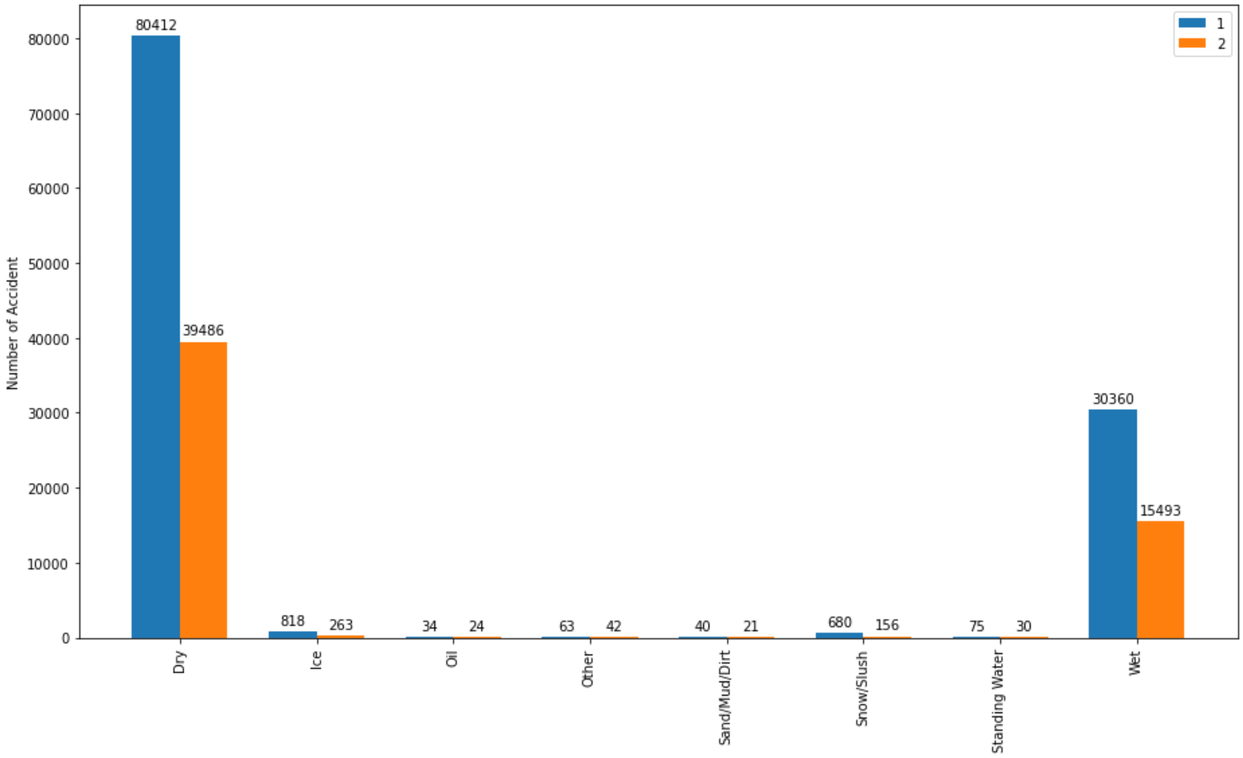


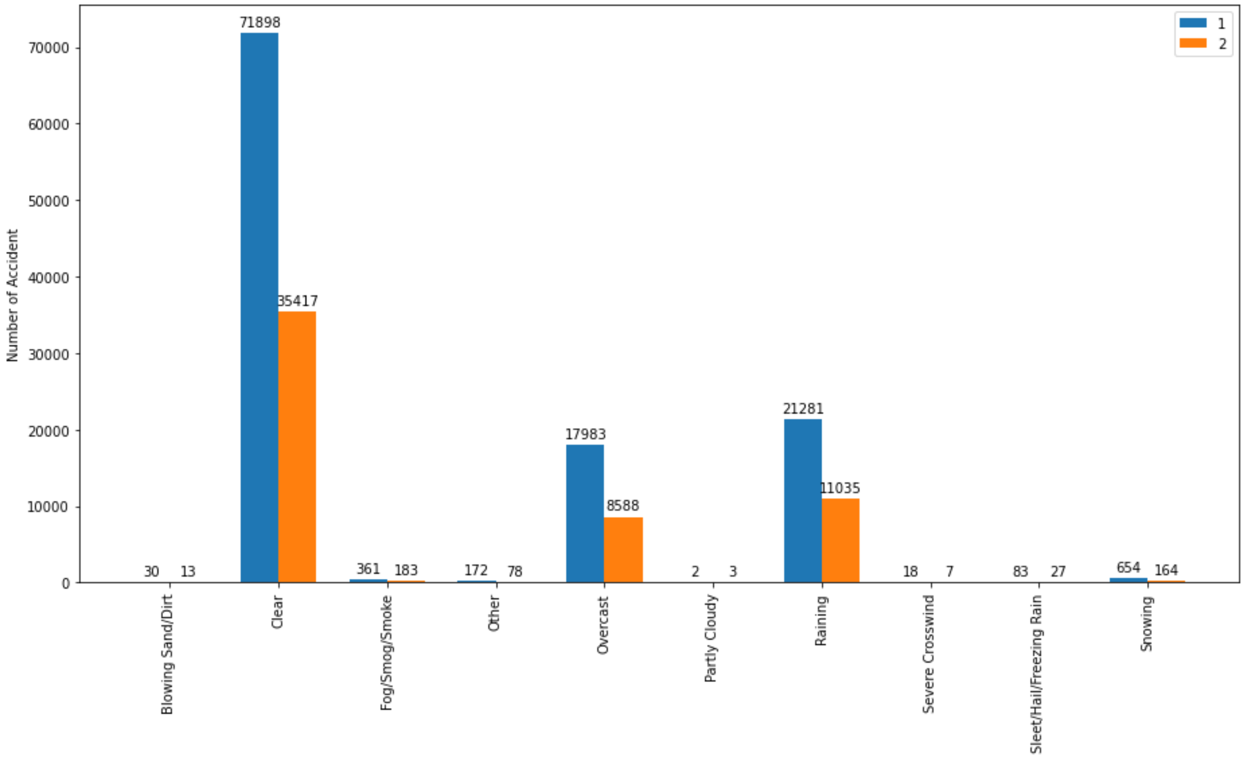
From figures above, we can see that our target data is unbalance, where we have half more severity code 1 than severity code 2. In addition, we also know that our initial thought about factors that influence the level of severity of an accident seems to be wrong. As we analyse, we found that, most of severity code 1 was happened in normal condition (Clear weather, dry road, bright light, good driver, normal speed). From this we can continue to dig deeper in order to increase our confidence about our newly findings.

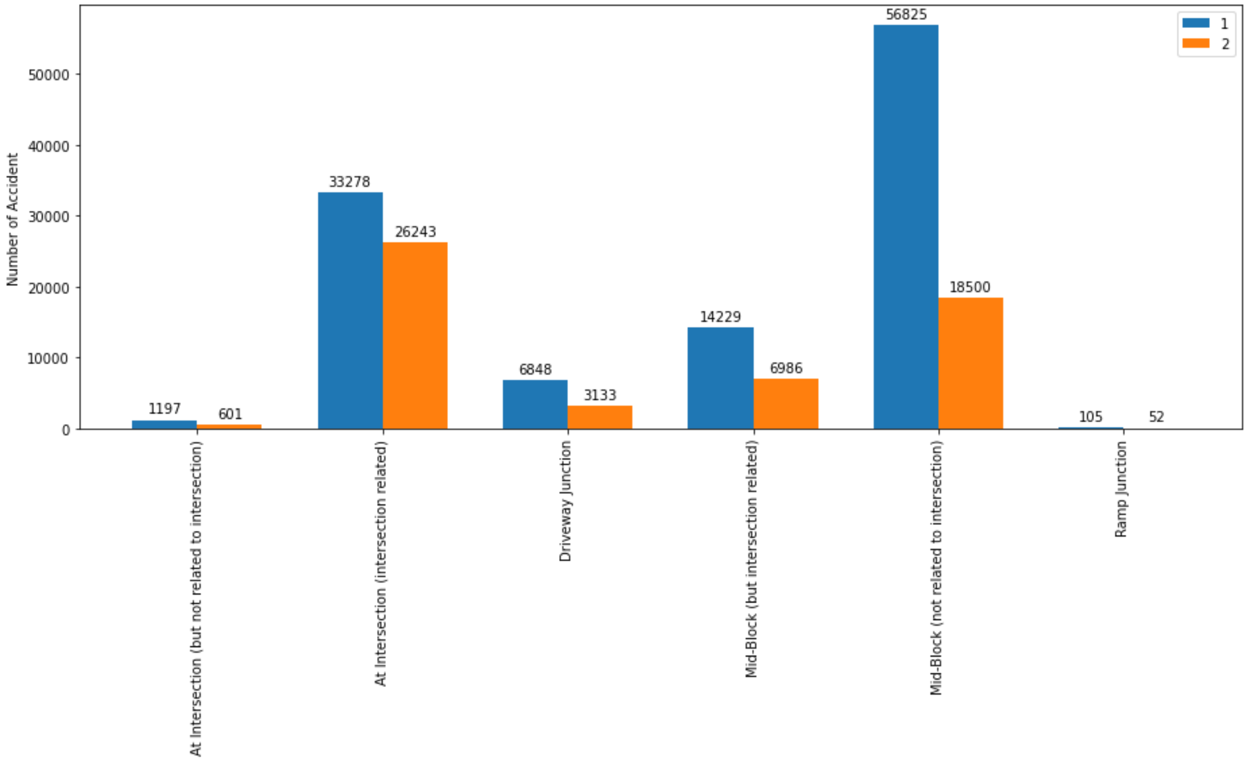
3.2 Bivariate Analysis

Bivariate analysis is mostly used to know the relationship between two variables. In our case, we are going to use this method to analyse the relationship between target variable and the features. This is what we found.



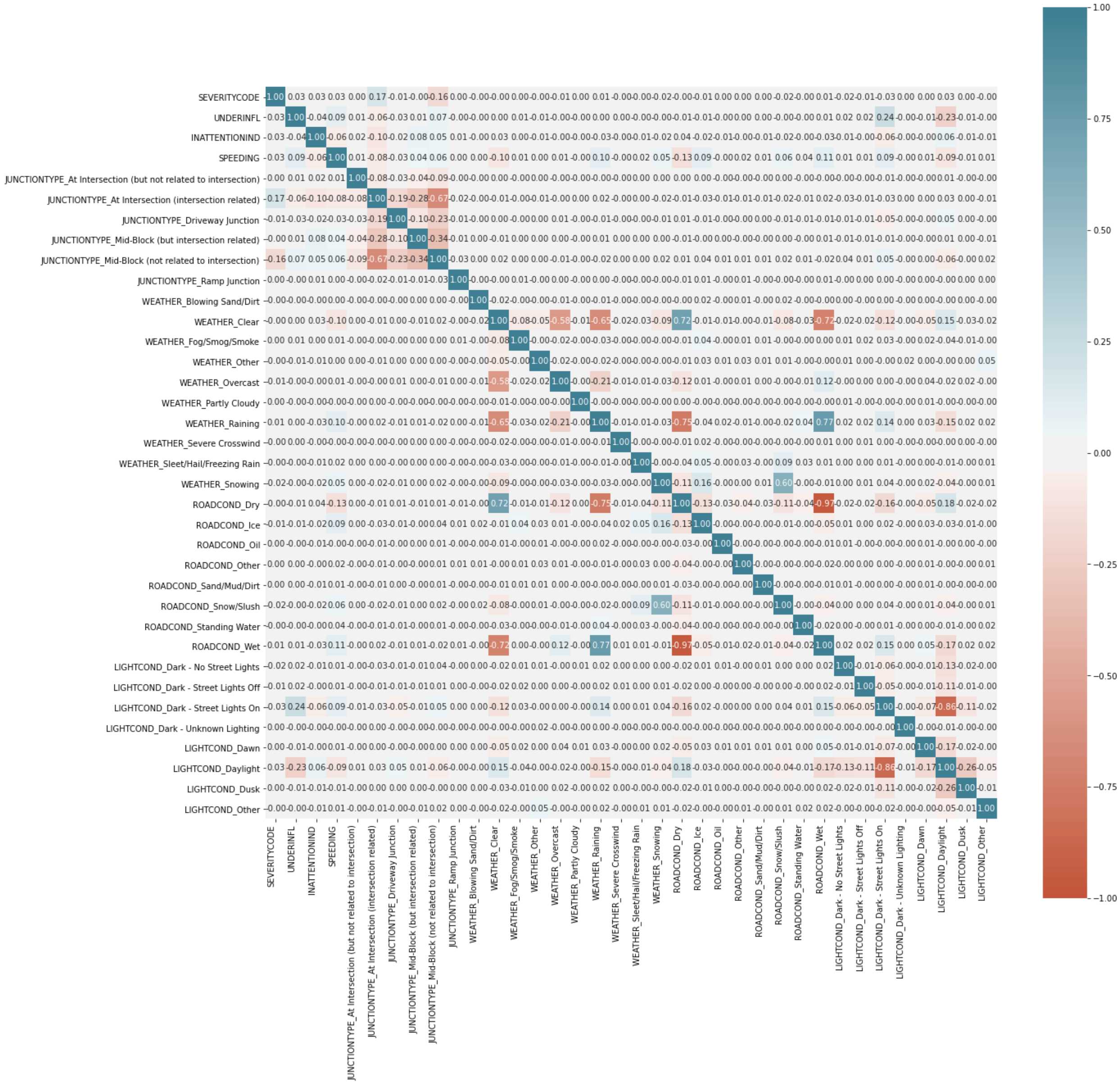

1
2
Number of Accident
0
10000
20000
30000
40000
50000
60000
70000
80000
80412
39486
818
263
34
24
63
42
40
21
680
156
75
30
30360
15493
Dry
Ice
Oil
Other
Sand/Mud/Dirt
Snow/Slush
Standing Water
Wet


1
2
Number of Accident
0
10000
20000
30000
40000
50000
60000
70000
30
13
71898
35417
361
183
172
78
17983
8588
2
3
21281
11035
18
7
83
27
654
164
Blowing Sand/Dirt
Clear
Fog/Smog/Smoke
Other
Overcast
Partly Cloudy
Raining
Severe Crosswind
Sleet/Hail/Freezing Rain
Snowing



From above result, we can clearly see that there is weak relationship between our target data and its features. This confirm our previous findings.

Another way to see the correlation between variables is by using pandas correlation method. But we also find weak relationship as we found before. This weak relationship will affect the accuracy of our model.



4. Predictive Modelling

4.1 Data Pre-processing

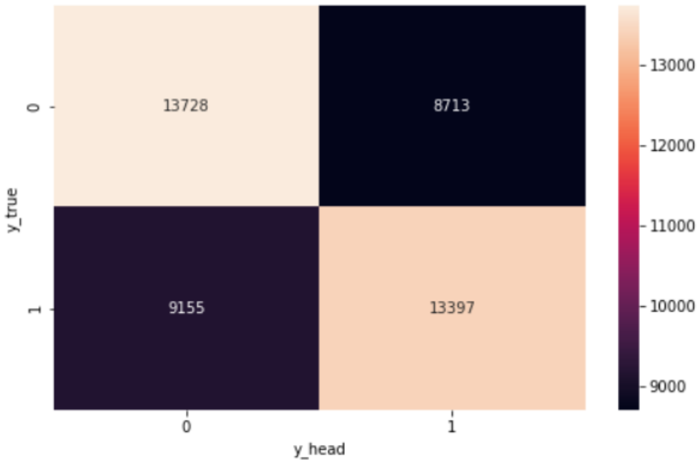
Machine learning model cannot process a data that has object or category format. So, data pre-processing exist for that problem. We use label encoder to change binary categorical data to 1 and 0 value. Other categorical data is transformed using dummy variable.

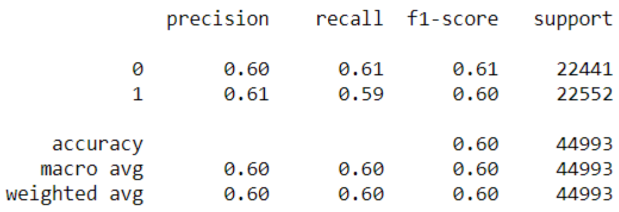
4.2 Model Building

Because our data mostly consist of categorical variable, we might consider to use tree-based model, such as random forest and decision tree to fit into our data. We also use logistic regression

4.3 Model Evaluation

We use confusion matrix as evaluation measurement of our model’s performance. Besides that, we also add accuracy score for our confidence. Here the figure for confusion matrix of random forest classifier.





```
              precision    recall  f1-score   support

           0       0.60      0.61      0.61     22441
           1       0.61      0.59      0.60     22552

    accuracy                           0.60     44993
   macro avg       0.60      0.60      0.60     44993
weighted avg       0.60      0.60      0.60     44993
```

4.4 Model Optimization

As it name, model optimization is used to maximize our model performance by adding some useful parameter. In this project, we use GridSearchCV as optimization method with logistic regression classifier. But unfortunately, we cannot increase the model accuracy as we hope.

5. Conclusion and Recommendation

After doing analysis of our dataset and train our model with that dataset, we conclude that those variable in our dataset cannot be used to train the model because it do not have enough correlation that we expect. Therefore, we recommend to try to apply another combination of variables.