Capstone Project - The Severity of Accident

1. Introduction

1.1 Problem Understanding

In this project, we focus on the subject of predicting the severity of an accident. people travel and transport by driving a lot in America, Sometimes, it is unfortunate but inevitable to be involved in a traffic accident. The purpose of this project is to make predictions of the severity of accidents based on some attributes like weather and road conditions. This prediction can offer a reference for drivers about the possibility of getting into a car accident and how severe it would be so that they would drive more carefully or even change the travel.

1.2 Target Audience

This project will be most useful for:

- People who travel by driving or people who are planning to drive.
- Works and administrators in the road transport industry, for example, the trucker driver.
- Police officers who want to reduce the accident rate and severity.

2. Data

2.1 Overview

The Collisions dataset includes records of collisions that happened on road from 2004 to Present. The dataset contains 194673 rows and 38 columns, each row is a record of the accident, and each column is an attribute. The first column "SEVERITYCODE" is the labeled data, which describes the fatality of an accident. The remaining 37 columns have different types of attributes. Some or all can be used to train the model.

There are some problems that need to be fixed in this dataset

- There are missing values.
- The data type of some attributes is not correct.
- Some attributes are not useful in building a machine learning model.
- The data has unbalanced labels. There are 136485 obs of label1 but only 58188 obs of label2.

Most of the observations are good to train and test the machine learning model, but a data preprocessing and data selection procedure needs to be conducted. A more detailed description of each attribute can be found on the website.

2.2 Data Preprocessing

In order to fix the problems mentioned above we need to conduct a data preprocessing procedure which includes following steps.

- Feature Selection

  Some features are meaningless, like `OBJECTID`, `COLDETKEY` and `STATUS`. `SEVERITYCODE.1` is duplicated with `SEVERITYCODE` The `INCDTTM` of many obs are missing or not completed we only select these features: `'SEVERITYCODE', 'X', 'Y', 'ADDRTYPE', 'SEVERITYDESC', 'COLLISIONTYPE','PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INCDATE', 'JUNCTIONTYPE', 'SDOT_COLCODE', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'ST_COLCODE', 'HITPARKEDCAR'`.

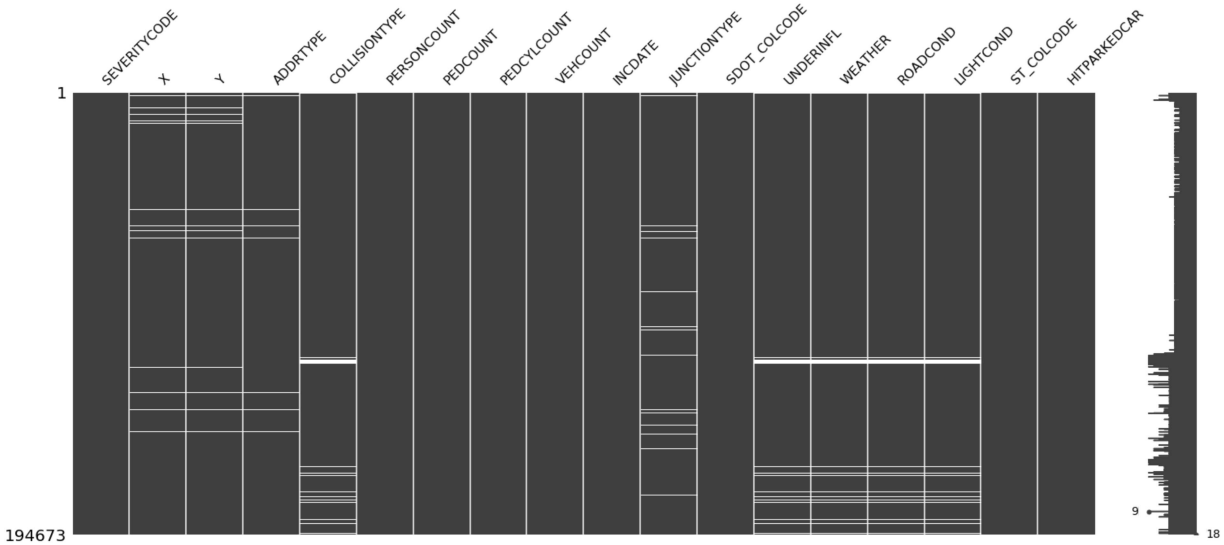
  After the selection, there are 19 attributes in total.

- Consistancy

  There are "Y"/"N" and "1"/"0" in `UNDERINFL`, we need to convert all "Y"/"N" to "1"/"0"

- Missing Values

  There are missing values, from the matrix plot of missing values, it seems that `UNDERINFL`, `WEATHER`, `ROADCOND`, `LIGHTCOND`, and `COLLISIONTYPE` usually miss at the same time, the location information `X`, `Y` miss at the same time, the other values miss randomly. We decide to drop all rows with missing values. After dropping all the missing values, we still have 180067 obs left.



- Correct Data Format

Convert `SDOT_COLCODE` from int to object, `convert` INCDATE` to datetime

Creat new variable `year` , `month` and `weekday` from `INCDATE`

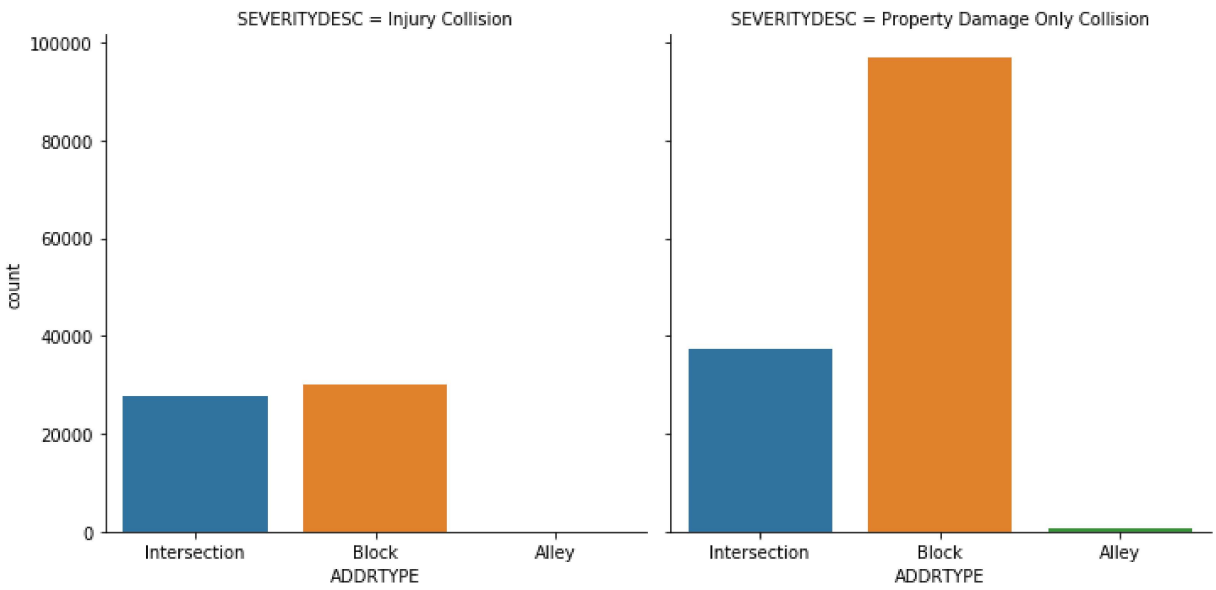
After cleaning the data, there are 22 columns and 180067 obs, These data will be used for exploratory data analysis and modeling.

3. Methodology

3.1 Exploratory Data Analysis

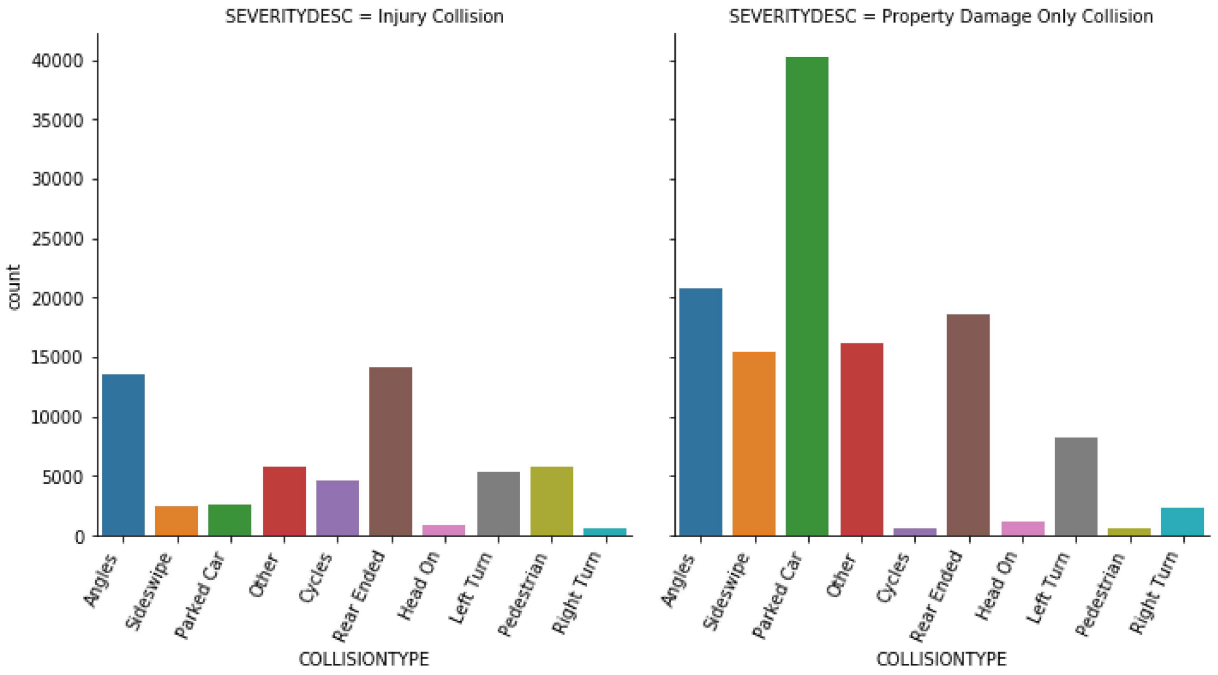
From the summary of the cleaned data, most of the features are categorical data, we will use barplot to see how type1 and type2 accidents were distributed in different conditions..

- **Severity of Accident VS Collision Address Type**



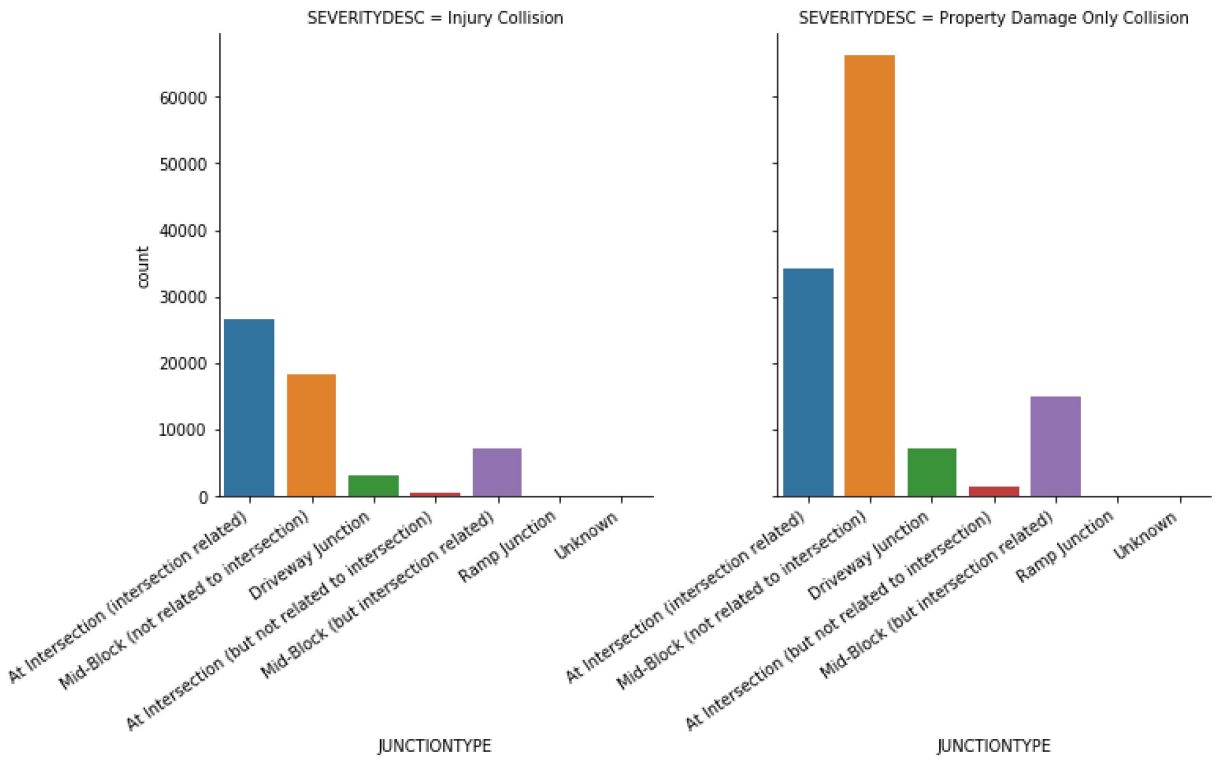
From the plot, we can see that most damage only collision happened in the block while only fewer than half happened in the intersection. For injury collision, the accident happened in intersection and Block is almost the same. This means intersection is more dangerous for people because when you encounter a collision in the intersection, there's a higher probability that you will be injured. However, there is more damage only collisions happen in blocks. Collisions that happened on the alley are very rare.

- **Severity of Accident VS Collision Type**



From the plot, we can see that the most common collision type is parked car, but most of them are the property only collisions. Other types like Angles and Rear-Ended also occur frequently， these two types of collisions also more likely to cause injuries. Cycle and pedestrian collisions not happen too much, but they are very dangerous to people because most of these accidents result in injuries.

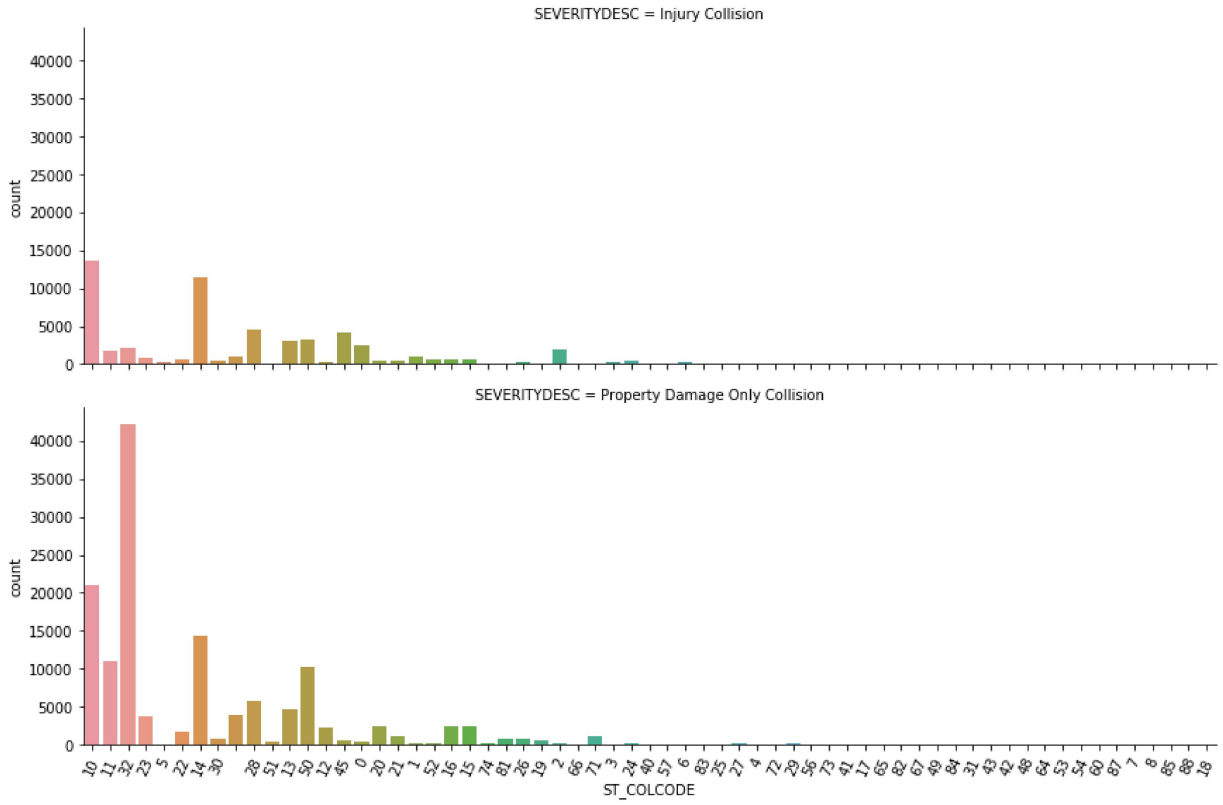
- **Severity of Accident VS Junction Type**



Most collisions happen in mid-blocks or interactions the result is similar to what we had known

from Severity of Accident VS Collision Address Type, collisions happen at interactions are more likely to cause injuries.

- **Severity of Accident VS ST_COLCODE**



From the plot, the most common collision type are:

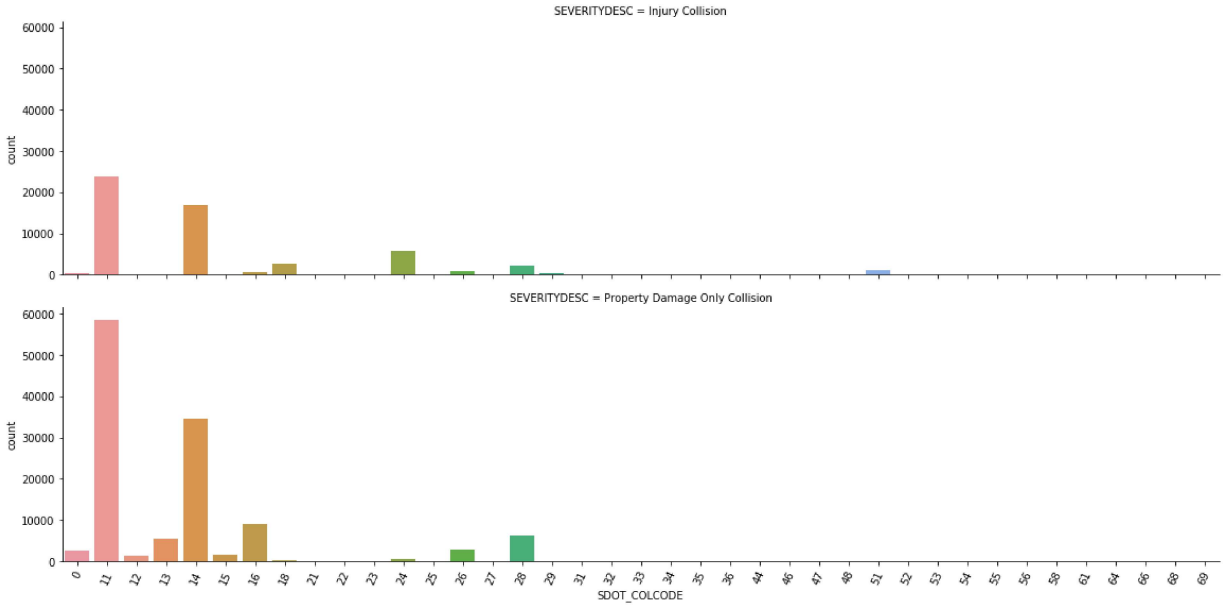
10-Entering at an angle,

32-One parked--one moving,

14-From same direction - both going straight - one stopped - rear-end

10 and 14 are very dangerous for humans.

- **Severity of Accident VS SDOT_COLCODE**



From the plot, the most common collisions type are:

11-MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END AT ANGLE,

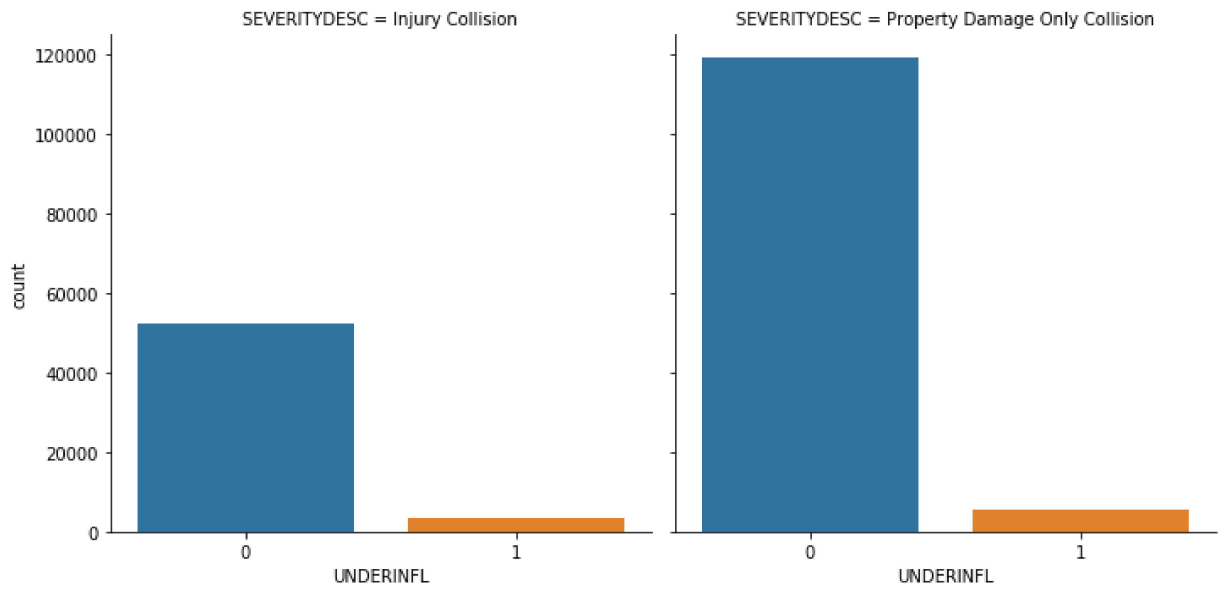
14-MOTOR VEHICLE STRUCK MOTOR VEHICLE, REAR END,

16-MOTOR VEHICLE STRUCK MOTOR VEHICLE, LEFT SIDE SIDESWIPE,

24-MOTOR VEHICLE STRUCK PEDESTRIAN

Where type 11 and 14 collisions happen most frequently, while 24 not happen too much but always cause injuries.

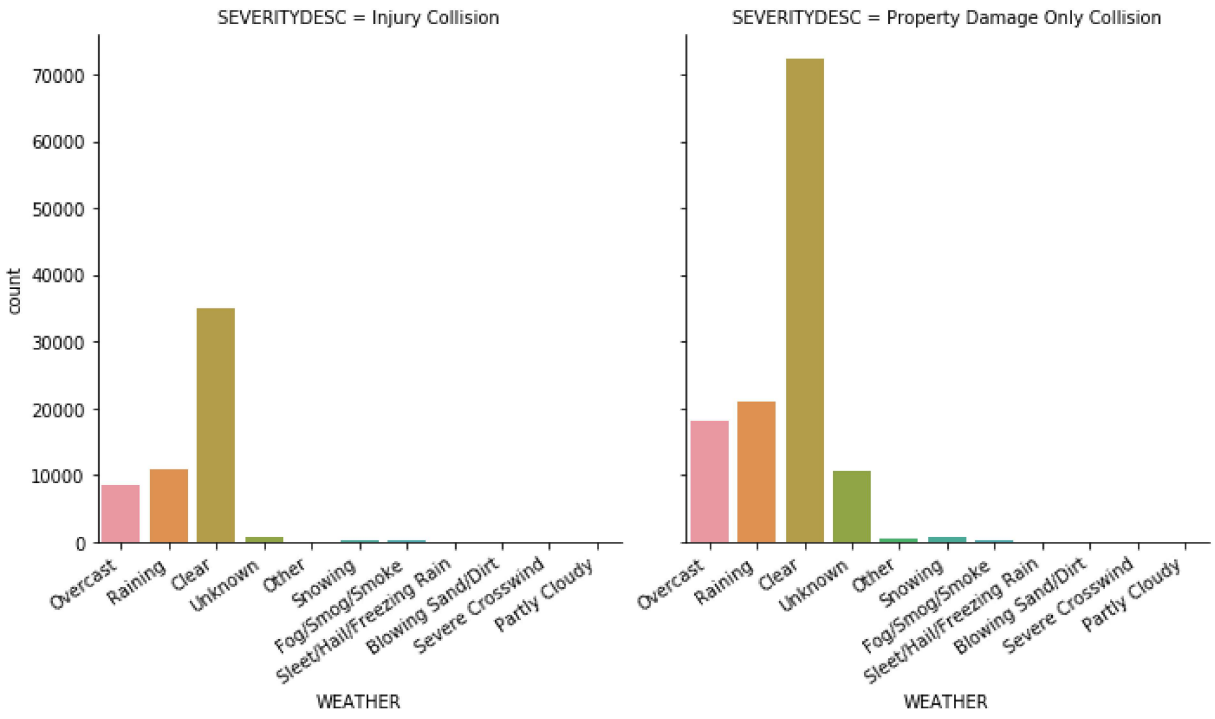
- **Severity of Accident VS Drugs or Alcohol**



Most people do not use drugs or alcohol, but in injury collision, a higher percentage of people use

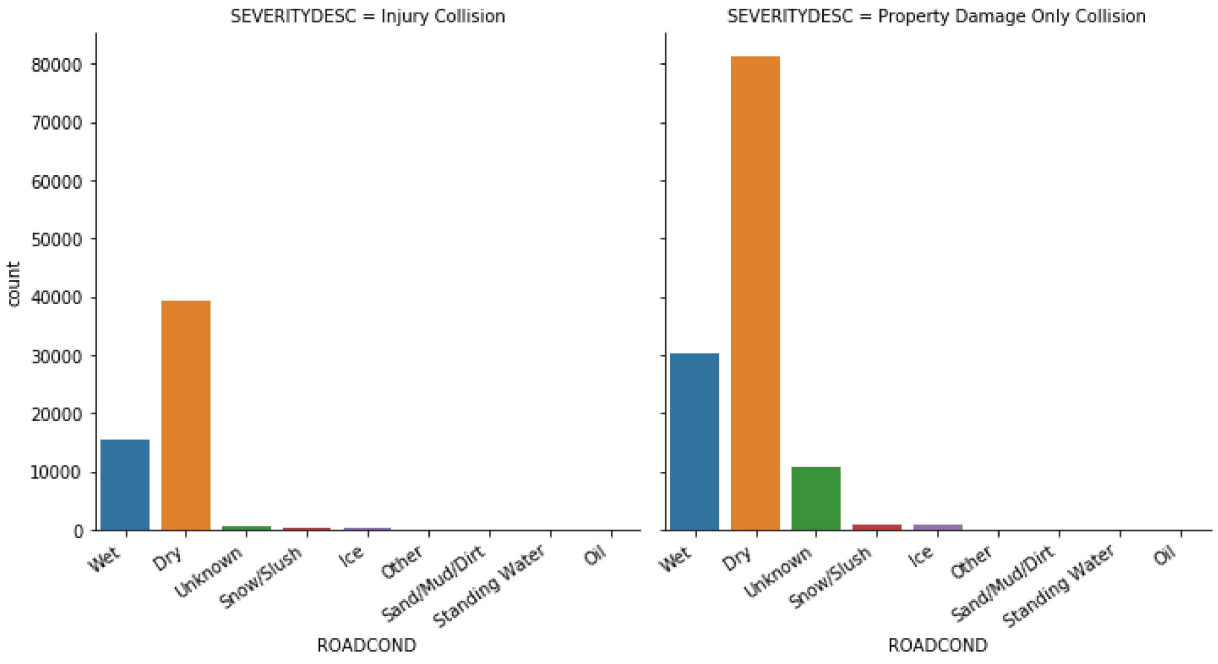
drugs or Alcohol.

- **Severity of Accident VS Weather**



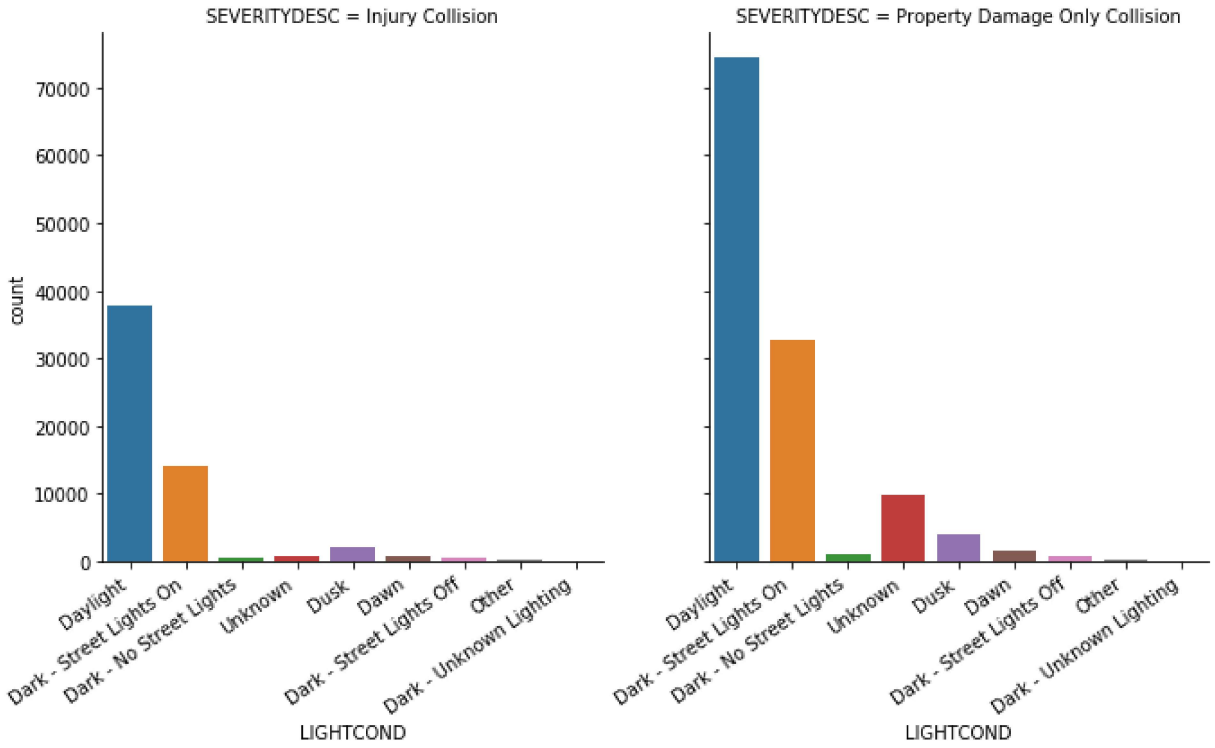
Most collisions happened in the clear weather, which indicates that bad weather do notleads to more accidents.

- **Severity of Accident VS Road Condition**



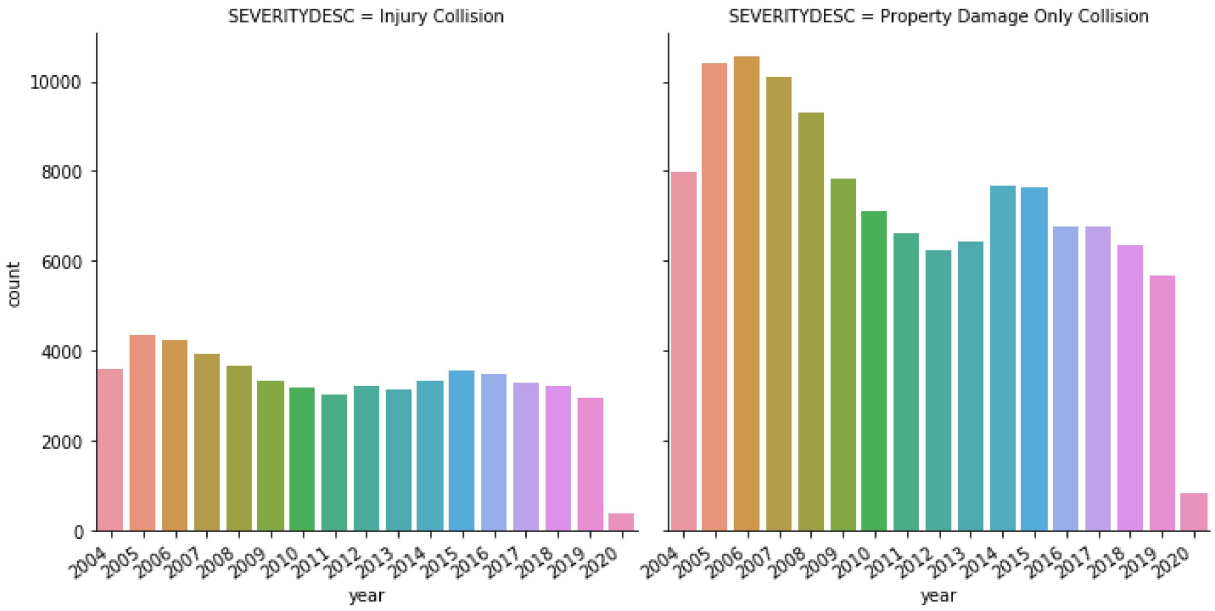
Most collisions happened in Dry Road, which indicates that bad road conditions do not lead to more accidents.

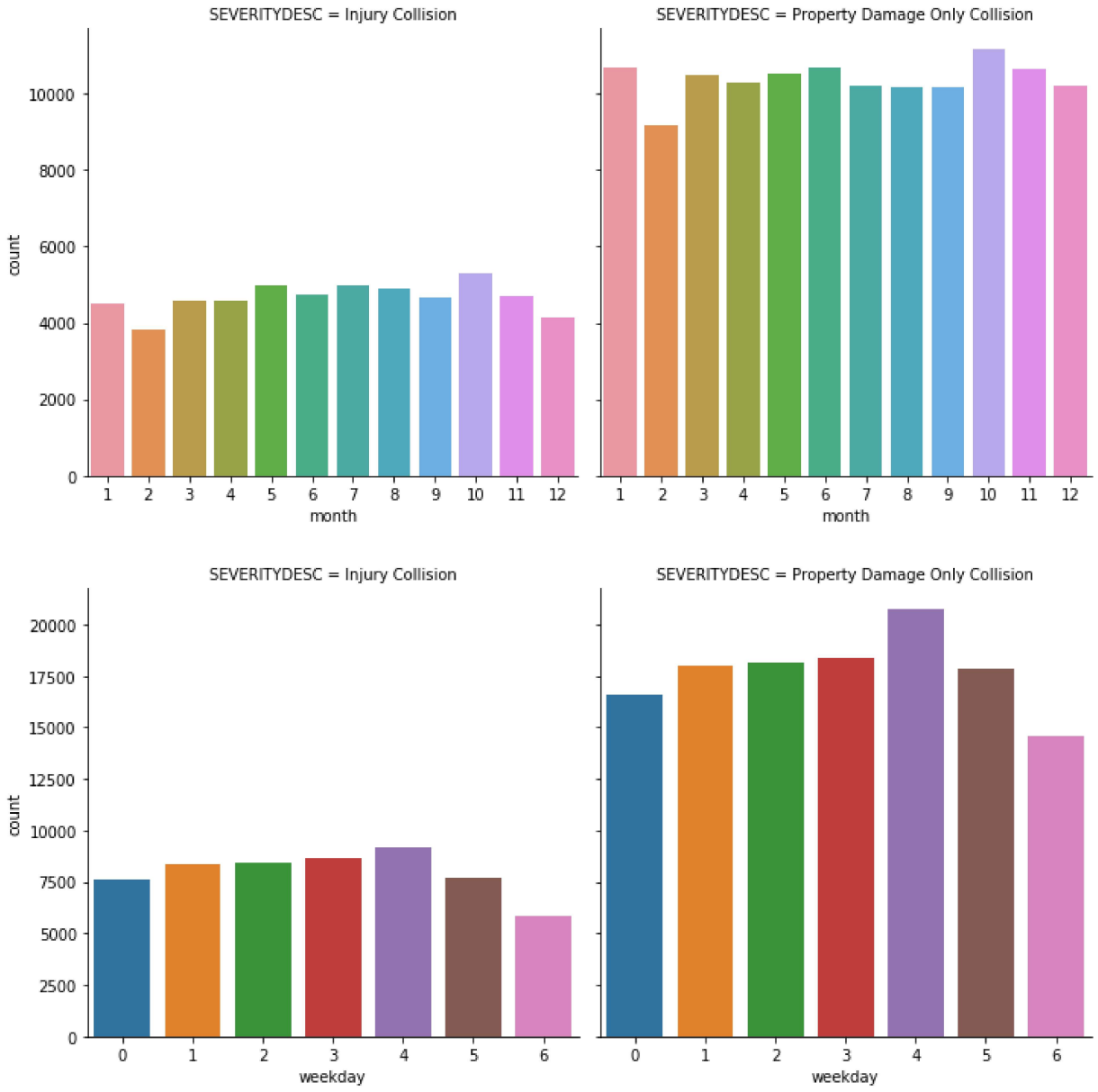
- **Severity of Accident VS Light Condition**



Most collisions happened in the condition when daylight or street lights on, which indicates that light conditions do not lead to more accidents.

- **Severity of Accident VS Time**





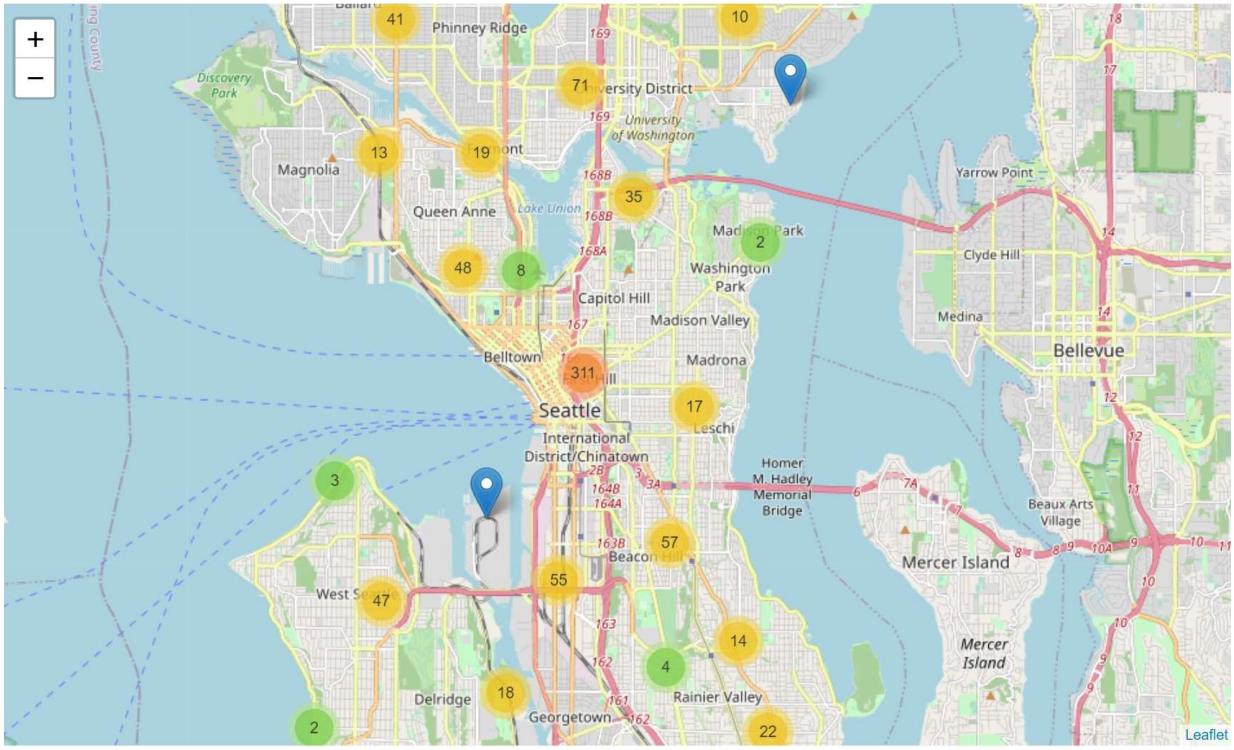
The number of accidents has fallen science 2005-2009, increase again from 2013, then drop again since 2016.

The number of cases among different months is almost the same, but the number that happened in January and October is slightly higher.

Thursday has the most accidents in the whole week.

3.2 Spatial Analysis

With the location data, we can plot accidents on the map, since there are too may data, we only select the data of 2020.



Most of the collisions are concentrated The accident was concentrated on the downtown of Seattle, the zoomable map can be found in the code file.

The accident rate in southern Seattle is slightly lower than the northern Seattle.

Most accidents located close to the state highway.

3.3 Machine Learning Modeling

3.3.1 Data Prepareing

- **Drop unnecessary attributes for modeling**

In the data preprocessing section, we kept some attributes for explatory data analysis they are not useful for modeling, so we drop attributes: `SEVERITYDESC` , `INCDATE` , `year` , `month` , `weekday` .

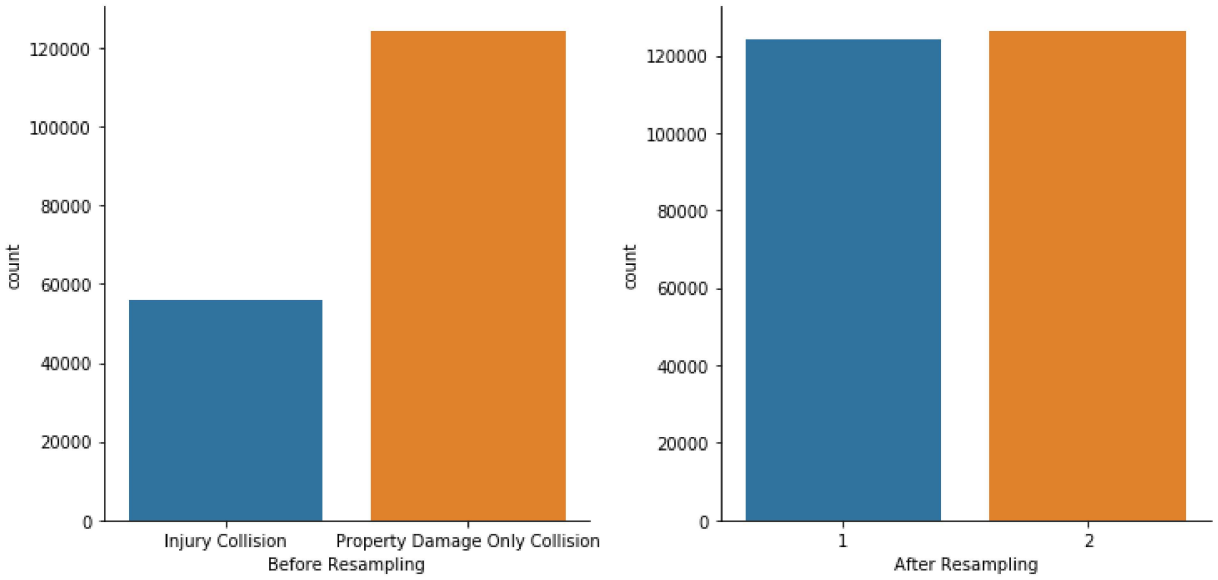
- **Creat Dummy variables**

For building the decision tree model we need create dummy variables for categorical attributes, convert them to the format of 0/1.

- **Solve problem of unbalanced labels**

Imbalanced class is a common problem in machine learning classification where there is a disproportionate ratio of observations in each class. Class imbalance can be found in many different areas including medical diagnosis, spam filtering, and fraud detection. In our dataset the labeled response value is imbalanced, there are 136485 obs of label1 and only 58188 obs of label2, we need to oversample the label2 data and add more copies of the minority class.

After the oversampling, the label is balanced and we can move on to the modeling



- **Split data into training and testing dataset**

In order to evaluate the performance of the different models, we split the data into training data and test data, where training data takes 70% and testing data takes 30%.

3.3.2 Decision Tree model

I applied Decision Tree models to train the dataset in the first place. A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from the root to leaf represent classification rules.

Decision Tree is a powerful tool for solving classification problems. In this project, I build a decision tree and set the max depth as 10. Then use the trained model to make predictions for the testing data, compare the true value and predicted value. then use Accuracy, F1-Score, and AUC to evaluate the performance of the model.

3.3.3 Logistic Regression model

Logistic regression is a statistical model that uses a logistic function to model a binary dependent variable, it is also a very powerful tool for classification problem.

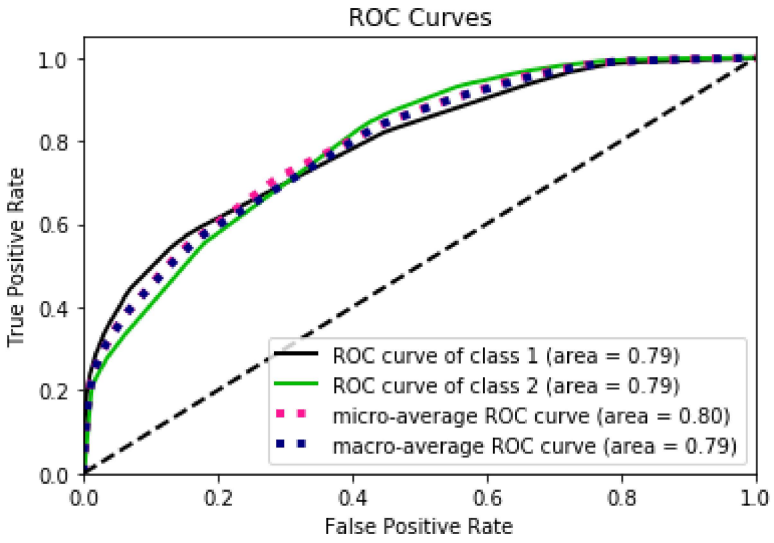
I use the training data to fit the logistic regression model, then make predictions based on the trained model, compare the true value and predicted value. then use Accuracy, F1-Score and AUC to evaluate the performance of the model.

4. Results

4.1 The Performance of Decision Tree model

DecisionTree's Accuracy: 0.7103074814726114

DecisionTree's F1-Score: 0.7051338029889037

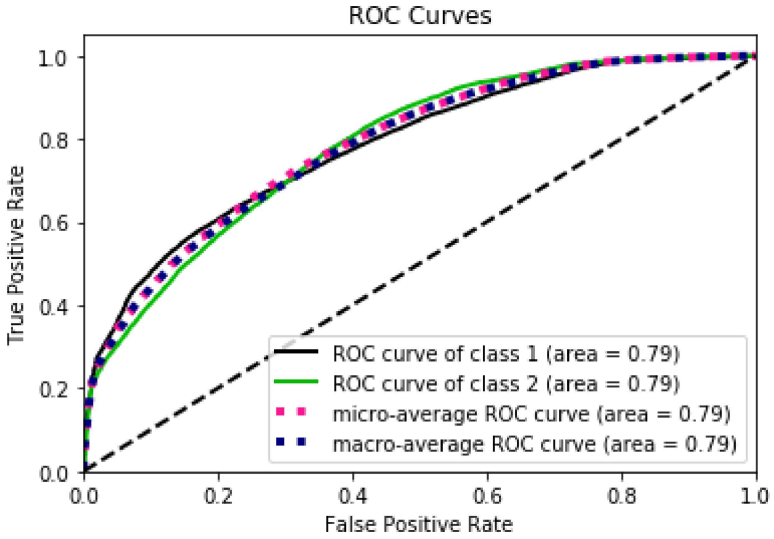


4.2 The Performance of Logistic Regression model

Logistic Regression's Accuracy: 0.7038012746311154

Logistic Regression's F1-Score: 0.7004139410689796

Logistic Regression's LogLoss : 0.5445183257641134



The result shows that these two methods have similar performance in predicting the severity of accidents, the Accuracy and F1 score are about 0.7, AUC is about 0.8, which is fairly high for a machine learning model.

5. Discussion

From the result of exploratory data analysis and machine learning modeling, we have some interesting findings.

- More accidents happened in blocks but accidents that happened in intersections result in injuries.
- Features like road condition, light condition, weather do not influence the accident rate as we expected, most accidents happened in good weather and clear road.
- In a whole week, more accidents happened on Thursday, maybe because after 3 days of hard work, drivers got tired. Less accident happened on Saturday because more people take rest at home and do not drive to work.
- More accidents happened downtown, that's reasonable because there are more people and more cars, also accidents located along state highways, because lots of people drive along these roads or live close these roads.
- With the data we have, we can predict the severity of an accident with about 70% accuracy, this can be applied in the real world, for example, given the features in our dataset, we can make predictions about the severity of the accident.

6. Conclusion

In this project, I analyzed the relationship between the severity of accidents that happened in Seattle from 2004 to present and their features like location, collision type, weather, light, etc. I also build two different models for predicting the severity of accidents, one is the decision tree model and the other is the logistic regression model, they have similar performance inaccuracy. These models can be very useful in helping drivers and police to avoid potential accidents. For example,

if the possibility of an injury accident is high, the driver should be vigilant when driving, the government can make warning signs in high accident area, it also gives a reference about how should we optimize the road planning and traffic system to avoid accidents, for example, change some road to one direction, build sidewalks, etc. In summary, this is a meaningful study for protecting the safety of our lives and properties, hope everyone drive safely on the road.