Applied Data Science Capstone Project

Car Accident Severity Analysis: Seattle, Washington

Miller Palaniswamy |IBM-Coursera| 15-Sep-20

Table of Contents

1.Introduction/Business Problem

1.1 Background

For an individual & the family dependent on him, his life & health is more important for survival. There are numerous car accidents happen across the world every day. All car accidents will create damage to cars involved or even worse, take lives. In fact, there are many factors that contribute to the severity of a car accident & human safety.

At the same time growth of transportation plays an indispensable role in our society and their ever-increasing dependency to enables communication, trade and exchange of goods and services across the globe, is unstoppable. But it has very unfortunate impact on the society in terms of accidents.

Therefore, it is advantageous for related departments to accurately predict the severity of car accidents under those conditions. For example, does bad road conditions involves in large number of car accidents? If it does, the prediction provides solid reason for better road constructions. From that, warning information like road signs to the drivers will lead to positive impacts. Since transportation industry is ripe for advancement, data science can bring about an evolution in this sector.

Data that might contribute to a car accident including locations, weathers, road conditions, light conditions, vehicles or pedestrians involved, speeding, whether the driver is involved was under the influence of drugs or alcohol, etc. This project aims to predict the car accident severity based on these data.

1.2 Case Study – Seattle traffic data

Washington State's largest city, Seattle, is a home for large tech giants such as Microsoft and Amazon headquartered in its metropolitan area. As of 2020, it has a total metro area population of 3.4 million (https://www.macrotrends.net)The total number of personal vehicles in Seattle in the year 2016 hit a new high of nearly 444,000 vehicles. In one South Lake Union census tract, the car population has more than doubled since 2010(http://www.seattletimes.com). The increase in car ownership rates can lead to higher numbers of accidents on the road because of a simple probability. Worldwide, approximately 1.35 million people die in road crashes each year, on average 3,700 people lose their lives every day on the roads and an additional 20-50 million suffer non-fatal injuries, often resulting in long-term disabilities. We will be trying to find possible insights from this data about the reasons as to why accidents occur, which area is more prone to accidents and what are the aftereffects of accidents on traffic flow.

The world suffers due to car accidents, including the USA. National Highway Traffic Safety Administration of the USA suggests that the economical and societal harm from car accidents can cost up to $871 billion in a single year. According to 2017 WSDOT data, a car accident occurs every 4 minutes and a person dies due to a car crash every 20 hours in the state of Washington while Fatal crashes went from 508 in 2016 to 525 in 2017, resulting in the death of 555 people. The project aims to predict how severity of accidents can be reduced based on a few factors.

1.3 Beneficiaries & Stakeholders

Data analytics can provide an in-depth knowledge of methods for analyzing and implementing intelligent transportation systems. For example: connected vehicles can be a way forward in the transportation industry. Sensors placed around cities that are then connected to apps can help drivers find parking spots faster, reducing traffic and emissions. The benefits of big data and analytics helps transportation firms to precisely enhance the model capacity, demand, revenue, pricing, customer sentiments, cost. Some of the assets include implementing real time monitoring system for enhancing operational efficiency, public transit system and traffic management system to improve bus transportation and reduce traffic congestion. By developing this project, we can help society with implementing some laws needed for transportation which can prevent the accidents and helps them to know at where they must be safer on roads which are more prone to accidents.

2.Data description, acquisition & Preprocessing

2.1 Data Sources

The dataset of all the collisions shared by SPD – Traffic department can be downloaded from https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv and the Metadata could be downloaded from https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf .

This dataset is a supervised with labeled severity of car collisions with numerous attributes like locations, weathers, road conditions, light conditions, etc. This dataset, however, needs to clean since there are empty inputs in some attributes like road conditions.

2.2 Data Cleansing

There are 194,673 observations and 38 variables in this data set. Since we would like to identify the factors that cause the accident and the level of severity, we will use SEVERITYCODE as our dependent variable Y, and try different combinations of independent variables X to get the result. Since the observations are quite large, we may need to filter out the missing value and delete the unrelated columns first. Then we can select the factor which may have more impact on the accidents, such as address type, weather, road condition, and light condition.

The target Data to be predicted under (SEVERITYCODE 1-prop damage 2-injury) label.

Other important variables include: (Not limited to)

- ADDRTYPE: Collision address type: Alley, Block, Intersection
- LOCATION: Description of the general location of the collision
- PERSONCOUNT: The total number of people involved in the collision helps identify severity level
- PEDCOUNT: The number of pedestrians involved in the collision helps identify severity level
- PEDCYLCOUNT: The number of bicycles involved in the collision helps identify severity level

• VEHCOUNT: The number of vehicles involved in the collision identify severity level

• JUNCTIONTYPE: Category of junction at which collision took place helps identify where most collisions occur

• WEATHER: A description of the weather conditions during the time of the collision

• ROADCOND: The condition of the road during the collision

• LIGHTCOND: The light conditions during the collision

• SPEEDING: Whether speeding was a factor in the collision (Y/N)

• SEGLANEKEY: A key for the lane segment in which the collision occurred

• CROSSWALKKEY: A key for the crosswalk at which the collision occurred

• HITPARKEDCAR: Whether the collision involved hitting a parked car

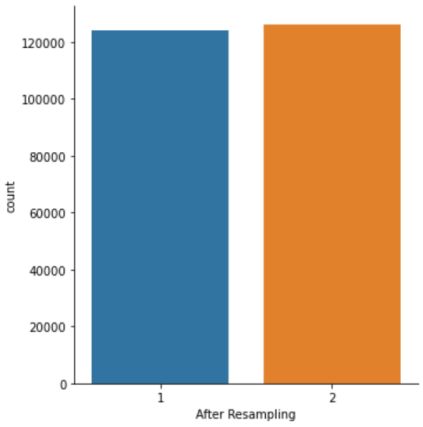
There are two major problems with the dataset. First, the labels are unbalanced. After getting the details of the dataset, I found this dataset only provides two different labels for 'SEVERITYCODE'. They are '1' for 'prop damage' and '2' for 'injury'. However, the number of these accidents for each has huge difference. The originally dataset has double size of label 1 accidents than that of label 2 accidents. Therefore, it's necessary to shuffle and resample to create a balanced dataset before training. Otherwise, a biased predication will be generated.

The second problem I need to deal with is that for 'ROADCOND', 'WEATHER', 'LIGHTCOND' feature sets use categorical variables. Hence the conversion from categorical to numeric is necessary before employing machine learning methods. Beyond that, some inputs in these categorical columns are empty, which should be dropped from the dataset in order to provide better results.

Therefore, I first dropped all the rows with empty inputs in those categorical columns, converted the categorical variables into numeric variables and Finally, shuffled the dataset and resample randomly to create a dataset with equal size rows for both label 1 and label 2 to ensure an unbiased predication.

```
ax2 = sns.catplot(x = "SEVERITYCODE", data = df6, kind = "count")
ax2.set(xlabel='After Resampling')
```

```
<seaborn.axisgrid.FacetGrid at 0x7fd17722a198>
```



Even this sample values can still be pruned and reduced by another 5000+, if we consider removing columns with more than 20% values missing such as (INATTENTIONIND, PEDROWNOTGRNT, SPEEDING) & rows for columns with less than 20% values missing (COLLISIONTYPE,JUNCTIONTYPE,UNDERINFL,WEATHER,ROADCOND,LIGHTCOND etc.

2.3 Data Preprocessing

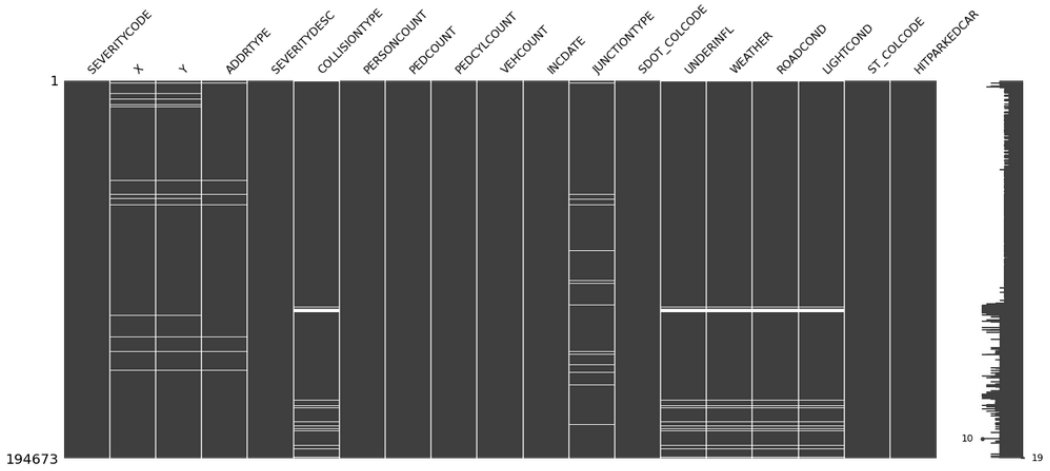
In order to fix the problems mentioned above we need to conduct a data preprocessing procedure which includes following steps.

- Feature Selection Some features are meaningless, like OBJECTID, COLDETKEY and STATUS. SEVERITYCODE.1 is duplicated with SEVERITYCODE The INCDTTM of many objects 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.
- Consistency
  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 objects left.

```
msno.matrix(df3)
```

```
<AxesSubplot:>
```



- Correct Data Format
  Convert SDOT_COLCODE from int to object, convertINCDATE` to datetime
  Create new variable year, month and weekday from INCDATE

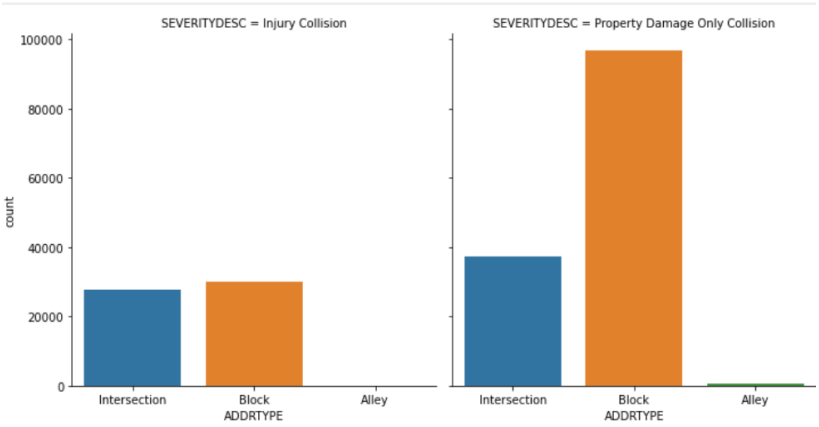
After cleaning the data, there are 22 columns and 180067 objects, these data will be used for exploratory data analysis and modeling.

3.Methodology

3.1 Exploratory Data Analysis

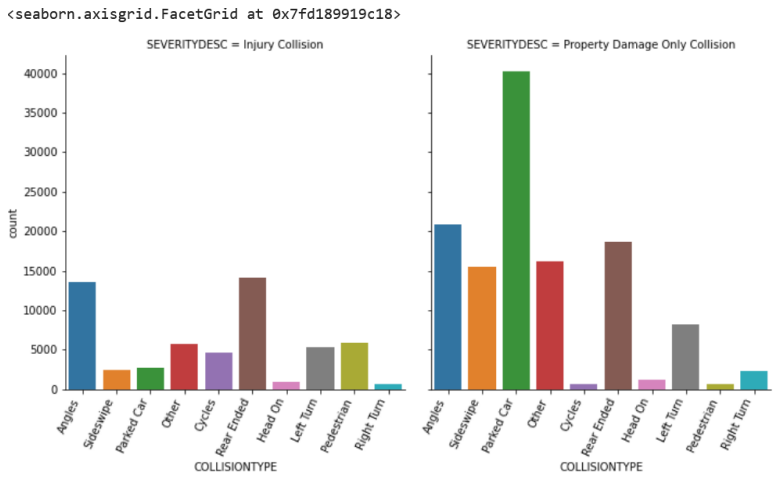
From the summary of the cleaned data, most of the features are categorical data, we will use bar plot to see how type1 and type2 accidents were distributed in different conditions.

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



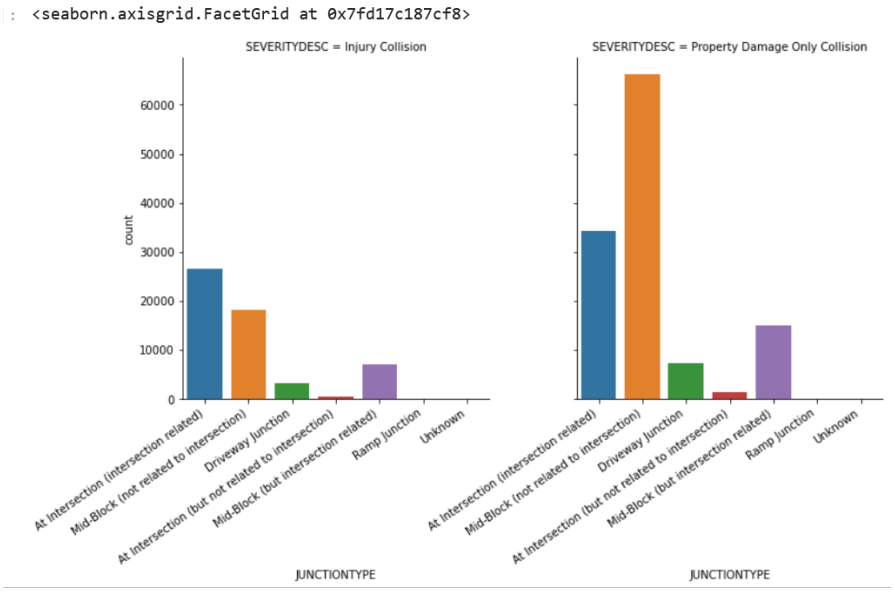
We can observe from the plot that the damage only accidents are almost double in block, whereas the injury collisions are similar in both intersections and blocks. Accidents at alley are comparatively minimal. Hence intersections are more dangerous for people, because the probability of more injuries is higher at intersections.

- **Severity of Accident VS Collision Type**



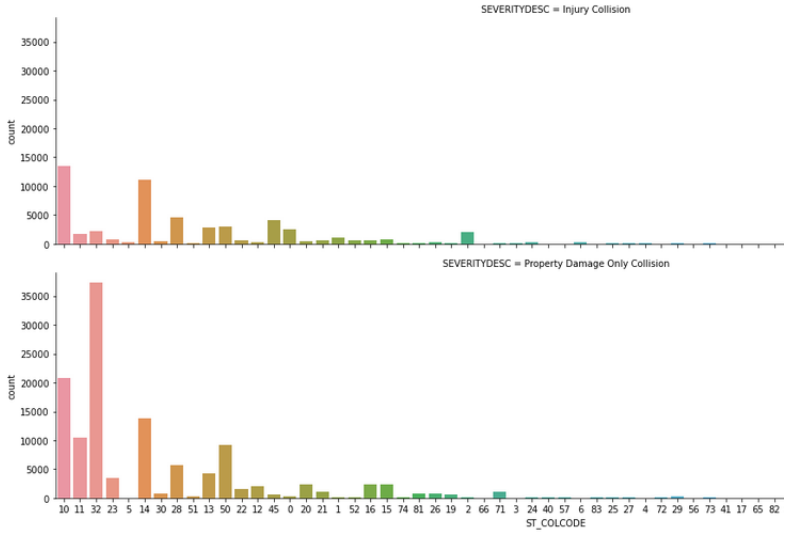
Referring to the chart we can see that Angles and Rear-ended collisions are happening frequently which is critical for people. Cycle and pedestrian collisions are not much but again they are dangerous for people. Though collisions related to parked car is high they are property damages only.

- **Severity of Accident VS Junction Type**



From the chart we can observe that Injury collisions are happening at mid-blocks and intersections which is again like collision address type which is dangerous to people.

- **Severity of Accident VS ST_COLCODE**

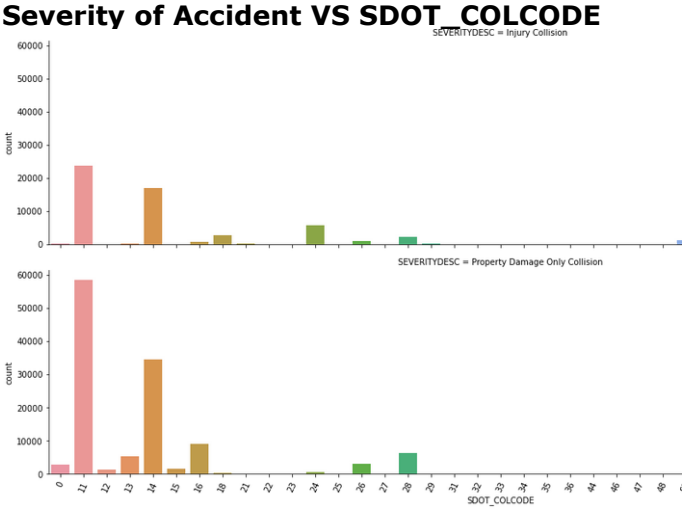


As per the chart, the most common collision types are:

Entering at an angle (10), One parked--one moving (31), From same direction - both going straight - one stopped - rear-end (14).

10 and 14 are very dangerous for humans.

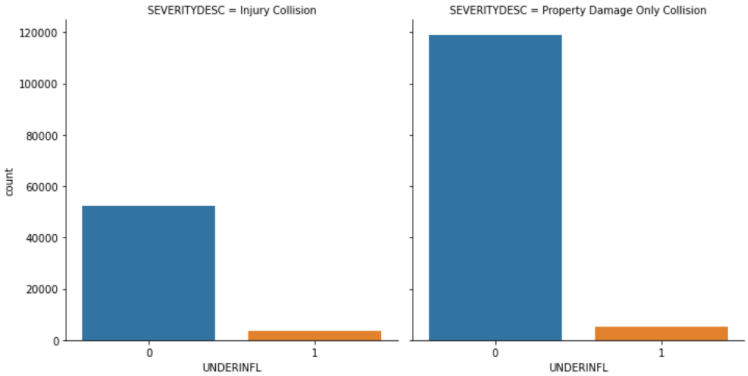
- **Severity of Accident VS SDOT_COLCODE**



From the plot, the most common collisions type are, motor vehicle struck motor vehicle- front end at angle(11), motor vehicle struck motor vehicle-rear end(14),motor vehicle struck motor vehicle-left side sideswipe(16), motor vehicle struck pedestrian(24).

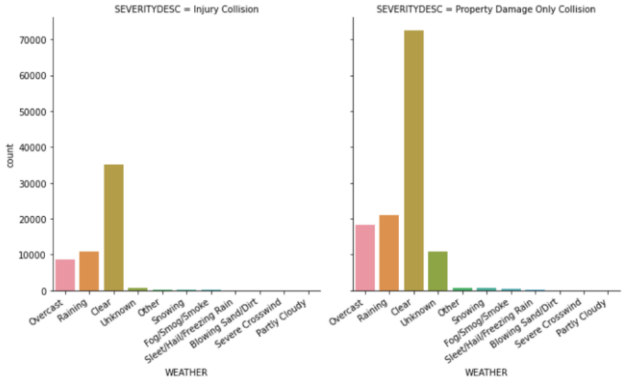
Where type 11 and 14 collisions happen most frequently, while 24 not happen too much but always cause injuries.

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



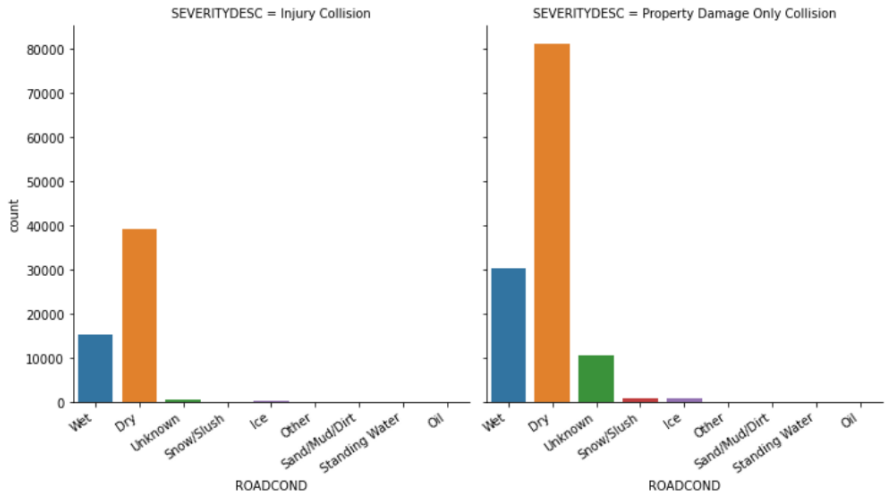
We can observe that in injury collision the drug influence is higher.

- **Severity of Accident VS Weather**



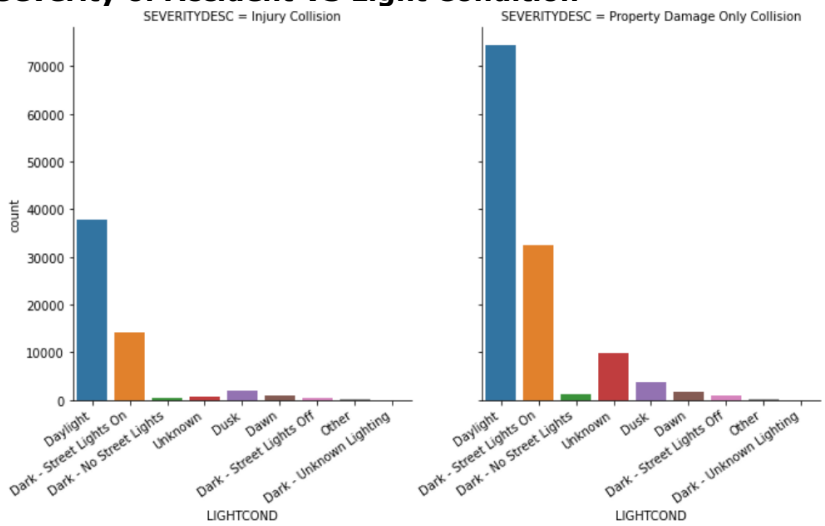
Weather does not play role here as the accidents, both property damage and injury are happening in clear weather conditions.

- **Severity of Accident VS Road Condition**



Most collisions happened in Dry Road, irrespective of the bad road conditions.

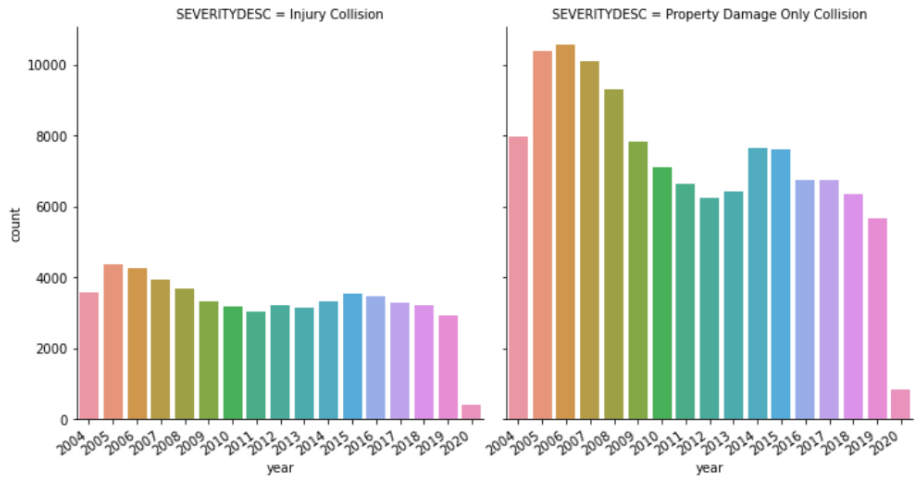
- **Severity of Accident VS Light Condition**



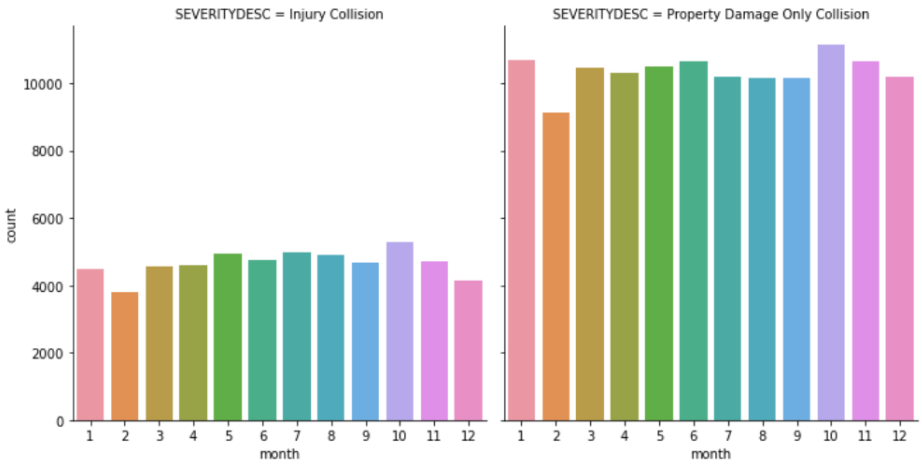
Light conditions do not lead to more accidents as the high number of collisions are recorded under daylight or Dar-Streetlights are on.

- **Severity of Accident VS Time**

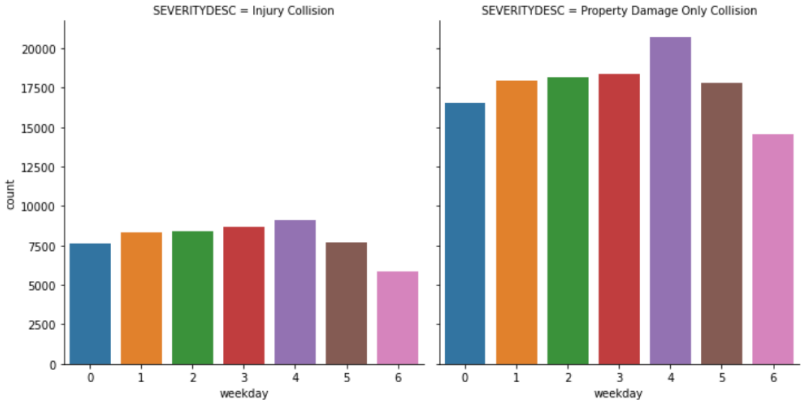
```
<seaborn.axisgrid.FacetGrid at 0x7fd1775b60f0>
```



The number of accidents has fallen since 2005-2009, increases again from 2013, then drops from 2016.



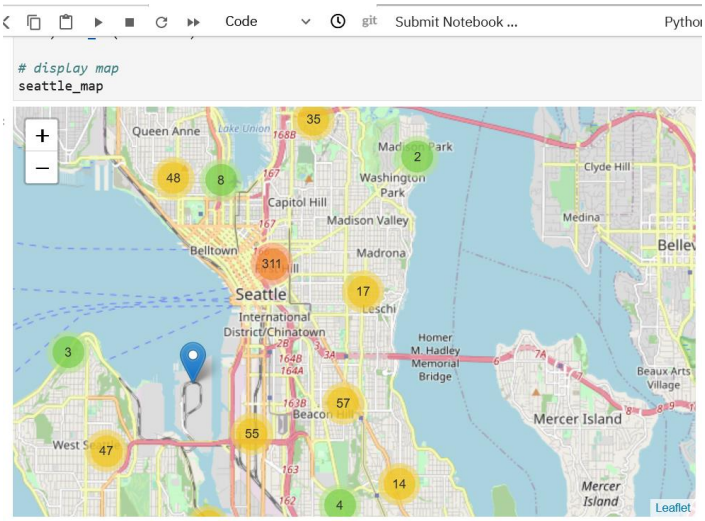
Except in Jan & October number of cases among different months are almost the same.



Most of the accidents seems to be happening around Thursday, 4th working day of the week.

3.2 Spatial Analysis

Since the recorded data is for 16years we can consider only year 2020.
(Zoomable map code is available in the code file.



Collisions are highly concentrated on the downtown of Seattle.

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 Preparing

- **Drop unnecessary attributes for modeling**

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

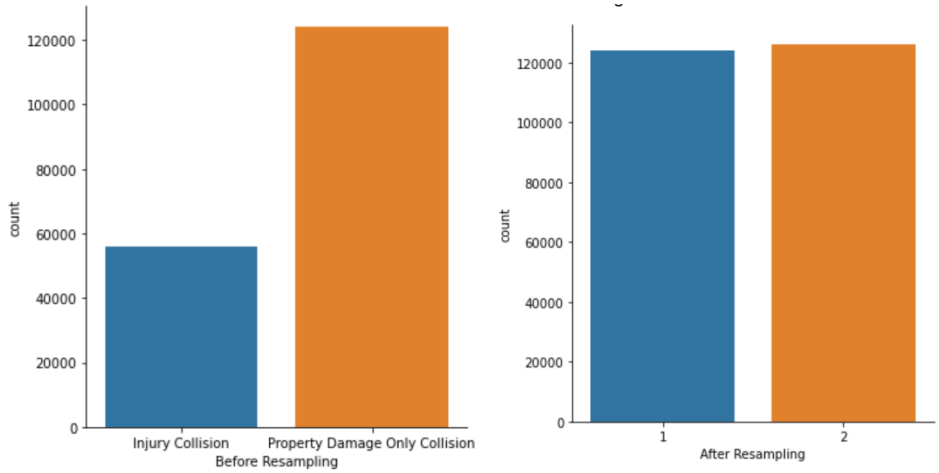
- **Create Dummy variables**

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 objects of label1 and only 58188 objects 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.

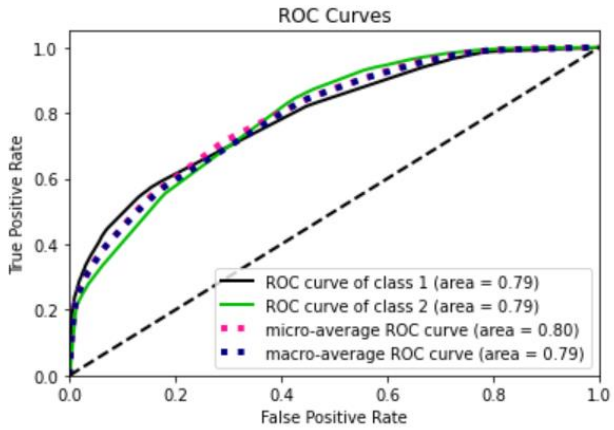
I use the training data to fit the logistic regression model, then make predictions based on the trained model, compare the true value and predicted value. then use Accuracy, F1-Score and AUC to evaluate the performance of the model.

4. Results

4.1 The Performance of Decision Tree model

Accuracy of Decision Tree: 0.7103740071049375

F1-Score of Decision Tree: 0.7051898592419864

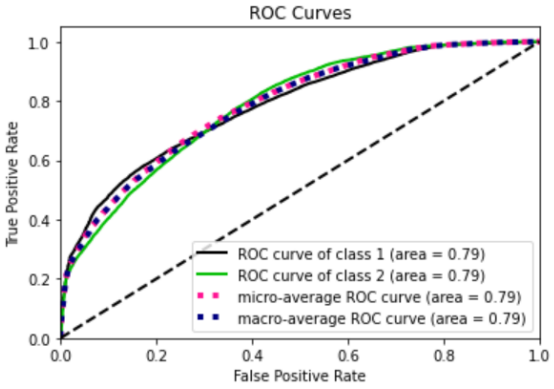


4.2 The Performance of Logistic Regression model

Logistic Regression's Accuracy: 0.7038012746311154

Logistic Regression's F1-Score: 0.7004139410689796

Logistic Regression's LogLoss : 0.5445183510374939



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 high for a machine learning model.

5. Observations

Findings from both exploratory data analysis & machine learning model are summarized below:

- Collision accidents happened at intersections are more dangerous because they result in injury to people rather than the one at block which is high but mostly property damage only.
- As per this recorded data on accidents, environmental conditions such as weather, light & road are not critical factors as surprisingly the number of accidents are high in good weather and clear road.
- It is observed that more accidents are on the 4th workday of the week, i.e. specifically Thursday which may be due to the high work stress of the drivers/people. At the same time collisions on Saturdays are less since people may be taking rest at home.
- 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.

7. References

Seattle Metro Data from "https://www.macrotrends.net"

Seattle Car Owners data from "https://www.seattletimes.com"

Call accidents analysis from ASIRT "https://www.asirt.org "

National Highway safety information from "https://www.nhtsa.gov/"

WA - DOT "https://wsdot.wa.gov/"