Introduction/Business Problem

This project aims to predict the accident “severity”. Which factors have more impact on the accidents such as weather, road condition, light condition, speeding or any other type of accidents.

Background Discussion

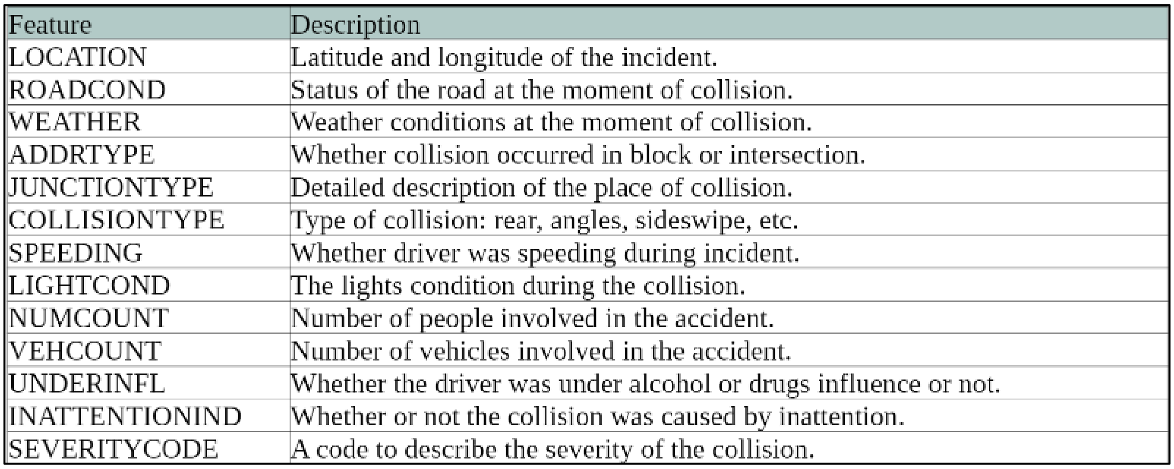
The society as a whole — the accident victims and their families, their employers, insurance firms, emergency and health care personal and many others — is affected by motor vehicle crashes in many ways. It would be great if real-time conditions can be provided to estimate the trip safeness. In this way, it can be decided beforehand if the driver will take the risk, based on reliable information.

Data

The data was collected by Seattle SPOT Traffic Management Division and provided by Coursera via a link. This dataset is updated weekly and is from 2004 to present. It contains information such as severity code, address type, location, collision type, weather, road condition, speeding, among others.

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 attributes are describes in the following tables:



Feature	Description
LOCATION	Latitude and longitude of the incident.
ROADCOND	Status of the road at the moment of collision.
WEATHER	Weather conditions at the moment of collision.
ADDRTYPE	Whether collision occurred in block or intersection.
JUNCTIONTYPE	Detailed description of the place of collision.
COLLISIONTYPE	Type of collision: rear, angles, sideswipe, etc.
SPEEDING	Whether driver was speeding during incident.
LIGHTCOND	The lights condition during the collision.
NUMCOUNT	Number of people involved in the accident.
VEHCOUNT	Number of vehicles involved in the accident.
UNDERINFL	Whether the driver was under alcohol or drugs influence or not.
INATTENTIONIND	Whether or not the collision was caused by inattention.
SEVERITYCODE	A code to describe the severity of the collision.

Our dependent variable is severitycode which is categorized as 1 if there is only property damage and 2 if it includes personal injuries.

Data Cleaning

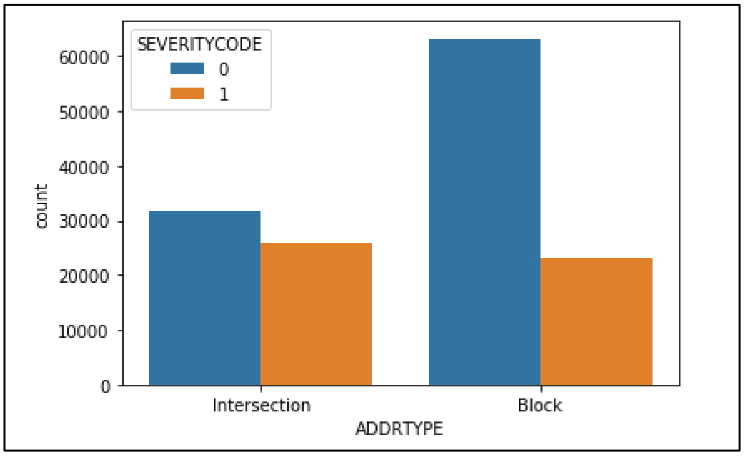
Some of the observations include 'Other' and 'Unknown' which does not provide enough information. Therefore, we dropped this entries from our dataset.

The UNDERINFL, INATTENTIONID and SPEEDING columns has Y and N which are transformed to 0 or 1.

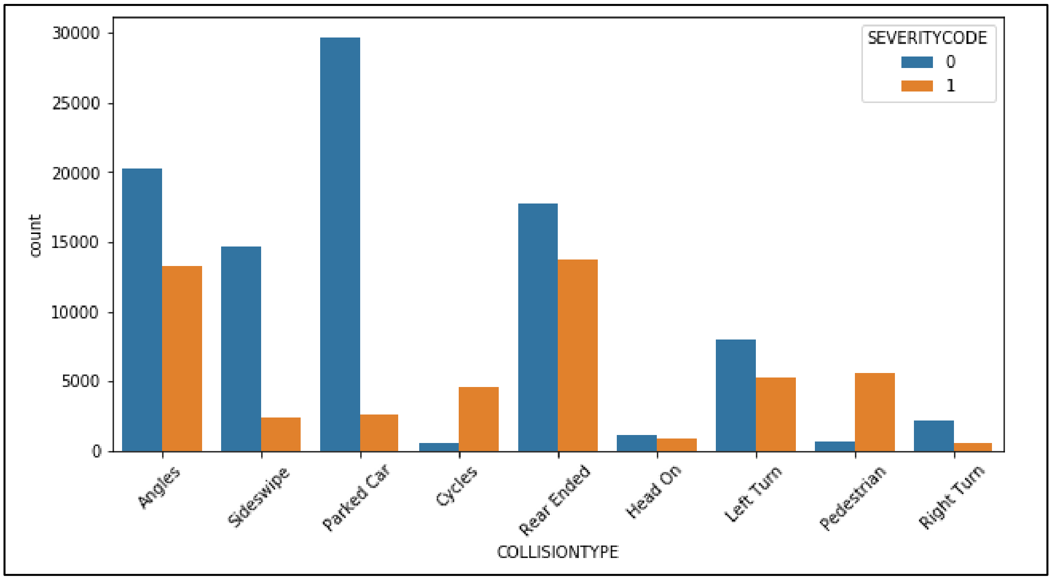
SEVERITYCODE column has 1 and 2 we converted them 0 and 1 for the sake of prediction.

Analysis

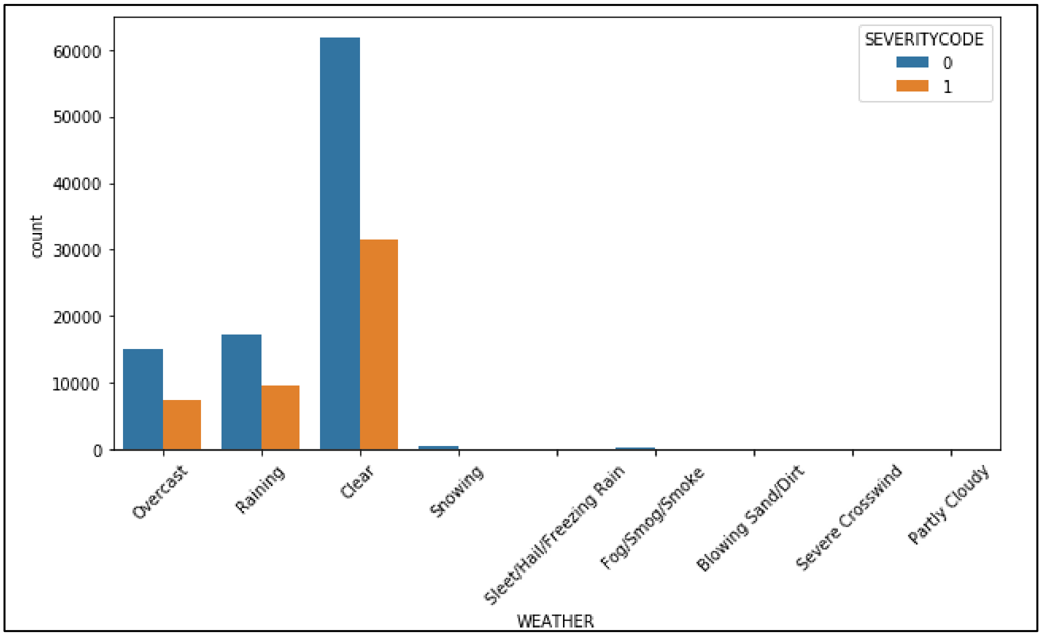
Before introducing the data to the classifier algorithms, we will explore he data if we can find some knowledge and insights from it.



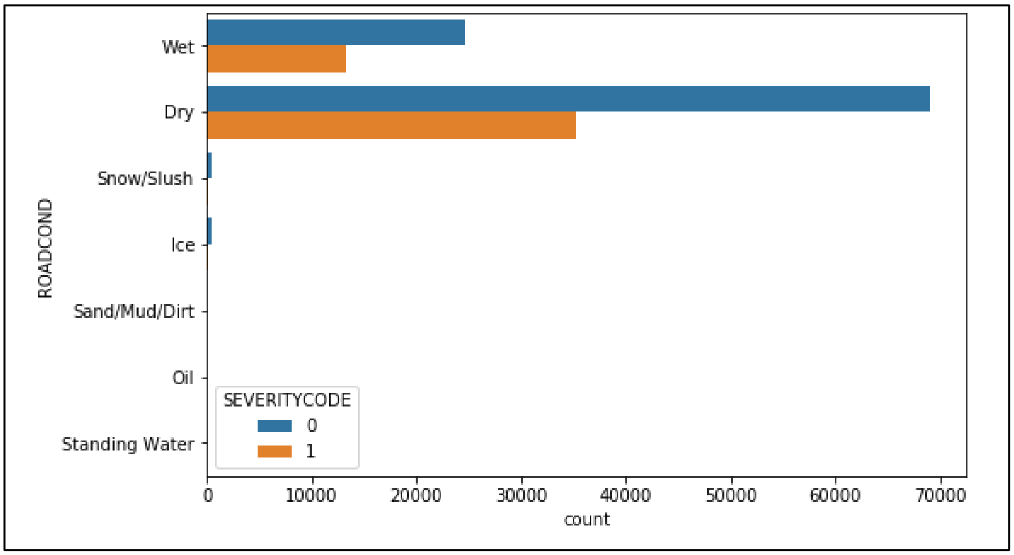
Looking at the below chart, we can observe a higher frequency for severe accidents in block rather than intersection.



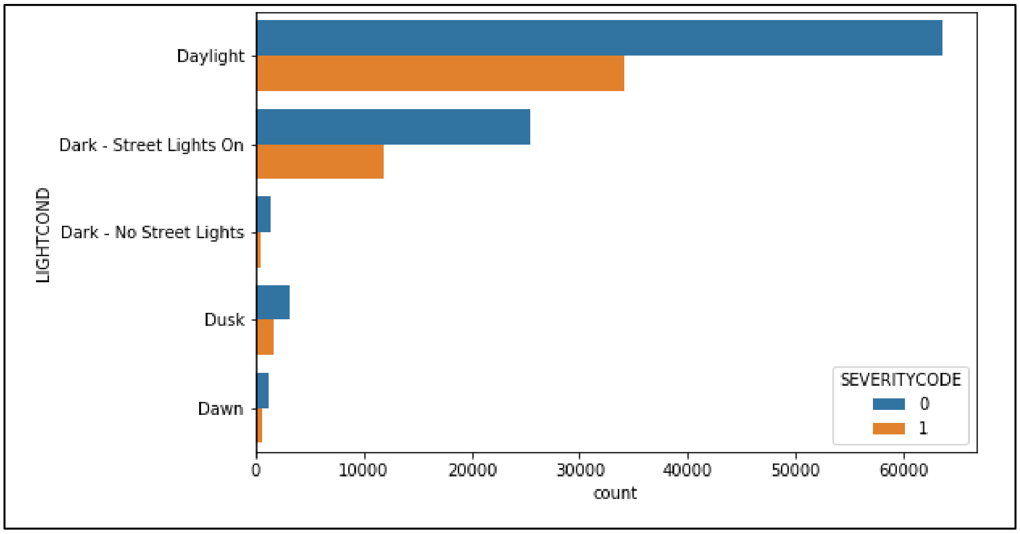
The most property damage happened in parked cars. Also we can conclude from this graph, left turn is riskier than right turn.



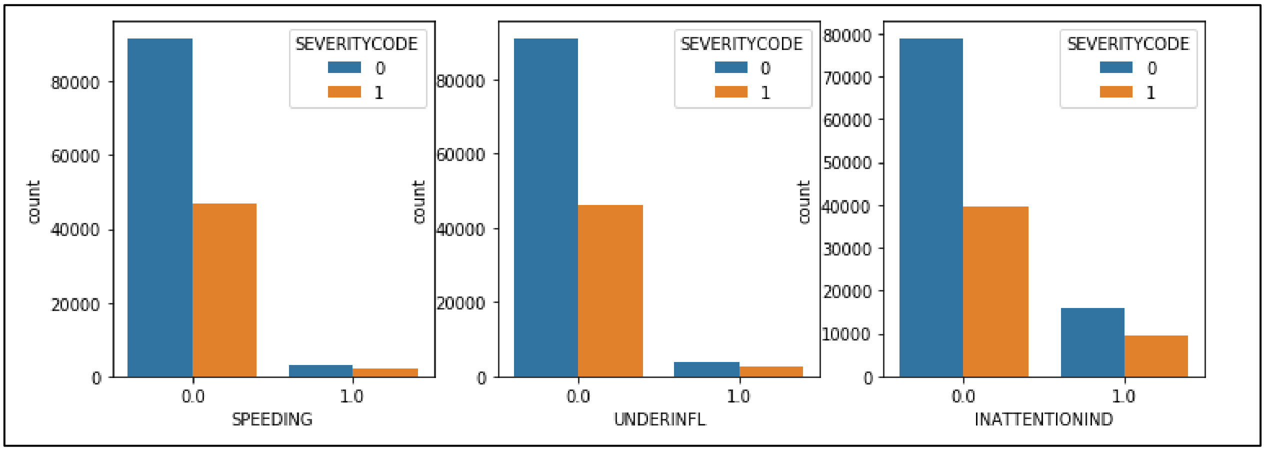
SEVERITYCODE
0
1
count
60000
50000
40000
30000
20000
10000
0
Overcast
Raining
Clear
Snowing
Sleet/Hail/Freezing Rain
Fog/Smog/Smoke
Blowing Sand/Dirt
Severe Crosswind
Partly Cloudy
WEATHER

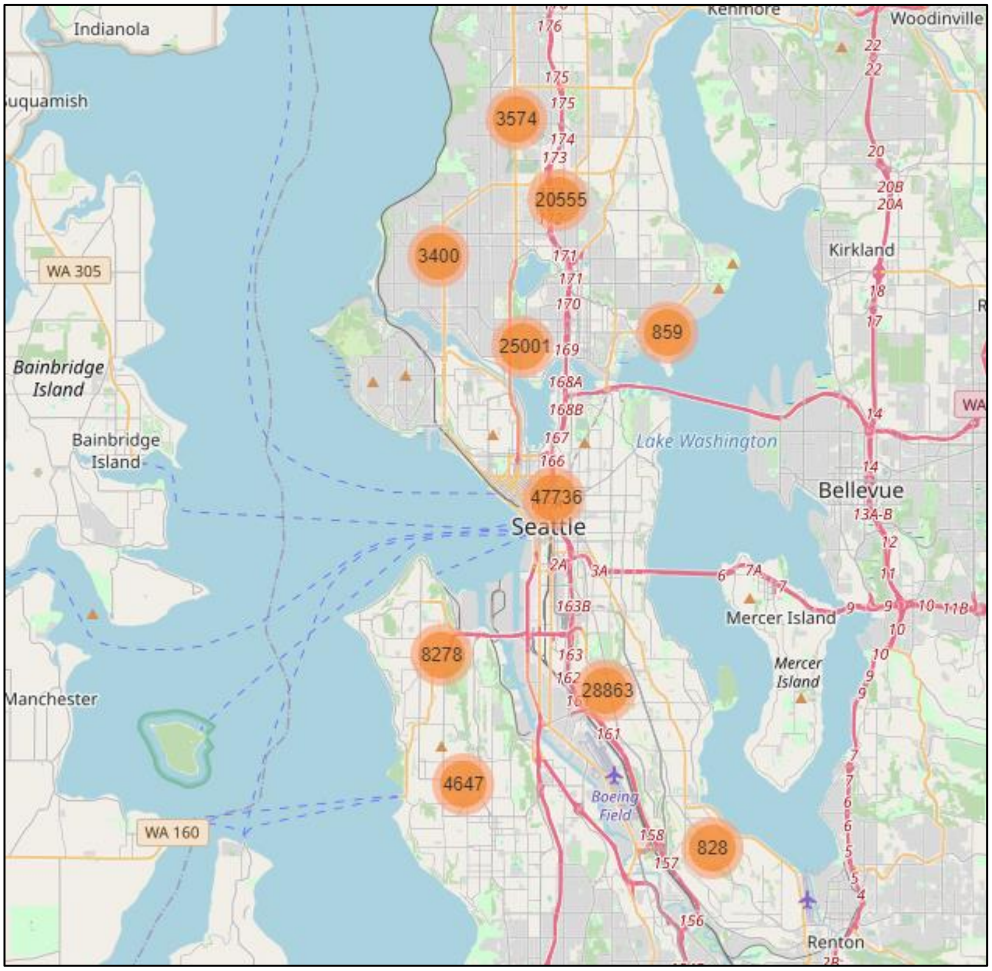


Wet
Dry
Snow/Slush
Ice
Sand/Mud/Dirt
Oil
Standing Water
ROADCOND
SEVERITYCODE
0
1
0
10000
20000
30000
40000
50000
60000
70000
count

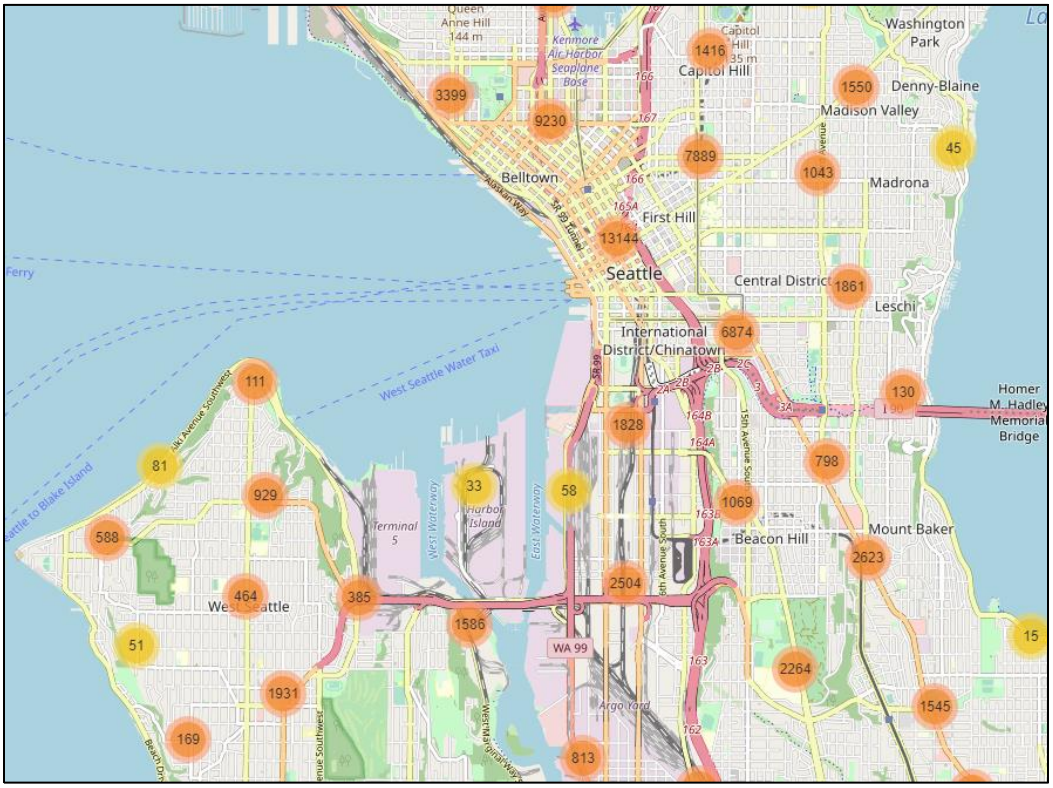


Surprisingly, the most property damage happened in clear weather. Snowing, rainy and other weather conditions have very low. Inside light conditions, daylight has more property damage. Moreover, dry road condition has more property damage than other road conditions.





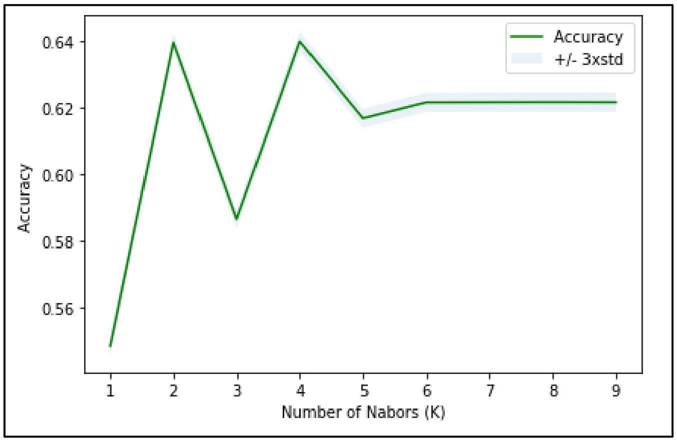
To see clustering collisions in different areas, we can look at map above. To have a look closer, we can conclude that the most of the accidents occurs in Pioneer Square, University Street and Belltown. These neighbors should take more attention and further evaluation from the local government and transportation division to increase infrastructure and to reduce the collision incidences.



Classification Models

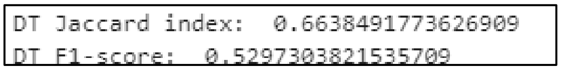
1- KNN Modeling

Firstly, we find the best k parameters to have best accuracy. The best accuracy was with 0.6898 with k=4.



2- Decision Tree Modeling

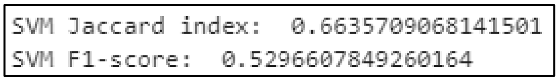
We indicated to use the entropy criterion to create the branches. We also used Jaccard and F1-score for accuracy.



```
DT Jaccard index:  0.6638491773626909
DT F1-score:  0.5297303821535709
```

3- SVM Modeling

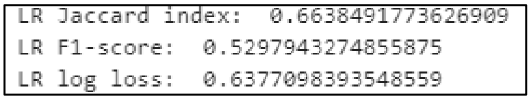
The results are the same as the ones obtained with the decision tree model. However, this is just a coincidence.



```
SVM Jaccard index:  0.6635709068141501
SVM F1-score:  0.5296607849260164
```

4- Logistic Regression

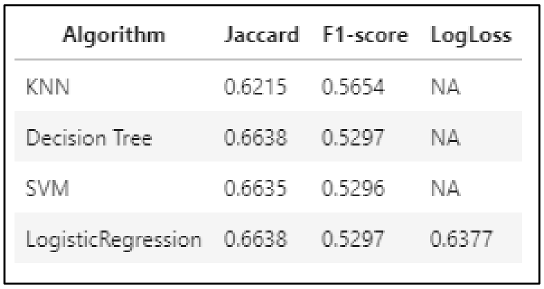
For the LR model, we have chosen the inverse of the regularization strength in C=0.01. The other parameters were left as default.



```
LR Jaccard index:  0.6638491773626909
LR F1-score:  0.5297943274855875
LR log loss:  0.6377098393548559
```

Summary

Based on the results obtained after training the different models, they are displayed in the table below.



Algorithm	Jaccard	F1-score	LogLoss
KNN	0.6215	0.5654	NA
Decision Tree	0.6638	0.5297	NA
SVM	0.6635	0.5296	NA
LogisticRegression	0.6638	0.5297	0.6377

The jaccard index which measures the similarity between the test and predicted set and ranges from 0 to 1, give us in the best approximation 0.64.

The f1-score which is determined based on the precision and the recall average of the sets, while predicting both target scenarios throws a similar number as the other scores, without showing too much dissimilarity between each prediction model.

Discussion

This project and analysis are quite helpful for the Seattle transportation department. Before I did the analysis, I thought that maybe weather, road, and light condition may cause more accidents, the results showed that it was not correct. However, we do figure out that the accidents are highly related to some specific locations. Thus, the traffic management division could try to improve the safety instructions or some other factors that could reduce the accidents.

Furthermore, there are some places which has more accidents during the dark time. For those places, adding lights might be a good solution to reduce the collisions.

Conclusion

Purpose of this project was to predict the accident severity in Seattle which conditions have higher impact of those accidents in order to aid stakeholders or government in narrowing down the search for optimal solution for reduce collisions. We used 4 different algorithms to predict severity. We found out that Decision Tree and Logistic Regression have more accuracy score than others.

Final decision on optimal solution will be made by stakeholders based on specific characteristics of conditions and locations.