CAR ACCIDENT SEVERITY IN SEATTLE

DEBASHRITA ROY

10/13/20

IBM Capstone Project

Contents

1. Introduction

1.1 Background

Car accidents are one of the most common hazards the world face in daily life. From minor fender-benders to catastrophic, multi-car pileups, getting into an accident in a motor vehicle can end a life or change it forever.

The seaport city of Seattle is the largest city in the state of Washington, as well as the largest in the Pacific Northwest. As of the latest census (https://www.theurbanist.org/2019/07/01/seattle-pushing-750000-with-steady-growth/) , there are 747,300 people living in Seattle. Seattle residents get around by car, trolley, streetcar, public bus, bicycle, on foot, and by rail. With such bustling streets, it's no surprise that much accidents take place there, each day.

According to 2019 data (https://www.colburnlaw.com/seattle-traffic-accidents/) **for Seattle from the Washington State Department of Transportation (WSDOT)**

- Fatal car accidents: 22
- Suspected Serious Injury collisions: 190
- Suspected Minor Injury: 834
- Possible Injury : 2,612
- No Apparent Injury: 6,657
- Total Crashes 10,315

1.2 Problem

Some of the causes for car accident in Seattle are as follows:

- Bad weather condition like foggy or rainy
- Tough road condition like wet or muddy
- Extreme Speeding
- Distracted Driving, busy with phones, drinking, fiddling with radio
- Failure to Obey Traffic Safety Devices due to poor light condition
- Unsafe Lane Changes

The project aims to predict how severity of accidents can be reduced based on a few factors.

1.3 Stakeholders

The reduction in severity of accidents can be beneficial to the Public Development Authority of Seattle which works towards improving those road factors and the car drivers themselves who may take precaution to reduce the severity of accidents.

2. Understanding Data and Exploratory Analysis

The dataset has total observations of 194673 with variation in number of observations for every feature. First of all, the total dataset was high variation in the lengths of almost every column of the dataset. The dataset had a lot of empty columns which could have been beneficial had the data been present there.

2.1 Feature Selection

Around 34 features dropped/deleted are: **OBJECTID', 'SEVERITYCODE.1', 'REPORTNO', 'INCKEY', 'COLDETKEY', 'X', 'Y', 'STATUS','ADDRTYPE’, TKEY', 'LOCATION', 'EXCEPTRSNCODE’, 'EXCEPTRSNDESC', 'SEVERITYDESC', 'INCDATE', 'INCDTTM', 'JUNCTIONTYPE', 'SDOT_COLCODE', 'SDOT_COLDESC', 'PEDROWNOTGRNT', 'SDOTCOLNUM', 'ST_COLCODE', 'ST_COLDESC', 'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR', 'PEDCOUNT', 'PEDCYLCOUNT', 'PERSONCOUNT', 'VEHCOUNT', 'COLLISIONTYPE', 'SPEEDING', 'UNDERINFL', 'INATTENTIONIND'**.

The selected features from the data set are:

Feature Variables	Description
WEATHER	Weather condition during time of collision (Overcast/Rain/Clear)
ROADCOND	Road condition during the collision (Wet/Dry..)
LIGHTCOND	Light conditions during the collision (Lights On/Dark with light on)
SEVERITY CODE	Property Damage and Injury Collision

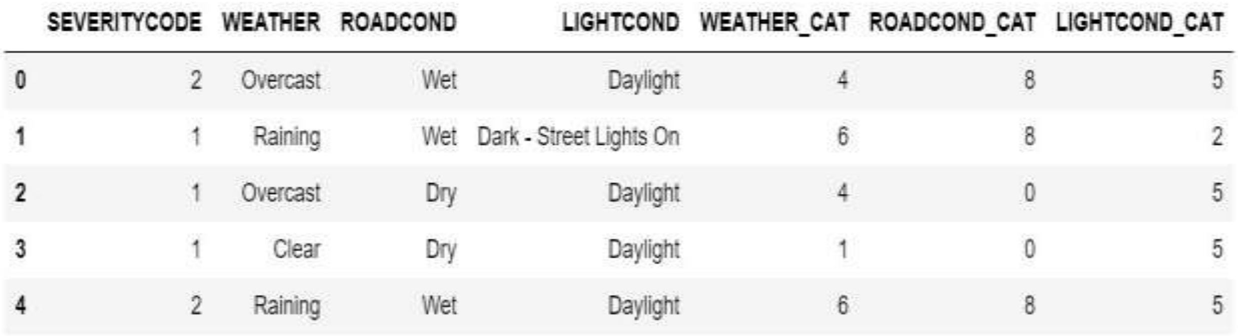
2.2 Label Encoding

The models aim is to predict the severity of an accident, considering that, the variable of Severity Code is in the form of 1 (Property Damage Only) and 2 (Injury Collision).

For lighting condition, Daylight is assigned 5 and Dark - Street Lights On is assigned 2.

For Road Condition, Dry is assigned 0 and Wet is assigned 8.
As for Weather Condition, 1 is Clear, 6 is Raining.

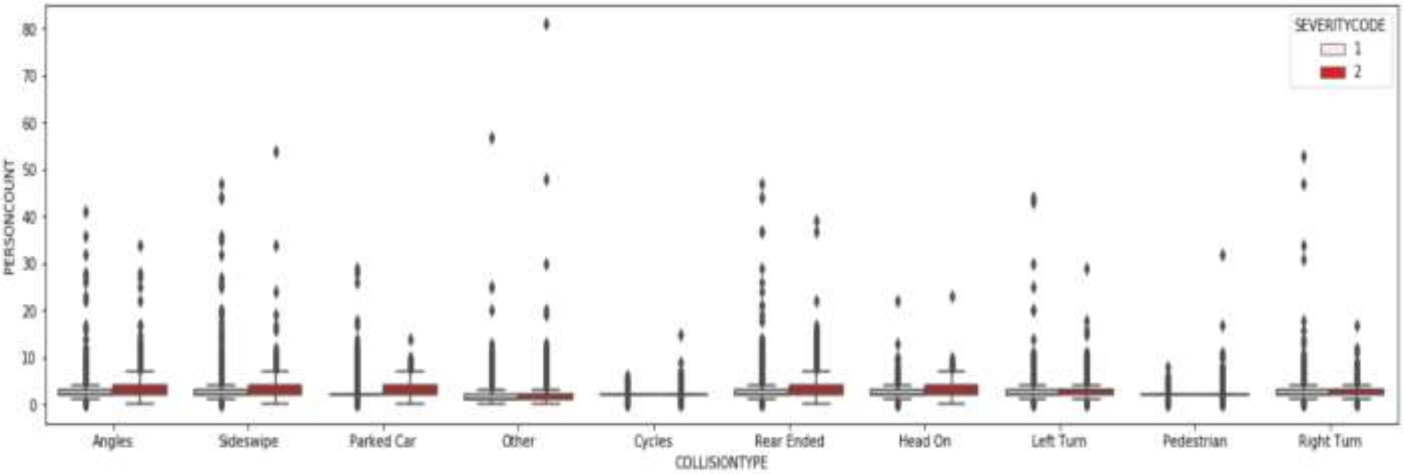
Whereas, there were unique values for every variable which were either ‘Other’ or ‘Unknown’, deleting those rows entirely would have led to a lot of loss of data which is not preferred.

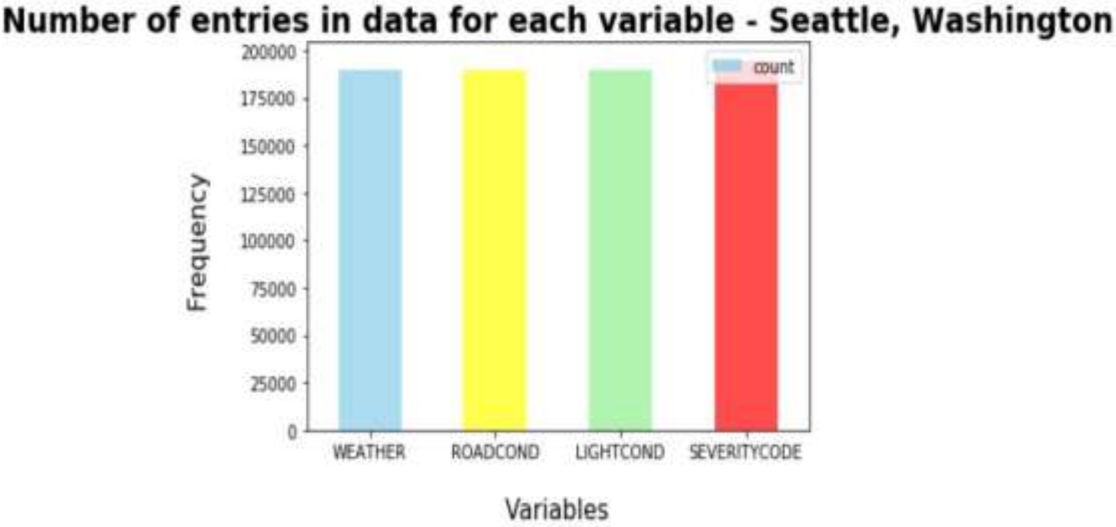


	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND	WEATHER_CAT	ROADCOND_CAT	LIGHTCOND_CAT
0	2	Overcast	Wet	Daylight	4	8	5
1	1	Raining	Wet	Dark - Street Lights On	6	8	2
2	1	Overcast	Dry	Daylight	4	0	5
3	1	Clear	Dry	Daylight	1	0	5
4	2	Raining	Wet	Daylight	6	8	5

2.3 Plotting

Using boxplot, bar plot, violin plot from matplotlib and seaborn libraries are used to depict, COLLISION TYPE and frequency for each chosen feature variable like WEATHER, LIGHTCOND, SEVERITY CODE etc.





Few other plots on basis of LOCATION and JUNCTIONTYPE are also shown.

2.4 Data Wrangling

- Resampling of data is done to balance Severity Code.
- For location data to map the accident, more than 5000 records who don't have X and Y data are dropped.
- Normalization of the data is performed.

3. Methodology

3.1 Data Collection

The dataset used for this project is based on car accidents which have taken place within the city of Seattle, Washington from the year 2004 to 2020. This data is regarding car accidents the severity of each car accidents along with the time and conditions under which each accident occurred. The data set used for this project can be found here.

3.2 Machine Learning Model Selection

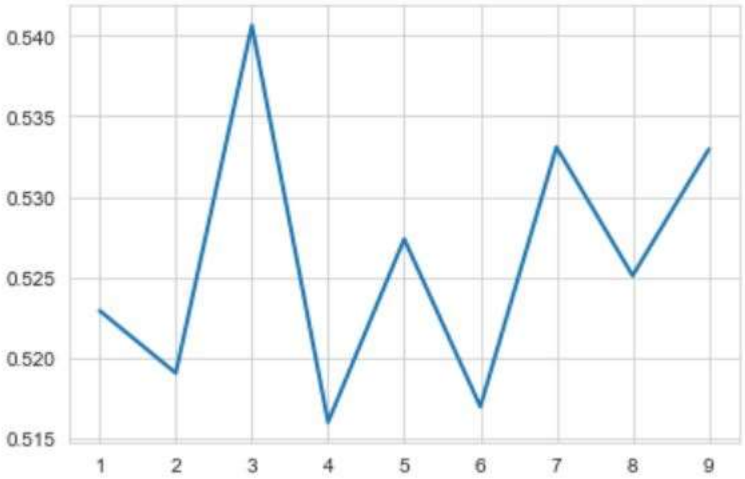
The machine learning models used are **k-Nearest Neighbor, Logistic Regression, Decision Tree Analysis and Support Vector Machine**. K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (based on distance). Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. The Decision Tree Analysis breaks down a data set into smaller subsets while at the same time an associated decision tree is incrementally developed. Support Vector Machines are supervised learning models with associated learning algorithms that analyze data used for classification.

Jaccard score, f1 score and log loss score are calculated with depiction of **classification report** and **confusion matrix plot** for each model.

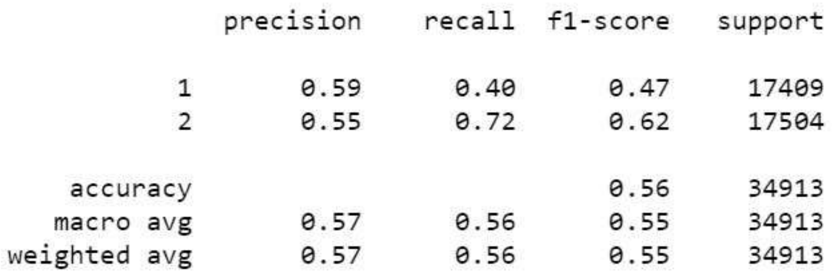
4. Results

4.1 k-Nearest Neighbor

The best K, as shown below, for the model where the highest elbow bend exists is at 3.



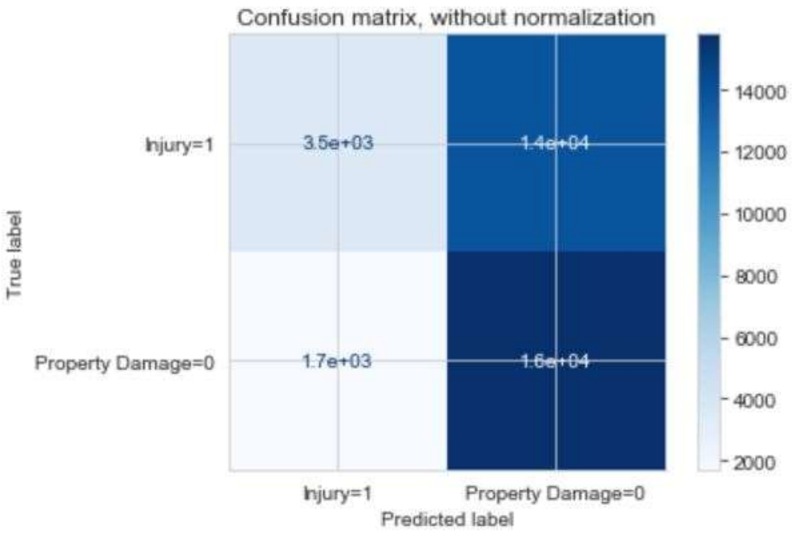
Classification Report is as follows:



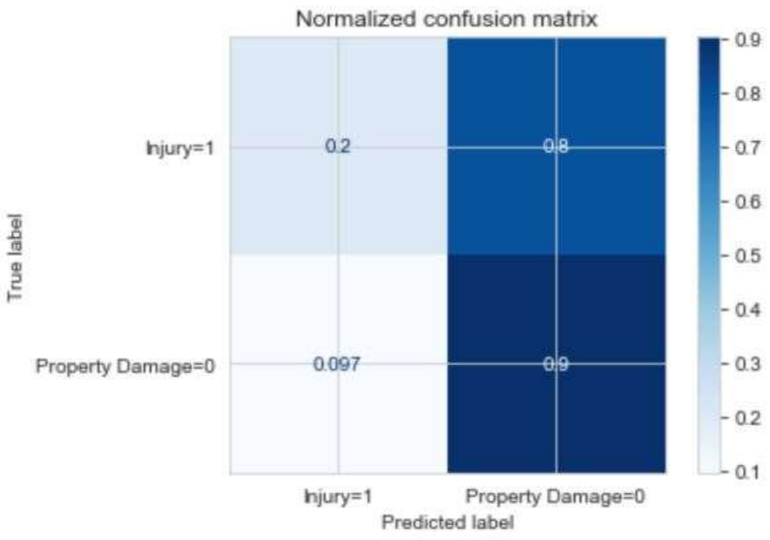
```
              precision    recall  f1-score   support

           1       0.59      0.40      0.47     17409
           2       0.55      0.72      0.62     17504

    accuracy                           0.56     34913
   macro avg       0.57      0.56      0.55     34913
weighted avg       0.57      0.56      0.55     34913
```

The confusion matrix with and without normalization are shown as:

```
Confusion matrix, without normalization
[[ 3543 13866]
 [ 1700 15804]]
```

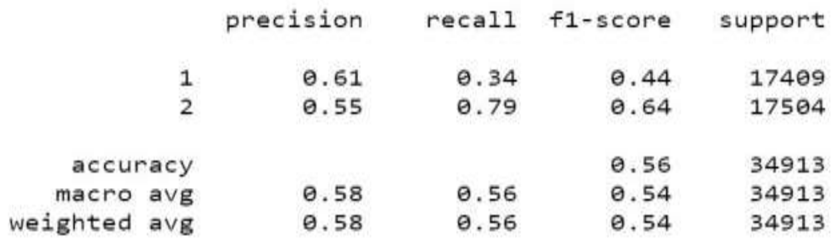


```
Normalized confusion matrix
[[0.20351542 0.79648458]
 [0.09712066 0.90287934]]
```



4.2 Decision Tree Analysis

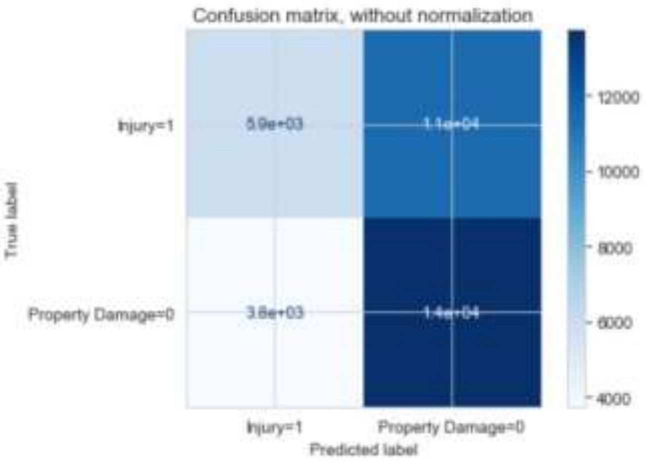
Classification Report is as follows:



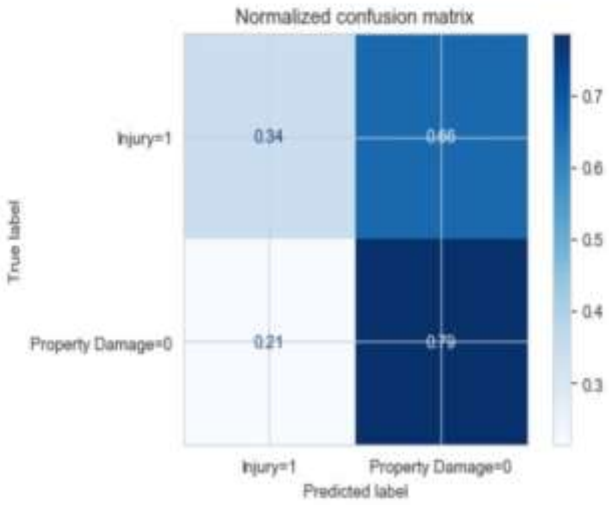
```
              precision    recall  f1-score   support

           1       0.61      0.34      0.44     17409
           2       0.55      0.79      0.64     17504

    accuracy                           0.56     34913
   macro avg       0.58      0.56      0.54     34913
weighted avg       0.58      0.56      0.54     34913
```

The confusion matrix with and without normalization are shown as:

```
Confusion matrix, without normalization
[[ 5937 11472]
 [ 3752 13752]]
```

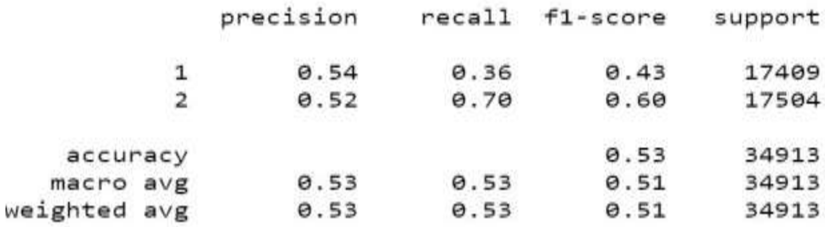


```
Normalized confusion matrix
[[0.3410305  0.6589695 ]
 [0.21435101 0.78564899]]
```



4.3 Logistic Regression

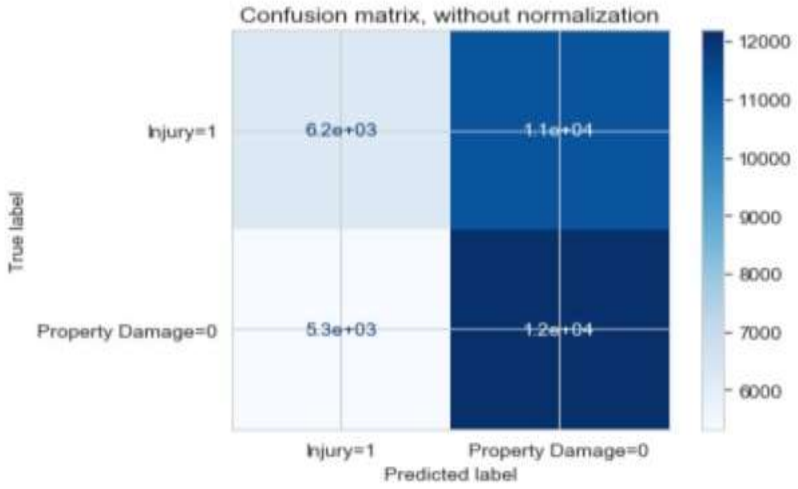
Classification Report is as follows:



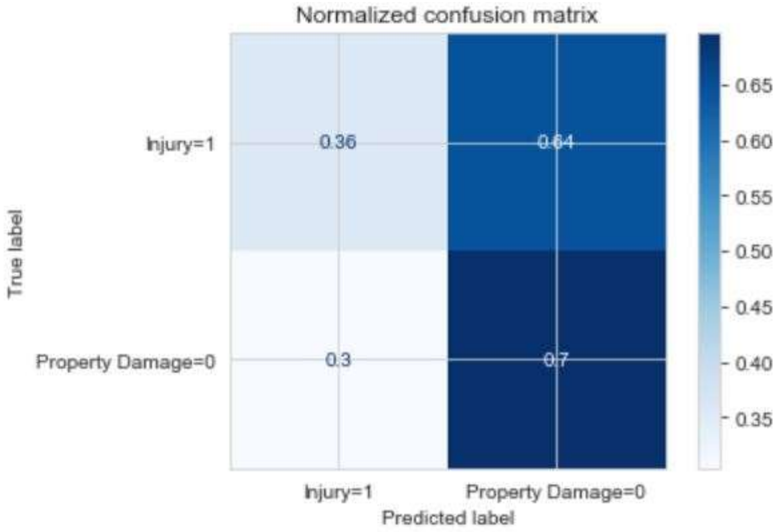
```
              precision    recall  f1-score   support

           1       0.54      0.36      0.43     17409
           2       0.52      0.70      0.60     17504

    accuracy                           0.53     34913
   macro avg       0.53      0.53      0.51     34913
weighted avg       0.53      0.53      0.51     34913
```

The confusion matrix with and without normalization are shown as:

```
Confusion matrix, without normalization
[[ 6183 11226]
[ 5322 12182]]
```

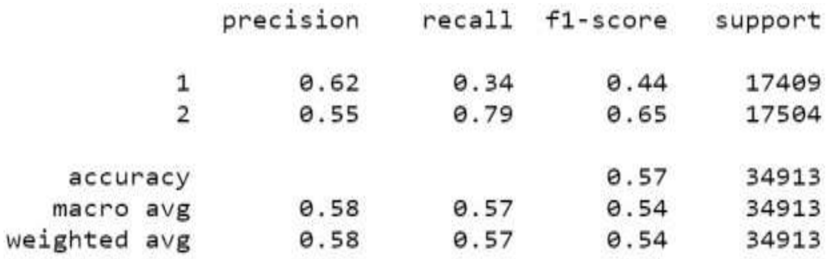


```
Normalized confusion matrix
[[0.35516112 0.64483888]
 [0.30404479 0.69595521]]
```



4.4 Support Vector Machine

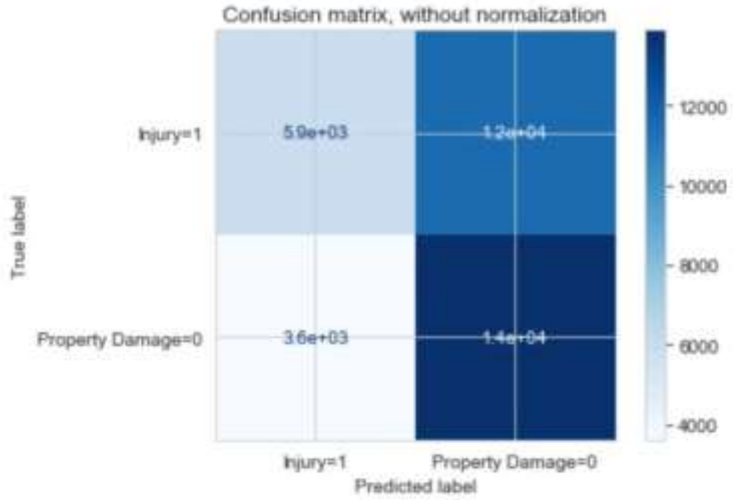
Classification Report is as follows:



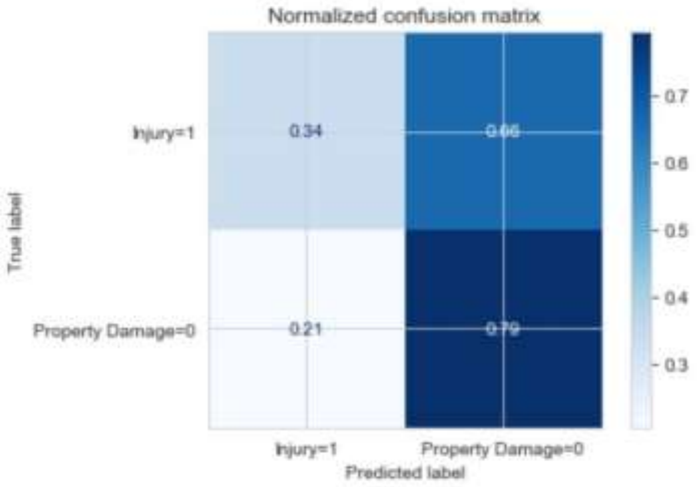
```
              precision    recall  f1-score   support

           1       0.62      0.34      0.44     17409
           2       0.55      0.79      0.65     17504

    accuracy                           0.57     34913
   macro avg       0.58      0.57      0.54     34913
weighted avg       0.58      0.57      0.54     34913
```

The confusion matrix with and without normalization are shown as:

```
Confusion matrix, without normalization
[[ 5867 11542]
 [ 3611 13893]]
```



```
Normalized confusion matrix
[[0.33700959 0.66299041]
 [0.2062957  0.7937043 ]]
```

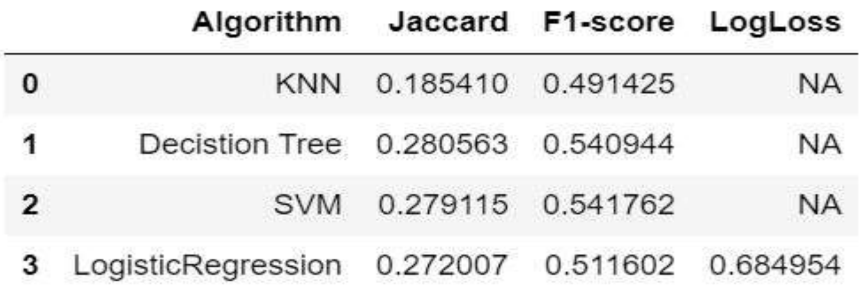


5. Discussion

Accuracy Score for each model are:

ALGORITHM	ACCURACY SCORE
K-Nearest Neighbor	0.554148884369719
Decision Tree Analysis	0.5639446624466531
Logistic Regression	0.5260218256809784
Support Vector Machine	0.5659782888895254

The final score table is:



	Algorithm	Jaccard	F1-score	LogLoss
0	KNN	0.185410	0.491425	NA
1	Decistion Tree	0.280563	0.540944	NA
2	SVM	0.279115	0.541762	NA
3	LogisticRegression	0.272007	0.511602	0.684954

Algorithm	Type	Precision	Recall	F1-score
k-Nearest Neighbor	Property Damage	0.68	0.20	0.31
	Injury Collision	0.53	0.90	0.67
Decision Tree	Property Damage	0.61	0.34	0.44
	Injury Collision	0.55	0.79	0.64
Logistic Regression	Property Damage	0.54	0.36	0.43
	Injury Collision	0.52	0.70	0.60
Support Vector Machine	Property Damage	0.62	0.34	0.44
	Injury Collision	0.55	0.79	0.65

5.1 PRECISION

Precision refers to the percentage of results which are relevant, in simpler terms it can be seen as how many of the selected items from the model are relevant. Mathematically, it is calculated by dividing true positives by true positive and false positive.

5.2 RECALL

Recall refers to the percentage of total relevant results correctly classified by the algorithm. In simpler terms, it tells how many relevant items are selected. It is calculated by dividing true positives by true positive and false negative.

5.3 AVERAGE F1-SCORE

f1-score is a measure of accuracy of the model, which is the harmonic mean of the model's precision and recall. Perfect precision and recall is shown by the f1-score as 1, which is the highest value for the f1-score, whereas the lowest possible value is 0 which means that either precision or recall is 0.

5.4 ACCURACY SCORE

In multilabel classification, the accuracy score function computes subset accuracy: the set of labels predicted for a sample must exactly match the corresponding set of labels in y_true.

5.5 JACCARD SCORE

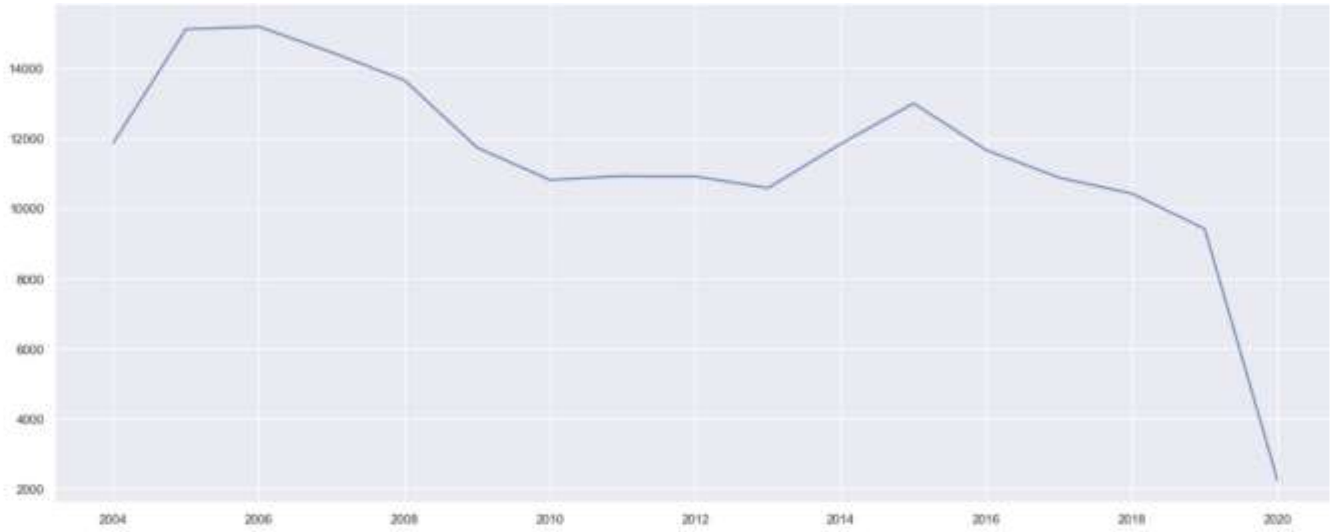
The Jaccard index [1], or Jaccard similarity coefficient, defined as the size of the intersection divided by the size of the union of two label sets, is used to compare set of predicted labels for a sample to the corresponding set of labels in y_true.

5.6 LOG LOSS SCORE

Log loss, aka logistic loss or cross-entropy loss is the loss function used in (multinomial) logistic regression and extensions of it such as neural networks, defined as the negative log-likelihood of a logistic model that returns y_pred probabilities for its training data y_true. The log loss is only defined for two or more labels.

6. Conclusion

The analysis of car accident severity in Seattle is performed.



This shows the year wise accident rate where a sharp rise is observed during 2005 but a decrease in can be seen at the present time.

Four classification models to predict what condition cause damage and collision is built.

These models can be very useful in helping the society in a number of ways.
For example:

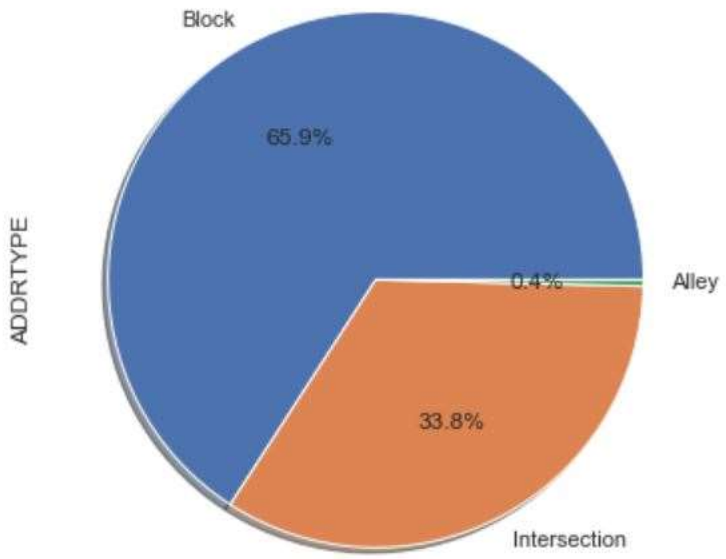
- Drivers can use this model to adjust their behavior to avoid injury;
- insurance company can use this model to adjust auto insurance premium level;
- Traffic management department can use the model to help decrease future injury collision by providing better light and road infrastructure.

.

7. Recommendation

This data to assess when to take extra precautions on the road under the given circumstances of light condition, road condition and weather, in order to avoid a severe accident, if any.

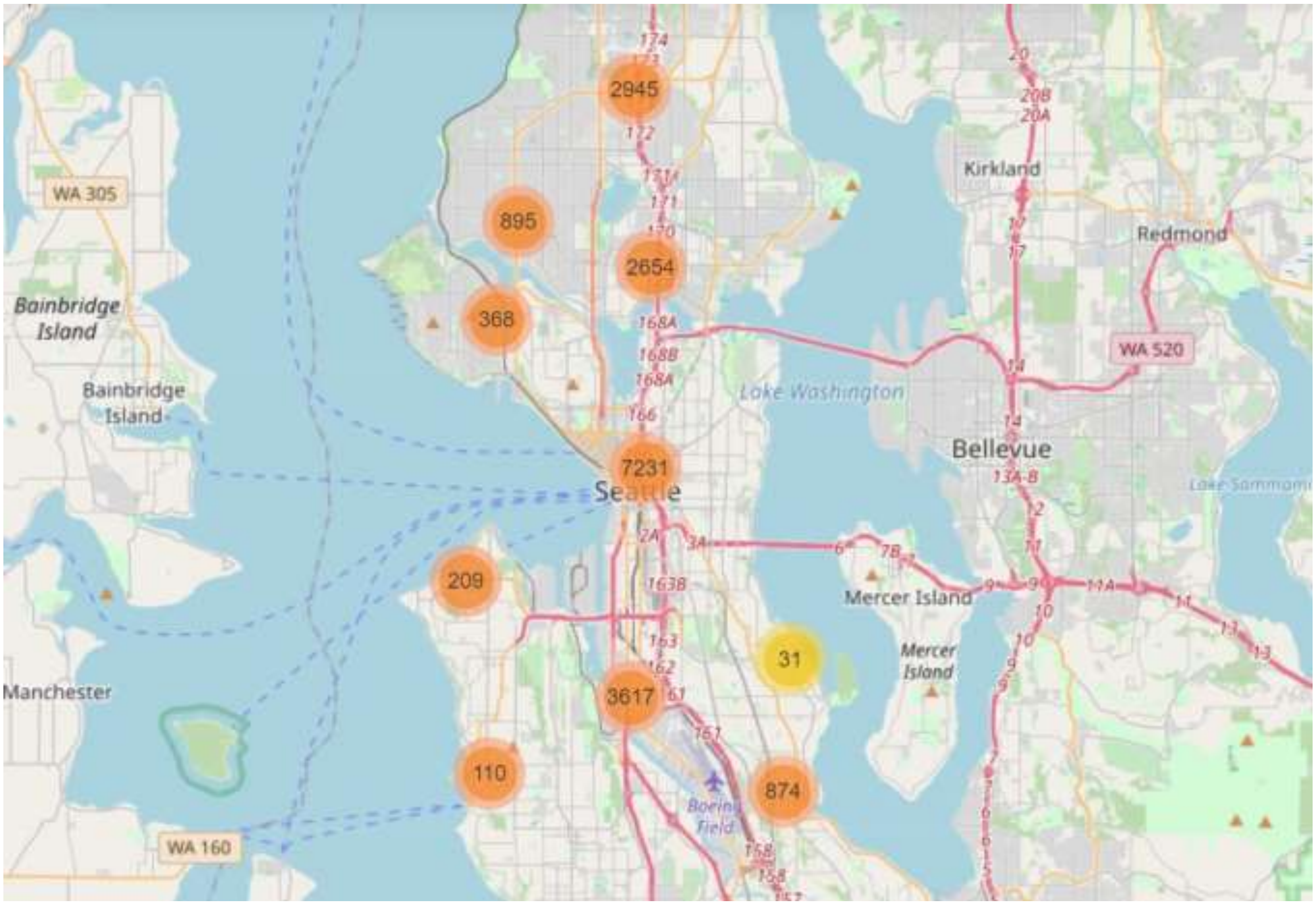
7.1 Public Development Authority of Seattle (PDAS)



Almost all of the accidents recorded have occurred on either a block or an intersection, the PDAS can take the following measures in response car accidents:

- Launch development projects for those areas where most severe accidents take place in order to minimize the problem.
- Increased investment towards improving lighting and road conditions of the area which have high instances recorded.
- Install safety signs on the roads and ensure that all precautions are being taken by people within the area.

7.2 Car Drivers



A higher concentration of accidents can be mostly seen on the main roads of the city, specifically near the highway in the city center.

Most incidents occur under adverse weather, road and light conditions. Precautions should be taken under such circumstances, for e.g. driving slow on a wet road which may lead to loss of control.

THANK YOU!