Final project report

1.1 Introduction :

It may surprise you to know that Seattle is the same as New York when it comes to traffic congestion. According to the TomTom Index Traffic Report, which was based on 2014 data, Seattle is tied with the Big Apple for having the fourth worst traffic congestion in the United States. This may not sound so bad – after all, Seattle is a big city. However, given that Seattle’s population is less than one-tenth that of New York’s, the fact that the two cities have the same amount of traffic congestion should make city planners in Seattle sit up and take notice.

1.2 Problem :

The world as whole 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 Stockholders :

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 the data:

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 in https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

2.1 data cleaning :

There are a lot of problems with this given dataset. The dataset has total observations of 194673 with variation in the number of observations for each parameter. First , 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. The models aim was to predict the severity of an accident.

2.2 feature selection :

A total of 5 features were selected for this project along with the target variable being Severity Code

FEATURE	DESCRIPTION
INATTENTIONIND	WHETHER OR NOT THE DRIVER WAS INATTENTIVE (Y/N)
UNDERINFL	WHETHER OR NOT THE DRIVER WAS UNDER THE INFLUENCE (Y/N)
WEATHER OVERCAST/RAIN/CLEAR	WEATHER CONDITION DURING TIME OF COLLISION
ROADCOND	ROAD CONDITION DURING THE COLLISION (WET/DRY..)
LIGHTCOND	LIGHT CONDITIONS DURING THE COLLISION (LIGHTS ON/DARK WITH LIGHT ON)
SPEEDING	WHETHER THE CAR WAS ABOVE THE SPEED LIMIT AT THE TIME OF COLLISION (Y/N)

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.

3.2 Exploratory Analysis :

Considering that the feature set and the target variable are categorical variables with the likes of

weather, road condition and light condition being an above level 2 categorical variables whose values

are limited and usually based on a particular finite group whose correlation might depict a different

image then what it is. Generally, considering the effect of these variables in car accidents are

important hence these variables were selected

3.3- Machine learning & Model selection :

The machine learning models used are Logistic Regression & Decision Tree Analysis. 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. The result is a tree with decision nodes and leaf nodes.

The reason why Decision Tree Analysis & Logistic Regression classification methods were chosen

is because the Support Vector Machine (SVM) model is inaccurate for large data sets, while this data set

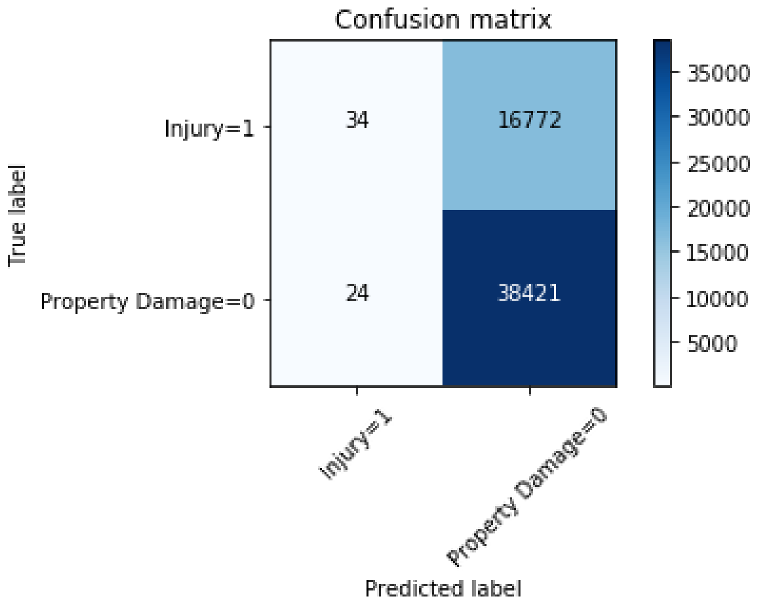
has more than 180,000 rows filled with data. Furthermore, SVM works best with dataset filled with text and images.

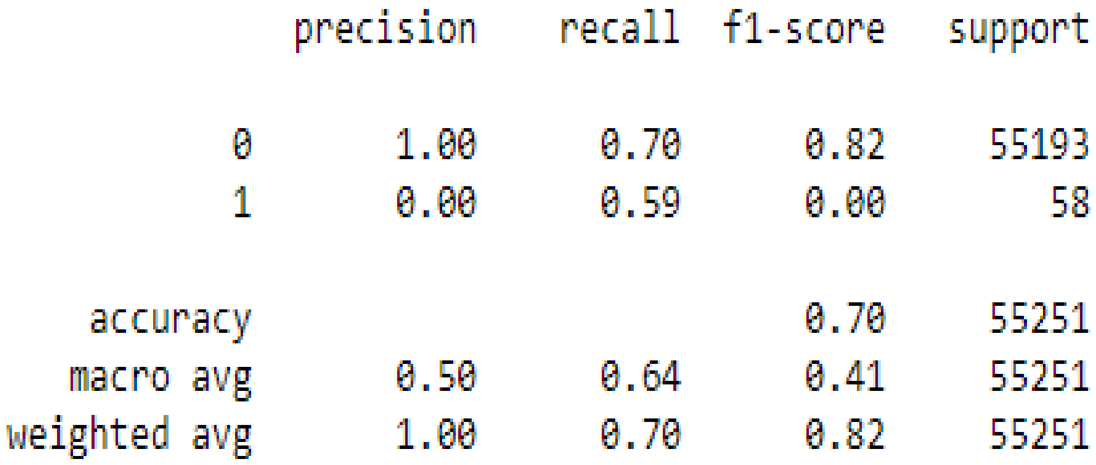
4- Result :

4.1 Decision Tree Analysis

Decision Tree Classifier from the scikit-learn library was used to run the Decision Tree Classification model on the Car Accident Severity data. The criterion chosen for the classifier was ‘entropy’ and the max depth was ‘6’.

The accuracy score for the decision tree classification model is : 0.6960050216

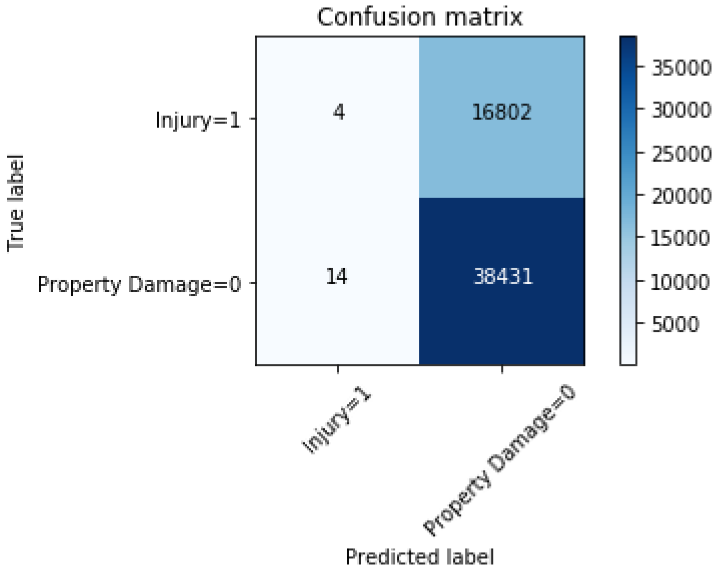


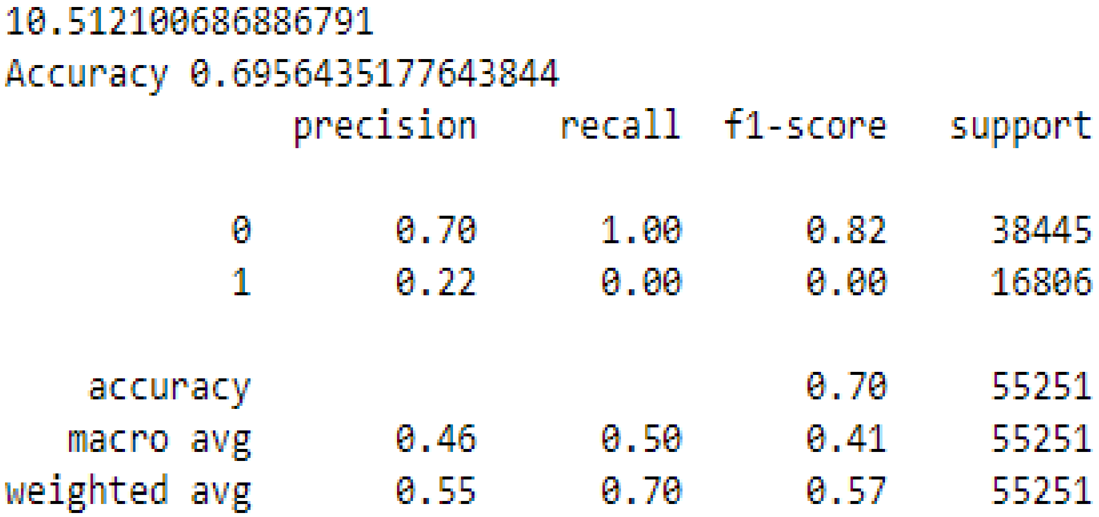


```
              precision    recall  f1-score   support

           0       1.00      0.70      0.82     55193
           1       0.00      0.59      0.00        58

    accuracy                           0.70     55251
   macro avg       0.50      0.64      0.41     55251
weighted avg       1.00      0.70      0.82     55251
```

4.2 Logistic Regression

Logistic Regression from the scikit-learn library was used to run the Logistic Regression Classification model on the Car Accident Severity data.





```
10.512100686886791
Accuracy 0.6956435177643844
              precision    recall  f1-score   support

           0       0.70      1.00      0.82     38445
           1       0.22      0.00      0.00     16806

    accuracy                           0.70     55251
   macro avg       0.46      0.50      0.41     55251
weighted avg       0.55      0.70      0.57     55251
```

5- Discussion :

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.

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. For the Decision Tree the precision of 0 is 1 and for 1 it is 0. for the Logistic Regression model, for 0 it is at 0.7 and for 1 it is 0.22.

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

6- conclusion :

When comparing all the models by their f1-scores, Precision and Recall, we can have a clearer picture in terms of the accuracy of the two models individually as a whole and how well they perform Car Accident Severity – Seattle, Washington .

When comparing these scores, we can see that the f1-score is highest for Logistic Regression at 0.7. However, later when we compare the recall for each of the model, we can see that the Decision tree model performs poorly in the recall. while when we compare the precision for each of the model we can find that the Logistic Regression performs poorly.

It can be concluded that the both the models can be used side by side for the best performance.