PROJECT REPORT

Harichandana Yeleswaram (yxh180027), Raisaat Rashid (rar150430)

Problem

The goal of this project was to predict the type of crime that is likely to happen at a specific location on the UTD campus at a given time of a day.

Dataset Generation

We have taken the UTD crime logs available online (here) in pdf format from year 2017 to 2020. We have automated the process of reading the pdf files using a python script for extracting the features that are needed and writing into a csv file.

Features that we have taken are:

1. **Day** when the crime has taken place i.e. Sunday through Saturday and converted them into numerical discrete values ranging from 1 to 7 respectively.
2. We put each crime into one of the following **crime type** categories after carefully looking at the dataset:
    a. Monetary Crimes
    b. Violent Crimes
    c. Possession or use of controlled substances
    d. Driving Violences
    e. Property (vandalization) crimes
    f. Other Agency Arrest Warrant
    g. False information to the police
    h. cyber crime
    i. mental illness and so on

    We converted these crime types into numerical discrete values ranging from 1 to 9 respectively.
3. **Time** when the crime has taken place and categorized into Early morning (12:00 am to 5:59 am), Morning from (6 am to 11:59 pm), Afternoon (12 pm to 3:59 pm), Evening (4 pm to 7:59 pm), Night (8 pm to 11:59 pm) and have given them numerical discrete values ranging from 1 to 5.
4. Location of the crime, i.e. where the crime has taken place. We have converted the location of the crime into numerical discrete values ranging from 1 to 111 (the same location was given the same number)

Our dataset (new_data_1.csv file) contains 653 examples. The csv file has 653 lines and each row has four feature columns. We did not collect data from years before 2017 because some of

the locations, more specifically buildings on the UTD campus, that were present before 2017 have ceased to exist. In addition, new buildings have been built on campus after 2017. As a result, the amount of data obtained could not be more than 653.

Initially before collecting the dataset, we wanted to predict the location where the crime has taken place given Day, Type of crime and Time but after testing with different algorithms it has given very low accuracy and it is expected as it has many feature labels.

Now, We are predicting the type of crime given Day, Time and Location of the crime.

Algorithms considered and selection of parameters

We tested two algorithms on our dataset - Scikit's Gradient Boosting and Scikit's SVM. To select the best parameters for each of these models, we used cross validation with 5 folds and 10 repetitions.

The gradient boosting algorithm was tested on learning rates 0.05, 0.075, 0.1 and 0.25 and number of estimators 20, 30, 45, 60 and 100. The following diagram shows the mean accuracy across the 5 folds of each combination of learning rate and number of estimators for gradient boosting:

```
Gradient Boosting:
Mean accuracy for learning rate =  0.05  num of estimators =  20 : 53.16%.
Mean accuracy for learning rate =  0.05  num of estimators =  30 : 54.12%.
Mean accuracy for learning rate =  0.05  num of estimators =  45 : 56.19%.
Mean accuracy for learning rate =  0.05  num of estimators =  60 : 57.46%.
Mean accuracy for learning rate =  0.05  num of estimators =  100 : 58.45%.
Mean accuracy for learning rate =  0.075  num of estimators =  20 : 54.24%.
Mean accuracy for learning rate =  0.075  num of estimators =  30 : 56.37%.
Mean accuracy for learning rate =  0.075  num of estimators =  45 : 57.40%.
Mean accuracy for learning rate =  0.075  num of estimators =  60 : 58.33%.
Mean accuracy for learning rate =  0.075  num of estimators =  100 : 59.68%.
Mean accuracy for learning rate =  0.1  num of estimators =  20 : 55.97%.
Mean accuracy for learning rate =  0.1  num of estimators =  30 : 57.40%.
Mean accuracy for learning rate =  0.1  num of estimators =  45 : 58.28%.
Mean accuracy for learning rate =  0.1  num of estimators =  60 : 59.25%.
Mean accuracy for learning rate =  0.1  num of estimators =  100 : 59.97%.
Mean accuracy for learning rate =  0.25  num of estimators =  20 : 57.70%.
Mean accuracy for learning rate =  0.25  num of estimators =  30 : 59.03%.
Mean accuracy for learning rate =  0.25  num of estimators =  45 : 59.26%.
Mean accuracy for learning rate =  0.25  num of estimators =  60 : 59.19%.
Mean accuracy for learning rate =  0.25  num of estimators =  100 : 58.07%.

Maximum accuracy obtained with learning rate = 0.1 and num of estimators = 100 and the accuracy is 59.97%.
```

The SVM algorithm was tested on C values 0.01, 0.1, 1, 10, 100 and 1000 and gamma values 10, 100 and 1000. The following diagram shows the mean accuracy across the 5 folds of each combination of C value and gamma value for SVM:

```
SVM:
Mean accuracy for C =  0.01 gamma = 10 : 47.78%.
Mean accuracy for C =  0.01 gamma = 100 : 47.78%.
Mean accuracy for C =  0.01 gamma = 1000 : 47.78%.
Mean accuracy for C =  0.1 gamma = 10 : 47.78%.
Mean accuracy for C =  0.1 gamma = 100 : 47.78%.
Mean accuracy for C =  0.1 gamma = 1000 : 47.78%.
Mean accuracy for C =  1 gamma = 10 : 64.44%.
Mean accuracy for C =  1 gamma = 100 : 64.44%.
Mean accuracy for C =  1 gamma = 1000 : 64.44%.
Mean accuracy for C =  10 gamma = 10 : 64.44%.
Mean accuracy for C =  10 gamma = 100 : 64.44%.
Mean accuracy for C =  10 gamma = 1000 : 64.44%.
Mean accuracy for C =  100 gamma = 10 : 64.44%.
Mean accuracy for C =  100 gamma = 100 : 64.44%.
Mean accuracy for C =  100 gamma = 1000 : 64.44%.
Mean accuracy for C =  1000 gamma = 10 : 64.44%.
Mean accuracy for C =  1000 gamma = 100 : 64.44%.
Mean accuracy for C =  1000 gamma = 1000 : 64.44%.

Maximum accuracy obtained with C = 1 and gamma = 10 and the accuracy is 64.44%.
```

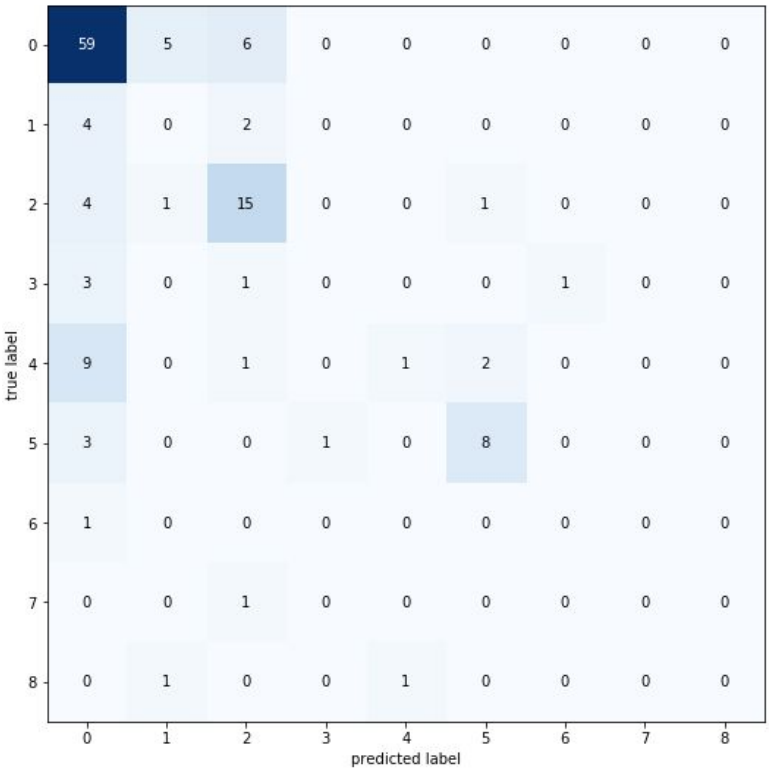
Final Models

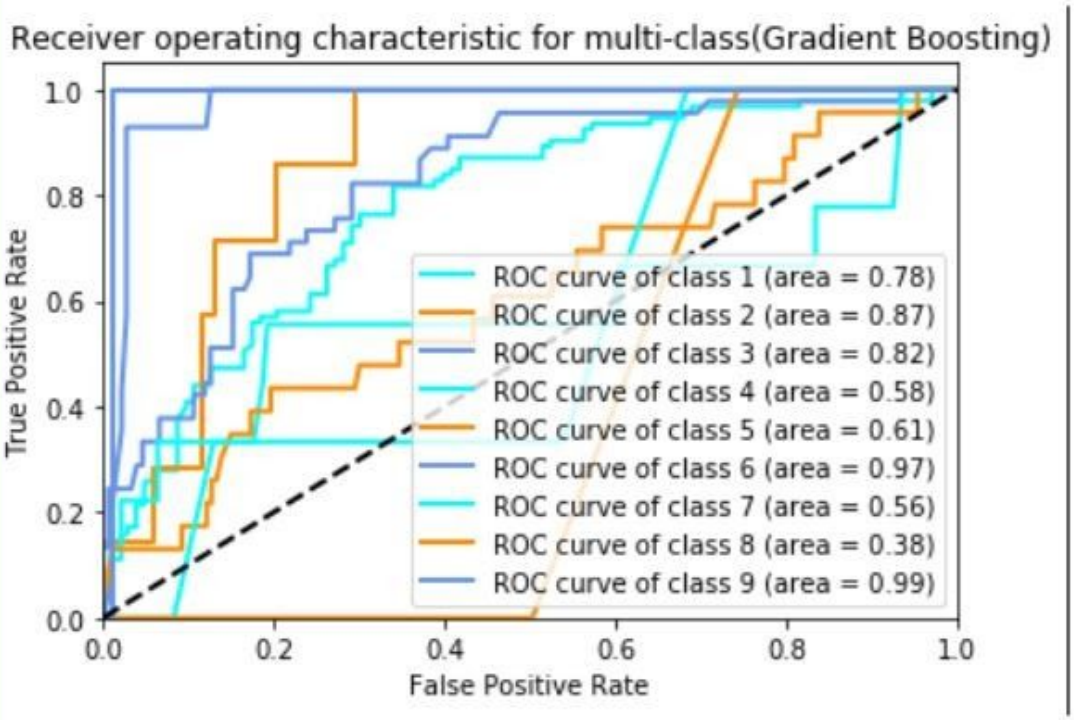
To test the final models, we trained the models on the same training set and tested them on the same testing set. The training and testing sets were created randomly with a specific seed value.

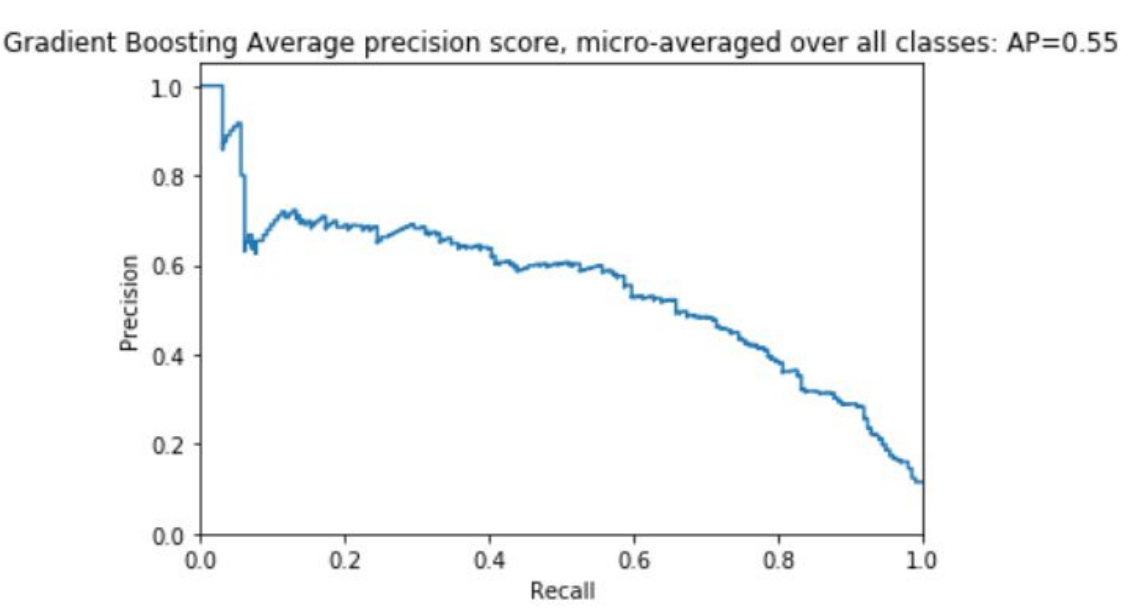
Gradient Boosting:

For gradient boosting, learning rate = 0.1 and number of estimators = 100, gave the maximum mean accuracy. So the gradient boosting model was trained using these parameter values.

The following diagram shows the confusion matrix obtained for gradient boosting:



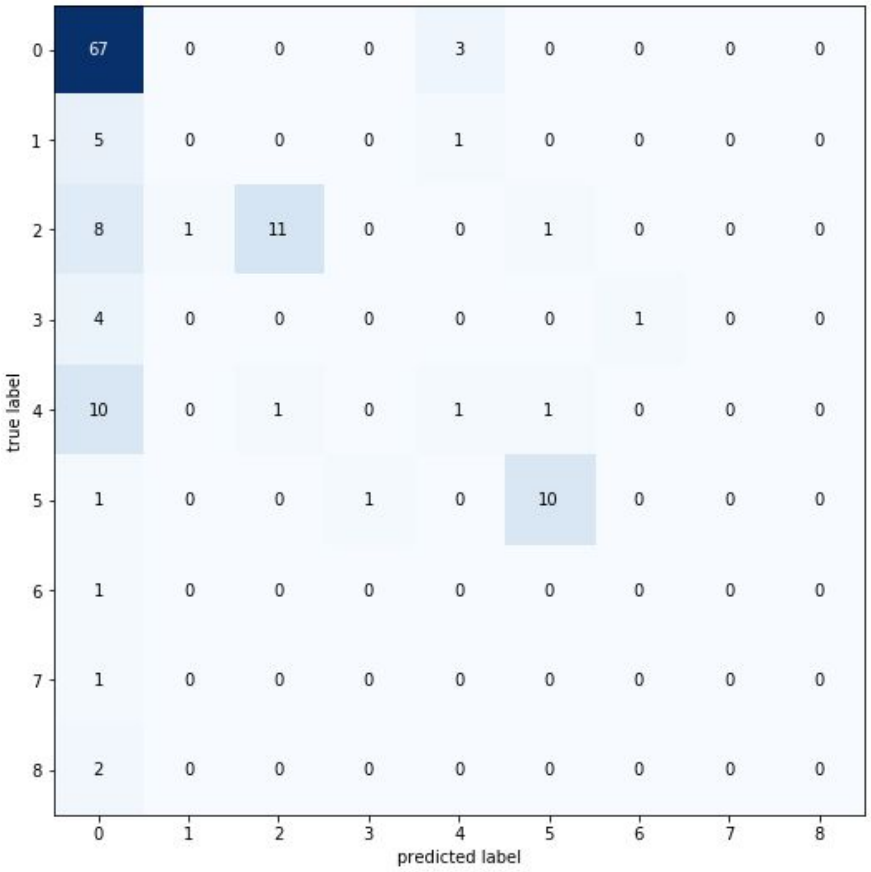


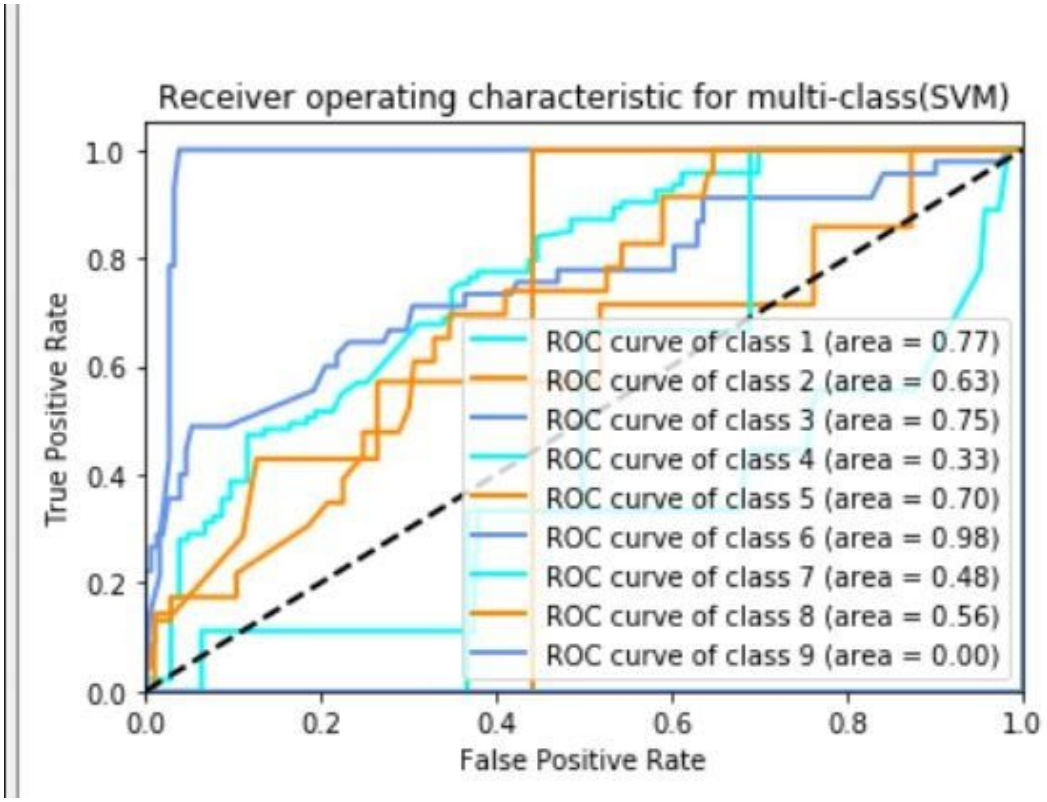


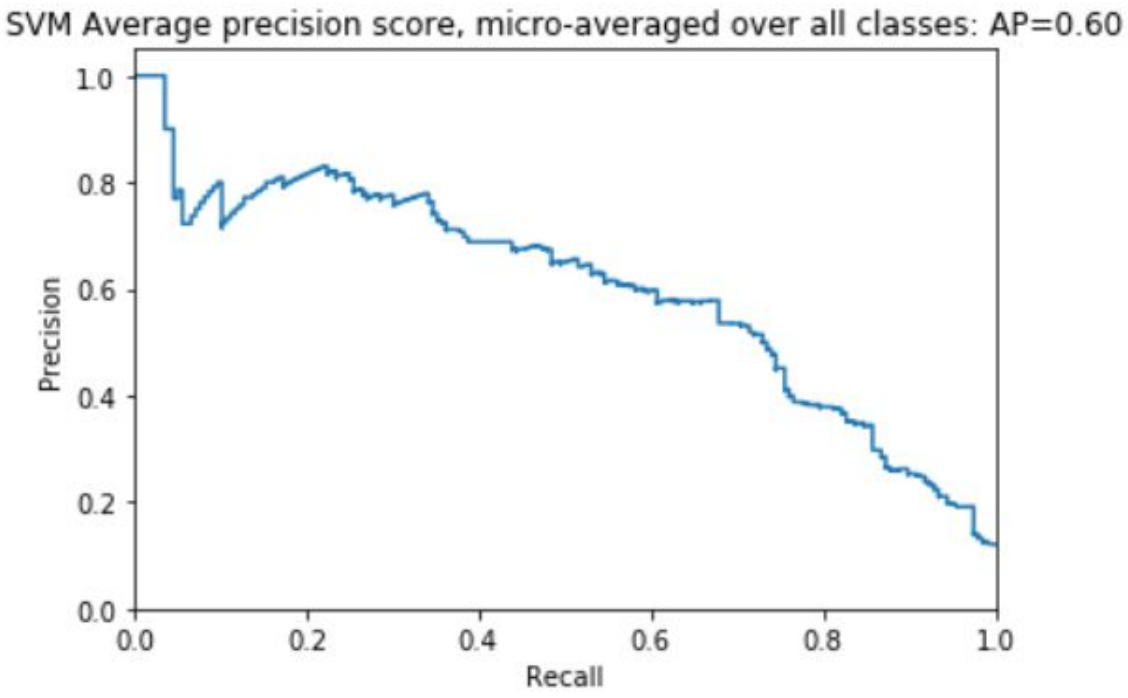
SVM:

For SVM, C = 1 and gamma = 10, gave the maximum mean accuracy. So the SVM model was trained using these parameter values.

The following diagram shows the confusion matrix obtained for SVM:







*The code for plotting ROC curves was obtained from here and the code for plotting precision-recall graphs was obtained from here.