Machine Learning Assignment 2 - Lovey Dovey

Objective

We are training a binary classifier, attempting to get good generalization performance.

Resources Used
Python 3.7
Numpy
Pandas
Sklearn
Matplotlib

Algorithms Tried

1. Logistic Regression
2. K means - Unsupervised Algorithm
3. Random Forest
4. Support Vector Machine
5. Convolutional Neural Network
6. Multi Layer Perceptron - Best among all

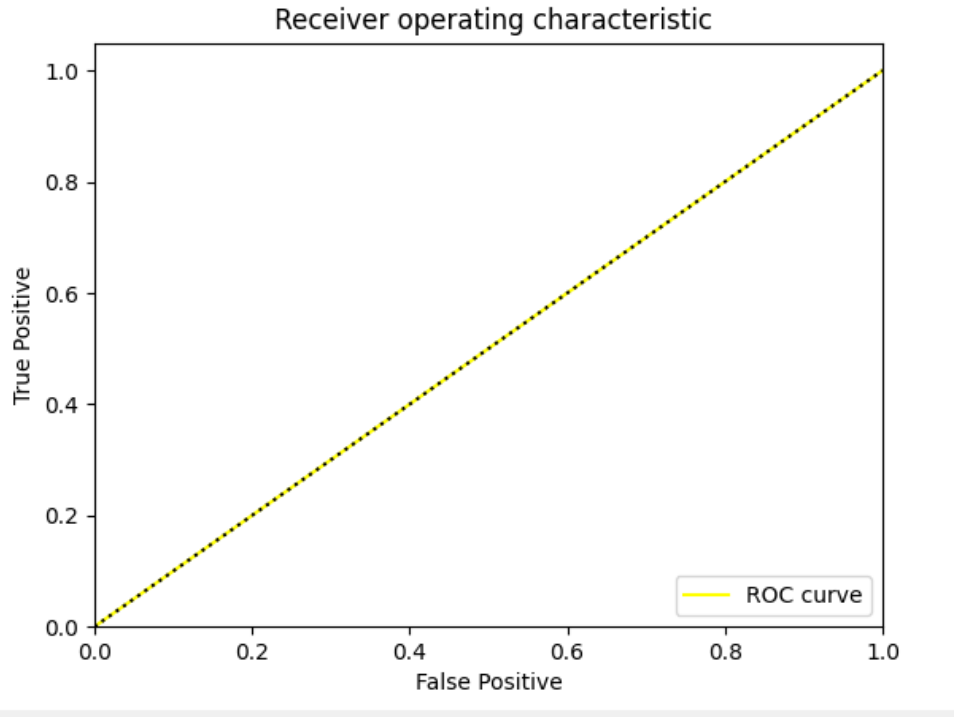
Detailed Analysis of Algorithms

Here we have detailed analysis of all given algorithms including **UNSUPERVISED K-MEANS** with Accuracy, Confusion Matrix and ROC

1. **Logistic Regression**

Accuracy : 0.50005
Confusion matrix:
[[9947 52]
[9947 54]]

ROC

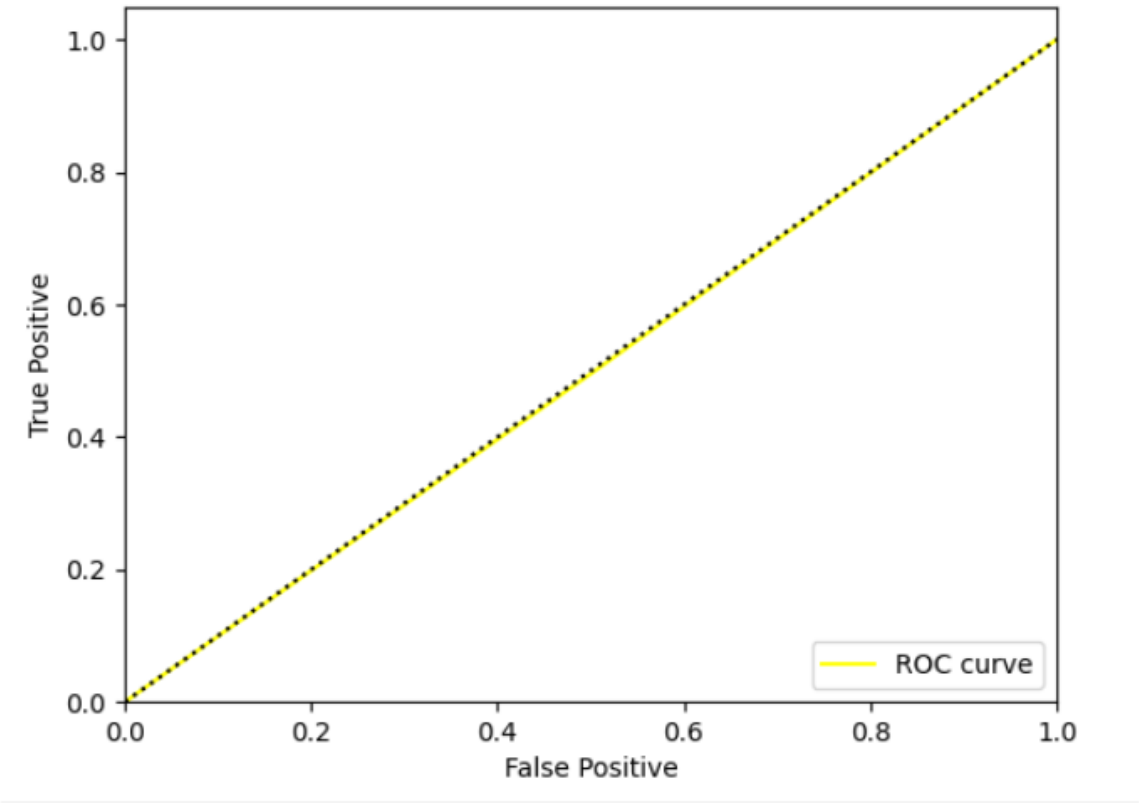


There accuracy is just over 0.5 and indicates that this model is only average at its performance with AUC around 0.5

2. **K means - Unsupervised Algorithm**

Accuracy : 0.49765
confusion matrix:
[[4965 5034]
[5013 4988]]

ROC for K-Means unsupervised algorithm

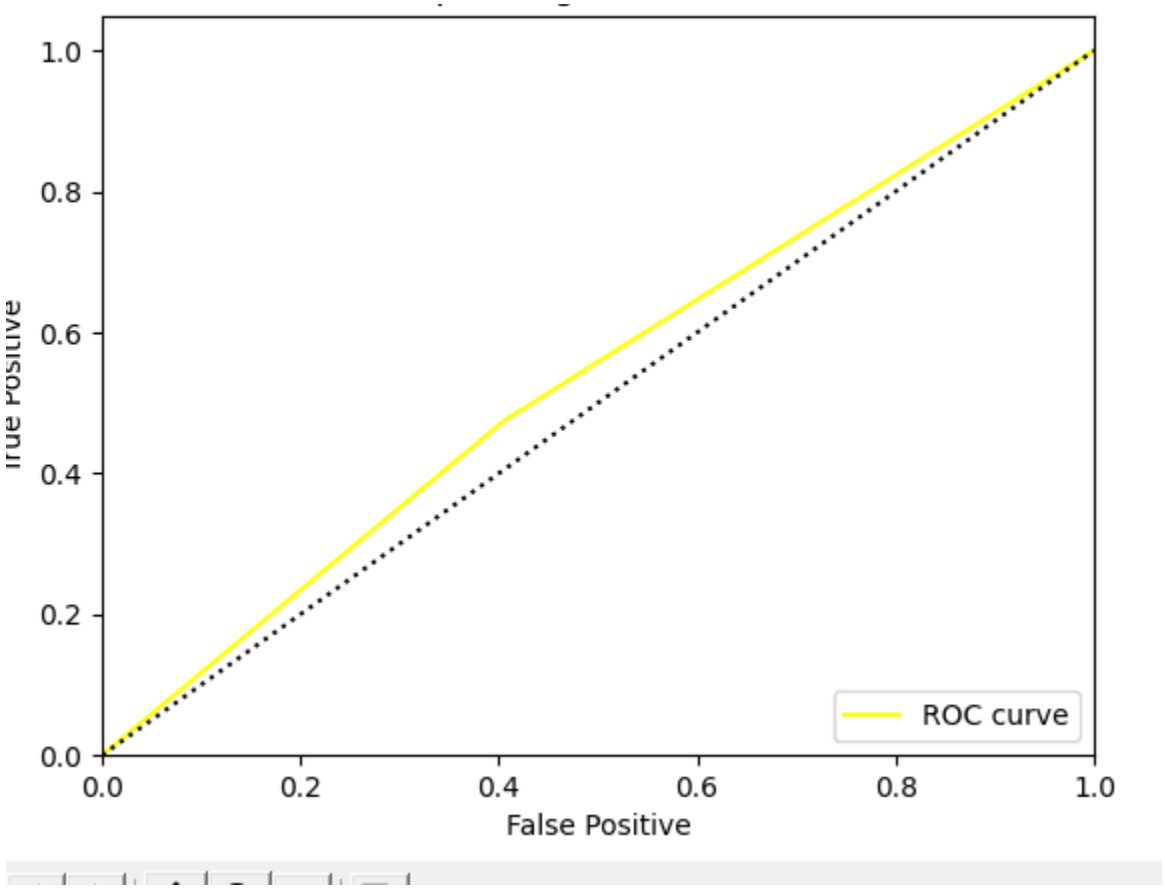


Unsupervised Learning gives AUC around 0.5 only. Confusion matrix also proves there are significant number of false positives and false negatives

3. Random Forest

Accuracy : 0.53455
confusion matrix:
[[5947 4052]
[5257 4744]]

ROC for Random Forest



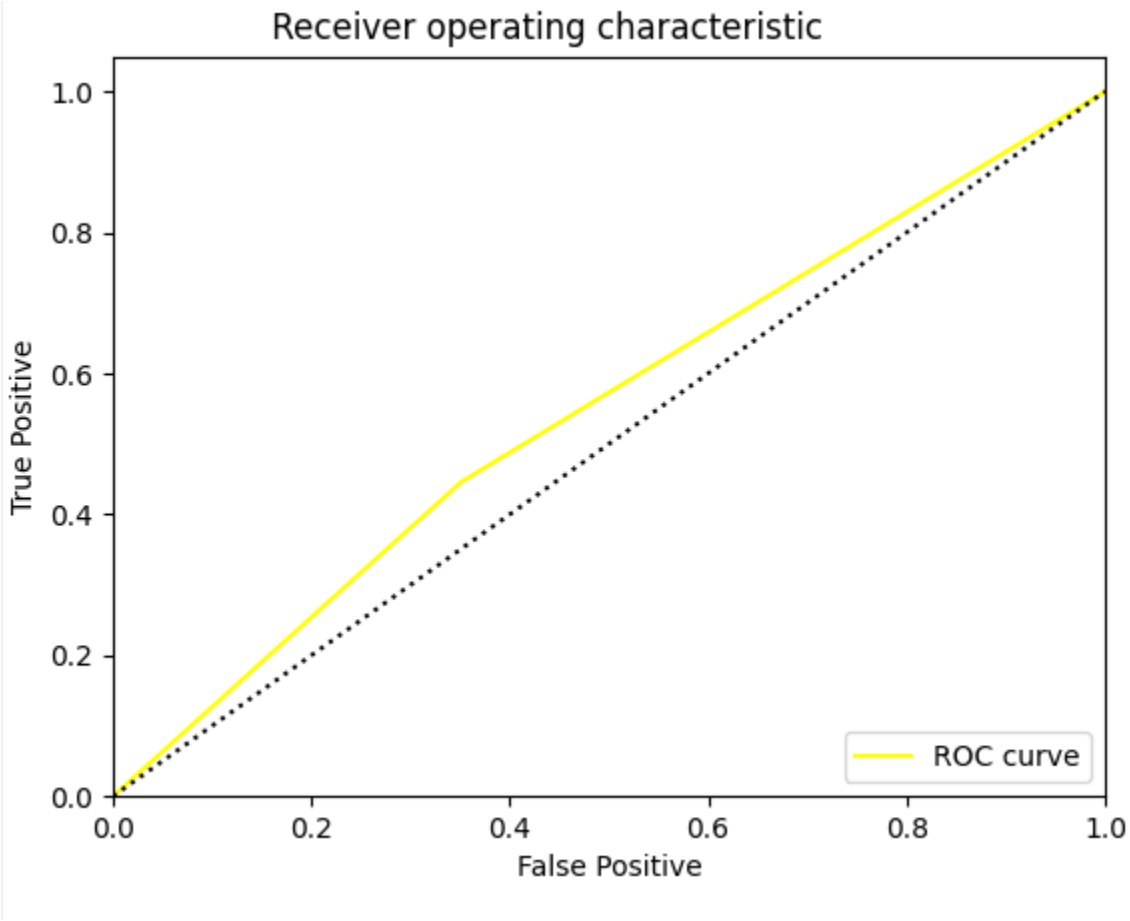
Random forest also has a better accuracy with 0.53 and AUC around 0.5. Still confusion matrix also proves there are significant number of false positives and false negatives

4. **Support Vector Machine**

For the support vector machine we used the Radial basis function as the kernel to get the maximum accuracy. We can see the SVN has a slightly better analysis the previous ones

accuracy			0.55	20000
macro avg	0.55	0.55	0.54	20000
weighted avg	0.55	0.55	0.54	20000

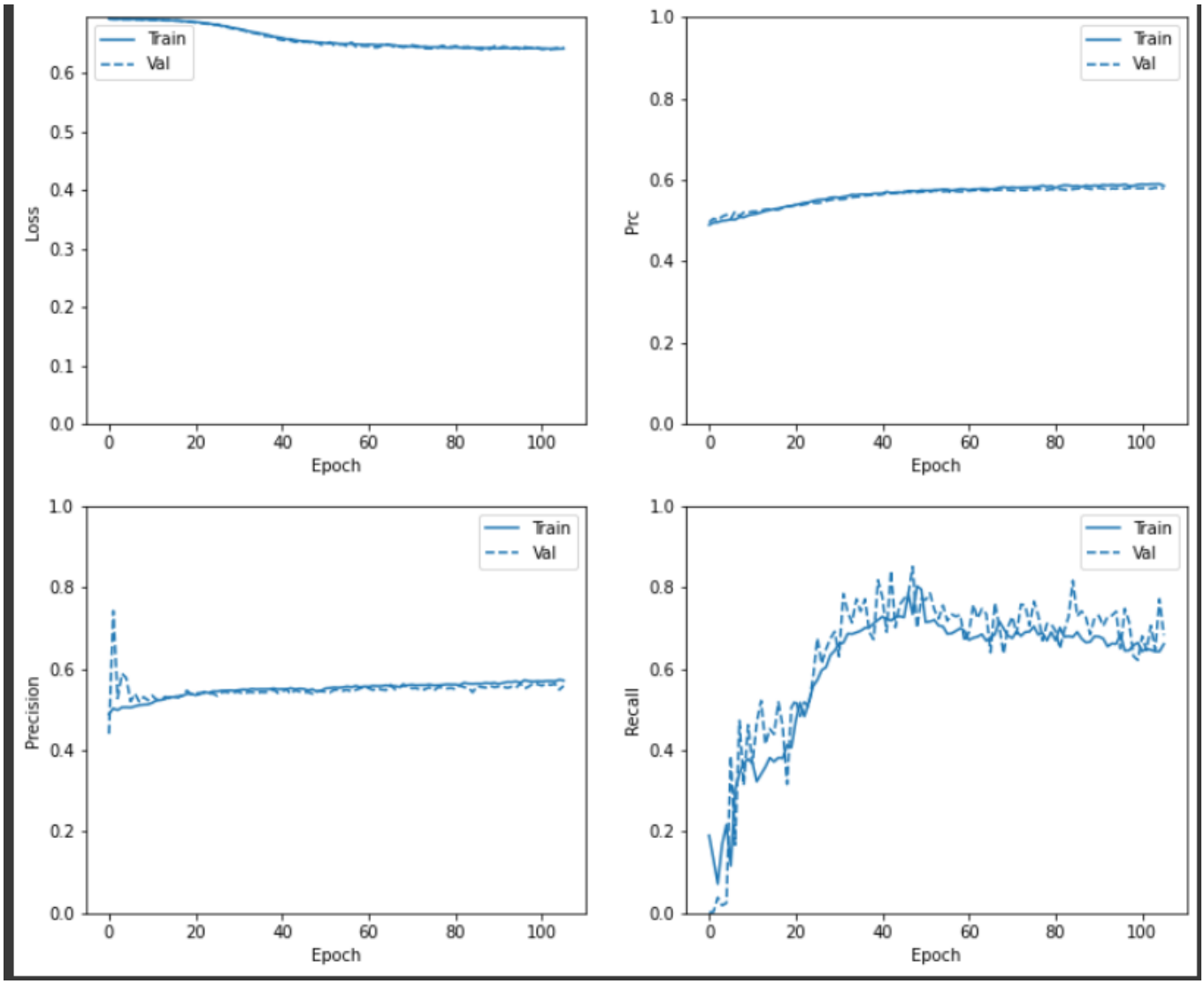
Accuracy : 0.5475
confusion matrix:
[[6490 3509]
[5541 4460]]

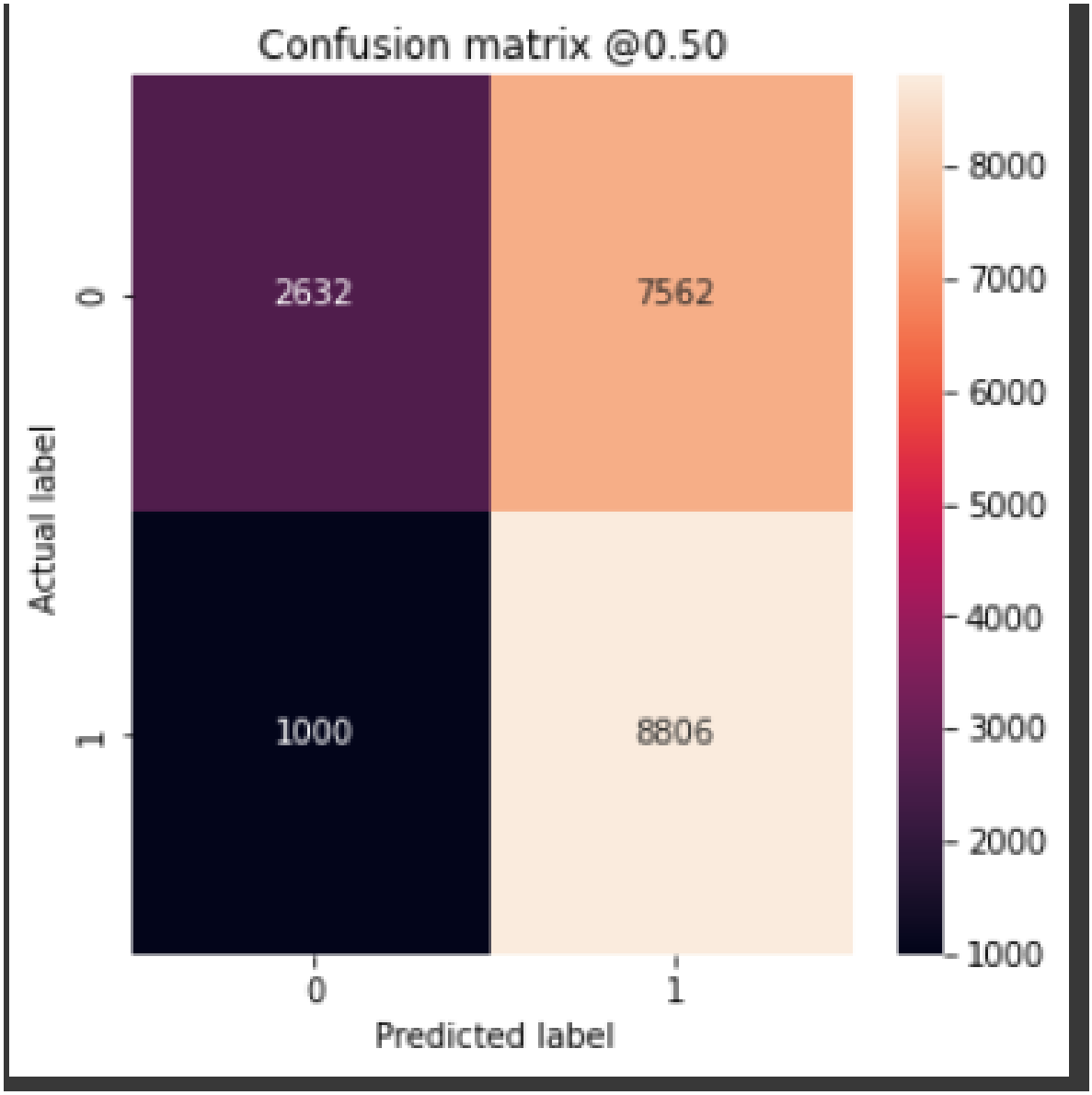


Accuracy if higher than previous ones with 0.55 indicating a better model than previous ones .

5. Convolutional Neural Network

loss : 0.6439611911773682
tp : 6340.0
fp : 4861.0
tn : 5253.0
fn : 3546.0
accuracy : 0.5796499848365784
precision : 0.5660209059715271
recall : 0.6413109302520752
auc : 0.6209748983383179
prc : 0.5785715579986572

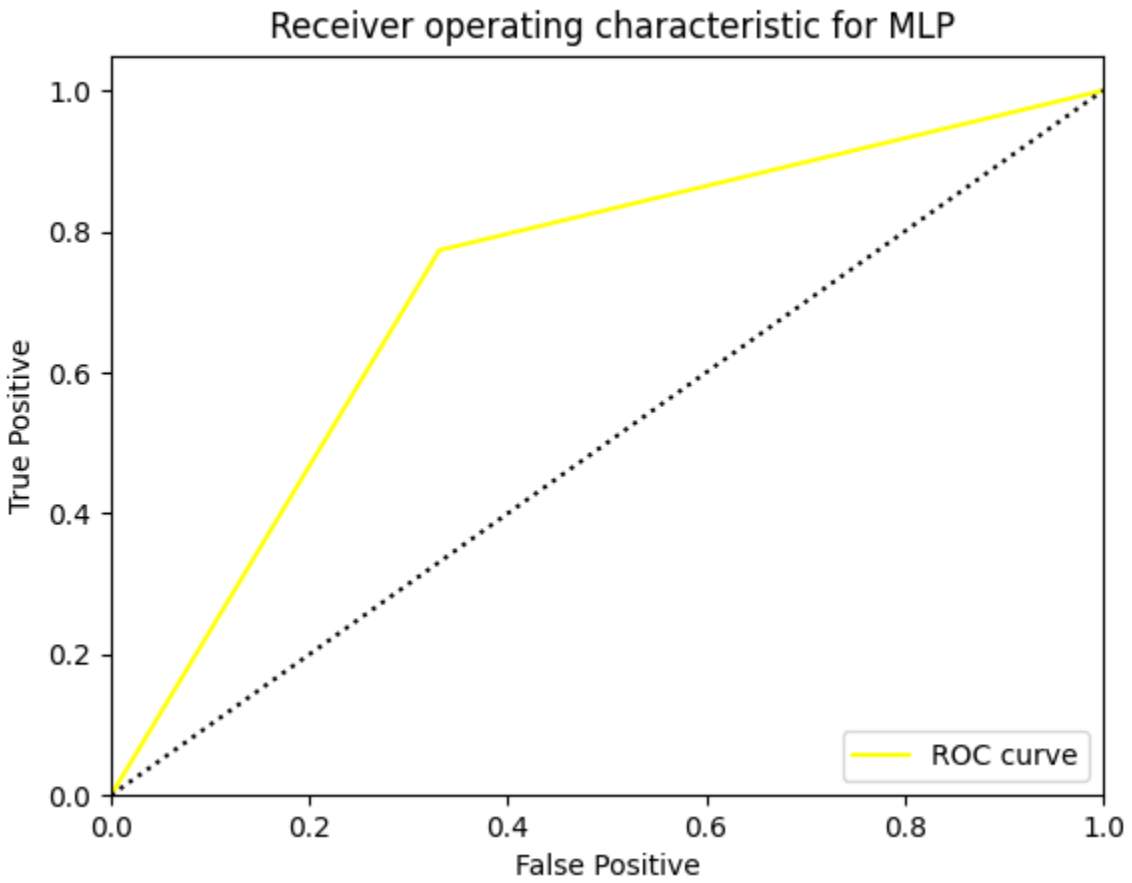




Best among so far of all algorithms For CNN we got accuracy of 0.58 and it has an AUC 0.62 indicating a fairly ok model.

6. Multi Layer Perceptron

Accuracy: 0.72115
confusion matrix:
[[6689 3310]
[2267 7734]]



As 0.5 < AUC < 1, there is a high chance compared to other classifiers that the classifier will be able to distinguish the positive class values from the negative class values.

Conclusion

We have analysed six algorithms where 5 are supervised and 1 is an unsupervised algorithm. Most of them gave average performance like Logistics Regression and Random Forest. Among the best with highest accuracy was for Multi Layer Perceptron with 0.72 followed by cnn with .57 Thus I would Multi Layer Perceptron Model as the best among the tested algorithms