SML ASSIGNMENT -1

NAME: SWAGATAM CHAKRABORTI

ROLL: MT18146

1. Fashion Dataset Classification

1.

Dataset: FMNIST dataset which has 60K training data points and 10K testing

Assumptions: Taking the trousers class as the positive class for the analysis purpose.

Methodology:

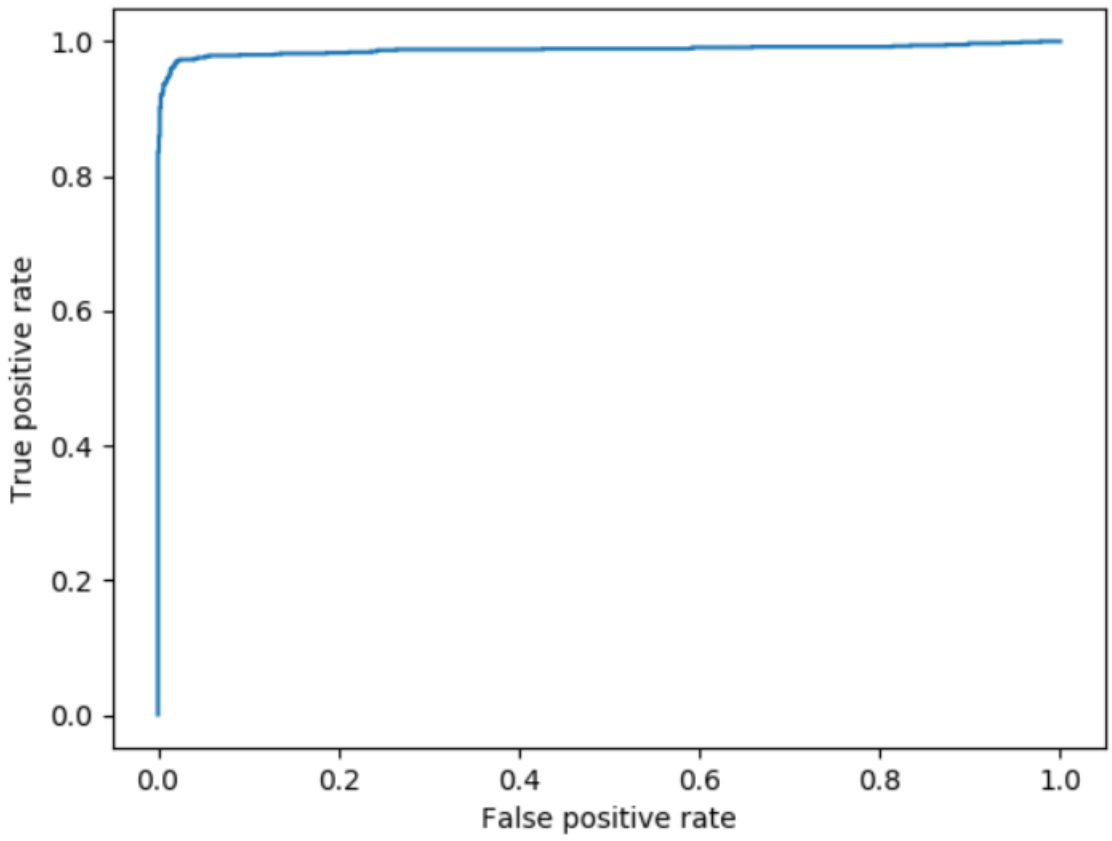
a. Binarize the images based on the threshold set at 80 for the pixel values
b. Each pixel values of a image represents a feature
c. Calculate the mean and variance for the trouser class and pullover class from the training dataset
d. For the testing dataset calculate the likelihood using the gaussian normal density formula
e. Using naïve bayes classify the images are classified

Results:

accuracy	97.3
precision	98.86363636363636
recall	95.7

confusion matrix:

	Predicted (Trouser class)	Predicted (Pullover Class)
Actual (Trouser class)	957	11
Actual (Pullover class)	43	989



ROC Curve: taking trouser as the positive class

2. Classification considering all the 10 classes in the fashion dataset

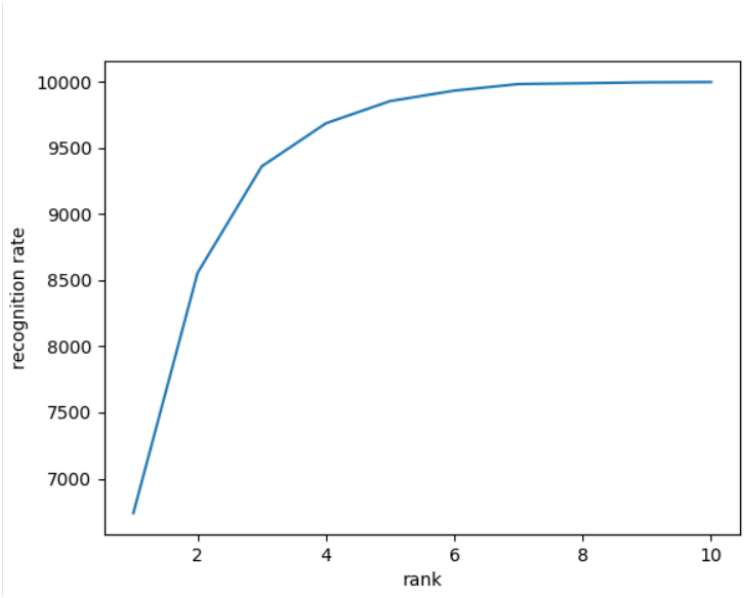
Results:

Confusion Matrix

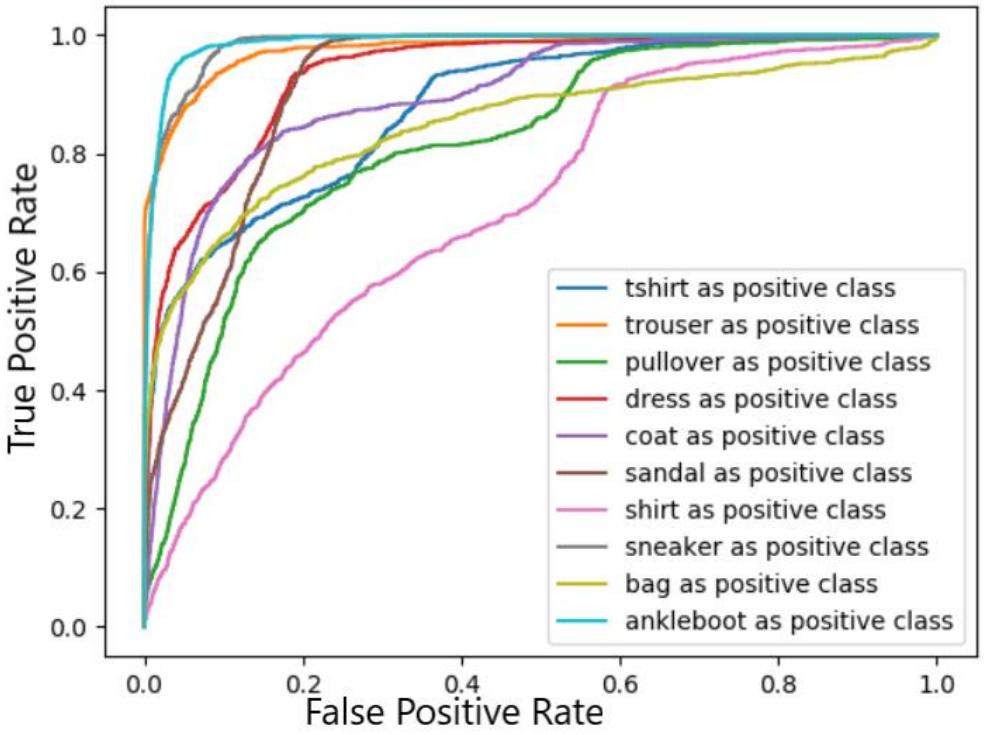
(predicted vs actual)	tshirt	trouser	pullover	dress	coat	sandal	shirt	sneaker	bag	ankleboot
tshirt	732	2	24	144	13	16	38	3	28	0
trouser	16	892	11	66	6	0	7	0	2	0
pullover	12	1	445	29	333	1	153	1	25	0
dress	25	91	5	811	30	0	30	2	6	0
coat	2	9	110	122	670	4	71	1	11	0
sandal	0	0	1	0	0	503	14	458	12	12
shirt	204	5	118	113	307	12	176	1	64	0
sneaker	0	0	0	0	0	27	0	935	1	37
bag	1	0	11	70	8	5	123	4	778	0
ankleboot	0	0	0	1	0	48	13	133	8	797

Positive class	Precision	recall
tshirt	73.79032258064516	73.2
trouser	89.2	89.2
pullover	61.37931034482759	44.5
dress	59.80825958702065	81.1
coat	49.012435991221653	67
sandal	81.65584415584416	50.3
shirt	28.16	17.6
sneaker	60.79323797139142	93.5
bag	83.20855614973262	77.8
ankleboot	94.2080378250591	79.7

Accuracy: 67.3



CMC Curve



ROC Curve comparison of taking each class as the positive

2. MNIST Dataset Classification

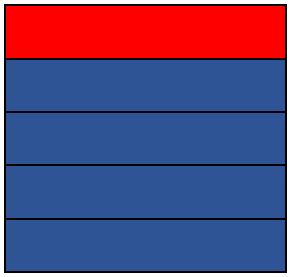
1. Taking only 1 and 8 in the dataset

RANDOM 5 FOLD MODELS:

TRAIN DATASET

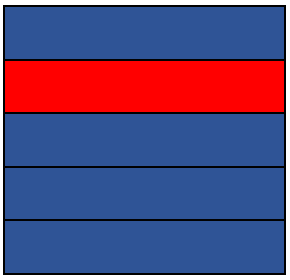
TEST DATASET

MODEL: 1



precision: 92.755

MODEL: 2

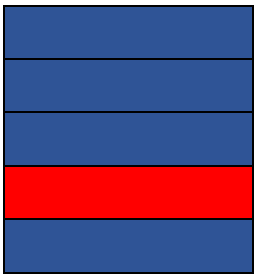


precision: 92.48

MODEL: 3

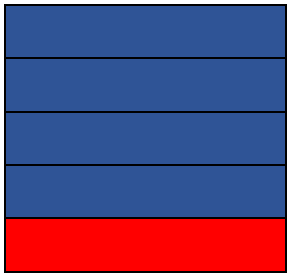
precision: 93.29

MODEL: 4



precision: 92.18

MODEL: 5



Precision: 93.74

MEAN: 92.87

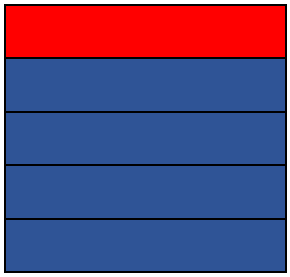
S.D : 0.005

STRATIFIED 5 FOLD MODELS:

TRAIN DATASET

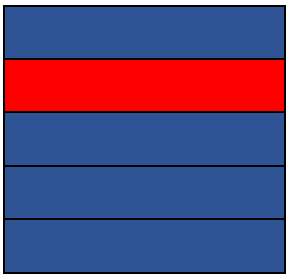
TEST DATASET

MODEL: 1



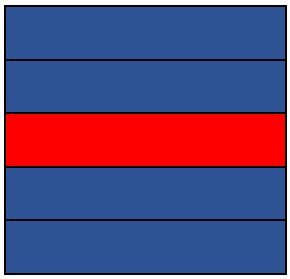
precision: 92.687

MODEL: 2



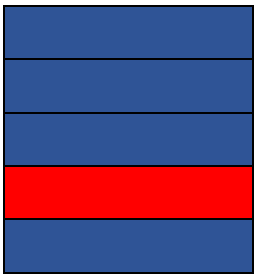
precision: 92.721

MODEL: 3



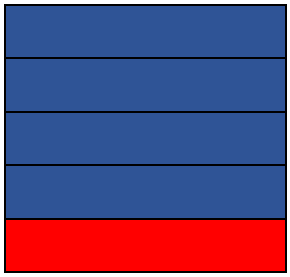
precision: 93.29

MODEL: 4



precision: 92.07

MODEL: 5



Precision: 92.80

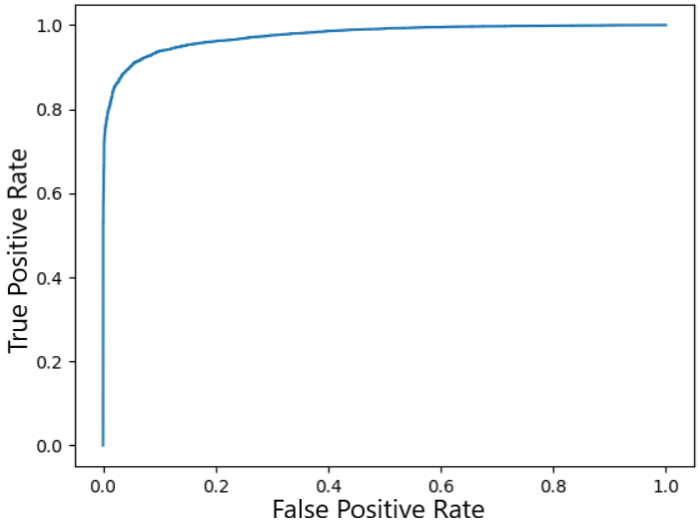
MEAN: 92.71

S.D : 0.003

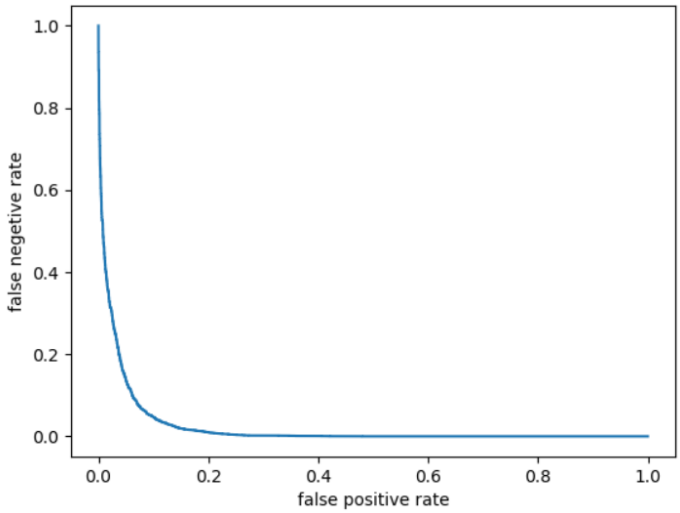
As per above results of the different models, random 5 fold model 5 having the best accuracy is chosen to be the best model.

Results:

a. On testing on the initial training set.



ROC Curve on initial train dataset



DET Curve tested on training data

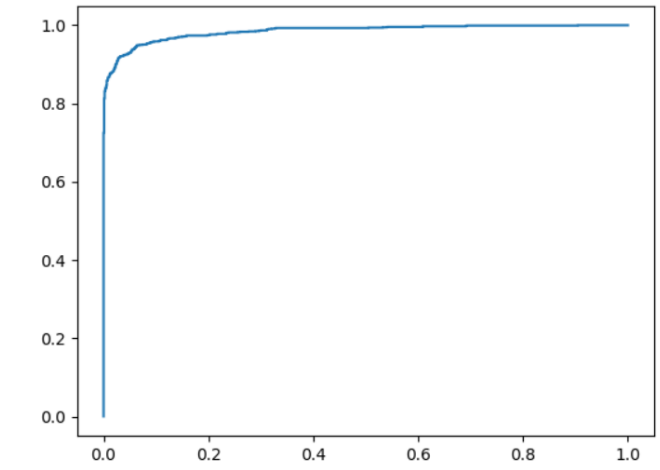
Confusion matrix

	Predicted 1's	Predicted 8's
Actual 1's	6431	663
Actual 8's	311	5188

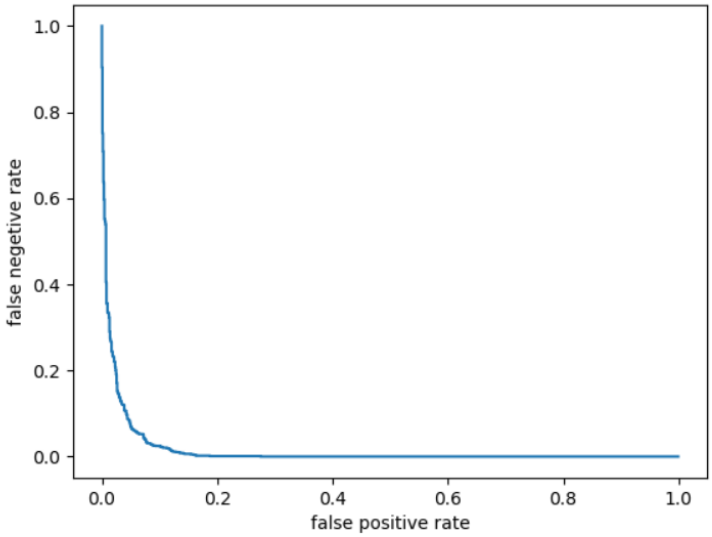
EQUAL ERROR RATE: 0.074

Accuracy: 92.26

b. On testing on initial testing data set



ROC Curve on initial testing dataset



DET Curve on initial test data

Confusion matrix

	Predicted 1's	Predicted 8's
Actual 1's	1087	64
Actual 8's	48	910

Equal Error Rate: 0.058

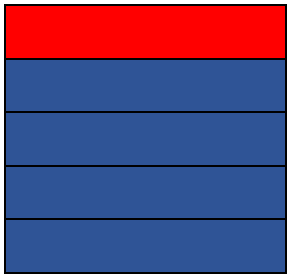
Accuracy: 94.62

1. Taking only 3 and 8 in the dataset
RANDOM 5 FOLD MODELS:

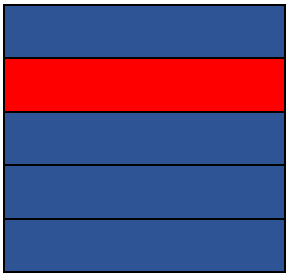
TRAIN DATASET

TEST DATASET

MODEL: 1



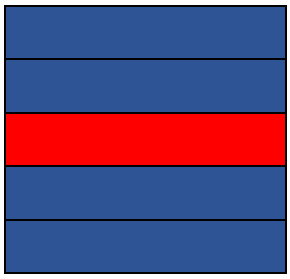
MODEL: 2



precision: 92.687

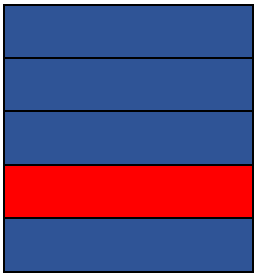
precision: 92.721

MODEL: 3



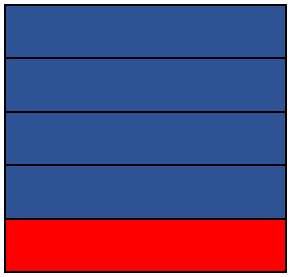
precision: 93.29

MODEL: 4



precision: 92.07

MODEL: 5



Precision: 92.80

MEAN: 92.713

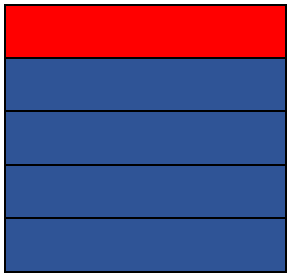
S.D : 0.003

STRATIFIED 5 FOLD MODELS:

TRAIN DATASET

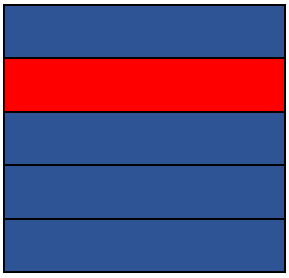
TEST DATASET

MODEL: 1



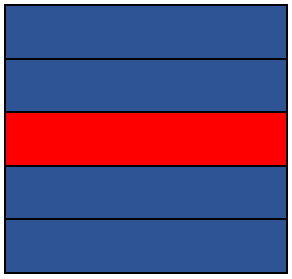
precision: 91.407

MODEL: 2



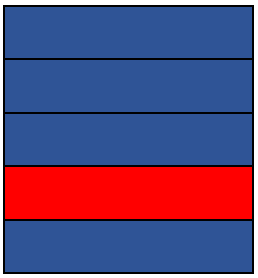
precision: 90.798

MODEL: 3



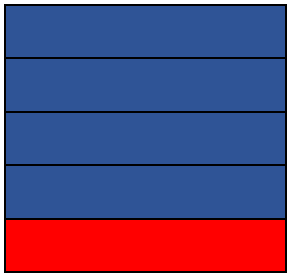
precision: 90.79

MODEL: 4



precision: 90.79

MODEL: 5

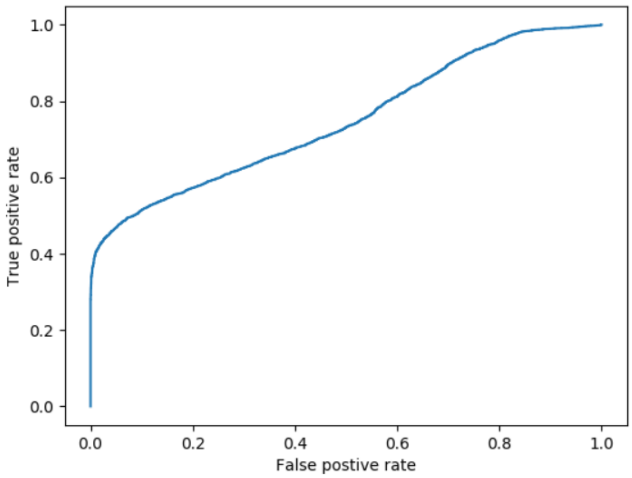


Precision: 93.53 MEAN: 91.26 S.D : 0.006

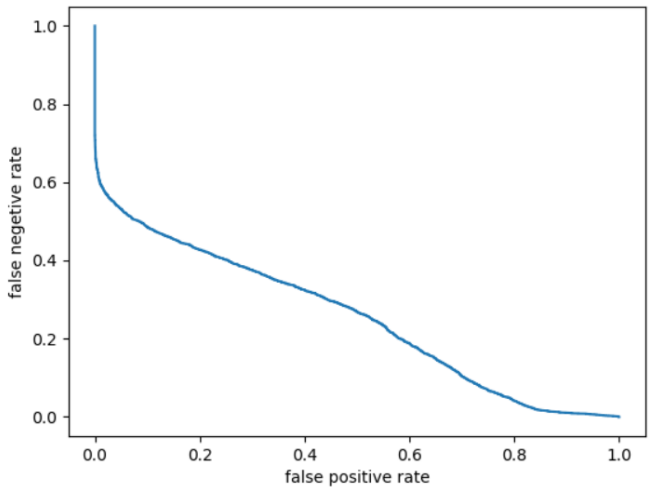
From the precision values stratified 5 fold Model 5 shows the best accuracy. So it is chosen as the best model.

Results:

a. On testing on the initial training set.



ROC Curve on testing over initial training data



DET Curve on testing over initial training data

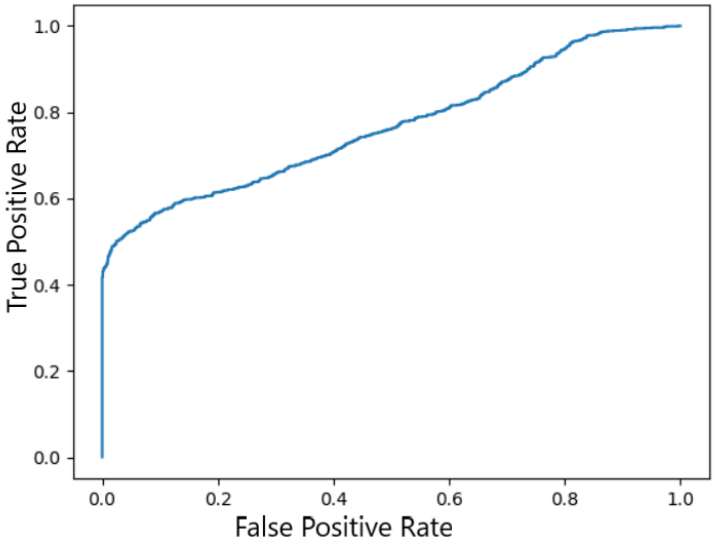
Confusion matrix

	Predicted 3's	Predicted 8's
Actual 3's	1087	64
Actual 8's	48	910

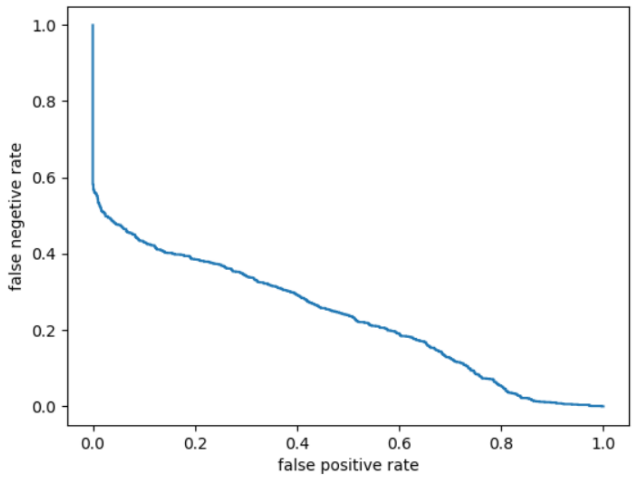
Equal Error Rate = 0.34

Accuracy = 91.64

b. On testing on the initial test dataset



ROC Curve on testing over initial test dataset



DET Curve tested on initial test dataset

Confusion matrix

	Predicted 1's	Predicted 8's
Actual 1's	933	85
Actual 8's	77	889

Equal Error Rate : 0.32 Accuracy: 91.8

OBSERVATION: As the two symbols are much similar in shape so there has been a decrease in the accuracy rate and there is distortion in the ROC and the DET curve