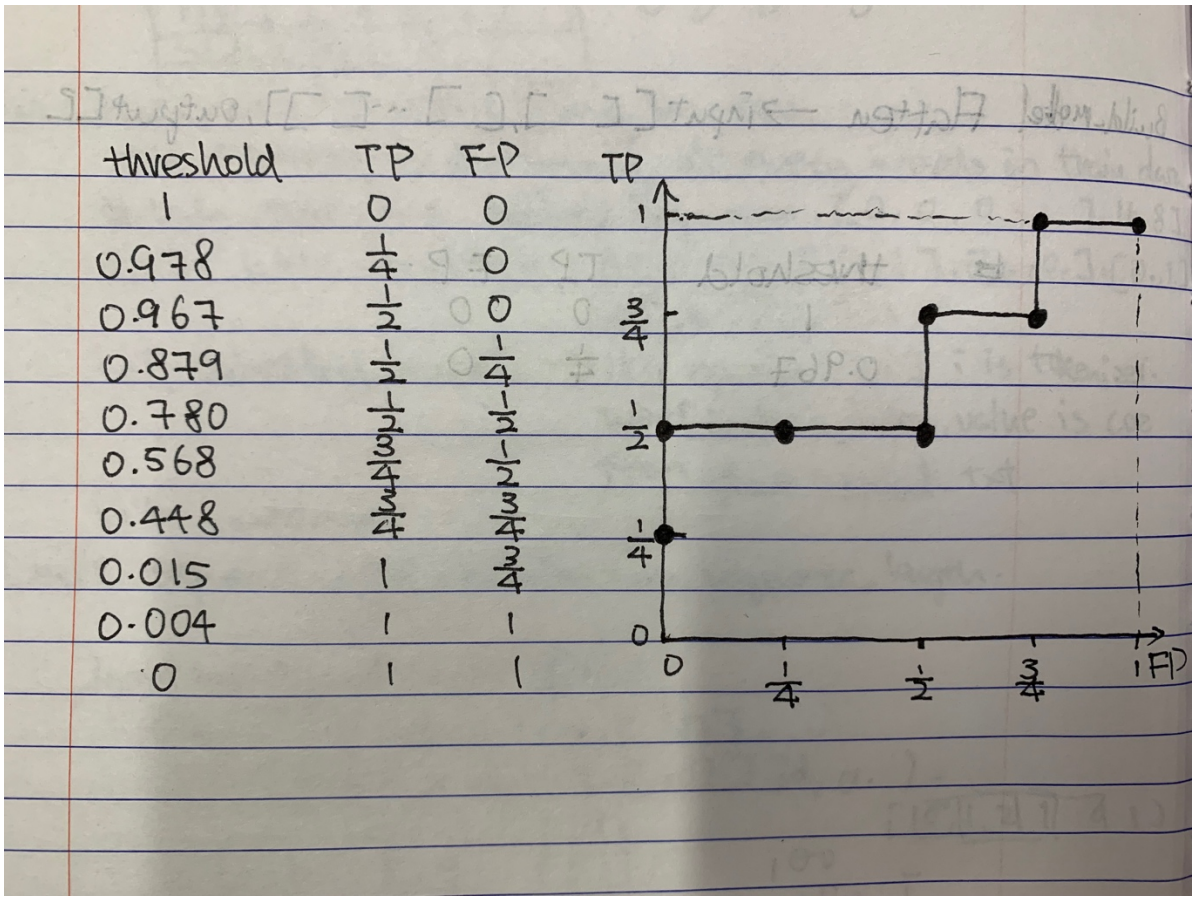
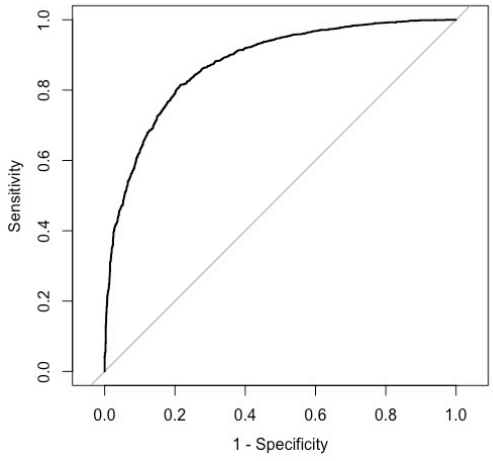
Part 1:



Part 2:

Task 1:

```
Data = read.csv("EE627A_HW3_DataSet1.csv", header = F)

sapply(Data, sd)
logit <- glm(V477~., data = Data, family = "binomial")
prob <- predict(logit, type = "response")
Data$prob = prob
g <- roc(V477 ~ prob, data = Data)
plot(g, legacy.axes = 1)
auc(g)
```

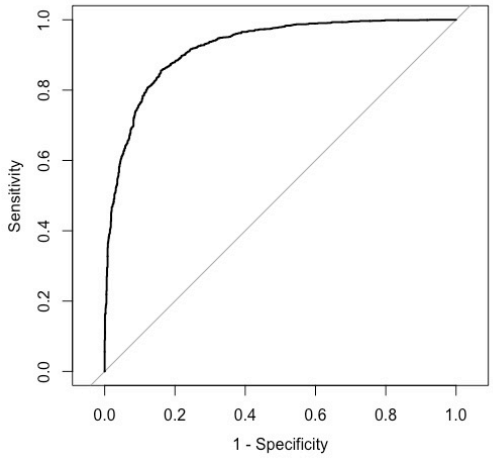


And area under the curve(AUC): 0.8748

Task2:

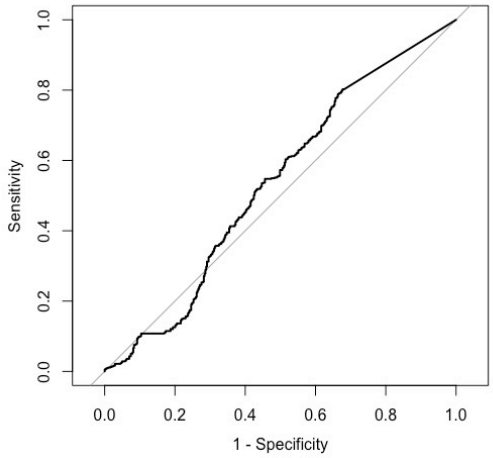
```
Data2 <- Data[1:3000,]
logit2 <- glm(V477~., Data2, family = "binomial")
prob2 <- predict(logit2, type = "response")
Data2$prob = prob2
g1 <- roc(V477 ~ prob, data = Data2)
plot(g1, legacy.axes = 1)
auc(g1)

Data3 <- Data[3001:4000,]
prob3 <- predict(logit2, Data3, type='response')
Data3$prob = prob3
g2 <- roc(V477 ~ prob, data = Data3)
plot(g2, legacy.axes = 1)
auc(g2)
```

Use training set data (first 3000 data) to apply the logistic regression:



Area under the curve: 0.9214

Apply the logistic regression coefficients in training set data to validation set (rest 1000 data)



Area under the curve: 0.5361

As first AUC shows, use training data itself to test result, we get very high value, which make sense because it is its own training data. For the second one, use non-training data to validate the model, and only get 0.5361, which is not good performer. Hence, this model should be improved or the first 3000 data and rest 1000 data are not too related.