**IDS 572 Business Data Mining**

**Fall 2016**

**Assignment #4**

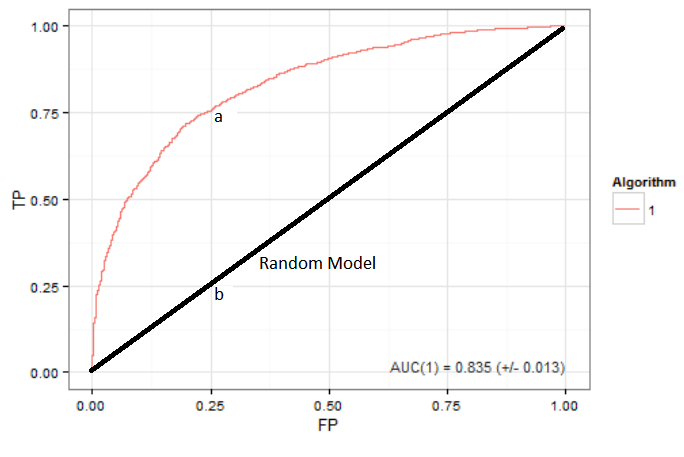
**Team members:**

**Kritika Chauhan**

**Eun Jung Ko**

**Priyanka Pathak**

Problem 1. Take a look at three examples below and answer the following questions:



(a) What can you say about this ROC curve? How this classifier differs from a random guess? Pick one point on a curve and interpret it using examples and illustrations. For example, this point represents a classifier that can detect x% of all patients, who have a disease, but y% those who have not, are classified incorrectly....

AUC(perfect prediction model)=1

AUC(1)= 0.835

AUC(random model) = 0.5

The area under the curve for this classifier is closer to 1 so it’s a good model which gives a prediction closer to the perfect prediction model. As compared to a random guess this classifier has a higher Recall and a lower False-alarm rate translating as higher AUC of 0.835 than AUC of 0.5 for a random model.

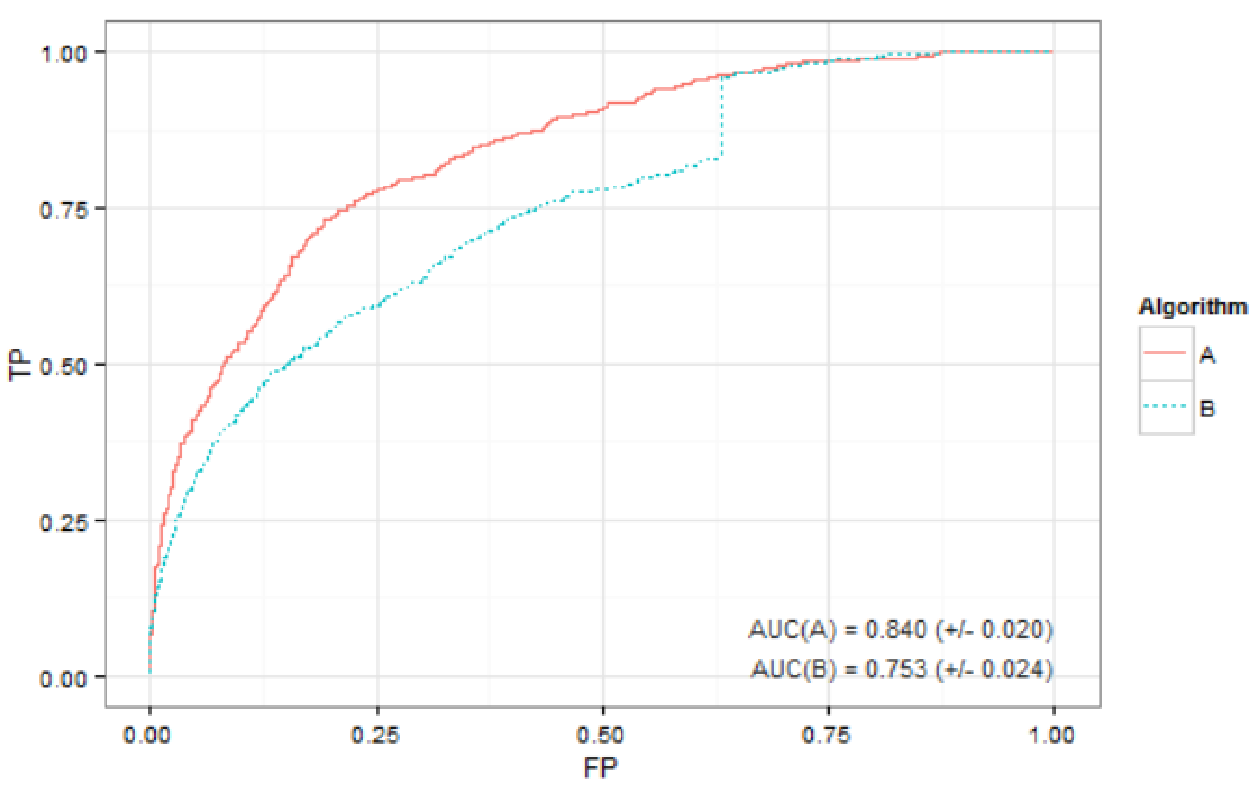
For example, as per our classifier at point (a): -

False Positive as 0.25 represents patients who have do not have a disease but 25% of the times they are predicted having a disease.

The Recall or True Positive Rate as 0.75 represents the patients who have a disease and 75% of the times are predicted as having a disease.

For a random classifier at point (b): -

False Positive and the Recall both would be at 0.25 indicating the classifier cannot distinguish between patients having the disease and patients not having the disease.



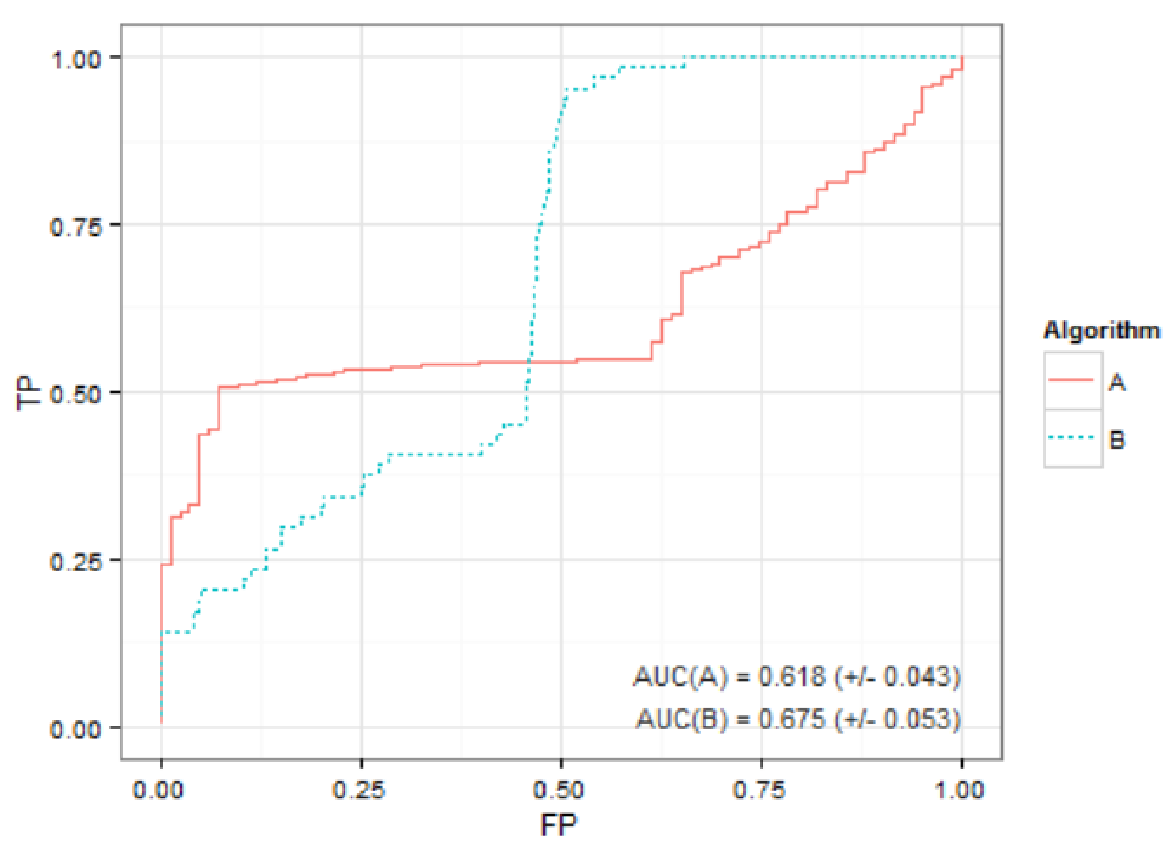
(b) Compare two ROC curves above. Which one is a better model and why?

AUC(A) = 0.840

AUC(B) = 0.753

The Recall of algorithm A is higher than that of algorithm B for False Positive rate less than 67% and for False Positive rate of more than 67% the recall rate of A almost coincides with than of B.

Moreover, the Area Under the Curve for A(0.840) has a higher value than the Area Under the Curve for B making it a better model than B(0.753).



(c) Compare two ROC curves above. When algorithm A would be preferred over algorithm B?

AUC(A)=0.618

AUC(B)=0.675

For False Positive rate, less than 49% the Recall for algorithm B is lower than that of Algorithm A making it a preferable choice than Algorithm B for same False Positive rate values.

For False Positive rate, more than 49% the Recall for algorithm A is lower than that of Algorithm B making it a preferable choice than algorithm A for same False Positive rate values.

Considering both the scenarios when we compare the Area Under the Curve values for both Algorithm A and Algorithm B we find it higher for B. Hence, as per AUC values algorithm B would be preferred over algorithm A.

Problem2.

Calculate confusion matrix, precision and recall under threshold of 0.5

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted | | |
| Actual |  | 1 | 0 |
| 1 | 5 ( True Positive) | 0 (False Negative) |
| 0 | 2 (False Positive) | 3 (True Negative) |

Precision: P(Actual = + | Predicted = +) = TP / (TP+FP) = 5 / (5+2) = 5/7 = 0.7143

Recall: P(Predicted = + | Actual = +) = TP / (TP+FN) = 5 / (5+0) = 5/5 o = 1

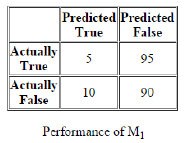
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| True Class | Prediction | T = 1 | T= 0.8 | T= 0.7 | T= 0.6 | T = 0.5 | T = 0.4 | T = 0.2 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 0.9 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 0.8 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 0.7 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| 1 | 0.6 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| 1 | 0.6 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| 0 | 0.6 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| 0 | 0.4 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 0 | 0.4 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 0 | 0.2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
|  | Recall | 0.2 | 0.6 | 0.6 | 1 | 1 | 1 | 1 |
|  | False-Alarm | 0 | 0 | 0.2 | 0.4 | 0.4 | 0.8 | 1 |

Draw a ROC curve

Problem3.

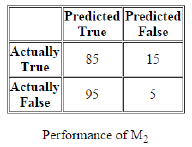
1. Precision= True Positive/(True Positive+ False Positive)

For M1:



Precision = 5/(5+10)= 5/15= 0.3333 i.e. 33.33%

For M2:



Precision = 85/(85+95) = 85/180 = 0.4722 i.e. 47.22%

We will buy Model M2 because the precision of M2 is 47.22% which is higher than M1’s precision that is 33.33%.

b) Let the cost of labeling as True something that is actually False (FP)=x

and the cost of labeling as False something that is actually True (FN)=y

Given that x>y.

So for M1:

Cost of False Positive= 10x

Cost of False Negative= 95y

and for M2:

Cost of False Positive= 95x

Cost of False Negative= 15y

We will buy Model M1 because the cost of False Positive of M1 is 85x which less than the cost of M2. Although, the cost of False Negative of M1 is 80y which is higher than the cost of M2. But 85x is much larger than 80y as it is given that x is far exceed than y.