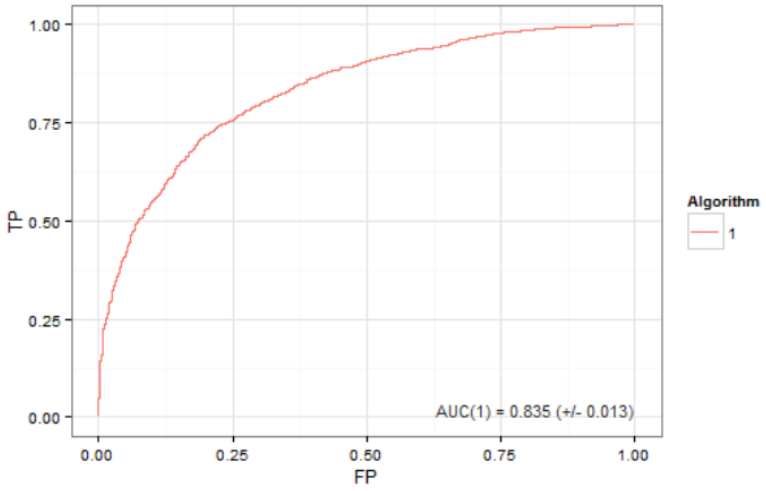
BUSINESS DATA MINING (IDS 572)

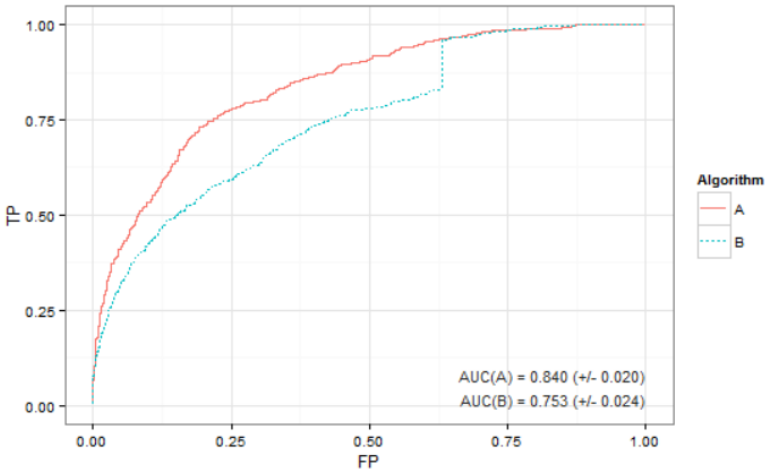
HOMEWORK 4
DUE DATE: WEDNESDAY, SEPTEMBER 28 AT 03:00 PM

- Please provide succinct answers to the questions below.
- You should submit an electronic pdf or word file in blackboard.
- Please include the names of all team-members in your write up and in the name of the file.

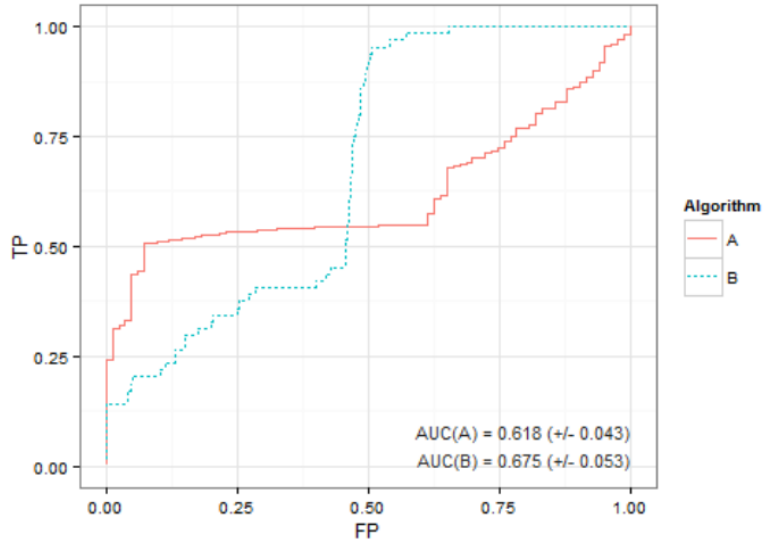
Problem 1. Take a look at three examples below and answer the following questions:



(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....



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



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

Problem 2. Calculate (on paper) confusion matrix, precision and recall for the given dataset under threshold of 0.5.

	True class	Prediction
1.	1	0.6
2.	1	0.8
3.	0	0.4
4.	1	0.9
5.	0	0.7
6.	1	0.6
7.	1	1.0
8.	0	0.2
9.	0	0.4
10.	0	0.6

Draw a ROC curve.

Problem 3. Assume that two individuals offer to sell you their predictive models M1 and M2. The confusion matrices produced by each model are as follows.

	Predicted True	**Predicted False**
Actually True	5	95
Actually False	10	90

Performance of M_1

	Predicted True	**Predicted False**
Actually True	85	15
Actually False	95	5

Performance of M_2

(a) Assuming that precision is of paramount importance in your application, which of the two models would you buy? Why?

(b) Assuming that the cost of labeling as True something that is actually False far exceeds the cost of labeling as False something that is actually True, which of the two models would you buy? Why?