Business Data Mining (IDS 572)

Homework 4-Solution

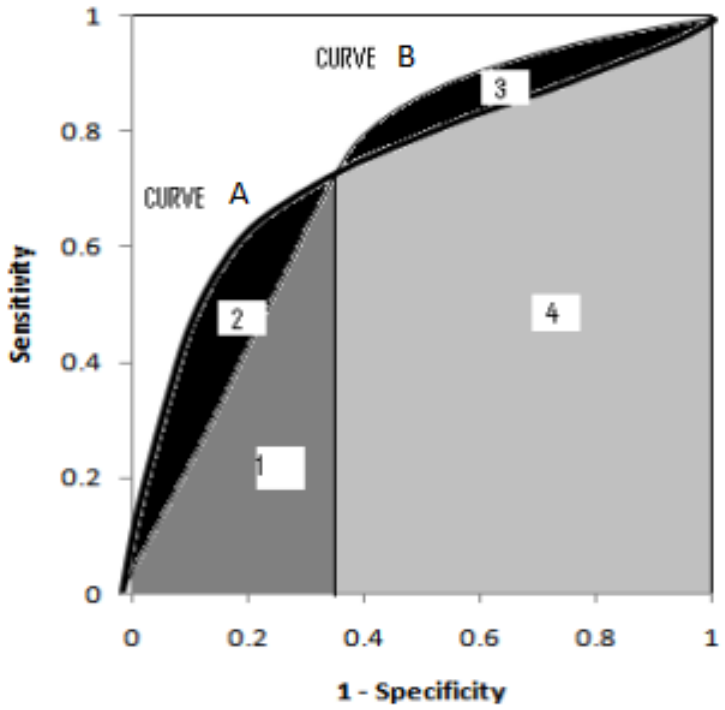
Question 1

(a) ROC curve illustrates the performance of a binary classifier. The curve is created by plotting the True Positive Rate (Recall or Sensitivity) against the False Positive Rate (False-Alarm or 1- Specificity). Each prediction result or instance of a confusion matrix represents one point in the ROC curve. The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). The (0,1) point is also called a perfect classification. A completely random guess would give a point along a diagonal line from the left bottom to the top right corners. The diagonal divides the ROC space. Points above the diagonal represent good classification results (better than random), points below the line poor results (worse than random).
Let assume this model represents prediction of disease among a group of people. Consider a point (0.25, 0.75) in the curve. This point indicates that 75% of the all the people who have the disease are predicted correctly and 25% of the people who do not have the disease are classified as patients with the disease.

(b) Total area under ROC curve (AUC) is a single index for measuring the performance a model. The larger the AUC, the better is overall performance of the model. Therefore, **Model A** with AUC = 0.840 is a better model than Model B with AUC = 0.753.

(c) In this graph we have two ROC curves crossing each other with nearly same area. Suppose these are two hypothetical ROC curves of two medical tests A and B applied on the same subjects to assess the same disease. Test A and B have nearly equal area but cross each other. Test B performs better than test A where high sensitivity is required, and test A performed better than B when high specificity is needed.

In such cases and in some other situations, the interest may be restricted to specific values of sensitivity or specificity. You may be interested in a test with high specificity as for a disease with grave prognosis (cancer). Then the interest will be in test A and that too for specificity ≥ 0.65 or (1 – specificity) < 0.35. In that case, the area of interest is 1+2 as shown in the figure below. This is called partial area under the curve. Some soft-wares calculates this area too.



Source: http://www.medicalbiostatistics.com/roccurve.pdf

Question 2

	True Class	Sorted Scores
7	1	1.0
4	1	0.9
2	1	0.8
5	0	0.7
1	1	0.6
6	1	0.6
10	0	0.6
3	0	0.4
9	0	0.4
8	0	0.2

Threshold value is given as 0.5

- If the score of prediction is greater than the threshold, the predicted class is '1'.
- If the score of prediction is less than the threshold, the predicted class is '0'.

	True Class	Sorted Scores	Predicted class
7	1	1.0	1
4	1	0.9	1
2	1	0.8	1
5	0	0.7	1
1	1	0.6	1
6	1	0.6	1
10	0	0.6	1
3	0	0.4	0
9	0	0.4	0
8	0	0.2	0

From the above table we obtain the following confusion matrix.

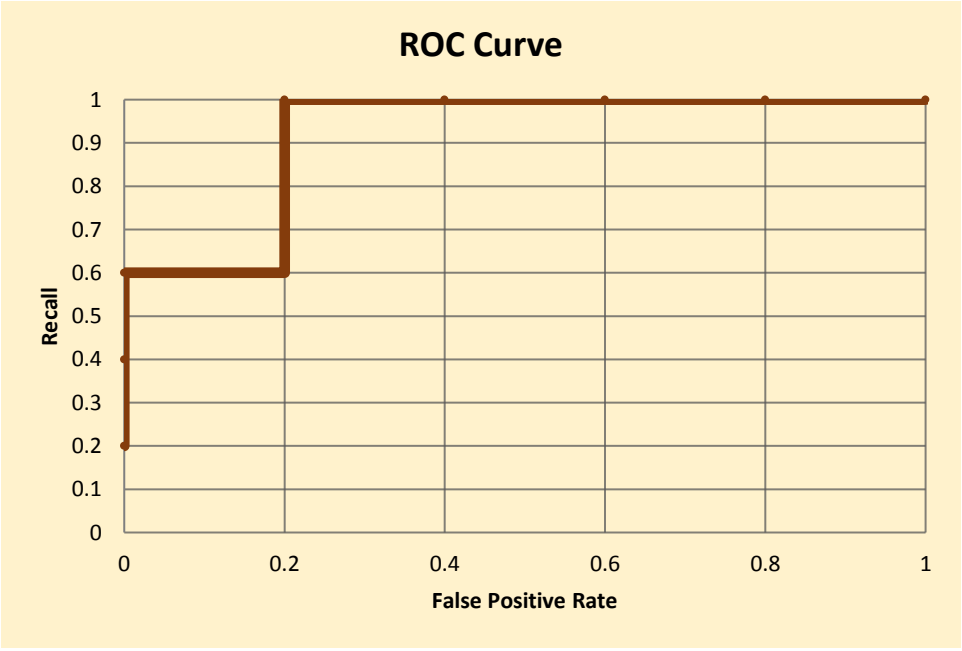
	Predicted	
Actual	1	0
1	5 (TP)	0 (FN)
0	2 (FP)	3 (TN)

Recall: TP/ (TP+FN) = 5/5 = 1

Precision: TP/ (TP+FP) = 5/7 = 0.71

To draw the ROC curve, we can consider the following table:

	True Class	Sorted Scores	TPR	FPR
7	1	1.0	1	0
4	1	0.9	2	0
2	1	0.8	3	0
5	0	0.7	3	1
1	1	0.6	4	1
6	1	0.6	5	1
10	0	0.6	5	2
3	0	0.4	5	3
9	0	0.4	5	4
8	0	0.2	5	5



Question 3

(a) Precision of model M1: 5/ (5+10) = 5/15 = 0.33
Precision of model M2: 85/ (85+95) = 85/180 = 0.47
We choose **Model 2** because it has a higher precision value compared to Model 1.

(b) Cost of labeling as True of an instance that is actually False (False Positive - FP) is much higher than the cost of labeling as False of an instance that is actually True (False Negative – FN). We have
Model 1: FP = 10, FN = 95
Model 2: FP = 95, FN = 15
Since the number of FP in Model 2 is significantly higher than the number of FP in Model 1, then **Model 1** is better for the prediction.