Spring, 2020: 90-760 Management Science II: Homework #2. Due Tuesday March 31st at noon

Problem #1:

a) Fill in the boxes below the following confusion matrix. Report the TPR, FPR, precision, recall, and specificity as fractions. Note: This matrix is oriented the way Ragsdale typically does, and how I did in class.

Assume the condition being tested for is group #2.

Classification Table: Cost-Weighted Cut-Off

	Predict 1	Predict 2	Total
Actual 1	40	20	60
Actual 2	10	80	90
Total	50	100	150

80	**True Positive**
10	**False Negative**
40	**True Negative**
20	**False Positive**

4/5	**True Positive Rate (TPR)**
2/5	**False Positive Rate (FPR)**

30	**# of Errors**

4/5	**Sensitivity**
4/5	**Specificity**

b) Fill in the boxes below the following confusion matrix. Report the TPR, FPR, precision, recall, and specificity as fractions. Note: This matrix is ***NOT*** oriented the way Ragsdale typically does, and how I did in class, including that it is group #1 not group #2 that has the condition.

Assume the condition being tested for is GROUP #1.

Classification Table: Cost-Weighted Cut-Off

	Actual 1	Actual 2	Total
Predict 1	40	20	60
Predict 2	10	80	90
Total	50	100	150

40	**True Positive**
10	**False Negative**
80	**True Negative**
20	**False Positive**

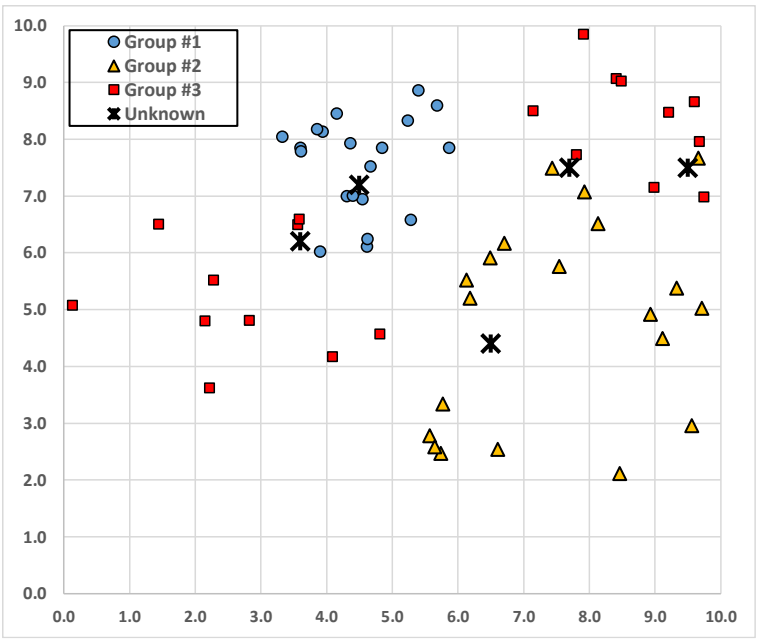
2/3	**True Positive Rate (TPR)**
2/9	**False Positive Rate (FPR)**

30	**# of Errors**

2/3	**Sensitivity**
8/9	**Specificity**

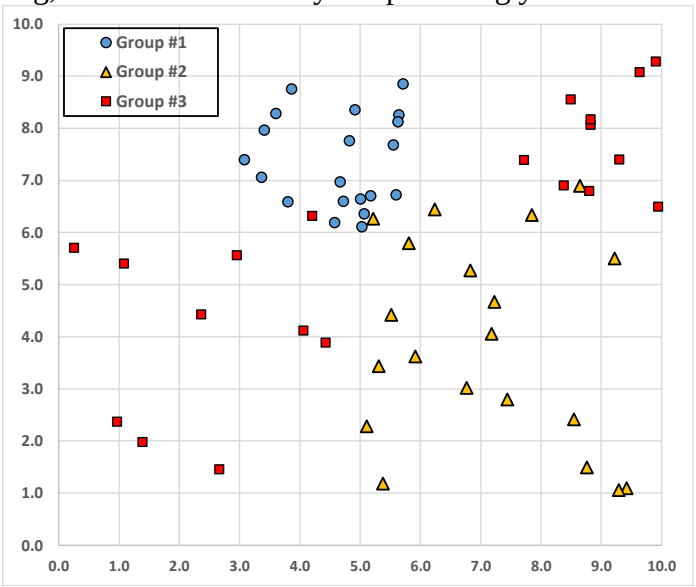
Problem #2

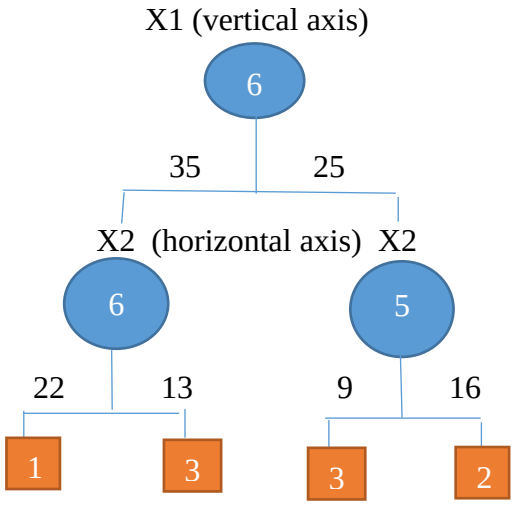
a) Use a 3-KNN rule to classify the five unknown observations. (Do this visually, not with software. It is a test of your conceptual understanding.)



Classify these new observations			
Obs	Group	X1	X2
1	3	3.6	6.2
2	1	4.5	7.2
3	2	7.7	7.5
4	3	9.5	7.5
5	2	6.5	4.4

b) Create the best possible classification tree for this problem. You are permitted two layers and all splits must be done on integer values of the variables. Again, I intend for you to do this visually, not with the software, as a test of conceptual understanding, so I am intentionally not providing you with the data file.





c) Create the confusion matrix based on your tree, with actual groups indicated by the row and predicted groups by the column. (Hint: There are 20 observations in each group.)

	Gp 1 – Predict	Gp 2- Predict	Gp 3 – Predict	Total actual
Gp 1 –Actual	20	0	0	20
Gp 2 – Actual	1	16	3	20
Gr 3 - Actual	1	0	19	20
Total predict	22	16	22	

Problem #3

This problem is a variation on Ragsdale's Problem 10.7 in which a university wants to use undergraduate GPA and standardized test scores (GMAT) to predict which students would do well (group #1) and which would do badly (group #2) if admitted.

a) Use the spreadsheet template provided to run a linear discriminant analysis by filling in the appropriate formulas in columns E and G. [Nothing to submit]

b) Report the centroids for each group (show two digits of precision beyond the decimal point).

1	3.19	597.97
2	2.69	597.49

c) Find the cut-off that would lead to 200 students being admitted if the actual applicant pool were identical to these 350 students in the training data. Report the cut-off (to two digits past the decimal) and the corresponding confusion matrix,

TPR, and FPR. If all 200 admitted students chose to attend, what proportion of the incoming class would be "good" (Group 1) students?

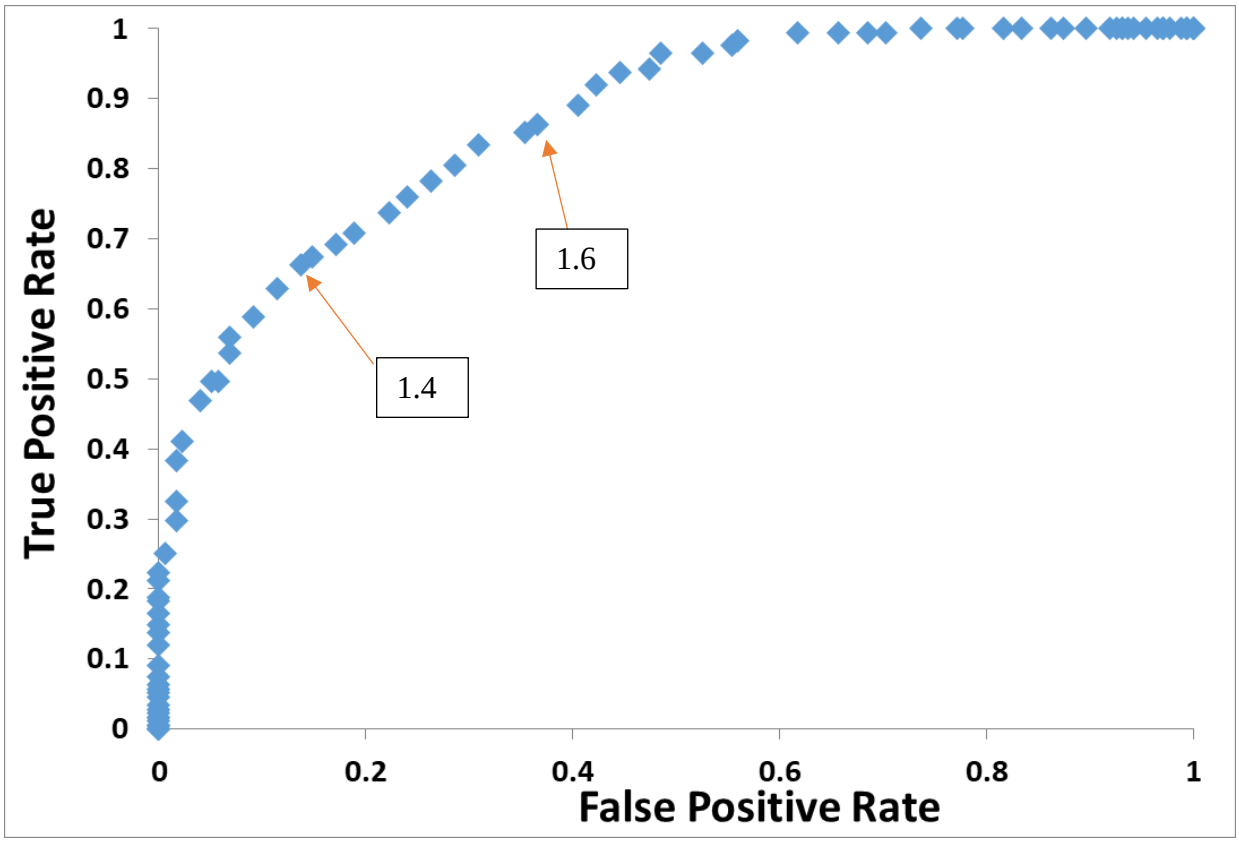
Cutoff: 1.56

	Predict 1	Predict 2	Total
Actual 1	146	29	175
Actual 2	54	121	175
Total	200	150	350

TPR: 0.8343
FPR: 0.3086
146/200 would be "good" students.

d) Show a picture of the ROC curve augmented with labels and arrows indicating the points on the curve corresponding to cut-offs of 1.4 and 1.6.

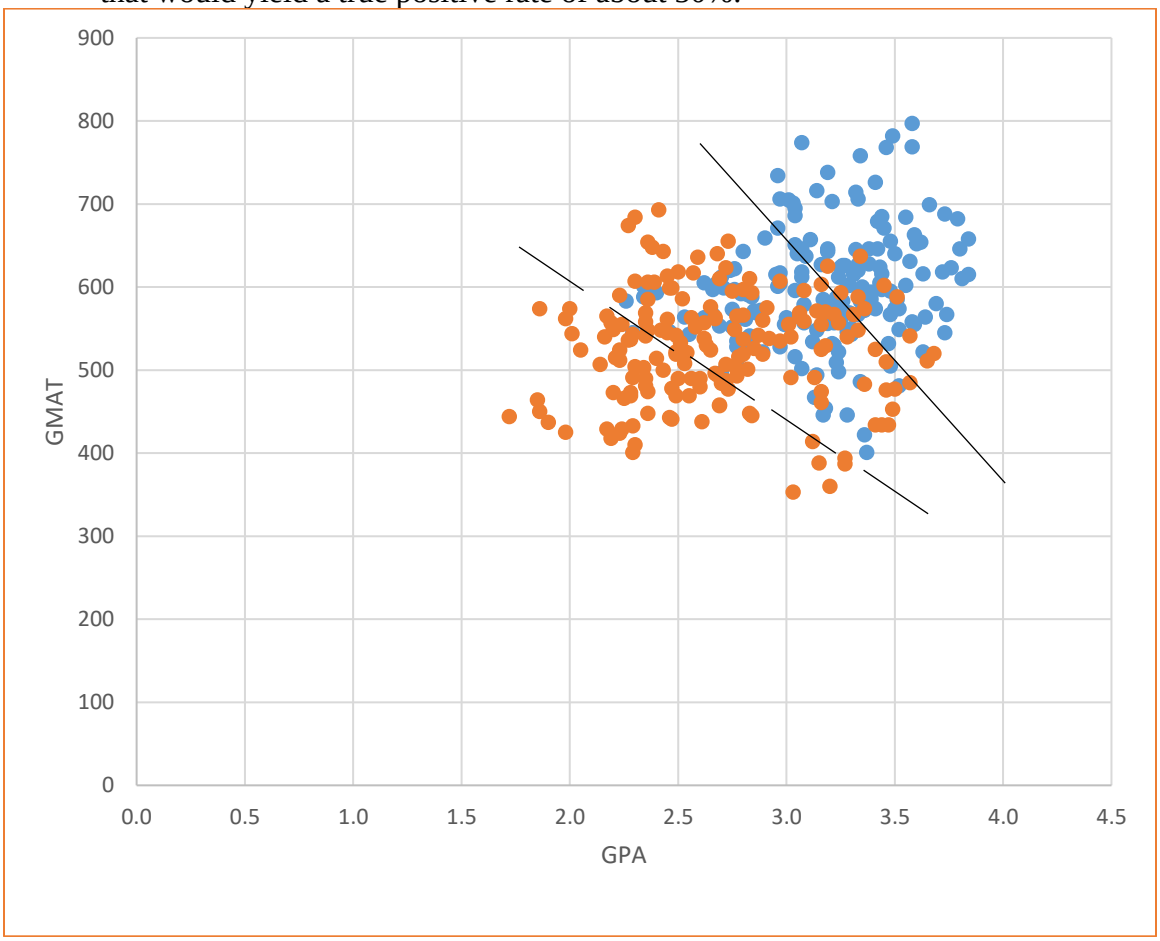
-



Problem #4

Continuing with the previous problem:

a) Scatter plot GMAT vs. GPA with separate plotting points for the two groups. (Adjust the axes so zoom in only on the relevant ranges of values, e.g., 1.5-4.0 for GPA and 300-800 for GMAT). By eye, add a dashed discriminant line that would achieve a 100% true positive rate with the lowest possible FPR and a solid line that would yield a true positive rate of about 50%.



b) On a single set of TPR vs. FPR axes, plot the ROC curves when using (i) Just GPA (X1), (ii) Just GMAT (X2), and (iii) Both variables. You can do that by copying and pasting the FPR and TPR rates from three variants of the template to a separate sheet and creating the scatter plot there. Show your combined ROC curve graph.

c) What kind of school (selective/elite or non-selective) could get by reasonably well using just one variable, and which variable would that be? Write a sentence or two justifying your answer.

A non-selective school could use GMAT scores only, because there is more variation in GMAT data. The TPR is better with GMAT.

Problem #5

Using the Analytical Solver platform's k-nearest neighbor tab under "Data Mining, Classify" run k-nearest neighbor with k = 1, 3, 5, 7, and 9 on the data from Problems #3 and #4 and also the 20 observations in the holdout sample (on sheet "Additional data for Problem #5"). Keep all settings at their defaults except that in the bottom right of the first part of the dialog box you'll need to specify that the success group is 1 not 2 (Group 1 is the strong students). Report the following:

a) Screen shot of the confusion matrix for the training data when k = 7.

Confusion Matrix		
Actual\Predicted	1	2
1	138	37
2	48	127

b) ROC curve for k = 7. (It is produced by the software when you check the box asking for it. You just need to cut and paste.)

c) The predictions for the 20 observations in the holdout sample with k = 7. (Again, you can cut and paste from the computer output.)

d) Create a scatter plot of the accuracy (% correct) for the training sample and the % correct for the holdout sample vs. k, for k = 1, 3, 5, 7, and 9.

e) You'll see a consistent gap between the proportions correct for for the training and hold out sample in that plot. Explain where it comes from and what it suggests about the importance of a holdout sample.

Problem #6

Create a classification tree for the data from Problem #5 (i.e., train with the 350 observations and then apply to the holdout sample of 20). Set the maximum tree depth to 5. Show the following:

a) A screenshot of the fully grown tree

b) Confusion matrix for the training data

c) Number and proportion correct on the training and holdout samples.

d) Explain step by step how the classification tree would classify someone (i) with a GPA of 3.0 and a GMAT of 550 and (ii) with a GPA of 2.6 and a GMAT of 550. Describe each step through the tree and the whether the final classification is to Group 1 (strong students) or Group 2 (weak students).

e) The results in 6d are counterintuitive. Briefly explain conceptually how this relates to overfitting?