Pruthvinath J V
654202834

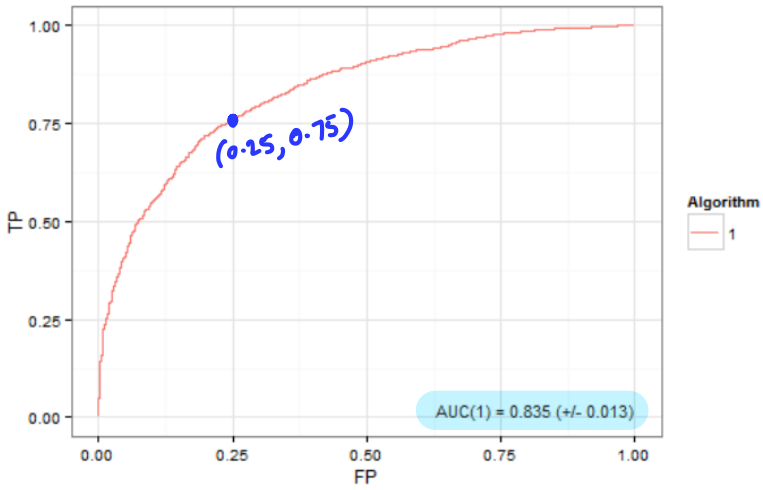
Ujjayini Dey Chowdhury
665035331

BUSINESS DATA MINING (IDS 572)

HOMEWORK 1
DUE DATE: WEDNESDAY, SEPTEMBER 28 AT 11:59 PM

Problem 1. Consider a binary classification problem where we want to classify all patients into "risky" and "low risk" groups. We apply two different models, algorithm A and algorithm B for this purpose. Below, we check the performance of these models using ROC curves.

(a) What can you say about the ROC curve below? How this classifier differs from a random guess? Pick one point on a curve and interpret it. For example, what does the point (0.25, 0.75) indicate?



From the above ROC Curve, algorithm is classifying better. Because the AUC is ~83.5%.

The point (0.25,0.75) says that among the total predicted positives I.e. predicted risky patients, 25% are falsely risky whereas 75% are truly risky.

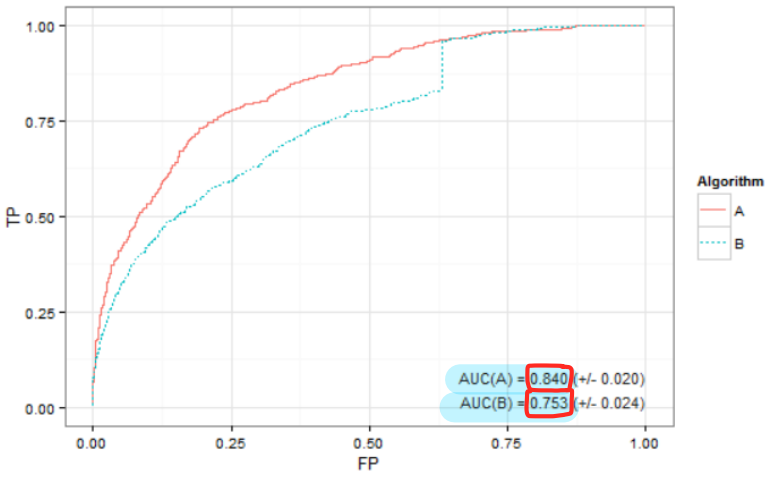
Positive class: classifying a patient as risky.
True Positive (TP): risky patient is classified as risky.
False Positive (FP): low risky patient is classified as risky.

Here, cost of FPR is not much high. Even after misclassification as positive, physician can test again and can prove the patient is not risky.

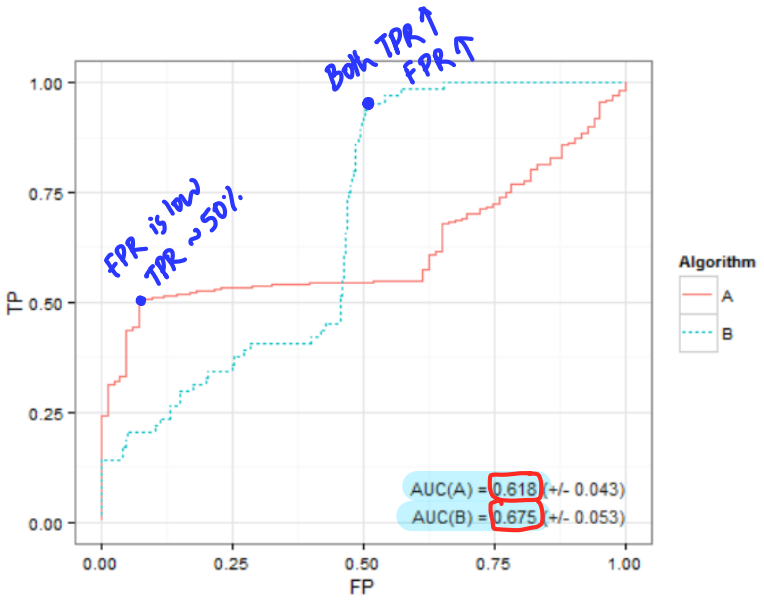
(b) Compare two ROC curves below. Which one is a better model and why?

Model A is better than Model B because of area under the curve.

AUC (A) > AUC (B)



(c) Compare two ROC curves below. Which model is better, A or B? When A would be preferred over B?



Model B is better than Model A because -

1. AUC (B) > AUC (A)
2. Here, cost of FPR is not so high. So model B is better to choose. Because even with more FPR, we get more TPR.

Problem 2. Consider a classifier that classifies documents as being either relevant or non-relevant.

(a) Which evaluation measure do we use for this classifier? Accuracy, precision, and/or recall? Justify your answer.

Positive class: classify as relevant.

False Positive (FP): classify as relevant which is not relevant actually.
False Negative (FN): relevant docs classified as irrelevant.

We have to reduce both FP and FN because -
- with FN, relevant information is being missed
- with FP, more irrelevant info creates junk

Here we have to used Accuracy as evaluation measure.

(b) Suppose that we have a collection of 10 documents - named D1, ... D10 - and two different classifiers A and B. Give an example of two result sets of documents, A_q and B_q, assumed to have been returned by two different systems in response to a query q, constructed such that A_q has clearly higher precision than B_q, but A_q and B_q have the same accuracy.

Precision: $A_q > B_q$
Accuracy ~ F-score: $A_q = B_q$
→ Recall : $A_q < B_q$

$$\text{F-score} = \frac{1}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

$$\rightarrow F = \frac{2PR}{P+R} \rightarrow \frac{2P_A R_A}{P_A + R_A} = \frac{2P_B R_B}{P_B + R_B}$$

Given, $P_A > P_B$
$\Rightarrow R_A < R_B$

(c) Suppose a classifier returns 3 relevant documents and 2 irrelevant documents to a search query. There are a total of 8 relevant documents in the collection. What is the precision of the system on this search, and what is its recall?

Actuals \ Predicted	+	−	
+	3 (TP)	5 (FN)	8
−	2 (FP)	(TN)	

$$\text{Precision} : \frac{TP}{TP+FP} : \frac{3}{3+2} = \frac{3}{5}$$

$$\text{Recall} : \frac{TP}{TP+FN} : \frac{3}{3+5} = \frac{3}{8}$$

Problem 3. We are looking to develop a machine learning algorithm to predict whether someone will pay a loan back or not.

- What is the positive class?
- What would a recall of 75% mean?
- What would a precision of 85% mean?
- Which measure do you choose to evaluate your model?

Positive class: classifying as someone is able to pay the loan.

False Positive (FP): classify the unable person as able to pay.
False Negative (FN): classify the able person as unable to pay.

Here, cost of FP is more. So need to reduce FP.

————————————————————————————

Recall of 75% means among the actual able persons, model can only predict 75% of them as able.

Precision of 85% means among the predicted able persons, only 85% are truly able to pay.

As the cost of FP is more, we need to use Precision as the evaluation measure.

Problem 4

(a) Before thinking about modeling, have a look at your data. Try to understand variables' distributions and their relationships with the target variable. Which variables do you think could be major predictors of diagnosis? Also clean your data appropriately: Are there highly correlated variables? Are there any missing values or outliers? If yes, how you handle them?

Variables we think important are:

- texture (standard deviation of gray scale values)
- radius
- concavity (severity of concave portions of the contour)
- concave points (Number of concave portions of the contour)

There are some highly correlated variables such as:

1. Radius variables with perimeter and area variables ——> removed mean perimeter, mean area, largest radius, largest perimeter and largest area.
2. Texture mean and texture large are correlated ——> removed texture large variable from analysis

(b) Create a decision tree (using "information" for splits) to its full depth. How many leaves are in this tree?

There are 10 leaves in our decision tree

(c) What are the major predictors of diagnosis suggested by your tree? Please justify your reasoning. Do these major predictors are the same as the ones you observed in part [a]?

1. Mean # of concave points
2. Largest # of concave points
3. Mean concavity
4. Mean radius
5. Mean compactness
6. Std dev of area

Guessed almost correct except compactness. Major predictors guessed are radius, concave points and concavity.

(d) Give two strong rules that describe who is likely to have cancer. Please justify your choices.

1. If mean # of concave points >= 0.49 +
largest # of concave points >= 0.15 +
std deviation of concavity < 0.14
=> then 99% probability is Cancer.

2. If mean # of concave points < 0.49 +
mean radius >= 15 +
Mean texture >= 16 +
Std deviation of compactness < 0.023
=> Then 100% probability is Cancer.

(e) What is the accuracy of your decision tree model on the training data? What is the accuracy of this model on the test data?

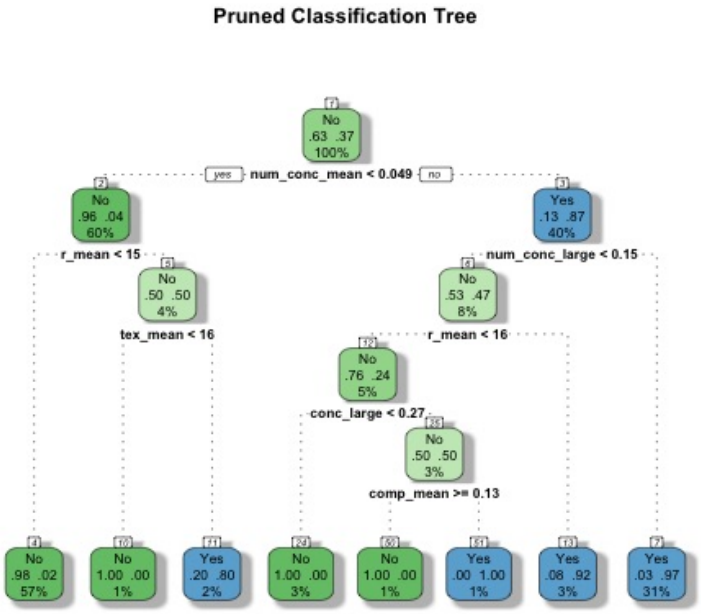
Accuracy of pruned decision tree on:
Training set: 97.58%
Test data: 96.49%

(f) Construct the "best possible" decision tree to predict the Y labels. Explain how you construct such tree and how you evaluate its performance.

After making cross-validation, minimise complexity parameter i.e. cp.

If it makes the model too complex and biased then it won't perform good on test data. Then using next least cp value.

(g) Plot your final decision tree model and write down all decision rules that you will consider for predictions.



No
- 0.00 when num_conc_mean < 0.049 & r_mean >= 15 & tex_mean < 16
- 0.00 when num_conc_mean >= 0.049 & r_mean < 16 & num_conc_large < 0.15 & conc_large < 0.27
- 0.00 when num_conc_mean >= 0.049 & r_mean < 16 & num_conc_large < 0.15 & conc_large >= 0.27 & comp_mean >= 0.13
- 0.02 when num_conc_mean < 0.049 & r_mean < 15

Yes
- 0.80 when num_conc_mean < 0.049 & r_mean >= 15 & tex_mean >= 16
- 0.92 when num_conc_mean >= 0.049 & r_mean >= 16 & num_conc_large < 0.15
- 0.97 when num_conc_mean >= 0.049 & num_conc_large >= 0.15
- 1.00 when num_conc_mean >= 0.049 & r_mean < 16 & num_conc_large < 0.15 & conc_large >= 0.27 & comp_mean < 0.13

Problem 5. (Decision tree and random forest in R) Phishing attacks are the most common type of cyber-attacks used to trick users into clicking on phishing links, stealing user information, and ultimately using user data to fake logging in with related accounts to steal funds. Phishing attacks have been affecting individuals as well as organizations across the globe. In this question, we want to build a machine learning model to detect website phishing. To do so, download the "Phising websites" data set from UCI Machine Learning Repository. Before constructing any models, explore this data set and clean it if required. Construct a decision tree, and a random forest model to detect phishing websites from the legitimate. Use cross-validation to check the performance of your models. Which model will be chosen as your final model? What evaluation measure(s) do you use to select the best model? Justify your answer.

Positive class: classifying site as phishing.
TP: classifying phishing site as phishing.
FP: classifying legitimate site as phishing.
FN: classifying phishing site as legitimate.

Both FP and FN should be minimum => Accuracy should be the evaluation measure.

Accuracy of cross validated RF on training set (98.58%) > Pruned DT (90.3%) AND
Cross validated RF on testing data (90.84%) > pruned DT (89.73%)

FINAL MODEL: Cross validated Random Forest