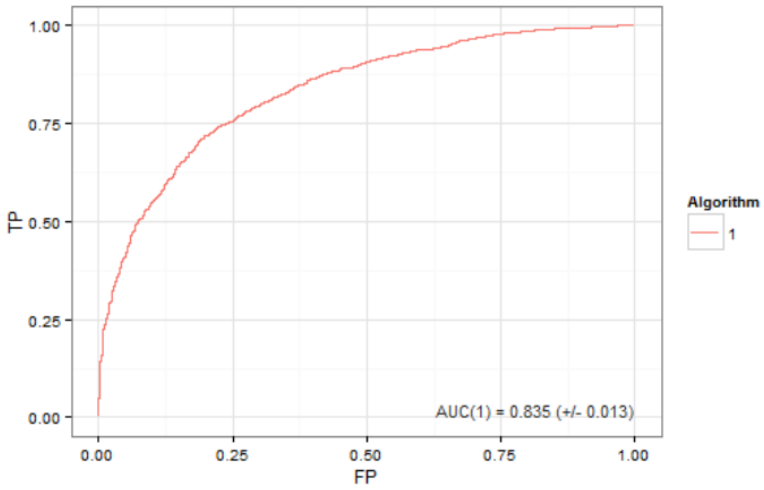
BUSINESS DATA MINING (IDS 572)

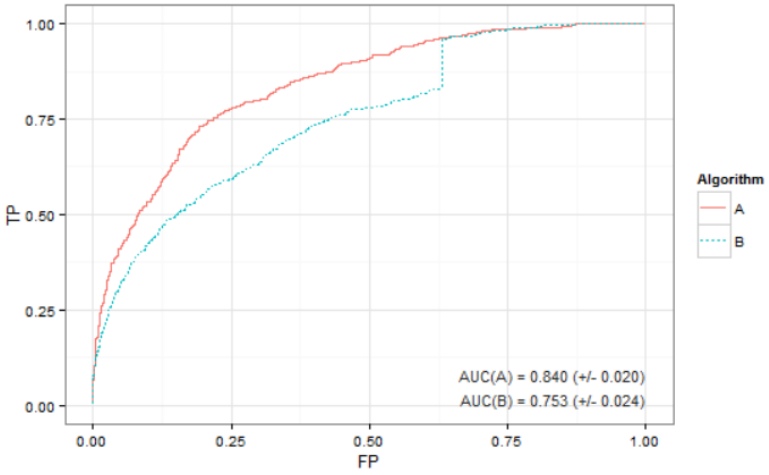
HOMEWORK 1
DUE DATE: WEDNESDAY, SEPTEMBER 28 AT 11:59 PM

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.

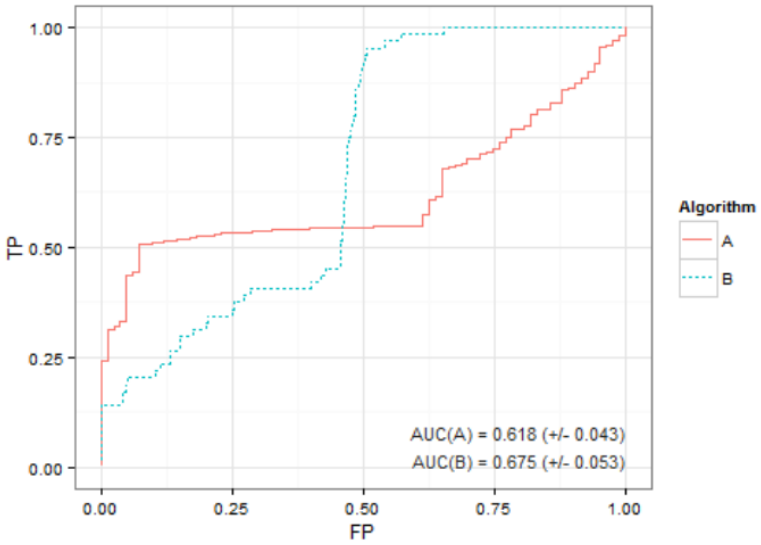
(a) What can you say about the ROC curve below? How this classifier differs from a random guess? Pick one point on a curve and interpret it. For example, what does the point (0.25, 0.75) indicate?



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



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



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.

(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.

(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?

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?

Problem 4. (Decision tree in R) One very interesting application area of machine learning is in making medical diagnoses. In this problem you will train and test a decision tree to detect breast cancer using real world data. To do so use the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. This dataset consists of 569 samples of biopsied tissue. The tissue for each sample is imaged and 10 characteristics of the nuclei of cells presenting each image are characterized. These characteristics are

(1) Radius
(2) Texture
(3) Perimeter
(4) Area
(5) Smoothness
(6) Compactness
(7) Concavity
(8) Number of concave portions of contour
(9) Symmetry
(10) Fractal dimension

Each of the 569 samples used in the dataset consists of a feature vector of length 30. The first 10 entries in this feature vector are the mean of the characteristics listed above for each image. The second 10 are the standard deviation and last 10 are the largest value of each of these characteristics present in each image. Each sample is also associated with a label. A label of value 1 indicates the sample was for malignant (cancerous) tissue. A label of value 0 indicates the sample was for benign tissue.

This dataset has already been broken up into training and test sets for you and is available on blackboard. The names of the files are "trainX.csv", "trainY.csv", "testX.csv" and "testY.csv." The file names ending in "X.csv" contain feature vectors and those ending in "Y.csv" contain labels. Each file is in comma separated value format where each row represents a sample.

(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?

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

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

(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?

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

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

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.