EAS 508 (Section 002): Homework 3

Due October 22, 2021 by 11:59 PM

Notes: The homework should be submitted via UBLearns. 25% of the grade will be removed for each day late (that includes 12:00 AM October 23, so make sure that you give sufficient time to submit). We do not want to see code. That is, the coding is not the end result, but is a means to answer the questions. We are interested in the answers and how you understand the techniques.

Questions:

We have learned several classification learning techniques so far in the class: Logistic Regression, CART, Bagging, Boosting, Random Forest, K-NN, and Naive Bayes. In this HW, we will use the "spam" data set from the "kernlab" package and create models using these seven classification learning techniques. The "spam" data set has 4601 observations and 58 variables. Among these 58 variables (columns), the "type" column represents the target (or dependent) variable. I have uploaded the data dictionary for this data set so that we can understand the data set properly and interpret the meaning of each column. Do not forget to set your seed. For the sake of consistency, we will use: set.seed(123). Answer the following questions:

1. Assume that we can only use five features to develop a spam detection classification model, albeit the availability of 57 columns (or independent variables). Which five columns (or independent variables) will you select as your features? Discuss your rationales. You may support your rationales with statistical evidence. **(5 points)**
2. Develop spam detection classification models using Logistic Regression, CART, Bagging, Boosting, Random Forest, K-NN, and Naive Bayes. Compare these seven models in the lens of accuracy, sensitivity, and specificity when deployed on an unseen test set. Also indicate how generalizable each model is. The answer needs to be succinct. A comparison table will work fine. **(20 points)**

Note that there may have several acceptable answers as long as you justify your rationales.