Stat 5330 Mini Project: Bank Marketing Data

This Project is due at 9pm, EST, Tuesday, April 21st, 2020. A well-written report should be submitted with your code onto Collab. In the paper report, the results should be clearly stated with some reasonable explanations (including necessary plots). Please DO NOT just copy and paste the raw outputs obtained from your software. Also please write down both your full name and the computing ID at the first page.

1. The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

   Exploratory Data Analysis

   (a) Briefly summarize the response variable, i.e., proportion of each class of observations (yes vs no) in both training and testing data sets.

   (b) Briefly summarize the predictors (in training data set) using appropriate tools (**figures or tables**).

   Modeling and Data Analysis

   For a classification problem, we usually have following measurements to evaluate the model performance:

   classification accuracy =num of obs that are correctly classified / total num of obs;

   sensitivity =num of "yes" obs which are correctly classified as "yes"/ num of totoal "yes" obs

   specificity =num of "no" obs which are correctly classified as "no"/ num of totoal "no" obs

   (a) Perform LDA and QDA and apply your trained model to the testing set. Report the corresponding **accuracy, sensitivity and specificity** for testing set predictions.

   (b) Perform a logistic regression model and a linear SVM model and apply your trained model on the testing set. Please report the corresponding **accuracy, sensitivity and specificity**, respectively, for testing set predictions.

(c) Try a non-linear SVM model and apply your trained model on the testing set. Please report the corresponding **accuracy, sensitivity and specificity**, respectively, for testing set predictions. Compare your results to those based on a linear SVM model above.

(d) Compare the prediction results of all above models, which one is better? Do you think the LDA and QDA models are appropriate here? Why?

(e) Perform the LDA, QDA, Logistic regression model and the linear SVM model again with **continuous predictors only** and report their **accuracy, sensitivity and specificity** values for testing data set, respectively. How are they compared with those above using all predictors.

(f) Suppose we just make a random prediction on the testing set, that is, for each obs in the testing set, we randomly toss a coin, and assign the label 1 to this obs if we get a ”head” while assign 0 if get a ”tail”. Calculate the corresponding classification **accuracy, sensitivity and specificity**. You might generate the prediction on the testing set by using the following codes to generate random binary (Bernoulli) outcomes:

```r
# n is the sample size of the testing set, and let prob=0.5
y.test.prediction=rbinom(n, 1, prob)
```

(g) Perform a random-forest model and report the **accuracy, sensitivity and specificity** on the testing predictions. Use a five-folder cross-validation to perform the tuning regarding the ”node size”. For the other two important parameters, we set total number of trees to be 500 and set the number of selected variables at each split as $mtry = \sqrt{p} \approx 7$.

(h) In the above random-forest model, we also want to check the importance of each covariate regarding the classification. Please take advantage of the ”importance” argument in the ”randomForest” function provided in the sample code and make a barplot of the MeanDecreaseGini values with respect to all covariates. (See Sample Code!) Based on what you find, do you think we can build a reduced model with fewer number of predictors?

(i) Run a boosting model with the exponential loss (using ”adaboost” for the argument ”distribution” in the ”gbm” function). Tune the number of trees and the shrinkage factor by five-folder cross-validation. The number of trees is tuned over (100, 200, 500, 1000), and the shrinkage factor is tuned over 0.01 to 0.1 with a step size 0.01. Report the selected tuning parameter values and the corresponding prediction **accuracy, sensitivity and specificity** on the testing set.