Fall 2022

MIS 572 Introduction to Big Data Analytics

Homework 2

- Graded out of **110** points. Please typeset your homework, save as an R source code file with title "your student ID_Homework_2.R" (e.g. B024020001_Homework_2.R).
- Please submit your code to NSYSU Cyber University before **1/3 11:59pm**. **No late submission**.
- DO NOT use any loops in your answers. Also notice that your code must follow the suggested programming and data analysis styles discussed in the class.

1. Please load the given dataset “redwinequality.csv” and answer the following data management questions.

   1.1 **[10 pts]** Please split the data into training set (70%) and testing set (30%), make sure you set "set.seed(2022)". Fit linear regression models that predict "quality" and calculate Mean Absolute Error(MAE) for both training and testing data.

   1.2 **[10 pts]** Please use forward and backward selections to find your best subsets of variables (up to two-way interactions). Did these selections choose the same variables? Also, report the training/testing MAEs of with variables selected by each method.

   1.3 **[10 pts]** Refer to Chapter 6 (Linear Model Selection and Regularization) of the textbook (ISLR). Please fit a LASSO with all variables and R package “glmnet”, use cv.glmnet() to find the best lambda**.** Also report the training/testing MAEs.

   1.4 **[10 pts]** Fit a random forest with all variables and report training/testing MAEs. Print out the importance of variance of random forest and order by the importance. Are random forest rankings different from those generated by linear models?

2. Please load the given dataset “diabetes.csv”. Answer the following questions.

   2.1 **[10 pts]** Split the data into training (70%) and testing (30%) datasets with set.seed(2022). Fit a Logistic regression model to predict “Outcome” and report both the training and testing AUC.

   2.2 **[10 pts]** Tree-based algorithms, such as CART, are able to identify non-linear relationships from data and represent the relationships in interpretable rules. Please use R package “rpart” and “rpart.plot” to fit a classification tree” that predict “Outcome”. Select (based on cross-validation error) and plot your best tree. Then report both the training and testing AUC.

2.3 **[10 pts]** Ensemble learning methods like random forest usually outperform the other simple models alone in terms of accuracy of prediction, because they reduce the variance of prediction by aggregating multiple models. Please fit a random forest model and report training/testing AUC.

2.4 **[10 pts]** Which variable you believe is the most important to predict the "Outcome"? Justify your finding with your analysis result.

3. Please load the given data "body signal of smoking" dataset (smoking.csv). Refer to here for more information.

3.1 **[10 pts]** Use R package caret to split the data into training (70%) and testing (30%) datasets with set.seed(2022). Fit Logistic regression models that predict "smoking". Use or create any variables that may better predict the target. What are the accuracy of predictions on both the training and the testing datasets, given the default cutoff value 0.5?

3.2 **[10 pts]** We can see that there is a class imbalance problem with the target (smoking). We understand that adjusting predicted class probability cutoff may help predict the rare cases. What is the optimal cutoff value based on Youden's J index? Please also report your model True Positive Rates (Sensitivities) with different cutoff values (0.5 and the "optimal" value).

3.3 **[10 pts]** Plot the ROC curves of your models for both training and testing datasets. Compare and report your model performance in terms of AUCs.