Buskirk, DATA 7440

Final Assignment, Spring 2023

This assignment can be completed in groups of 2 to 3 students and you are encouraged to submit your assignment in groups. The assignment is worth 100 points and accounts for **40% of your final grade (taking the place of individual quizzes and a group project).** Your group should submit only one final assignment and you should plan to use R for your submissions via an Rmarkdown file that generates a pdf document. Be sure to include your code chunks as part of the solution. The assignment is due via canvas submission Tuesday April 25, 2023 by 2pm.

Problem 1: [50 Points]

We will consider the API data avaialble from the Survey package in R (see: https://cran.r-project.org/web/packages/survey/survey.pdf). The Academic Performance Index is computed for all California schools based on standardized testing of students. The data sets contain information for all schools with at least 100 students. We want to predict whether or not a school has qualified for it's performance awards (awards) using various school, teacher and parent information. The data for this problem can be found in the apipop1.RData workspace.

Part 1: Using the information in the apipop_train1 dataset within the apipop1.RData workspace:

a. construct a LASSO model that predicts whether or not a school qualified for it's award program. Use 5-fold cross validation on the apipop_train1 dataset to determine the penalty parameter and specify the seed 7440 prior to running the model.

b. Now apply a one rule algorithm to determine the feature that most explains the award outcomes.

c. Now apply a sequential covering algorithm to expand the rule set derived in B.

d. Now apply a rule-fit algorithm to determine features and possible interaction effects that are important for explaining award qualification. For this model set the seed to be 2021 and also use 5 folds.

Part 2: Using the respective models create a table that provides overall accuracy, sensitivity and specificity for predicting if a school qualified for its awards program using the apipop_test1 as the test set for each of the models you derived in parts a-d of Part 1.

Part 3: For every model for which it is appropriate, derive the variable importance measures (feature specific importances, not rule specific importances) and create another table that lists the top 4 most important features for predicting whether a school qualified for it's award program.

For rule fit, provide the graphs that illustrate the top three most important RULES.

Comment on where there are consistencies in the most important features and where there aren't. Which model gives you the best sense of what is going on between the features and qualifying for the awards programs.

Part 4: Now compute a random forest model to predict whether or not a school qualifies for its awards program (using the training data: apipop_train1) with 750 trees and default value of the mtry parameter. Prior to running the model, set the seed to be 429. Using your model:

a. create an overall interaction plot that displays the overall H statistic for each of the features.

b. Using the **feature with the highest overall H statistic**, create an additional plot that shows the H statistics related to this specific feature.

c. Is the information in the second plot consistent with the some of the rules the Rule Fit model is reporting? Why or why not?

d. What is the H-statistic value corresponding to the variables included in the top-most important rule in the rule fit model that involves only two predictors? (Be sure to indicate what the top-most important rule with two features is from your rule fit model, along with it's importance) as well as the plot of that rule from the RuleFit model.

Problem 2: [50 points]

We will use the api data again for this problem, but this time we are interested in predicting the api test scores for 2000 (i.e. api00) based on a battery of school, teacher and parent information. This time the data and models are available in the apipopProb2.RData R workspace. Note the data sets here are different than for Problem 1.

For this problem, you are given three models that aim at predicting api00 (R objects model1 – model3), using different methods (i.e., CART, XGBoost, OLS regression). Utilize the following interpretation techniques with the training data to learn more about how those models were trained. Specifically for each of the models you should:

a. Produce Partial Dependence plots for the most important predictor of each model.

b. Produce Partial Dependence plots for the two most important predictors of each model (i.e., 3D plots or heatmaps or contour plots).

c. Produce Accumulated Local Effects plots for the most important predictor in each model. How do these plots vary from those produced in part b? Given the correlation structure in the apipop_train data is it surprising or not that these plots differ or are similar to each other.

d. Compute the overall interaction H statistics as well as the feature specific statistics for the ELL variable for all of the models. Provide a table that includes as rows the predictors in the apipop_train data file and the H statistics for each model as columns. For the ELL specific H statistics, provide an interaction plot per model.

e. Evaluate the prediction performance of each model using the test set that is included in the workspace (e.g. apipop_test).

f. Considering the results for the tasks above, which model object belongs to which method? Explain your choice!