CART. Homework #3

Name:_____SBU ID:_____

Please include (1) Rmd file; (2) Output from Rmd with answers to all the questions asked

Please upload your homework solutions to the Brightspace by 8:30am on Monday, February 20, 2023.

CART with the Banknote Data – Classification Task

The banknote.csv data (see attached) were extracted from images that were taken from genuine and forged banknote-like specimens. Yes, this is a ***Catch Me if You Can*** story. For digitization, an industrial camera usually used for print inspection was used. The final images have 400x 400 pixels. Wavelet Transform tool were used to extract features from images. There are 1372 banknotes, and 5 variables:

1. variance of Wavelet Transformed image (continuous)
2. skewness of Wavelet Transformed image (continuous)
3. curtosis of Wavelet Transformed image (continuous)
4. entropy of image (continuous)
5. class (binary) - this is the response variable of interest, 1 (forged) or 0 (genuine).

First, please check whether there is any missing value and if so, please delete those observations with missing values. Now after the data cleaning step, your task is to split the data randomly into training (80%) and testing (20%), first build the full tree and then establish an optimal model with pruning and 10-fold cross validation using the training data, and then use that model to predict whether each check in the testing data is forged or not. Please use *rpart* to build the classification tree and *rattle* to plot the tree.

Please review the following websites for related methods and concepts:
http://www.sthda.com/english/articles/35-statistical-machine-learning-essentials/141-cart-model-decision-tree-essentials/
https://trevorstephens.com/kaggle-titanic-tutorial/r-part-3-decision-trees/

1. For the entire dataset, please perform the data cleaning as instructed before; namely, check and delete missing values, if any, in the dataset. Please note that if there is missing value in any of the variables, then that entire observation must be deleted. (There are ways to impute the missing values but we shall not use those techniques at this point.) Please report how many observations are left after this step. Then please use the random seed 123 to divide the cleaned data into 80% training and 20% testing.

2. Please first build a fully grown tree using the training data, and draw the tree plot using *rattle*. Next please use this tree to predict whether each check is forged or not in the testing data. Please compute the Confusion matrix and report the sensitivity (that is, a forged check is found

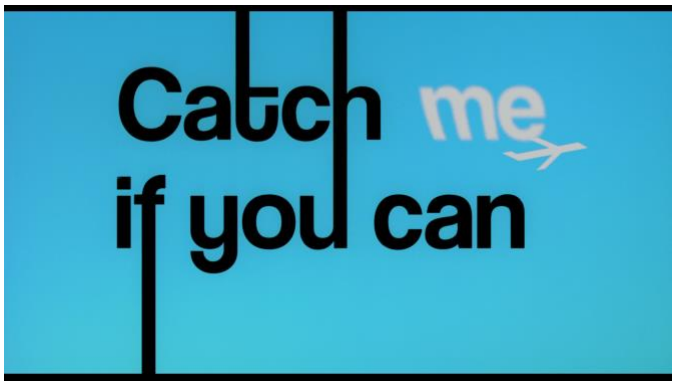
to be forged), specificity (that is, a genuine check is found to be genuine), and the overall accuracy for the testing data.

3. To make the tree more robust, we will prune the fully grown tree using the training data with 10-fold cross-validation. Please (1) show the complexity plot, (2) report the best CP value, and (3) draw the pruned tree using *rattle*.

4. Please use this optimal pruned tree to predict whether each check is forged or not in the testing data. Please compute the Confusion matrix and report the sensitivity, specificity and the overall accuracy for the testing data.

5. Please fit a logistic regression model using the training data and use this model to predict whether each check is forged or not in the testing data. Please compute the Confusion matrix and report the sensitivity, specificity and the overall accuracy for the testing data.

6. Please output the test data along with the prediction results from (1) the Full Tree, (2) the Optimal Pruned Tree, and (3) the logistic regression, in an excel file. Please generate the majority vote using these three prediction methods. Please compute the Confusion matrix and report the sensitivity, specificity and the overall accuracy for the testing data based on the majority vote.



https://wallpapercave.com/catch-me-if-you-can-wallpapers