Assignment 3

Background

A national veterans' organization wishes to develop a predictive model to improve the cost-effectiveness of their direct marketing campaign. The organization, with its in-house database of over 13 million donors, is one of the largest direct-mail fundraisers in the United States. According to their recent mailing records, the overall response rate is 5.1%. Out of those who responded (donated), the average donation is $13.00. Each mailing, which includes a gift of personalized address labels and assortments of cards and envelopes, costs $0.68 to produce and send.

Using these facts, we take a sample of this dataset to develop a classification model that can effectively capture donors so that the expected net profit is maximized. Weighted sampling is used, under-representing the non-responders so that the sample has equal numbers of donors and non-donors.

Data

The file Fundraising.csv contains 3120 records with 50% donors (TARGET_B = 1) and 50% non-donors (TARGET_B = 0). The amount of donation (TARGET_D) is also included but is not used in this case. Refer to the book for the descriptions for the 22 variables (including two target variables) in page 522.

Use "Fundraising.csv" to solve the below questions in Python and submit only the Python script file for your group thru BB. Give me your answers with # as comments in the code.

1. Provide me with a frequency table of donors VS no donors with the average donation for all donors.
2. Drop non-useful columns ['Row Id', 'Row Id.', 'TARGET_D']
3. Partition the dataset into 60% training and 40% validation with specifying "TARGET_B" is the outcome variable, while others are input variables in form of X and Y objects.
4. Run the following models with providing confusion matrix f
    a. Logistic Regression
    b. Classification Tree
    c. Random Forest
    d. Neural Network
    e. Linear Discriminant Analysis
5. What do you think is the best model?