Art History of Florence

A new title: The Art History of Florence, is ready for release. CBC sent a test mailing to a random sample of 4000 customers from its customer base. The customer responses have been collated with past purchase data. Each row (or case) in the spreadsheet (other than the header) corresponds to one market test customer. Each column is a variable, with the header row giving the name of the variable. The variable names and descriptions are given below

Variable Name	Description
Seq#	Sequence number in the partition
ID#	Identification number in the full (unpartitioned) market test dataset
Gender	0=Male,1=Female
M	Monetary—Total money spent on books
R	Recency—Months since last purchase
F	Frequency—Total number of purchases
FirstPurch	Months since first purchase
ChildBks	Number of purchases from the category child books
YouthBks	Number of purchases from the category youth books
CookBks	Number of purchases from the category cookbooks
DoItYBks	Number of purchases from the category do-it-yourself books
RefBks	Number of purchases from the category reference books (atlases, encyclopedias, dictionaries)

ArtBks	Number of purchases from the category art books
GeoBks	Number of purchases from the category geography books
ItalCook	Number of purchases of book title Secrets of Italian Cooking
ItalAtlas	Number of purchases of book title Historical Atlas of Italy
ItalArt	Number of purchases of book title Italian Art
Florence	= 1 if The Art History of Florence was bought; = 0 if not

Data Mining Techniques

Various data mining techniques can be used to mine the data collected from the market test. No one technique is universally better than another. The particular context and the particular characteristics of the data are the major factors in determining which techniques perform better in an application. For this assignment, we focus on four fundamental techniques:

- K Nearest Neighbors
- Classification Trees
- Random Forest
- Boosted Trees
- Logistic Regression

In the direct marketing business, the most commonly used variables are the RFM variables:

- R = recency, time since last purchase
- F = frequency, number of previous purchases from the company over a period
- M = monetary, amount of money spent on the company's products over a period

The assumption is that the more recent the last purchase, the more products bought from the company in the past, and the more money spent in the past buying the company's products, the more likely the customer is to purchase the product offered.

For *The Art History of Florence*, CBC wants to use the following all the explanatory variables (including R, F, and M) in the data to predict whether a customer would order the book.

Assignment:

1. Data partitioning and exploration (10 points)
    - Partition the data into training (60%) and validation (40%). Use seed = 1.
    - What is the response rate for the training data customers taken as a whole?
    - Plot the response rate by the Recency variable?
    - Plot the response rate by the Monetary variable?
    - Plot the response rate by the Frequency variable?
2. k-Nearest Neighbors (20 points): Use the training set to construct a k-nearest-neighbor approach to classify cases with k = 1, 2, ..., 11, using Florence as the outcome variable. Remember to normalize all the explanatory variables.
    - Based on the validation set, find the best k.
    - Calculate the confusion matrix for the best k model on the validation set and report the Sensitivity, Specificity, Positive Prediction Value, and the overall Accuracy.
3. Classification Tree (20 points): Use the training set to construct a classification tree of optimal depth with Florence as the outcome variable and all the explanatory variables.
    - Calculate the confusion matrix for the classification tree on the validation set and report the Sensitivity, Specificity, Positive Prediction Value, and the overall Accuracy.
4. Random Forest (20 points): Use the training set to construct a random forest with Florence as the outcome variable and all the explanatory variables.
    - Calculate the confusion matrix for the random forest on the validation set and report the Sensitivity, Specificity, Positive Prediction Value, and the overall Accuracy.
5. Boosted Trees (20 points): Use the training set to construct a boosted trees model with Florence as the outcome variable and all the explanatory variables.
    - Calculate the confusion matrix for the classification tree on the validation set and report the Sensitivity, Specificity, Positive Prediction Value, and the overall Accuracy.

6. Logistic Regression (20 points): Use the training set to construct a logistic regression model with Florence as the outcome variable and all the explanatory variables.
    - If the cutoff criterion for a campaign is a 30% likelihood of a purchase, calculate the confusion matrix on the validation set and report the Sensitivity, Specificity, Positive Prediction Value, and the overall Accuracy.
7. Final Recommendation (10 points): Create a table with the Sensitivity, Specificity, Positive Prediction Value, and the overall Accuracy for each of the four data mining techniques above. Based on these results, which technique (if any) would you recommend that CBC use for to determine their mailing list for *The Art History of Florence?*