Assignment 4

(To be done individually. **Due 11:59 PM March 13**.)

Please read the Bookbinders Book Club case in the Class 9 module before doing this assignment.

Gary Lilien, BookBinder's marketing director, has experimented with several database marketing approaches to improve BBBC's mailing yields and profits. His first market test relied on RFM analysis. Direct marketers have used RFM approaches to predict customer behavior for more than 50 years. Despite this initial success, Lilien would like to evaluate alternative approaches. BBBC offers books in many different genres and records every customer's previous book purchases by genre in the database. RFM analysis, however, does not use this or other customer information such as gender. Lilien suspects that a more sophisticated modeling approach could yield superior results to RFM.

In particular, Lilien wants to try logistic regression, which offers a powerful method for modeling response. For a recent mailing, the company selected 20,000 customers in Pennsylvania, New York and Ohio from its database and included with their regular mailing a specially produced brochure for the book *The Art History of Florence*. BBBC then developed a database to calibrate a response model to identify the factors that influenced these purchases. For this assignment, we will use a subset of the database available to BBBC, collected in the file BBBCData.xlsx. This file has one worksheet called "Estimation sample" and another called "Validation sample."

Part I: Calibrating models using the Estimation Sample

Q1. Estimate a logistic regression model using "Choice" as the dependent variable and the following as explanatory variables: Gender, Amt_purchased, Frequency, Last_Purchase, First_purchase, P_Child, P_Youth, P_Cook, P_DIY, and P_Art. Report the regression results.

Q2. Summarize and interpret the results (so that a marketing manager can understand them). Which variables are statistically significant? Which seem to be economically "important"? Interpret the economic importance for some of the explanatory variables. (Hint: Use `mfx<-logitmfx()` function to estimate the marginal probability change. Then report the results by `mfx$mfxest`. Here, you can calculate either marginal probability change at mean of x or average marginal probability change.)

Q3. Run a "pure R" model (i.e., a RFM model without F and M) with **deciles** and predict response rate for each decile group.

Part II: Validate the two models on the Validation Sample

Q4. Check the overall classification performance of the logistic regression model above on the validation sample. Create a table showing the fraction of observations which are correctly predicted by the model. (Hint: use *table ()* command shown in the lecture slides.)

Q5. Using your logistic regression result, assign each customer to a decile based on his/her predicted probability of purchase. (Hint: use .bincode function as we did in RFM.)

Q6. Using your logistic regression result, report the number of customers, the number of buyers of "The Art History of Florence," and the average response rate to the offer by decile for the 1200 customers in the validation dataset, respectively. (Hint: To show the number of observations by decile, use `table()` function. To show the number of buyers by decile, use `tapply(x,decile,sum)`, and to show the average predicted response rate by decile, use `tapply(x,decile,mean)`. You need to fill in x and decile.)

Q7. Using your pure-R model, report the number of customers and the number of buyers of "The Art History of Florence," by decile for the 1200 customers in the validation dataset.

Part III: Lift and Cumulative Lift in the Validation Sample

Q8. Use the computations above to create a table showing the lift and cumulative lift for each decile, for both logistic regression results and R(FM) results. You may want to use Excel for these calculations.

Q9. Use the computations above to create a table showing the gains and cumulative gains for each decile, for both logistic regression results and R(FM) results. You may want to use Excel for these calculations.

Q10. Create a chart showing the cumulative gains by decile along with a reference line corresponding to "no model," for the logistic regression and R(FM).

Part IV: Targeting

Use the following cost information to assess the profitability of using logistic regression to determine which of the remaining customers should receive a specific offer.

- Cost of each solicitation: $1
- Margin on each sale of *The Art History of Florence*: $10.20

Q11. What is the breakeven response rate? (Hint: No R is required.)

Q12. Using your logistic regression result, for the customers in the validation dataset, create a new variable (call it "**target**") with a value of 1 if the customer's predicted probability is greater than or equal to the breakeven response rate and 0 otherwise.

Q13. Using your logistic regression result, compute the following numbers

- Percentage of customers you will send a mail to
- Average response rate among the customers you target
- Expected number of buyers
- Gross profit
- Gross sales
- Total marketing cost
- Marketing ROI

Q14. Using your R(FM) result, compute the following numbers

- Percentage of customers that you are going send a mail

- Average response rate among the customers you target
- Expected number of buyers
- Gross profit
- Gross sales
- Total marketing cost
- Marketing ROI

Q15. Compare the results of mass-marketing, pure-R model and logistic regression in terms of Marketing ROI, based on the calculations above.