Kellogg School of Management
Northwestern University

BookBinders: Predicting Response with Logistic Regression

As a direct marketer of specialty books, the BookBinders Book Club has achieved steady growth in their customer base. Yet while sales have grown steadily, profits began falling when the database got larger and when the company diversified its book selection and increased the number of offers sent to customers. The falling profits have led Dave Lawton, BookBinders' marketing director, to experiment with different database marketing approaches in order improve BookBinders' mailing yields and profits.

Dave began a series of live market tests, each involving a random sample of customers from the database. An offer for the current book selection is sent to the sample and then the sample customers' responses, either purchase or no purchase, are recorded and used to calibrate a response model for the current offering. The response model's results are then used to "score" the remaining customers in the database and select customers from the full customer database for the 'rollout' mailing campaign.

Dave's first market tests relied on RFM (recency – frequency – monetary) analysis. Direct marketers have used this approach to predict customer behavior for more than 50 years. The approach is intuitive, easy to implement, and produced significant improvements in response rates and profits compared with mass mailings to BookBinders' full database. Despite this initial success, Dave is eager to evaluate the effectiveness of alternate approaches. BookBinders offers books in different categories including cooking, art and children's' books – and the number of previous book purchases in each category is recorded in each customer's record in the database. RFM analysis does not use this or other customer information such as gender and Dave suspects that a more sophisticated modeling approach could yield superior results to the RFM approach.

Logistic Regression offers a powerful method for modeling response. Logistic regression is similar to linear regression – the key difference is that the dependent variable is binary (for example, purchase or no purchase) rather than continuous. For each customer, logistic

Professor Charlotte Mason from the University of North Carolina prepared this case to provide material for class discussion rather than to illustrate either effective or ineffective handling of a business situation. Names and data may have been disguised to assure confidentiality. The assistance of the Direct Marketing Educational Foundation in supplying the data used for this case is gratefully acknowledged. Professor Florian Zettelmeyer modified this case for use in his course and for use with R instead of SPSS..

regression predicts a probability, between 0 and 1, of purchase or response, which can be used for targeting and prediction decisions. Like linear regression, it can accommodate both continuous and categorical predictors, including interaction terms. Its use in database marketing has grown as software becomes more readily available and as familiarity with the approach grows.

The company currently has 550,000 customers who are being mailed catalogs. Dave has just received a dataset containing the responses of a random sample of 50,000 customers to a new offering from BookBinders titled “The Art History of Florence.” Dave is eager to assess the potential value of logistic regression as a method for predicting customer response and has asked you to complete the following analyses.

Part I: Logistic Regression *(10 points)*

1. Estimate a logistic regression model using “buyer” as the dependent variable and the following as predictor variables:

   gender
   last
   total
   child
   youth
   cook
   do_it
   reference
   art
   geog

   > Technical Note:
   > *purch* is excluded from the set of predictor variables – including it will lead to perfect collinearity since *purch* (the number of books purchased) is equal to the sum of the number of books purchased in the 7 categories. By including the number of purchases in each category, there is no need to include the total number of purchases.

2. Use “plot_model(…, show.values = TRUE, transform = NULL)” to display the coefficients and confidence intervals. Which variables are statistically significant and which ones seem to be economically ‘important’?

3. Use the “plot_model(…, type=”eff”)” command to plot marginal effects. For which variables does your assessment of the importance of a variable change and why?

4. Add the predicted values of the logistic regression model to the “bbb” data frame. For the first few observations in the data, visually compare the “buyer” variable to the predicted values. Next, for the full dataset, compare the average of the predicted values with the average of the “buyer” variable. What do you notice? Why is that?

Part II: Decile Analysis of Logistic Regression Results *(6 points)*

1. Assign each customer to a decile based on his or her predicted probability of purchase. Assign those with the highest predicted probability of purchase to decile 1 and those with the lowest predicted probability of purchase to decile 10.

2. Create a bar chart plotting response rate by decile (as just defined above).
   Hint: The “response rate” is not the same as the "predicted probability of purchase" that the model generated. Instead, it is the actual percentage of customers in a given group (for example a decile) that have bought "The Art History of Florence."

3. Generate a report showing number of customers, the number of buyers of “The Art History of Florence’ and the response rate to the offer by decile for the random sample (i.e. the 50,000) customers in the dataset.

Part III: Lifts and Gains *(5 points)*

1. Use the information from the report in II.3 above to create a table showing the lift and cumulative lift for each decile. You may want to use Excel for these calculations.

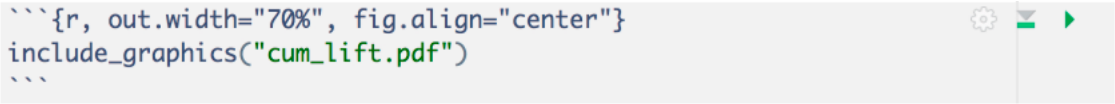
2. In Excel, create a chart showing the cumulative lift by decile.

3. Use the information from the report in II.3 above to create a table showing the gains and cumulative gains for each decile. You may want to use Excel for these calculations.

4. In Excel, create a chart showing the cumulative gains by decile along with a reference line corresponding to ‘no model’.

Hint: Please integrate the Excel-generated charts into the R Notebook you are using for the rest of this assignment. Here is how:

- Save the graphs and tables in Excel as pdf files
- Place the pdf files into the same directory as your R Notebook for the assignment
- Use the “`include_graphics()`” command to insert each pdf.
- For example, suppose your pdf is called “`cum_lift.pdf`”, then insert the code block:



~~~
```{r, out.width="70%", fig.align="center"}
include_graphics("cum_lift.pdf")
```
~~~

- Please note the header of the code block. There you can change the width of the chart (here 70% of page width) and how it is aligned (here centered).

Part IV: Profitability Analysis *(9 points)*

Use the following cost information to assess the profitability of using logistic regression to target customers:

Cost to mail offer to customer:	$.50
Selling price (shipping included):	$18.00
Wholesale price paid by BookBinders:	$9.00
Shipping costs:	$3.00

1. What is the breakeven response rate?

2. For the customers in the 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. (Hint: in mutate() multiply the TRUE/FALSE expression with “1” to get a 0/1 variable).

3. For the customers in the dataset, if had you used the model to select which customer to target, what percentage of customer would you have targeted? Of those customers you would have targeted, what percentage would have purchased the "Art History of Florence?"

4. For the 500,000 remaining customers, what would the expected profit (in dollars) and the expected return on marketing expenditures have been if BookBinders had mailed the offer to buy "The Art History of Florence" only to customers with a predicted probability of buying that was greater than or equal to the breakeven rate? Make the calculations in R?

5. For the 500,000 remaining customers, calculate the incremental profit of having used the logistic regression model instead of a mass mailing?

(Please see the next page for a description of the data)

Exhibit 1

The BookBinders Book Club
R Dataset

Summary information about the BookBinders Book Club's customers' purchasing history and demographics is in the R dataset called *bbb.Rdata*

Below is a listing of the variable names and descriptions of the data types:

The contents of bbb.Rdata is one data frame "bbb" which contains records for 50,000 customers			
Variable name	Type	Size	Description
acctnum	Numeric	5	Customer account number
gender	String	1	Customer gender – M=male, F=female
state	String	2	State where customer lives (2-character abbreviation)
zip	String	5	ZIP code (5-digit)
zip3	String	3	First 3 digits of ZIP code
first	Numeric	3	Number of months since first purchase
last	Numeric	3	Number of months since most recent purchase
book	Numeric	8	Total dollars spent on books
nonbook	Numeric	8	Total dollars spent on non-book products
total	Numeric	8	Total dollars spent
purch	Numeric	5	Total number of books purchased
child	Numeric	5	Total number of children's books purchased
youth	Numeric	5	Total number of youth books purchased
cook	Numeric	5	Total number of cook books purchased
do_it	Numeric	5	Total number of do-it-yourself books purchased
reference	Numeric	5	Total number of reference books purchased
art	Numeric	5	Total number of art books purchased
geog	Numeric	5	Total number of geography books purchased
buyer	Numeric	1	Did the customer buy "The Art History of Florence?" (1=yes, 0=no)
training	Numeric	1	Dummy variable that splits the dataset into a training ("1") and validation ("0") dataset. This variable is used only later in the course.