Rady School of Management
University of California, San Diego

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 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 targeting 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 mailing campaign rollout.

Dave's first market tests relied on RFM (Recency – Frequency – Monetary) analysis. Direct marketers have used this approach to predict customer response for decades. 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.

Logistic regression is a powerful tool to model response. It is similar to linear regression but the key difference is that the response variable is binary (e.g., purchase or no purchase) rather than continuous. For each customer, logistic regression predicts a probability of purchase, which can be used for targeting. Like linear regression, it can accommodate both continuous and categorical predictors, including interaction terms.

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 value of logistic regression as a method to predict 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 response variable and the following as explanatory variables:

   gender, last, total, child, youth, cook, do_it, reference, art, geog

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 Professor Florian Zettelmeyer modified this case for use in his course and for use with Stata and Professor Vincent Nijs modified it for use with R and Radiant.

See “?radiant.model::logistic” for details. Create a new variable called “purch_prob” with the predicted purchase probability for each customer.

2. Summarize and interpret the logistic regression results. Which explanatory variables are statistically significant? Which seem to be most “important”? Make sure your model evaluation includes an interpretation of the odds-ratios estimated for each of the explanatory variables.

Part II: Decile Analysis of Logistic Regression Results *(10 points)*

1. Assign each customer to a decile based on his or her predicted probability of purchase. Assign the deciles to a new variable “prob_dec”. **Note**: The first decile should have the highest average predicted probability of purchase.

2. Create a bar chart of response rates per decile (i.e., use “prob_dec” as the x-variable and “buyer” as the y-variable). Note that the “response rate” is **not** the same as the "predicted probability of purchase." Response rate captures the proportion of customers in a given group (e.g., in a decile) that actually bought "The Art History of Florence." **Note**: Always check what the first level of a {factor} is. By default these levels will be in alphabetical order unless explicitly specified otherwise.

3. Report the number of customers, the number of buyers of “The Art History of Florence”, and the response rate to the offer per decile for the sample of customers we have data for (i.e., the 50,000).

4. Estimate a logistic regression model with “buyer” as the response variable and “child” as the only explanatory variable. Why is the odds ratio for “child” different compared to the logistic regression in Part I? Please be specific and investigate beyond simply stating the statistical problem.

Part III: Lifts and Gains (5 points)

1. Use the information reported in II.3 above to create a table with lift and cumulative lift numbers for each decile. Please use R for these calculations.

2. Create a ggplot chart showing the cumulative lift per decile.

3. Use the information reported in II.3 above to create a table with gains and cumulative gains numbers for each decile. Please use R for these calculations.

4. Create a ggplot chart showing the cumulative gains per decile along with a reference line for the “no model”.

Part IV: Profit, ROME, and Confusion (5 points)

1. Use the information reported in II.3 above and the cost information shown in section V below to create a table of profit and ROME numbers for each decile. Please use R for these calculations.

2. Create a ggplot chart showing the Profit per decile.

3. Create a ggplot chart showing the ROME per decile.

4. Create the confusion matrix using the cost and net revenue numbers shown in section V below. Also calculate model "accuracy" (see http://lab.rady.ucsd.edu/sawtooth/RBusinessAnalytics/logit_models.html for an example).

Part V: Profitability Analysis *(5 points)*

Use the following cost information to assess the profitability of using logistic regression to determine which of the remaining 500,000 customers should receive a specific offer:

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

1. What is the break-even response rate?

2. For customers in the dataset, create a new variable "mailto_logit" that is TRUE if the customer's predicted purchase probability is greater than the break-even response rate and FALSE otherwise.

3. Considering that there are 500,000 remaining customers, generate a report summarizing the number of customers, the expected number of buyers of "The Art History of Florence", and the expected response rate to the offer when "mailto_logit" is TRUE

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

5. The calculations in V.2 through V.4 above assume that the predicted probabilities are estimated without error. Calculate the standard error of the prediction by adding "se = TRUE" to the call to predict for the logistic regression model. Now redo the calculations from V.2 through V.4 adjusting for these errors. How do the results change?

Part VI: Model comparison (10 points)

1. Compare the model performance of (1) Sequential RFM, (2) Naïve Bayes, (3) Logistic regression, and (4) Logistic regression adjusted for prediction standard errors. Use profit calculations as in V.1 through V.4, Lift, Gains, Profit and ROME charts, and Confusion matrices. What model would you recommend that Dave Lawton use and why?

Information about the random sample of 50,000 BookBinders Book Club's customers' purchasing history and demographics is in the R dataset *bbb.rda* on Dropbox in the data/ directory. Below is an overview of the variables:

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