Marketing Data Science

Modeling Techniques in Predictive Analytics with R and Python

THOMAS W. MILLER

Publisher: Paul Boger
Editor-in-Chief: Amy Neidlinger
Executive Editor: Jeanne Glasser Levine
Operations Specialist: Jodi Kemper
Cover Designer: Alan Clements
Managing Editor: Kristy Hart
Manufacturing Buyer: Dan Uhrig

Published by Pearson Education, Inc.
Old Tappan New Jersey 07675

For information about buying this title in bulk quantities, or for special sales opportunities (which may include electronic versions; custom cover designs; and content particular to your business, training goals, marketing focus, or branding interests), please contact our corporate sales department at corpsales@pearsoned.com or (800) 382-3419.

For government sales inquiries, please contact governmentsales@pearsoned.com.

For questions about sales outside the U.S., please contact international@pearsoned.com.

Printed in the United States of America

First Printing May 2015

ISBN-10: 0-13-388655-7
ISBN-13: 978-0-13-388655-9

Pearson Education LTD.
Pearson Education Australia PTY, Limited.
Pearson Education Singapore, Pte. Ltd.
Pearson Education Asia, Ltd.
Pearson Education Canada, Ltd.
Pearson Educación de Mexico, S.A. de C.V.
Pearson Education—Japan
Pearson Education Malaysia, Pte. Ltd.
Library of Congress Control Number: 2015937911

Contents

3 Targeting Current Customers

"Listen, I—I appreciate this whole seduction scene you've got going, but let me give you a tip: I'm a sure thing. OK?"

—JULIA ROBERTS AS VIVIAN WARD IN *Pretty Woman* (1990)

Mass marketing treats all customers as one group. One-to-one marketing focuses on one customer at a time. Target marketing to selected groups of customers or market segments lies between mass marketing and one-to-one marketing. Target marketing involves directing marketing activities to those customers who are most likely to buy.

Targeting implies selection. Some customers are identified as more valuable than others and these more highly valued customers are given special attention. By becoming skilled at targeting, a company can improve its profitability, increasing revenues and decreasing costs.

Targeting is best executed by companies that keep detailed records for individuals. These are companies that offer loyalty programs or use a customer relationship management system. Sales transactions for individual customers need to be associated with the specific customer and stored in a customer database. Where revenues (cash inflows) and costs (cash outflows) are understood, we can carry out discounted cash-flow analysis and compute the return on investment for each customer.

A target is a customer who is worth pursuing. A target is a profitable customer—sales revenues from the target exceed costs of sales and support. Another way to say this is that a target is a customer with positive lifetime value. Over the course of a company's relationship with the customer, more money comes into the business than goes out of the business.

Managers want to predict responses to promotions and pricing changes. They want to anticipate when and where consumers will be purchasing products. They want to identify good customers for whom sales revenues are higher than the cost of sales and support.

For companies engaging in direct marketing, costs may also be associated with individual customers. These costs include mailings, telephone calls, and other direct marketing activities. For companies that do not engage in direct marketing or lack cost records for individual customers, general cost estimates are used in estimating customer lifetime value.

In target marketing, we need to identify factors that are useful and determine how to use those factors in modeling techniques. A response variable is something we want to predict, such as sales dollars, volume, or whether a consumer will buy a product. Customer lifetime value is a composite response variable, computed from many transactions with each customer, and these transactions include observations of sales and costs.

Explanatory variables are used to predict response variables. Explanatory variables can be continuous (having meaningful magnitude) or categorical (without meaningful magnitude). Statistical models show the relationship between explanatory variables and response variables.

Common explanatory variables in business-to-consumer target marketing include demographics, behavioral, and lifestyle variables. Common explanatory variables in business-to-business marketing include the size of the business, industry sector, and geographic location. In target marketing, whether business-to-consumer or business-to-business, explanatory variables can come from anything that we know about customers, including the past sales and support history with customers.

Regression and classification are two types of predictive models used in target marketing. When the response variable (the variable to be predicted) is continuous or has meaningful magnitude, we use regression to make the prediction. Examples of response variables with meaningful magnitude are sales dollars, sales volume, cost of sales, cost of support, and customer lifetime value.

When the response variable is categorical (a variable without meaningful magnitude), we use classification. Examples of response variables without meaningful magnitude are whether a customer buys, whether a customer stays with the company or leaves to buy from another company, and whether the customer recommends a company's products to another customer.

To realize the benefits of target marketing, we need to know how to target effectively. There are many techniques from which to choose, and we want to find the technique that works best for the company and for the marketing problem we are trying to solve.

All other things being equal, the customers with the highest predicted sales should be the ones the sales team will approach first. Alternatively, we could set a cutoff for predicted sales. Customers above the cutoff are the customers who get sales calls—these are the targets. Customers below the cutoff are not given calls.

When evaluating a regression model using data from the previous year, we can determine how close the predicted sales are to the actual/observed sales. We can find out the sum of the absolute values of the residuals (observed minus predicted sales) or the sum of the squared residuals.

Another way to evaluate a regression model is to correlate the observed and predicted response values. Or, better still, we can compute the squared correlation of the observed and predicted response values. This last measure is called the coefficient of determination, and it shows the proportion of response variance accounted for by the linear regression model. This is a number that varies between zero and one, with one being perfect prediction.

If we plotted observed sales on the horizontal axis and predicted sales on the vertical axis, then the higher the squared correlation between observed sales and predicted sales, the closer the points in the plot will fall along a straight line. When the points fall along a straight line exactly, the squared correlation is equal to one, and the regression model is providing a perfect prediction of sales, which is to say that 100 percent of sales response is accounted for by the model. When we build a regression model, we try to obtain a high value for the proportion of response variance accounted for. All other things being equal, higher squared correlations are preferred.

The focus can be on predicting sales or on predicting cost of sales, cost of support, profitability, or overall customer lifetime value. There are many possible regression models to use in target marketing with regression methods.

To develop a classification model for targeting, we proceed in much the same way as with a regression, except the response variable is now a category or class. For each customer, a logistic regression model, for example, would provide a predicted probability of response. We employ a cut-off value for the probability of response and classify responses accordingly. If the cut-off were set at 0.50, for example, then we would target the customer if the predicted probability of response is greater than 0.50, and not target otherwise. Or we could target all customers who have a predicted probability of response of 0.40, or 0.30, and so on. The value of the cut-off will vary from one problem to the next.

To illustrate the targeting process we consider the Bank Marketing Study from appendix C (page 356). The bank wants its clients to invest in term deposits. A term deposit is an investment such as a certificate of deposit. The interest rate and duration of the deposit are set in advance. A term deposit is distinct from a demand deposit, which has no set rate or duration.

The bank is interested in identifying factors that affect client responses to new term deposit offerings, which are the focus of the marketing campaigns. What kinds of clients are most likely to subscribe to new term deposits? What marketing approaches are most effective in encouraging clients to subscribe?

We begin by looking at the subset of bank clients who are approached with a call for the first time. Part of the challenge in target marketing is dealing with low rates in response to sales and promotional efforts. In this problem only 71 of 3,705 bank clients responded affirmatively by subscribing to the term deposit being offered by the bank.

We examine the relationships between each demographic variable and response to the bank's offer. The demographic variables include age, job type, marital status, and level of education. We also examine relationships between banking experience variables and response to the bank's offer. These variables include the client's average yearly balance, and whether or not the client defaulted on a loan, has a housing loan, or has a personal loan.

Figures 3.1 through 3.5 provide mosaic and lattice plots for selected relationships. Bank clients who subscribe to term deposit offers are older, more highly educated, more likely to be white collar workers than blue collar workers, and more likely to be single, divorced, or widowed than married. They are also less likely to have a housing loan with the bank.

We define a linear predictor using the eight explanatory variables and fit a logistic regression model to the training data. With logistic regression, the left-hand side of the model is a mathematical rendering of the probability of ordering a system upgrade—the logarithm of an odds ratio.

Even though the name of the method is called "logistic regression," the method involves classification, not regression—the response in this problem is categorical, whether or not the client accepts the bank's offer.

We can judge model performance by statistical criteria. After selecting a modeling technique—logistic regression in this case—we employ a probability cut-off to identify target customers. The model provides a predicted probability of response, and we use the cut-off to convert the probability of response into a choice prediction.

When observed binary responses or choices are about equally split between *yes* and *no*, for example, we would use a cut-off probability of 0.50. That is, when the predicted probability of responding *yes* is greater than 0.50, we predict *yes*. Otherwise, we predict *no*.

Logistic regression provides a means for estimating the probability of a favorable (yes) response to the offer. The density lattice in figure 3.6 provides a pictorial representation of the model and a glimpse at model performance.

To evaluate the performance of this targeting model, we look at a two-by-two contingency table or confusion matrix showing the predicted and observed response values. A 50 percent cut-off does not work in the Bank Marketing Study, given the low base rate of responses to the offer.

Figure 3.1. Age and Response to Bank Offer

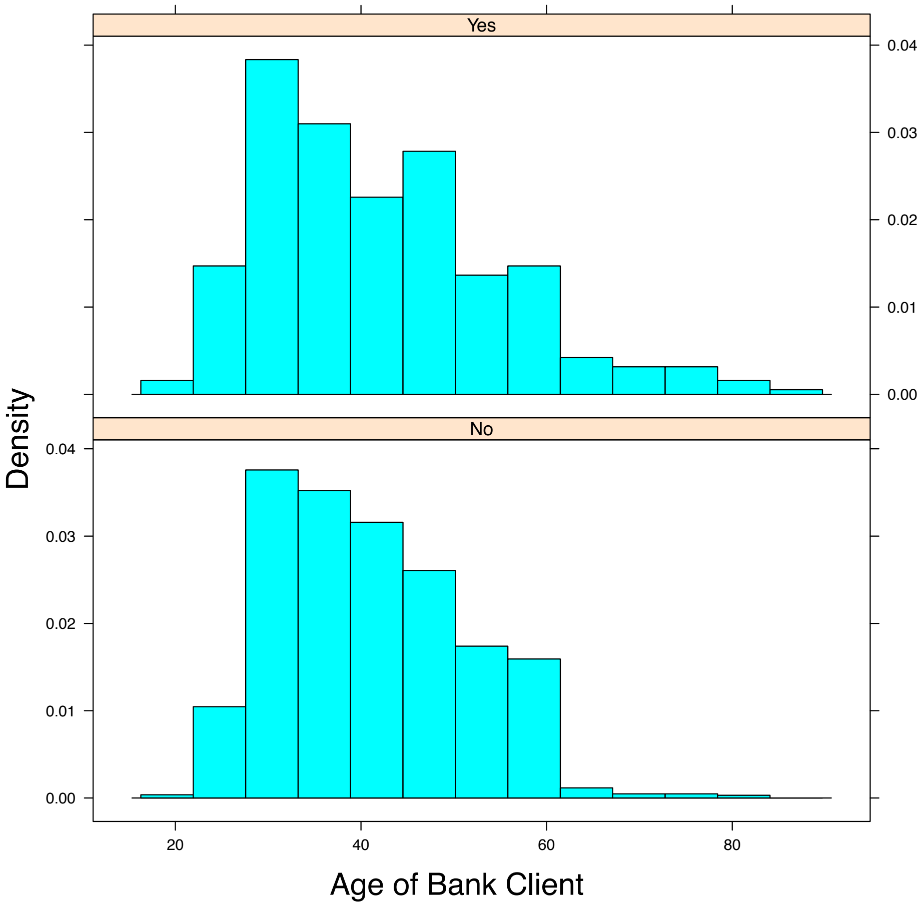


***Figure* 3.2.** *Education Level and Response to Bank Offer*

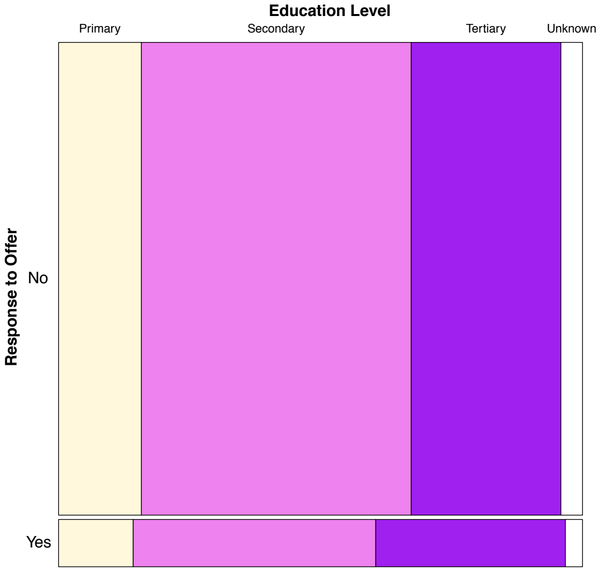


***Figure* 3.3.** *Job Type and Response to Bank Offer*

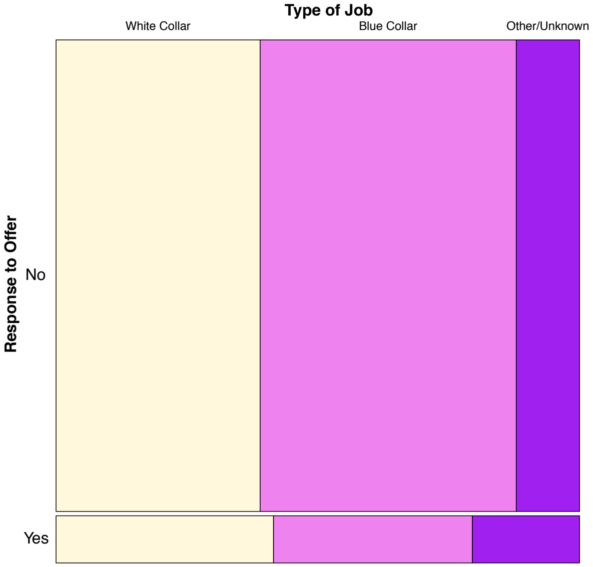


Figure 3.4. Marital Status and Response to Bank Offer

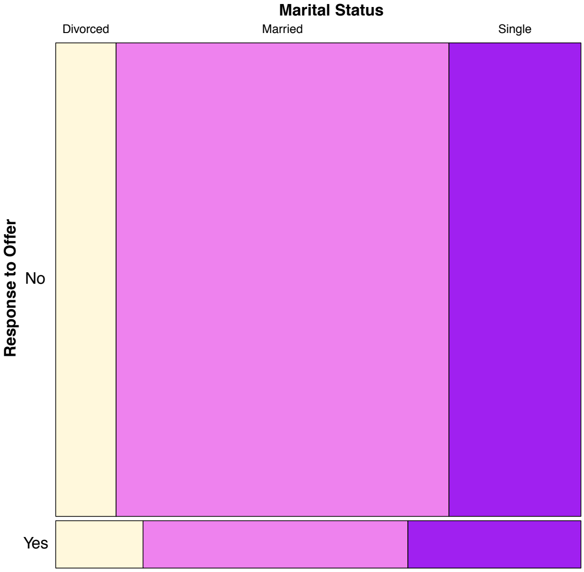


Figure 3.5. Housing Loans and Response to Bank Offer

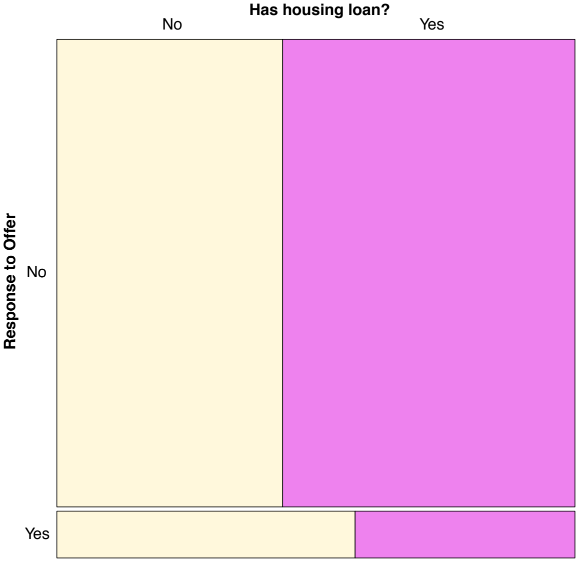


Figure 3.6. *Logistic Regression for Target Marketing (Density Lattice)*

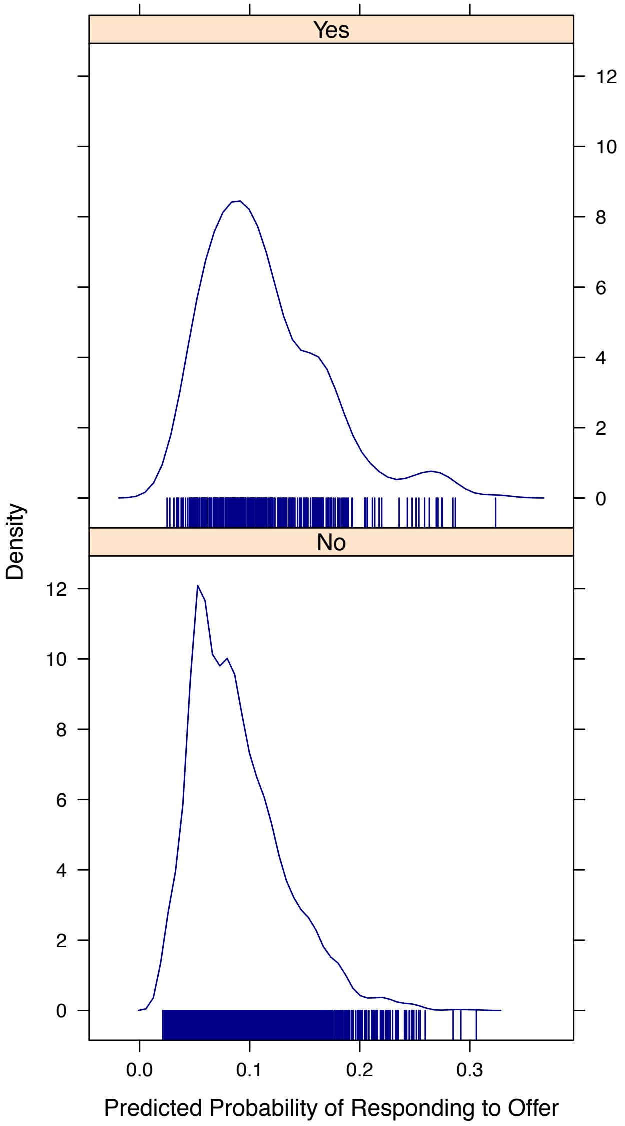
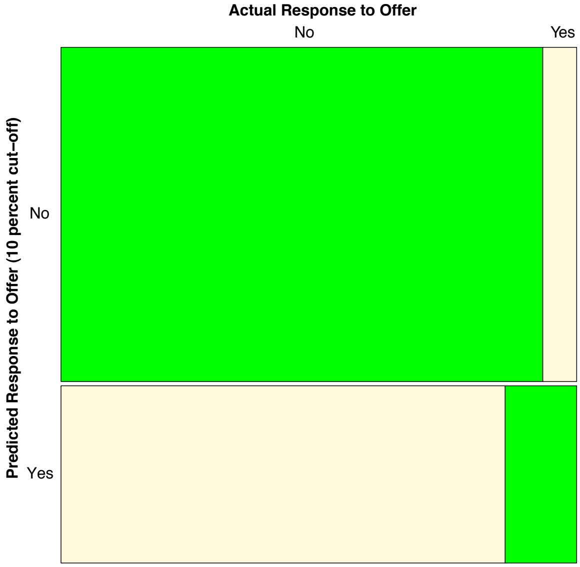


Figure 3.7. *Logistic Regression for Target Marketing (Confusion Mosaic)*

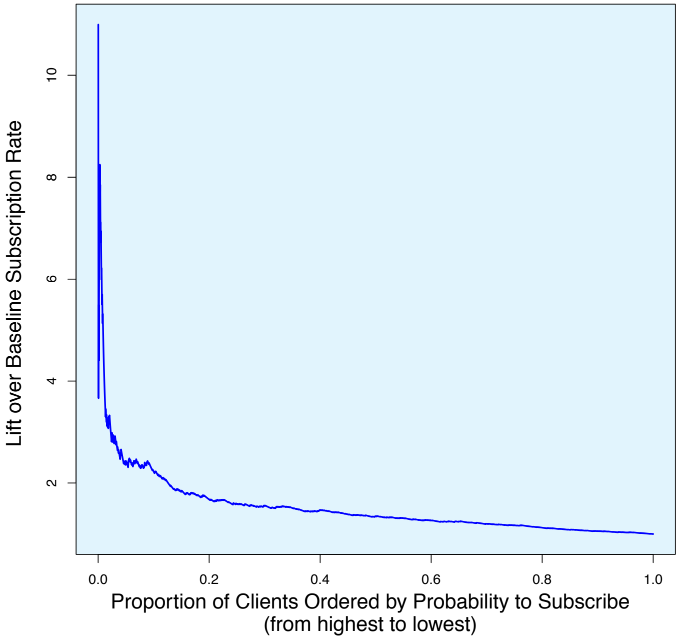


A 50 percent cut-off will not work for the bank, but using a 10 percent cut-off for the response variable (accepting the term deposit offer or not), yields 65.9 percent accuracy in classification. The confusion matrix for the logistic regression and 10 percent cut-off is shown as a mosaic in figure 3.7.

The Bank Marketing Study is typical of target marketing problems. Response rates are low, much lower than 0.50, so a 50 percent cut-off performs poorly. In fact, if bank analysts were to use a 50 percent cut-off, they would predict that every client would respond *no*, and the bank would target no one. Too high a cut-off means the bank will miss out on many potential sales.

Too low a cut-off presents problems as well. Too low a cut-off means the bank will pursue sales with large numbers of clients, many of whom will never subscribe to the term deposit offer. It is wise to pick a cut-off that maximizes profit, given the unit revenues and costs associated with each cell of the confusion matrix. Target marketing, employed in the right situations and with the right cut-offs, yields higher profits for a company.

Figure 3.8. *Lift Chart for Targeting with Logistic Regression*



The analyst or data scientist sets the cut-off probability, and the cut-off affects the financial performance of the targeting model. One approach to picking the right cut-off is to compute *lift* or the response rate that the predictive model provides over the response rate observed in the entire customer base. We order customers by their predicted probability of responding to an offer and then note how much this predicted probability is greater than the base rate of responding to the offer. Lift is a ratio of these probabilities or rates of responding.

Figure 3.8 displays the lift chart for the Bank Marketing Study. The horizontal axis shows the proportion of clients ordered by their probability to subscribe, from highest to lowest, and the vertical axis shows the associated values of lift.

To set a cut-off value for the probability of responding, we might determine that we want to contact clients who are at least twice as likely to subscribe than clients at large. Then we would choose the probability cut-off that corresponds to a lift value of two. Lift does not directly translate into revenues and costs, however, so it may make more sense to perform financial calculations to choose a probability cut-off for targeting.

When we engage in target marketing, we review data from current customers, particularly sales transaction data. We also assess the costs of sales and support for current customers. We can think of each customer as an investment, and compute the return on investment for each customer. If the expected lifetime value of a customer is positive, then it makes sense to retain that customer.

Customer lifetime value analysis draws on concepts from financial management. We evaluate investments in terms of cash in-flows and out-flows over time. Before we pursue a prospective customer, we want to know that discounted cash in-flows (sales) will exceed discounted cash out-flows (costs). Similarly, it makes sense to retain a current customer when discounted future cash flows are positive. To review discounted cash flow and investment analysis, see financial management references such as Higgins (2011) and Brealey, Myers, and Allen (2013).

Customer lifetime value is computed from our experience with each customer. For the cash in-flows, we note a customer's purchasing history as recorded in sales transactions. For the cash out-flows, we note past sales and support costs as recorded in customer relationship management systems. Customer lifetime value analysis is best executed when detailed records are maintained for customers.

Data for valuing customers may be organized as panel or longitudinal data. Rows correspond to customers and columns correspond to time periods. Data from past transactions may be incomplete, and future cash-flows are unknown. So, we use predictive models to estimate cash-flows. We draw on available data to impute missing observations from the past, and we use observations from the past to forecast observations in the future.

Direct marketers are the quintessential target marketers. Their work involves contacting prospective and former customers directly through telephone, mail, e-mail, and online channels. Direct marketers collect and maintain information about past contacts, mailings, in-coming and outgoing communications, and business transactions. And they do this on a customer-by-customer basis. These data are used to guide sales promotions and direct marketing programs. Direct mailings and outgoing communications include product brochures and announcements, as well as coupons and information about product prices, bundles, and promotions.

Each direct marketing promotion can be evaluated in terms of its contribution to the profit of the firm. There are costs associated with mailings and online activities. There are revenues coming from people who order. The hit-rate or proportion of people who respond to a direct mail or online offer is a critical number to watch because it determines the success or failure of the promotion.

Let us consider revenues and costs in the Bank Marketing Study. Bank term deposits provide money for loans. A bank generates revenue by charging a higher rate of interest on loans than it pays on deposits. Suppose the difference in interest rates across average deposit and loan amounts is 100 Euros—this corresponds to the revenue from one client term deposit. Suppose the associated sales and marketing costs for each client contact (including mailings and telephone calls) is 5 Euros. Furthermore, suppose that post-deposit/post-sale support costs are 25 Euros. Then a financial analysis of target marketing using logistic regression with a 10 percent cut-off would play out as shown in figure 3.9. We can see that target marketing is financially beneficial to the bank. The money it saves in sales and marketing costs exceeds the money lost as a result of having fewer term deposit subscriptions.

Direct marketing promotions, properly constructed, represent field experiments. Rarely is it wise to mail to an entire list at once. Better to divide the list into sections and vary the direct mail offer or advertising copy across sections. The conditions that yield the highest profit on the test mailing set the stage for subsequent mailings. Numerous treatment conditions can be examined for each direct marketing promotion in a phased rollout of the marketing effort. Methods that have been practiced for many years across physical mail channels are now being employed across online channels.

***Figure* 3.9.** *Financial Analysis of Target Marketing*

Logistic Regression Model
Confusion Matrix (10 percent cut-off)

Predicted	Actual Response		
Response	No	Yes	
No	2,262	159	2,421
Yes	1,106	178	1,284
	3,368	337	3,705

Pursue All Clients

Number of Subscriptions	Revenue per Subscription		
337	100	33,700	Revenue
Supported Clients	Unit Cost of Support		
337	25	8,425	Expense
Clients Pursued by Sales and Marketing	Unit Cost of Sales and Marketing		
3,705	5	18,525	Expense
		6,750	Profit

Pursue Only Targeted Clients

Number of Subscriptions	Revenue per Subscription		
178	100	17,800	Revenue
Supported Clients	Unit Cost of Support		
178	25	4,450	Expense
Clients Pursued by Sales and Marketing	Unit Cost of Sales and Marketing		
1,284	5	6,420	Expense
		6,930	Profit

Direct and database marketers build models for predicting who will buy in response to marketing promotions. Traditional models, or what are known as *RFM models*, consider the recency (date of most recent purchase), frequency (number of purchases), and monetary value (sales revenue) of previous purchases. More complicated models utilize a variety of explanatory variables relating to recency, frequency, monetary value, and customer demographics.

Useful reviews of traditional direct marketing are provided by Wunderman (1996), and Nash (2000, 1995). Hughes (2000) discusses strategies associated with database marketing and online direct marketing. Direct and database marketing is a rich area of application in marketing data science. Anand and Büchner (2002) discuss applications in cross-selling, finding prospects for additional products from an existing customer list. Blattberg, Kim, and Neslin (2008) provide a comprehensive review modeling methods in direct and database marketing, including extensive discussion of RFM models, lift charts, and alternative methods for setting probability cut-offs.

Lift charts and ROC curves are common tools in direct and database marketing. Area under the ROC curve is a good way to evaluate the statistical accuracy of classifiers, especially when working on a low-base-rate problem as observed in the Bank Marketing Study. There are many other ways to evaluate the statistical accuracy of a classifier. See appendix A (page 271).

Target marketing (in particular, one-to-one target marketing) has been bolstered by the emergence of hierarchical Bayesian methods. Bayesians use the term *consumer heterogeneity* to refer to individual differences across customers. The thinking is that describing consumers in terms of their positions along underlying attribute parameters is more informative than describing them as being members of segments. Bayesian methods in marketing, reviewed by Rossi, Allenby, and McCulloch (2005), have been implemented in R packages by Rossi (2014) and Sermas (2014).

Exhibit 3.1 shows the R program for identifying target customers in the Bank Marketing Study. The program draws on R packages provided by Meyer, Zeileis, Hornik, and Friendly (2014), Sarkar (2008, 2014), and Sing et al. (2015). The corresponding Python program is on the website for the book.

Exhibit 3.1. *Identifying Customer Targets (R)*

```
# Identifying Customer Targets (R)

# call in R packages for use in this study
library(lattice)  # multivariate data visualization
library(vcd)  # data visualization for categorical variables
library(ROCR)  # evaluation of binary classifiers

# read bank data into R, creating data frame bank
# note that this is a semicolon-delimited file
bank <- read.csv("bank.csv", sep = ";", stringsAsFactors = FALSE)
# examine the structure of the bank data frame
print(str(bank))

# look at the first few rows of the bank data frame
print(head(bank))

# look at the list of column names for the variables
print(names(bank))

# look at class and attributes of one of the variables
print(class(bank$age))
print(attributes(bank$age))  # NULL means no special attributes defined
# plot a histogram for this variable
with(bank, hist(age))

# examine the frequency tables for categorical/factor variables
# showing the number of observations with missing data (if any)

print(table(bank$job , useNA = c("always")))
print(table(bank$marital , useNA = c("always")))
print(table(bank$education , useNA = c("always")))
print(table(bank$default , useNA = c("always")))
print(table(bank$housing , useNA = c("always")))
print(table(bank$loan , useNA = c("always")))

# Type of job (admin., unknown, unemployed, management,
# housemaid, entrepreneur, student, blue-collar, self-employed,
# retired, technician, services)
# put job into three major categories defining the factor variable jobtype
# the "unknown" category is how missing data were coded for job...
# include these in "Other/Unknown" category/level
white_collar_list <- c("admin.","entrepreneur","management","self-employed")
blue_collar_list <- c("blue-collar","services","technician")
bank$jobtype <- rep(3, length = nrow(bank))
bank$jobtype <- ifelse((bank$job %in% white_collar_list), 1, bank$jobtype)
bank$jobtype <- ifelse((bank$job %in% blue_collar_list), 2, bank$jobtype)
bank$jobtype <- factor(bank$jobtype, levels = c(1, 2, 3),
    labels = c("White Collar", "Blue Collar", "Other/Unknown"))
with(bank, table(job, jobtype, useNA = c("always")))  # check definition

# define factor variables with labels for plotting
```

```
bank$marital <- factor(bank$marital,
    labels = c("Divorced", "Married", "Single"))
bank$education <- factor(bank$education,
    labels = c("Primary", "Secondary", "Tertiary", "Unknown"))
bank$default <- factor(bank$default, labels = c("No", "Yes"))
bank$housing <- factor(bank$housing, labels = c("No", "Yes"))
bank$loan <- factor(bank$loan, labels = c("No", "Yes"))
bank$response <- factor(bank$response, labels = c("No", "Yes"))

# select subset of cases never perviously contacted by sales
# keeping variables needed for modeling
bankwork <- subset(bank, subset = (previous == 0),
    select = c("response", "age", "jobtype", "marital", "education",
               "default", "balance", "housing", "loan"))

# examine the structure of the bank data frame
print(str(bankwork))
# look at the first few rows of the bank data frame
print(head(bankwork))
# compute summary statistics for initial variables in the bank data frame
print(summary(bankwork))

# -----------------
# age  Age in years
# -----------------
# examine relationship between age and response to promotion
pdf(file = "fig_targeting_customers_age_lattice.pdf",
    width = 8.5, height = 8.5)
lattice_plot_object <- histogram(~age | response, data = bankwork,
    type = "density", xlab = "Age of Bank Client", layout = c(1,2))
print(lattice_plot_object)  # responders tend to be older
dev.off()
# -----------------------------------------------------------
# education
# Level of education (unknown, secondary, primary, tertiary)
# -----------------------------------------------------------
# examine the frequency table for education
# the "unknown" category is how missing data were coded
with(bankwork, print(table(education, response, useNA = c("always"))))
# create a mosaic plot in using vcd package
pdf(file = "fig_targeting_customers_education_mosaic.pdf",
    width = 8.5, height = 8.5)
mosaic( ~ response + education, data = bankwork,
  labeling_args = list(set_varnames = c(response = "Response to Offer",
  education = "Education Level")),
  highlighting = "education",
  highlighting_fill = c("cornsilk","violet","purple","white",
      "cornsilk","violet","purple","white"),
  rot_labels = c(left = 0, top = 0),
  pos_labels = c("center","center"),
  offset_labels = c(0.0,0.6))
dev.off()
```

```
# ---------------------------------------------------------------
# job status using jobtype
# White Collar: admin., entrepreneur, management, self-employed
# Blue Collar: blue-collar, services, technician
# Other/Unknown
# ---------------------------------------------------------------
# review the frequency table for job types
with(bankwork, print(table(jobtype, response, useNA = c("always"))))
pdf(file = "fig_targeting_customers_jobtype_mosaic.pdf",
    width = 8.5, height = 8.5)
mosaic( ˜ response + jobtype, data = bankwork,
  labeling_args = list(set_varnames = c(response = "Response to Offer",
  jobtype = "Type of Job")),
  highlighting = "jobtype",
  highlighting_fill = c("cornsilk","violet","purple",
      "cornsilk","violet","purple"),
  rot_labels = c(left = 0, top = 0),
  pos_labels = c("center","center"), offset_labels = c(0.0,0.6))
dev.off()
# ----------------------------------------------
# Marital status (married, divorced, single)
# [Note: ‘‘divorced’’ means divorced or widowed]
# ----------------------------------------------
# examine the frequency table for marital status
# anyone not single or married was classified as "divorced"
with(bankwork, print(table(marital, response, useNA = c("always"))))
pdf(file = "fig_targeting_customers_marital_mosaic.pdf",
    width = 8.5, height = 8.5)
mosaic( ˜ response + marital, data = bankwork,
  labeling_args = list(set_varnames = c(response = "Response to Offer",
  marital = "Marital Status")),
  highlighting = "marital",
  highlighting_fill = c("cornsilk","violet","purple",
      "cornsilk","violet","purple"),
  rot_labels = c(left = 0, top = 0),
  pos_labels = c("center","center"),
  offset_labels = c(0.0,0.6))
dev.off()
# -----------------------------------------
# default  Has credit in default? (yes, no)
# -----------------------------------------
with(bankwork, print(table(default, response, useNA = c("always"))))
pdf(file = "fig_targeting_customers_default_mosaic.pdf",
    width = 8.5, height = 8.5)
mosaic( ˜ response + default, data = bankwork,
  labeling_args = list(set_varnames = c(response = "Response to Offer",
  default = "Has credit in default?")),
  highlighting = "default",
  highlighting_fill = c("cornsilk","violet"),
  rot_labels = c(left = 0, top = 0),
  pos_labels = c("center","center"),
  offset_labels = c(0.0,0.6))
dev.off()
```

```
# ------------------------------------------
# balance  Average yearly balance (in Euros)
# ------------------------------------------
# examine relationship between age and response to promotion
pdf(file = "fig_targeting_customers_balance_lattice.pdf",
    width = 8.5, height = 8.5)
lattice_plot_object <- histogram(~balance | response, data = bankwork,
    type = "density",
    xlab = "Bank Client Average Yearly Balance (in dollars)",
    layout = c(1,2))
print(lattice_plot_object)  # responders tend to be older
dev.off()
# ------------------------------------
# housing  Has housing loan? (yes, no)
# ------------------------------------
with(bankwork, print(table(housing, response, useNA = c("always"))))
pdf(file = "fig_targeting_customers_housing_mosaic.pdf",
    width = 8.5, height = 8.5)
mosaic( ~ response + housing, data = bankwork,
  labeling_args = list(set_varnames = c(response = "Response to Offer",
  housing = "Has housing loan?")),
  highlighting = "housing",
  highlighting_fill = c("cornsilk","violet"),
  rot_labels = c(left = 0, top = 0),
  pos_labels = c("center","center"),
  offset_labels = c(0.0,0.6))
dev.off()
# ----------------------------------
# loan  Has personal loan? (yes, no)
# ----------------------------------
with(bankwork, print(table(loan, response, useNA = c("always"))))
pdf(file = "fig_targeting_customers_loan_mosaic.pdf",
    width = 8.5, height = 8.5)
mosaic( ~ response + loan, data = bankwork,
  labeling_args = list(set_varnames = c(response = "Response to Offer",
  loan = "Has personal loan?")),
  highlighting = "loan",
  highlighting_fill = c("cornsilk","violet"),
  rot_labels = c(left = 0, top = 0),
  pos_labels = c("center","center"),
  offset_labels = c(0.0,0.6))
dev.off()
# ----------------------------------
# specify predictive model
# ----------------------------------
bank_spec <- {response ~ age + jobtype + education + marital +
    default + balance + housing + loan}
# ----------------------------------
# fit logistic regression model
# ----------------------------------
bank_fit <- glm(bank_spec, family=binomial, data=bankwork)
print(summary(bank_fit))
print(anova(bank_fit, test="Chisq"))
```

```
# compute predicted probability of taking the train
bankwork$Predict_Prob_Response <- predict.glm(bank_fit, type = "response")

pdf(file = "fig_targeting_customer_log_reg_density_evaluation.pdf",
    width = 8.5, height = 8.5)
plotting_object <- densityplot( ~ Predict_Prob_Response | response,
               data = bankwork,
               layout = c(1,2), aspect=1, col = "darkblue",
               plot.points = "rug",
               strip=function(...) strip.default(..., style=1),
               xlab="Predicted Probability of Responding to Offer")
print(plotting_object)
dev.off()

# predicted response to offer using using 0.5 cut-off
# notice that this does not work due to low base rate
# we get more than 90 percent correct with no model
# (predicting all NO responses)
# the 0.50 cutoff yields all NO predictions
bankwork$Predict_Response <-
    ifelse((bankwork$Predict_Prob_Response > 0.5), 2, 1)

bankwork$Predict_Response <- factor(bankwork$Predict_Response,
    levels = c(1, 2), labels = c("NO", "YES"))

confusion_matrix <- table(bankwork$Predict_Response, bankwork$response)
cat("\nConfusion Matrix (rows=Predicted Response, columns=Actual Choice\n")
print(confusion_matrix)
predictive_accuracy <- (confusion_matrix[1,1] + confusion_matrix[2,2])/
                        sum(confusion_matrix)
cat("\nPercent Accuracy: ", round(predictive_accuracy * 100, digits = 1))

# this problem requires either a much lower cut-off
# or other criteria for evaluation... let’s try 0.10 (10 percent cut-off)
bankwork$Predict_Response <-
    ifelse((bankwork$Predict_Prob_Response > 0.1), 2, 1)
bankwork$Predict_Response <- factor(bankwork$Predict_Response,
    levels = c(1, 2), labels = c("NO", "YES"))
confusion_matrix <- table(bankwork$Predict_Response, bankwork$response)
cat("\nConfusion Matrix (rows=Predicted Response, columns=Actual Choice\n")
print(confusion_matrix)
predictive_accuracy <- (confusion_matrix[1,1] + confusion_matrix[2,2])/
                        sum(confusion_matrix)
cat("\nPercent Accuracy: ", round(predictive_accuracy * 100, digits = 1))
# mosaic rendering of the classifier with 0.10 cutoff
with(bankwork, print(table(Predict_Response, response, useNA = c("always"))))
pdf(file = "fig_targeting_customers_confusion_mosaic_10_percent.pdf",
    width = 8.5, height = 8.5)
mosaic( ~ Predict_Response + response, data = bankwork,
  labeling_args = list(set_varnames =
  c(Predict_Response =
      "Predicted Response to Offer (10 percent cut-off)",
       response = "Actual Response to Offer")),
  highlighting = c("Predict_Response", "response"),
  highlighting_fill = c("green","cornsilk","cornsilk","green"),
  rot_labels = c(left = 0, top = 0),
  pos_labels = c("center","center"),
  offset_labels = c(0.0,0.6))
dev.off()

# compute lift using prediction() from ROCR and plot lift chart
bankwork_prediction <-
    prediction(bankwork$Predict_Prob_Response, bankwork$response)
bankwork_lift <- performance(bankwork_prediction , "lift", "rpp")
pdf(file = "fig_targeting_customers_lift_chart.pdf",
    width = 8.5, height = 8.5)
plot(bankwork_lift,
col = "blue", lty = "solid", main = "", lwd = 2,
    xlab = paste("Proportion of Clients Ordered by Probability",
    " to Subscribe\n(from highest to lowest)", sep = ""),
    ylab = "Lift over Baseline Subscription Rate")
dev.off()

# direct calculation of lift
baseline_response_rate <-
    as.numeric(table(bankwork$response)[2])/nrow(bankwork)
prediction_deciles <- quantile(bankwork$Predict_Prob_Response,
    probs = seq(0, 1, 0.10), na.rm = FALSE)
# reverse the deciles from highest to lowest
reordered_probability_deciles <- rev(as.numeric(prediction_deciles))
lift_values <- reordered_probability_deciles / baseline_response_rate
cat("\nLift Chart Values by Decile:", lift_values, "\n")

# Suggestions for the student:
# Try alternative methods of classification, such as neural networks,
# support vector machines, and random forests. Compare the performance
# of these methods against logistic regression. Use alternative methods
# of comparison, including area under the ROC curve.
# Ensure that the evaluation is carried out using a training-and-test
# regimen, perhaps utilizing multifold cross-validation.
# Check out the R package cvTools for doing this work.
# Examine the importance of individual explanatory variables
# in identifying targets. This may be done by looking at tests of
# statistical significance, classification trees, or random-forests-
# based importance assessment.
```