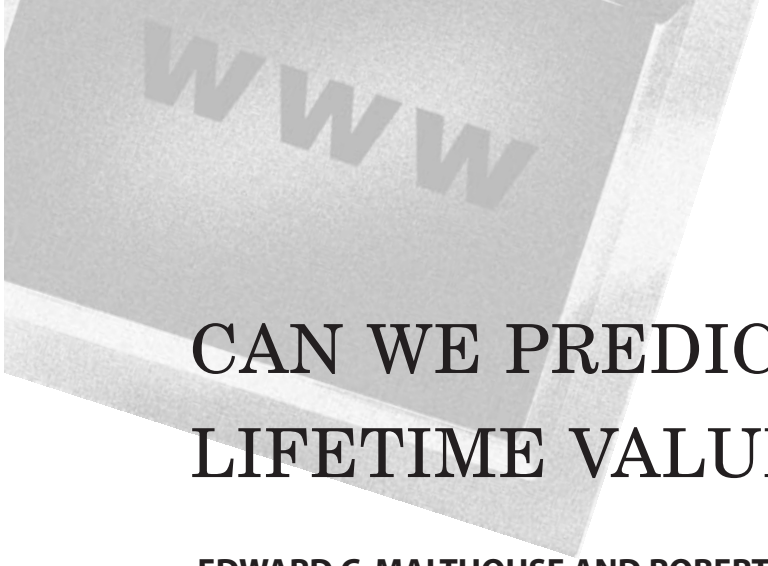

CAN WE PREDICT CUSTOMER LIFETIME VALUE?

EDWARD C. MALTHOUSE AND ROBERT C. BLATTBERG

EDWARD C. MALTHOUSE
is an Associate Professor, Integrated Marketing Communications, Medill School of Journalism, Northwestern University, Evanston, IL; e-mail: ecm@northwestern.edu

ROBERT C. BLATTBERG
is Polk Bros. Distinguished Professor of Retailing, Kellogg School of Management, Northwestern University

The authors are grateful to Karsten Hansen and Kay Peters for helpful discussions, and several anonymous reviewers for helpful comments. They also thank Experian for the Z-24 data set.

Relationship marketing assumes that firms can be more profitable if they identify the most profitable customers and invest disproportionate marketing resources in them. While intuitive, such strategies presume that a firm can accurately predict the *future* profitability of customers. In particular, we argue that the feasibility of such strategies depends on the *probabilities* and *costs* of misclassifying customers. This paper presents a detailed empirical evaluation of how accurately the future profitability of customers can be estimated. We evaluate a firm's ability to estimate the future value of customers using four data sets from different industries. Out-of-sample estimates of predictive accuracy are provided. We examine (1) the accuracy of predictions, (2) how accuracy depends on the length of time over which estimates are made, and (3) the predictors of the firm's best customers. We propose the 20–55 and 80–15 rules. Of the top 20%, approximately 55% will be misclassified (and not receive special treatment). Of the future bottom 80%, approximately 15% will be misclassified (and receive special treatment). Thus, a firm cannot assume that high-profit customers in the past will be profitable in the future nor can they assume that historically low-profit will be low-profit customers in the future.

JOURNAL OF INTERACTIVE MARKETING VOLUME 19 / NUMBER 1 / WINTER 2005
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/dir.20027

INTRODUCTION

The long-term value (CLV) of a customer "represents the present value of the expected benefits (e.g., gross margin) less the burdens (e.g., direct costs of servicing and communicating) from customers" (Dwyer, 1997, p. 7). CLV has become central to relationship marketing (e.g., Sheth, Mittal, & Newman, 1999) and customer equity approaches to marketing (e.g., Blattberg, Getz, & Thomas, 2001; Rust, Zeithaml, & Lemon, 2000). "In relationship marketing, relationships with single customers are interpreted as capital assets requiring appropriate management and investment (e.g., Hennig-Thurau & Hansen, 2000, p. 16)." Such approaches to marketing contend that a firm can ultimately be more profitable by evaluating the profitability of customers and then designing marketing programs for its best customers. Disproportionate marketing resources should be allocated to retaining best customers and keeping them loyal. This strategy would seem to make obvious sense, since it is common for a small percentage of customers to account for a large percentage of revenues and profits (Mulhern, 1999).

Using CLV or predictors of CLV (e.g., historical purchasing behavior) to allocate marketing resources assumes that the *future* value of a customer can be estimated accurately. This assumption is rarely discussed and there is little empirical evidence evaluating it. The accuracy with which the future value of a customer can be predicted falls along a continuum. One extreme is where future behavior can be predicted perfectly given the customer's past behavior and the firm's marketing actions (in regression terms this would correspond to $R^2 = 1$). The other extreme is where the future behavior of customers is independent of their past behavior and the firm's marketing actions (in regression this would correspond to $R^2 = 0$). As Mulhern (1999, p. 28) notes, "models incorporating predicted future purchases are subject to a great deal of forecasting error," but he does not quantify how much forecasting error.

The firm considering whether or not to practice such relationship marketing and customer equity strategies must understand where it falls along this continuum. Investing disproportionate resources in specific customers makes unquestionable sense when their future behavior can be predicted perfectly, but no sense when future behavior is unpredictable ($R^2 = 0$). In the latter case, an egalitarian strategy where all customers are treated equally or the quid-pro-quo incentives discussed below should be used.

Suppose a firm offers two levels of treatment: "best-customer" treatment and "normal" treatment. Assuming the firm cannot predict the future behavior of customers perfectly, the firm can misclassify customers in two possible ways. It could misclassify a future normal customer as a future best customer—a false positive using the language of hypothesis testing—or misclassify a future best customer as future normal customer—a false negative. There are costs associated with both types of misclassifications. When a firm makes a false positive misclassification it is spending scarce marketing resources to deliver best-customer treatment to a future "normal" customer whose behavior does not justify such treatment. It is more difficult to quantify the costs of a false negative. The customer who deserves best-customer treatment but receives normal treatment could switch part or all of its future expenditures to a competitor, spread negative word of mouth, etc. Whether or not a firm should make disproportionate marketing investments across customers depends on the probabilities and costs of misclassifying customers. The costs of misclassification have not be quantified in either the literature or by business practitioners, to our knowledge.

Some examples illustrate our point. An executive who has been using a credit card to spend a large amount of money on expensive clothing, airline tickets, car rentals, hotel rooms, cellular phone service, etc. may retire and spend far less in these categories. This executive goes from being a "best customer" of the companies that provide these products or services to a non-best customer. Showering this executive with discretionary marketing investments after retirement may not be an optimal strategy. This is an example of a false positive. Alternatively, someone who is not so valuable today can, for example, take a new job and become a star customer tomorrow—a false negative.

In using historical information to allocate marketing investments a firm may be relying on chance purchases. There will always be a certain level of randomness in a customer's purchases. Are the customers who receive special treatment really better customers? Or, are they customers who just happened to be "better" during some recent period and will "regress" back to

their true, non-best-customer behavior in the future? For example, a consultant who is normally an occasional flyer on some airline may be assigned to a job in the airline's hub city. The consultant may fly on the airline every week during the job, but resume the occasional-flyer status when the job is completed. Giving this consultant special perks will not be a good strategy.

This paper provides a direct evaluation of how accurately the future behavior of customers can be estimated. The focus of this paper is on companies that maintain databases of customer/end-user information on a substantial percentage of customers and that can customize marketing "investments," at least to some extent, across customers. Such companies include hotels, airlines, credit card companies, banks and financial service providers, companies that sell over the internet, telecommunications companies, catalogers, retail stores with "loyalty/frequent-shopper" programs, publishers, computer companies that sell direct to consumers, and many more. We shall refer to such companies as *database marketing companies*. The discussion here is not as applicable to organizations that do not know their specific end-users, e.g., most producers of consumer package goods.

TYPES OF MARKETING INVESTMENTS

Day (2000) discusses different types of exchanges between customers and companies. Value-adding exchanges involve "giving continuing incentives for the customer to concentrate most of their purchases with them . . . Some customers are more equal than others when it comes to deciding how close a relationship will be formed (p. 25)." In some industries, this is accomplished through a loyalty program, which is "designed to build customer loyalty by providing incentives to profitable customers (Yi & Jeon, 2003, p. 230)." We introduce a distinction between types of value-adding exchanges.

Our thesis, that disproportionate marketing investments should depend on the firm's ability to forecast future profits and the costs of misclassification, has varying levels of relevance for different types of marketing investments. Our thesis is most applicable to marketing investments *without* quid-pro-quo terms. The firm has discretion over which customers will receive these investments and how much it invests in individual customers. We call these *discretionary marketing investments*. There are numerous examples. Direct communication with customers is usually discretionary. Catalog companies decide how many catalogs each customer receives over some period of time. More generally, any organization using direct mail decides the number of contacts to make with each customer, and thus the level of investment. Firms can also customize investments for in-bound communication. Day (2000, p. 25) describes how Hertz has a dedicated phone line for preferred customers so that they do not have to wait so long to make reservations. Likewise, some credit card companies use caller ID to route incoming calls from best customers to shorter phone queues.

Communication is not the only form of discretionary marketing investment. Collinger (2002, p. 32) defines *surprises and delights* as "the unexpected and unpromised benefits that enhance the product or service." Credit card companies waive late-payment fees of certain customers and banks waive checking overdraft fees of some customers. Hotels might unexpectedly leave a bouquet of flowers, bottle of wine, or some other gift in the room of a best customer. Hotels will occasionally upgrade best customers to a larger room. Airlines might give best customers priority for upgrades and have even delayed a flight so that some very important passenger could make a connection. Airlines offer shorter check-in queues for their very best customers. Some catalog companies send an unexpected holiday gift to their best customers. The concept of *customer delight* (Rust & Oliver, 2000), which refers to "a profoundly positive emotional state generally resulting from having one's expectations exceeded to a surprising degree (p. 86)," is closely related. A company that spends resources to customize the product itself without full compensation from the buyer is also making a discretionary investment.

Discretionary investments are often extras or perks, intended to cause the recipient to have positive affect towards the company. In a different context, Geyskens, Steenkamp, Scheer, and Kumar (1996) discusses *affective* commitment. "An affectively committed channel member *desires* to continue its relationship because it likes the partner and enjoys the partnership . . . It experiences a sense of loyalty and belongingness (p. 304)."

Our thesis is less relevant to marketing investments having explicit quid-pro-quo terms. The company and buyer explicitly agree on the terms of such "investments" at the time of purchase, although they are not part of the product/service being purchases. To a large extent, loyalty/reward programs fall into this category. For these programs, many, if not all, benefits that end-users receive and investments that firms make are on explicit quid-pro-quo terms. For example, most frequent flyer programs offer explicit rewards/incentives such as "if I fly X miles/trips, I get a free flight." Hotel programs usually have explicit terms such as "if I stay X nights, I get Y." Credit card programs typically offer explicit rewards such as miles or cash-back bonuses for usage; "for every dollar I spend on this card I get X." Promotions such as "buy two get one free" and negotiated price breaks to high-volume customers are of the same ilk. Such programs, at one level, attempt to increase share of wallet and/or consumption. The firm *rewards* the buyer for behaving in a certain way, usually involving multiple purchases over time. The free flight is an explicit *incentive* for the buyer to fly often with a particular airline. Loyalty programs that offer proportionally larger rewards to best customers such as an airline that awards best customers with 1.5 times the actual mileage flown is practicing a hybrid between discretionary and quid-pro-quo.

Our thesis is less central to quid-pro-quo investments because (1) the investments are available to all customers and (2) the firm is not directly choosing to "invest" more in one customer than another. Any customer can join the program, get the two-for-one special, or the free flight—customers self-select into the programs. The customer's behavior directly determines the level of awards; e.g., a customer who flies more will get more free tickets. The important question with these investments is whether the customer would consume at the same level without the reward/incentive.

METHODOLOGY

As stated above our objective here is to estimate the *future* CLV for individuals or households using past purchase behavior and other available information. For a review of CLV models see Jain and Singh (2002). To evaluate the accuracy of estimates of future CLV, we use a study design that "turns back the clock." The process is illustrated in Figure 1. Assume a long time series of contributions and expenses are available for a sample of customers. For example, we might have data from January 1, 1994 until December 31, 2000. Pretend that "now" is some moment in the past such as the beginning of January 1, 1997. The universe of customers will be those who were on file as of "now," January 1, 1997. The objective of our analysis is to predict the discounted value of a customer from January 1, 1997 through 2000, hereafter called the *target period*, using information from the period 1994–1996, hereafter called the *base period*. The length of the target and base periods will be denoted by T and B, respectively. It will be convenient to think of having $T + B$ discrete time periods. Similar designs are commonly used for direct marketing scoring models, where the target is some measure of response to an offer.

The central empirical question addressed here is whether a customer's value can be estimated over some long period of time rather than a customer's entire lifetime. Firms periodically evaluate a customer's value and adjust marketing investments accordingly. Hotels and airlines, for example, evaluate customer tier membership every calendar year. Since these adjustments occur periodically, the firm should be primarily interested in estimating *long-term*, rather than *lifetime* value. The approach used here does exactly this in a direct way.

Statistical Modeling

Assume a sample of n customers on file as of "now" and measurements of the contributions each

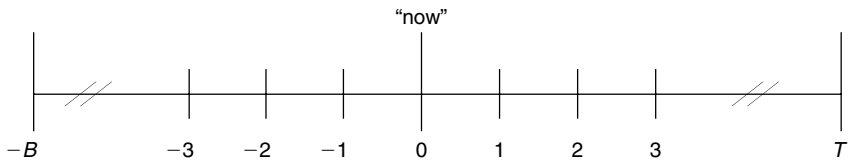

FIGURE 1
Illustration of Target and Base Periods for CLV Estimation

customer makes during the target period. Let c_{it} denote the net contribution of customer i during time period t. By net contribution we mean some appropriate measure of profit attributable to a transaction without consideration of fixed costs, e.g., gross sales less cost of goods sold, direct marketing costs, order processing, return processing, etc. We also have p measurements, $\mathbf{x}_i$, on customer i that are known as of "now," time 0, aggregated from base-period information. Denote the discount rate by d. The CLV of customer i is $y_i = \sum_{t=1}^{T} c_{it}(1 + d)^{-t}$.

CLV is related to the predictor variables with some "regression" function f

$$g(y_i) = f(\mathbf{x}_i) + e_i,$$

where e_i are independent random variables with mean 0 and (ideally) homoscedastic error variance $V(e_i) = \sigma^2$. Because the dependent variable (y_i) is an amount, its variance often increases with its mean, violating the assumption of homoscedasticity. Invertible function g is a variance stabilizing transformation (e.g., Carroll & Ruppert, 1988) such as the logarithm or square root, usually assumed to be known prior to the estimation of f.

We consider three regression methods for f in this paper. The first is a linear regression with variance stabilizing transformations estimated with ordinary least squares. Residual plots inform our selection of a variance stabilizing transformation (Cook & Weisberg, 1982), which we select from the Box-Cox family (Neter, Kutner, Nachtsheim, & Wasserman, 1996, p. 132). Further discussion of the variance stabilizing transformation is given in the empirical results section. To address possible nonlinearities (e.g., diminishing marginal returns to scale) with predictor variables that are amounts or counts, we compute square root and logarithm "first-aid" transformations (Mosteller & Tukey, 1977, p. 109). The influence of outliers of untransformed count and amount variables is reduced with 1% Winsorization, i.e., values greater than the 99th percentile are set equal to the 99th percentile.

The second regression method is linear regression estimated with iteratively re-weighted least squares (IRLS), as described in Neter et al. (1996, pp. 403–405). IRLS is another way of addressing the problem of heteroscedasticity. We use the same predictor variables as in the OLS model. When the dependent variable (CLV) is highly right skewed, we apply the logarithm transformation to symmetrize its distribution, increase the density of observations in the right tail, and reduce the influence of outliers. We implement IRLS by initially estimating the model with OLS. Next, we estimate the absolute value of the residuals (following the recommendation of Neter et al. after equation 10.15) using the same predictor variables. We then re-estimate the original regression equation using weighted least squares, with the reciprocal of the squared residual estimates as weights (equation 10.16a in Neter et al., 1996). We iterate between estimating residuals (and thus weights) and CLV.

The estimates from the IRLS regression model are also estimates of the following random coefficient model, which accounts for unobserved heterogeneity:

$$\begin{aligned} y_i &= (\alpha + a_i) + \sum_j (\beta_j + b_{ij}) x_{ij} + e_i \\ &= \alpha + \sum_j \beta_j x_{ij} + \left(a_i + \sum_j b_{ij} x_{ij} + e_i \right), \end{aligned}$$

where a_i is a random variable with mean 0 and standard deviation σ_a, b_{ij} is a random variable with mean 0 and standard deviation σ_b, and e_i is a random variable with mean 0 and standard deviation σ. We assume that a_i and b_{ij} are independent of e_i. We group all of the random components into a single error term. We do not give separate estimates of the variance components, because comparing the variance components is not relevant to the thesis of this paper.

The third method is a feedforward neural network (Venables & Ripley, 1999, section 9.4), estimated using S-Plus Version 6.0.2. The conclusions we make in this paper depend on the predictive accuracy of the regression model. Neural networks are universal approximators and thus provide a bound for predictive accuracy. They can uniformly approximate any continuous function over compact sets (Ripley, 1996, section 5.7).

Estimating Predictive Accuracy

All examples give out-of-sample estimates of predictive accuracy. Evaluating predictive accuracy using the same data that were used to estimate f is problematic because the estimated model could capture sampling idiosyncrasies of the data set. The problem of making "honest" estimates (e.g., ones that are not

subject to overfitting) has been thoroughly studied (e.g., see Efron & Tibshirani, 1993, Ch. 17; Ripley, 1996, sections 2.6–7). Two common ways of making out-of-sample estimates of predictive accuracy are to use holdout samples and k-fold cross validation. When data are plentiful, a good solution is to use an independent data set to evaluate f. Prior to estimation we partition the available data into *estimation* and *holdout* samples of roughly equal size. The estimation sample is used to estimate the free parameters of the model while the holdout sample is "kept in a locked safe where it has rested untouched and unscanned during all the choices and optimizations" (Mosteller & Tukey, 1977, p. 38) involved in estimating f. The estimated model is then applied to the holdout sample and summaries of predictive accuracy are computed.

When data are scarce, k-fold cross validation is a good way to get an "honest" estimate of prediction accuracy (Efron & Tibshirani, 1993, p. 240). For the smaller data sets we use 10-fold cross validation. We assign each observation randomly to one of 10 groups and estimate the model 10 times. First, we estimate the model using all but the first group, and then apply the estimated model to the first group. Second, we estimate the model using all but the second group and apply the estimated model to this group, etc. Estimates of predictive accuracy are computed on the left-out groups.

EMPIRICAL RESULTS

We examine how accurately CLV can be predicted with four case studies from organizations. None of the companies offered "special treatment" to any of its customers during the study periods. The objective here is to evaluate the best tools that companies currently have to predict CLV. Two of the data sets are available to other researchers; SAS code related to these examples is available from the first author's Web site. To calculate CLV we used an annual discount rate of 15% ($d = .15$). We also have data from 131 catalog companies, which though less extensive, will be used to confirm some of the results from this study. To compute CLV we use a discount rate of $d = 15\%$, consistent with Reinartz and Kumar (2000, p. 23).

Description of Organizations

Service Company. We have a simple random sample of 150,000 customers from a company that offers its members a single service. Customers enroll in the service by signing a one-, six-, or 12-month contract, where cancellation is not allowed. Lapsed customers sometimes re-enroll in the service during a later time period. We have five years of membership history and the date of the first purchase (for those who were customers prior to the five-year period). For each customer and month, we have the following information. First, we have the length of the current contract (0, 1, 6, or 12 months), where 0 indicates that the customer was not a member during a particular month. Second, we have a measure of the quantity of involvement during the month, defined as the number of times that a customer uses the service varies. A customer who likes and enjoys the service will use it more often. The monthly charge is the same, regardless of the level of usage. Third, we have a measure of quality of involvement. Think of the service as providing a lesson. The quality measure indicates how well the customer is learning the lesson. This data set is interesting because of its simplicity; it offers only a single product line and price and has no outliers. We expect it to provide an upper bound for predictive accuracy.

The first two years of data constitute the base period and the last three years the target period. The universe consists of the 71,381 customers who were active at least one month during the base period. The 71,381 observations were randomly split into estimation and holdout samples of roughly equal size. The predictor variables were recency, frequency, the involvement quantity measure averaged over all months during the base period, the involvement quality measure averaged over the base period, dummies for different contract types, and longevity (months on file). All variables were examined for outliers and high skewness.

Not-for-Profit Organization. Each year the Knowledge Discovery and Data mining Special Interest Group (SIG-KDD) of the Association for Computing Machinery (ACM) sponsors a data mining competition. SIG-KDD provides contestants with a data set and a data-mining task. In 1998, the data set was from a not-for-profit organization and the task was to determine which "one-year lapsed donors" should be sent a solicitation during 6/1997.[1] A one-year lapsed donor was one who had not responded to any solicitations since 6/1996.

[1] See http://kdd.ics.uci.edu.

KDD98 contains the promotion and donation history for the two years prior to the lapsed period (6/1994–6/1996). KDD98 contains all 191,779 one-year lapsed donors.

There are several features of this data set that make it interesting. First, the data have extensive overlays at both the household and five-digit zip code level. These overlay variables are often the only information a company or organization has on prospective donors/customers. This data set will allow us to evaluate the predictive power of various levels of overlay variables vis-à-vis behavioral variables (e.g., RFM). Second, this data set is available to the general public so that it can be used for benchmarking and comparing methods. If other researchers develop alternative methods to the one proposed here, the performances can be compared directly on this data set. Third, there are very large, defined estimation and holdout samples (the documentation calls them "learning" and "validation" sets) of 95,412 and 96,367 donors, respectively. In total there are 481 variables.

We use this data set in a different way than it was used for the data-mining contest. Details of our analysis and SAS code for preparing the data are available from the first author's Web site. Define "now" as June 1, 1994. Our universe of donors is all who were on file before this date, reducing the sizes of the estimation and holdout samples to 68,026 and 68,804, respectively. The organization sent out 22 "card promotions" during 6/94–6/96; the data set contains the date each solicitation was mailed, the date a donation in response to a particular solicitation was received, and the dollar amount. The (discounted) sum of these 22 amounts is the revenue during the target period. The mean is \$37.43, the minimum \$0, the 99th percentile \$141, and the maximum \$8,137. The estimation sample has a 99th percentile of \$142 and a maximum of \$1,686 while the holdout sample has a 99th percentile of \$140 and a maximum of \$8,137. The difference in the maximum values emphasizes the importance of paying close attention to outliers. The dependent variable of our analysis is the square root of revenues less costs. The square root variance stabilizing transformation ($g(y) = y^{1/2}$) is used to reduce heteroscedasticity, make the distribution more symmetric, and reduce the influence of observations in the right tail.

We constructed predictor variables following the examples given in the Direct Marketing Educational Foundation (DMEF) data sets. These data sets have a large number of variables capturing interactions between RFM, product category, and purchase channel. Interactions between recency and the frequency and monetary variables are captured with variables such as orders (or dollars) within the most recent year, orders (dollars) last year, orders (dollars) two years ago, orders (dollars) three years ago. Interactions between frequency and monetary are captured by dividing monetary by frequency giving "average order amounts." Interactions between purchase channel and the frequency and monetary variables are captured with variable such as orders (or dollars) from category A, orders (dollars) from category B, etc. Likewise for purchase channel.

Business-to-Business Company. We have a simple random sample of 100,000 "small-business" customers of a large company. For each of these customers we have the transaction history over a seven-year period and Dunn and Bradstreet overlays. The transaction file gives the customer ID, date, price, quantity, and SKU of every transaction. SKUs are categorized into five main product lines, and several other small ones accounting for a very small percentage of transactions and dollars.

This data set is interesting for several reasons. First, it is a very complicated data set, with many SKUs, multiple product lines, strong seasonal buying patterns, a large number of extreme outliers, and multiple delivery channels. This empirical study thus spans a wide range of CLV situations from simple (service company) to complex (this company). Second, many, but not all, of the products have long inter-purchase times. If a customer buys one of these products today, the customer will not need to buy the product again for several years, unless the business expands. Between purchases of one of these products, a customer may buy from other product lines, or complementary SKUs from the same product line. Third, prior to acquiring a customer, often companies know only the information contained in the Dunn and Bradstreet overlays about prospects. This data set will allow us to evaluate the predictive power of such overlays, compared with behavioral data such as RFM. Fourth, this is an unusually long time series. Many companies would not be able assemble information at this level of detail from seven years ago.

For this evaluation we use the first two years as the base period and the last five years as the target period. We have tried other splits and found similar conclusions, e.g., three-year base and four-year target, four-year base and three-year target, etc. The universe for the analysis described here is all customers "on file" as of the beginning of year 3, giving a sample of size 24,047. Using the transaction file, we computed 61 variables from base-period transactions including RFM variable overall and by product category. We computed frequency in terms of items and orders (one order can contain multiple items), and monetary value during the most recent year and the most recent two years. We also computed square root and logarithm transformations for variables where we expected diminishing marginal returns.

Catalog Company. The Direct Marketing Educational Foundation (DMEF) has made available four real data sets for academic research and teaching. We use the "DMEF3" data set here, which is from a long-time specialty catalog company that mails both full-line and seasonal catalogs to its customer base. The data set is a random sample 106,284 customers who have bought before from the company and were being considered for a mailing in Fall, 1995. The data set has 12 years of purchase history through July 31, 1995.

The DMEF3 data set is interesting for several reasons. First, it is from a retail consumer catalog company, an industry not represented by the other detailed data sets in this paper. Second, it contains an exceptionally long time series: 12 years. Third, it is available to all researchers from the DMEF.

For this analysis, we define "now" as August 1, 1990 and select as our universe all customers who were on file before this date. This gives a sample of 41,669 customers, with six-year base and target periods. These observations are randomly assigned to estimation and holdout samples of approximately equal size. We are able to construct RFM variables, time on file, first purchase amount, and indicators of product classes and sub-classes. We Winsorized (1%) and applied the square root transformation to all amount and count variables.

Experian Z-24 Catalog Data. The Z-24 database, which is owned by Experian, allows catalog companies to exchange mailing lists. Hundreds of catalog companies periodically provide their mailing lists to Experian along with RFM in exchange for names from other lists to be used in prospecting for new customers. We have a random sample of 1 million households from this database with RFM information as of January 1, 2001, 2002, and 2003. For this analysis, we take "now" to be 1/1/2001, the target period to be the two years 1/1/2001–1/1/2003, and the base period to be the time prior through 1/1/2001. We analyze companies for which we have at least 3,000 households in our sample; using this criterion we have 131 companies. Sample sizes range from 3,005 to 94,523 for individual companies. We model the logarithm of CLV as a linear function of the logarithms of RFM.

This data set is interesting because it allows us to study the variation *across* catalog companies in how the RFM variables affect CLV and how accurately CLV can be predicted from RFM. By having data from 131 distinct catalog companies, which we consider representative of all catalogs, we can make strong statements concerning the generalizability of our conclusions.

Marketing Cost Data. We do not have variables measuring the level of marketing investment for three of the four organizations or for the Z-24 catalog companies. The service organization gives all of its customers the same contacts, so marketing investment is irrelevant for this company. While it is desirable to have such information, our impression is that very few companies currently keep it. The goal of this paper is to evaluate what companies are doing today and we can achieve this goal with the present data. If companies tracked marketing contacts and they could be included in the model, perhaps our conclusions would change, but our conclusions apply to what companies are currently doing.

Accuracy of Predictions

We developed regression and neural network models for each of the data sets. We also estimate IRLS for the business-to-business company. The final model was the one that gave the best fit, measured by R^2 computed on the holdout sample. Fit is the criterion suggested in the data-mining literature (e.g., Breiman, 2001, p. 204, 205, 229; Breiman, 1996; Hastie, et al., 2001, section 2.9, ch. 7) for problems where the primary objective is making predictions that are as accurate as possible, as it is here.

Once a final model has been selected, we evaluate its predictive accuracy in two ways. The first is the familiar coefficient of determination (R^2). The second comes from a classification table. Part of the goal of CLV is to separate "best" customers from others. For simplicity, we assume that the top 20% based on *actual* CLV values in the *target period* are "best" customers. We use the estimated regression models to rank customers from best to worst. The 20% with the largest predicted values are assigned "best-customer" status and would receive perks. The classification table is a cross tabulation of actual group versus predicted group.

Table 1 gives an example cross-tabulation for the Service Company. The false positive and false negative rates give us more details about the accuracy of the predictions. These terms are usually applied to hypothesis tests. Define the null hypothesis H_0 to be that a customer is part of the (actual) bottom 80% and does not deserve special treatment. The alternative hypothesis, H_1 is that a customer is in the top 20% and "deserves special treatment." The *false positive rate* is $P(\text{Reject } H_0 \mid H_0 \text{ True}) = 3{,}832/28{,}615 = 13.4\%$. Of the customers who do not deserve special treatment, 13.4% would receive it if this model were used. The *false negative rate* is $P(\text{Do not reject } H_0 \mid H_1) = 3{,}891/7{,}154 = 54.4\%$. Of the (actual) best customers in the future, 54.4% would not be identified by this model. The *power* of the model is $P(\text{Reject } H_0 \mid H_1) = 45.6\%$.

TABLE 1 Classification Tables for the Service Company Predicting 20–80 Group for 3-Year CLV

	ACTUAL		
PREDICTED	**BOTTOM 80**	**TOP 20**	**TOTAL**
Bottom 80	24,783	3,891	28,674
Col Pct	86.6%	54.4%	80.2%
Top 20	3,832	3,263	7,095
Col Pct	13.4%	45.6%	19.8%
Total	28,615	7,154	35,769
Row Pct	80.0%	20.0%	100.0%

We estimated models for each of the data sets and also varied the length of the future time horizon. The false positive and negative rates for the estimation and holdout samples are summarized in Table 2.

TABLE 2 Measures of Predictive Accuracy

LENGTH		FALSE NEGATIVE		FALSE POSITIVE		*R*-SQUARED	
COMPANY	**FUTURE (T)**	**ESTIMATION**	**HOLDOUT**	**ESTIMATION**	**HOLDOUT**	**ESTIMATION**	**HOLDOUT**
Service	1 year	0.2482	0.2466	0.3021	0.3063	0.4800	0.4850
Service	2 years	0.1632	0.1625	0.4641	0.4678	0.3861	0.3870
Service	3 years	0.1339	0.1366	0.5439	0.5448	0.3362	0.3374
Nonprofit	2 years	0.1431	0.1443	0.5722	0.5774	0.1303	0.1332
B2B	1 year	0.1326	0.1345	0.5305	0.5379	0.2812	0.2868
B2B	2 years	0.1261	0.1288	0.5050	0.5151	0.3135	0.3104
B2B	3 years	0.1252	0.1281	0.5008	0.5126	0.3200	0.3126
B2B	4 years	0.1233	0.1254	0.4933	0.5019	0.3538	0.3375
B2B	5 years	0.1213	0.1268	0.4854	0.5073	0.3617	0.3442
Catalog	1 year	0.1386	0.1394	0.5545	0.5574	0.2077	0.1983
Catalog	2 years	0.1339	0.1349	0.5356	0.5395	0.2326	0.2205
Catalog	3 years	0.1322	0.1366	0.5289	0.5442	0.2452	0.2249
Catalog	4 years	0.1310	0.1345	0.5244	0.538	0.2465	0.2198
Catalog	5 years	0.1306	0.135	0.5226	0.5401	0.2423	0.2175
Catalog	6 years	0.1336	0.1366	0.5345	0.5465	0.2402	0.2152

The third row of the table summarizes the service company example discussed above from Table 1. The fifth row, "B2B 1 year," gives summarizes the misclassifications for the one-year predictions of the business-to-business company.

20–55 and 80–15 Rules

It is striking how similar the results are across the data sets. With the exception of the one- and two-year estimates for the service company, the false negative rates are all approximately 51–55% and the false positive rates are all approximately 13–15%. We posit two new empirical rules of thumb based on these results.

The 20–55 Rule. Of the *actual* best customers (top 20%), approximately 55% will be misclassified (and not receive special treatment).

The 80–15 Rule. Of the actual normal customers (bottom 80%), 15% will be misclassified (and receive special treatment).

We evaluate whether these rules generalize across catalog companies with the Z-24 data. Figure 2 shows the predictive accuracy of 131 regression models with boxplots. We estimated separate multiple regression models for each of the companies. The distribution of R^2 values is concentrated among small values, with $R^2 < 17.5\%$ for three-fourths of the catalogs. There are outliers in R^2 values, indicating that a few companies can predict the future values of customers with greater accuracy, but these are exceptions. We conclude that most companies will not be able to predict the future behavior of customers accurately, as measured by high R^2 values.

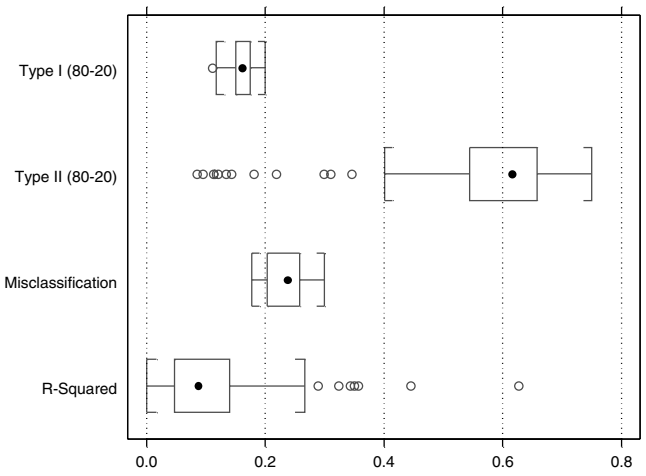

FIGURE 2

Boxplots Showing Predictive Accuracy for 131 Catalog Companies From Z24 Database

The top boxplot shows the distribution of Type I error rates across the catalogs. The median Type I error rate (dot in middle of box) is 16.1%, which is approximately equal to the 15% posited by our rule. The lower quartile is 15.0% (left end of box) and the upper quartile is 17.5% (right end of box), so half of these catalog companies have Type I error rates between 15% and 17.5%. The range extends from 11.1% to 20.0%. The 80–15 rules thus holds fairly consistently across catalog companies.

There is more variation in Type II error rates across companies. The mean is 57%, the median is 62%, and the quartiles are 54% and 66%. There are several outliers in the left tail, indicating that some exceptional companies have substantially lower Type II error rates. Thus, the 20–55 rule appears to hold "on average" for catalogs, although there is more variation across companies and some exceptions.

SENSITIVITY TO VARIANCE STABILIZING TRANSFORMATION AND METHOD OF ESTIMATION

Using the Business-to-Business data set, we evaluate whether our conclusions change when different variance stabilizing transformations are used or when the model is estimated with IRLS. Box-Cox transformations have the form $g(y) = (y + a)^p$. When $p = 0$, the logarithm transformation is used. Constant a is added to every value to avoid, for example, taking the logarithm or inverse ($p = -1$) of 0. Table 3 gives the results using a $T = 1$ year future period and $T = 9$ year future period. We estimate the same model with the following transformations: none ($p = 1,\ a = 0$), square root ($p = 1/2, a = 0$), cube root ($p = 1/3, a = 0$), fourth root ($p = 1/4,\ a = 0$), and logarithm ($p = 0, a = 1$). The "For 6" row gives the results when forward selection is used and only the first six variables are allowed to enter. The IRLS row gives the results when iteratively re-weighted least squares is used to estimate the model parameters, as described

TABLE 3 Performance and Fit Measures for the Business-to-Business Company Using Different Variance Stabilizing Transformations and Estimation Methods

			FALSE NEGATIVE		FALSE POSITIVE		*R*-SQUARE	
TRANS	**EST**	***x***	**TRAIN**	**TEST**	**TRAIN**	**TEST**	**TRAIN**	**TEST**
				***T* = 1 Year Future Period**				
None	OLS	All	0.1341	0.1327	0.5363	0.5309	0.6714	0.5973
$\sqrt{}$	OLS	All	0.1274	0.1301	0.5096	0.5205	0.5371	0.5193
$3\sqrt{}$	OLS	All	0.1299	0.1336	0.5196	0.5346	0.3659	0.3668
$4\sqrt{}$	OLS	All	0.1326	0.1345	0.5305	0.5379	0.2812	0.2868
Log	OLS	All	0.1346	0.1364	0.5384	0.5458	0.1971	0.2052
Log	OLS	For 6	0.1361	0.1358	0.5442	0.5433	0.1837	0.1983
Log	IRLS	All	0.1349	0.1377	0.5396	0.5508	0.1946	0.2022
				***T* = 5 Year Future Period**				
None	OLS	All	0.1219	0.1267	0.4875	0.5068	0.7583	0.6692
$\sqrt{}$	OLS	All	0.1172	0.1221	0.5313	0.4886	0.6281	0.5866
$3\sqrt{}$	OLS	All	0.1189	0.1244	0.4758	0.4977	0.4704	0.4452
$4\sqrt{}$	OLS	All	0.1213	0.1268	0.4854	0.5073	0.3617	0.3442
Log	OLS	All	0.1335	0.1389	0.5338	0.5557	0.2401	0.247
Log	OLS	For 6	0.1404	0.1423	0.5618	0.5694	0.2193	0.2366
Log	IRLS	All	0.1340	0.1403	0.5359	0.5611	0.2384	0.2450

above. Table 4 gives the parameter estimates for the six-variable model. The variables were selected using stepwise selection. There is a healthy mixture of RFM variables in the models.

First, note the similarity between the false negative rates, which are all within less than a percent of each other. For a $T = 1$ year future, the test-set values vary between 13.01% with a square-root transformation to 13.77% with the IRLS estimate. Our rounding to the "80–15" rule introduces more error than the variance stabilizing transformation or method of estimation. The same is true for the false positive rates, in that they range from 52.05% to 55.08%. Predictive accuracy does not change much across these models. The R-squared values are not comparable across these models because the variance stabilizing transformation changes the denominator $\text{SST} = \sum(y_i - \bar{y})^2$ (e.g., see Scott & Wild, 1991). We report their values to highlight the importance of paying close attention to outliers. There is a large difference between the training and test set values when no transformation is used (0.6714 versus 0.5973). The differences are smaller when the influence of outliers is reduced. Outliers and the long right tail inflate the value of SST for the identity and other weak transformations, but the influence of these values are reduced for, e.g., the logarithm. Extreme outliers can exert a strong influence on the estimation even when the sample sizes are large.

It is not surprising that the method of addressing heteroscedasticity does not matter. Having heteroscedastic error variance implies that the OLS estimates are no longer the best linear unbiased estimates (BLUE). Estimates from heteroscedastic data are still unbiased, but do not have the lowest variance across all unbiased estimates. The variance of a slope estimate, however, is also a function of the sample size used to estimate the model. Finding the transformation that gets closest to homoscedasticity will have more effect on the variance of the slope estimates when the sample size is small, but when the sample size is very large—as it is here—the variance

TABLE 4 Parameter Estimates for the 6-Variable Forward-Selection Models ("For 6") Using the Business-to-Business Data ($n = 11{,}979$)

VARIABLE	ESTIMATE	STD ERR	*T* VALUE
	***T* = 1 Year Future Period**		
Intercept	−0.1946	0.1639	−1.19
√(number orders)	0.9236	0.0580	15.95
√(dollars most recent year)	0.0397	0.0030	13.43
log(dollars product line 2)	−0.1802	0.0093	−19.33
log(total dollars)	0.3148	0.0366	8.61
√(total dollars product line 3)	−0.0332	0.0041	−8.04
Dollars most recent year	−0.0001	0.00001	−8.25
	***T* = 5 Year Future Period**		
Intercept	2.8419	0.04262	66.66
√(number items purchased)	0.1970	0.0012	19.75
√(dollars most recent year)	0.0233	0.0012	19.75
Indicator first order product line 3	−0.3651	0.0535	−6.82
log(total number orders)	0.7543	0.0638	11.82
√(total dollars product line 2)	−0.0284	0.0027	−10.68
√(total dollars product line 3)	−0.0148	0.0021	−6.95

of the estimate will be small regardless of whether the square root or logarithm (or even identity) was used.[2]

MANAGERIAL IMPLICATIONS

The managerial implications of the empirical rules are important. If a company were to start offering special treatment for its best 20% of customers, it would have to reward customers based on their past behavior, since that is all that would be known. The 20–55 rule suggests that such a company would be wrong about 55% of the time in deciding who deserves the perks. It would give perks to the wrong customers. The customer who deserves best-customer treatment but receives normal treatment could switch part or all of its future expenditures to a competitor or spread negative word of mouth.

A false negative can also be a missed opportunity to develop a best customer if the customer were responsive to best-customer interventions. This raises several questions requiring further research. How does a customer being misclassified affect that customer's attitude and commitment towards the company? Is the customer who deserves perks but does not receive them more likely to defect? We conjecture that these issues are particularly problematic when the perks are visible. Duncan and Moriarty (1998, p. 8) note that "everything a company does (and sometimes does not do) sends a message that can strengthen or weaken relationships." When customers know what perks other customers are receiving, we conjecture the misclassified best customer will be more likely to defect and have a more negative attitude towards the company. If there is an

[2] Some justification for these claims can be easily seen from the simple linear regression formulas. Suppose that $y_i = \alpha + \beta x_i + e_i$, where e_i is normal with mean 0 and standard deviation σ_i. Assume also that e_i is independent of e_j for $i \neq j$. Let $S_{xx} = \sum(x_i - \bar{x})^2$. The OLS estimate of β is $b = \sum(x_i - \bar{x})(y_i - \bar{y})/S_{xx}$. It is easy to show that

$$V(b) = \frac{\sum \sigma_i^2 (x_i - \bar{x})^2}{S_{xx}^2}.$$

Note that under homoscedasticity there is cancellation and $V(b) = \sigma^2/S_{xx}$, which is the formula given in textbooks. As the sample size grows, the denominator should become larger and the variance of the slope estimate decreases. As we get more data our estimates become more precise.

interaction between the perk and customer status (e.g., the perk works better on best customers), the firm also loses the additional revenue due to the interaction.

The managerial implications of the 80–15 rule include the fact that the company is spending scarce marketing resources on the wrong customers. It also highlights that best customers do not remain best customers forever. Reinartz and Kumar (2002) hint at something similar when they recommend "let butterflies fly." When a company discovers that a former best customer is no longer deserving of perks and stops giving them, does this customer become more likely to defect? Does revoking a perk cause a declining customer to decline faster? We have anecdotal evidence that this is true in the airline industry, where former frequent fliers avoid an airline after their "executive status" has been taken away. If this is true, the lost revenue due to an accelerated decline must be taken into consideration when deciding whether to offer perks. Companies should have a plan in place to keep the loyalty of customers who have had their level of perks lowered.

These ideas can be incorporated into a profit function. Let P_B be the baseline profit from a best customer, i.e., the profit that a best customer would produce without any additional perks. Let P_N be the baseline profit from a normal customer. Let C be the cost of a perk, I_B be the incremental profit generated by giving an actual best customer perks, and I_N be the incremental profit generated by giving an actual normal customer perks. When $I_B \neq I_N$, the perks have a different effect on a best customer than on a normal customer; we conjecture that for most companies $I_B > I_N$. Let C_{II} be the cost of a type II error, i.e., not giving perks to someone who deserves them. Then profit is

$$\begin{aligned} P &= .2 \times .45(P_B + I_B - C) + .2 \times .55(P_B - C_{II}) \\ &\quad + .8 \times .85P_N + .8 \times .15(P_N + I_N - C) \\ &= .2P_B + .09I_B - .21C + .8P_N + .12I_N - .11C_{II} \end{aligned}$$

Giving perks is thus sensible when $.09I_B + .12I_N > .21C + .11C_{II}$. Companies will want to evaluate this function based on their specific costs, benefits, and misclassification probabilities.

DISCUSSION

Neils Bohr wrote "prediction is very difficult, especially about the future." This quote applies to making CLV estimates for the four organizations examined here. Historical value is not a very accurate predictor of future value. In situations where the future cannot be predicted accurately, an organization that invests a disproportionate amount of marketing resources in historically valuable customers may be investing in the wrong customers.

Relationship marketing and customer equity strategies suggest that firms should determine the value of customers and invest disproportionately in better customers. These approaches to marketing should emphasize the importance of the accuracy of value estimates. Our empirical work suggests that if a firm offers its alleged best 20% of customers special treatment, it will frequently misclassify customers. Of the actual top 20%, approximately 55% will be misclassified (and not receive special treatment). Of the actual bottom 80%, 15% will be misclassified (and receive special treatment). Misclassifying customers has potential costs. The best customer who is misclassified as normal could defect to a competitor, develop a negative attitude towards the firm, or not consume as much as it would if given best-customer treatment. The now-normal customer who receives perks is not as deserving as others.

Should organizations invest *discretionary marketing resources* in alleged best customers? The answer depends on the probabilities and costs of misclassifying customers, the additional revenue generated as a result of the special treatment, and the cost of the special treatment itself. In some cases this accuracy could be adequate, while in others it could be inadequate. Our point is that these misclassification rates and costs must be considered. This thought process is not currently emphasized—or even mentioned—by writers and speakers on the subject. Offering premium treatment to a select group of customers may improve that group's CLV, but could it have a negative effect on other customer groups? Does the percentage of true positives increase substantially by offering special treatment? Rust and Oliver (2002, p. 92) ask "What happens if a firm delights the customer in one period and then reverts to the former level of quality?" They label this "hit-and-run delight." If a customer stops receiving

discretionary marketing investments, is the customer more likely to defect to a competitor than if the customer had never received any such investments? These are important research questions that need to be addressed in future research. Future research should also examine how including information about contacts affects predictive accuracy. Management judgment, absent empirical research, may not provide adequate intuition to answer this question.

We have provided evidence that firms will have difficulty predicting future behavior of their customers with much accuracy. One might ask why customers' future behavior is not very predictable? Some reasons have been hypothesized in the relationship marketing literature. As Day (2000, p. 24) notes "a strategy of investing in or building close relationships is neither appropriate nor necessary for every market, customer, or company. Some customers want nothing more than the timely exchange of the product or service with a minimum of hassles. And because close relations are resource intensive, not every customer is worth the effort." Diller (2000, pp. 39–43) suggests classes of "demotivators of loyalty." *Opportunism* means that customers are willing to "take any opportunity to get more value for the money, to be fully flexible when shopping and to only be interested in their own personal benefit (p. 40)." *Variety seeking* is a second reason. *Autonomy* "means freedom from others and decision-making independence (p. 42)." Clearly there are many potential explanations.

Our discussion so far has focused on discretionary marketing investments. Our position on quid-pro-quo investments is different because the amount of a quid-pro-quo investment depends on *actual future behavior*, whereas the discretionary investments are made based on *predicted future behavior*. For example, the customer who actually flies more miles in the future will receive more free flights—the number of free flights is roughly in proportion to future miles flown. The free flight is offered as a carrot to reward desirable future behavior. The important question when deciding to offer carrots is whether the customer would behave in the same way without the carrot. See Humby, Hunt, and Phillips (2003, especially pp. 29 and 215–216) for excellent discussion on this topic and hybrid approaches where better customers are offered proportionally larger carrots than less profitable customers.

Based on reading this paper, we expect that firms would be highly circumspect about their targeting and CRM strategies based on predicted customer value. The 20–55 rule means that treating lower valued customers poorly may cause defectors of potentially future high valued customers.

REFERENCES

Blattberg, R.C., Getz, G., & Thomas, J.S. (2001). Customer Equity: Building and Managing Relationships As Valuable Assets. Boston: Harvard Business School Press.

Breiman, L. (1996). Heuristics of Instability and Stabilization in Model Selection. Annals of Statistics, 24(6), 2350–2383.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. Statistical Science, 16(3), 199–231.

Carroll, R.J., & Ruppert, D. (1988). Transformation and Weighting in Regression. New York: Chapman and Hall.

Collinger, T. (2002). The Tao of Customer Loyalty: Getting to "My Brand, My Way." In D. Iacobucci & B.J. Calder (Eds.), Kellogg on Integrated Marketing (pp. 16–38). Hoboken, NJ: Wiley.

Cook, R.D., & Weisberg, S. (1982). Residuals and Influence in Regression. New York: Chapman and Hall.

Day, G. (2000). Managing Market Relationships. Journal of the Academy of Marketing Science, 28(1), 24–30.

Diller, H. (2000). Customer Loyalty: Fata Morgana or Realistic Goal? Managing Relationships With Customers. In T. Hennig-Thurau & U. Hansen (Eds.), Relationship Marketing: Gaining Competitive Advantage Through Customer Satisfaction and Customer Retention (pp. 29–48). New York: Springer.

Duncan, T., & Moriarty, S. (1998). A Communication-Based Marketing Model for Managing Relationships. Journal of Marketing, 62, 1–13.

Dwyer, F.R. (1997). Customer Lifetime Valuation to Support Marketing Decision Making. Journal of Direct Marketing, 11(4), 6–13.

Efron, B., & Tibshirani, R.J. (1993). An Introduction to the Bootstrap. New York: Chapman and Hall.

Geyskens, J.-B., Steenkamp, E.M., Scheer, L.K., & Kumar, N. (1996). The Effects of Trust and Interdependence on Relationship Commitment: A Trans-Atlantic Study. International Journal of Research in Marketing, 13(4), 303–317.

Hastie, T., Tibshirani, R.J., & Friedman, J.F. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer.

Hennig-Thurau, T., & Hansen, U. (2000). Relationship Marketing—Some Reflections on the State-of-the-Art of the Relational Concept. In T. Hennig-Thurau and

U. Hansen (Eds.), Relationship Marketing: Gaining Competitive Advantage Through Customer Satisfaction and Customer Retention (pp. 3–27). New York: Springer.

Humby, C., Hunt, T., & Phillips, T. (2003). Scoring Points: How Tesco is Winning Customer Loyalty. London: Kogan Press.

Jain, D., & Singh, S.S. (2002). Customer Lifetime Values Research in Marketing: A Review and Future Directions. Journal of Interactive Marketing, 16(2), 34–46.

Mosteller, F., & Tukey, J. (1977). Data Analysis and Regression. New York: Addison-Wesley.

Mulhern, F. (1999). Customer Profitability Analysis: Measurement, Concentration, and Research Directions. Journal of Interactive Marketing, 13(1), 25–40.

Neter, J., Kutner, M., Nachtsheim, C., & Wasserman, W. (1996). Applied Linear Statistical Models (4th ed.). Chicago: Irwin.

Reinartz, W., & Kumar, V. (2000). On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing. Journal of Marketing, 64, 17–35.

Reinartz, W., & Kumar, V. (2002). The Mismanagement of Customer Loyalty. Harvard Business Review, 86–94.

Ripley, B. (1996). Pattern Recognition and Neural Networks. Cambridge: Cambridge University Press.

Rust, R.T., & Oliver, R.L. (2000). Should We Delight the Customer? Journal of the Academy of Marketing Science, 28(1), 86–94.

Rust, R.T., Zeithaml, V.A., & Lemon, K.N. (2000). Driving Customer Equity: How Customer Lifetime Value Is Reshaping Corporate Strategy. New York: Free Press.

Scott, A., & Wild, C. (1991). Transformations and R^2. The American Statistician, 45(2), 127–129.

Sheth, J., Mittal, B., & Newman, B.I. (1999). Customer Behavior: Consumer Behavior and Beyond. Forth Worth: Dryden.

Venables, W., & Ripley, B. (1999). Modern Applied Statistics with S-PLUS. New York: Springer.

Yi, Y., & Jeon, H. (2003). Effects of Loyalty Programs on Value Perception, Program Loyalty, and Brand Loyalty. Journal of the Academy of Marketing Science, 31(3), 229–240.