Size and Share of Customer Wallet

Rex Y. Du1
Assistant Professor of Marketing
Terry College of Business
University of Georgia
Athens, GA, 30602
rexdu@terry.uga.edu
Tel: 706.542.3540

Wagner A. Kamakura
Ford Motor Company Professor of Global Marketing
Fuqua School of Business
Duke University
Durham, NC, 27708
kamakura@duke.edu
Tel: 919.660.7855

Carl F. Mela
Professor of Marketing
Fuqua School of Business
Duke University
Durham, NC, 27708
mela@duke.edu
Tel: 919.660.7767

May 9, 2006

1 This paper is adapted from part of the first author’s dissertation, and he is grateful for Professor Rick Staelin, for his continued guidance and mentoring throughout the development of the dissertation. The authors also wish to thank the support provided by the Teradata Center for Customer Relationship Management at Duke University.
Size and Share of Customer Wallet

Abstract

Many companies collect substantial information regarding their interactions with their customers. Yet information regarding their customers’ transactions with competing firms is often sparse or non-extant. As a result, firms are often compelled to manage customer relationships based on an inward view of their customers. However, our empirical analysis indicates: (i) the volume customers transact within a firm has little correlation with the volume they transact with the firm’s competitors; and (ii) a small percentage of customers account for a large portion of all the external transactions, suggesting considerable potential to increase sales if these customers can be correctly identified and incented to switch. Thus, we argue for a more outward view in customer relationship management and develop a list-augmentation based approach to augment firms’ internal records with insights regarding their customers’ relationships with competing firms, including the size of each customer’s wallet and the firm’s share of it.

Key Words: Customer Relationship Management, Share of Wallet, Share of Category

Requirements, List Augmentation
Size and Share of Customer Wallet

Introduction

Motivation

With continued advances in information technology, firms are capturing an expanding array of detailed records regarding their interactions with their individual customers. Such information has been used to generate not only behavioral insights but also important metrics such as customer lifetime value, customer equity, etc., and is becoming indispensable in guiding firms’ customer relationship management initiatives (Rust, Zeithaml, and Lemon 2000; Blattberg, Getz, and Thomas 2001). Yet this abundance of internal (within the firm) information is often accompanied by dearth of information on external (outside the firm) customer activities. As a result, firms are often compelled to manage customer relationships using a view of their customers that is based mostly on internal records. However, such an inward focus could provide misleading measures of a customer’s market potential. For example, customers who appear to have high value based on internal records may have modest growth potential if they have only limited requirements served by competing firms. In contrast, customers who have a high transactional volume with competing firms may be good targets for growth to the extent one can attract a larger share of their business. Bell et al. (2002) indicate this lack of individual-level, industry-wide consumer data as a primary barrier to customer relationship management. In the absence of such information, customer loyalty programs (Dowling and Uncles 1997), cross- and up-selling applications (Knott, Hayes, and Neslin 2002), targeted promotions (Rossi, McCulloch, and Allenby 1996) and many other marketing efforts face difficulties achieving the best return on investment.
Our study attempts to redress this limitation by developing an approach to determine how much business a customer transacts not only with a focal firm, but also its competitors, or, in industry parlance, to estimate the firm’s share of the customer’s total wallet. Such an approach could prove useful in managing customer relationships. For example, without information pertaining to a customer’s demand from competing firms, it is not possible to discern a customer with high firm share and low category volume from another with low firm share and high category volume. Yet the marketing prescriptions for each differ considerably. For the former customer the marketing prescription might be to generate new primary demand in the category (if possible), while for the latter customer encouraging switching to the firm’s existing products might be more appropriate.

Overview of Our Approach

One way to address the problem of not knowing how much business a customer does with the competition is through a procedure commonly known as “list augmentation” or “database augmentation,” which overlays data obtained from customer surveys or secondary sources with existing databases (e.g., Wyner 2001; Crosby, Johnson, and Quinn 2002; Kamakura and Wedel 2003; Kamakura, Wedel, de Rosa, and Mazzon 2003; Cole 2005). A typical list augmentation exercise involves the following steps. First, the focal firm (often anonymously) surveys a random sample of its customers, collecting information that is not available from the firm’s internal database. The information from the survey is then linked to information already stored in the internal database (e.g., transaction history, demographics, etc.) to form a sample with a complete set of records. Based on this sample, predictive models can be developed leveraging the correlation patterns between the survey results and the internal data. Finally, the

---

2 We later formalize these notions by defining the concept of “total category requirements,” and “share of category requirements,” and “share of wallet.”
best performing model is applied to the remainder of the database, producing individual-level estimates of the survey results for the entire customer base, as shown by Kamakura and Wedel (2003). Relative to the costs of developing and maintaining the customer database, the costs of implementing list augmentation are fairly small.

Central to the general framework described above is the development of an effective predictive model. In this study, we present three models for estimating a customer’s total purchases in a category and how large a share of these purchases comes from the focal firm, by augmenting the firm’s internal database with survey information on a sample of customers’ purchases from the firm’s competitors. To our knowledge, this is the first study in the marketing literature using list-augmentation to impute wallet size and share of wallet.

We apply our models to a proprietary dataset provided by a major consumer bank in the U.S. containing information for over 34,000 customers on their uses of 10 categories of financial products offered by both the bank and its competitors. We first calibrate the models on a sub-sample with information on account balances both inside and outside the bank. We then use the calibrated models to predict total and share of requirements in each category for a validation sample using only inside balances (with the outside balances providing the basis of validation), thereby emulating the application of the models to the rest of the customer database, where information regarding outside balances is unavailable.

**Key Findings**

One of our three models stands out in terms of its predictive performance and managerial interpretability. It predicts correctly seventy-two percent of times whether a customer uses a product offered by the focal bank’s competitors, and offers the most accurate estimates of total and share of category requirements. Furthermore, it yields insights regarding customers’ share
decisions that can be used in guiding future relationship management initiatives. Several findings are highlighted below, some of which, we conjecture, may be idiosyncratic to the bank under study (e.g., 3), while others may be generalizable across firms and industries (e.g., 1, 2, and 4).

1. Longer relationships do not necessarily translate into larger share of wallet -- customer tenure is positively correlated with share of requirements in only six out of ten categories we analyze; the correlation is neutral or negative in the other four categories. This is consistent with the argument that relationship duration and customer share should be considered as two separate dimensions of customer relationship (Reinartz and Kumar 2003).

2. Customers with high share in one category also tend to have high share in another, indicating customers’ share decisions are positively correlated and hence the potential for positive externalities across categories.

3. Customers’ share and total purchase decisions are sometimes negatively correlated, suggesting that, for some categories, customers with small shares within the focal firm tend to transact large volume outside of it. These customers might represent significant opportunities for volume growth to the extent the focal firm can induce them to switch.

4. Customers with higher income tend to balance share of requirements across firms. This may suggest that either the focal firm is not serving such customers well or that customers with higher incomes have incentives to allocate business across firms.

To further investigate the managerial value of our proposed approach, we conduct a series of customer targeting simulations and find that substantial lifts in targeting efficiency can be obtained by using estimated total wallet and share of wallet. For example, 13% of customers in the validation sample are identified as high-potential customers as their estimated total wallet is in the top quintile yet their estimated share of wallet at the focal firm is below average. These customers turn out to account for 53% of the validation sample’s financial requirements that are fulfilled outside the focal bank (suggesting considerable potential for increasing revenue, to the extent the focal firm can induce them to switch).

In the rest of the paper we proceed as follows. The next section discusses in more detail the value of share of category requirements in managing relationships with individual customers. We then present the three models for estimating total and share of category requirements,
followed by our empirical illustration, where we present the managerial problem and data, the models’ predictive performances and parameter estimates, and the results from several customer targeting simulations. Finally, we summarize and discuss directions for future research.

**Share of Category Requirements as an Outward-looking Relationship Metric**

Customer relationship management efforts are often targeted towards a firm’s best customers, defined as those who contribute the most to the firm’s bottom line. Such strategies are effective when firms desire to strengthen relationships with “high value” customers thereby mitigating the likelihood these customers will defect to competing firms (Peppers and Rogers 2004). However, it could be myopic to target such customers for relationship developments, because profitability measures are blind to the relationships a customer might maintain with competitors, and there may be little correlation between the customer’s growth potential and her current or past net contributions. Such low correlation may predominate in industries where customers maintain concurrent relationships with multiple providers. Yet the volume of business a customer does with the competition represents an important source of the customer’s potential profitability. Indeed, a number of researchers have proposed or shown the link between customer share and profitability. For example, Garland (2004) examined the role of share-of-wallet in predicting customer profitability, finding it to be the single relationship-based measure with the most impact on customer contribution. Bowman and Narayandas (2004) and Keiningham, Perkins-Munn, Aksoy, and Estrin (2005) propose that share-of-wallet mediates the relationship between satisfaction and profits, an effect empirically confirmed by Bowman and Narayandas (2004). Reinartz, Thomas, and Kumar (2005) use share-of-wallet as a covariate when balancing acquisition and retention efforts to maximize customer profitability, finding that it positively affects customer tenure and profitability. Zeithaml (2000) proposes a conceptual model in which
increased share of wallet is one of four factors mediating the effect of customer retention on firm profits.

As scant empirical evidence exists regarding the degree to which a customer’s purchases inside and outside a firm correlate, we exemplify this issue using information we obtained from a large sample of customers of a major US bank (for a more detailed description of the data see the Data section). For this sample, the bank knows each customer’s financial portfolio with both the bank and its competitors. We find that 81% of the sample’s internal assets (i.e., the sum of financial assets the sample customers have deposited in the bank) come from customers who are the top 20% in terms of internal asset (a.k.a. the “80/20” rule). Yet these same customers account for only 34% of the sample’s external assets (i.e., the sum of financial assets the sample customers have deposited with the bank’s competitors). Moreover, the bank’s high volume customers are often not the customers with the greatest growth potential -- the correlation between internal assets and external assets is just 0.13.

For debt products (e.g., credit card, personal loan, mortgage, etc.), the pattern is more striking: 96% of the sample’s debts with the focal bank come from customers who are the bank’s top 20% clientele in terms of internal debt; however, these customers account for only 20% of the sample’s external debts. The correlation between internal debts and external debts is −0.04. In short, we find little correlation between a customer’s internal and external requirements. Thus, any targeted relationship development efforts predicated solely upon internal information would miss many high-potential customers who have large sums of assets and debts outside the bank. This suggests it is especially desirable to consider the bank’s share of a customer’s total wallet in gauging the customer’s potential for growth.

**Share of Category Requirements and Share of Wallet**
To elaborate on this notion of customer share, we distinguish between “share of category requirements” and “share of wallet.” We define share of category requirements as the ratio of i) a customer’s requirements for a particular category of products/services from a focal supplier to ii) the customer’s total requirements for products/services from all suppliers in the category (i.e., total category requirements). For a firm offering multiple categories of products/services to its customers, we define the firm’s share of wallet of a customer as the share of total requirements across all the product categories offered by the focal firm. Thus, share of category requirements is defined at the category level, and share of wallet is defined as an aggregate measure across all the categories in which the focal firm competes. We use share of category requirements because of its long-standing status in the marketing literature, and we use share of wallet because of its popularity among practitioners.

Share of category requirements has long been used as a metric of brand loyalty in the context of consumer packaged goods (Fader and Schmittlein 1993), and is becoming an important metric of customer relationship strength under different names, depending on the industry using it (Malthouse and Wang 1998). For example, financial services companies call it “share of wallet”; the auto industry calls it “share of garage”; the fashion industry calls it “share of closet”; the media industry calls it “share of eyeballs”; the non-profit industry calls it “share of heart.” Several papers in marketing have studied factors that can affect customer share. For example, Bhattacharya et al. (1996) explore the relationship between marketing mix variables and brand-level share of category requirements. Bowman and Narayandas (2001) assess the impact of customer-initiated contacts on share of category requirements. Keiningham, Perkins-Munn, and Evans (2003) analyze the impact of customer satisfaction on share of wallet in a B2B environment. Verhoef (2003) investigates the differential effects of relationship perceptions and
marketing instruments on customer retention and customer share. In the financial services area, researchers have looked into the relationship between customer characteristics and share-of-wallet; for example, Baumann, Burton, and Elliott (2005) use survey data to identify customer characteristics that are associated with high share-of-wallet in retail banking; Garland and Gendall (2004) use share-of-wallet as a factor in predicting customer behavior.

**Share of Category Requirements as a Basis for Customer Segmentation**

A number of researchers have proposed to use share-of-wallet as a segmentation basis. For example, Anderson and Narus (2003) propose a framework for strategically pursuing a customer’s business by selecting those with significant incremental share available and superior projected growth in need. Beaujean, Cremers, and Pereira (2005) propose to use loyalty and share-of-wallet as the bases for customer segmentation. Similarly, Reinartz and Kumar (2003) propose segmenting customers based upon their length of relationship with the company and share-of-wallet while Keiningham, Vavra, Aksoy, and Wallard (2005, pp.200-228) propose to combine share-of-wallet and customer lifetime value for the same purpose. Though this work is largely theoretical, below we demonstrate empirically how share-of-wallet as a segmentation basis can yield actionable strategic insights.

By combining customers’ share-of-wallet with their total wallet and using the bank data we shall later describe in more detail, Figure 1 depicts a 5-by-5 segmentation scheme that offers an outward-looking view of the bank’s customer base. Specifically, the first dimension represents the focal bank’s share of a customer’s total financial needs (i.e., assets plus debts) divided into quintiles: (0, 20%), (20%, 40%), (40%, 60%), (60%, 80%), and (80%, 100%). The second dimension is the customer’s total financial needs by quintiles. The size of each circle in the figure corresponds to the percentage of customers falling into the corresponding segment.
Under this scheme, customers in different segments can be viewed as a portfolio of assets with different growth prospects.

Customers in the upper-right corner are “ideal” from the bank’s perspective, as they have large requirements for financial products and fulfill most of those requirements with the bank’s offerings. However, these customers represent a small fraction of the bank’s customer base, as indicated by the size of their circles in Figure 1. Two growth strategies are indicated: first is to migrate customers in the upper-left corner towards the right (i.e., gaining a larger share of their business), and second is to migrate those in the lower-right corner upward (i.e., increasing their category requirements). We conjecture that the former strategy might prove more fruitful in the short run, because a customer’s category requirements is to a large degree driven by her intrinsic needs and constrained by financial resources. In our data, the two largest cells in the upper-left corner of Figure 1 account for 28% of customers. In terms of short-term growth potential in assets and debts, customers in these cells account for 72%. Thus, we contend that customers in the upper-left corner (i.e., those with small share of wallet but large total wallet) could be the best prospects for marketers seeking near-term growth, even though they are also likely to be targeted by the competition.

Unfortunately, it is difficult to implement such an outward-looking segmentation and targeting strategy in practice, as firms rarely have accurate customer-level measures for share of wallet and total wallet. Without such information, the upper-left customer segments in Figure 1 are indistinguishable from those in the lower-right by just looking at requirements fulfilled inside the focal firm. The main purpose of our study is to propose an approach that helps firms estimate these two important measures for each of their customers.
Share of Category Requirements as a Detector for “Silent Attrition”

Aside from being a segmentation basis, share of wallet can also be viewed as an indicator of relationship strength that can be used in early detection of customer attrition. While customers who close accounts or move all business to another supplier are clearly defecting, those whose purchases represent a smaller share of their total expenditures are also “defectors” (Reichheld 1996). For example, Coyles and Gokey (2002) found that 5% of customers at a bank close their checking accounts annually, taking with them 3% of the bank’s total balances. But every year, the 35% of customers who reduced their shares with the bank significantly cost the bank 24% of its total balances. This effect obtained in all 16 industries they studied (including airlines, banking and consumer products), and was dominant in two-thirds of them. This suggests that “partial defection” or “silent attrition” (Malthouse and Wang 1998) caused by decreasing share of wallet could be more serious than attrition, which is detected only when a customer has decided to no longer use the product or service of a firm. Accordingly, marketers need to monitor share of wallet on an ongoing basis and decreasing share of wallet should be viewed as an early warning signal that a relationship is gradually decaying. Compared with marketing interventions aimed at preventing attrition, marketing efforts that attempt to stop share of wallet from decreasing could be more proactive and therefore more effective. Unfortunately, most firms’ databases do not have share of wallet estimates for individual customers.

Share of Category Requirements and Cross-selling

Since the early 1990’s (Kamakura, Ramaswami, and Srivastava 1991), a growing literature has evolved on the topic of cross-selling (Jarrar and Neely 2002; Lau, Chow, and Liu 2004; Lau, Wong, Ma, and Liu 2003). This literature is focused on identifying the next product to offer a customer, based on the products previously purchased, and on the patterns of purchase
incidence across all customers. More recently (Li, Sun, and Wilcox 2005; Kamakura, Wedel, de Rosa, and Mazzon 2003), researchers recognized that customers have multiple relationships, proposing list augmentation approaches to make cross-selling recommendations that consider the possibility that the customer might already have purchased the product elsewhere. To our knowledge, most (if not all) cross-selling models in the marketing literature focus only on the purchase incidence decision. While cross-selling is an important tool for developing customers, the identification of cross-selling prospects covers only one aspect (purchase incidence) of customer development, overlooking other avenues to enhance the value of a customer. This is particularly true for industries such as banking and retailing, where customers maintain relationships with multiple vendors in the same product category. In these industries, customer relationships can be developed not only by having the customer make purchases in categories she has not bought before, but also by increasing the firm’s share of the customer’s requirements in categories where she has already made purchases. The list augmentation models we propose and test next consider the customer’s decision to adopt a product category as in these cross-selling models, but will also impute the total volume consumed in the product category, and the share of volume the customer devotes to a particular firm. In other words, we extend the cross-selling models from predicting only the purchases incidence decision to predicting the quantity as well as the incidence decision, thus providing a more informative estimate of a customer’s growth potential.

Summary

In light of the foregoing discussion regarding the advantages and challenges of using share-of-wallet in managing customer relationships, the goal of this paper is to develop a predictive model through which a firm can use its internal records, supplemented with a small
sample of external records to estimate total and share of requirements in all categories (and thus total and share of wallet) for all customers. This enables firms to extend existing customer development initiatives such as cross-selling and benefit from the segmentation, targeting and attrition detection strategies discussed above. Given the magnitude of these problems in practice, such an approach could prove quite useful to firms in many industries. Next, we present three models to achieve this aim.

**Models for Estimating Total and Share of Category Requirements**

**The Modeling Task**

Consider a firm that offers products in $J$ categories, with transaction records for $N$ customers. The firm knows i) $Y_{nj}^1 \in Y^1$ -- how much requirements it serves customer $n$ in category $j$ (superscript 1 denotes the focal firm), and ii) $X_n \in X$ -- a vector of other customer characteristics (for instance, in our empirical illustration $X$ consists of customer income and length of relationship with the focal firm). The firm does not have information on $Y_{nj}^0 \in Y^0$ -- the size of customer $n$’s requirements in category $j$ served by the firm’s competitors (superscript 0 denotes outside the focal firm). To learn about customers’ outside requirements $Y^0$, the firm conducts a survey among a random sample, $I$, of its $N$ customers (it is unlikely that the firm can survey all its millions of customers, but it is able to survey a sample of thousands). The goal of such a survey is to collect two pieces of information in each product category for each customer $i$ in sample $I$: the self-reported inside requirements $Y_{ij}^1$ and the self-reported outside requirements $Y_{ij}^0$. By obtaining self-reported inside requirements, the firm can assure they are consistent with their internally recorded counterparts. Once data are cleaned to ensure accuracy, the self-reported outside requirements can be linked to records in the firm’s customer database. In sum, the firm
has complete information -- \((Y^1, Y^0, X)\) -- for only a sample \((I)\) of customers. For the other \(N-I\) customers, information on outside requirements, \(Y^0\) is missing.

The objective is to develop a model that uses inside requirements, \(Y^1\) and customer characteristics, \(X\), to estimate share of category requirements \((S)\) and total category requirements \((T)\) for the \(N-I\) customers for whom information regarding outside requirements is missing. Specifically, the firm first splits the \(I\) customers with complete information into a calibration sample and a validation sample. Then various predictive models are estimated with the calibration sample, and their predictive performances compared in the validation sample. The model with the best predictive performance is then applied to the remainder of the customer database to impute the unknown total and share of category requirements, thereby augmenting the database with estimates regarding the customers’ relationships with competing firms and hence their potential for growth.

We develop three models for the above task. Model A predicts whether a customer will transact in a category with the focal firm’s competitors (i.e., the outside incidence decision), and if so, what is the transaction volume (i.e., the outside quantity decision). Model B extends Model A by simultaneously modeling the customers’ incidence and quantity decisions both inside and outside the focal firm across multiple product categories, thereby explicitly allowing for the possibility that these decisions might be correlated. Model A and B can be used to infer the focal firm’s share of a customer’s category requirements by dividing the customer’s inside purchases over the sum of her inside and (predicted) outside purchases. In contrast, Model C predicts the share allocation directly. It makes three simultaneous predictions: 1) whether a customer will buy in a category, regardless of the vendor (i.e., the category ownership decision), and if so, 2) the amount to buy in the category (i.e., the total decision), and 3) how large a portion from the focal
firm (i.e., the share decision), which can be 0%, 100%, or anything in-between. All three models are calibrated with the same data, and compared based on their predictive performances and managerial interpretability in the empirical section of this study.

Model A -- Modeling Purchase Incidence and Quantity Allocated to Competitors in a Product Category

In Model A, we use purchases within the focal firm, along with other customer data available internally, to predict the volume transacted outside the firm. Since the decisions regarding whether to buy and how much to buy from competitors may be driven by different underlying processes, we propose the following Type-2 Tobit regression model of incidence and quantity conditional on incidence (Amemiya 1985):

\[
\begin{align*}
\eta_{ij1}^* &> 0, \text{ then } Y_{ij}^0 = \exp(\eta_{ij2}^*); \text{ else } Y_{ij}^0 = 0 \\
\eta_{ij1}^* = \alpha_{j1} + x_i' \beta_{j1} + Ind(Y_i^1) \gamma_{j01} + \ln(Y_i^1) \gamma_{j11} + \varepsilon_{ij1} \\
\eta_{ij2}^* = \alpha_{j2} + x_i' \beta_{j2} + Ind(Y_i^1) \gamma_{j02} + \ln(Y_i^1) \gamma_{j12} + \varepsilon_{ij2}
\end{align*}
\]

where,

- \( \eta_{ij1}^* \) is a latent variable that captures the likelihood of customer \( i \) buying in category \( j \) from a competing firm. In other words, this latent variable determines the incidence whether customer \( i \) has a relationship outside the focal firm in category \( j \);
- Conditional on that the customer has a relationship with a competitor, \( \eta_{ij2}^* \) is another latent variable determining the quantity purchased by the customer from the competitor. As the empirical distributions of conditional quantities are often skewed, they are assumed to be an exponential function of the latent variable, and thus enter the likelihood function on a log-transformed scale;
- \( x_i \) is a \( K \)-element vector of observable characteristics of customer \( i \);
- \( Y_i^1 = (Y_{i1}^1, \ldots, Y_{ij}^1)' \), and \( Ind(\cdot) \) is an incidence indicator function, such that \( Ind(Y_{ij}^1) = 1 \) if \( Y_{ij}^1 > 0 \), else \( Ind(Y_{ij}^1) = 0 \);
• $\varepsilon_{ij1}$ and $\varepsilon_{ij2}$ are normally distributed with the covariance being 
\[
\begin{pmatrix}
\sigma^2_{j1} & r_{jj1}\sigma_{j1}\sigma_{j2} \\
r_{jj2}\sigma_{j1}\sigma_{j2} & \sigma^2_{j2}
\end{pmatrix},
\]
and it is assumed that $\sigma_{j1} = 1$ for identification purposes (this covariance structure implies that the incidence and quantity decisions might be correlated, which can occur if there are unobserved factors that affect both decisions);

• $\alpha, \beta, \gamma, r$ and $\sigma$ are the model parameters, and there are $(2 + 2K + 4J + 2)J$ of them to be estimated.

Model A is relatively simpler than the subsequent ones in that it models the customer’s purchase decision in each category independently. Of note, there are likely to be unobserved factors (therefore can not be included in Model A as predictors) affecting customers’ purchases from both inside ($Y_{ij}^1$, $s$) and outside ($Y_{ij}^0$, $s$) the focal firm. If the impact of these unobserved factors is substantial, it can lead to biased parameter estimates and thus poor predictions of total and share of category requirements. That said, we still view Model A as a highly practical solution and a strong benchmark for the other two models proposed next, since it can be easily estimated as a Type-2 Tobit regression and it leads to straightforward data imputations. Model B extends Model A by correcting these potential biases. It does this by modeling purchase decisions inside and outside the focal firm simultaneously, allowing them to be correlated across categories as well as with each other.

Model B -- Modeling Incidence and Quantity Inside and Outside the Focal Firm across Multiple Categories

Unlike Model A, which ignores the process governing purchases within the focal firm, Model B assumes that customers make four decisions in each product category: 1) whether to buy from the focal firm; 2) if so, how much; 3) whether to buy from competitors; and 4) if so, how much. Model B allows these decisions to be driven by four different underlying processes that might be correlated not only with each other, but also across categories, due to the impact of
a common set of unobserved customer-specific factors. For example, a recent divorce (for which we do not have data) might lead to reduced purchases from all firms in all categories, suggesting that these decisions might be positively correlated. Formally, we model the incidence and quantity decisions in each category, both inside and outside the focal firm, as follows,

\[
\begin{align*}
\text{if } \eta_{ij}^s > 0, & \text{ then } Y_{ij}^1 = \exp(\eta_{ij}^s); \text{ else } Y_{ij}^1 = 0 \\
\text{if } \eta_{ij}^3 > 0, & \text{ then } Y_{ij}^0 = \exp(\eta_{ij}^4); \text{ else } Y_{ij}^0 = 0.
\end{align*}
\]

Equation 2 indicates that customer \(i\)'s incidence and quantity decisions in category \(j\) are determined by four latent variables, \(\eta_{ij}^s\) 's. The top half of Equation 2 implies that customer \(i\) will only buy a product in category \(j\) from the focal firm if \(\eta_{ij}^1\) is greater than zero. When that happens, the customer’s requirements for the focal firm’s product is an exponential function of \(\eta_{ij}^2\). The bottom half of Equation 2 states that customer \(i\) will buy a product in category \(j\) from other firms if \(\eta_{ij}^3\) is greater than zero, and if so, the customer’s requirements served outside the focal firm is an exponential function of \(\eta_{ij}^4\), which implies that the conditional quantities of customers requirements \(Y_{ij}^0\), like \(Y_{ij}^1\), enter the likelihood function on a log-transformed scale.

As indicated previously, we seek to capture the impact of unobserved customer-specific factors that simultaneously affect customers’ incidence and quantity decisions made across product categories, and inside and outside the focal firm. Formally, we adopt the following structure on the four latent variables \(\eta_{ij}^1\), \(\eta_{ij}^2\), \(\eta_{ij}^3\), and \(\eta_{ij}^4\):

\[
\begin{align*}
\eta_{ij}^1 & = \alpha_{j1} + x_i \beta_{j1} + z_i \gamma_{j1} + \epsilon_{ij1} \\
\eta_{ij}^2 & = \alpha_{j2} + x_i \beta_{j2} + z_i \gamma_{j2} + \epsilon_{ij2} \\
\eta_{ij}^3 & = \alpha_{j3} + x_i \beta_{j3} + z_i \gamma_{j3} + \epsilon_{ij3} \\
\eta_{ij}^4 & = \alpha_{j4} + x_i \beta_{j4} + z_i \gamma_{j4} + \epsilon_{ij4}
\end{align*}
\]
where,

- $x_i$ is defined as earlier, a $K$-element vector of observable characteristics of customer $i$;
- $z_i$ is a $P$-element vector that captures unobserved, individual-specific factors affecting the incidence and quantity decisions of customer $i$; each element of $z_i$ is assumed to be i.i.d. standard normal;
- $\varepsilon_{ij1}, \varepsilon_{ij2}, \varepsilon_{ij3},$ and $\varepsilon_{ij4}$ are the stochastic components in each decision that are normally distributed with variances being $\sigma_{j1}^2$, $\sigma_{j2}^2$, $\sigma_{j3}^2$, and $\sigma_{j4}^2$, respectively; for identification purposes, it is assumed that $\sigma_{j1} = \sigma_{j3} = 1$;
- $\alpha$, $\beta$, $\gamma$ and $\sigma$ are the model parameters; $P$, the dimensionality of unobserved customer factors, is to be determined empirically; given $K$, $P$, and $J$, $((1 + K + P) \times 4 + 2) \times J$ parameters need to be estimated.

The above structure provides a parsimonious representation of the correlation pattern between the $4^*J$ (i.e., the # of decisions in each category times the # of categories) latent variables, $\eta_i^*$'s. The gain in parsimony will be substantial when the number of categories $J$ is large and decisions across these categories are inter-related. Like Model A, a caveat of Model B is that its parameter estimates cannot be directly interpreted to learn about how customers decide to allocate their business across vendors. Accordingly, in Model C, we model customers’ share decisions explicitly.

**Model C -- Modeling Category Ownership, Total and the Focal Firm’s Share**

In Model C we attempt to develop an approach that can be used to directly predict a customer’s category demand and the focal firm’s share, when only inside purchases are recorded. Our motivation for developing such a model is twofold. First, we would like to understand how observed customer characteristics affect the total and share of category requirements decisions. This is important because different customer characteristics may be related to a customer’s category demand and share allocation in different ways, and marketers interested in
understanding these relationships can use such insights to devise customer development strategies accordingly.

The second reason for modeling total and share of category requirements explicitly is that such a specification aligns itself well with the choice modeling literature that decomposes consumer purchase decisions into “incidence, quantity and choice” (Gupta 1988), wherein consumers decide whether to buy in a category and, if so, how much, and finally, which brand(s) to choose. In Model C, we propose to decompose customers’ category requirements decisions in a very similar fashion --“ownership, total and share.” Of note, our approach and context is quite different than that of Gupta (1988) and others inasmuch as i) the goal is to impute these decisions with incomplete information, ii) we consider these decisions made across multiple categories, and iii) the choice and share decisions require different treatments.

Consequently, Model C assumes that a customer faces decisions of Ownership -- whether or not to own a product category, Total -- the total category requirements if she decides to own, and Share -- the share of that total requirements, if any, to be served by the focal firm. Formally, we have:

\[ \begin{align*}
&\text{if } \eta_{ij1} > 0, \text{ then } T_{ij} = \exp(\eta_{ij2}^*); \text{ else } T_{ij} = 0 \\
&\text{if } \eta_{ij3} \leq 0, \text{ then } S_{ij} = 0; \text{ else if } \eta_{ij3} > 1, \text{ then } S_{ij} = 1; \text{ else } S_{ij} = \eta_{ij3}^*.
\end{align*} \]

Equation 4 posits that customer i’s purchase decisions in category j are governed by three latent variables, \( \eta_{ij}^* \)’s. The top portion of Equation 4 states that the customer will only buy a product in the category if \( \eta_{ij1}^* \) is greater than zero (i.e., category ownership), and when that happens the customer’s total requirements \( T_{ij} \) is an exponential function of \( \eta_{ij2}^* \), which implies that the conditional quantities of total requirements \( T_{ij} \) enter the likelihood function log-transformed. The bottom portion of Equation 4 implies that the customer will allocate some of...
her purchases to the focal firm if $\eta^*_ij3$ is greater than zero. When that happens, the customer might have all her requirements served by the focal firm (if $\eta^*_ij3$ is greater than or equal to one), or a share of them that is equal to $\eta^*_ij3$. Thus, we model share of category requirements as a distribution truncated at 0 and 1. Further, we assume that the three latent variables $\eta^*_ij1$, $\eta^*_ij2$, $\eta^*_ij3$ are functions of a common set of factors, observed as well as unobserved, with a structure as follows:

$$
\begin{align*}
\eta^*_ij1 &= \alpha_{ij1} + x_i' \beta_{ij1} + z_i' \gamma_{ij1} + \epsilon_{ij1} \\
\eta^*_ij2 &= \alpha_{ij2} + x_i' \beta_{ij2} + z_i' \gamma_{ij2} + \epsilon_{ij2} \\
\eta^*_ij3 &= \alpha_{ij3} + x_i' \beta_{ij3} + z_i' \gamma_{ij3} + \epsilon_{ij3}
\end{align*}
$$

(5)

where,

- $x_i$ and $z_i$ are defined as earlier, denoting, respectively, $K$ observed characteristics and $P$ unobserved factors associated with customer $i$, with each element of $z_i$ assumed to be i.i.d. standard normal; $z_i$ can also be interpreted in the terminology of factor analysis as $i$’s scores on $P$ latent factors, with $\gamma_j$ as the factor loadings (which we shall demonstrate how to interpret in the Results section);

- $\epsilon_{ij1}$, $\epsilon_{ij2}$, and $\epsilon_{ij3}$ are the stochastic components in each decision, normally distributed with variances being $\sigma^2_{j1}$, $\sigma^2_{j2}$, and $\sigma^2_{j3}$, respectively; for identification purposes, it is assumed that $\sigma_{j1} = 1$;

- $\alpha$, $\beta$, $\gamma$ and $\sigma$ are the model parameters; $P$, the dimensionality of unobserved customer factors, is to be determined empirically; given $K$, $P$, and $J$, $((1 + K + P) \times 3 + 2) \times J$ parameters need to be estimated.

Denote the variance-covariance matrix of the latent vector $\eta^*_i$ as $\Lambda (3J \times 3J)$, arising from 3 decisions by $J$ categories. Thus, for the $j^{th}$ and $g^{th}$ categories ($j$ and $g$ in $J$), and the $h^{th}$ and

---

3 White tests for heteroskedasticity on the log-transformed non-zero balances show that the residuals of the model are not a function of the predicted log total balances, which indicates that the homoscedasticity assumption with regard to the stochastic components conforms to the data.
\( i^{th} \) decisions (ownership, total or share), the foregoing structure implies that the variance and covariance between \( \eta_{ijh}^* \) and \( \eta_{igl}^* \) can be expressed as:

\[
\text{var}(\eta_{ijh}^*) = \gamma_{jh}^\prime \gamma_{jh} + \sigma_{jh}^2 \quad \text{or} \quad \text{cov}(\eta_{ijh}^*, \eta_{igl}^*) = \gamma_{jh}^\prime \gamma_{gl}^\prime.
\]

Equation 6 depicts the role of \( \gamma \), the \( 3J \times P \) factor loadings, in reducing a large \( 3J \times 3J \) variance-covariance matrix into a smaller \( 3J \times P \) factor space as \( \Lambda = \gamma \times \gamma' + \Sigma \), where \( \Sigma \) is a diagonal matrix with \( \sigma_{jh}^2 \) as entries. As such, the structure depicted in Equation 5 provides a parsimonious representation of the nature and strength of the correlations between \( 3*J \) decisions, namely ownership, total and share decisions made by each customer in \( J \) product categories. Moreover, the estimates of factor loadings lend themselves to substantive interpretation and graphical display, which we will illustrate later in our application. Detailed procedures for estimating this model and for imputation can be found in Appendix A and B, respectively.

### Data for Empirical Illustration

#### Data Collection

The data used for our empirical illustration are provided by a major U.S. bank. This bank provided us with information for 34,142 of their customers regarding their balances outside the bank in 10 categories such as non-interest checking, interest checking, savings, CDs, investments, car loan, personal loan, line of credit, credit card loan, and mortgage. These data were compiled by a syndicated research supplier via an ongoing market audit study. Each observation from this audit details when it was collected, and only households who had at least one account with positive balance with the bank were included to ensure these customers were active. For the periods during which the market audits were conducted (i.e., from 1999 Q3 through 2002 Q2, with about 4,500 households each quarter), this information on customer
external relationship was aligned with records on the balances these households had inside the bank in the 10 categories \( Y^1, \cdots, Y^{10} \), thereby affording us with a unique dataset that includes both share of category requirements \( S^1, \cdots, S^{10} \) and total category requirements \( T^1, \cdots, T^{10} \).

The dataset provided by the bank also contains two observed customer characteristics, annual household income and customer tenure, which we use as predictor variables \( x_1 \) and \( x_2 \), respectively) in the empirical illustration our proposed models\(^4\). Household income is pre-tax and includes income from not only salaries but also interests and investment returns. Given its skewed nature, household income is log-transformed throughout our illustration. Customer tenure is measured in the number of years since the first account was opened at the focal bank.

As most firms in other industries would not have the benefit of a syndicated service such as the one the bank has, they would therefore need to conduct a survey (often anonymously) on their own with a sample of their customers to obtain data on the volume of businesses these customers have with competitors. For this sample of customers, with complete information on their relationships with the firm and its competitors, our proposed models can be calibrated, and then used to impute total and share of category requirements for all customers. To improve the accuracy of the imputation results, it is desirable to check for self-reporting errors in the survey data. Since firms can not obtain transactional data from their competitors, it is difficult to check directly the error rates in the sample customers’ self-reported data on the business they transact outside the focal firm. One indirect approach is to collect self-reported data on the business the sample customers transact inside the focal firm, and to ensure these data align with internal records. Accordingly, in Appendix C we compare the account balances inside the bank obtained

\( ^4 \) We seek to develop a modeling framework that generalizes beyond the banking industry, and these two characteristics are not unique to banking customers. In practice, our proposed models can readily extend to include other industry-specific customer characteristics, whenever they are available. Adding those variables as predictors could only serve to increase the effectiveness of our approach.
from internal records with the customers’ self-reports of those balances (obtained through the market audit study). We find the discrepancies to be negligible, and therefore presume that the balances outside the bank from the market audit study will also exhibit little systematic error.

**Study Design**

We apportion our data into a calibration sample comprised of 23,957 subjects and a validation sample of 10,185 subjects. This apportionment is intended to reflect the data available to firms in practice, where they have complete data (analogous to the calibration sample) for only a subset of their customers, and incomplete data for the rest of their customer base (analogous to the validation sample). Accordingly, for the calibration sample we use all the information available, including account balances inside the focal bank \(Y^1\), household income \(X_1\) and customer tenure \(X_2\) as well as account balances outside the focal bank \(Y^0\). The validation sample mimics the remainder of the customer base for whom the focal bank does not know balances the customers might keep with competitors. We first use the calibration sample to fit the proposed model. Then, in the validation sample, we apply the calibrated models to the internal data (i.e., inside balances, household income and customer tenure) to impute the outside balances, and therefore predict total \(T\) and share of category requirements \(S\). Finally, we evaluate the models’ predictive performances in the validation sample by comparing the imputed \(T\) and \(S\) against their observed counterparts. Table 1 summarizes the above design.

--- INSERT TABLE 1 ABOUT HERE ---

---

\(^5\) Many firms may not be able to afford a large scale survey on customers’ external transactions. In those cases, because of its efficiency in using data, K-fold cross-validation could be more preferable than the holdout sample cross-validation.
Summary Statistics

Table 2 reports several summary statistics of our data for both the calibration and the validation samples. Not surprisingly, the two sub-samples are quite similar. The second column in the table shows the average category ownership of each type of product. The third column shows the average total balances conditional on ownership. Among those who have positive total balances, the fourth column shows the percentage of customers who have zero balance with the focal bank (hence share = 0). The fifth column reports the percentage of customers who have positive balances both inside and outside the bank (hence 0 < share < 1). About 60% of the bank’s customers have financial assets inside, as well as outside, the bank. Similarly, about 50% of the bank’s customers have financial debts both inside and outside the bank. The sixth column shows the percentage of customers who are exclusive customers of the focal bank (hence share = 1). About one in five of the bank’s customers have 100% of their requirements for financial assets served by the bank. Less than one in 10 of the bank’s customers do so when it comes to fulfilling requirements for financial debts. Together, columns four through six suggest the importance of treating the distribution of share decisions as truncated at both 0 and 1.

Finally, the last column shows the bank’s average share of requirements in each category. The bank’s deposit products (interest and non-interest checking, savings, CDs) perform best in terms of obtaining a large share of customers’ business. Loan products, except for line of credit, have much smaller shares than deposit products. The bank performs most poorly with respect to attracting customers’ investment dollars. Taking as a whole, the bank has only about 20% of its existing customers’ total wallet. Having accurate estimates of share of category requirements for its millions of customers should be a helpful first step towards determining which customers should be targeted first, in what product categories, with what kinds of marketing treatments.
The other two variables included in our empirical illustration are the customers’ household income and tenure with the focal bank. The mean, median, and standard deviation of income are, respectively, $54,566, $48,086, and $36,198 for the calibration sample, and $54,859, $48,776, and $35,684 for the validation sample. The mean, median, and standard deviation of customer tenure (in years) are, respectively, 9.4, 10, and 3.8 for the calibration sample, and 9.2, 10, and 3.9 for the validation sample.

-- INSERT TABLE 2 ABOUT HERE --

Results

Model Comparison

In this section we first compare the predictive performances of Models A, B, and C in the validation sample. Specifically, we evaluate the ability of each model to predict total and share of category requirements for each separate category, as well as for all assets together, all debts together, and all categories as a whole (i.e., share of wallet and total wallet). We consider three dimensions: outside product ownership, total category requirements (wallet size), and share of category requirements (share of wallet). Organizing our discussion by each dimension, we find:

• *Predicting Outside Product Ownership*. Table 3, Part a (columns 2-4), compares the models’ ability to predict outside product ownership in terms of the *hit rate*, i.e., the percentage of times a model correctly predicts the ownership or non-ownership of the product offered by the *competition*\(^6\). The results show that *all three models perform equally well* in terms of predicting whether a customer has positive balances outside the bank in a particular category, with hit rates averaging about 72~73%.

• *Predicting Total Category Requirements*. To compare the performance of the models in terms of their ability to predict total category requirements, we calculate the *Mean Absolute Deviation* (MAD) of the predicted total category requirements\(^7\). To reflect the variability inherent in our data, we also calculate the MAD for a naïve model where the predicted total

---

\(^6\) A hit is counted when either the customer has positive balances outside the bank and the model predicts the likelihood of this incidence is in excess of 50%, or the customer has zero balances outside the bank and the model predicts the likelihood of having positive balances outside the bank is less than 50%.

\(^7\) We also calculated the MAD of the log-transformed predicted total category requirements, and the performance of Model C remained much better than the other two models.
category requirements is simply the account balances inside the bank plus the average account balances outside the bank. Table 3, Part b (column 5-8) reports the naïve model’s MAD and the percentage improvement (reduction) in MAD relative to the naïve model for the three predictive models (which is equal to 1 - MAD_Model / MAD_Naïve). The results show that Model C performs much better than Model A and B in predicting the exact sizes of total category requirements (or, equivalently, requirements served outside the focal bank). This is the case for total wallet, total assets, total debts, as well as for all 10 individual product categories.

- **Predicting Share of Category Requirements.** For this comparison, we calculate the Mean Absolute Deviation of the predicted share of category requirements. By definition, when account balance in the focal bank is zero, share of category requirements is either zero or not defined; consequently, accuracy in predicting share of category requirements is relevant only with respect to observations where there is positive account balance in the focal bank. Again, as a reference, we use a naïve model in which predicted share of category requirements is equal to account balances inside the bank divided by the sum of account balances inside the bank and the average account balances outside the bank. Table 3, Part c (column 9-12) reports the MAD of the naive model, along with the percentage improvement (reduction) in MAD relative to the naïve model for the three predictive models. Again, Model C outperforms Models A and B in predicting share of category requirements, which is the case for total wallet, total assets, total debts, as well as for 9 out of 10 individual product categories. To our surprise, in 8 out of 10 categories, the simpler Model A does better than the more sophisticated Model B.

Taking the model comparison results reported in Table 3 as a whole, we conclude that if the goal is simply to predict whether a customer uses a product or service offered by the competition (i.e., the cross-selling prediction proposed by Kamakura et al. 2003), any of the three models would be sufficient. On the other hand, if the goal is to predict the sizes of total and share of category requirements, Model C is more accurate than the other models. Given that i) total and share of category requirements are central to assessing firm profits (as customer value is typically proportional to their purchases summed across various product offerings), ii) Model C’s parameter estimates can be readily interpreted to gain insights into customers’ total and share decisions, and iii) it provides the best prediction overall, we believe Model C to have the greatest
utility for imputing total and share of category requirements. As such, we subsequently focus upon the estimation results for Model C to conserve space. Parameter estimates for Model A and B are available upon request.

**Estimation Results for Model C**

Estimation of Model C from the calibration sample for $P = 0$ to 3 latent factors leads us to choose the model with $P = 2$. The BIC’s for Model C with $P = 0, 1, 2, 3$, and for the other two models are reported in Table 4. Parameter estimates of the two-factor ($P = 2$) Model C are reported in Table 5a (which includes the intercepts, $\alpha$, and the coefficients on customer income and tenure, $\beta$) and 5b (which reports the factor loadings, $\gamma$).

**Income.** Table 5a indicates that customer income positively and significantly correlates with both ownership and total decisions across the product categories. Customers with higher income are more likely to own assets and debts (e.g., $\beta_{1,\text{income}} = .49$ for savings, with $p < .01$), and higher requirements when they own them (e.g., $\beta_{2,\text{income}} = .86$ for savings, with $p < .01$). Customers with higher income also have a significantly smaller share of their requirements for financial products with the bank under study (e.g., $\beta_{3,\text{income}} = -.54$ for savings, with $p < .01$). We conjecture that there could be many causes for this result. First, with greater requirements for financial products, high-income customers stand to gain more (e.g., lowered risk exposure) by spreading their “nest eggs” across more financial institutions. Second, competition may be more intense for the high income households’ wallets, giving them more incentives and opportunities to fulfill their requirements at multiple competing institutions. Third, the bank under study may

---

8 We also estimated Model C with customer income and tenure omitted. Results indicate that most of the Model’s predictive power arises from historical purchase behavior, as opposed to demographic variables such as income or tenure.
be comparatively less preferred by its higher income clientele due to, for example, a particular perceived market position. In any case, the bank should attend to its lower share of its higher income customers’ wallets, and see to what extent better targeted marketing treatments could redress this unfavorable position.

**Tenure.** The relation between customer tenure and the customer ownership, total and share decisions are less clear. Longer-tenures are associated with an *increased* share of customers’ *checking, CDs, investments, personal loans,* and *line of credits balances* (as reflected in the positive and significant $\beta_{\text{tenure}}$’s). In contrast, there is *no significant correlation* between a customer’s tenure and the bank’s share of their *savings, car loans,* and *credit card balances* (as reflected in the insignificant $\beta_{\text{tenure}}$’s). For customers with longer tenures, the bank has a *smaller* share of their mortgages (where $\beta_{\text{tenure}} = -.61$, with $p < .01$). The lack of positive relation between customer tenure and customer share in certain categories may deserve the bank’s close attention, as it may imply that lengthening relationship with their customers is not necessarily being translated into a larger share of their customers’ business in these categories.

**Factor Loadings.** Out of the 60 estimated factor loadings reported in Table 5b, 52 are significant at $p < .01$ level, indicating the latent factor structure captures unobserved individual-specific factors that have caused the correlations between ownership, total and share decisions in the various product categories. To visualize these correlations, Figure 2a and 2b plot the factor loadings from Table 5b. The rendering of correlation patterns yields managerial insights regarding customers’ ownership, total wallet and share of wallet decisions among the bank’s services. For example, a positive correlation between two ownership decisions means a customer more likely to own one service is also more likely to own another. In Figure 2a and 2b, any two vectors pointing in the same (or opposite) direction indicate that decisions represented by these
vectors are positively (or negatively) correlated; two vectors orthogonal to each other imply that the corresponding decisions are independent to each other. All the factor loadings plotted in the figures have been standardized so that the length of the vectors, as well as the angle between them, directly reflect the strength of correlation between the decisions they represent.

-- INSERT FIGURE 2a AND 2b ABOUT HERE --

Figure 2a captures correlations between the “Ownership of a Service” and “Total Requirements for a Service” decisions. Several things are noteworthy:

a) There is an “Assets vs. Debts” dimension in terms of the “Total Requirements” decisions. *Interest checking, savings, CDs and investments* decisions point in one direction, while *credit card, mortgage* and *personal loans* decisions point in the other direction. Thus, all else equal, customers with more financial assets have less debts (except for *line of credit*). This also provides face validity for the correlation results.

b) In terms of the “Ownership of a Service” decision, customers who own a *non-interest checking* and/or a *debt* account are less likely to own an *interest checking* account. Ownership of *interest checking, savings, CDs* and *investments* accounts are positively correlated, thus reflecting a propensity to save on the part of some customers. Similarly, customers with one type of debt are more likely to have other types of debt and are less likely to have financial assets (except for *non-interest checking*).

c) The same factors that induce customers to assume debt lead them to be deeper in debt (as vectors representing debt ownership and total debt point in the same direction). A similar relationship exists for assets.

More germane to targeting customers with high growth potential, the factor loadings plotted in Figure 2b depict the correlations between “Total Requirements” and “Share of Requirements” decisions across various product categories, suggesting that:

a) Share decisions are typically positively correlated across product categories, except for *car loan*. Thus, customers who do the bulk of transactions with the bank in one category also tend to do so in others. This might reflect general goodwill towards the bank.

b) In 5 out of 10 categories (i.e., *interest checking, savings, CDs, investments*, and *car loan*), the bank has been efficient in targeting in the sense that the bank has larger shares of those customers with larger requirements.
c) However, for line of credit, share of requirements and total requirements are independent. This implies that the bank might not be targeting high-volume customers to promote their line of credit.

d) For personal loans, credit card and mortgage, share of category requirements and total category requirements are negatively correlated. This suggests that the focal bank has a low share among high requirement customers in these three categories. Thus, these customers represent potentially attractive targets.

**Testing the Targeting Efficiency of Our Recommended Approach**

To ascertain the efficacy of Model C for targeting customers with “large total requirements” but “small current shares” (hence having the highest potential for short-term growth), we assess Model C’s ability to identify those customers whose total assets (i.e., sum of checking, savings, CDs and investments) are in the top quintile but the bank’s shares are lower than average. We do this by applying the calibrated model to the validation sample in order to predict who is in this “high-opportunity” segment (i.e., the top quintile of total assets but lower than average shares with the focal bank). We then compare these predictions to the actual data. Figure 3a shows the predicted distribution of customers by total assets and share of assets while Figure 3b portrays the actual customer distribution. These figures indicate that the model produces a fairly accurate estimate of the true distribution in the validation sample.

-- INSERT FIGURE 3 ABOUT HERE --

Though comprising only 12% of the validation sample, customers predicted to be “high-potential” (i.e., large wallet, small share of wallet) account for 37% of the validation sample’s assets outside the bank. This leads to a lift ratio of 37%/12% = 3.1 for Model C. Comparable lift ratios for Model A, and Model B are 2.6 and 2.5, respectively. Using the observed data to determine the high potential group (i.e., perfect hindsight and maximum lift achievable) yields a lift of 4.7. Together, these results suggest that our list-augmentation approach allows firms to
effectively target customers with high potential for short-term growth, and that Model C performs better than the two benchmarks at this task.

Targeting exercises similar to the one above can also be performed for total debts, or, on a category-by-category basis. For example, relative to other categories, the focal bank has the smallest share of its customers’ investment dollars (see Table 2). Model C suggests that if the bank targets customers whose requirements for investment products is predicted to be in the top quintile and yet the share allocated to the bank is predicted to be below average, the bank will be addressing 18% of its customer base that accounts for only 1% of the investment dollars already inside the bank but 48% of the investment dollars that are currently outside the bank.

Finally, consider the 5-by-5 segmentation scheme previously illustrated in Figure 1. Using estimated share of wallet and total wallet to segment its customer base, the bank can target the segment on the upper and left corner, namely those with total wallet in the top quintile and yet share of wallet below average. Our approach based on Model C yields 13% of the validation sample customers for targeting. There is an overlap of 85% between the customers predicted to be in this segment and the customers classified into this segment using the actual data. Moreover, these predicted targets account for 51% of the validation sample’s financial requirements that are currently served outside the focal bank. Again, this suggests substantial targeting efficiency (with a lift ratio of 51%/13% = 3.9).

**Concluding Remarks**

As a result of limited data, many CRM initiatives ignore customers’ transactions with competing firms, which in our view reflects a tendency of enterprises taking the new customer-centric paradigm of marketing to an inward-looking extreme. We provide direct evidence that internal transaction records alone are largely uninformative about a customer’s total market
potential (a crucial piece of information for assessing potential customer profitability). This highlights the risks of gauging a customer’s potential value by relying solely on levels of transaction or profitability recorded in the internal database. Both internal and external data are necessary to distinguish customers with large total expenditures and small share from those with small total expenditures and large share. Even though these two types of customers are indistinguishable from internally recorded purchases alone, they call for different relationship development strategies.

Firms lack individual-level, industry-wide customer data because they seldom have information regarding their customers’ relationships with competitors. Though it is generally infeasible to obtain transaction records from competing firms, firms can obtain external relationship data for a small sample of their customers via customer surveys or syndicated services. We present a list-augmentation approach that allows firms to combine internal records with these external data collected for a sample of customers, to calibrate a predictive model that can then be used to estimate total and share of requirements for all customers in all categories.

In our empirical illustration, we apply three such models to a proprietary dataset provided by a major U.S. bank, containing information regarding over 34,000 customers’ holdings of financial products in 10 categories, both inside and outside the bank. These data allow us to test the performance of these models in imputing wallet size and share of wallet based only on internal data. One of our three proposed models stands out in several ways: i) the parsimony of the parameter space, ii) the ease in interpreting the parameter estimates for behavioral insights, and iii) the better performance in predicting the sizes of total and share variables. Additionally, this model extends extant approaches for imputing data missing due to sub-sampling (Little and Rubin 2002; Schafer 1997; Kamakura and Wedel 2000) to the context wherein multiple
unobserved customer decisions -- a) which categories to select, b) how much to expend in these categories and c) how large a share for the focal vendor in each category, are simultaneously imputed from the observed joint outcomes of these decisions, i.e., how much, if any, to expend on the focal vendor’s offerings. Together, these measures enable firms to segment their internal customers based on share of wallet and total expenditures, and to discriminate between high share low expenditure customers and low share high expenditure customers.

We demonstrate that it is possible to use our approach to predict, via internal records, who has high expenditures outside the firm. Such information is useful in targeting customers with high market potential but low share of requirements. Generally speaking, we believe, variations in share of requirements -- across customers, product categories, and time periods -- should prove to be a fertile ground for response modeling and for evaluating the impacts of different marketing instruments and competitive actions. Insights from these studies can be used to identify the lead indicators and drivers behind those variations, which can then be used to optimize allocation of marketing resources in customer relationship management.

Our study also leads to some interesting behavioral findings on customers’ share decisions. For example, we find the bank under study has a smaller share of its higher income customers’ wallets. Apparently the bank may want to invest more in its high income clientele to gain a larger share of their business. Another intriguing finding is that longer customer tenure does not necessarily turn into larger customer share, which runs counter to the conventional wisdom that the longer a customer stays with a company, the more she buys from the company. Related to tenure, another avenue for further exploration is to ascertain whether decreasing share of wallet can be used as an early warning signal to prevent customer attrition, which inevitably entails examining longitudinal share of wallet movements.
Though the approach we propose is illustrated using data from the financial services sector (7.3% of U.S. GDP in 2003, according to *Survey of Current Business*, Jan 2005), it can be readily adapted to any industry where consumers routinely fulfill their category requirements by purchasing, simultaneously, a variety of products and services from multiple competing suppliers (e.g., retailing, direct marketing, etc.), rendering any particular firm uncertain about how large a share of a customer’s business it has⁹. We believe share of category requirements can potentially play a more prominent role in improving our understanding and managing of customer relationship in a broad range of industries, and our modeling framework could be viewed as a step towards achieving that goal, by generating insights regarding customers’ allocation of their business across vendors and by augmenting firms’ customer databases with reasonably accurate estimates of total and share of category requirements for each individual customer.

---

⁹ Even for Consumer Packaged Goods industries where panel data are commonly available, the non-disclosure of information by major retailers such as Wal-Mart is becoming increasingly a problem of incomplete customer purchase information, and our approach is useful in that context as well.
Appendix A. Model Estimation

In this Appendix we focus on the estimator for Model C. First, let
\[ \lambda_{ijk} = \alpha_{jk} + \gamma_i + z_j' \beta_j + \gamma_k' \beta_k \] for \( k = 1, 2, \text{ or } 3 \). For all possible scenarios of \( T_{ij} \) and \( S_{ij} \), we have the following densities conditional on \( z_i \):

When \( T_{ij} = 0 \), \( f(T_{ij}, S_{ij}) = f(\varepsilon_{ij1} \leq -\lambda_{ij1}) \), or formally,
\[
(A.1) \quad f(T_{ij}, S_{ij} | z_i, x_i, \Theta) = \Phi(-\lambda_{ij1}) = \text{Prob}_{ij[1]}(z_i);
\]

When \( T_{ij} > 0 \) and \( S_{ij} = 1 \), \( f(T_{ij}, S_{ij}) = f(\varepsilon_{ij1} > -\lambda_{ij1}, \varepsilon_{ij2} = \ln(T_{ij}) - \lambda_{ij2}, \varepsilon_{ij3} \geq 1 - \lambda_{ij3}) \), or,
\[
(A.2) \quad f(T_{ij}, S_{ij} | z_i, x_i, \Theta) = \Phi(\lambda_{ij1}) \times \frac{1}{\sigma_{j2}} \varphi \left( \frac{\ln(T_{ij}) - \lambda_{ij2}}{\sigma_{j2}} \right) \times \Phi \left( \frac{\lambda_{ij3} - 1}{\sigma_{j3}} \right) = \text{Prob}_{ij[2]}(z_i);
\]

When \( T_{ij} > 0 \) and \( S_{ij} = 0 \), \( f(T_{ij}, S_{ij}) = f(\varepsilon_{ij1} > -\lambda_{ij1}, \varepsilon_{ij2} = \ln(T_{ij}) - \lambda_{ij2}, \varepsilon_{ij3} \leq -\lambda_{ij3}) \), or,
\[
(A.3) \quad f(T_{ij}, S_{ij} | z_i, x_i, \Theta) = \Phi(\lambda_{ij1}) \times \frac{1}{\sigma_{j2}} \varphi \left( \frac{\ln(T_{ij}) - \lambda_{ij2}}{\sigma_{j2}} \right) \times \Phi \left( \frac{-\lambda_{ij3}}{\sigma_{j3}} \right) = \text{Prob}_{ij[3]}(z_i);
\]

When \( T_{ij} > 0 \) and \( 1 > S_{ij} > 0 \), \( f(T_{ij}, S_{ij}) = f(\varepsilon_{ij1} > -\lambda_{ij1}, \varepsilon_{ij2} = \ln(T_{ij}) - \lambda_{ij2}, \varepsilon_{ij3} = S_{ij} - \lambda_{ij3}) \), or,
\[
(A.4) \quad f(T_{ij}, S_{ij} | z_i, x_i, \Theta) = \Phi(\lambda_{ij1}) \times \frac{1}{\sigma_{j2}} \varphi \left( \frac{\ln(T_{ij}) - \lambda_{ij2}}{\sigma_{j2}} \right) \times \frac{1}{\sigma_{j3}} \varphi \left( \frac{S_{ij} - \lambda_{ij3}}{\sigma_{j3}} \right) \equiv \text{Prob}_{ij[4]}(z_i).
\]

\( \Phi(\cdot) \) and \( \varphi(\cdot) \) represent, respectively, cumulative and probability density functions of the standard normal distribution. Since \( z_i \) is unobservable from a modeler’s perspective, we need to integrate it out in deriving the likelihood contribution of individual \( i \), hence:
\[
(A.5) \quad L_i = \prod_{1} \text{Prob}_{ij[1]}(z_i) \prod_{2} \text{Prob}_{ij[2]}(z_i) \prod_{3} \text{Prob}_{ij[3]}(z_i) \prod_{4} \text{Prob}_{ij[4]}(z_i) \prod_{p=1}^{p} \varphi(z_{ip}) \ dt_1 \cdots dt_p,
\]
where \( \prod_1 \prod_2 \prod_3 \prod_4 \) and \( \prod_4 \) denote, respectively, the product over four different types of observations across all product categories -- an observation belongs to type 1 if \( T_{ij} = 0 \), type 2 if \( T_{ij} > 0 \) and \( S_{ij} = 1 \), type 3 if \( T_{ij} > 0 \) and \( S_{ij} = 0 \), or type 4 if \( T_{ij} > 0 \) and \( 1 > S_{ij} > 0 \).

For a given \( P \), the dimensionality of the unobserved customer characteristics, we estimate Model C by maximizing the sample likelihood function, which is defined as the product of the individual likelihood functions in Equation A.5 across all \( i \)'s in the calibration sub-sample of \( I \). We use simulation to evaluate the integrals (see Gourieroux and Montfort 2002 for an introduction). With a Simulated Maximum Likelihood (SML) estimator, the individual likelihood contributions in Equation A.5 are approximated as:

\[
L_i \approx \frac{1}{R} \sum_{r=1}^{R} \prod_{1} \prod_{2} \prod_{3} \prod_{4} \text{Prob}_{r(1)}(z_i^r) \text{Prob}_{r(2)}(z_i^r) \text{Prob}_{r(3)}(z_i^r) \text{Prob}_{r(4)}(z_i^r),
\]

where \( z_i^r \) is the \( r^{th} \) draw from a \( P \)-dimensional multivariate standard normal distribution, and \( R \) is the total number of draws taken. An appealing aspect of the SML estimator is that the simulated likelihood function in Equation A.6 is twice differentiable, simplifying likelihood maximization with gradient-based search algorithms. We use a unique Halton sequence (Train 2003) to simulate each dimension of \( z_i \), assigning a different set of draws to each subject. In determining the appropriate \( R \), we adopt the following heuristic: for any given \( P \), start with \( 50P \) draws, and then double the number of draws until no significant improvements can be obtained in estimator efficiency (determined by comparing the standard errors of parameter estimates based on different numbers of draws, as well as the associated likelihoods).

The foregoing discussion indicates how our model can be estimated for a given \( P \). To determine the number of factors, we use the Bayesian Information Criterion (BIC). More specifically, we start with \( P = 1 \), then \( P = 2 \), and so on, until the resulting BIC stops decreasing.
Appendix B. Imputation Procedure

After calibrating Model C, we are interested in making inferences about the unobserved customer heterogeneities, i.e., imputing $z_i$, the latent factor scores of each individual customer, by combining model parameter estimates and information available for individual $i$. Depending on the availability of information at the individual level, the formula for imputing $z_i$ varies. In the following we detail the imputation procedures for two scenarios of data availability.

Scenario 1 – Both $T_{ij}^*$’s and $S_{ij}^*$’s are observed (or, equivalently, both $Y_{ij}^1$’s and $Y_{ij}^0$’s are observed).

Recall $\prod_k Prob_{g[k]}(z_i)$ for $k = 1, 2, 3, \text{ or } 4$, as defined in Appendix A. We have the marginal density function for $T_{ij}^*$’s, $S_{ij}^*$’s and $z_i$ as follows:

$$f(T_{ij}, S_{ij}, \ldots T_{ij}, S_{ij}, z_i; x_i, \Theta) = \prod_{j=1}^{J} f(T_{ij}, S_{ij} | z_i; x_i, \Theta) \prod_{p=1}^{P} \phi(z_{ip}) =$$

$$\prod_{1}^{4} Prob_{g[1]}(z_i) \prod_{2}^{4} Prob_{g[2]}(z_i) \prod_{3}^{4} Prob_{g[3]}(z_i) \prod_{4}^{4} Prob_{g[4]}(z_i) \prod_{p=1}^{P} \phi(z_{ip}) \equiv Q_i(z_i)$$

Since $T_{ij}^*$’s and $S_{ij}^*$’s are observed, the above density function can be calculated for any $z_i$, which allows us to impute customer $i$’s latent factor scores by maximizing $Q_i(z_i)$ over $z_i$. Or, formally, $\hat{Z}_i(T_{ij}, S_{ij}, \ldots T_{ij}, S_{ij}, x_i, \Theta) = \arg \max Q_i(z_i)$.

Scenario 2 – Only $Y_{ij}^1$’s is observed, while $T_{ij}^*$’s, $S_{ij}^*$’s and $Y_{ij}^0$’s are unobserved.

Again, the marginal density function for $Y_{ij}^1$’s and $z_i$ are:

$$f(Y_{ij}^1, \ldots Y_{ij}^1, z_i; x_i, \Theta) = \prod_{j=1}^{J} f(Y_{ij}^1 | z_i; x_i, \Theta) \prod_{p=1}^{P} \phi(Z_{ip}) .$$

When $f(Y_{ij}^1 | z_i; x_i, \Theta)$ is available for $\forall j$, $f(Y_{ij}^1, \ldots Y_{ij}^1, z_i; x_i, \Theta)$ is defined:
\[
\hat{Z}_i(Y_i^1, \cdots Y_j^1, z_i; x_i, \hat{\Theta}) = \arg \max \left( f(Y_i^1, \cdots Y_j^1, z_i; x_i, \hat{\Theta}) \right) = \\
\arg \max \left( \prod_{j=1}^{J} f(Y_j^1 | z_i; x_i, \hat{\Theta}) \prod_{p=1}^{P} \varphi(Z_{yp}) \right).
\]  

(B.3)

Thus, when only \( Y_i^1 \)'s is observed, the central task in imputing \( z_i \) is calculating conditional density \( f(Y_i^1 | z_i; x_i, \hat{\Theta}) \). Depending on whether \( Y_i^1 = 0 \) or \( Y_i^1 > 0 \), different formulas are needed in calculating \( f(Y_i^1 | z_i; x_i, \hat{\Theta}) \). When \( Y_i^1 = 0 \), it can result from two kinds of situations, i.e., 1) \( T_{ij} = 0 \), or 2) \( T_{ij} > 0 \) and \( S_{ij} = 0 \). Formally, when \( Y_i^1 = 0 \), we have:

\[
f(Y_i^1 | z_i; x_i, \hat{\Theta}) = f(T_{ij} = 0 | z_i; x_i, \hat{\Theta}) + f(T_{ij} > 0, S_{ij} = 0 | z_i; x_i, \hat{\Theta})
\]

(B.4)

\[
= \Phi(-\lambda_{y_1}) + \varphi(\lambda_{y_1}) \times \varphi \left( \frac{-\lambda_{y_2}}{\sigma_{j2}} \right) = U_{[1]}(z_i)
\]

where \( \lambda_{y_k} = \alpha_{yk} + x_i^k \beta_{yk} + z_{y_k} \gamma_{yk} \) for \( k = 1, 2, \) or 3.

When \( Y_i^1 > 0 \), it can also result from two kinds of situations: 1) \( T_{ij} = Y_i^1 \) and \( S_{ij} = 1 \), or 2) \( T_{ij} \times S_{ij} = Y_i^1 \) and \( 1 > S_{ij} > 0 \). Formally, when \( Y_i^1 > 0 \), we have:

(B.5)

\[
f(Y_i^1 | z_i; x_i, \hat{\Theta}) = f(T_{ij} = Y_i^1, S_{ij} = 1 | z_i; x_i, \hat{\Theta}) + f(T_{ij} \times S_{ij} = Y_i^1, 1 > S_{ij} > 0 | z_i; x_i, \hat{\Theta}) =
\]

\[
f(T_{ij} = Y_i^1, S_{ij} = 1 | z_i; x_i, \hat{\Theta}) + \int_{1 > S > 0} f(T_{ij} = \frac{Y_i^1}{S}, S_{ij} = S, z_i; x_i, \hat{\Theta})f(S_{ij} = S | z_i; x_i, \hat{\Theta})dS =
\]

\[
\Phi(\lambda_{y_1}) \times \left[ \frac{1}{\sigma_{j2}} \varphi \left( \frac{\ln(Y_i^1) - \lambda_{y_2}}{\sigma_{j2}} \right) \frac{\lambda_{y_3} - 1}{\sigma_{j3}} + \int_{1 > S > 0} \frac{1}{\sigma_{j2}} \Phi \left( \frac{\ln(Y_i^1) - \lambda_{y_2}}{S \sigma_{j2}} \sigma_{j2} \right) \frac{1}{\sigma_{j3}} \varphi \left( \frac{S - \lambda_{y_3}}{\sigma_{j3}} \right) dS \right] = U_{[1]}(z_i)
\]

Combining the above:

39
(B.6)

\[
f(Y_{ij}^1, \cdots Y_{ij}^j, z_i, x_i, \hat{\Theta}) = \prod_{j=1}^{J} f(Y_{ij}^1 | z_i, x_i, \hat{\Theta}) \prod_{p=1}^{P} \phi(Z_{ip}) = \prod_{0} U_{i[0]}(z_i) \prod_{1} U_{i[1]}(z_i) \prod_{p=1}^{P} \phi(Z_{ip}) \equiv V_i(z_i)
\]

where \( \prod_{0} \) and \( \prod_{1} \) denote, respectively, the product over two different types of observed \( Y_{ij}^1 \)'s across all product categories -- an observation belongs to type 1 if \( Y_{ij}^1 = 0 \), or type 2 if \( Y_{ij}^1 > 0 \). To impute the latent factor scores when only \( Y_{ij}^1 \)'s is observed, we can maximize \( V_i(z_i) \) over \( z_i \).

Formally: \( \hat{Z}_i(Y_{i1}^1 \cdots Y_{ij}^1, x_i, \hat{\Theta}) = \arg \max (V_i(z_i)) \).

**Prediction of \( T_{ij} \)'s and \( S_{ij} \)'s**

Given \( \hat{\Theta} \), the estimated model parameters, and \( \hat{z}_i \)'s, the imputed latent factor scores, we can predict \( T_{ij} \)'s, customer \( i \)'s total category requirements, and \( S_{ij} \)'s, shares of category requirements that are served by the focal firm, conditional on: a) observed \( Y_{ij}^1 \)'s -- levels of category requirements that are served by the focal firm, and b) \( x_i \) -- observable customer characteristics.\(^{10}\) The rest of this section demonstrates the formulas for making these predictions.

**For categories where \( Y_{ij}^1 = 0 \), we can predict, for example --**

The expected value of \( T_{ij} \) conditional on \( T_{ij} > 0 \)

\[
E(T_{ij} | T_{ij} > 0, Y_{ij}^1 = 0, \hat{z}_i; x_i, \hat{\Theta}) = \exp \left( \lambda_{ij2} + \frac{\sigma^2_{j2}}{2} \right);
\]

The expected value of \( S_{ij} \) conditional on \( T_{ij} > 0 \)

\[\]

---

\(^{10}\) Instead of plugging in \( \hat{z}_i \), an alternative is to integrate \( z_i \) out, which in practice leads to predictions that are of no significant differences, but could take significantly more time to implement when the dimensionality of \( z_i \) is high.
Similarly, for categories where $Y^i > 0$, we can make the above predictions as –

The expected value of $T_{ij}$ conditional on $T_{ij} > 0$

$$E(T_{ij} | T_{ij} > 0, Y^i = Y > 0, z_i; x_i, \Theta) =$$

$$\left(\frac{Y^i}{1}\right) \frac{1}{\sigma_j} \varphi\left(\frac{\ln\left(\frac{Y^i}{S}\right) - \lambda_{y_j}}{\sigma_j}\right) \Phi\left(\frac{\lambda_{y_j} - 1}{\sigma_j}\right) + \int_{1 > S > 0} \frac{Y^i}{S} \frac{1}{\sigma_j} \varphi\left(\frac{\ln\left(\frac{Y^i}{S}\right) - \lambda_{y_j}}{\sigma_j}\right) \frac{1}{\sigma_j} \varphi\left(\frac{S - \lambda_{y_j}}{\sigma_j}\right) dS$$

(B.9)

The expected value of $S_{ij}$ conditional on $T_{ij} > 0$

$$E(S_{ij} | T_{ij} > 0, Y^i = Y > 0, z_i; x_i, \Theta) =$$

$$1 \times \frac{1}{\sigma_j} \varphi\left(\frac{\ln\left(\frac{Y^i}{S}\right) - \lambda_{y_j}}{\sigma_j}\right) \Phi\left(\frac{\lambda_{y_j} - 1}{\sigma_j}\right) + \int_{1 > S > 0} S \times \frac{1}{\sigma_j} \varphi\left(\frac{\ln\left(\frac{Y^i}{S}\right) - \lambda_{y_j}}{\sigma_j}\right) \frac{1}{\sigma_j} \varphi\left(\frac{S - \lambda_{y_j}}{\sigma_j}\right) dS$$

(B.10)
## Appendix C. Market Audit Data and Reporting Errors

<table>
<thead>
<tr>
<th>Account Balances Inside the Focal Bank</th>
<th>From Market Audit</th>
<th>From Internal Records</th>
<th>Reporting Error = Market Audit − Internal Records</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Standard Deviation</td>
<td>Mean</td>
</tr>
<tr>
<td>Non-interest Checking</td>
<td>1939</td>
<td>4125</td>
<td>1906</td>
</tr>
<tr>
<td>Interest Checking</td>
<td>2650</td>
<td>8356</td>
<td>2581</td>
</tr>
<tr>
<td>Savings</td>
<td>5686</td>
<td>16456</td>
<td>5152</td>
</tr>
<tr>
<td>CDs</td>
<td>3165</td>
<td>19513</td>
<td>3132</td>
</tr>
<tr>
<td>Investments</td>
<td>2978</td>
<td>22111</td>
<td>2924</td>
</tr>
<tr>
<td>Car Loan</td>
<td>902</td>
<td>3871</td>
<td>919</td>
</tr>
<tr>
<td>Personal Loan</td>
<td>837</td>
<td>7148</td>
<td>845</td>
</tr>
<tr>
<td>Line of Credit</td>
<td>860</td>
<td>5025</td>
<td>847</td>
</tr>
<tr>
<td>Credit Card</td>
<td>549</td>
<td>1590</td>
<td>542</td>
</tr>
<tr>
<td>Mortgage</td>
<td>11540</td>
<td>45000</td>
<td>11645</td>
</tr>
</tbody>
</table>

*: The mean reporting error is significant at p < .01 level.
Table 1
Structure of the Dataset Used in Empirical Illustration

<table>
<thead>
<tr>
<th>Inside the Bank</th>
<th>Outside the Bank</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Calibration Sample with 23,957 Subjects</em></td>
<td></td>
</tr>
<tr>
<td>Total and share of category requirements available by combining internal balances with external balances from a market audit study conducted by a syndicated service</td>
<td></td>
</tr>
<tr>
<td>Internal Balances in 10 Accounts</td>
<td>External Balances in 10 Accounts from Syndicated Service</td>
</tr>
<tr>
<td>Household Income</td>
<td>(Hence Total and Share of Category Requirements Observed)</td>
</tr>
<tr>
<td>Length of Relationship with the Bank</td>
<td></td>
</tr>
<tr>
<td><em>Validation Sample with 10,185 Subjects</em></td>
<td></td>
</tr>
<tr>
<td>Mimicking the remainder of the customer base where information on total and share of category requirements is unavailable</td>
<td></td>
</tr>
<tr>
<td>Internal Balances in 10 Accounts</td>
<td>Total and Share of Category Requirements to Be Imputed</td>
</tr>
<tr>
<td>Household Income</td>
<td>(Performance Evaluated in Relation to Available Data from Syndicated Service)</td>
</tr>
<tr>
<td>Length of Relationship with the Bank</td>
<td></td>
</tr>
</tbody>
</table>
## Table 2

**Sample Summary Statistics**

<table>
<thead>
<tr>
<th>Product Category</th>
<th>% of Subjects with Positive Total Category Requirements</th>
<th>Average Total Category Requirements ($)</th>
<th>% of Subjects with Share of Category Requirements = 0</th>
<th>% of Subjects with 0 &lt; Share of Category Requirements &lt; 1</th>
<th>% of Subjects with Share of Category Requirements = 1</th>
<th>Average Share of Category Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sub-sample</strong>*</td>
<td>Calibration Validation</td>
<td>Calibration Validation</td>
<td>Calibration Validation</td>
<td>Calibration Validation</td>
<td>Calibration Validation</td>
<td>Calibration Validation</td>
</tr>
<tr>
<td>Total Assets</td>
<td>99.8 99.7</td>
<td>95,191 (226315) **</td>
<td>91,936 (212861) **</td>
<td>20.3 20.1</td>
<td>60.0 59.4</td>
<td>19.7 20.5</td>
</tr>
<tr>
<td>Non-interest Checking</td>
<td>60.4 60.4</td>
<td>5,697 (9264)</td>
<td>5,810 (10694)</td>
<td>29.6 29.2</td>
<td>18.8 18.8</td>
<td>51.5 52.0</td>
</tr>
<tr>
<td>Interest Checking</td>
<td>52.0 51.7</td>
<td>10,315 (20862)</td>
<td>9,420 (17317)</td>
<td>38.9 38.4</td>
<td>17.3 16.9</td>
<td>43.8 44.6</td>
</tr>
<tr>
<td>Savings</td>
<td>79.5 79.4</td>
<td>19,912 (56922)</td>
<td>18,800 (44414)</td>
<td>43.2 42.8</td>
<td>23.6 24.6</td>
<td>33.2 32.6</td>
</tr>
<tr>
<td>CDs</td>
<td>19.4 20.2</td>
<td>49,126 (94765)</td>
<td>48,396 (99631)</td>
<td>52.7 54.6</td>
<td>13.2 11.4</td>
<td>34.1 34.0</td>
</tr>
<tr>
<td>Investments</td>
<td>55.8 55.3</td>
<td>108,940 (242226)</td>
<td>105,996 (231246)</td>
<td>89.0 89.0</td>
<td>6.5 6.2</td>
<td>4.5 4.9</td>
</tr>
<tr>
<td>Total Debts</td>
<td>87.3 87.2</td>
<td>79,857 (112534)</td>
<td>78,931 (108140)</td>
<td>43.4 43.7</td>
<td>48.4 48.3</td>
<td>8.2 8.0</td>
</tr>
<tr>
<td>Car Loan</td>
<td>46.3 46.7</td>
<td>14,106 (12095)</td>
<td>14,065 (11986)</td>
<td>80.9 81.6</td>
<td>5.9 5.6</td>
<td>13.3 12.8</td>
</tr>
<tr>
<td>Personal Loan</td>
<td>24.8 24.7</td>
<td>18,746 (29687)</td>
<td>18,637 (26938)</td>
<td>81.3 80.5</td>
<td>4.2 4.5</td>
<td>14.5 15.0</td>
</tr>
<tr>
<td>Line of Credit</td>
<td>18.7 18.1</td>
<td>19,521 (24737)</td>
<td>19,578 (32399)</td>
<td>50.0 50.7</td>
<td>7.2 7.5</td>
<td>42.8 41.8</td>
</tr>
<tr>
<td>Credit Card</td>
<td>66.3 65.7</td>
<td>4,713 (7621)</td>
<td>4,682 (6901)</td>
<td>60.8 60.5</td>
<td>26.0 26.2</td>
<td>13.1 13.3</td>
</tr>
<tr>
<td>Mortgage</td>
<td>52.1 52.3</td>
<td>102,546 (117592)</td>
<td>100,760 (111205)</td>
<td>74.0 73.5</td>
<td>2.0 2.0</td>
<td>24.0 24.5</td>
</tr>
</tbody>
</table>

*: For subjects whose total account balances are greater than zero; **: standard deviation in parentheses; ***: # of subjects in the calibration sample = 23,957; # in the validation sample = 10,185
Table 3
Predictive Performance Comparisons in the Validation sample

<table>
<thead>
<tr>
<th>Measure</th>
<th>Predicting Outside Product Ownership (a)</th>
<th>Predicting Total Category Requirements (b)</th>
<th>Predicting Share of Category Requirements (c)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Naïve MAD</td>
<td>Percentual Improvement in MAD</td>
<td>Naïve MAD</td>
</tr>
<tr>
<td></td>
<td>Model A</td>
<td>Model B</td>
<td>Model C</td>
</tr>
<tr>
<td>Total Wallet</td>
<td>0.73</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>Total Assets</td>
<td>0.73</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>Non-Interest Checking</td>
<td>0.71</td>
<td>0.69</td>
<td>0.70</td>
</tr>
<tr>
<td>Interest Checking</td>
<td>0.74</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>Savings</td>
<td>0.66</td>
<td>0.64</td>
<td>0.64</td>
</tr>
<tr>
<td>CDs</td>
<td>0.87</td>
<td>0.87</td>
<td>0.86</td>
</tr>
<tr>
<td>Investments</td>
<td>0.68</td>
<td>0.68</td>
<td>0.67</td>
</tr>
<tr>
<td>Total Debts</td>
<td>0.73</td>
<td>0.71</td>
<td>0.73</td>
</tr>
<tr>
<td>Car Loan</td>
<td>0.63</td>
<td>0.62</td>
<td>0.63</td>
</tr>
<tr>
<td>Personal Loan</td>
<td>0.79</td>
<td>0.79</td>
<td>0.79</td>
</tr>
<tr>
<td>Line of Credit</td>
<td>0.89</td>
<td>0.89</td>
<td>0.89</td>
</tr>
<tr>
<td>Credit Card</td>
<td>0.62</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>Mortgage</td>
<td>0.71</td>
<td>0.65</td>
<td>0.71</td>
</tr>
</tbody>
</table>
## Table 4
Fit Measures for the Predictive Models in the Calibration Sample

<table>
<thead>
<tr>
<th></th>
<th>Model A</th>
<th>Model B</th>
<th>Model C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Univariate Type-2 Tobit Regression Model</td>
<td>Multivariate Latent Factor Model</td>
<td>Multivariate Share*Total Latent Factor Model</td>
</tr>
<tr>
<td># of Factors</td>
<td>Not Applicable</td>
<td>P = 0</td>
<td>P = 1</td>
</tr>
<tr>
<td># of Parameters</td>
<td>460</td>
<td>140</td>
<td>180</td>
</tr>
<tr>
<td>BIC</td>
<td>529695</td>
<td>895825</td>
<td>878446</td>
</tr>
<tr>
<td>CAIC</td>
<td>530155</td>
<td>895965</td>
<td>878626</td>
</tr>
</tbody>
</table>

*: The smallest BIC and hence the number of factors chosen.
Table 5a
Model C Parameter Estimates for Observed Customer Characteristics (# of Parameters = 90)*

<table>
<thead>
<tr>
<th>Category</th>
<th>Ownership Decision (Conditional on Ownership)</th>
<th>Total Decision (Conditional on Ownership)</th>
<th>Share Decision (Conditional on Ownership)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha_1$ Intercept</td>
<td>$\beta_1$ (income)</td>
<td>$\beta_1$ (tenure)</td>
</tr>
<tr>
<td>Assets</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-interest</td>
<td>0.78</td>
<td>0.07</td>
<td>-0.09</td>
</tr>
<tr>
<td>Checking</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interest</td>
<td>0.16</td>
<td>1.12</td>
<td>0.37</td>
</tr>
<tr>
<td>Checking</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Savings</td>
<td>0.91</td>
<td>0.49</td>
<td>0.06</td>
</tr>
<tr>
<td>CDs</td>
<td>-0.96</td>
<td>0.20</td>
<td>0.24</td>
</tr>
<tr>
<td>Investments</td>
<td>0.14</td>
<td>0.79</td>
<td>0.09</td>
</tr>
<tr>
<td>Debts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Car Loan</td>
<td>-0.09</td>
<td>0.51</td>
<td>-0.20</td>
</tr>
<tr>
<td>Personal Loan</td>
<td>-0.70</td>
<td>0.28</td>
<td>-0.20</td>
</tr>
<tr>
<td>Line of Credit</td>
<td>-0.96</td>
<td>0.41</td>
<td>0.11</td>
</tr>
<tr>
<td>Credit Card</td>
<td>0.45</td>
<td>0.24</td>
<td>-0.08</td>
</tr>
<tr>
<td>Mortgage</td>
<td>0.03</td>
<td>0.78</td>
<td>0.11</td>
</tr>
</tbody>
</table>

*: Parameter estimates that are significant at p < .01 level are in bold. ** Income was log-transformed.
Table 5b
Model C Parameter Estimates for Factor Loadings and Variances (# of Parameters = 80)*

<table>
<thead>
<tr>
<th>Category</th>
<th>Ownership Decision</th>
<th>Total Decision (Conditional on Ownership)</th>
<th>Share Decision (Conditional on Ownership)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\gamma_{1,1}$</td>
<td>$\gamma_{2,1}$</td>
<td>$\gamma_{1,2}$</td>
</tr>
<tr>
<td>Non-interest Checking</td>
<td>-2.76</td>
<td>-0.30</td>
<td>-0.01</td>
</tr>
<tr>
<td>Interest Checking</td>
<td>4.01</td>
<td>1.15</td>
<td>0.67</td>
</tr>
<tr>
<td>Savings</td>
<td>0.16</td>
<td>0.27</td>
<td>0.50</td>
</tr>
<tr>
<td>CDs</td>
<td>0.34</td>
<td>0.15</td>
<td>0.40</td>
</tr>
<tr>
<td>Investments</td>
<td>0.10</td>
<td>0.16</td>
<td>0.42</td>
</tr>
<tr>
<td>Debts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Car Loan</td>
<td>-0.25</td>
<td>-0.17</td>
<td>-0.03</td>
</tr>
<tr>
<td>Personal Loan</td>
<td>-0.21</td>
<td>-0.17</td>
<td>-0.02</td>
</tr>
<tr>
<td>Line of Credit</td>
<td>-0.03</td>
<td>-0.18</td>
<td>0.09</td>
</tr>
<tr>
<td>Credit Card</td>
<td>-0.21</td>
<td>-0.22</td>
<td>-0.12</td>
</tr>
<tr>
<td>Mortgage</td>
<td>-0.14</td>
<td>-0.23</td>
<td>-0.02</td>
</tr>
</tbody>
</table>

*: Parameter estimates that are significant at p < .01 level are in bold.
Figure 1
Segmentation Based on Share of Wallet and Total Wallet

Inside: 23K
Outside: 498K

Top Quintile: 14.3
2nd Quintile: 10.1
3rd Quintile: 12.2
4th Quintile: 13.5
Bottom Quintile: 4.6

Small: Share of Wallet (0, 0.2] (0.2, 0.4] (0.4, 0.6] (0.6, 0.8] (0.8, 1]
Large: Share of Wallet (0, 0.2] (0.2, 0.4] (0.4, 0.6] (0.6, 0.8] (0.8, 1]
Figure 2a
Factor Loadings on OWNERSHIP and TOTAL Decisions

Decisions
○: ownership (_o)
Δ: total (_t)

Assets
non: non-interest checking
int: interest checking
sav: savings
cds: CDs
invest: investments

Debts
car: car loan
pl: personal loan
loc: line of credit
cc: credit card
mor: mortgage
Figure 2b

Factor Loadings on SHARE and TOTAL Decisions

Decisions
- ○: share (\_s)
- △: total (\_t)

Assets
- non: non-interest checking
- int: interest checking
- sav: savings
- cds: CDs
- invest: investments

Debts
- car: car loan
- pl: personal loan
- loc: line of credit
- cc: credit card
- mor: mortgage
Figure 3a
Surface plot of customer distribution by *predicted* total assets percentile and share of assets in the validation sample

Figure 3b
Surface plot of customer distribution by *actual* total assets percentile and share of assets in the validation sample
References


