

國立交通大學

管理科學系碩士班

碩士論文

Customer Base Analysis: A Non-contractual Online
Retail Purchase Process Application

顧客基礎分析：以一非契約型之線上零售商的顧客
購買資訊為例

研究生：江品儀

指導教授：姜齊 教授

唐瓊璋 教授

中華民國九十六年六月

Customer Base Analysis: A Non-contractual Online Retail Purchase
Process Application

顧客基礎分析：以一非契約型之線上零售商的顧客購買資訊為例

研究生：江品儀

Student : Ping-Yi Chiang

指導教授：姜齊

教授

Advisor : Chi Chiang

唐瓊璋


教授

Edwin Tang

國立交通大學

管理科學系碩士班

碩士論文



A Thesis
Submitted to Department of Management Science
College of Management
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in
Management Science

June 2007

Hsinchu, Taiwan, Republic of China

中華民國九十六年六月

Abstract

In this research, we combine the BG/NBD (Fader et al., 2005) and the Extended SMC model (Schmittlein et al., 1994) to simultaneously and completely incorporate the past purchase behavior of customers to do some effective forecasts based on customer base analysis. Differed from the “entire” extended SMC model (based on Pareto/NBD), this research preserves and advocates the easy implementing of the BG/NBD and consider the past dollar volume spent by customers simultaneously by adding the Extended SMC model. Hence, our model is more suitable to be a basis for doing further CLV research than the “pure” BG/NBD, which doesn’t consider any “monetary” information of customers. Furthermore, based on the BG/NBD, we derive the equation of expected active probability for a random chosen customer. It could help us to understand the individual active probability and the true customer base of a firm after summing the active probabilities of all customers. We also empirically validate our model by using a database from an online VCD retailer and try to anticipate the possible purchase patterns of customers in the future both individually and collectively. And we also validate our model results through 1-way MANOVA to test and then we have statistical evidence to approve the differentiation capabilities of the key expected values. Finally, we transform the worksheet of the BG/NBD to a more user-friendly form. With this new worksheet, we only should put the basic purchase history of all customers into and then we could get the expected values of interests at one time. It could save a lot of time to implement this model especially when the base of customers is huge. And the other purpose is that we wish it could improve the utility rate of our model.

Key words: Pareto/NBD; BG/NBD; RFM; customer lifetime value

Abstract in Chinese

顧客關係管理與一對一行銷不僅是近年來行銷研究領域的熱門議題之一；實務上，許多企業也紛紛致力於深度經營顧客關係，以減少顧客流失率。本研究基於顧客關係管理中的顧客基礎分析(Customer Base Analysis)，利用已知的顧客購買行為(如一定期間內，顧客的購買次數、購買金額、最近一次購買日)等資訊，透過兩個主要的機率模型的結合：BG/NBD模型(Fader et al., 2005 a)與Extended SMC模型(Schmittlein et al., 1994)，加上基於BG/NBD模型的假設，我們推導出原模型並未導出的公式---任一隨機抽取顧客的期望存活率，試圖去預測未來某段期間內，個別與整體顧客的購買次數、平均單次購買金額及顧客存活率。同時，我們將使用一家線上日本動畫及影音VCD零售商店的資料庫，去從事實證分析；並透過單因子多變量分析的驗證，我們發現本模型的三種主要的預測結果，也可以成為良好的區別龐大顧客群的變數，顯示本研究除了可以成為未來顧客一生價值分析(Customer Lifetime Value)的基礎模型外，也同時是企業在落實一對一行銷的良好工具之一。最後，我們將BG/NBD模型原先提供的工作試算表格式，轉換成為一個更有效率、更方便使用的試算表格式，透過新的試算表，在輸入已知的顧客過去購買行為後，我們便可在同一時間、一次得到所有顧客在未來某期間內的預期購買模式。除了可方便模型的使用外，也期待此舉能更模型廣泛被使用並提高模型的價值。

關鍵字: Pareto/NBD; BG/NBD; RFM; 顧客一生價值

Acknowledgement

兩年的碩士生涯其實過的很快，在這兩年的學習中，很幸運的認識到一群好同學，尤其是同門的師姐妹，在論文撰寫過程中給予的相互支持與鼓勵。而最重要的，要十分感謝唐瓔璋老師，在這兩年中總共修習了唐老師所開設的四門行銷領域的課程，這些課程與大學時期所修習過的行銷課程相比，跳脫千篇一律的基礎行銷理論介紹，也不再僅以陳腔濫調的公司個案為例，「環球行銷」提供了一個更具國際觀的視野；「行銷研究專題」介紹了各種行銷相關之統計分析工具；「行銷工程」也介紹了不同行銷學域的重要理論與基礎模型，而這篇論文的研究動機，便是來自「行銷工程」中老師所介紹的貝氏統計概念與顧客關係管理領域中的預測模型。

同時，在論文撰寫的過程中，也感謝老師的細心指導與鼓勵，讓我能夠克服許多挫折與障礙，並順利找到研究的方向。

最後，我要感謝我的父母以及待在我身邊最重要的人，謝謝他們的支持，我才能夠奢侈地以非應屆的身份追求這個的碩士學位；當然，我也要將這個碩士學位，獻給我的父母，謝謝他們賜與我幸福溫馨的家庭生活，讓我無後顧之憂地放心追逐我的理想；也希望未來我能獻給他們更高的榮耀。

Table of Contents

ABSTRACT	I
ABSTRACT IN CHINESE	II
ACKNOWLEDGEMENT	III
TABLE OF CONTENTS	IV
LIST OF FIGURES	VI
LIST OF TABLES	VII
1. INTRODUCTION	1
2. LITERATURE REVIEW	3
2.1 CUSTOMER RELATIONSHIP MANAGEMENT	3
2.2 DATABASE MARKETING	4
2.3 ONE-TO-ONE MARKETING.....	4
2.4 CUSTOMER BASE ANALYSIS	5
2.5 CUSTOMER LIFETIME VALUE.....	5
2.6 THE BENEFITS OF STOCHASTIC CHOICE MODELS	6
2.7 PROBABILITY MIXED MODEL (SCHMITTLEIN, 1989)	7
2.8 THE PARETO/NBD MODEL (SCHMITTLEIN ET AL, 1987).....	8
2.8.1 Assumptions.....	8
2.8.2 Model Specification.....	9
2.8.3 Key Mathematical Results.....	10
2.8.4 Related Key Points about the Pareto/NBD Model.....	12
2.9 THE BG/NBD MODEL (FADER ET AL, 2005 A).....	12
2.9.1 Assumptions.....	12
2.9.2 Key Mathematical Results.....	14
2.9.3 Related Key Points about the BG/NBD Model	15
2.10 THE EXTENDED SMC MODEL (SCHMITTLEIN ET AL, 1994).....	16
2.10.1 Assumptions.....	16
2.10.2 Key Mathematical Results.....	17
2.10.3 Related Key Points about the Extended SMC Model.....	18
3. CONCEPTUAL FRAMEWORK	19
3.1 ASSUMPTIONS	19
3.1.1 Terms Definition.....	19
3.1.2 Assumptions.....	20
3.2 MODEL DEVELOPMENT FOR A RANDOMLY CHOSEN CUSTOMER.....	22
3.2.1 Expected Number of Transactions	22
3.2.2 Expected Active Probability.....	22
3.2.3 Expected Dollar Volume per Reorder	24
4. EMPIRICAL ANALYSIS	25
4.1 DATA DESCRIPTION	25
4.2 PARAMETER ESTIMATION	26

4.2.1 <i>Transaction/Dropout Process</i>	26
4.2.2 <i>Transaction Dollar Volume</i>	28
4.3 MODEL RESULTS	29
4.3.1 <i>Expected Number of Reorders (Conditional Expectation)</i>	29
4.3.2 <i>Expected Active Probability</i>	33
4.3.3 <i>Expected Dollar Volume per Reorder</i>	35
4.4 INTEGRATED INDIVIDUAL CUSTOMER FORECASTS	36
4.5 COLLECTIVE CUSTOMER FORECASTS	37
4.6 MODEL VALIDATION	37
4.6.1 <i>1-Way MANOVA</i>	38
4.6.2 <i>Validation Result</i>	41
5. CONCLUSION AND DISCUSSION	42
5.1 RESEARCH CONTRIBUTION.....	42
5.2 RESEARCH LIMITATION	42
5.3 FUTURE RESEARCH DIRECTION.....	43
REFERENCES	45



List of Figures

Figure1. Screenshot of Excel Worksheet of Raw Data.....26
Figure2. Screenshot of Excel Worksheet of Parameter Estimation.....27
Figure3. Screenshot of Excel Worksheet of Conditional Expectation.....31
Figure4. Screenshot of Excel Worksheet of Conditional Expectation—New Version.....32
Figure5. Screenshot of Excel Worksheet of Expected Active Probability.....34
Figure6. Screenshot of Excel Worksheet of Expected Dollar Volume.....35
Figure7. 1-Year Predictions for Illustrative Customer Accounts.....36



List of Tables

Table1. Correlation between Forecast Period Transaction Numbers.....	16
Table2. Model Estimation Results of Transaction/Dropout Process.....	27
Table3. Model Estimation Results of Transaction Dollar Volume.....	29
Table4. 1-way MANOVA Test Result---Overall Test.....	38
Table5. 1-way MANOVA Test Result---Marginal Test---EY.....	39
Table6. 1-way MANOVA Test Result---Marginal Test---EZ.....	40
Table7. 1-way MANOVA Test Result---Marginal Test---PRO.....	40
Table8. 1-Way MANOVA Results without Zero Class.....	41



1. Introduction

Customer relationship management (CRM) is becoming a central research paradigm in the marketing channel literatures (Heide, 1994). Within the so many research regions, database marketing is one facet of CRM. And it's also one instrument to implement CRM. If a company can address and target individual customers to implement one-to-one marketing, it can improve its profitability by serving these customers differently (Niraj et al. 2001). Customer base analysis is one part of database marketing and is also one tool to implement one-to-one marketing. In this research, we utilize some stochastic choice models to analyze a customer base of an online retailer and hope to predict future purchase behavior of the customers. Besides, if we want to calculate the lifetime value (LTV) of a customer, customer profitability models of current period costs and revenues as well as forecasting models of future revenues and costs are required (Niraj et al., 2001). Because our model is belonged to a forecasting model, it would be help to calculate the LTV of customers.

In this research, based on customer base analysis, we will implement two stochastic choice models by using a non-contractual online retail database. In a non-contractual setting, the time at which a customer becomes inactive is unobserved. The big challenge that faces non-contractual marketers is how to differentiate some customers who have indeed ended their relationship with a firm from other customers who are just in the midst of a long hiatus between transactions. In conclusion, we will use a transaction database from an online retail to empirically validate several stochastic choice models and intend to solve the following questions:

- ◆ Which individuals in this database are most likely to be active or inactive in the future?
- ◆ What level of transactions measured individually or collectively should be expected in the future?
- ◆ How much dollar volume attained individually or collectively could be

expected in the future?

Although the first question was the key issue of the Pareto/NBD model (Schmittlein et al., 1987), it could not be solved by the BG/NBD model (Fader et al, 2005 a) because the authors of this model did not derive the active probability function for a randomly chosen customer. So we will try to derive it based on the BG/NBD model and answer the first question. The last one could not be solved either by the BG/NBD model alone. Thus we will incorporate dollar volume by using the normal-normal mixture model jointly (also called the extended SMC model) (Schmittlein and Peterson, 1994) to increase the practical utility of the original BG/NBD model.

This paper is organized as follows. Section 2 presents an overview of literature about many constructs which have been mentioned above and three probability models for customer base analysis. Section 3 specifies our conceptual framework. In our framework, the main stochastic model is based on the BG/NBD model to capture the flow of transactions over time and a normal-normal mixture for spend per transaction. We will integrate the two models and try to develop an equation to capture the customer active probability based on BG/NBD model. Of course, we expected it could also be utilized in Microsoft Excel as well and it's also one of the contributions of the BG/NBD model especially for marketing practitioners. Section 4 empirically implements the models with a non-contractual online retail database and utilizes 1-way MANOVA to validate our model results. In section 5, we conclude with several issues that arise from this work, review several limitations in our research and suggest some future research directions.

2. Literature Review

2.1 Customer Relationship Management

The definitions of CRM in many literatures are a little bit different. One definition of CRM is utilizing software and other related technologies to automate and improve business processes in areas of sales, marketing, customer service and etc. Another one is that CRM is a measure for an organization to acquire new customers, retain original customers and increase the profitability of customers through continuous communications to understand and influence the behavior of customers. To sum it up in a sentence, we can say that CRM is an approach which uses data mining, information technology and integrated marketing communication to understand and communicate with customers and then influence their behavior. The ultimate objectives are increasing market share, decreasing churn rate of customers, recapturing lost customers and enhancing customer lifetime value.

CRM research can be organized along the customer lifecycle, including *customer acquisition, development* and *retention* strategies (Kamakura et al., 2005). *Customer acquisition* extends from the channels customers use to first access the firm (Ansari et al., 2004) to the promotion that bring them to the firm. If a firm uses appropriate *development* strategies such as delivering customized products (Ansari and Mela, 2003) and cross-selling (Kamakura et al., 1991 and 2003), it can enhance the value of a customer from the firm. Early detection and prevention of customer churn can also enhance the total lifetime of the customer, if a firm can put efforts on the *retention* of valuable customers (Kamakura et al., 2005).

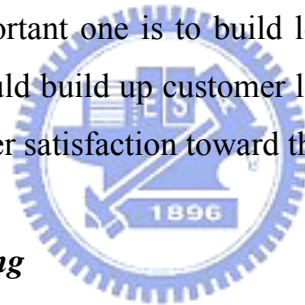
Why has CRM caught so much attention? One reason is that customer relationship management is a natural evolution from well-established market segmentation and target-marketing activities (Essential readings in marketing, MBR). Last but not least, if a firm implements CRM, profits will have some increase.

Profits can increase because of several reasons (Reichheld, 1996). First, by applying retention programs, customers are confronted with increasing switching

costs so they have fewer incentives to change their current behavior (Jones et al., 2000). Secondly, the longer a customer stays, the more he spends at the company. Otherwise they may be likely to convince others about the positive value the company offers (word-of-mouth effect). And they tend to be less price sensitive (Zeithaml et al., 1996) and would be less responsive to competitive pull (Stum and Thiry, 1991). All these could greatly promote the firm's profitability.

2.2 Database marketing

Database marketing is using information technology to construct and maintain a database system consisting of related information of current and latent customers. Then a company could exploit this database to provide customers better products or services. The most important one is to build long term relationships with them. In this way, a company could build up customer loyalty, decrease wasting of resources and enhance the customer satisfaction toward the company (Hughes, 1994).



2.3 One-to-one marketing

One-to-one marketing means being willing and able to change on what the customer tells you and what else you know about that customer (Peppers et al., 1999). Practiced correctly, one-to-one marketing can increase the value of the customer base of firms (Niraj et al. 2001). One-to-one marketing focuses on customer satisfaction and is customer-oriented (Weng and Liu, 2004). To implement one-to-one marketing, it is necessary to (a) identify customers; (b) differentiate customers; (C) interact with customers; and (d) personalize products or services to tailor-suit customers (Peppers et al., 1999).

2.4 Customer base analysis

Customer base analysis is concerned with using the observed past purchase behavior of customers to understand their current and likely future purchase patterns (Schmittlein and Peterson, 1994). Importantly, many choice modelers often find purchase history to be much more predictive than marketing mix variables such as price or promotions (Fader and Lattin, 1993; Guadagni and Little, 1983). In so many aspects that CRM puts emphasis on; customer base analysis especially focuses on retained customers. Retained customers could produce higher revenues and margin than new customers (Reichheld and Sasser, 1990). And a 1% improvement in retention can increase firm value by 5% (Gupta et al., 2004). Therefore it's reasonable and supported that firms should spend more marketing resources to retain existing customers rather than acquiring new ones (Rust and Zahorik, 1993; Mozer et al., 2000).



2.5 Customer Lifetime Value

In this customer-centric era, firms should focus on building and managing customer equity and not just brand equity. Customer equity is the sum of lifetime values (LTVs) of customers, where each customer's LTV is the sum of the properly discounted stream of net profits from the customer over the lifetime of the customer-firm relationship (Blattberg and Deighton, 1996). Customers generally interact with a firm over multiple periods. If we want to calculate the LTV of a customer, customer profitability models and forecasting models are required (Niraj et al., 2001). Within the two aspects, it's more complex to estimate future revenues and costs. Forecasting models, also called stochastic choice models, have been advanced to predict the likelihood of future events based on past history in marketing literature. Because they can be used to identify the likelihood of current customers being active in the future and to predict future revenues, they play an

important role in the calculation of LTV of customers.

2.6 The benefits of stochastic choice models

First of all, the development and application of stochastic choice models could help to formulate and adopt customer *retention* strategies, which belong to one scope of CRM research. Secondly, through utilizing stochastic choice models, it's one of the methods to identify and differentiate customers, which are the first two steps to implement one-to-one marketing. Thirdly, because of the forecasting capability of the stochastic choice models, they also take a big responsibility for the expectation and calculation of LTV of customers.

Although in our framework the main stochastic model is based on the BG/NBD model, the creative motivation behind the BG/NBD model was looking forward to be approximated to the Pareto/NBD model. Thus, the behavioral stories of the two models are almost close to each other.

The Pareto/NBD model (Schmittlein et al., 1987) is a benchmark model for customer-base analysis in a non-contractual setting. But its empirical application can be challenging, especially in terms of parameter estimation (Fader and Hardie, 2005 a). It's also one of the reasons for us to choose the BG/NBD model as the main stochastic model.

In this following, before the complete introduction of the models, we will show the basic form of a probability mixed model. Because all the three following models are categorized to probability mixed models. And then we will respectively introduce the Pareto/NBD model, the BG/NBD model and the normal-normal mixture (also called the extended SMC model) (Schmittlein and Peterson, 1994) in order.

2.7 Probability Mixed Model (Schmittlein, 1989)

Some models have been called probability mixture models, since they envision a mixture across individuals of heterogeneous probabilistic processes. Such a mixture model consists of two components. First, events for individuals are assumed to follow some stochastic process whose form is specified up to some parameter θ (which may be a vector), and second, the stochastic process observed may vary across individuals. And usually the general form of the process is assumed to remain the same over individuals, so the variation can be thought of as a distribution for the latent trait θ over the population. Then, letting a random variable X associated with an individual have a cumulative distribution function $F(X|\theta)$ and the variation in θ over the population have a distribution function $G(\theta|\gamma)$.

The distribution of X for an individual chosen at random is

$$H(X|\gamma) = \int_{-\infty}^{\infty} F(X|\theta) dG(\theta|\gamma)$$

And the population mean is

$$u = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x dF(x|\theta) dG(\theta|\gamma)$$

Thus, for an individual, the expectation of X conditioned on the latent trait θ is

$$J(\theta) \equiv E[X|\theta] = \int_{-\infty}^{\infty} x dF(x|\theta)$$

Therefore, given an initial observation that $X=x$, the conditional expectation of this characteristic (denoted by X^*) is

$$\begin{aligned} E[X^*|X=x] &= E_{\theta}[J(\theta)|X=x] \\ &= \int_{-\infty}^{\infty} J(\theta) dG(\theta|\gamma, X=x) \end{aligned}$$

where $G(\theta|\gamma, X=x)$ is an updated distribution of θ given the observation $X=x$.

Within a Bayesian framework, $G(\theta|\gamma)$ is the prior distribution while $G(\theta|\gamma, X=x)$ is the posterior. Consequently, the researchers may observe X -values of many

heterogeneous individuals and use these values to estimate the distribution $G(\theta|\gamma)$ (Morris, 1983).

After understanding the prototype of a probability mixed model, we hope it could help to comprehend the model developments of the three following models. Hereafter, we will specify the assumptions and development of the three models separately.

2.8 The Pareto/NBD Model (Schmittlein et al, 1987)

Before we introduce the Pareto/NBD model, we'll make some definitions for the concepts we'll mention later. This model will consider only purchasing or transaction events. That is, the amounts purchased by customers won't be considered in the model development. Each transaction event will be termed a "purchase." A customer who is still active will be termed "Alive," while a customer who has left for whatever reason will be termed "Dead." The observation time of a customer who is still alive at time 0 is T . During T of a customer, the customer will have made X purchases with the last purchase coming at t , $0 < t \leq T$. Hence, the information on this customer contains 3 elements:

$$\text{Information} = (X, t, T).$$

After these concepts have been defined, we will introduce the basic assumptions and the development processes of the Pareto/NBD model.

2.8.1 Assumptions

(1) Transactions by active customers

While active, the number of transactions X , made by each customer in time period of length t is a **Poisson** random variable with a purchase rate λ :

$$P[X = x/\lambda, \tau > T] = e^{-\lambda T} \frac{(\lambda T)^x}{x!}; \quad x = 0, 1, 2, \dots$$

$$E[X/\lambda, \tau > T] = \lambda T, \quad Var[X/\lambda, \tau > T] = \lambda T$$

(2) *Individual customer retention/dropout*

Each customer has an unobserved “lifetime” of length τ which is an **exponential** random variable with a dropout rate μ :

$$f(\tau/\mu) = \mu e^{-\mu\tau}; \quad \tau > 0$$

$$E[\tau/\mu] = \frac{1}{\mu} \quad Var[\tau/\mu] = \frac{1}{\mu^2}$$

(3) *Heterogeneity in purchase rates*

The purchase rate λ of different customers follows a **gamma** distribution across the population of customers with shape parameter γ and scale parameter α :

$$g(\lambda/\gamma, \alpha) = \frac{\alpha^\gamma}{\Gamma(\gamma)} \lambda^{\gamma-1} e^{-\alpha\lambda}; \quad \lambda > 0; \quad \gamma, \alpha > 0$$

(4) *Heterogeneity in dropout rates*

The dropout rate μ of different customers follows a **gamma** distribution across the population of customers with shape parameter s and scale parameter β :

$$h(\mu/s, \beta) = \frac{\beta^s}{\Gamma(s)} \mu^{s-1} e^{-\beta\mu}; \quad \mu > 0; \quad s, \beta > 0$$

(5) *Rates λ and μ are independent*

The purchase rate λ and dropout rate μ vary **independently** across customers.

2.8.2 Model Specification

(1) *Purchase event model---NBD model*

Based on assumption (1) & (3), purchases made by a sample of customers while they are active follow the NBD model (Poisson mixed with Gamma)

(Ehrenberg, 1972).

$$P[X = x/\gamma, \alpha, \tau > T] = C_x^{\alpha+\gamma-1} \left(\frac{\alpha}{\alpha+T} \right)^r \left(\frac{T}{\alpha+T} \right)^x; \quad x = 0, 1, 2, \dots$$

(2) *Duration model---Pareto model*

Based on assumption (2) & (4), the lifetime “ τ ” of a sample of customers follows the Pareto distribution of the second kind (Exponential mixed with Gamma) (Johnson and Kotz, 1970).

$$f(\tau/s, \beta) = \frac{s}{\beta} \left(\frac{\beta}{\beta + \tau} \right)^{s+1}, \quad \tau > 0;$$

$$E(\tau/s, \beta) = \frac{\beta}{s-1}, \quad s > 1$$

(3) *Thus, this combined purchase event model and duration model will be called the Pareto/NBD model. It has four parameters: γ , α , s and β*

2.8.3 Key Mathematical Results

There are the key results that derived from the above-cited distributions below.

(1) *The expected number of purchases in a time period of length T for a randomly chosen customer is*

$$E[X/\gamma, \alpha, s, \beta, T] = \frac{\gamma\beta}{\alpha(s-1)} \left[1 - \left(\frac{\beta}{\beta+T} \right)^{s-1} \right]$$

(2) *The probability for a randomly chosen customer being active in a time period of length T is*

$$P[\tau > T/\gamma, \alpha, s, \beta, X = x, t, T]$$

$$= \int_0^\infty \int_0^\infty P[\tau > T/\lambda, \mu, X = x, t, T] f(\lambda, \mu/r, \alpha, s, \beta, X = x, t, T) d\lambda d\mu$$

The result varies depending on the values of α and β . There are three conditions.

Case 1: $\alpha > \beta$

$$P[\tau > T/\gamma, s, \alpha > \beta, X = x, t, T] \\ = \left\{ 1 + \frac{s}{\gamma + x + s} \left[\left(\frac{\alpha + T}{\alpha + t} \right)^{\gamma + x} \left(\frac{\beta + T}{\alpha + t} \right)^s F(a_1, b_1; c_1; z_1(t)) - \left(\frac{\beta + T}{\alpha + T} \right)^s F(a_1, b_1; c_1; z_1(t)) \right] \right\}^{-1}$$

where $a_1 = \gamma + x + s$; $b_1 = s + 1$;
 $c_1 = \gamma + x + s + 1$; $z_1(y) = \frac{\alpha - \beta}{\alpha + y}$

Case 2: $\alpha < \beta$

$$P[\tau > T/\gamma, s, \alpha < \beta, X = x, t, T] \\ = \left\{ 1 + \frac{s}{\gamma + x + s} \left[\left(\frac{\alpha + T}{\beta + t} \right)^{\gamma + x} \left(\frac{\beta + T}{\beta + t} \right)^s F(a_2, b_2; c_2; z_2(t)) - \left(\frac{\alpha + T}{\beta + T} \right)^{\gamma + x} F(a_2, b_2; c_2; z_2(t)) \right] \right\}^{-1}$$

where $a_2 = \gamma + x + s$; $b_2 = \gamma + x$;
 $c_2 = \gamma + x + s + 1$; $z_2(y) = \frac{\beta - \alpha}{\beta + y}$

Case 3: $\alpha = \beta$

$$P[\tau > T/\gamma, s, \alpha = \beta, X = x, t, T] \\ = \left\{ 1 + \frac{s}{\gamma + x + s} \left[\left(\frac{\alpha + T}{\alpha + t} \right)^{\gamma + x + s} - 1 \right] \right\}^{-1}$$

In case 2&3, $F(a, b; c; z)$ is the Gauss hyper-geometric function (Abramowitz and Stegun, 1972, p.558). It can be computed using either numerical integration or the algorithms. The authors of the Pareto/NBD model used numerical integration to compute the Gauss hyper-geometric function.

(3) *The expected number of purchases in a future period of length T^* for a randomly chosen customer is*

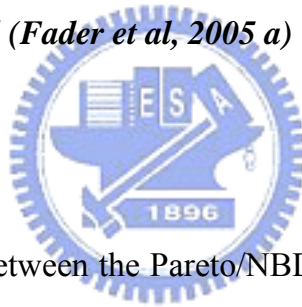
$$E[X^* | \gamma, \alpha, s, \beta, X = x, t, T, T^*] = E[X^* | \gamma + x, \alpha + T, s, \beta + T, T^*] \times P[\tau > T | \gamma, \alpha, s, \beta, X = x, t, T]$$

2.8.4 Related Key Points about the Pareto/NBD Model

- ◆ The likelihood function associated with the Pareto/NBD model is very complex because it involves many evaluations of the Gaussian hyper-geometric function. These multiple evaluations of the Gaussian hyper-geometric function are not only unfamiliar to most researchers working in the areas of database marketing and CRM analysis but are quite demanding from a computational standpoint (Fader et al., 2005).
- ◆ In the only published paper which successfully implemented the Pareto/NBD model by Reinartz and Kumar in 2003, the estimations of parameters have ever been commented on the associated computational burden as well(Reinartz and Kumar, 2003).

2.9 The BG/NBD Model (Fader et al, 2005 a)

2.9.1 Assumptions



The only difference between the Pareto/NBD model and the BG/NBD model lies in the behavior story about how or when customers become inactive. The Pareto/NBD model assumes that dropout can occur at any point in time whether or not actual purchases are taking place. However the BG/NBD model should be based on an assumption that dropout could only occur after an occurrence of an actual purchase and then we could model this process using the beta-geometric (BG) model.

The concepts such as “purchase”, “Alive”, “Dead” and “ T ” are defined the same as them in the Pareto/NBD model. The observation time of a customer who is still alive at time 0 is T .

There are also five assumptions behind the BG/NBD model.

(1) Transactions by active customers

While active, the number of transactions X , made by each customer in time

period of length t is a Poisson random variable with a purchase rate λ . It equals to assume that the time between transactions is distributed exponential with the same purchase rate λ :

$$f(t_j | t_{j-1}; \lambda) = \lambda e^{-\lambda(t_j - t_{j-1})}, \quad t_j > t_{j-1} \geq 0$$

(2) *Individual customer retention/dropout*

After any transaction, a customer could become inactive with a dropout rate p . So the point at which a customer becomes inactive follows a geometric distribution with dropout rate p :

$$\begin{aligned} &P(\text{inactive immediately after } j\text{th transaction}) \\ &= p(1-p)^{j-1}, \quad j = 1, 2, 3, \dots \end{aligned}$$

(3) *Heterogeneity in purchase rates*

The purchase rate λ of different customers follows a gamma distribution across the population of customers with shape parameter γ and scale parameter α :

$$f(\lambda | \gamma, \alpha) = \frac{\alpha^\gamma \lambda^{\gamma-1} e^{-\lambda\alpha}}{\Gamma(\gamma)}, \quad \lambda > 0$$

(4) *Heterogeneity in dropout rates*

The dropout rate p of different customers follows a beta distribution across the population of customers with two parameters a and b :

$$f(p | a, b) = \frac{p^{a-1}(1-p)^{b-1}}{B(a, b)}, \quad 0 \leq p \leq 1$$

(5) *Rates λ and p are independent*

The purchase rate λ and dropout rate p vary independently across customers.

Thus, based on assumption (1) and (3), the transaction model also belongs to the NBD model. And based on assumption (2) & (4), the retention model belongs to the BG model. This small and relatively inconsequential change to the original Pareto/NBD assumptions does not require any different psychological theories, nor does it have any

noteworthy managerial implications (Fader et al., 2005 a).

2.9.2 Key Mathematical Results


(1) The expected number of purchases in a time period of length t for a randomly chosen customer is

$$E(X(t)|\gamma, \alpha, a, b) = \frac{a+b-1}{a-1} \left[1 - \left(\frac{\alpha}{\alpha+t} \right)^\gamma {}_2F_1 \left(\gamma, b; a+b-1; \frac{t}{\alpha+t} \right) \right]$$

(2) The probability for a customer being still active at a point of time t is

$$\begin{aligned} P(\tau > t) &= P(\text{active at } t | \lambda, p) = \sum_{j=0}^{\infty} (1-p)^j \frac{(\lambda t)^j e^{-\lambda t}}{j!} \\ &= e^{-\lambda p t} \end{aligned}$$

(3) The expected number of purchases in a future period of length t for a randomly chosen customer is



$$\begin{aligned} &E(Y(t)|X = x, t_x, T, \gamma, \alpha, a, b) \\ &= \frac{\frac{a+b+x-1}{a-1} \left[1 - \left(\frac{\alpha+T}{\alpha+T+t} \right)^{\gamma+x} {}_2F_1 \left(\gamma+x, b+x; a+b+x-1; \frac{t}{\alpha+T+t} \right) \right]}{1 + \delta_{x>0} \frac{a}{b+x-1} \left(\frac{\alpha+T}{\alpha+t_x} \right)^{\gamma+x}} \end{aligned}$$

(4) The likelihood function for a randomly chosen customer with purchase history (X, t_x, T)

$$\begin{aligned}
L(\gamma, \alpha, a, b | X = x, t_x, T) &= \frac{B(a, b+x)}{B(a, b)} \frac{\Gamma(\gamma+x)\alpha^\gamma}{\Gamma(\gamma)(\alpha+T)^{\gamma+x}} + \delta_{x>0} \frac{B(a+1, b+x-1)}{B(a, b)} \frac{\Gamma(\gamma+x)\alpha^\gamma}{\Gamma(\gamma)(\alpha+t_x)^{\gamma+x}} \\
&= A_1 \cdot A_2 \cdot (A_3 + \delta_{x>0} A_4)
\end{aligned}$$

where

$$A_1 = \frac{\Gamma(\gamma+x)\alpha^\gamma}{\Gamma(\gamma)}, A_2 = \frac{\Gamma(a+b)\Gamma(b+x)}{\Gamma(b)\Gamma(a+b+x)}, A_3 = \left(\frac{1}{\alpha+T}\right)^{\gamma+x}, A_4 = \left(\frac{a}{b+x+1}\right)\left(\frac{1}{\alpha+t_x}\right)^{\gamma+x}$$

In order to implement the likelihood function in Microsoft Excel, the authors rewrote the original one and used A1~A4 to replace it.

(5) The sample log-likelihood function

Then suppose we have a sample of N customers, where customer i has transactions $X_i=x_i$ in the time period $(0, T_i]$ and their last transaction occurs at t_{xi} . Thus the sample log-likelihood function is

$$LL(\gamma, \alpha, a, b) = \sum_{i=1}^N \ln \left[L(\gamma, \alpha, a, b | X_i = x_i, t_{xi}, T_i) \right]$$

By this function, the authors of the BG/NBD model use the method of maximum likelihood to estimate the four parameters γ, α, a and b .

2.9.3 Related Key Points about the BG/NBD Model

- ◆ It's easy to implement the BG/NBD model because we could operate the whole model with a standard spreadsheet package (Excel), even the estimation of parameters.
- ◆ In the BG/NBD model, the authors didn't derive the formula about the probability that a random chosen customer is still active at a point of time t ($P(\tau > t) = P(\text{active at } t | \gamma, \alpha, a, b)$). It will somewhat diminish its usefulness if we want to know the size of the currently active customer pool and the rate at which that pool's size is increasing or decreasing. Therefore, to derive the formula is one of our main contributions in this research paper.
- ◆ One of the reasons that the authors wanted to develop the BG/NBD model

was to approximate the Pareto/NBD model. From the table 1 we could make a conclusion that the approximation effect is impressively good. (The table comes from (Fader et al.,2005 a))

Table 1 Correlation Between Forecast Period Transaction Numbers

	Actual	BG/NBD	Pareto/NBD
Actual	1.000		
BG/NBD	0.626	1.000	
Pareto/NBD	0.630	0.996	1.000

2.10 The Extended SMC Model (Schmittlein et al, 1994)

The Extended SMC model is a normal-normal mixture model and it incorporates dollar volume of transactions made by customers. This model got this name just because one of the builders of the Extended SMC model was the same with the one of the Pareto/NBD model, also called SMC model, and this model could also compensate the drawback of the Pareto/NBD model that didn't consider the dollar volume of transactions. This model has three assumptions shown below.

2.10.1 Assumptions

For a customer observed to have X reorders in a time period, they let Z_i denote the dollar volume of per order i ($i=1, \dots, X$)

(1) Individual customer level

The set of Z_i are *i.i.d.* normal random variables with mean θ and variance σ_w^2 which represents the variance in the dollar volume spent across reorders for a customer and is constant across customers.

$$Z_i \sim N(\theta, \sigma_w^2)$$

(2) *Heterogeneity*

Mean θ is assumed to vary across customers according to a normal distribution with mean $E[\theta]$ and variance σ_A^2 which represents the variance in average dollar volume spent across customers.

$$\theta \sim N(E[\theta], \sigma_A^2)$$

(3) *Rates λ , μ , and θ are independent from each other*

The average amount spent θ , the purchase rate λ and the dropout rate μ vary independently across customers where λ and μ have been clearly defined in the Pareto/NBD model.

Thus, based on assumption (1) and (2), the Extended SMC model belongs to a normal-normal mixture model.

2.10.2 Key Mathematical Results

(1) The reliability coefficient that the confidence one could place in a single observed dollar volume Z_i of past order is (relative to rely on the population average dollar volume $E[\theta]$):

$$\rho_1 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_W^2}$$

(2) If the observed number of reorder equaled to 1, $X=1$, the best estimate for θ is (Schmittlein, 1989):

$$E[\theta|Z_1] = \rho_1 Z_1 + (1 - \rho_1) E[\theta]$$

(3) If $X > 1$, the reliability coefficient of $\bar{Z} = \frac{1}{X} \sum_{i=1}^X Z_i$ is:

$$\rho_X = \frac{\sigma_A^2}{\sigma_A^2 + (\sigma_W^2/X)}$$

(4) The expected future volume per reorder is:

$$E[\theta|Z_1, \dots, Z_X] = \left(\frac{X\sigma_A^2}{X\sigma_A^2 + \sigma_W^2} \right) \bar{Z} + \left(\frac{\sigma_A^2}{X\sigma_A^2 + \sigma_W^2} \right) E[\theta]$$

(5) For a customer with a given purchase information (X, t, T, \bar{Z}) , the expected future dollar volume could be multiply “the expected future volume per reorder” (Schmittlein et al, 1994) by “the expected number of purchases in a future period of length T^* ” (Schmittlein et al, 1987).

2.10.3 Related Key Points about the Extended SMC Model

- ◆ The first two assumptions in the Extended SMC model were convenient but unnecessary. Because the formula, “the expected future volume per reorder”, could also be derived from the standpoint of minimizing squared prediction error in θ without the normality assumptions (Gerber, 1979, Chapter 6).
- ◆ Thus the third assumption is a “true” constraint when we use the model. In the original paper (Schmittlein et al, 1994).
- ◆ Although the Extended SMC model was named like this, the estimation of the parameter θ and the model development had no relationship with the four parameters in the Pareto/NBD model (SMC model). The three basic assumptions of the Extended SMC model could also be hold the same while we use the BG/NBD model to substitute for the Pareto/NBD model in the recommended computation formula of “the expected future dollar volume”.

Thus in our following framework, we will use the BG/NBD model to capture the transaction/dropout process of customers and the Extended SMC model to incorporate the dollar volume of transactions.

3. Conceptual Framework

This research will be based on customer base analysis and simultaneously use the three important customer purchase information: R (recency), F (frequency), and M (monetary) to predict the three aspects of purchase patterns.

First of all, we will use the BG/NBD model to get “the expected transactions”. Secondly, we will derive the formula of “the projected active rate” by ourselves and utilize it. Third, we will use the extended SMC model to compute “the expected dollar volume spent per reorder”. After multiplying “the expected transactions” by “the expected dollar volume spent per reorder”, we could get “the expected future dollar volume” of each customer. So far we could solve the three questions above in the Introduction.

Lastly, we will utilize 1-way MANOVA to test the model validation. We will use 80/20 rules to classify the sample of customers into two kinds, where one kind of customer whose total number is only 20% of the entire sample but contributes toward 80% of total sales volumes. The two kinds of customers are named class 1 and 2 separately. We use class 1 & 2 as independent variables and use “the expected number of transactions”, “the expected active probability” and “the expected dollar volume spent per reorder” as dependent variables simultaneously to employ a 1-way MANOVA analysis. If we have statistical evidences to support the three dependent variables, it could explain the total variances a lot no matter if it’s collectively or individually. We will say with much confidence that “the expected number of transactions”, “the expected active probability” and “the expected dollar volume spent per reorder” could be treated as successful discrimination variables to differentiate customers and be a good starting point to achieve so called one-to-one marketing and implement other CRM strategies.

3.1 Assumptions

3.1.1 Terms Definition

- (1) Each transaction event is being termed as a “purchase”.

- (2) Each customer who is active is being termed as “Alive”, while each customer has left for any reason is being termed “Dead”.
- (3) The observation time of each customer who is alive at time 0 is T .
- (4) During the observation time “ T ”, each customer will have made X purchases and his last purchase will occur at t_x , $0 < t_x < T$.
- (5) If a customer observed to have X purchases in a time period “ T ”, we could denote Z_i as the dollar volume of per order i ($i=1, \dots, X$), and denote $\bar{Z} = \frac{1}{X} \sum_{i=1}^X Z_i$ as the average dollar volume spent by the customer.
- (6) ρ is denoted a reliability coefficient. While estimating the expected dollar volume of a customer, we could use it to represent our confidence in utilizing observed past order amount relative to relying on the population average order amount. If $X=1$, we use ρ_1 . If $X > 1$, we use ρ_x .
- (7) Hence, the information on each customer contains 4 elements:

$$\text{Information} = (X, t_x, T, Z_i).$$

Within the 4 elements, we could also correspond the three basic CRM variables “Recency” (R), “Frequency” (F), “Monetary” (M) to t_x , X , Z_i respectively.

3.1.2 Assumptions

There are seven assumptions in our framework. The first four are identical to the ones of the BG/NBD model and the last three are the same with the ones of the Extended SMC model.

(1) Transactions by active customers follow **Poisson** distribution

$$f(t_j | t_{j-1}; \lambda) = \lambda e^{-\lambda(t_j - t_{j-1})}, \quad t_j > t_{j-1} \geq 0$$

While active, the number of transactions X , made by each customer in time period of length t is a Poisson random variable with a purchase rate λ . It means that the time between transactions is distributed exponential with the

same purchase rate λ .

- (2) *Individual customer retention/dropout probability follows **Geometric** distribution*

$$\begin{aligned} P(\text{inactive immediately after } j\text{th transaction}) \\ = p(1-p)^{j-1}, \quad j=1, 2, 3, \dots \end{aligned}$$

After any transaction, a customer could become inactive with a dropout rate p . So the point at which a customer becomes inactive follows a geometric distribution with dropout rate p .

- (3) *Heterogeneity in purchase rates λ which follows **Gamma** distribution*

$$f(\lambda | \gamma, \alpha) = \frac{\alpha^\gamma \lambda^{\gamma-1} e^{-\lambda\alpha}}{\Gamma(\gamma)}, \quad \lambda > 0$$

The purchase rate λ of different customers follows a gamma distribution across the population of customers with shape parameter γ and scale parameter α .

- (4) *Heterogeneity in dropout rates p which follows **Beta** distribution*

$$g(p | a, b) = \frac{p^{a-1}(1-p)^{b-1}}{B(a, b)}, \quad 0 \leq p \leq 1$$

The dropout rate p of different customers follows a beta distribution across the population of customers with two parameters a and b .

- (5) *The dollar volume of per order (Z_i) of individual customer follows **Normal** distribution*

$$Z_i \sim N(\theta, \sigma_w^2)$$

The set of Z_i are *i.i.d.* normal random variables with mean θ and variance σ_w^2 which represents the variance in the dollar volume spent across reorders for a customer and is constant across customers.

- (6) *Heterogeneity in average amount spent per order θ which follows **Normal** distribution*

$$\theta \sim N(E[\theta], \sigma_A^2)$$

Mean θ is assumed to vary across customers according to a normal distribution with mean $E[\theta]$ and variance σ_A^2 which represents the variance in average dollar volume spent across customers.

(7) Rates λ , p , and θ are independent from each other

The average amount spent per order θ , the purchase rate λ and the dropout rate p vary independently across customers.

3.2 Model Development for a Randomly Chosen Customer

3.2.1 Expected Number of Transactions

The expected number of transactions in the time period $(T, T+t]$ for a randomly chosen individual with observed purchase behavior (X, t_x, T) (The BG/NBD model, Equation (10))

$$E(Y(t)|X = x, t_x, T, \gamma, \alpha, a, b) = \frac{a+b+x-1}{a-1} \left[1 - \left(\frac{\alpha+T}{\alpha+T+t} \right)^{\gamma+x} {}_2F_1 \left(\gamma+x, b+x; a+b+x-1; \frac{t}{\alpha+T+t} \right) \right] = \frac{1 + \delta_{x>0} \frac{a}{b+x-1} \left(\frac{\alpha+T}{\alpha+t_x} \right)^{\gamma+x}}{1} \quad (1)$$

3.2.2 Expected Active Probability

(1) The active probability at the individual level

For the case where purchases were made in the period $(0, T]$, the probability that a customer with purchase behavior (X, t_x, T) , is still active at T , conditional on λ and p , is simply the probability that he did not drop out at t_x and made no purchases in $(t_x, T]$, divided by the probability of making no purchases in $(0, T]$. (The BG/NBD model, Appendix)

$$P(\text{active at } T|X = x, t_x, T, \lambda, p) = \frac{(1-p)e^{-\lambda(T-t_x)}}{p + (1-p)e^{-\lambda(T-t_x)}} \quad (2)$$

(2) For easily calculation, multiplying equation (2) by $\frac{[(1-p)^{x-1} \lambda^x e^{-\lambda t_x}]}{[(1-p)^{x-1} \lambda^x e^{-\lambda t_x}]}$

(3) The result is: (The BG/NBD model, Equation (A2))

$$P(\text{active at } T|X = x, t_x, T, \lambda, p) = \frac{(1-p)^x \lambda^x e^{-\lambda T}}{L(\lambda, p|X = x, t_x, T)} \quad (3)$$

(4) As the purchase rate λ and dropout rate p are unobserved, we compute

$P(\text{active at } T|X = x, t_x, T)$ for a randomly chosen customer by taking the expectation of (3) over the distribution of λ and p , updated to take into account the information $X=x, t_x, T$:

$$\begin{aligned} & P(\text{active at } T|X = x, t_x, T, \gamma, \alpha, a, b) \\ &= \int_0^1 \int_0^\infty P(\text{active at } T|X = x, t_x, T, \lambda, p) \cdot f(\lambda, p|\gamma, \alpha, a, b, X = x, t_x, T) d\lambda dp \end{aligned} \quad (4)$$

(5) By Bayes theorem, the joint posterior distribution of λ and p is given by

$$\begin{aligned} & f(\lambda, p|\gamma, \alpha, a, b, X = x, t_x, T) \\ &= \frac{L(\lambda, p|X = x, t_x, T) \cdot f(\lambda|\gamma, \alpha) \cdot g(p|a, b)}{L(\gamma, \alpha, a, b|X = x, t_x, T)} \end{aligned} \quad (5)$$

(6) Substituting equation (3) and (5) in (4), we get

$$P(\text{active at } T|X = x, t_x, T, \gamma, \alpha, a, b) = \frac{A}{L(\gamma, \alpha, a, b|X = x, t_x, T)} \quad (6)$$

where

$$\begin{aligned} A &= \int_0^1 \int_0^\infty (1-p)^x \lambda^x e^{-\lambda T} f(\lambda|\gamma, \alpha) g(p|a, b) d\lambda dp \\ &= \frac{B(a, b+x)}{B(a, b)} \frac{\Gamma(\gamma+x)\alpha^\gamma}{\Gamma(\gamma)(\alpha+T)^{\gamma+x}} \end{aligned} \quad (7)$$

3.2.3 Expected Dollar Volume per Reorder

Case 1: $X=1$ (The Extended SMC model, Equation (10), (11))

$$E[\theta|Z_1] = \rho_1 Z_1 + (1 - \rho_1)E[\theta] \quad (8)$$

$$\text{Where } \rho_1 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_W^2}$$

Case 2: $X > 1$ (The Extended SMC model, Equation (12), (13), (14))

$$E[\theta|Z_1, \dots, Z_X] = \left(\frac{X\sigma_A^2}{X\sigma_A^2 + \sigma_W^2} \right) \bar{Z} + \left(\frac{\sigma_A^2}{X\sigma_A^2 + \sigma_W^2} \right) E[\theta] \quad (9)$$

$$\text{Where } \bar{Z} = \frac{1}{X} \sum_{i=1}^X Z_i$$

$$\rho_X = \frac{\sigma_A^2}{\sigma_A^2 + (\sigma_W^2/X)}$$



4. Empirical Analysis

4.1 Data Description

The database applied is retrieved from an online CD/VCD retailer. This database consists of the purchase history that customers have made in the whole year of 2006. In this database, we have user IDs, the purchase date, dollar volume spent, ZIP codes, and order quantities of each purchase event. For utilizing our model, we only need the first three transaction records and importantly only choose the customers who have ever made purchases at the online retailer in the first quarter of 2006 to organize our dataset. In this way, we could have a data covering their initial (trial) and subsequent (repeat) purchases occasions for the period January 2006 through December 2006. After restructuring, we changed the dataset from covering 2856 to 1003 purchase events and the total number of customers was 368 in the end.

Although the information in this dataset is complete enough for us to utilize our model, we still need to reorganize the dataset into five columns which are ID, X, t_x , T, and \bar{Z} . ID numbers could represent different customers. X is the total number of transactions made by each customer in the period April 2006 through December 2006. T represents the total observation time of each customer and it may be different between customers. As the BG/NBD model, we get the value of T by the equation “ $T = 52 - \text{time of first purchase}$ ”. “52” means that we have 52 weeks in 2006. For example, if a customer made his first purchase in January 21, we could transform the date into “3/week” through dividing “21” by “7”. So the unit of T is 49 for this customer. However, the method to get t_x is a little complicate. In first, we also need to transform the date of the last purchase of a customer into Y with unit of a week. Then through the equation “ $t_x = T - (52 - Y)$ ” we could finally get the value t_x of the customer. \bar{Z} represents the average dollar volume spent by a customer and we could easily use the “AVERAGE” function in Excel to get it. (If $X_i = 0$, $t_{x_i} = 0$ and $\bar{Z} = 0$.) The raw data is shown in Figure 1.

Figure1. Screenshot of Excel Worksheet of Raw Data

	A	B	C	D	E	F
1	ID	x	t_x	T	Z bar	
2	34	20	48.28571	49.14286	2243.5	
3	35	12	49.71429	50.71429	978.3	
4	36	1	7.857143	50.71429	200.0	
5	37	10	44.14286	48.85714	1020.0	
6	38	3	33.14286	50.71429	986.7	
7	41	14	46.85714	47	586.4	
8	42	0	0	50.57143	0.0	
9	43	1	4.571429	50.57143	220.0	
10	45	0	0	50.57143	0.0	
11	46	2	28.14286	50.14286	1200.0	
12	47	21	45.57143	47.85714	349.0	

4.2 Parameter Estimation

4.2.1 Transaction/Dropout Process

To estimate the four parameters γ , α , a , and b of the BG/NBD model, we use the method of maximum likelihood estimation (MLE). One of the biggest advantages of the BG/NBD model is that we could use an easy way to estimate parameters. As we mentioned in Introduction, the likelihood function of the BG/NBD model is:

$$\begin{aligned}
 L(\gamma, \alpha, a, b | X = x, t_x, T) &= \frac{B(a, b+x)}{B(a, b)} \frac{\Gamma(\gamma+x)\alpha^\gamma}{\Gamma(\gamma)(\alpha+T)^{\gamma+x}} + \delta_{x>0} \frac{B(a+1, b+x-1)}{B(a, b)} \frac{\Gamma(\gamma+x)\alpha^\gamma}{\Gamma(\gamma)(\alpha+t_x)^{\gamma+x}} \\
 &= A_1 \cdot A_2 \cdot (A_3 + \delta_{x>0} A_4)
 \end{aligned}$$

where

$$A_1 = \frac{\Gamma(\gamma+x)\alpha^\gamma}{\Gamma(\gamma)}, A_2 = \frac{\Gamma(a+b)\Gamma(b+x)}{\Gamma(b)\Gamma(a+b+x)}, A_3 = \left(\frac{1}{\alpha+T}\right)^{\gamma+x}, A_4 = \left(\frac{a}{b+x+1}\right)\left(\frac{1}{\alpha+t_x}\right)^{\gamma+x} \quad (10)$$

After rewriting the original function to have $A_1 \sim A_4$ form the new equation, we could easily code it in Excel by taking the “log” of the whole equation, adding the four elements as “ln(all)”, summing “ln(all)” of 368 customers, and using the “Linear Programmer” of the Solver tool in Microsoft Excel to get a constrained biggest “LL”. However, before using the “Linear Programmer”, we have to set the initial value of the four parameters as “1”. We then find the four parameters. The following figure shows the sheet of parameter estimation.

Figure2. Screenshot of Excel Worksheet of Parameter Estimation

	A	B	C	D	E	F	G	H	I	J
1	r	0.523								
2	alpha	7.791								
3	a	0.027								
4	b	0.219								
5	LL	-3349.1332								
6										
7	ID	x	t_x	T	ln(all)	ln(A_1)	ln(A_2)	ln(A_3)	ln(A_4)	
8	34		20	48.28571	49.14286	-41.7074	41.4427	-0.2018	-82.9502	-89.2140
9	35		12	49.71429	50.71429	-31.8068	19.3340	-0.1879	-50.9559	-56.7770
10	36		1	7.857143	50.71429	-5.2400	0.4239	-0.1152	-6.1955	-6.2902
11	37						14.5362	-0.1829	-42.4782	-47.4046
12	38		5	55.14286	50.71429	-12.8710	1.7695	-0.1490	-14.3338	-17.4922
13	41		14	46.8571			9	-0.1921	-58.1416	-64.3047
14	42		0				9	0.0000	-2.1251	0.0000
15	43		1	4.571429			9	-0.1152	-6.1918	-5.9313
16	45		0						-2.1251	0.0000
17	46		2	28.14286					10.2399	-12.8527
18	47		21	45.5714					86.5004	-92.2237
19	48		0	47.57143					-2.0976	0.0000
20	49		2	34.57143	50.57143	-9.5031	0.8443	-0.1370	-10.2585	-13.2679
21	51		6	38.42857	42.85714	-19.5133	6.2467	-0.1687	-25.6005	-30.2754
22	52		3	44	46.28571	-12.4220	1.7695	-0.1490	-14.0565	-18.3209
23	53		0	0	50.42857	-1.0510	1.0729	0.0000	-2.1239	0.0000
24	54		3	47	50.42857	-12.6811	1.7695	-0.1490	-14.3165	-18.5193
25	57		0	0	50.42857	-1.0510	1.0729	0.0000	-2.1239	0.0000
26	58		5	45.42857	46.85714	-17.7139	4.5378	-0.1636	-22.0954	-27.0081

We now decompose the equation of “ln(A_1)” for easy understanding.

$$A_1 = \frac{\Gamma(\gamma + x)\alpha^\gamma}{\Gamma(\gamma)}$$

Because $A_1 = \frac{\Gamma(\gamma + x)\alpha^\gamma}{\Gamma(\gamma)}$, then the equation of log of A₁ in cell F8 is:

$$\begin{aligned} \ln(A_1) &= \ln[\Gamma(\gamma + x)] - \ln[\Gamma(\gamma)] + \gamma \ln(\alpha) \\ &= \text{GAMMALN}(B\$1 + B8) - \text{GAMMALN}(B\$1) + B\$1 * \text{LN}(B\$2) \end{aligned} \quad (11)$$

The second equation above shows the function we code in Microsoft Excel.

Therefore, the model estimation results are shown in Table 2.

Table2. Model Estimation Results of Transaction/Dropout Process

	0.523
	7.791
a	0.027
b	0.219
LL	-3349.1332

4.2.2 Transaction Dollar Volume

To estimate the three parameters σ_w^2 , σ_A^2 and $E(\theta)$ of the Extended SMC model, we could also utilize Microsoft Excel to perform the work.

(1) σ_w^2 Estimation

We have denoted σ_w^2 as the variance in the dollar volume spent across reorders for a customer in Conceptual Framework. Thus, in order to estimate σ_w^2 , we should choose a pool of customers who have ever made at least 2 repurchases in their observation time period and the sample size is 159. At first, we use the “VARP” function in Microsoft Excel to compute the variances of the dollar volume spent across reorders by each customer respectively. Then we use X (the number of reorders) of each customer as the weighted values to compute the weighted average σ_w^2 . The value of the weighted average σ_w^2 is 117789.9.

(2) σ_A^2 Estimation

The denotation of σ_A^2 is the variance in average dollar volume spent across customers (θ). Nevertheless, because the value of θ could not be observed, we have complication to compute σ_A^2 directly. However, the total variance (σ^2) in amount spent across both reorders and customers (which equals $\sigma_w^2 + \sigma_A^2$) is available and equals to 756038.6. We compute this value for the same 159 customers by using the recent dollar volume spent (the dollar volume of t_x) by each customer. As a result, the value of σ_A^2 is estimated as $756038.6 - 117789.9 = 638248.6$.

(3) $E(\theta)$ Estimation

$E(\theta)$ is denoted the expected value of θ which is assumed to vary across customers according to a normal distribution. We use the “AVERAGE” function in Microsoft Excel to compute the expected value of \bar{Z}_i of the

customers ($i=1\sim 368$). In the end, we use the “AVERAGE” function again and the value $E(\theta)$ is estimated as 614.10. We use the following table to show the values of the three parameters.

Table3. Model Estimation Results of Transaction Dollar Volume

σ_w^2	117789.9
σ_A^2	638248.6
$E(\theta)$	614.10

4.3 Model Results

4.3.1 Expected Number of Reorders (Conditional Expectation)

The expected number of reorders is computed using Equation (1). However, there is a value created from a complex function we need to estimate. This complicated function is called the Gaussian hyper-geometric function. In Literature Review, we have briefly mentioned this kind of function. Usually, we could have two methods to compute this function. One is numerical integration and the other is the algorithms. We will choose the method of numerical integration to compute it as the authors of the BG/NBD model. Therefore, before forecasting the expected number of reorders, we will introduce this function simply and the process to estimates its value in Microsoft Excel.

(1) *Gaussian hyper-geometric function* ${}_2F_1(\cdot)$ (Fader et al., 2005 b)

A. The prototype of ${}_2F_1(\cdot)$

$$\begin{aligned}
 {}_2F_1(a, b; c; z) &= \sum_j \frac{(a)_j (b)_j}{(c)_j} \frac{z^j}{j!}, \quad c \neq 0, -1, -2... \\
 &= \sum_j u_j, \quad \text{where } u_j = \frac{(a)_j (b)_j}{(c)_j} \frac{z^j}{j!}
 \end{aligned}$$

where $(a)_j$ is Pochhammer’s symbol, which denotes the ascending factorial $a(a+1)\cdots(a+j-1)$. The series converges for $|z| < 1$ and is divergent for

$|z| > 1$; if $|z| = 1$, the series converges for $c - b - a > 0$.

And we could have the following recursive expression for each term of the series:

$$\frac{u_j}{u_{j-1}} = \frac{(a + j - 1)(b + j - 1)}{(c + j - 1)j} z, \quad j = 1, 2, 3, \dots$$

where $u_0 = 1$.

B. The Numerical Integration Method Employed in Microsoft Excel

It's easy and simple for us to estimate ${}_2F_1(\cdot)$ by utilizing the above series. We just need to continue adding terms to the series until u_j is less than “machine epsilon” (the smallest number that a specific computer recognizes as being bigger than zero). In Microsoft Excel, it's easier to compute the series to a fixed number of terms, as a result, we will evaluate the first terms ($j = 0, 1, 2 \dots 150$).

In terminating the adding process of the series at $j=150$, whether have we evaluated too few terms? Because the speed with which $u_j \rightarrow 0$ depends on the magnitude of z and the observed number of reorders X in Equation (1), where $z = \frac{t}{\alpha + T + t}$.

The only unfixed variable in Z is T ($t = 52$; $\alpha = 7.791$). Under our dataset sampling rule, which is choosing the customers who have ever made their first purchase in the first quarter of 2006, the smallest T equals to 40.14. Thus the biggest z equals to 0.52. However, there is no point in going beyond $j = 40$ for $z < 0.5$ (Fader et al., 2005 b). As a result, based on the biggest z value, it's feasible for us to terminating the series at $j = 150$.

As to X , the biggest X in our dataset equals 50. We have tried to continue adding the terms to the biggest extent that $j = 240$, and we found the final result is the same as while $j = 150$.

(2) Computing conditional expectation of number of reorders

After estimating the parameters and the value of ${}_2F_1(\cdot)$, we could also get the conditional expectation of number of reorders in Microsoft Excel. Figure 3 shows the screenshot of excel worksheet of conditional expectation.

Figure3. Screenshot of Excel Worksheet of Conditional Expectation

	A	B	C	D	E	F	G	H	I
1	r	0.523		2F1	3.49191E+14				
2	alpha	7.791		a	50.523	=SUM(E7:E157)			
3	a	0.027		b	50.219	=B1+B6			
4	b	0.219		c	49.246	=B4+B6			
5				z	0.477978611				
6	x	50		Terms					
7	t_x	48.71429		0	1	=B3+B4+B6-1			
8	T	49		1	24.6259326	=B9/(B2+B8+B9)			
9	t	52.000		2	309.1006585				
10				3	2635.744518				
11	E(Y(t) X=x,t_x,T)		45.775	4	17171.33268	=E10*(E\$2+D11-1)*(E\$3+D11-1)/((E\$4+D11-1)*D11)*E\$5			
12				5	91134.9297				
13				6	410330.6833				
14				7	1611573.461	=E11*(E\$2+D12-1)*(E\$3+D12-1)/((E\$4+D12-1)*D12)*E\$5			
15				8	5634520.615				
16				9	17810141.62				
17				10	51517397.01				
18				11	137709204.7				
19				12	342913088.4				
20				13	800815849.9				
21				14	1763919129				
22				15	3682470988				
23				16	7317253276				
24				17	13890141283				
25				18	25271163724				
26				19	44193041350				

In this worksheet, we first place the four model parameters we have estimated in cells B1:B4. Then we place the purchase history ($X = x, t_x, T$) of a particular customer in cells B6:B9. For example, we choose the customer whose $ID=34$, $X=20$, $t_x=48.285$ and $T=49.142$. For all customers, t equals to 52 because the length of time over which we wish to make the conditional forecast is a year. Furthermore, we use the method outlined above to compute ${}_2F_1(\cdot)$ which is central to Equation (1). Corresponding to the prototype of ${}_2F_1(\cdot)$, the function parameters (a, b, c) are given in cells E2:E4 and the function argument (z) are setting in cell E5. In last, we evaluate the first 151 terms of the series (cell E7:E157) and these terms are summed in cell E1. Then we could

finally get the value $E(Y(t)|X = x, t_x, T, \gamma, \alpha, a, b)$ in cell C11. For this customer, we expect that he might have almost 19 reorders in next year 2007 conditioned on his purchase behavior in 2006.

(3) Use-friendly worksheet

The original worksheet developed by the authors of the BG/NBD is easily operated even for marketing practitioners. However, looking at figure 3, we could only get one conditional expectation of a customer once at a time. Thus, based on Equation (1), we redesign a worksheet which is more user-friendly for customer base analysers. This worksheet is shown in Figure 4.

Figure4. Screenshot of Excel Worksheet of Conditional Expectation—New Version

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	r	0.523		ID	x	t _x	T	t	a	b	c	z	$E(Y(t) X=x, t_x, T)$	2F1	Term
2	alpha	7.791		34	20	48.28571	49.14286	52.000	20.523	20.219	19.246	0.477352	18.520	1176199.577	
3	a	0.027		35	12	49.71429	50.71429	52.000	12.523	12.219	11.246	0.470564	10.985	5614.979	
4	b	0.219		36	1	7.857143	50.71429	52.000	1.523	1.219	0.246	0.470564	0.697	16.481	
5				37	10	44.14286	48.85714	52.000	10.523	10.219	9.246	0.478607	9.488	1898.727	
6				38	3	33.14286	50.71429	52.000	3.523	3.219	2.246	0.470564	2.966	21.962	
7				41	14	46.85714	47	52.000	14.523	14.219	13.246	0.48693	13.608	32391.267	
8				42	0	0	50.57143	52.000	0.523	0.219	-0.754	0.471173	0.444	0.596	
9				43	1	4.571429	50.57143	52.000	1.523	1.219	0.246	0.471173	0.580	16.543	
10				45	0	0	50.57143	52.000	0.523	0.219	-0.754	0.471173	0.444	0.596	
11				46	2	28.14286	50.14286	52.000	2.523	2.219	1.246	0.47301	2.081	13.809	
12				47	21	45.57143	47.85714	52.000	21.523	21.219	20.246	0.483053	19.841	2876980.449	
13				48	0	0	47.57143	52.000	0.523	0.219	-0.754	0.484339	0.467	0.561	
14				49	2	34.57143	50.57143	52.000	2.523	2.219	1.246	0.471173	2.113	13.625	
15				51	6	38.42857	42.85714	52.000	6.523	6.219	5.246	0.506583	6.554	223.242	
16				52	3	44	46.28571	52.000	3.523	3.219	2.246	0.490209	3.297	26.280	
17				53	0	0	50.42857	52.000	0.523	0.219	-0.754	0.471783	0.445	0.594	
18				54	3	47	50.42857	52.000	3.523	3.219	2.246	0.471783	3.062	22.223	
19				57	0	0	50.42857	52.000	0.523	0.219	-0.754	0.471783	0.445	0.594	
20				58	5	45.42857	46.85714	52.000	5.523	5.219	4.246	0.487582	5.155	87.923	
21				60	0	0	50.42857	52.000	0.523	0.219	-0.754	0.471783	0.445	0.594	
22				62	0	0	50.28571	52.000	0.523	0.219	-0.754	0.472396	0.446	0.593	
23				63	1	25.28571	50.28571	52.000	1.523	1.219	0.246	0.472396	1.041	16.669	
24				64	3	45.57143	49	52.000	3.523	3.219	2.246	0.477979	3.138	23.497	
25				65	0	0	50.28571	52.000	0.523	0.219	-0.754	0.472396	0.446	0.593	

In this new worksheet, the four parameters we have estimated are still placed in cell B2:B4. And the three parameters (a, b, c) and the argument (z) of ${}_2F_1(\cdot)$ are given in column I to column L corresponding to each customers. Then we equally evaluate the 151 terms of series from column P to column FJ. And the value of ${}_2F_1(\cdot)$ are summed in column N. In this way, as long as we put into the purchase information of each customer (column E to column G) and set the value of t over which we want to forecast, we could get the value

$E(Y(t)|X = x, t_x, T, \gamma, \alpha, a, b)$ in column M respectively.

4.3.2 Expected Active Probability

The expected active probability is computed using Equation (6) and (7) that we have derived. In coding these equations in Microsoft Excel, we luckily find that the elements of Equation (7) are similar to the log-likelihood function (Equation (13)) of the BG/NBD). In Equation (7), we could rewrite the Equation (7) with the Beta-Gamma transformation function $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ and $A_1, A_2,$ and A_3 of Equation (10) to

$$\begin{aligned}
 A &= \frac{B(a, b+x)}{B(a, b)} \frac{\Gamma(\gamma+x)\alpha^\gamma}{\Gamma(\gamma)(\alpha+T)^{\gamma+x}} \\
 &= \frac{\Gamma(a+b)\Gamma(b+x)}{\Gamma(b)\Gamma(a+b+x)} \frac{\Gamma(\gamma+x)\alpha^\gamma}{\Gamma(\gamma)} \left(\frac{1}{\alpha+T} \right)^{\gamma+x} \\
 &= A_2 \cdot A_1 \cdot A_3
 \end{aligned} \tag{12}$$

To combine Equation (6) and (12), we get Equation (13):

$$P(\text{active at } T | X = x, t_x, T, \gamma, \alpha, a, b) = \frac{A_1 \cdot A_2 \cdot A_3}{L(\gamma, \alpha, a, b | X = x, t_x, T)} \tag{13}$$

With Equation (13), we could also easily code it in Excel by taking “log” first, summing these elements together and then taking “exponent” to get the final expected active probability. In order to explain the coding process completely, we have the following figure to show the screenshot worksheet.

Figure5. Screenshot of Excel Worksheet of Expected Active Probability

	A	B	C	D	E	F	G	H	I	J	K	
1	r	0.523										
2	alpha	7.791										
3	a	0.027										
4	b	0.219										
5												
6												
7	ID	x	t _x	T	LL	ln(all)	ln(A_1)	ln(A_2)	ln(A_3)	ln(A_4)	ln(A_1)+ln(A_2)+ln(A_3)-ln(all)	Active Rate
8	34	20	48.28571	49.14286	-41.70707	41.4456	-0.2028	-82.95177	-89.21207			=EXP(I8)
9	35	12	49.71429	50.71429	-31.8067	19.3368	-0.1888	-50.9576	-56.7749			
10	36	1	7.857143	50.71429	-5.2375	0.4255	-0.1159	-6.1973	-6.2853			
11	37	10	44.14286	48.85714	-28.1176	14.5388	-0.1838	-42.4799	-47.4025			
12	38											
13	41	14	46.85714	47	-33.8654	24						
14	42	0	0	50.57143	-1.0532	1.0737	0.0000	-2.1269	0.0000			
15	43	1	4.571429	50.57143	-5.0483	0						
16	45	0	0	50.57143	-1.0532	1						
17	46	2	28.14286	50.14286	-9.4621	0.8462	-0.1377	-10.2416	-12.8502			
18	47	21	45.57143	47.85714	-42.2357	44.4672	-0.2041	-86.5020	-92.2216			
19	48	0	0	47.57143	-1.0256	1.0737	0.0000	-2.0993	0.0000			
20	49	2	34.57143	50.57143	-9.5034	0.8462	-0.1377	-10.2602	-13.2654			

Likewise, we place the four estimated parameters in B1:B4. We could also observe that this worksheet is almost identical to the one of Parameter Estimation (Figure 2) till Column I. And the formula of Column J is $[\ln(A_1)] + [\ln(A_2)] + [\ln(A_3)] - [\ln(all)]$. Finally, taking “exponent” for all cells of Column J in Column K respectively, we could get the expected active probability of each customer.

There is one thing we must pay attention to. When some customers have zero reorder in their observation time period ($X=0, t_x=0, T$), their active probabilities are expected to 1. Some people may argue that whether it’s reasonable or dependable.

However, in the model development of BG/NBD, the concept of the active probability has been simultaneously considered in deriving the conditional expectation of number of reorders (BG/NBD, Equation (A3)). And comparing the performance of the BG/NBD and Pareto/NBD on conditional expectation for the customers who have made zero reorder in T (called “zero class”), the forecasting capability of BG/NBD is better, especially for the zero class (Fader et al., 2005). Therefore, it could raise us more confidence on the expected active probability for the zero class.

On the other hand, because the zero class could be easily recognized, we also

could focus these customers who at least have made one reorder in T and then take appropriate response to these expected active probabilities while implementing one-to-one marketing afterward.

4.3.3 Expected Dollar Volume per Reorder

The expected dollar volume per reorder is computed using Equation (8) and (10). As the BG/NBD, we have already coded these equations in Microsoft Excel. The following figure shows the screenshot.

Figure6. Screenshot of Excel Worksheet of Expected Dollar Volume

	A	B	C	D	E	F	G	H	I	J
1	ID	Z bar	x	平均 σ_w^2	σ^2	σ_A^2	ρ_1	$E[\theta]$	$E[\theta/Z_{1-1}]$	ρ_x
2	42	0.0	0	117789.93	756038.55	638248.62	0.84	614.10	95.676	
3	45	0.0	0						95.676	
4	48	0.0	0						95.676	
5	53	0.0	0						95.676	
6	57	0.0	0						95.676	
7	60	0.0	0						95.676	
8	62	0.0	0						95.676	
9	65	0.0	0						95.676	
10	68	0.0	0						95.676	
11	69	0.0	0						95.676	
12	79	0.0	0						95.676	
13	80	0.0	0						95.676	
204	464	1920.0	1						1716.542	
205	476	1010.0	1						948.319	
206	480	3780.0	1						3286.756	
207	488	820.0	1						787.921	
208	489	1920.0	1						1716.542	
209	506	670.0	1						661.290	
210	520	670.0	1						661.290	
211	46	1200.0	2						1150.503	0.916
212	49	970.0	2						939.933	0.916
213	67	1410.0	2						1342.762	0.916
214	70	640.0	2						637.812	0.916
215	77	390.0	2						408.932	0.916

There are two cases while computing the expected dollar volume per reorder. If $X = 1$, we use Equation (8); if $X > 1$, we use Equation (10). One of the differences between them is the formulas of reliability coefficient ρ (Equation (9) and (12)). The other one is Z vs. \bar{z} . After estimating parameters, the values of weighted average σ_w^2 (cell D2), σ^2 (cell E2), σ_A^2 (cell F2), $E[\theta]$ (H2), \bar{z} (column B), ρ_1 (cell G2), and ρ_x (column J) have been also computed. Finally, we could get the expected

dollar volume per reorder in Column I respectively.

4.4 Integrated Individual Customer Forecasts

To generalize the model results of the three aspects above, we use the following figure to advocate the benefits of the whole model. (We only show ten customers for short.)

Figure7. 1-Year Predictions for Illustrative Customer Accounts

	A	B	C	D	E	F	G	H	I
1	ID	x	t_x	T	Observed Average Reorder Dollar Volume	Expected 1-Year # Reorders	Expected Dollar Volume per Reorder (\$)	Expected Active Probability	Expected 1-Year Dollar Volume
2	34	20	48.286	49.143	2243.5	18.53	2228.6	0.998	41285.7
3	35	12	49.714	50.714	978.3	10.99	972.8	0.997	10688.6
4	36	1	7.857	50.714	200.0	0.69	264.5	0.522	183.7
5	37	10	44.143	48.857	1020.0	9.49	1012.6	0.993	9610.4
6	38	3	33.143	50.714	986.7	2.97	965.1	0.959	2862.7
7	41	14	46.857	47.000	586.4	13.61	587.1	0.998	7992.2
8	42	0	0.000	50.571	0.0	0.44	95.7	1.000	42.4
9	43	1	4.571	50.571	220.0	0.58	281.4	0.433	162.7
10	45	0	0.000	50.571	0.0	0.44	95.7	1.000	42.4
11	46	2	28.143	50.143	1200.0	2.08	1150.5	0.931	2393.3

In Figure 7, the first five columns are the basic information in the dataset; the next three columns are the forecasting results from the three aspects above. And we multiply Column F to Column G and report in Column I respectively. In this way, we could concretely forecast the future dollar volume we will earn in a specific time period. In our research, our forecasting time period is setting 1 year equivalent to the length of the observation period. Of course, we could also set different time period of interest for forecasting. Besides, the expected future dollar volume could also be an important basis for us to compute the customer lifetime value (CLV) if introducing a discount rate into consideration.

Furthermore, the expected individual active probability can help a firm to pare the mailing list or the huge database of inactive customers with low active probabilities to reduce database management costs, mailing costs or other unnecessary marketing expenses.

4.5 Collective Customer Forecasts

In addition to individual customer forecasts, we also could forecast the number of reorders, the dollar volume and the active probability collectively. Based on the same dataset, the total number of 1-year expected reorder is 1074.24; the total 1-year expected dollar volume is \$980025.9. Therefore we could expect that the average expected dollar volume per reorder in 2007 is \$912.3 ($\$980025.9/1074.24$). Besides, although the size of our dataset is 368 customers, the sum of active probabilities of the customer base is only 335. It means that the number of active customers is not usually as many as the actual size of the customer base. Especially, with the expected active probabilities, we then could more correctly anticipate the future growth of firms and diagnose the basic health of them.

4.6 Model Validation

Even though the forecasting performances of the BG/NBD and the Extended SMC model have been validated in the original papers, we still want to utilize 1-way MANOVA to verify the differentiation capabilities of these three expected values: expected number of reorders (which will be called “EY” for short), expected dollar volume per reorder (which will be called “EZ” for short) and expected active probability (which will be called “PRO” for short). Before employing 1-way MANOVA, we will use the 80/20 rule to classify the 368 customers into two classes (1 & 2) and to be as independent variable and let these three expected values be dependent variables at first. Therefore, because we have one independent variable and we want to consider three independent variables simultaneously, we utilize 1-way MANOVA to validate our model. If the testing results are statistically significant, that means these dependent variables could account for the variances between the two classes and, more importantly, it’s more feasible for us to use these expected values as the basis to implement one-to-one marketing and other CRM strategies to retention customers.

So-called 80/20 rule means that almost 80% sales volume of a firm is contributed

by 20% of the firm's customers. Before implementing 1-way MANOVA, we use 80/20 rule to classify the 368 customers into two classes. At first, the total dollar volume spent by each customer in their observation time period could be observed and then we could sort them from the highest to the lowest. And the size of 20% of customers nearly equals to 74 (368*20%). Therefore we could sum the sorted dollar volume spent by the first 74 customers (\$732,250) and let this value divide by the sum of total dollar volume spent by 368 customers (\$732,250/\$982400 =74.54%). This value is close to 80% so we could say that 80/20 rule is almost realized in our dataset. Importantly, we let the first 74 customers to be class 1 and the others to be class 2. We then use the two classes as independent variables and "EY", "EZ" and "PRO" as dependent variables simultaneously to implement 1-way MANOVA

4.6.1 1-Way MANOVA

Because we have three dependent variables and one independent variable, we use 1-way MANOVA to proceed with hypothesis testing. Among the different kinds of statistical software, we use SAS to perform the work. The hypotheses and testing results are shown below.

(1) Overall Test

$$H_0 : u_{EY_1} = u_{EY_2} = u_{EZ_1} = u_{EZ_2} = u_{PRO_1} = u_{PRO_2}$$

$$H_1 : H_0 \text{ is not all equal}$$

Table 4. 1-way MANOVA Test Result---Overall Test

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Class Effect
H = Type III SSCP Matrix for Class
E = Error SSCP Matrix

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.41233299	172.93	3	364	<.0001
Pillai's Trace	0.58766701	172.93	3	364	<.0001
Hotelling-Lawley Trace	1.42522436	172.93	3	364	<.0001
Roy's Greatest Root	1.42522436	172.93	3	364	<.0001

According to the p-value of the statistics “Wilks’ Lambda”, we have enough evidences to reject H_0 . That is there are indeed some significant differences between the two classes. In order to find the sources that cause the differences, we then employ three marginal tests.

(2) Marginal Test

- ◆ EY--- expected number of reorder

$$H_0 : u_{EY_1} = u_{EY_2}$$

$$H_1 : u_{EY_1} \neq u_{EY_2}$$

Table 5. 1-way MANOVA Test Result---Marginal Test---EY

Dependent Variable: EY

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2943.726089	2943.726089	277.83	<.0001
Error	366	3877.985623	10.595534		
Corrected Total	367	6821.691712			

R-Square	Coeff Var	Root MSE	EY Mean
0.431524	111.3986	3.255078	2.922011

According to the p-value and the F-value, we have enough evidences to reject H_0 . That is there are indeed some significant differences between EY_1 and EY_2 . And from the relatively large value of R-Square, we could conclude that EY plays a better variable to explain variances.

- ◆ EZ--- expected dollar volume per reorder

$$H_0 : u_{EZ_1} = u_{EZ_2}$$

$$H_1 : u_{EZ_1} \neq u_{EZ_2}$$

Table6. 1-way MANOVA Test Result---Marginal Test---EZ

Dependent Variable: EZ

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	44100723.9	44100723.9	149.59	<.0001
Error	386	107904068.1	294819.9		
Corrected Total	387	152004791.9			

R-Square	Coeff Var	Root MSE	EZ Mean
0.290127	85.97506	542.9732	631.5473

The same as EY, the testing result in Table 6. is rejecting H_0 . And the value of R-Square tells us that EZ is the next best variable to explain the variances between the two classes.

◆ PRO--- expected active probability

$$H_0 : u_{PRO_1} = u_{PRO_2}$$

$$H_1 : u_{PRO_1} \neq u_{PRO_2}$$

Table7. 1-way MANOVA Test Result---Marginal Test---PRO

Dependent Variable: PRO

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.35004133	0.35004133	13.51	0.0003
Error	386	9.48599466	0.02591802		
Corrected Total	387	9.83603599			

R-Square	Coeff Var	Root MSE	PRO Mean
0.035588	17.65723	0.160991	0.911755

We could see that the P-value is small enough for us to reject H_0 . However, from the relatively small value of R-Square, the variance explaining capability of PRO is poor. It seems that the “PRO” variable itself could only explain a little difference between the two classes.

However, how about we temperately pare the zero class from the hypothesis testing, since the “PRO” value of the zero class seems make a little conflict

between others. Hence, we employ 1-way MANOVA again for the dataset without zero class (The sample size is 232 now). The hypotheses are all equal; the results are simplified and organized in Table 8.

Table8. 1-Way MANOVA Results without Zero Class

	Variables	Statistics “Wilks’ Lambda”	F Value	p-value	R-Square (without Zero Class)	Results	R-Square (with Zero Class)
Overall Test		0.48	82.08	<.0001	-	Reject H ₀	-
Marginal Test	EY	-	122.52	<.0001	0.347563	Reject H ₀	0.431524
	EZ	-	47.27	<.0001	0.170470	Reject H ₀	0.290127
	PRO	-	47.35	<.0001	0.170726	Reject H ₀	0.035588

Although H₀ are all rejected as well as the former tests, the R-Square of “PRO” has enhanced a lot. Thus the variance accountability of “PRO” increases while paring the zero class. However the most part of variances is still explained by the “EY”.

4.6.2 Validation Result

After employing 1-way MANOVA, we could have more confidence to use “EY”, “EZ” and “PRO”, especially “EY”, while *identifying* and *differentiating* customers which are the two bases for a firm to implement one-to-one marketing, we have mentioned in Introduction.

Therefore the three expected values not only could help us forecast future transaction behavior of customers, but could play differentiation or discrimination variables further to help firms implement more tailor-made marketing strategies and decrease unnecessary resources waste as well.

5. Conclusion and Discussion

5.1 Research Contribution

In this research, we combine the BG/NBD and the Extended SMC model to simultaneously and completely incorporate the past purchase behavior of customers (X, t_x, T, Z_i) to do some effective forecasts based on customer base analysis. Differed from the entire extended SMC model (based on Pareto/NBD), our research preserve and advocate the easy implementing of the BG/NBD and consider the past dollar volume spent by customers in the meanwhile through combining the Extended SMC model. Hence, our model is more suitable to be a basis to do further CLV research than the “pure” BG/NBD. We also empirically validate our model by using a database from an online VCD retailer and try to anticipate the possible purchase patterns of customers in the future both individually and collectively. And then we validate our model through 1-way MANOVA to ensure the differentiation capabilities of the important expected values. In this way we could not only use our model to do forecasting but use the model results to differentiate customers for further one-to-one marketing putting into practice.

Furthermore, based on the BG/NBD, we have derived the equation of expected active probability for a random chosen customer. It could help us to understand the individual active probability and the true customer base of a firm after summing the active probabilities of all customers.

Besides, we have transformed the worksheet of the BG/NBD to a more user-friendly form. With this new worksheet, we only should put the basic purchase history of all customers into and then we could get the expected values of interests at one time. It could save a firm a lot of time to implement this model especially when its base of customers is huge. And also we wish to improve the utility rate of our model.

5.2 Research Limitation

The biggest limitation to implement our model is that the database must be non-contractual setting. The non-contractual setting means that the transactions

occur continuously other than discretely and the time point at which a customer becomes inactive is unobserved. In this setting, the value of the expected active probability exists.

Another limitation is that the authors of these models have made assumptions that also in a non-contractual setting, the basic model assumptions, where the transaction process follows NBD and the dropout process follows BG, are satisfied.

Besides, the accuracy of our model's forecasting ability is influenced by the future marketing activities targeted at the same group of customers. If there are many differences between present and future marketing activities, the expected values only based on past purchase histories may have some distances from the actual future conditions. Therefore while comparing the expected values with actual ones and employ a correlation analysis; we should compare the marketing activities and expenditures in advance.

5.3 Future Research Direction

There are some future research directions when using our model. Because our model has incorporated dollar volume of customers, it could be the basis to project the customer lifetime value. Furthermore our model results (three important expected values) could be the variables to help us discriminate and select customers who could bring more sales volume and profits. If the original size of the customer base is too large or very divergent within, we could utilize demographic variables or other maybe "RFM" variables to segment the customer base first and then implement our model respectively. In this way, the estimated parameters may be fitter for different segments and could improve the forecasting accuracy. If we could extend our observation time and get a database covering more than one year, we could do cross-validation to ensure and validate whether our model results could successfully predict future purchase behavior or not. Moreover, if the database applied could provide more complete information, such as demographic variables, we could not only utilize these variables to classify our customers other than the

monetary variable which combined with 80/20 rule and to have more reliable hypothesis testing results while doing model validation, but also we could combine the model results (three important expected values) with the demographic variables and we could have better understanding about the profile of every individual customer to practically implement one-to-one marketing.



References

- Abramowitz, M. & Stegun, I. A.. 1972. Handbook of Mathematical Functions, Dover Publications Inc., New York.
- Ansari, Asim & Carl F. M. 2003. "E-Customization," J. Marketing Research 40(2), 131–145.
- Ansari, Asim, Carl Mela, & Scott A. Neslin. 2004. "Customer Channel Migration," Working Paper, Columbia University School of Business, NY.
- Blattberg Robert C. & John Deighton. 1996. "Manage Marketing by the Customer Equity Test." Harvard Business Review 74, 136-144.
- Fader, P. S. & Lattin, J. M. 1993. "Accounting For Heterogeneity and Nonstationarity In A Cross-sectional of Consumer Purchase Behavior." Marketing Science 12(3), 304-317.
- Fader, P. S., B. G. S. Hardie, & K. L. Lee. 2005 a. "Counting Your Customers': the Easy Way: An Alternative to the Pareto/NBD Model", Marketing Science 24(2), 275-284.
- Fader, P. S., B. G. S. Hardie, & K. L. Lee. 2005 b. "Implementing the BG/NBD Model for Customer Base Analysis in Excel". (Spreadsheet Note) (<http://brucehardie.com/notes/004/>)
- Ehrenberg, A. S. C. 1972. Repeat Buying, North-Holland, Amsterdam.
- Gerer, H. U. 1979. An Introduction to Mathematical Risk Theory, Monograph No. 8, Philadelphia, PA: S. S. Heubner Foundation for Insurance Education, The Whartn School, University of Pennsylvania.
- Guadagni, P. M. & J. D. C. Little. 1983. "A Logit Model of Brand Choice Calibrated on Scanner Data." Marketing Science (2), 203-238.
- Gupta, Sunil, Donald R. Lehmann & Jennifer Ames Stuart. 2004. "Valuing Customers" J. Marketing Research 41(1), 7–18.
- Heide, J. B. 1994. "Interorganizational Governance in Marketing Channels." J. Marketing 58(1), 71-85.
- Hughes Arthur M. 1994. Strategic Database Marketing. Probus Publishing, Chicago.
- Johnson, N. L. & Singer, B. 1970. Continuous Univariate Distributions---1, John Wiley, New York, 233.
- Jones, M.A., Mothersbaugh, D.L. & Beatty, S.E. 2000. "Switching Barriers and Repurchase Intentions in Services." Journal of Retailing 76 (2), 259 – 274.
- Kamakura, W. A., S. Ramaswami & R. Srivastava. 1991. "Applying Latent Trait Analysis in

- the Evaluation of Prospects for Cross-Selling of Financial Services,” *International Journal of Research in Marketing* 8, 329–349.
- Kamakura, Wagner A, Michel Wedel, Fernando de Rosa & Jose A. Mazzon. 2003. “Cross-Selling Through Database Marketing: A Mixed Data Factor Analyzer for Data Augmentation and Prediction,” *International Journal of Research in Marketing* 20, 45–65.
- Kamakura, Wagner A, Ansari, Asim., Bodapati, A., Fader, P. S., Raghuram Iyengar, Naik, P., Scott A. Neslin, Baohong Sun, Verhoef, P. C., Michel Wedel & Wilcox, R. 2005. “Choice Models and Customer Relationship Management” *Marketing Letters* 16:3/4, 279–291.
- Leigh McAlister, Ruth N. Bolton, & Ross Rizley. 2006. *Essential Readings in Marketing*. Marketing Science Institute.
- Moms, C. N. 1983. “Parametric Empirical Bayes Inference: Theory and Applications.” *Journal of the American Statistical Association* 78, 47-55.
- Mozer, M.C., Wolniewicz, R., Grimes, D.B., Johnson, E. & Kaushansky, H. 2000. “Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry.” *IEEE Transactions on Neural Networks* 11(3), 690-696.
- Niraj, Rakesh, Mahendra Gupta, & Chakravarthi Narasimhan. 2001. “Customer profitability in a supply chain.” *J. Marketing* 65(July), 1-16.
- Peppers, D., Rogers, M. & Dorf, B. 1999. “Is Your Company Ready for One-to-one Marketing?” *Harvard Business Review* 77(1), 151-156.
- Rakesh Niraj, Mahendre Gupta & Chakravarthi Narasimhan. 2001. “Customer Profitability in a Supply Chain.” *J. Marketing* 65(July) 1-16.
- Reichheld, F.F. 1996. “Learning from customer defections” *Harvard Business Review* 74 (2), 56-69.
- Reichheld, F.F. & Sasser, W.E. 1990. “Zero defections: Quality Comes to Services.” *Harvard Business Review* 68(5), 105-111.
- Reinartz, Werner, V. Kumar. 2000. “On the Profitability of Long-life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing.” *J. Marketing* 64(October) 17-35.
- Rust, Roland & Anthony Zahorik. 1993. “Customer Satisfaction, Customer Retention, and Market Share” *Journal of Retailing* 69(2).
- Schmittlein, David, Donald Morrison, and Richard Colombo. 1987. “Counting Your

- Customers: Who Are They and What Will They Do Next?" Management Science 33(1).
- Schmittlein, David. 1989. "Surprising Inferences from Unsurprising Observations: Do Conditional Expectations Really Regress to the Mean?," The American Statistician 43, 171-183.
- Schmittlein, David & Robert A. Peterson. 1994. "Customer Base Analysis: An Industrial Purchase Process Application" Marketing Science 13(1), 41-67.
- Stum, D. & Thiry, A. 1991. "Building customer loyalty." Training and Development Journal 45(4), 34-36.
- Sung-Shun Weng & Mei-Ju Liu. 2004. "Feature-based Recommendations for One-to-one Marketing." Expert Systems with Applications 26, 493-508.
- Zeithaml, V.A., Berry, L.L., Parasuraman, A. 1996. "The Behavioural Consequences of Service Quality." J. Marketing 60(2), 31-46.

