Optimal threshold analysis of segmentation methods for identifying target customers

著者	Mizuno Makoto, Saji Akira, Sumita Ushio, Suzuki Hideo
year	2005-11
シリーズ	Department of Social Systems and Management Discussion Paper Series ~ no. 1136
URL	http://hdl.handle.net/2241/11479

Department of Social Systems and Management

Discussion Paper Series

No. 1136

Optimal Threshold Analysis of Segmentation Methods for Identifying Target Customers

by

Makoto Mizuno, Akira Saji, Ushio Sumita and Hideo Suzuki

November 2005

UNIVERSITY OF TSUKUBA
Tsukuba, Ibaraki 305-8573
JAPAN

Optimal Threshold Analysis of Segmentation Methods for Identifying Target Customers

Makoto Mizuno†, Akira Saji†, Ushio Sumita† and Hideo Suzuki†

† Graduate School of Systems and Information Engineering,
University of Tsukuba, 1-1-1 Tennou-dai, Tsukuba,
Ibaraki 305-8573, Japan.
{mizuno, saji00, sumita, hsuzuki}@sk.tsukuba.ac.jp

November 8, 2005

Abstract

In CRM (Customer Relationship Management), the importance of a segmentation method for identifying good customers has been increasing. For evaluation of different segmentation methods, Accuracy often plays a key role. This indicator, however, cannot distinguish the following two types of errors: Type I Error for misidentifying a good customer as a bad customer and Type II Error for misinterpreting a bad customer as a good customer. In order to analyze the financial effectiveness of various segmentation methods, it is crucial to capture the distinction between Type I and Type II Errors since the former represents the opportunity cost while the latter results in the inefficient use of the promotion budget. The purpose of this paper is to overcome this pitfall by introducing two different indicators: Recall and Precision, which have been prevalent in the area of Information Retrieval. A mathematical model is developed for describing a generic segmentation method. Assuming that a promotion is addressed exclusively to the selected target customers, the financial effectiveness of the underlying segmentation method is expressed as a function of Recall and Precision. An optimization problem is then formulated so as to maximize the financial measure by finding the optimal threshold level in terms of the severeness for estimating the target set of good customers. By introducing a functional form which represents correctness and mistakes about the target set, the unique optimal solution is derived explicitly. Using real customer purchase data, the proposed approach is validated where Logistic Regression Model and Support Vector Machine are employed as segmentation methods. The methodology developed in this paper may provide a foundation for understanding and comparing the performance characteristics of various segmentation methods from a new perspective.

Keywords: CRM, Identifying target customers, Segmentation method, Optimal threshold level, Recall, Precision

0 Introduction

During the past decade, the Internet has impacted the way businesses are conducted across all industries, where emerging new business models collectively constitute e-business. Accordingly, marketing has been going through the Internet revolution with a thrust of e-marketing. Before the Internet, the mass marketing through TV, newspapers, radio and other media was directed one way from the media to customers, while the one-to-one marketing was laborious, time-consuming and costly, and was conducted only in a limited way using hearing via telephones, interviews at exits of stores, and the like. One of the most significant features of e-marketing is that the mass marketing and the one-to-one marketing can be done simultaneously with little cost and high speed.

Along this new trend, CRM (Customer Relationship Management) has become increasingly important, where a corporation and its customers constantly engage themselves in two way communications and exchange information valuable to each other. Through such practices, customer profiles are captured and purchase data are accumulated in almost real time. Consequently, customer databases of huge magnitude are created and processed for understanding the market better and implementing effective marketing strategies. In this regard, segmentation of potentially profitable customers, whom we call good customers, becomes significantly important. Through ranking tools such as RFM (Recency, Frequency, Monetary) Analysis and Cross Analysis, customers are classified based on their past purchase behaviors. The target set consisting of potentially good customers is then identified, to whom concentrated promotional efforts may be applied so as to keep such customers loyal and increase the profit.

For this purpose, a variety of segmentation methods can be found, including Decision Tree and Neural Network prevalent in data mining, Discriminant Analysis and LRM (Logistic Regression Model) based on statistical analysis, and SVM (Support Vector Machine) which has recently attracted attention of many researchers and practitioners in such areas as machine learning and medical sciences. Because of availability of many segmentation methods, it is desirable to establish a methodological approach for understanding and comparing the performance characteristics of such methods. Traditionally, the key criterion has been Accuracy which measures the ratio of correctly identified good or bad customers. Accuracy, however, fails to capture separately two types of errors: Type I Error for misidentifying good customers as bad customers and Type II Error for misunderstanding bad customers as good customers. Assuming that a promotional campaign is addressed exclusively to the estimated good customers, the former represents some opportunity loss while the latter results in the inefficient use of the campaign budget. Accordingly, it is important to incorporate the distinction between the two types of errors in evaluating the financial effectiveness of segmentation methods for marketing campaign.

The purpose of this paper is to achieve this objective by introducing two performance measures called Recall and Precision which have been prevalent in the field of Information Retrieval. Recall is the ratio of the good customers who were included in the target set, while Precision is the ratio of the target customers who actually turned out to be good customers. Utilizing the two performance measures, a financial measure is then established, which enables one to assess the trade-off between the opportunity loss and the ineffective use of the campaign budget mentioned above. This trade-off is parameterized by specifying a threshold level concerning the severeness for estimating good customers. Accordingly, given a segmentation method, an optimization problem can be formulated so as to maximize the financial measure by finding the optimal threshold level.

In general, this optimization problem cannot be solved since it involves the magnitudes of Type I Error and Type II Error represented by Recall and Precision, which are unknown until the purchasing records of the customers in the next future period become available. This difficulty is overcome by introducing a functional structure with two parameters

α and β which represents correctness and mistakes in estimating the target set. Recall and Precision can then be expressed mathematically in terms of α and β, which in turn enables one to obtain the unique optimal solution in a closed form. Given a segmentation method, the two parameters can be estimated based on the past data available now, and consequently the optimal threshold level can be specified. The approach proposed in this paper may provide a foundation for understanding and comparing the performance characteristics of various segmentation methods from a new perspective.

The structure of this paper is as follows. In Section 1, a mathematical model is introduced for describing a general structure of segmentation methods. Utilizing two performance measures, Recall and Precision, a financial measure is established in Section 2. By evaluating the trade-off between the opportunity loss and the ineffective use of the campaign budget, an optimization problem is formulated so as to maximize the financial measure by finding the optimal threshold level. The optimal solution structure is discussed for some simple cases. Section 3 deals with development of a functional structure for representing correctness and mistakes in estimating the target set. The unique optimal threshold level is explicitly derived in terms of two parameters involved in the functional structure. In Section 4, the proposed approach is validated using real customer purchase data where LRM and SVM are employed as segmentation methods. Numerical results reveal that the use of the optimal threshold level outperforms the use of the default values in a consistent manner. It is also found that LRM is superior to SVM for the real purchase data employed in this paper. Finally, some concluding remarks are given in Section 5.

1 General Structure of Segmentation Methods

We consider a set of N customers $CS = \{c_1, \dots, c_N\}$. Associated with each coustomer c_i is the profile vector $\boldsymbol{x}_i$, typically describing c_i's basic information and his/her past purchasing behavior. The domain of the profile vectors is denoted by Ω, i.e. $\boldsymbol{x}_i \in \Omega, 1 \leq i \leq N$. According to a prespecified criterion, we suppose that a set of good customers will be determined by their purchasing outcome in the next future period. More specifically, let $D^* : CS \to \{-1, 1\}$ be a mapping describing this separation with

$$CS = B \cup G \tag{1.1}$$

where

$$B = \{c_i : D^*(c_i) = -1\};\ G = \{c_i : D^*(c_i) = 1\}. \tag{1.2}$$

Here B and G denote the set of bad customers and the set of good customers respectively. For notational convenience, we define

$$X_B = |B|;\ X_G = |G|, \tag{1.3}$$

where the cardinality of a set A is denoted by $|A|$. It should be noted that B and G will be realized only upon completion of the next future period. Accordingly, at the present time, of interest is to develop a segmentation method which would attempt to identify those customers in G by estimating customers' future purchasing behavior based on $\boldsymbol{x}_i, 1 \leq i \leq N$.

Let a mapping $D : \Omega \to [0, 1)$ be such that, for $z \in [0, 1]$,

$$CS = B(z) \cup G(z) \tag{1.4}$$

with

$$B(z) = \{c_i : D(\boldsymbol{x}_i) < z\};\ G(z) = \{c_i : D(\boldsymbol{x}_i) \geq z\}. \tag{1.5}$$

It should be noted that

$$\begin{cases} B(0) = \emptyset;\ B(1) = CS;\ z_1 < z_2 \Rightarrow B(z_1) \subset B(z_2). \\ G(0) = CS;\ G(1) = \emptyset;\ z_1 < z_2 \Rightarrow G(z_1) \supset G(z_2). \end{cases} \tag{1.6}$$

The mapping D describes a segmentation method, where those customers estimated to be in G are identified as $G(z)$ by calculating the value $D(\boldsymbol{x}_i)$ and comparing it with $z, 1 \leq i \leq N$. We call $G(z)$ the target set given a threshold level $z \in [0, 1]$.

In general, the target set $G(z)$ of the estimated good customers and the set G of the realized good customers may not necessarily coincide with each other. In order to capture such possible differences, we now introduce the four cell functions as below:

$$\begin{aligned} x_{BB}(z) &= |B(z) \cap B|; \quad x_{BG}(z) = |B(z) \cap G|; \\ x_{GB}(z) &= |G(z) \cap B|; \quad x_{GG}(z) = |G(z) \cap G|. \end{aligned} \tag{1.7}$$

The cell functions in (1.7) constitute the confusion matrix, the term often employed in the field of data mining, see e.g. Berry and Linoff [2]. The confusion matrix is given below in Table 1.1.

Table 1.1: Confusion Matrix

		Realization		
		B	G	Total
Pre-	$B(z)$	$x_{BB}(z)$	$x_{BG}(z)$	$X_{B(z)}$
diction	$G(z)$	$x_{GB}(z)$	$x_{GG}(z)$	$X_{G(z)}$
Total		X_B	X_G	N

Here, one has

$$\begin{aligned} x_{BB}(z) + x_{GB}(z) &= X_B; \qquad x_{BG}(z) + x_{GG}(z) = X_G; \\ x_{BB}(z) + x_{BG}(z) &= X_{B(z)}; \quad x_{GB}(z) + x_{GG}(z) = X_{G(z)} \end{aligned} \tag{1.8}$$

and

$$X_B + X_G = X_{B(z)} + X_{G(z)} = N. \tag{1.9}$$

From (1.6) and (1.7), $x_{ij}(z),\ i, j \in \{B, G\}$ should satisfy the following properties.

(1.10) Both $x_{BB}(z)$ and $x_{BG}(z)$ are nondecreasing in z and

$$\lim_{z \to 0+} x_{BB}(z) = 0,\ \lim_{z \to 1-} x_{BB}(z) = X_B,$$
$$\lim_{z \to 0+} x_{BG}(z) = 0,\ \lim_{z \to 1-} x_{BG}(z) = X_G.$$

(1.11) Both $x_{GB}(z)$ and $x_{GG}(z)$ are nonincreasing in z and

$$\lim_{z \to 0+} x_{GB}(z) = X_B,\ \lim_{z \to 1-} x_{GB}(z) = 0,$$
$$\lim_{z \to 0+} x_{GG}(z) = X_G,\ \lim_{z \to 1-} x_{GG}(z) = 0.$$

Traditionally, the performance of various segmentation methods has been evaluated by the overall accuracy

$$(1.12) \qquad A(z) = \frac{x_{BB}(z) + x_{GG}(z)}{N}.$$

This approach often ignores the distinction between the following two types of errors: to misidentify good customers as non-target customers (Type I Error captured by $x_{BG}(z)$) and to misidentify bad customers as target customers (Type II Error corresponding to $x_{GB}(z)$). When a marketing campaign is addressed to those in the target set $G(z)$, Type I Error represents some opportunity loss while Type II Error results in the inefficient use of the campaign budget. Hence, the two types of errors have different financial implications, and accordingly it is important to distinguish Type I Error from Type II Error in managerial decision making. In order to reflect the difference between the two types of errors explicitly in our analysis, we introduce two performance measures, Recall $R(z)$ and Precision $P(z)$, defined by

$$(1.13) \qquad R(z) = \frac{x_{GG}(z)}{X_G};\ P(z) = \frac{x_{GG}(z)}{X_{G(z)}}.$$

These performance measures $R(z)$ and $P(z)$ are prevalent in the field of Information Retrieval, see e.g. Rijsbergen [5]. We note that Recall $R(z)$ is the ratio of the good customers who were included in the target set, while Precision $P(z)$ is the ratio of the target customers who actually turned out to be good customers.

From (1.11) and (1.13), one sees that

$$(1.14) \qquad \begin{aligned} &R(z) = \frac{x_{GG}(z)}{X_G} \text{ is nonincreasing in } z \in [0,1] \\ &\text{with } R(0) = 1 \text{ and } R(1) = 0 \end{aligned}$$

and

$$(1.15) \qquad P(z) = \frac{x_{GG}(z)}{X_{G(z)}} = \frac{1}{1 + r_G(z)};\quad r_G(z) = \frac{x_{GB}(z)}{x_{GG}(z)} \text{ with } P(0) = \lambda$$

where

$$(1.16) \qquad \lambda = \frac{X_G}{N}.$$

The function $P(z)$ at $z = 1$ is not defined since $G(1) = \emptyset$ and consequently $X_{G(1)} = 0$. Its limiting value $P(1_-)$ and monotonicity depend on the behavior of $r_G(z)$. More specifically, one has

$$(1.17) \qquad \begin{aligned} &P(z) \text{ is nondecreasing in } z \text{ if } r_G(z) \text{ is nonincreasing in } z \\ &\text{and } P(1_-) = \frac{1}{1 + r_G(1_-)}. \end{aligned}$$

From (1.13), the following relationship exists between $R(z)$ and $P(z)$.

$$(1.18) \qquad \frac{R(z)}{P(z)} - \frac{X_{G(z)}}{X_G} = 0.$$

The next proposition then holds:

Proposition 1.1 Let λ be as in (1.16). Then

$$P(z) \geq \lambda R(z) \quad \text{for all } z \in [0,1). \tag{1.19}$$

Proof One sees from (1.18) that $R(z)/P(z) = X_{G(z)}/X_G$ which is nonincreasing from (1.8) and (1.11). Hence it follows from (1.14) and (1.15) that for all $z \in [0,1)$,

$$\frac{R(0)}{P(0)} = \frac{1}{\lambda} \geq \frac{R(z)}{P(z)},$$

proving the proposition. ∎

In Alvarez [1], it was shown that the following relationship exists between $A(z)$, $R(z)$ and $P(z)$

$$A(z) = \frac{1}{N}\left\{X_B + X_G R(z)\left(2 - \frac{1}{P(z)}\right)\right\}, \tag{1.20}$$

which can be obtained from (1.8), (1.9), (1.12) and (1.13). In this paper, the emphasis will be on the role of $R(z)$ and $P(z)$ in assessing the financial impact of the marketing campaign and analyzing the optimal threshold level z^* for a given segmentation method.

2 Impact of Segmentation Methods on Financial Effectiveness of Marketing Efforts

Given a segmentation method and a threshold level $z \in [0,1]$ discussed in the previous section, we consider a marketing strategy where the promotional efforts are concentrated on the target customers in $G(z)$ exclusively and not applied to those in $B(z)$. Of interest is to analyze the impact of the choice of a segmentation method and $z \in [0,1]$ on financial effectiveness of the marketing strategy.

Let θ_B be the expected revenue from each of those customers in B during the next future period. It is assumed that, under the influence of the promotional efforts, a customer in B increases his/her expenditure by a factor of $(1+\eta_B)$ where $\eta_B \in [0,\infty)$. θ_G and η_G are defined similarly. Clearly, those in G purchase more than those in B. It is also likely that the promotional efforts yield that the resulting incremental revenue for those in G is greater than that for those in B. Accordingly, throughout the paper, we assume that:

$$\theta_B < \theta_G \quad \text{and} \quad \theta_B\eta_B < \theta_G\eta_G. \tag{2.1}$$

The cost per customer for the promotional efforts is denoted by $\nu > 0$. The expected total revenue with the promotional efforts is then given by

$$\begin{aligned} V(z) = {} & \theta_B x_{BB}(z) + \{\theta_B(1+\eta_B) - \nu\}x_{GB}(z) \\ & + \theta_G x_{BG}(z) + \{\theta_G(1+\eta_G) - \nu\}x_{GG}(z). \end{aligned} \tag{2.2}$$

We note that, with $\eta_B = \eta_G = \nu = 0$, $V(z)$ is the expected total revenue without the promotional efforts. Consequently, the financial effectiveness of the promotional efforts can be measured by the difference between $V(z)$ with $\eta_B, \eta_G \geq 0, \nu > 0$ and $V(z)$ with $\eta_B = \eta_G = \nu = 0$, i.e.,

$$\Delta V(z) = (\theta_B \eta_B - \nu) x_{GB}(z) + (\theta_G \eta_G - \nu) x_{GG}(z). \tag{2.3}$$

From (1.8) and (1.11), Equation (2.3) can be rewritten in terms of Recall $R(z)$ and Precision $P(z)$ as

$$\Delta V(z) = X_G R(z) \left(\gamma_B \frac{1 - P(z)}{P(z)} + \gamma_G \right) \tag{2.4}$$

where

$$\gamma_B = \theta_B \eta_B - \nu \quad \text{and} \quad \gamma_G = \theta_G \eta_G - \nu. \tag{2.5}$$

One sees from (2.1) and (2.5) that

$$\gamma_B < \gamma_G. \tag{2.6}$$

Given a segmentation method, we are then interested in the following maximization problem for determing an optimal threshold level $z^* \in [0, 1]$:

Problem 2.1 Find $z^* \in [0, 1]$ satisfying

$$\Delta V(z^*) = \max_{0 \leq z \leq 1} \left[\Delta V(z) = X_G R(z) \left(\gamma_B \frac{1 - P(z)}{P(z)} + \gamma_G \right) \right].$$

While $R(z)$ and $P(z)$ are related to each other as given in (1.18), it is worthwhile to explore the structure of ΔV as a function of R and P by suppressing the variable z, as we will see. For notational convenience, we define

$$F(R, P) = \frac{\Delta V(R, P)}{X_G} = R \left(\gamma_B \frac{1 - P}{P} + \gamma_G \right) \tag{2.7}$$

where $R, P \in [0, 1]$. Let $F(R, P) = K > 0$. Substituting this into (2.7) and solving for P as a function of R, one finds that

$$P(R) = -\frac{hR}{R - g} \tag{2.8}$$

where

$$g = \frac{K}{\gamma_G - \gamma_B} \quad \text{and} \quad h = \frac{\gamma_B}{\gamma_G - \gamma_B}. \tag{2.9}$$

It then follows that

$$\frac{\mathrm{d}}{\mathrm{d}R} P(R) = \frac{gh}{(R - g)^2}, \quad \left(\frac{\mathrm{d}}{\mathrm{d}R} \right)^2 P(R) = -\frac{2gh}{(R - g)^3}. \tag{2.10}$$

From (2.10), structural properties of $P(R)$ can be characterized by considering the sign of γ_B and γ_G. We first note that γ_G cannot be negative for $F(R, P) = K > 0$ from (2.6) and (2.7). Accordingly, only the following two cases should be considered.

$$\begin{cases} \text{Case I:} & \gamma_B < 0 < \gamma_G, \\ \text{Case II:} & 0 < \gamma_B < \gamma_G. \end{cases} \tag{2.11}$$

For Case I, one has $h < 0 < g$ so that, from (2.10), $P(R)$ is strictly decreasing, and is strictly concave for $R < g$ and strictly convex for $R > g$. However the former case results in $P(R) < 0$ from (2.8) and should be discarded. In summary, one has $P(R)$ as in Figure 2.1. In Case II, both g and h are positive, and one has $R < g$ because otherwise $P(R)$ becomes negative from (2.8). It then follows from (2.10) that $P(R)$ is strictly increasing and is strictly convex, as depicted in Figure 2.2. Contours of $F(R, P)$ for the two cases are exhibited in Figures 2.3 and 2.4, where the arrow indicates the increasing direction. It should be noted that Problem 2.1 can be solved by plotting $(R(z), P(z))$ for $z \in [0, 1]$ as determined by (1.18) and then finding the point crossing with the highest contour.

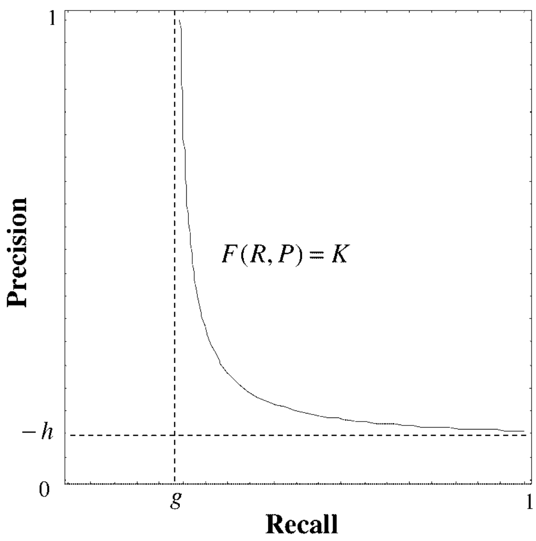


Figure 2.1: $F(R, P) = K$ for Case I

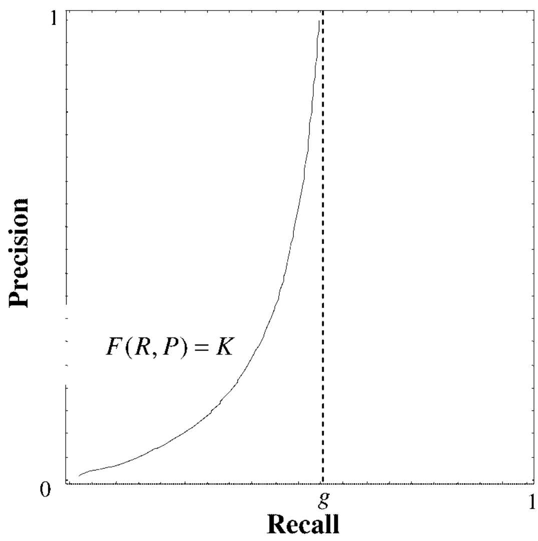


Figure 2.2: $F(R, P) = K$ for Case II

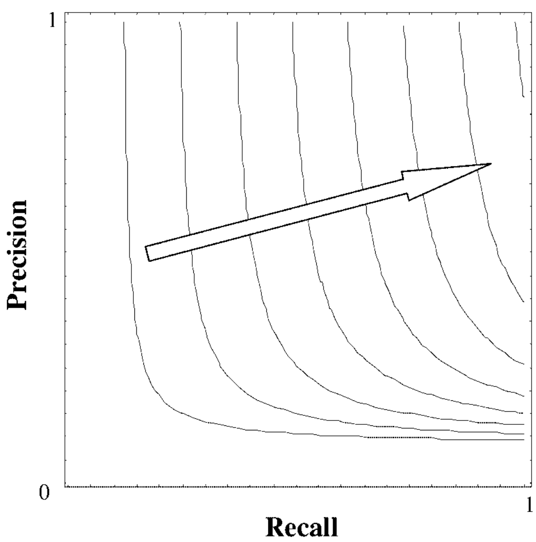


Figure 2.3: Contours of F for Case I

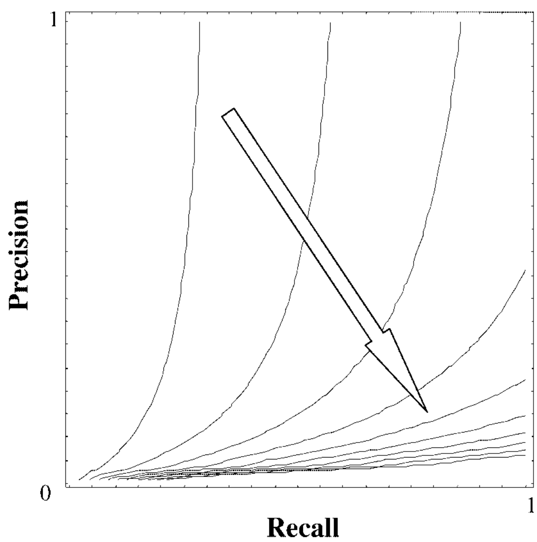


Figure 2.4: Contours of F for Case II

For Case II, one has $0 < \gamma_B < \gamma_G$ from (2.11). This means that the marketing campaign is effective for not only good customers in G but also bad customers in B. Accordingly it makes sense to target the whole customers by setting $z = 0$, $G(z) = CS$ and $B(z) = \emptyset$. This point may be observed graphically as shown in Figure 2.5. Here, from Proposition 1.1, $(R(z), P(z))$ should lie above the linear function $P = \lambda R$ as z moves from 0 to 1 with $(R(0), P(0)) = (1, \lambda)$ and $(R(1\text{-}), P(1\text{-})) = (0, \frac{1}{1+r_G(1-)})$. Since the contours of $\Delta V(R, P) = K$, $K > 0$, are as in Figure 2.4 for Case II, the maximum value of $\Delta V(R, P)$ is attained at $(R^*, P^*) = (1, \lambda)$. More formally, one has the following theorem.

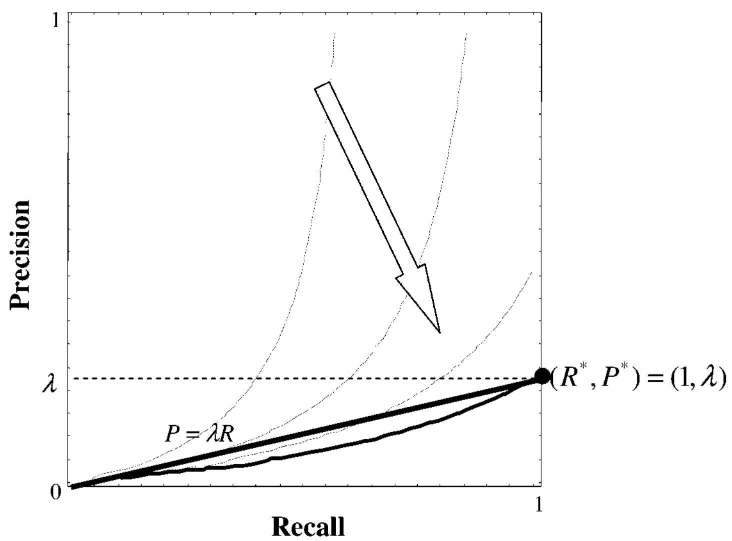


Figure 2.5: Optimal Structure for Case II

Theorem 2.2 For Case II of (2.11), Problem 2.1 has an optimal solution $z^* = 0$ with $(R^*, P^*) = (1, \lambda)$. This optimal solution is unique when either $x_{GB}(z)$ or $x_{GG}(z)$ in (1.7) is strictly decreasing.

Proof From (1.11), both $x_{GB}(z)$ and $x_{GG}(z)$ are noninreasing so that $x_{GB}(0) \geq x_{GB}(z)$ and $x_{GG}(0) \geq x_{GG}(z)$ for all $z \in [0, 1]$. It then follows from (2.3) and (2.5) that $\Delta V(0) - \Delta V(z) = \gamma_B(x_{GB}(0) - x_{GB}(z)) + \gamma_G(x_{GG}(0) - x_{GG}(z)) \geq 0$ since $0 < \gamma_B < \gamma_G$ for Case II. Clearly this inequality is strict when either $x_{GB}(z)$ or $x_{GG}(z)$ is strictly decreasing, completing the proof. ▮

For Case I of (2.11), one has $\gamma_B < 0 < \gamma_G$. This means that the positive returns of the promotional efforts from good customers in G are offset, to some extent, by the negative returns from bad customers in B, and the situation is more complicated. Since $R(z)$ and $P(z)$ cannot be determined until the purchasing records of all customers during the next future period become available, Problem 2.1 cannot be mathematically pursued further for Case I. In subsequent sections, we overcome this difficulty and analyze the above trade-off phenomenon by assuming certain functional forms of the cell functions.

3 Estimating Structure of Cell Functions and Optimal Threshold Level

In this section, we introduce a model where the structure of the cell functions $x_{ij}(z),\ i,j \in \{B,G\}$ in (1.7) is estimated so as to satisfy the conditions in (1.10) and (1.11). The model enables one to solve Problem 2.1 for Case I explicitly, yielding the unique optimal threshold level z^*. The usefulness of this approach will be validated in Section 4, where real customer purchase data are analyzed using LRM and SVM as segmentation methods.

While the cell functions $x_{ij}(z)$ take only discrete values, throughout this paper, we treat them as continuous functions in $z \in [0,1]$ for analytical simplicity. Consequently both $R(z)$ and $P(z)$ are also continuous. This approximation is reasonable when X_G and $X_{G(z)}$ are relatively large so that their reciprocals are small, say 10^{-3}. Clearly typical market data satisfy such conditions.

For $z \in [0,1]$, we estimate the four cell functions as

$$\begin{aligned} x_{BB}(z) &= X_B(1-\mathrm{e}^{-\alpha\frac{z}{1-z}}), & x_{BG}(z) &= X_G(1-\mathrm{e}^{-\beta\frac{z}{1-z}}), \\ x_{GB}(z) &= X_B\,\mathrm{e}^{-\alpha\frac{z}{1-z}}, & x_{GG}(z) &= X_G\,\mathrm{e}^{-\beta\frac{z}{1-z}} \end{aligned} \tag{3.1}$$

where $\alpha, \beta > 0$. It should be noted that all the conditions specified in (1.10) and (1.11) are satisfied by (3.1). In particular, the target customers are narrowed down more and more as $z \to 1-$. Accordingly both $x_{GB}(z)$ and $x_{GG}(z)$ decrease to 0 as $z \to 1-$. If an underlying segmentation method is trustworthy to some extent, it is then natural to assume that $x_{GB}(z)$ decreases to 0 faster than $x_{GG}(z)$ as $z \to 1-$, i.e. $r_G(z) = \frac{X_B}{X_G}\mathrm{e}^{-(\alpha-\beta)\frac{z}{1-z}} \to 0$ as $z \to 1-$. Consequently we assume that $\alpha > \beta > 0$ and hence

$$c = \frac{\alpha}{\beta} > 1. \tag{3.2}$$

From (1.13) and (3.1), one sees that Recall $R(z)$ can be written as

$$R(z) = \mathrm{e}^{-\beta\frac{z}{1-z}} \tag{3.3}$$

or equivalently

$$\frac{z}{1-z} = -\frac{1}{\beta}\log R(z). \tag{3.4}$$

For Precision $P(z)$, one sees from (1.18), (3.1) and (3.3) that

$$P(z) = R(z)\frac{X_G}{X_{G(z)}} = R(z)\frac{1}{\frac{X_B}{X_G}\,\mathrm{e}^{-\alpha\frac{z}{1-z}} + R(z)}.$$

Substituting (3.4) into the above equation, it then follows that

$$P(z) = \frac{1}{1 + w(\lambda)R(z)^{c-1}} \tag{3.5}$$

where λ is as given in (1.16) and $w(x)$ is defined by

$$w(x) = \frac{1-x}{x}. \tag{3.6}$$

Given a segmentation method, the assumed structure of the cell functions in (3.1) enables one to restate Problem 2.1 on a concrete basis for Case I. Furthermore, as we will see, the problem can be solved explicitly, yielding the unique optimal threshold level $z^* \in [0,1]$ which maximizes the financial effectiveness of the marketing campaign.

Problem 3.1

1. Find $(R^*, P^*) \in [0,1] \times [0,1]$ by solving

$$\begin{aligned} &\text{maximize} && F(R,P) = R\left(\gamma_B w(P) + \gamma_G\right) \\ &\text{subject to} && G(R,P) = P\left\{1 + w(\lambda) R^{c-1}\right\} - 1 = 0 \\ & && 0 \le R, P \le 1. \end{aligned} \tag{3.7}$$

2. Set

$$z^* = \left[1 - \frac{\beta}{\log R^*}\right]^{-1}.$$

We recall that $F(R,P)$ is as given in (2.7), while $G(R,P)$ of (3.7) is derived from (3.5). The value z^* is then obtained from (3.4). For Case I of (2.11), Problem 3.1 can be solved explicitly yielding the unique optimal solution. Two preliminary lemmas are needed. We define $P_{const}(R)$ as a function of $R \in [0,1]$ obtained by solving the constraint $G(R,P) = 0$ for P.

Lemma 3.2 Let

$$P_{const}(R) = \left[1 + w(\lambda) R^{c-1}\right]^{-1}. \tag{3.8}$$

Then, under (3.2), $P_{const}(R)$ is strictly decreasing in $R \in [0,1]$.

Proof One sees that

$$\frac{\mathrm{d}}{\mathrm{d}R} P_{const}(R) = -w(\lambda)\,(c-1)\,R^{c-2}\left[1 + w(\lambda) R^{c-1}\right]^{-2}.$$

From (1.16) and (3.2), one has $\frac{1}{\lambda} > 1$ and $c > 1$ so that $w(\lambda) > 0$ and $c - 1 > 0$. Consequently $\frac{\mathrm{d}}{\mathrm{d}R} P_{const}(R) < 0$ and the lemma follows. ∎

Lemma 3.3 For case I of (2.11), let

$$\tilde{P} = \frac{-\gamma_B \cdot c}{\gamma_G - \gamma_B \cdot c}; \;\; \tilde{R} = \left[\frac{w(\tilde{P})}{w(\lambda)}\right]^{\frac{1}{c-1}}, \tag{3.9}$$

where c is as given in (3.2). Then one has $0 < \tilde{P} < 1$ and $0 < \tilde{R} < 1$.

Proof Since $\gamma_B < 0 < \gamma_G$ for Case I and $c > 1$ from (3.2), one has

$$0 < \tilde{P} = \frac{-\gamma_B \cdot c}{\gamma_G - \gamma_B \cdot c} = \frac{c}{c - \frac{\gamma_G}{\gamma_B}} < 1.$$

For $\tilde{R}$, one obseves from Lemma 3.2 that $P_{const}(R)$ is strictly decreasing in R and hence $P_{const}(R) > \lambda = P_{const}(1)$ for all $R \in [0,1)$. In particular, we note that $\tilde{P} = P_{const}(\tilde{R}) > \lambda$.

Since $w(x)$ is strictly decreasing in $x \in [0,1]$ from (3.6), one has $w(\tilde{P}) < w(\lambda)$ and hence $w(\tilde{P})/w(\lambda) < 1$. It then follows from (3.9) that $0 < \tilde{R} < 1$ since $c > 1$, completing the proof. ∎

We are now in a position to prove the following theorem.

Theorem 3.4 For Case I of (2.11), suppose the cell functions are as given in (3.1). Then (R^*, P^*, z^*) is optimal for Problem 3.1 if and only if

$$P^* = -\frac{\gamma_B \cdot c}{\gamma_G - \gamma_B \cdot c}; \quad R^* = \left[\frac{w(P^*)}{w(\lambda)}\right]^{\frac{1}{c-1}}; \quad \text{and}$$
$$z^* = \left[1 - \frac{\beta}{\log R^*}\right]^{-1}. \tag{3.10}$$

Proof Suppose $(\hat{R}, \hat{P}, \hat{z})$ is an optimal solution of Problem 3.1. Let $L(R, P, \xi)$ be the Lagrangian function defined by

$$L(R, P, \xi) = F(R, P) + \xi G(R, P). \tag{3.11}$$

Then $(\hat{R}, \hat{P})$ should satisfy the first order necessary conditions for optimality written as

$$\begin{cases} \dfrac{\partial L}{\partial R} = \gamma_B \cdot w(P) + \gamma_G + \xi P w(\lambda)\,(c-1)\,R^{c-2} = 0, \\ \dfrac{\partial L}{\partial P} = -\dfrac{R}{P^2}\gamma_B + \xi\left\{1 + w(\lambda) R^{c-1}\right\} = 0, \\ \dfrac{\partial L}{\partial \xi} = G(R, P) = 0. \end{cases} \tag{3.12}$$

After a little algebra, one then finds that

$$\begin{cases} \hat{P} = \dfrac{-\gamma_B \cdot c}{\gamma_G - \gamma_B \cdot c} = P^*; \\ \hat{R} = \left[\dfrac{w(P^*)}{w(\lambda)}\right]^{\frac{1}{c-1}} = R^*; \\ \xi^* = \dfrac{R^* \gamma_B}{P^*}. \end{cases} \tag{3.13}$$

From Lemma 3.3, one has $0 < R^*, P^* < 1$ and z^* is obtained from (3.4).

Conversely, let (R^*, P^*) be as in (3.10). From Lemma 3.3, one has $0 < R^*, P^* < 1$. For notational convenience, we define

$$d = \frac{\gamma_G}{\gamma_B} < 0 \tag{3.14}$$

where $d < 0$ holds since $\gamma_B < 0 < \gamma_G$ for Case I. From (3.10) and (3.14), one finds that

$$P^* = -\frac{\gamma_B \cdot c}{\gamma_G - \gamma_B \cdot c} = \frac{c}{c-d}, \tag{3.15}$$

and hence

$$w(P^*) = \frac{1-P^*}{P^*} = \frac{1-\frac{c}{c-d}}{\frac{c}{c-d}} = \frac{-d}{c}. \tag{3.16}$$

Furthermore, from (3.8) with $P = P_{const}(R)$, one has

$$w(P) = \frac{1-P}{P} = w(\lambda)R^{c-1}. \tag{3.17}$$

We now show that, for $F(R,P)$ of (3.7),

$$F(R^*,P^*) - F(R,P) > 0 \quad \text{for all } (R,P) \in [0,1]\times[0,1], R \neq R^*, P \neq P^*. \tag{3.18}$$

It can be readily seen from (3.7), (3.10) and (3.17) that

$$F(R^*,P^*) - F(R,P) = \gamma_B \cdot H(R) \tag{3.19}$$

where

$$H(R) = \left[\frac{w(P^*)}{w(\lambda)}\right]^{\frac{1}{c-1}} \left\{w(P^*) + d\right\} - R\left\{w(\lambda)R^{c-1} + d\right\}. \tag{3.20}$$

Using (3.16), it then follows that

$$H(R) = -w(\lambda)R^c - dR + V; \quad V = -\left[\frac{-d}{cw(\lambda)}\right]^{\frac{1}{c-1}} dw(c). \tag{3.21}$$

We note that $w(c) < 0$ since $c > 1$ from (3.2) so that

$$H(0) = V < 0. \tag{3.22}$$

One also sees that

$$\frac{\partial}{\partial R}H(R) = -cw(\lambda)R^{c-1} - d = 0 \Leftrightarrow R = \left[\frac{-d}{cw(\lambda)}\right]^{\frac{1}{c-1}} = R^* \tag{3.23}$$

where the last equality holds since $w(P^*) = \frac{-d}{c}$ from (3.16). Furthermore

$$\left(\frac{\partial}{\partial R}\right)^2 H(R) = -c(c-1)w(\lambda)R^{c-2} < 0. \tag{3.24}$$

Hence $H(R)$ is strictly concave in R with the global maxima attained at R^*. In addition, from (3.17) and (3.21), one has

$$\begin{aligned} H(R^*) &= -w(\lambda)R^{*c} - dR^* - R^*dw(c) \\ &= R^*\left[-w(\lambda)R^{*c-1} - d\left\{1 + w(c)\right\}\right] \\ &= R^*\left[w(P^*) - \frac{d}{c}\right], \end{aligned}$$

so that, from (3.16),

$$H(R^*) = 0. \tag{3.25}$$

It then follows from (3.22),(3.23),(3.24) and (3.25) that $H(R) < 0$ for $R \in [0,1]$ with $R \neq R^*$. Since $\gamma_B < 0$ for Case I, (3.18) holds true from (3.19), completing the proof. ∎

Given a segmentation method, Theorem 3.4 enables one to set the optimal threshold level so as to maximize the financial effectiveness of the promotional efforts. Then, of interest is to examine the usefulness of this theorem by comparing the result using the optimal threshold value with the result using the default threshold value of the given segmentation method. In the next section, we demonstrate the validity of our approach based on real customer purchase data.

4 Model Validation Based on Real Customer Purchase Data

In this section, we examine the validity of our approach discussed in Section 3 based on real customer purchase data using LRM and SVM as segmentation methods. A succinct summary of the two segmentation methods LRM and SVM is given in Appendix A for the reader's convenience.

We first describe the basic features of the real customer purchase data used for this analysis.

- Basic data: purchasing records by category collected from several drugstores in Japan with customer IDs
- Time period: July 2002-June 2003
- Number of categories: 96
- Number of customer IDs: 64,453
- Number of purchase incidences: 2,362,163
- Total sales in the period: about 1.56 billion yen

For validation purposes, the total period is decomposed into the following 4 periods:

- Period I: July-September, 2002
- Period II: October-December, 2002
- Period III: January-March, 2003
- Period IV: April-June, 2003

The data set for Period J is denoted by Data J, J =I, II, III, IV.

The validation procedure is summarized in Figure 4.1. Here, firstly, Data I is used to identify 36 basic components. Considered are the basic customer profile components such as sex and age, and the purchasing records related to RFM (Recency, Frequency, Monetary) represented by the latest purchasing date, the purchasing frequency and the purchasing amount, as well as grouping product categories and aggregating related records. In addition, certain indices are formed using multiple components. The pair-wise correlation analysis is then conducted among the 36 components, eliminating one component for each pair with correlation coefficient greater than 0.76. As a result, 26 components are selected to construct customer profile vectors $\boldsymbol{x}_i(J), 1 \leq i \leq N, J = \mathrm{I}, \mathrm{II}, \mathrm{III}, \mathrm{IV}$. We note that any customer without purchasing records during Period J is excluded from analysis of Period J. A brief description of these 26 components is given in Appendix B.

We next define the set of good customers G_J in Period J as follows.

Definition 4.1 (The set of good customers in Period J)
When customers are listed in descending order of their purchasing amount during Period J, those within top $\pi\%$ constitute the set of good customers G_J in Period J. The determinant variable of c_i for Period J, denoted by $y_i(J)$, is then given as

$$(4.1) \qquad y_i(J) = \begin{cases} 1 & \text{if } c_i \in G_J \text{ in Period } J, \\ -1 & \text{if } c_i \in B_J \text{ in Period } J. \end{cases}$$

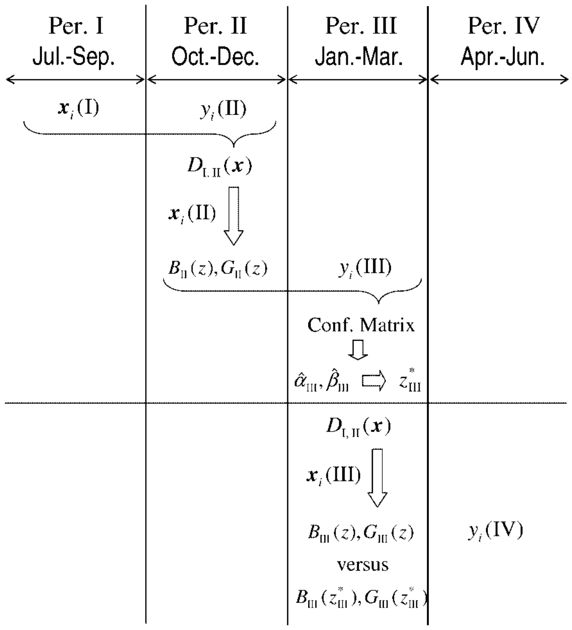


Figure 4.1: Validation Procedure

For $\boldsymbol{x}_i(\mathrm{I})$ and $y_i(\mathrm{II})$, each of the two segmentation methods can be applied, resulting in $B_{\mathrm{II}}(z)$ and $G_{\mathrm{II}}(z)$ for $z \in [0,1]$. More specifically, for LRM, the segmentation function based on $\boldsymbol{x}_i(\mathrm{I})$ and $y_i(\mathrm{II})$, denoted by $D_{L:\mathrm{I,II}}$, is given by

$$\text{(4.2)} \qquad \text{LRM:} \quad D_{L:\mathrm{I,II}}(\boldsymbol{x}_i(J)) = [1+\exp\{-(\hat{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{x}_i(J)+\hat{b}_0)\}]^{-1},$$

where $\hat{\boldsymbol{b}}$ and $\hat{b}_0$ are obtained as shown in Appendix A.1. $B_J(z)$ and $G_J(z)$ are then determined for $J = \mathrm{II}, \mathrm{III}$ by applying (4.2) to (1.5). In this paper, Clementine 7.1 by SPSS Inc. is employed for estimating parameter values, where all of 26 components given in Appendix B are forced to be used. For SVM, the optimal separating hyperplane is generated by determining the vector $\boldsymbol{w}^*$ and the constant w_0^* together with the relaxation vector $\boldsymbol{\xi}^*$ where $y_i(\mathrm{II})g(\boldsymbol{x}_i(\mathrm{I})) \geq 1-\xi_i^*$ with $g(\boldsymbol{x}) = \boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x}+w_0^*$, by following the procedure specified in Appendix A.2. In turn, the segmentation function $D_{S:\mathrm{I,II}}$ is obtained for $J = \mathrm{II}, \mathrm{III}$ as

$$\text{(4.3)} \qquad \text{SVM:} \quad D_{S:\mathrm{I,II}}(\boldsymbol{x}_i(J)) = \frac{g(\boldsymbol{x}_i(J)) - g(\boldsymbol{x}_{\min}(\mathrm{I}))}{g(\boldsymbol{x}_{\max}(\mathrm{I})) - g(\boldsymbol{x}_{\min}(\mathrm{I}))}$$

where $g(\boldsymbol{x}_{\max}(\mathrm{I})) = \max_{1\leq i\leq N}[g(\boldsymbol{x}_i(\mathrm{I}))]$ and $g(\boldsymbol{x}_{\min}(\mathrm{I})) = \min_{1\leq i\leq N}[g(\boldsymbol{x}_i(\mathrm{I}))]$. It should be noted that $D_{S:\mathrm{I,II}}(\boldsymbol{x}_i(J))$ may not belong to [0,1) for $J = \mathrm{II}, \mathrm{III}$. When the value exceeds 1, we replace it by $1-\epsilon$ for sufficiently small $\epsilon > 0$. If the value becomes negative, it is redefined as 0. With these modifications, one has $D_{S:\mathrm{I,II}}(\boldsymbol{x}_i(J)) \in [0,1)$, and $B_J(z)$ and $G_J(z)$ can be determined for $J = \mathrm{II}, \mathrm{III}$ from (1.5). In this paper, SVM^{light}5.00 [4] with parameter $C = N/\sum_{i=1}^N \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i$ by default is employed for constructing the segmentation function. Nonlinear segmentation is discarded after pre-testing.

Based on $B_{\mathrm{II}}(z)$ and $G_{\mathrm{II}}(z)$ together with $y_i(\mathrm{III})$, the confusion matrix can be constructed for each $z \in [0,1]$. By taking the values of z from 0 to 1 with a step size of 0.05, and using the cell funtion structure assumed in (3.1), the least square estimates $\hat{\alpha}_{\mathrm{III}}$ and $\hat{\beta}_{\mathrm{III}}$ can be obtained. These estimates, in turn, yield the optimal threshold level z^*_{III} via Theorem 3.4. Finally, in order to examine the financial impact of the optimal threshold level against nonoptimal threshold levels, the objective value $\Delta V(z^*_{\rm III})$ based on the confusion matrix generated by $B_{\rm III}(z^*_{\rm III})$, $G_{\rm III}(z^*_{\rm III})$ and $y_i(\rm IV)$ is compared with $\Delta V(z)$ calculated from the confusion matrix obtained by $B_{\rm III}(z)$, $G_{\rm III}(z)$ and $y_i(\rm IV)$ where $z = 0.5$ as the default value for LRM and $z = -g(\boldsymbol{x}_{\min}(\rm I))/\{g(\boldsymbol{x}_{\max}(\rm I)) - g(\boldsymbol{x}_{\min}(\rm I))\}$ for SVM.

Table 4.1 summarizes the least square estimates $\hat{\alpha}_{\rm III}$ and $\hat{\beta}_{\rm III}$ for LRM and SVM with $\pi = 5, 10, 15, 20\%$. These estimates are calculated using the confusion matrix generated from $B_{\rm II}(z)$, $G_{\rm II}(z)$ and $y_i(\rm III)$. In Figures 4.2 and 4.3, the estimated cell functions are compared with the actual numbers for LRM and SVM respectively where $\pi = 10\%$. One finds that the cell functions for SVM approximate the actual numbers extremely well. With respect to LRM, both $x_{BB}(z)$ and $x_{GB}(z)$ are well approximated, while some discrepancy can be observed for $x_{BG}(z)$ and $x_{GG}(z)$ for z not near 0, 0.5 or 1.0. Nevertheless, the use of the optimal z^* yields the substantial improvement over the use of the default values as we see next.

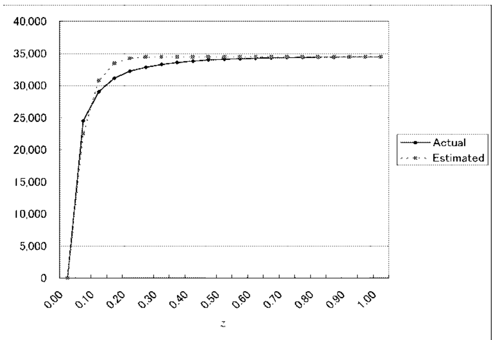
Table 4.1: Least Square Estimates

	LRM		SVM	
π	$\hat{\alpha}_{\rm III}$	$\hat{\beta}_{\rm III}$	$\hat{\alpha}_{\rm III}$	$\hat{\beta}_{\rm III}$
5%	34.67	1.89	23.02	4.38
10%	20.05	1.23	23.72	5.30
15%	12.46	1.14	19.22	5.46
20%	8.20	0.84	25.91	6.90

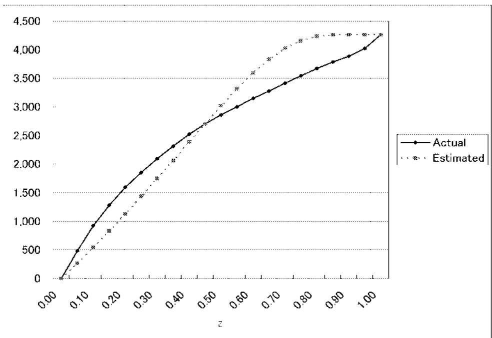
The expected revenues θ_B and θ_G from each of those in B and G respectively are estimated from Data III and Data IV as shown in Table 4.2, where θ_B and θ_G based on Data III are used to find the optimal threshold level $z^*_{\rm III}$ while those estimated from Data IV are employed to compute $\Delta V(z)$ for the actual validation. Computational experiments are then conducted extensively covering a wide range of parameter values of $(\eta_B,\ \eta_G,\ \nu)$ where η_B and η_G are the purchase increase factors of those in B and G respectively, and ν is the cost per customer for the promotional efforts. It is found that the use of the optimal z^* uniformly outperforms the use of the default values by a factor ranging from 1.5 to 2.8 for LRM and from 1.2 to 7.5 for SVM. It is also observed that LRM is consistently superior to SVM for this data set. Figures 4.4 and 4.5 exhibit some representative cases for LRM and SVM respectively.

Table 4.2: Expected Revenues (Unit: Yen)

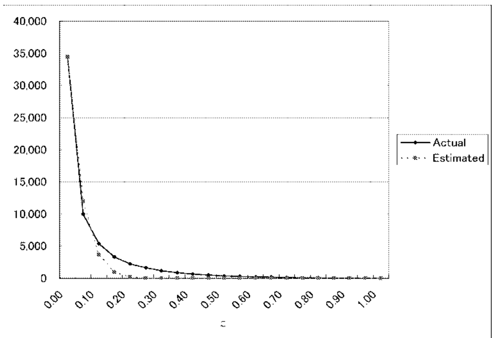
	Data III		Data IV	
π	θ_B	θ_G	θ_B	θ_G
5%	6,982.9	40,758.5	6,711.1	40,318.5
10%	6,066.4	32,121.1	5,817.4	31,554.8
15%	5,366.6	27,402.0	5,140.7	26,811.3
20%	4,787.5	24,207.8	4,584.5	23,617.7



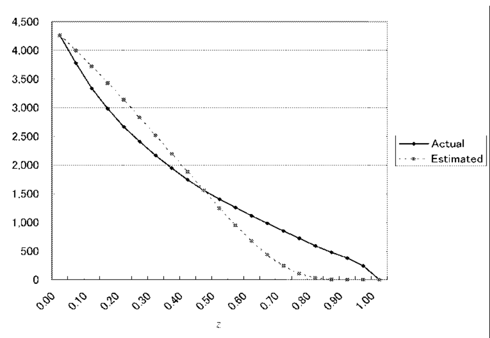
(a) $x_{BB}(z)$



(b) $x_{BG}(z)$

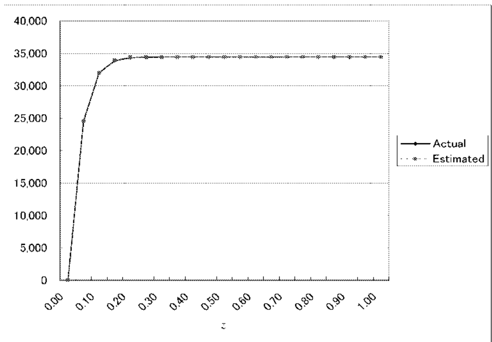


(c) $x_{GB}(z)$

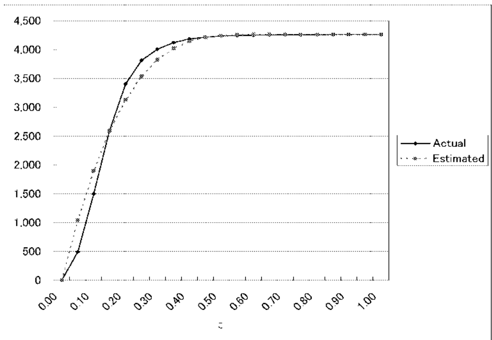


(d) $x_{GG}(z)$

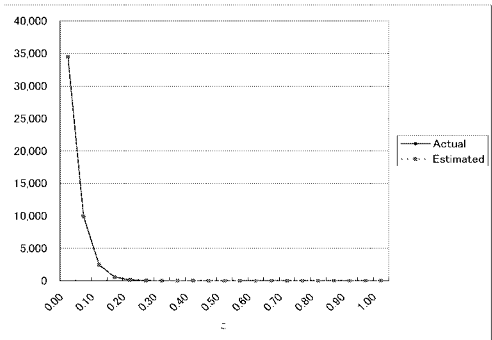
Figure 4.2: Cell Functions for LRM



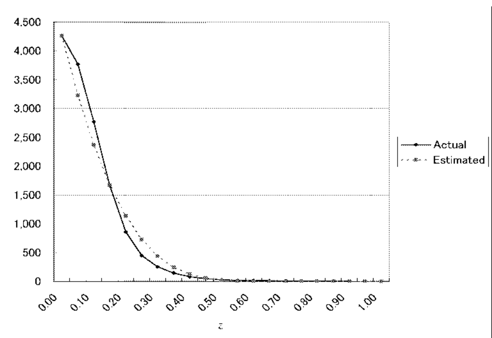
(a) $x_{BB}(z)$



(b) $x_{BG}(z)$

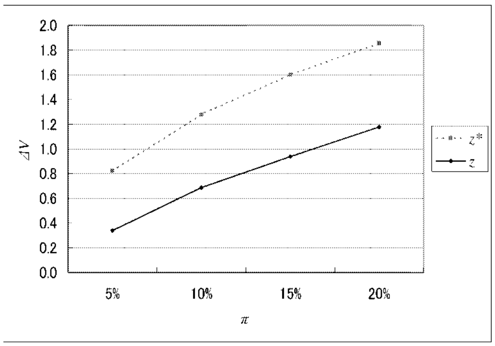


(c) $x_{GB}(z)$

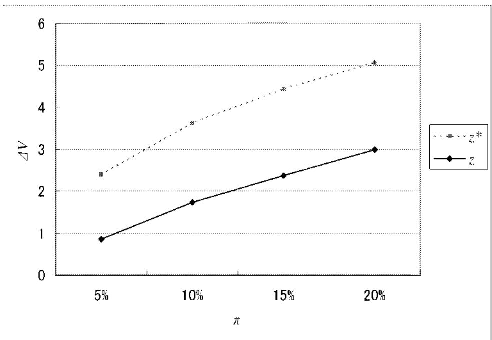


(d) $x_{GG}(z)$

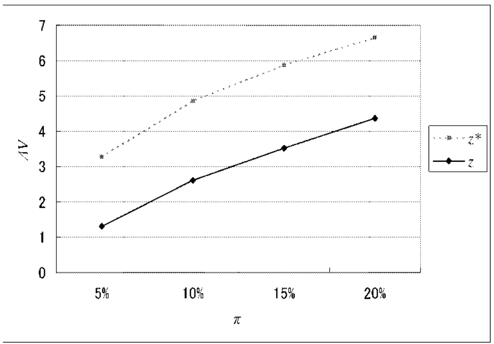
Figure 4.3: Cell Functions for SVM



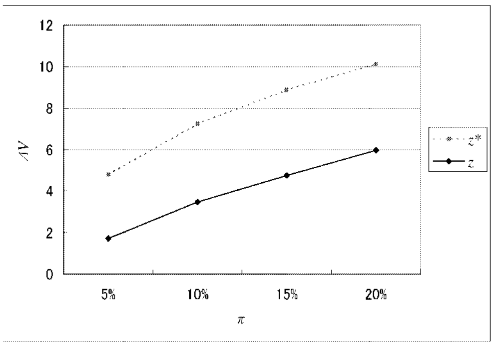
(a) $\eta_B = 0,\ \eta_G = 2,\ \nu = 100$ (Yen)



(b) $\eta_B = 1,\ \eta_G = 5,\ \nu = 250$ (Yen)

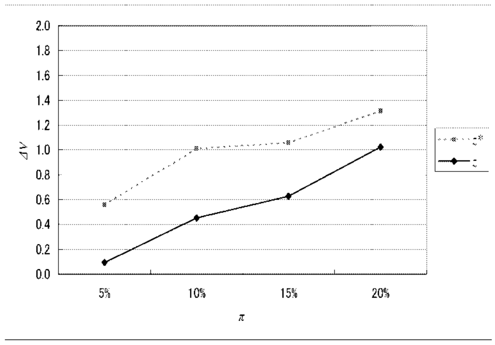


(c) $\eta_B = 2,\ \eta_G = 8,\ \nu = 500$ (Yen)

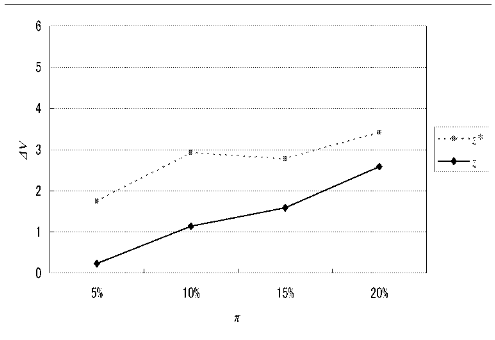


(d) $\eta_B = 2,\ \eta_G = 10,\ \nu = 500$ (Yen)

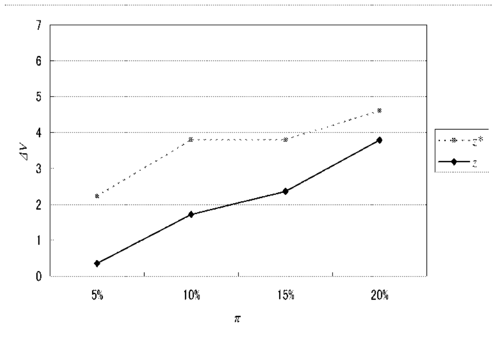
Figure 4.4: ΔV for LRM (Unit: Million yen)



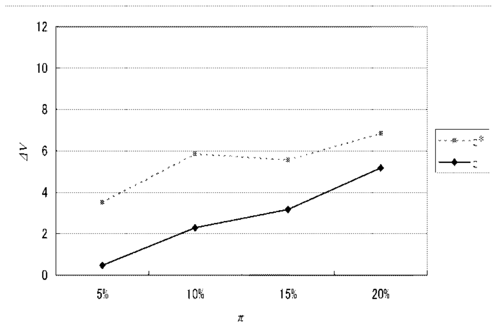
(a) $\eta_B = 0,\ \eta_G = 2,\ \nu = 100$ (Yen)



(b) $\eta_B = 1,\ \eta_G = 5,\ \nu = 250$ (Yen)



(c) $\eta_B = 2,\ \eta_G = 8,\ \nu = 500$ (Yen)



(d) $\eta_B = 2,\ \eta_G = 10,\ \nu = 500$ (Yen)

Figure 4.5: ΔV for SVM (Unit: Million yen)

5 Concluding Remarks

Traditionally, the performance of segmentation methods has been evaluated by Accuracy, which fails to capture separately two types of errors: Type I Error for misidentifying good customers as bad customers and Type II Error for misunderstanding bad customers as good customers. Assuming that a promotional campaign is addressed exclusively to the estimated good customers, the former represents some opportunity loss while the latter results in the inefficient use of the campaign budget. Accordingly, it is important to incorporate the distinction between the two types of errors in evaluating the financial effectiveness of segmentation methods for marketing campaign. By achieving this objective, one may establish a foundation for understanding and comparing the performance characteristics of various segmentation methods from a new perspective.

In order to reflect the effects of the two types of errors, this paper first develops a mathematical model for describing a general structure of segmentation methods. Utilizing two performance measures, Recall and Precision, a financial measure is then established, which enables one to assess the trade-off between the opportunity loss and the ineffective use of the campaign budget mentioned above. This trade-off is parameterized by specifying a threshold level concerning the severeness for estimating good customers. Accordingly, given a segmentation method, an optimization problem can be formulated so as to maximize the financial measure by finding the optimal threshold level.

In general, this optimization problem cannot be solved since it involves the magnitudes of Type I Error and Type II Error represented by Recall and Precision, which are unknown until the purchasing records of the customers in the next future period become available. This difficulty is overcome by introducing the structure of the four cell functions characterized by two parameters α and β. Recall and Precision can then be expressed mathematically in terms of α and β, which in turn enables one to obtain the unique optimal solution in a closed form. Given a segmentation method, the two parameters can be estimated based on the past data available now, and consequently the optimal threshold level can be specified.

The validity of the above approach is tested using real customer purchase data where LRM (Logistic Regression Model) and SVM (Support Vector Machine) are employed as segmentation methods. Extensive numerical experiments reveal that the use of the optimal threshold level uniformly outperforms the use of the default values for both LRM and SVM. For this set of data, it is also found that LRM is superior to SVM.

Further extension of this research is in progress, including multi-layer targeting, different functional forms for the four cell functions, endogenous responses to promotion and additional tests for different sets of purchasing records. These results will be reported elsewhere.

Acknowledgement

This paper has been partially supported by MEXT Grant-in-Aid for Fundamental Research (C) 17510114.

References

[1] S. A. Alvarez: An Exact Analytical Relation among Recall, Precision and Classification Accuracy in Information Retrieval. *Technical Report BCCS-02-01* (Computer Science Department, Boston College, 2002).

[2] M. J. A. Berry and G. Linoff: *Mastering Data Mining: Art and Science of Customer Relationship Management* (John Wiley & Sons, 1999).

[3] T. Hastie, R. Ribshirani and J. Friedman: *The Elements of Statistical Learning: data mining, inference, and prediction* (Springer, 2001).

[4] T. Joachims: SVM^{light} Support Vector Machine.
`http://svmlight.joachims.org`

[5] C. J. van Rijsbergen: *Information Retrieval Second Edition* (Butterworths, 1979).

[6] V. N. Vapnik: *Statistical Learning Theory* (John Wiley & Sons, 1998).

Appendix A Segmentation Methods

We consider data points spread over three periods I, II and III. A set of data points for Period J is denoted by $\boldsymbol{x}_i(J)$ $(i = 1, \ldots, N)$. Let the two sets B_J and G_J represent a segmentation of the entire set for Period J, e.g. G_J is a set of good customers defined by top $\pi\%$ in the revenue ranking of Period J as in Definition 4.1 of this paper. Given $\boldsymbol{x}_i(J)$, it is known that either $i \in B_J$ or $i \in G_J$ in Period J. Let $y_i(J) \in \{-1, 1\}$ be determinant variables defined by $y_i(J) = -1$ if $i \in B_J$ and $y_i(J) = 1$ if $i \in G_J$ in Period J. A segmentation method is an algorithmic procedure based on $\boldsymbol{x}_i(\mathrm{I})$ and $y_i(\mathrm{II})$. The algorithmic procedure is then applied to $\boldsymbol{x}_i(\mathrm{II})$, resulting in the estimate of $y_i(\mathrm{III})$. The data set for Periods I and II constitutes the learning data, while the data set for Periods II and III is called the test data.

A.1 LRM (Logistic Regression Model)

LRM is a statistical approach for segmentation of data points, where a functional value is computed from the components of each data point and then compared with a threshold level to see whether or not the underlying point ought to be selected. More specifically, let $\boldsymbol{x}_i(J)$ $(i = 1, \ldots, N)$ be a set of data points for Period J (J=I, II), and suppose that $y_i(\mathrm{II})$ $(i = 1, \ldots, N)$ are known for Period II. Using these data, the problem is how to predict the values of $y_i(\mathrm{III})$. In this regard, given $\boldsymbol{x}_i = \boldsymbol{x}_i(\mathrm{II})$, the probability of occuring $y_i(\mathrm{III}) = 1$, i.e. $i \in G_{\mathrm{III}}$, is assumed to take a form

$$\text{(A.1)} \qquad p(\boldsymbol{x}_i) = D(Z(\boldsymbol{x}_i)) = \frac{\exp(Z(\boldsymbol{x}_i))}{1 + \exp(Z(\boldsymbol{x}_i))} = \frac{1}{1 + \exp(-Z(\boldsymbol{x}_i))},$$

where $Z(\boldsymbol{x}) = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x} + b_0$. From (A.1), it follows that

$$\text{(A.2)} \qquad \log\frac{p(\boldsymbol{x}_i)}{1 - p(\boldsymbol{x}_i)} = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}_i + b_0.$$

In general, the maximum likelihood approach is employed to find the estimates $\hat{\boldsymbol{b}}$ and $\hat{b}_0$ using $\boldsymbol{x}_i(\mathrm{I})$ and $y_i(\mathrm{II})$, i.e.

$$\text{(A.3)} \qquad L(\hat{\boldsymbol{b}}, \hat{b}_0) = \max_{\boldsymbol{b}, b_0}\left[\prod_{i \in G_{\mathrm{II}}} p(\boldsymbol{x}_i(\mathrm{I})) \times \prod_{i \in B_{\mathrm{II}}} \{1 - p(\boldsymbol{x}_i(\mathrm{I}))\}\right].$$

Given a threshold level $z \in [0, 1]$, the segmentation set $G_{\mathrm{III}}(z)$ for Period III is then obtained as

$$\text{(A.4)} \qquad G_{\mathrm{III}}(z) = \{i : \hat{p}(\boldsymbol{x}_i) = D(\hat{Z}(\boldsymbol{x}_i)) \geq z\},$$

where $\hat{Z}(\boldsymbol{x}) = \hat{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{x} + \hat{b}_0$. In general, $z = 0.5$ is taken by default. The reader is refered to Hastie, Tibshirani and Friedman [3] for further details.

A.2 SVM (Support Vector Machine)

SVM is also a segmentation method and its function is similar to that of LRM. The basic tool for SVM is a separating hyperplane of the form $g(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + w_0$. Let $f(\boldsymbol{x}) = \mathrm{sign}\{g(\boldsymbol{x})\}$ where $\mathrm{sign}\{x\} = 1$ if $x \geq 0$ and $\mathrm{sign}\{x\} = -1$ otherwise. Given $\boldsymbol{x}_i = \boldsymbol{x}_i(\mathrm{I})$ and $y_i = y_i(\mathrm{II})$, the first step is to find $\boldsymbol{w}^*$ and w_0^* satisfying

$$y_i = f(\boldsymbol{x}_i) = \mathrm{sign}\{g(\boldsymbol{x}_i)\}. \tag{A.5}$$

In general, the existence of such $\boldsymbol{w}^*$ and w_0^* is not guaranteed. In order to overcome this difficulty, we introduce relaxation variables $\xi_i \geq 0$ $(i = 1, ..., N)$ and additional constraints given by

$$\begin{cases} g(\boldsymbol{x}_i) \geq 1 - \xi_i & \text{if } \; y_i = 1, \text{ i.e. } i \in G_{\mathrm{II}}, \\ g(\boldsymbol{x}_i) \leq -1 + \xi_i & \text{if } \; y_i = -1, \text{ i.e. } i \in B_{\mathrm{II}}. \end{cases} \tag{A.6}$$

Figure A.1 illustrates $\boldsymbol{x}_i(\mathrm{I})$, B_{II}, G_{II}, and a separating hyperplane $H_0 : g(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + w_0 = 0$ together with a relaxation variable $\xi_i > 0$.

It is clear that the distance between the two hyperplanes $H_1 : g(\boldsymbol{x}) = 1$ and $H_2 : g(\boldsymbol{x}) = -1$ can be obtained as $2/\|\boldsymbol{w}\|$. In SVM, it is desirable to maximize this distance while some penalties imposed on positive relaxation variables are contained in some way. Accordingly, the following optimization problem can be formulated so as to determine an optimal separating hyperplane.

$$\begin{array}{ll} \underset{\boldsymbol{w}, w_0, \boldsymbol{\xi}}{\text{minimize}} & \dfrac{1}{2}\|\boldsymbol{w}\|^2 + C \sum\limits_{i=1}^{N} \xi_i \\ \text{subject to} & \forall i, \; y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + w_0) - (1 - \xi_i) \geq 0 \\ & \forall i, \; \xi_i \geq 0. \end{array} \tag{A.7}$$

Here, C represents the magnitude of penalties for positive relaxzation variables. Publicly available software packages for SVM often employ $C = N / \sum_{i=1}^{N} \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i$ as a default value.

The primal problem of (A.7) has the dual problem given below:

$$\begin{array}{ll} \underset{\boldsymbol{\alpha}}{\text{maximize}} & \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} - \dfrac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{\alpha} \\ \text{subject to} & \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y} = 0 \\ & \mathbf{0} \leq \boldsymbol{\alpha} \leq C\mathbf{1}, \end{array} \tag{A.8}$$

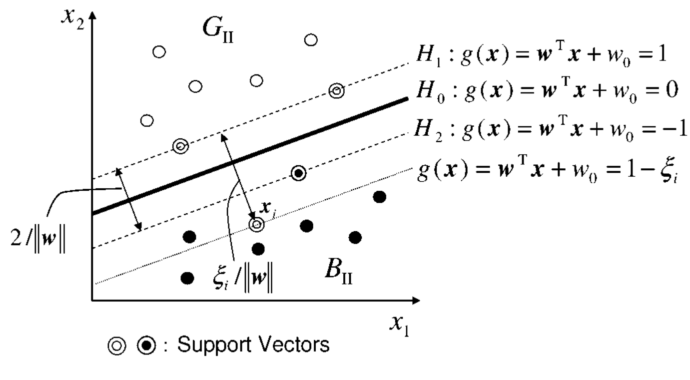


Figure A.1: Linear SVM

where $\mathbf{1} = [1, \ldots, 1]^{\mathrm{T}}$, $\boldsymbol{D} = \boldsymbol{y}_D \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{y}_D$ with $\boldsymbol{y}_D = \mathrm{diag}\{y_1, \ldots, y_N\}$, $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$ and $\boldsymbol{y} = [y_1, \ldots, y_N]^{\mathrm{T}}$. As can be seen, the dual problem is a concave quadratic programming problem which is much easier to solve than the primal problem.

Let $(\boldsymbol{w}^*, w_0^*)$ and $\boldsymbol{\alpha}^*$ be the optimal solution of the primal problem (A.7) and the dual problem (A.8) respectively. One then finds that

$$\boldsymbol{w}^{*\mathrm{T}} = \boldsymbol{\alpha}^{*\mathrm{T}} \boldsymbol{y}_D \boldsymbol{X}^{\mathrm{T}}; \quad w_0^* = y_k - \boldsymbol{\alpha}^{*\mathrm{T}} \boldsymbol{y}_D \boldsymbol{X}^{\mathrm{T}} \boldsymbol{x}_k, \tag{A.9}$$

where k is any index satisfying $0 < \alpha_k^* < C$. Let $g^*(\boldsymbol{x}) = \boldsymbol{w}^{*\mathrm{T}} \boldsymbol{x} + w_0^*$ and define the segmentation function $f^*(\boldsymbol{x}) = \mathrm{sign}\{g^*(\boldsymbol{x})\}$. It then follows that

$$G_{\mathrm{III}}(z) = \{i : f^*(\boldsymbol{x}_i(\mathrm{II})) = 1\}, \tag{A.10}$$

where this case is treated as the case of default. As is shown in Section 4, a general case for identifying good customers with a prespecified level $z \in [0, 1]$ can be described by

$$G_{\mathrm{III}}(z) = \{i : D(\boldsymbol{x}_i(\mathrm{II})) \geq z\}, \tag{A.11}$$

where D is given in (4.3). A vector $\boldsymbol{x}_i$ with $\alpha_i^* > 0$ is called a Support Vector. It is worth noting that the segmentation function $f^*(\boldsymbol{x})$ is contributed only by Support Vectors.

In some cases, it may be desirable to segmentize the data points by a nonlinear function. This becomes possible by introducing a Kernel function $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = h(\boldsymbol{x}_i)^{\mathrm{T}} \boldsymbol{\mu}_D h(\boldsymbol{x}_j)$, where $\boldsymbol{\mu}_D = \mathrm{diag}\{\mu_1, \ldots, \mu_M\}$ with $\mu_i \geq 0$ and $h : \mathbb{R}^L \to \mathbb{R}^M$ is a nonlinear function. Typically one has M much larger than L and the linear SVM approach can be applied in $\mathbb{R}^M$ which results in a nonlinear segmentation in $\mathbb{R}^L$. The reader is referred to Vapnik [6] for further details.

Appendix B Structure of Customer Profile Vectors

Table B.1: Components of Customer Profile Vector

Variable	Contents
1	the latest purchase date
2	the second latest purchase date
3	the total number of purchase incidents
4	the total purchase amount (in Yen)
5	RFM rank
6	the average purchase amount per purchase
7	the average number of purchase items per purchase
8	the total number of purchase incidents during weekends and holidays
9	Variable 8 / Variable 3
10	Index for similarity to the purchase pattern of the whole customers
11	the total purchase amount of items in the top 2 categories in terms of the total purchase amount of the whole customers
12	the number of categories purchased at least once in the top 2 categories within the period
13	the number of categories purchased at least once in Category A within the period
14	the total purchase amount of the items in Category B
15	the total number of purchase items in Category B
16	the number of categories purchased at least once in Category B within the period
17	the total purchase amount of the items in Category C
18	the number of categories purchased at least once in Category C within the period
19	the customer's decil value in Basic Cosmetics Category
20	the customer's decil value in Make-up Cosmetics Category
21	the customer's decil value in Non-Alcoholic Beverages Category
22	the customer's decil value in Medicine for Sensory Organ and Skin Category
23	the customer's decil value in Supplementary Nutrition Category
24	the customer's decil value in Sanitary Items Category
25	Gender
26	Age

Variables 1 and 2 are concerned with "Recency", while Variables 3, 8 and 9 are related to "Frequency". "Monetary" is expressed by Variables 4 and 6. Variable 5 called "RFM rank" is defined as the sum of the following three binary indices regarding the desirability of the customer in terms of Recency, Frequency and Monetary.

$$\text{R-desirability} = \begin{cases} 1 & (\text{Variable } 1 \leq 5) \\ 0 & (\text{Variable } 1 > 5) \end{cases}$$

$$\text{F-desirability} = \begin{cases} 1 & (\text{Variable } 3 \geq 7) \\ 0 & (\text{Variable } 3 < 7) \end{cases}$$

$$\text{M-desirability} = \begin{cases} 1 & (\text{Variable 4 belongs to the top 30\%}) \\ 0 & (\text{else}) \end{cases}$$

R-desirability and F-desirability are set in such a way that approximately 30% of the whole customers would have the value of 1 in Period J, J =I, II, III, IV.

Variable 10 is computed by multiplying the ratio of the purchase amount in each category against the total purchase amount of the customer by that of the whole customers, and then summing the result over all categories expressed in %.

The top 2 categories for Variables 11 and 12 are Basic Cosmetics Category and Make-up Cosmetics Category respectively throughout the four periods.

A, B and C Categories are determined based on ABC Analysis of the purchase amount of the whole customers in each period where categories that amount to top 75% constitute Category A, those contributing to the range between 75% and 90% are named Category B, with the rest called Category C.

The decil value used for Variables 19 through 24 is defined as $k/10$ when the customer belongs to the k-th decil of the specified category among the whole customers for each period with k =0, 1, 2, ... , 9 where decil 0 is the best. Those customers who do not purchase any item in the specified category are assigned the decil value of 1.0.

In this paper, SVM^{light} [4] by Joachims is used for implementing SVM. This software package requires the domain of customer-profile vectors to be approximately the L-dimensional cubic with the left lower corner at the origin. In order to meet this requirement, those variables whose domain is not $[0, 1]$ are scaled in the following manner. For $J =$ I, II, III, IV, let $V^*_{i:m}(J)$ be the value of Variable m for customer c_i in Period J after scaling. $V_{i:m}(J)$ is the corresponding value before scaling.

$$V^*_{i:m}(J) = \frac{V_{i:m}(J) - \min_{c_i \in CS} V_{i:m}(\mathrm{I})}{\max_{c_i \in CS} V_{i:m}(\mathrm{I}) - \min_{c_i \in CS} V_{i:m}(\mathrm{I})}.$$

Consequently, $V^*_{i:m}(J)$ are likely between 0 and 1 although it may become negative for $J \neq \mathrm{I}$. This scaling is expected to contain the effect of outliers.