Expert Systems with Applications 36 (2009) 5866–5871

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Knowledge discovery on RFM model using Bernoulli sequence

I-Cheng Yeh [a,*], King-Jang Yang [b,1], Tao-Ming Ting [c,2]

[a] Department of Information Management, Chung Hua University, No. 707, Section 2, WuFu Road, Hsin Chu 30067, Taiwan, ROC
[b] Department of Applied Mathematics, Chung Hua University, Hsin Chu 30067, Taiwan, ROC
[c] Department of Business and Entrepreneurial, Kainan University, Taoyuan 33857, Taiwan, R.O.C

ARTICLE INFO

Keywords:
Knowledge discovery
RFM model
Marketing
Bernoulli sequence

ABSTRACT

The objective of this paper is to introduce a comprehensive methodology to discover the knowledge for selecting targets for direct marketing from a database. This study expanded RFM model by including two parameters, time since first purchase and churn probability. Using Bernoulli sequence in probability theory, we derive out the formula that can estimate the probability that one customer will buy at the next time, and the expected value of the total number of times that the customer will buy in the future. This study also proposed the methodology to estimate the unknown parameters in the formula. This methodology leads to more efficient and accurate selection procedures than the existing ones. In the empirical part we examine a case study, blood transfusion service, to show that our methodology has greater predictive accuracy than traditional RFM approaches.

1. Introduction

Since the early 1980s, the concept of relationship management (CRM) in marketing area has gained its importance. Acquiring and retaining the most profitable customers are serious concerns of a company to perform more targeted marketing campaigns (Bult & Wansbeek, 1995; Hughes, 1994; Hwang, Jung, & Suh, 2004; Kahan, 1998; Malthouse & Blattberg, 2004; Schmittlein, Morrison, & Colombo, 1987; Stone, 1995). For effective customer relationship management, it is important to gather information on customer value. The most powerful and simplest model to implement CRM may be the RFM model – Recency, Frequency, and Monetary value. RFM is a behavior-based model, meaning it is used to analyze the behavior a customer is engaging in, and make predictions based on this behavior (Colombo & Jiang, 1999; Hughes, 1996).

RFM has a corollary: Customers who have purchased or visited more recently, more frequently, or created higher monetary values are much more likely to respond to your marketing efforts, compared with other customers who are less recent, less frequent, and create less monetary value.

Classic RFM implementation ranks each customer on the recency, frequency, and monetary value parameters against all the other customers, and creates an RFM "score" for each customer. Customers with high scores are usually the most profitable, the most likely to repeat a behavior (visit or purchase, for example), and the most highly responsive to promotions. The opposite is true for customers with low RFM scores. Moreover, some researchers pointed out that a high recency/frequency/monetary value customer who stops visiting is a customer who is finding alternatives to old site.

The advantage of RFM model is that its concept is extraordinarily intuition, its implementation is very simple, and its computation is rather easy. Marketing personnel can analyze customers without assistance of professional computer information systems, so it has already used for a long time in the business circles. But RFM model has some shortcomings (Colombo & Jiang, 1999; Hughes, 1996; Miglautsch, 2000):

(1) RFM model is not a precise quantitative predict model; using customers' historical trading data, it appraise customers' value based on näive method and subjective judgment, such as five partition score systems.
(2) Each RFM parameter has different importance to different industries. For example, parameter R may have very good evaluation ability in some industries, while parameter F and M may have better evaluation ability in other industries. RFM model is unable to integrate these three parameters into a single precise quantitative evaluation index.
(3) While RFM analysis is popular among marketing personnel, ad-hoc rules are often employed to judge whether customers are still active or not. Because customers do not declare explicitly when they have found alternative sites to the old one, a company infers a customer is out of their business if he/she did not make any purchase, for example, for over three months.

* Corresponding author. Tel.: +886 3 5186511.
E-mail addresses: icyeh@chu.edu.tw (I.-C. Yeh), kingjang@chu.edu.tw (K.-J. Yang), dmhk.tw@yahoo.com.tw (T.-M. Ting).
[1] Tel.: +886 3 5186388.
[2] Tel.: +886 3 5186388/3412500x1551.

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.07.018

Nomenclature

b	difference between the period when a customer remains active and the period from the first response and the most recent response
B_k	binary variable for the kth sorted customer, $B_k = 1$ means this customer took response, $B_k = 0$ means this customer did not take response
$E(X_L)$	expected value of times that a customer will respond in the next L marketing campaigns
$E_i(X_L \vert L = 1)$	expected value of times that the customer will respond at the next marketing campaign
F	number of responses in the period when the customer is active
G	period when a customer remains active
g	ratio of the response probability of the current campaign and the average of response probabilities of previous campaigns
M	monetary value, total monetary amount that a customer purchases up to now
m	number of data for moving average
n	times of recent consecutive marketing campaigns that a customer had no response
P	response probability, the probability that a specific customer respond to a marketing campaign
$P_{B,i}$	actual value of response probability for the customer in the next marketing campaign
Q	churn probability, the probability that a customer discontinue his/her use of a service for ever after a marketing campaign
R	recency, time since the most recent response for a customer
S	number of marketing campaigns in the period when the customer is active
s	period per marketing campaign
T	time since the first response for a customer
X_L	number of times that a customer will respond in the next L marketing campaigns
Y	status variable; $Y = 0$ denotes a customer was still active under the condition that this customer had no response for n consecutive marketing campaigns. $Y = 1$ denotes a customer had been inactive under the condition that this customer had no response for n consecutive marketing campaigns

To improve the above-mentioned shortcomings, and enable the model to be adjusted based on the data of the marketing databases of different industries, in this study, we will derive an augmented RFM model, called RFMTC model (Recency, Frequency, and Monetary value, Time since first purchase, and Churn probability), using Bernoulli sequence in probability theory. The model can automatically build the formula that can predict the probability that one customer will buy at the next time, and the expected value of the total number of times that the customer will buy in the future. The main contribution of this study is to develop predictive formulas based on probability theory instead of appraise customers with naive methods, such as RFM score method.

In this paper, section two will derive the model to estimate the probability that one customer will buy at the next time, and the expected value of the total number of times that the customer will buy in the future. Section three proposes the method to estimate the parameters in the model. Section four will prove that the above-mentioned theory is reliable using a real case study. Section five gives the conclusion.

2. RFMTC marketing model

In order to obtain the mathematically theoretical solution of the RFMTC marketing model, we set the following hypotheses:

Hypothesis 1. The probability that a specific customer respond to a marketing campaign is the constant P, called response probability.

Hypothesis 2. A customer having response to a marketing campaign means that he/she is still active, i.e., the probability that this customer is still active is one.

Hypothesis 3. A customer having no response to a marketing campaign means that the probability that this customer is still active is $(1-Q)^n$, where Q is the churn probability, and n is the times of recent consecutive marketing campaigns that a customer had no response. Churn probability is defined as the probability that a customer discontinue his/her use of a service for ever after a marketing campaign.

Theorem 1. *Under the condition that a customer had no response for n consecutive marketing campaigns, the expectation of times X_L that this customer will respond in the next L marketing campaigns is*

$$E(X_L) = (1-Q)^n P \cdot \left[\sum_{k=0}^{L-1} k(1-Q)^k Q + L(1-Q)^L \right] \tag{1}$$

Proof

Let

$Y = 0$ denotes a customer was still active under the condition that this customer had no response for n consecutive marketing campaigns.
$Y = 1$ denotes a customer had been inactive under the condition that this customer had no response for n consecutive marketing campaigns.

Then

$$\begin{aligned} E(X_L) &= E(X_L|Y=0)\cdot \Pr(Y=0) + E(X_L|Y=1)\cdot \Pr(Y=1) \\ &= E(X_L|Y=0)\cdot \Pr(Y=0) + E(X_L|Y=1)\cdot (1-\Pr(Y=0)) \\ &= E(X_L|Y=0)\cdot (1-Q)^n + 0\cdot [1-(1-Q)^n] \\ &= E(X_L|Y=0)\cdot (1-Q)^n \end{aligned} \tag{2}$$

In the next L marketing campaigns, a customer may become inactive at the ith marketing campaign, $i = 1,2,3,\ldots,L$, or may be still active.

Let

C_i denotes the event that a customer became inactive at the ith marketing campaign, $i = 1,2,3,\ldots,L$, and
C_{L+1} denotes the event that a customer was still active after next L consecutive marketing campaigns.

When a customer became inactive at the kth marketing campaign, he/she would face first $k-1$ marketing campaigns. Since the response probability of each marketing campaign for a customer is P, assuming the process can be described as a Bernoulli sequence, by the formula of expectation of a binomial random variable, we obtain

$$E(X_L|Y=0,C_k)=(k-1)P \quad k=1,2,3,\dots,L \tag{3}$$

The probability that a customer would become inactive at the kth marketing campaign is the product of probabilities of customer being active in the first $k-1$ marketing campaigns and the probability of customer being inactive at the kth marketing campaign, that is

$$\Pr(C_k)=\overbrace{(1-Q)(1-Q)\cdots(1-Q)}^{k-1}\cdot Q=(1-Q)^{k-1}Q \quad k=1,2,3,\dots,L \tag{4}$$

Finally, the probability that a customer was still active after the L marketing campaigns is the products of probabilities of customer being active in the L marketing campaigns, that is

$$\Pr(C_{L+1})=(1-Q)^L \tag{5}$$

Table 1 shows the probabilities of event C_k and the expectations $E(X_L|Y=0,C_k)$, where $k=1,2,3,\dots,L,L+1$.

Therefore

$$\begin{aligned}E(X_L|Y=0)&=\sum_{k=1}^{L+1}\Pr(C_k)\cdot E(X_L|Y=0,C_k)\\&=0\cdot Q+P(1-Q)Q+2P(1-Q)^2Q+\cdots\\&\quad+(L-1)P(1-Q)^{L-1}Q+LP(1-Q)^L\\&=\left[\sum_{k=0}^{L-1}kP(1-Q)^kQ+LP(1-Q)^L\right]\end{aligned} \tag{6}$$

Substituting (6) into (2), we obtain

$$\begin{aligned}E(X_L)&=(1-Q)^n\left[\sum_{k=0}^{L-1}kP(1-Q)^kQ+LP(1-Q)^L\right]\\&=(1-Q)^nP\left[\sum_{k=0}^{L-1}k(1-Q)^kQ+L(1-Q)^L\right]\end{aligned} \tag{7}$$

and the theorem is proved.

Consequence 1.1. Under the same condition as Theorem 1, the expectation of a customer that will respond at the next marketing campaign is

$$E(X_L|L=1)=(1-Q)^{n+1}P \tag{8}$$

Proof. The result comes from (1) with $L=1$. □

Consequence 1.2. Under the same condition as Theorem 1, the expectation of times X_L that this customer will respond in all the following marketing campaigns is

$$E(X_L|L=\infty)=(1-Q)^n\frac{P(1-Q)}{Q}=(1-Q)^{n+1}\frac{P}{Q} \tag{9}$$

Proof. By (6) and letting $L\to\infty$, we have □

$$\begin{aligned}E(X_L|Y=0)&=\sum_{k=0}^{\infty}kPQ(1-Q)^k\\&=PQ(1-Q)\sum_{k=1}^{\infty}k(1-Q)^{k-1}\\&=PQ(1-Q)\sum_{k=1}^{\infty}\left[-\frac{d}{dQ}(1-Q)^k\right]\\&=PQ(1-Q)\left[-\frac{d}{dQ}\sum_{k=1}^{\infty}(1-Q)^k\right]\end{aligned} \tag{10}$$

From the formula of infinite series, when $x<1$, then we know

$$\sum_{k=1}^{\infty}x^k=\frac{x}{1-x} \tag{11}$$

Therefore

$$\sum_{k=1}^{\infty}(1-Q)^k=\frac{1-Q}{1-(1-Q)}=\frac{1-Q}{Q} \tag{12}$$

Substituting (12) into (10), we obtain

$$\begin{aligned}E(X_L|Y=0)&=PQ(1-Q)\left[-\frac{d}{dQ}\frac{1-Q}{Q}\right]\\&=PQ(1-Q)\left[-\frac{d}{dQ}\left(\frac{1}{Q}-1\right)\right]\\&=PQ(1-Q)\frac{1}{Q^2}=\frac{P(1-Q)}{Q}\end{aligned} \tag{13}$$

Substituting (13) into (1), we get

$$E(X_L)=(1-Q)^n\frac{P(1-Q)}{Q}=(1-Q)^{n+1}\frac{P}{Q} \tag{14}$$

and the proof is completed.

3. Parameter estimation of RFMTC marketing model

3.1. Undetermined parameters

In previous section, we obtained from Consequence 1.1 that

$$E(X_L|L=1)=(1-Q)^{n+1}P \tag{8}$$

In this equation, we need to estimate three parameters P, Q and n. In order to get their estimation, we make the following hypotheses:

Hypothesis 1. The probability P for a customer responding a marketing campaign could be estimated by

$$P=g\frac{F}{S} \tag{15}$$

where

S is the number of marketing campaigns in the period when the customer is active, and
F is the number of responses in the period when the customer is active, and
g is the ratio of the response probability of the current campaign and the average of response probabilities of previous campaigns.

Hypothesis 2. The number of marketing campaigns S in the period when the customer is active could be estimated by

$$S=\frac{G}{s} \tag{16}$$

where

Table 1
Probabilities of C_k and expectations $E(X_L{-}Y=0,C_k)$

Event	$Pr(C_k)$	$E(X_L{-}Y=0,C_k)$
C_1	Q	$0P$
C_2	$(1-Q)^1Q$	$1P$
C_3	$(1-Q)^2Q$	$2P$
C_4	$(1-Q)^3Q$	$3P$
:	:	:
C_{L-1}	$(1-Q)^{L-2}Q$	$(L-2)P$
C_L	$(1-Q)^{L-1}Q$	$(L-1)P$
C_{L+1}	$(1-Q)^L$	LP

s is the period per marketing campaign, and

G is the period when a customer remains active. The exact value of G is unknown; however, it must be greater than $T - R$, where T is the time since first response, and R is the time since the most recent response. Hence assume G can be estimated by

$$G = (T - R) + b \tag{17}$$

Substituting (17) into (16), we have

$$S = (T - R + b)/s \tag{18}$$

Assuming the period per marketing campaign is one day (or one week, one month) and all the unit of time in this model for all time variables is one day (or one week, one month), then we may let $s = 1$ in order to simplify formulas. Therefore, by (18), we have

$$S = (T - R + b)/1 = T - R + b \tag{19}$$

Hypothesis 3. Assume that n is the times of recent consecutive marketing campaigns that this customer had no response, and could be estimated by

$$n = \frac{R}{s} \tag{20}$$

where R is the time since the most recent response for a customer. If we assume that $s = 1$, then

$$n = R \tag{21}$$

Theorem 2. *Under the same condition as Theorem 1, the expectation of times X_L that this customer will respond at the next marketing campaign is*

$$E(X_L|L = 1) = (1 - Q)^{R+1} g \frac{F}{T - R + b} \tag{22}$$

Proof. Substituting (15), (19), and (21) into (8), the result is obtained.

In (22), R, F and T are known, but Q, g, b are unknown and must be estimated.

3.2. Method of parameter estimation

In order to estimate parameters Q, g, b, we present the following model:

$$\text{Find } Q, g, b \tag{23a}$$

$$\text{Min} \sum_i (E_i(X_L|L = 1) - P_{B,i})^2 \tag{23b}$$

where

$P_{B,i}$ is the actual value of response probability for the customer in the next marketing campaign, and

$E_i(X_L|L = 1)$ is the expected value of times X_L that the customer will respond at the next marketing campaign estimated by Eq. (22). Because the maximum value of it is one, it can be regarded as the expected value of response probability.

Supposing that the actual value of response probability can be provided by marketing database, the optimal estimated parameter values can be found by Eq. (23b). However, in the marketing database a customer will either respond or not respond in the marketing campaign, thus $P_{B,i}$ is unknown. We proposed the novel "Sorting Moving Average Method" (SMA) to estimate it:

1. Assign one set of values to Q, g, b, then compute the expected value of response probability $E_i(X_L|L = 1)$ for each customer by Eq. (22).
2. Sort customer data by the value $E_i(X_L|L = 1)$ from large to small.
3. Use the moving average to estimate the actual value of response probability by

$$P_{B,i} = \frac{\sum_k B_k}{2m + 1}, \quad k = i - m, i - m + 1, \ldots, i - 1, i, i + 1, \ldots, i + m - 1, i + m \tag{24}$$

where

B_k is the binary variable for the kth sorted customer, $B_k = 1$ means this customer took response, $B_k = 0$ means this customer did not take response, and

m is the number of data for moving average. In (23c), we center at the ith sorted customer and use $2m$ samples around this customer to estimate $P_{B,i}$.

After the above steps, if the assuming values of Q, g, b make (23b) reach its minimum, then they are the optimal estimation of the parameter values.

4. Case studies: blood transfusion service

To demonstrate the RFMTC marketing model, this study adopted the donor database of Blood Transfusion Service Center in Hsin-Chu City in Taiwan. The center passes their blood transfusion service bus to one university in Hsin-Chu City to gather blood donated about every three months. To build a FRMTC model, we selected 748 donors at random from the donor database. These 748 donor data, each one included R (Recency – months since last donation), F (Frequency – total number of donation), M (Monetary – total blood donated in c.c.), T (Time – months since first donation), and a binary variable representing whether he/she donated blood in March 2007 (1 stand for donating blood; 0 stands for not donating blood). Table 2 shows the descriptive statistics of the data. We selected 500 data at random as the training set, and the rest 248 as the testing set.

Because formula (23b) is a three variables optimization problem, it is easy to solve it by numerical optimization methods. To improve the efficiency of the solving process, it is necessary to estimate the rational range of these three parameters, Q, g, b, first

- A greater Q stands for that the customer churn probability is high. $0 \leqslant Q \leqslant 1$. According to general marketing experience, when the period per marketing campaign is one month, it may be between 0.01 and 0.2.
- A greater g stand for the prevalence of the current marketing campaign compared to a common one. If it greater than 1, this campaign is superior to an average one. Its value domain is assumed in the range between 0.25 and 4.0.
- The b stands for that the difference between the period when a customer remains active and the period from the first response and the most recent response. A reasonable guess is that b is about the average period per response; hence, its value domain is assumed in the range as follows

$$0.25\left(\frac{T - R}{F - 1}\right)_{\text{Avg}} < b < 4\left(\frac{T - R}{F - 1}\right)_{\text{Avg}} \tag{25}$$

Note when we estimate the value domain of parameter b by (23d), only the data whose frequency $F \geqslant 2$ are adopted. In this study

$$\left(\frac{T - R}{F - 1}\right)_{\text{Avg}} = 7.85 \tag{26}$$

Table 2
Descriptive statistics of the data

Variable	Minimum	Maximum	Mean	Standard deviation
Recency (months)	0.03	74.4	9.74	8.07
Frequency (times)	1	50	5.51	5.84
Monetary (c.c. blood)	250	12500	1378.68	1459.83
Time (months)	2.27	98.3	34.42	24.32

Table 3
Estimated parameter values

Parameter	Estimated value domain	Estimated value
Q	0.01–0.2	0.111
g	0.25–4	3.72
b	1.96–31.4	9.52

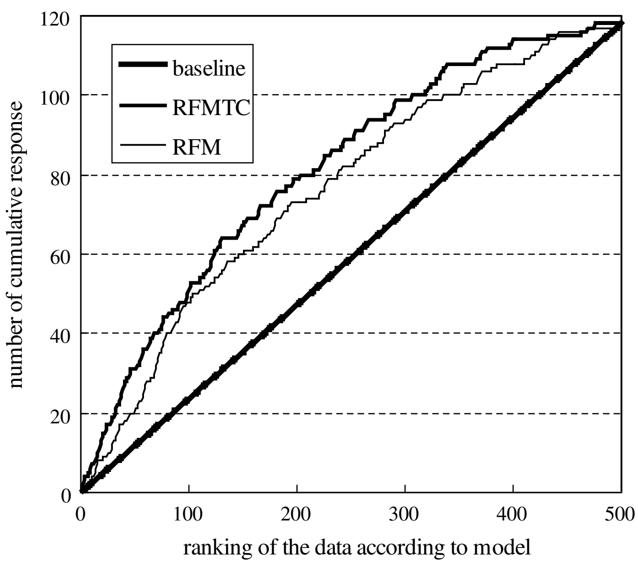

Fig. 1. Lift chart of RFMTC model and RFM model for training set.

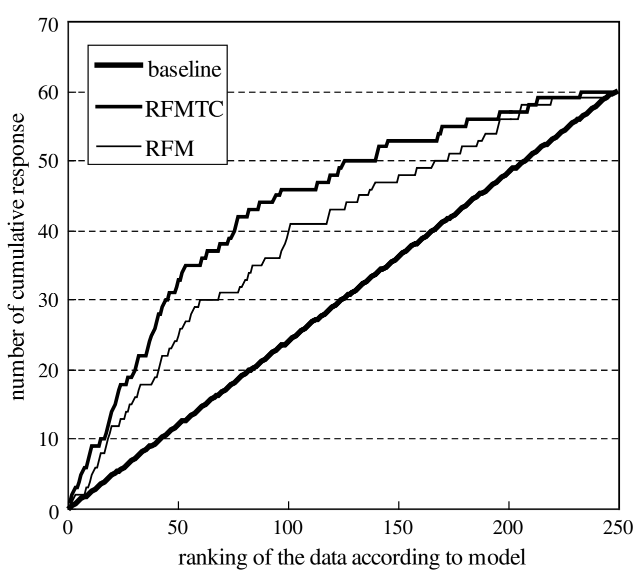

Fig. 2. Lift chart of RFMTC model and RFM model for testing set.

hence

$$1.96 < b < 31.4 \tag{27}$$

Using numerical optimization method, based on the above-mentioned value ranges, the most fitting parameters are shown in Table 3. No parameters reached their limits, which appears that the limits of these parameters may be reasonable.

To compare the performance of RFMTC and RFM model, lift chart was adopted. In the lift chart, the horizontal axis represents the ranking of the data according to predictive model, and the vertical axis shows the number of cumulative response. There are two curves, model curve and baseline curve, in the lift chart.

The RFM model is implemented based on the following scoring systems:

- R (months) score: 0–2 months = 5, 3 months = 4, 4–10 months = 3, 11–15 months = 2, 16 months and up = 1;
- F (times) score: 1 time = 1, 2 times = 2, 3–4 times = 3, 5–7 times = 4, 8 times and up = 5;
- M (c.c. blood) score: 250 c.c. = 1, 500 c.c. = 2, 750–1000 c.c. = 3, 1250–1750 c.c. = 4, 2000 c.c. and up = 5;
- RFM score = R score + F score + M score.

The lift chart obtained with RFMTC model and with RFM model is shown in Fig. 1 for the training set and Fig. 2 for the testing set. The RFMTC model curve is obviously above the RFM model curve in training set and in testing set, which shows the RFMTC model is really better than the RFM model in ranking targets from the population of the direct marketing campaign.

5. Conclusions

The most important issue for direct marketers is how to sample targets from a population for a direct marketing campaign. Although some selection methods are described in the literature, there seems to be few literatures discussing the analytical and statistical aspects. RFM model is a growing area of marketing practice, yet the academic journals contain very little research on this topic based on sound mathematic theory.

The objective of this paper is to introduce a comprehensive methodology for selecting targets for direct marketing from a database. This study expanded RFM model to RFMTC model by including two parameters, time since first purchase and churn probability. Using Bernoulli sequence in probability theory, we derive formulas that can predict the probability that one customer will buy at the next time, and the expected value of the total number of times that the customer will buy in the future.

This study also proposed the methodology to estimate the unknown parameters in the model. At least theoretically, this methodology leads to more efficient and accurate selection procedures than the existing ones. In the empirical part we show that our methodology has greater predictive accuracy than traditional RFM approaches.

Comparing RFMTC model with traditional RFM model, we find there are some advantages as follows:

(1) RFM model is not a precise quantitative prediction model; it appraises customers with rough methods, such as five partition systems, on the customers' historical trading data. RFMTC model can build the precise quantitative prediction model that can predict the probability that one customer will buy at the next time.
(2) RFM model cannot estimate the expected value of the total number of times that one customer will buy in the future. RFMTC model can deduce the formula to estimate it, using the infinite series skill.

(3) RFM model must adjust the weights of R, F, M parameter according to different industries, but this kind of adjustment is short of the systematic method. RFMTC model can build the optimum predictive model automatically based on the data of the marketing databases of different industries, and avoid the problem of adjusting the weights of parameters in a trial-and-error approach.
(4) RFM model needs to segment the customers into different groups in order to confirm the response rate of each group. RFMTC model don't need to segment the customers into different groups, and uses single customer group to confirm the response rate of each customer, so the necessary quantity of trials on customers can be greatly reduced.

References

Bult, J. R., & Wansbeek, T. (1995). Optimal selection for direct mail. *Marketing Science, 14*(4), 378–381.
Colombo, R., & Jiang, W. (1999). A stochastic RFM model. *Journal of Interactive Marketing, 13*(3), 2–12.
Hughes, A. M. (1994). *Strategic database marketing*. IL: Probus Publishing Company.
Hughes, A. M. (1996). Boosting response with RFM. *American Demographics, 5*, 4–9.
Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry. *Expert Systems with Applications, 26*(2), 181–188.
Kahan, R. (1998). Using database marketing techniques to enhance your one-to-one marketing initiatives. *Journal of Consumer Marketing, 15*(5), 491–493.
Malthouse, E. C., & Blattberg, R. C. (2004). Can we predict customer lifetime value? *Journal of Interactive Marketing, 19*(1), 2–16.
Miglautsch, J. (2000). Thoughts on RFM scoring. *Journal of Database Marketing, 8*(1), 67–72.
Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting your customers: Who are they and what will they do next? *Management Science, 33*(1), 1–24.
Stone, B. (1995). *Successful direct marketing methods* (Vol. 3). IL: NTC Business Books.