Cumhuriyet Üniversitesi Fen Fakültesi
Fen Bilimleri Dergisi (CFD), Cilt:36, No: 4 Ozel Sayı (2015)
ISSN: 1300-1949

Cumhuriyet University Faculty of Science
Science Journal (CSJ), Vol. 36, No: 4 Special Issue (2015)
ISSN: 1300-1949

Integrating RFM and Classification for Response Modeling Based on Customer Lifetime Value

Atefeh RABIEI[1], Hamid RASTEGARI[1*]

[1]*Department of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Iran*

Received: 22.03.2015; Accepted: 29.05.2015

Abstract. One of the most important challenges in direct marketing is finding differences between customers and identifies profitability of each customer for target marketing. Response modeling is an useful technique for this issue that predicts customer᾿s response to a campaign. Accuracy of response model is very important due to high cost and time of marketing process. Due to this, this paper has provided a framework for building an accurate model based on weighted RFM analysis and calculating customer lifetime value (CLV) for each segment of customers, then uses CLV as one of predictor features with demographical features in C5 algorithm. The experimental results show by compacting transactional behaviors of customers in CLV value and using it with demographical features concurrently as predictors of classification algorithm is an efficient method for building response model that is much more accurate than those methods that did not used demographical features and CLV for prediction.

Keywords: Customer lifetime value, RFM analysis, Response modeling, Data mining, classification

1. INTODUCTION

In today᾿s business, firms need to develop new strategies to improve customer acquisition and retention. In this regard, customer relationship management is a very important strategy .The main objective of CRM is to make long-lasting and profitable relationships with customers [1].

On the other hand, there are large databases that contain demographical and transactional information about customers. Data mining techniques are widely used information technology for extracting marketing knowledge and further supporting marketing decision from them [2].

In many marketing branches, including pricing, new product development and advertising diverse techniques and models have been proposed over last five decade[3]. The main task is identifying more profitable and loyal customers from marketing knowledge and makes relationships with them.

Response modeling is one of the most popular models for identifying potential customers for target marketing [3]. Response models predict customer᾿s response probability to a new campaign or product offer. This model is a classification model that classifies customers into two classes: respondents and non-respondents [4]. Various data mining techniques including statistical analysis and machine learning algorithms can be useful for building response model [5-9]. RFM is one of the most famous models that have been used for analyzing customer data. RFM relies on three customer transactional behavior (how long since the last purchase by customer, how often the customer purchase, how match the customer has bought). Hughes proposed a method for RFM scoring that based on RFM data, customers have been divided into five groups. Different marketing strategies could then be applied for different groups [10]. Stone suggested that different weights should be assigned to RFM variables depending on the characteristics of the industry [11].

*Corresponding author. *Email: Rastegari@iaun.ac.ir*

Special Issue: Technological Advances of Engineering Sciences

Another useful concept for analyzing customer behavior is CLV. CLV define as: "the present value of future profit stream expected over a given time horizon of transactional behavior of customer"[12, 13]. This concept can be applied for customer segmentation.

In previous studies that used RFM model for response modeling, demographical features of customers have been ignored [14], On the other hand response models that haven't used RFM model and are based on classification algorithms such as decision trees, neural networks and SVM although are more accurate there are some problems with them such as production of additional rules, feature plurality and feature selection method, relationship between features, increasing depth for decision tree and problem for applied on new data for neural network and complexity and non transparency of SVM [15, 16]. Due to this; current study provides a framework for building response model by segmenting customers by RFM features and calculating CLV values of different segments and finally builds response model with CLV value and demographical features of customer as predictor variables and applies them as inputs of classification algorithm. For classification task decision tree algorithms has been chosen because they are powerful and popular tools for classification and prediction and have the abilities for generating rules that can be translated into natural language in contrast to other model such as neural network [17]. The purpose of this study is building a response model that have high accuracy and readability for marketing decision such that increasing profits and decreasing costs of direct marketing. The reminder of this study is organized as follow. Section 2 describes the framework and methods. In Section3 experimental evaluations and results have been described. Finally section 5 draws conclusion and summarizes the contributions of this work.

2. FRAMEWORK OF BUILDING RESPONSE MODEL

Current study provides a framework for building response model by segmenting customers in homogenous segments and calculating CLV value for each segment and finally building model using CLV and demographical features of customers as predictor variables for decision tree algorithm. Figure (1) represents the framework.

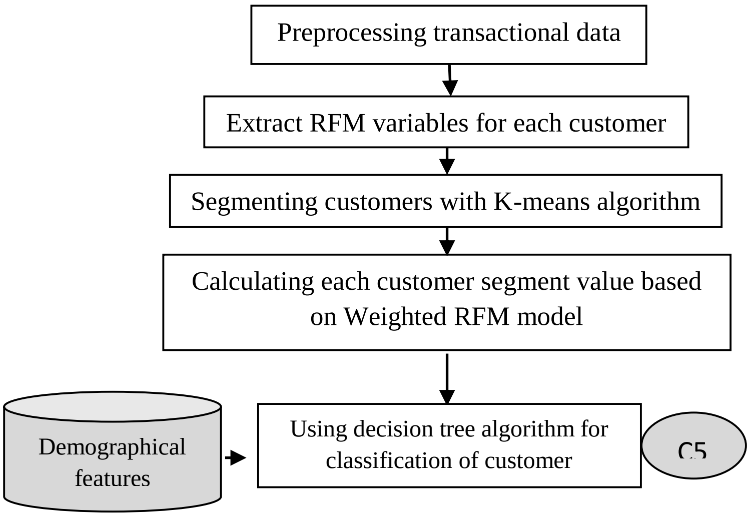
Preprocessing transactional data
Extract RFM variables for each customer
Segmenting customers with K-means algorithm
Calculating each customer segment value based on Weighted RFM model
Demographical features
Using decision tree algorithm for classification of customer
C5

Figure 1. Framework of building proposed response model.

This study as it shown in Figure (1) is summarized in five steps:

Step 1: In the first step data was collected from UCI KDD [18]. This data was in form of three dataset: transactional, demographical and campaign datasets. Transactional dataset including

69215 transactions of credit cards of 12589 customers from 2001 to 2007-06-06 (date of performing campaign) which was conclude card ID, date and amount of transaction. Demographical dataset including demographical features conclude number of children, marital status and average level of income and length of service. Campaign dataset conclude card ID of customers and their respond to the campaign. All records with missing values were deleted from datasets.

Step 2: RFM parameters were extracted for each customer. For extracting R, number of days from last transaction to date of performing campaign (2007-06-06) was calculated. M was obtained from sum of amounts of each customer transactions and finally F was calculated by counting number of transaction for each customer.

Step 3: In this step customer segmentation was done with K-means algorithm. For determining inputs of K-means algorithm we perform correlation analysis for RFM variables. Result of analysis is shown in table 1.

Table 1. RFM variables Correlation.

	R	**F**	**M**
R	1		
F	-0.207	1	
M	-0.245	0.473	1

In previous studies that used RFM model, clustering was done with three variable R, F, M [5], [19,[20]. But in this study because of strong correlation between M and F, We calculated a new variable V by
Eq. 1

$$V = \frac{M}{F} \qquad (1)$$

Where M is monetary value and F is frequency value for each customer. After computing V for each customer clustering with K-means algorithm based on R and V as inputs was done. The k parameter was set to 8, since eight (2*2*2) possible combinations of each RFM variables according to less or greater than overall average of RFM variables.

Step 4: For calculating CLV value of each segment normalized form of RFM parameters were needed. Since F and M positively influenced CLV and R negatively impacts CLV we use profit and cost form according to Eq. 2 and Eq. 3 respectively for normalizing RFM values.

$$x' = (x - x^s)/(x^L - x^s) \qquad (2)$$

$$x' = (x^L - x)/(x^L - x^s) \qquad (3)$$

Where x' normalized value and x is original value and x^s and x^L are smallest and greatest value of parameters respectively.
After normalizing RFM parameters we calculated CLV score of each cluster based on weighted RFM model as follow in Eq. 4:

$$clv^j = w_R c_R^j + w_F c_F^j + w_M c_M^j \qquad (4)$$

Where w_R, w_F, w_M are the relative importance of the RFM variables. The weighting (relative importance) of each RFM variable was evaluated using AHP. Data were gathered by interviewing the evaluators that are marketing managers and decision makers. Interviews were conducted using a questionnaire (Table 2) and the answers were expressed in the form of a pair wise comparison matrix (Table 3). According to the assessments obtained by AHP method based on expert people idea, the relative weights of the RFM variables are as follow: w_R=0.731, w_F=0.188 and w_M=0.081[19].

Table 2. AHP questionnaire for RFM.

Criteria	Comparative importance									Criteria
	9:1	7:1	5:1	3:1	1:1	3:1	5:1	7:1	9:1	
Recency	9	7	5	3	1	3	5	7	9	Frequency
Recency	9	7	5	3	1	3	5	7	9	Monetary
Frequency	9	7	5	3	1	3	5	7	9	Monetary

Also c_R^j , c_F^j , c_M^j are averaging normalized RFM value of customers in cluster j.

Table 3. Example of RFM pair wise comparison matrix.

	R	F	M
R	1	5	7
F	1/5	1	3
M	1/7	1/3	1

In table4 the normalized averaged R, F and M values for each cluster and CLV of them are shown.

Step 5: In this step we want to build response model based on CLV and demographical features. For classification task C5 decision tree algorithm was chosen. CLV and five demographical features of customer have been used as predictor inputs of C5 algorithms.

Table 4. Calculating CLV value for 8 clusters.

Cluster	Averaged normalized R	Averaged normalized F	Averaged normalized M	CLV
1	0.962	0.255	0.079	0.757
2	0.405	0.008	0.017	0.311
3	0.202	0.024	0.031	0.155
4	0.894	0.028	0.211	0.676
5	0.766	0.105	0.065	0.585
6	0.447	0.169	0.048	0.363
7	0.624	0.089	0.070	0.479
8	0.884	0.091	0.086	0.670

3. RESULTS

Experimental setup

For evaluation our response model we compared it with two other methods: The WRFM clustering based on C5 algorithm and non-demographical RFM based on C5 algorithm. In the first method all steps of building response model are similar to proposed model but for clustering used three variable R, F and M as inputs of k-means algorithm. In second method transactions are summarized in three RFM variables but demographical features have not been used as predictors of C5 algorithm. The dataset was divided into a 66% training set and a 34% testing set.

In this study, the two non-overlapping classes are buyers and non-buyers of a particular investment product of a bank. The four count, which constitute a confusion matrix as seen in Table 5 for binary classification are: the number of correctly recognized positive class examples (true positives), the number of correctly recognized examples that belong to the negative class (true negatives), and examples that either were incorrectly assigned to the positive class (false positives) or that were not recognized as positive class examples (false negatives).

Most often performance measures based on the values of the confusion matrix that are used to evaluate the performance of a classification model include precision, accuracy, recall and lift. The performance criteria differ in their assumptions about the costs of misclassification errors and the types of errors that are used to measure the performance of classifier[21].

Precision is number of correctly classified positive examples divided by the number of examples labeled by model as positive:

$$\text{Precision}=TP/\,(TP+FP) \qquad (5)$$

Accuracy measures the overall effectiveness of a classifier:

$$\text{Accuracy}=(TP+TN)\,/\,(TP+FN+FP+TN) \qquad (6)$$

Recall is the number of correctly classified positive examples divided by the number of positive examples in data. This rate determines the effectiveness of a classifier to identify positive labels:

$$\text{Recall}=TP/\,(TP+FN) \qquad (7)$$

An F-metric could be used to balance the tradeoff between precision and recall and given by Eq.8.

$$F = \frac{2 \times precision \times recall}{precision + recall} \qquad (8)$$

Each metrics computed for proposed method and two other methods.

Table 5. The confusion matrix.

		Predicted class	
		Positive	**Negative**
Actual class	**Positive**	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

Experimental results

All three methods were conducted using the IBM SPSS Modeler 14.2. We computed four metrics for three methods and compared them. Results of experiments are shown in table 6.

By comparing proposed method with second method that difference with our method just in step of clustering customers we conclude that our clustering method based on new variable V and R was efficient. That led to increasing precision and recall and F-measure and finally accuracy. Moreover By comparing proposed method with first method that didn't used demographical features as inputs of C5 algorithms and didn't clustering customers we conclude that using demographical feature increase the precision and accuracy of response model.

In Figure (2) and Figure (3) the results of this study are shown as Gain and Lift charts respectively.

Table 6. Evaluation metrics for three models.

Method	Precision (%)	Recall (%)	F-measurement (%)	Accuracy (%)
Proposed method	97	80	87	92.83
RFM-based C5 (1)	85	88	86	91.02
WRFM clustering (2)	86	57	68	82.2

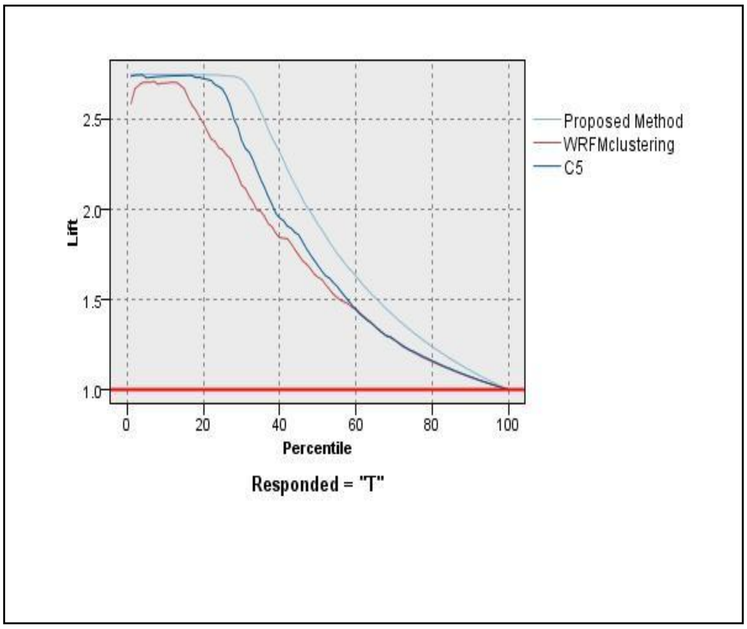

Figure 2. Gain chart of three methods.

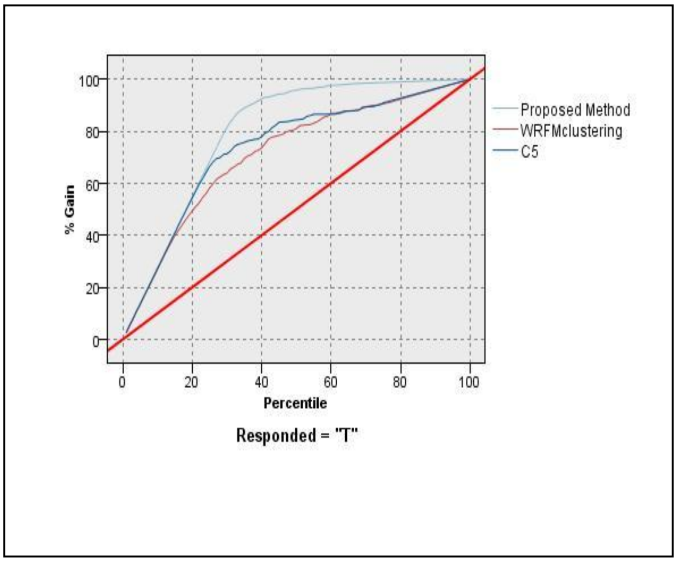

Figure 3. Lift chart of three methods.

DISCUSSION

Finding differences between customers to make marketing decisions more profitability is very important issue for firms. Using customer segmentation based on transactional history of customer and predicting their response could be useful in this area. RFM analyze is an useful model for summarizing transactional data in three variables. This study proposed a framework for building response model based on RFM parameters and clustering customers in exact clusters for calculating CLV of each cluster and uses this profitable variable with demographical features of customers as predictors of classification algorithm.

Results of this study shows using V and R for clustering customers is more efficient than clustering based on R, F and M. Moreover considering demographical features has a positive impact on efficiency of response models.

REFERENCES

[1] Kim, Y., Toward a successful CRM: variable selection, sampling, and ensemble. Decision Support Systems, 2006. 41(2): p. 542-553.

[2] Huan-Ming, C. and S. Chia-Cheng. A study on the applications of data mining techniques to enhance customer lifetime value &#x2014; based on the department store industry. in Machine Learning and Cybernetics, 2008 International Conference on. 2008.

[3] Berger, P. and T. Magliozzi, The effect of sample size and proportion of buyers in the sample on the performance of list segmentation equations generated by regression analysis. Journal of Direct Marketing, 1992. 6(1): p. 13-22.

[4] Chen, W.-C., C.-C. Hsu, and J.-N. Hsu, Optimal selection of potential customer range through the union sequential pattern by using a response model. Expert systems with applications, 2011. 38(6): p. 7451-7461.

[5] Cheng, C.-H. and Y.-S. Chen, Classifying the segmentation of customer value via RFM model and RS theory. Expert systems with applications, 2009. 36(3): p. 4176-4184.

[6] Wong, K.W., et al., Mining customer value: From association rules to direct marketing. Data Mining and Knowledge Discovery, 2005. 11(1): p. 57-79.

[7] Cui, G., M.L. Wong, and G. Zhang, Bayesian variable selection for binary response models and direct marketing forecasting. Expert Systems with Applications, 2010. 37(12): p. 7656-7662.

[8] F.Burstein and C.Holsapple, Systems for supporting marketing decisions, in Handbook on Decision Support Systems. 2008, Springer. p. 395-418.

[9] Trenkler, G., Regression analysis, theory, methods and applications: Sen, A. and Srivastava, M. Springer, Berlin (1990), 347 pp, ISBN3-540-97211-0,DM 88. Computational Statistics & Data Analysis, 1992. 13(1): p. 109-110.

[10] Hughes, A., Strategic database marketing. 1994, Chicago: Probus Publishing Company.

[11] Stone, B. and R. Jacobs, Successful direct marketing methods. 2008: McGraw Hill Professional.

[12] Fader, P.S., B.G.S. Hardie, and K.L. Lee, RFM and CLV: Using Iso-Value Curves for Customer Base Analysis. Journal of Marketing Research, 2005. 42(4): p. 415-430.

[13] Khajvand, M., et al., Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. Procedia Computer Science, 2011. 3: p. 57-63.

[14] Olson, D.L. and B.K. Chae, Direct marketing decision support through predictive customer response modeling. Decision Support Systems, 2012. 54(1): p. 443-451.

[15] D’Haen, J., D. Van den Poel, and D. Thorleuchter, Predicting customer profitability during acquisition: Finding the optimal combination of data source and data mining technique. Expert Systems with Applications, 2013. 40(6): p. 2007-2012.

[16] Sing'oei, L. and J. Wang, Data Mining Framework for Direct Marketing: A Case Study of Bank Marketing. International Journal of Computer Science Issues (IJCSI), 2013. 10(2).
[17] van den Berg, B. and T. Breur, Merits of interactive decision tree building: Part 1. J Target Meas Anal Mark, 2007. 15(3): p. 137-145.
[18] Hettich. S, B.S., 1999, Irvine, CA: University of California, Department of Information and Computer Science: The UCI KDD Archive [http://kdd.ics.uci.edu].
[19] Liu, D.-R. and Y.-Y. Shih, Integrating AHP and data mining for product recommendation based on customer lifetime value. Information & Management, 2005. 42(3): p. 387-400.
[20] Hsieh, N.-C., An integrated data mining and behavioral scoring model for analyzing bank customers. Expert systems with applications, 2004. 27(4): p. 623-633.
[21] Duman, E., Y. Ekinci, and A. Tanrıverdi, Comparing alternative classifiers for database marketing: The case of imbalanced datasets. Expert Systems with Applications, 2012. 39(1): p. 48-53.