2011 International Conference on Innovation, Management and Service
IPEDR vol.14(2011) © (2011) IACSIT Press, Singapore

A Case Study of Using Classification and Regression Tree and LRFM Model in A Pediatric Dental Clinic

Shih-Yen Lin [1], Jo-Ting Wei [2], Chih-Chien Weng [3] and Hsin-Hung Wu [3+]

[1] Department of Leisure Studies and Tourism Management, National Chi Nan University, Nantou, Taiwan, R.O.C.

[2] Department of Business Management, National Sun Yat-Sen University, Kaohsiung, Taiwan, R.O.C.

[3] Department of Business Administration, National Changhua University of Education, Changhua, Taiwan, R.O.C.

Abstract. A case study in a pediatric dental clinic was presented. The data were transformed into LRFM (Length, Recency, Frequency, and Monetary) format with fixed M covered by National Health Insurance program in Taiwan, where the data were categorized into 1 to 5 for L, R, and F variables. Later, gender was classified into two types, and age was grouped into four categories. The target in this study was frequency, while L, R, gender, and age were the input variables when classification and regression tree was performed. The overall accuracy is about 60% but the prediction accuracies for lost patients and very important patients were very effective. Therefore, the dental clinic can pay much attention to those who might be considered as very important patients.

Keywords: LRFM model, pediatric dental clinic, classification and regression tree

1. Introduction

The medical care industry in Taiwan is very competitive when the National Health Insurance (NHI) program, which is a mandatory, single-payer social health insurance system, was launched by the Government in Taiwan since March 1995 (Shieh et al., 2010). Under such system, each citizen should have equal access to health care services, health care of acceptable quality, comprehensive benefits, and convenient access to treatment with low premiums and health care expenditures. Thus, citizens are able to freely choose health care providers and medical institutions with low reimbursement (Shieh et al., 2010).

In Taiwan's dental services, dental care is covered as part of the benefit package in NHI program with this cost-containment mechanism and global budgeting system (Lee and Shih, 2009). A dentist's income is limited when serving a dental care covered by NHI program since NHI program will cover most of cost except for the co-payment and registration fee per visit (Lee and Shih, 2009). This scenario shows that it is critically important for Taiwanese dentists to identify profitable customers and then to retain those important customers as many as possible.

In this case study, the transactions data in a pediatric dental clinic in Taiwan were used. The data were first transformed into LRFM (Length, Recency, Frequency, and Monetary) format as well as gender and age were classified into two and four clusters, respectively, and then analyzed by classification and regression tree (CART) to classify each patient into appropriate group with rules and to predict the frequency of a particular patient based on these transactions data.

2. Literature Review

+ Corresponding author. Tel.: +886-4-723-2105 ext. 7412; fax: +886-4-721-1292.
E-mail address: hhwu@cc.ncue.edu.tw.

2.1. LRFM model

RFM (Recency, Frequency, and Monetary) model is a behavior-based model to analyze the behavior of a customer and then make predictions based on the behavior in the database (Hughes, 1996; Yeh et al., 2009). Specifically, recency represents the length of a time period since the last purchase; frequency denotes the number of purchase within a specified time period; and monetary means the amount of money spent in this specified time period (Wang, 2010). The traditional approach to quantify customer behavior based on RFM model is as follows. First, sort the database by each dimension of RFM and then divide the customer list into five equal segments. The top 20% segment is coded as 5. The next 20% segment is assigned as a code 4 and so forth (Hughes, 1996; Kahan, 1998; Tsai and Chiu, 2004). Therefore, all customers can be presented by 555, 554, 553, …, 111 with possible 125 (5×5×5) RFM cells.

Chang and Tsay (2004) added length into original RFM model to become LRFM (Length, Recency, Frequency, and Monetary) model since length measures the time period between the first visit and the last visit of a particular customer. Reinartz amd Kumar (2000) stated that RFM model cannot segment which customers have long-term or short-term relationship with the company. With the introduction of length, the relationship between the customer and the company can be determined from numerical viewpoint. In order to divide the customer into five equal segments, prioritize the length values by descending order and then select the top 20% values as the number of 5. The next 20% values can be assigned as a value of 4 and so forth. Therefore, the numerical length values can be transformed into 5 to 1.

RFM models have been widely applied to a wide variety of areas, such as nonprofits, financial organizations (Hsieh, 2004; Sohrabi and Khanlari, 2007), government agencies (King, 2007), on-line industry (Li et al., 2010), telecommunication industry (Li et al., 2008), travel industry (Ha and Park, 1998; Lumsden et al., 2008) and marketing industry (Spring et al., 1999; Jonker et al., 2006). In addition, RFM model can be used to segment customers, calculate customer value and customer lifetime value (CLV), observe customer behavior, estimate the response probability for each offer type and evaluate on-line reviewers (Wei et al., 2010).

2.2. Classification and regression tree

Solomon et al. (2006) summarized that the decision tree model is a powerful and popular tool to classify and predict data patterns since rules are generated more straightforward and relatively easy to be interpreted. In addition, from business management viewpoint, decision trees can generate a set of rules from the classified data set which can be applied to the unclassified data set and predict the outcomes by aiding the future decision-making process (Lee and Siau, 2001). Classification and regression tree (CART) is one of the most commonly seen algorithms in decision tree models (Solomon et al., 2006; Rokach et al., 2007).

Classification and regression tree is one of the decision tree algorithms for classification by constructing a flowchart-like structure where each internal node represents a test on an attribute, each branch denotes an outcome of the test, and each external node means a class prediction (Han and Kamber, 2007; Wu and Guo, 2011). Razi and Athappilly (2005) and Witten and Frank (2005) depicted that the characteristic of CART is to use a set of "if-then" conditions to perform predictions or classification of cases such that CART is very suitable to either large problems or small data set with both continuous and categorical variables. Moreover, the attribute that is not appeared in the tree is assumed to be irrelevant in the analysis because CART ranks the attributes by giving their respective weights (Loh, 2011). Therefore, the set of attributes appearing in the tree forms the reduced subset of attributes.

The major advantages of CART are below (Han and Kamber, 2007; Hill and Lewicki, 2007). First, the interpretation of the results in a flowchart-like tree is very simple to explain why observations are classified into a particular manner. Second, there is no implicit assumption to be made when the underlying relationships between the predictor variables and the dependent variables are to be linear or follow some specific non-linear link function because CART possesses both non-parametric and non-linear properties. Finally, CART is very suitable for data mining because little knowledge or any coherent set of theories or predictions regarding which variables are related and how are to be known in advance (Delen, 2009).

3. Research Method

This children's dental clinic begins its operation since September 17, 1995. By definition, the patients must be less than 18 years old to be classified as children. When the patients become 18 years old, they are no longer to be considered as children. In this study, it is necessary to consider the "age" in this children's dental clinic in order to identify and analyze profitable patients. This study collects the data set with 2,061 patients who visited this clinic from January 1, 2009 to July 15, 2010 with complete needed information such as gender and birth date.

The profile for each patient is composed of the membership number, gender, birth date, the days from the first visit date to the last visit date, the last visit date, and visit frequency. Monetary value for each patient is excluded in the profile since the majority of the costs were covered by NHI program and thus is proposed to be fixed. Gender equals 1 if the patient is the male and 0 if the patient is the female. The birth date was classified into four age groups. The age of 5 and below is coded as 4, the age of 6-9 is coded as 3, the age of 10-13 is coded as 2 and the age of 14-18 is coded as 1. The definition of LRFM model is depicted in Table 1.

Table 1: The definition of LRFM variables.

Variables	Definitions
Length (L)	refers to the number of days from the first visit date to the last visit date since September 17, 1995
Recency (R)	refers to the number of days since the last visit from January 1, 2009 to July 15, 2010
Frequency (F)	refers to the number of visit in a specified time period (January 1, 2009 to July 15, 2010)
Monetary (M)	refers to the co-payment and registration fee per visit, which is proposed to be fixed

In order to transform the LRFM data into categorical data for CART, L, R, and F values were divided into five equal segments with each segment containing 20% of the entire data for each variable. The specific information is summarized in Table 2. Since M is proposed to be fixed, the target for CART was chosen to be the frequency. The higher value the frequency, the much money the dental clinic would make under NHI program. The input variables of CART included gender, age, L, and R. Clementine 12.0 was used to perform CART with "expert" mode by using the default values for all parameters. The maximum surrogates and minimum change in impurity were set to 5 and 0.0001, respectively. The impurity measure for categorical targets was Gini. The stopping criteria were based on the percentage with minimum records in parents branch (%) of two and minimum records in child branch (%) of one. Furthermore, training partition size and testing partition size were set to 55 and 45 percents, respectively.

Table 2: The transformed data with codes for classification and regression tree analysis.

Code	Gender	Age	L (days)	R (days)	F (times)
1	Male	14-18	1-1079	1-185	0
2	Female	10-13	1080-2159	186-370	1
3		6-9	2160-3238	371-555	2
4		1-5	3239-4318	556-740	3-4
5			4319-5397	741-926	5 and above

4. Results

The tree depth was six, and the variable importance values of gender, age, L, and R were 0.006, 0.034, 0.003, and 0.957, respectively. The prediction accuracy of the model, stated in Table 3, was well over 58% for both training and testing data sets. To further examine the coincidence matrix as shown in Table 4, where row shows actual category while column makes the prediction of the category, the second, third, and fourth categories of frequency were less than 50%. Only the first and fifth categories of frequency could be accurately predicted for both training and testing data sets.

Table 3: Prediction accuracy of the model.

Partition	Training	Testing
Correct	665 (60.62%)	561 (58.2%)
Wrong	432 (39.38%)	403 (41.8%)

Total	1,097	964

Table 4: Coincidence matrix for both training and testing data sets.

Training	1	2	3	4	5	Prediction Accuracy	Testing	1	2	3	4	5	Prediction Accuracy
1	338	5	0	0	0	98.54%	1	303	5	0	0	0	98.38%
2	0	111	11	35	71	48.68%	2	0	91	12	52	51	44.17%
3	0	24	18	46	60	12.16%	3	0	29	8	45	56	5.80%
4	0	23	12	77	97	36.84%	4	0	15	7	63	87	36.63%
5	0	9	3	36	131	73.18%	5	0	8	3	33	96	68.57%

To further discuss the managerial implications from Table 4, the second, third, and fourth categories of frequency tended to be overestimated. In contrast, this model could predict the first and fifth categories relatively better. By observing ten rules generated by CART, only one rule was for the first category of frequency. The rule said that the predicted frequency was to be the first category when the recency code is in 1, 2, and 3. This rule also indicated that these patients have not been to the dental clinic for more than one year, i.e., 926-555 = 371 days. Thus, when the recency value becomes far away from the current value as time goes by, the information might indicate that the patient might be lost at least 98 out of 100 times. On the other hand, the fifth category of frequency shows that the patients have been to the dental clinic five or more times in this specified time period. The more the patients come to the dental service, the much money the dental clinic would make. Three rules were related to fifth category, namely, R in 5, age in 3 and 4, and L in 2 and 3; R in 5, age in 3 and 4, L in 1, and gender in 2; and R in 5, age in 3, L in 1, and gender in 1. These patients can be viewed as important patients for this dental clinic.

5. Results

This study uses the patients' transactions data from a pediatric dental clinic. First, the data were categorized into 1 to 5 for L, R, and F variables, while M is proposed to be fixed covered by NHI program. The gender was classified into two types, whereas age has four groups. Classification and regression tree was performed for both classification and prediction. The overall accuracy is about 60% but the prediction accuracies of first and fifth categories were much better than those of the other three categories. That is, this model could be applied effectively to predict the lost patients as well as the very important patients for the dental clinic. Therefore, the dental clinic can pay much attention to those who might be considered as very important patients and try to reallocate resources to those who might be the lost patients.

6. Acknowledgements

This study was partially supported by the National Science Council in Taiwan with the grant number of NSC 99-2221-E-018-012-MY2.

7. References

[1] H. H. Chang, and S. F. Tsay. Integrating of SOM and K-Mean in data mining clustering: an empirical study of CRM and profitability evaluation. J. Inf. Manage. 2004, 11 (4): 161-203.

[2] D. Delen. Analysis of cancer data: a data mining approach. Expert Syst. 2009, 26 (1): 100-112.

[3] S. H. Ha, and S. C. Park. Application of data mining tools to hotel data mart on the Intranet for database marketing. Expert Syst. Appl. 1998, 15: 1-31.

[4] J. Han, and M. Kamber. Data Mining: Concepts and Techniques, Second Edition. Morgan Kaufmann Publishers, 2007.

[5] T. Hill, and P. Lewicki. STATISTICS Methods and Applications. StatSoft, 2007.

[6] N. C. Hsieh. An integrated data mining and behavioral acoring model for analyzing bank customers. Expert Syt. Appl. 2004, 27: 623-633.

[7] A. M. Hughes. Boosting reponse with RFM. Market. Tools. 1996, 5: 4-10.

[8] J. J. Jonker, N. Piersma, and R. Potharst. A decision support system for direct mailing decisions. Decis. Support Syst. 2006, 42: 915-925.

[9] R. Kahan. Using database marketing techniques to enhance your one-to-one marketing initiatives. J. Consumer Market. 1998, 15 (5): 491-493.

[10] S. F. King. Citiznes as customers: Exploring the future of CRM in UK local government. Gov. Inform. Q. 2007, 24: 47-63.

[11] S. J. Lee, and K. Siau. A review of data mining techniques. Ind. Manage. Data Syst. 2001, 101 (1): 41-46.

[12] W. I. Lee, and B. Y. Shih. Application of neural networks to recognize profitable customers for dental services marketing – a case of dental clinics in Taiwan. Expert Syst. Appl. 2009, 36: 199-208.

[13] S. T. Li, L. Y. Shue, and S. F. Lee. Business intelligence approach to supportng strategy-making of ISP service management. Expert Syst. Appl. 2008, 35: 739-754.

[14] Y. M. Li, C. H. Lin, and C. Y. Lai. Identifying influential reviewers for word-of-mouth marketing. Electron. Commerce Res. Applications. 2010, 9: 294-304.

[15] W. Y. Loh. Classification and regression trees, WIREs Data Min. Knowl. Disc. 2011, 1 (1): 14-23.

[16] S. A. Lumsden, S. Beldona, and A. M. Morison. Customer value in an all-inclusive travel vacation club: an application of the RFM framework. J. Hospit. Leisure Market. 2008, 16 (3): 270-285.

[17] M. A. Razi and K. Athappilly. A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. Expert Syst. Appl. 2005, 29 (1): 65-74.

[18] W. J. Reinartz, and V. Kumar. On the profitability of long-life customers in a noncontractual setting: an empirical investigation and implications for marketing. J. Market. 2000, 64 (4): 17-35.

[19] L. Rokach, and O. Maimon. Data Mining with Decision Trees Theory and Applications. World Scientific Publishing, 2007.

[20] J. -I Shieh, H. -H. Wu, and K. -K. Huang. A DEMATEL method in identifying key success factors of hospital service quality. Knowl-Based Syst. 2010, 23 (3): 277-282.

[21] B. Sohrabi, and A. Khanlari. Customer lifetime value (CLV) measurment based on RFM model. Iranian Acc. Aud. Rev. 2007, 14 (47): 7-20.

[22] S. Solomon, H. Nguyen, J. Liebowitz, and W. Agresti. Using data mining to improve traffic safety programs. Ind. Manage. Data Syst. 2006, 106 (5): 621-643.

[23] P. Spring, P. S. H. Leeflang, and T. Wansbeek. The combination strategy to ptimal target selection and offer segmentation in direct mail. J. Market. Focused Manage. 1999, 4: 187-203.

[24] C. Y. Tsai, and C. C. Chiu. A purchase-based market segmentation metholology. Expert Syst. Appl. 2004, 27: 265-276.

[25] C. H. Wang. Apply robust segmentation to the service industry using kernel induced fuzzy clustering techniques. Expert Syst. Appl. 2010, 37: 8395-8400.

[26] J. -T. Wei, S. -Y Lin, and H. -H. Wu. The review of the application of RFM model Afr. J. Bus. Manage. 2010, 4 (19): 4199-4206.

[27] I. H. Witten, and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. Morgan Kaufmann Publishers, 2005.

[28] S. Wu, and J. Guo. A data mining analysis of the Parkinson's disease. iBus. 2011, 3: 71-75.

[29] I. C. Yeh, K. J. Yang, and T. M. Ting. Knowledge discovery on RFM model using Bernoulli sequence. Expert Syst. Appl. 2009, 36: 5866-5871.