Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2000 Proceedings

Americas Conference on Information Systems (AMCIS)

2000

A Data Mining System for Managing Customer Relationship

Edward H. Ip
University of Southern California, eip@bus.usc.edu

Katsutoshi Yada
Osaka Industrial University

Yukinobu Hamuro
Osaka Industrial University

Naoki Katoh
Kyoto University

A Data Mining System for Managing Customer Relationship

Edward H. Ip, University of Southern California, eip@bus.usc.edu
Katsutoshi Yada, Osaka Industrial University, Japan
Yukinobu Hamuro, Osaka Industrial University, Japan.
Naoki Katoh, Kyoto University, Japan.

Abstract

This paper presents a data mining study that aims to identify potential high-value visitors for a drugstore chain in Japan. Our purpose is to provide timely decision support to the marketing and service departments for managing customer relationship. The conceptualization of customer value is discussed and is differentiated from a more commonly used construct, customer loyalty. We briefly describe the data mining system that supports the study. Our result show two supervised learning methods are comparable in terms of predictive accuracy.

Introduction

The global competitive landscape of the retail business has changed radically during the past several years. The availability of detailed customer data and advances in technology for warehousing and mining data (e.g. Spangler, May, and Vargas 1999, Sung, Chang, and Lee 1999) enable companies to better understand and service their customers. The Internet, on the other hand, accelerates this trend as online transaction information and visitor click streams data can be readily captured, processed and analyzed, thus creating a new frontier for new business processes and models in the retail industry. While also raising consumer concerns over privacy issues, early adopters of information technology had successfully gained competitive advantage by exploiting the tremendous amount of customer data that can be made available, whether in cyberspace or the physical market place. For example, recognizing that "credit cards aren't banking, they're information" (Fishman 1999), Capital One revolutionized the credit card industry by using data mining techniques such as decision tree to effectively manage their customer relationship life cycle (Berry and Linoff 2000), and as a result enjoyed a high success rate of keeping their most profitable customers. In the retail business, the issue of leveraging information for customer retention and loyalty management is becoming increasingly important. Numerous studies have shown that the retaining the right customers is a determinant for long term profit (Reichheld 1993).

This article presents a study on a data mining system that is designed to manage customer relationship, with a focus on managing loyal and highly profitable customers. The data mining system was deployed by a nationwide drugstore chain in Japan. The drugstore chain, Pharma, has an annual revenue of 70 billion yen. There are 1,230 Pharma membership retail stores across Japan. The structure of the chain store is similar to that of a franchise in the American system, but has substantial differences. For example, each store can use its own name. However, all stores operate under a centralized information system. This system handles daily transaction data from its 2.3 million customers, and monitors inventory and processes replenishing orders for its membership stores. Pharma had been systemically collecting detailed transaction data and customer information since the early 1990's. By leveraging its information asset, Pharma generates, besides its membership fee income, an alternative source of revenue by acting as an information broker. For example, it provides research reports on consumer taste and behavior to manufacturers, and conducts marketing research for manufacturers using tools such as customer checkout interview that is supported by its sophisticated information system.

Although online retail businesses are still in infancy in Japan at this time point, Pharma faces a high level of competition from other retailers in the physical market place. During the decade-long recession period of the 90's, Japanese consumers were becoming more price-sensitive and as a result a lot of drugstore chains suffered from decreasing profits and revenues. Pharma has to continue to strive in retaining its high-value customers from defecting to other retailers that compete on price. The data mining activities studied in this article concerns the early identification of potential high-value new visitors to the store. The successful identification of potential high-value new customers early on is important to Pharma because the company can use this information to establish a close relationship to this selected group of customers, thus building a tighter bond and reduces the chance that they will leave. Unlike its American counterparts, Japanese drugstore chains in general enjoy a closer tie with their customers. For example, clerks at Pharma cash registers could have substantial interactions with customers at the point of checkout. They sometimes provide free medicine samples, ask for feedback on the use of free samples, and briefly interview customers using online market research questionnaire form. By closely monitoring the purchasing behavior of relatively new visitors to the store and applying data mining tools on pertinent data, the company can provide decision support to clerks and the marketing department for relationship building. For example, sales campaign information, customized coupons, and free samples can be directly mailed to the targeted group.

In the next section, we provide a discussion of how we conceptualize customer value, as differentiated from customer loyalty. Then we describe the data mining system and report the results of directed data mining activities that aim to manage value in the customer relationship lifecycle.

Conceptualizing customer value

Customer value is closely related to customer loyalty. There is nonetheless a subtle difference between the two constructs. The reason is partly cultural. For a drugstore chain such as the Pharma, a loyal customer may not necessarily generate positive profit over a sustained period of time. Empirical evidence from the Pharma showed that a non ignorable proportion of their loyal customers are actually generating negative profits. Japanese housewives are well known to be highly selective in their buying habits. Because they are often full-time housekeepers and are purse string keepers as well, they can afford time to research and wait for the best buy. For example, it is a tradition for Japanese retail stores to hold sales on certain week days. Bargain hunters often visit the store only on these store promotion days and target products that are competitively priced, sometimes under cost. This sector of the customer population, even if it continues its patronage to the store, does not necessarily contribute to the long term profitability of the company. Indeed, Pharma estimated about 7% of the population were consistent bargain hunters. They generated -1.31% in total profit share. Traditional measures of customer loyalty such as customer retention rate or share of purchase (Jones and Sasser 1995) are therefore not necessarily applicable in pinpointing the "right" customer. It should be pointed out, however, that over generalizing about the "right" customer may fail to take into account that a customer who is of low value for one company may be valuable for another (Reichheld, 1990). But we shall not pursue the issue in this paper.

Based upon discussions with the management at Pharma, we proposed a two-dimensional measure for customer value. The two dimensions are customer profitability per visit and frequency of visit. In this short paper we will not developed a theory for the construct of customer value such that the construct can be fully conceptualized, operationalized and evaluated. We do, however, provide an operational measurement of customer value: customers who have high observed profitability per visit and high observed frequency of visits will be classified as high-value customers in the customer database. Note that customers who repeatedly made purchases but had low or negative profitability, and those who made a large purchase and seldom came back are excluded according to this definition. Both types of customers do exist.

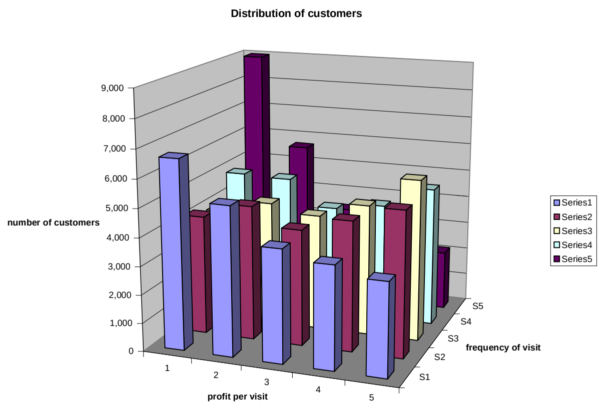

Figure 1. Distribution of customers with respect to profit per visit and frequency of visits

Figure 1 shows the distribution of a sample of 114,069 customers on the two dimensions. The variables profitability per visit and frequency of visit are computed using data from the year of 1998. The two variables are each categorized into 5 classes (Table 1). The criteria for forming these classes are based upon consultation with marketing personnel from Pharma.

Figure 1 indicates that the distribution of low profitability customers (class 1 on y-axis) is bimodal. A large group of one-time shopper shopped for bargain items and never returned, while an even larger group of "bargain hunters" consistently shopped at the store but created poor return. On the other hand, the distribution of the high profitability group approximately follows a normal distribution. A relatively small group of one-time shopper, presumably just made a stop to purchase items out of convenience, generated a high one-time profit, but never returned.

First dimension (x-axis)	
Class	Profit/visit (in yen)
5	566 to 28091
4	315 to 565
3	170 to 314
2	41 to 169
1	-87440 to 40
Second dimension (y-axis)	
Class	Freq. of visits in 1998
5	13 to 323
4	7 to12
3	4 to 6
2	2 to 3
1	1

Table 1. Categorization of variables

The graph shows that the two proposed dimensions are rather independent and therefore are both meaningful in characterizing high-value customers. Our previous study showed that a one-dimensional definition such as total profit (profit per visit times frequency) was not effective in characterizing customer value. For short periods of observation, the group classified as high-value customer consisted of a high proportion of one-time shoppers when the one-dimensional definition was used.

In this study, shoppers who both score a 4 or above on both dimensions were classified as high-value customers.

The Data Mining System Architecture

The purpose of a Data Mining information system is to "identify valid, novel, potentially useful and understandable correlations and patterns in existing data" (Chung and Gray 1999). The data mining system developed at Pharma includes tools for undirected data mining as well as for directed data mining. For example, undirected data mining includes OLAP tools such as Cognos Powerplay (www.cognos.com) and visualization tools such as Spotfire (www.spotfire.com) that generate different views on the data for exploratory purpose. In this article, we only report results of directed data mining experiments -- predictive modeling for early identification of potential high-value new visitors.

Before business transaction data can be input into any useful data mining algorithm, they need to undergo a transformation process, often referred to as Extraction, Transform, Load (ETL) in data warehousing. The Pharma currently has not installed a data warehouse, in the traditional sense, to handle its Point-of-Sale (POS) data. It does have a separate relational system that supports day to day accounting and reporting. On the other hand, instead of storing POS data in a relational database using the star schema, Pharma stores all its transaction data as flat files on a single Unix platform with back up data in CD ROM (Hamuro, Katoh, Matsuda, and Yada 1998). The transaction data are relatively clean and unstructured. Pharma relies on a relatively simple but highly scalable system for managing its transaction data (approximately 60 Gigabytes a year). The data mining system (Figure 2) consists of several components, the first of which includes a scalable set of data manipulation commands that were written in Unix scripts. This tool, tentatively named P/S Transformer, enables the transformation of process-oriented data to subject-oriented data, which are more usable and amenable to data mining. For example, the P/S Transformer can directly act on large compressed files of transaction data to return information arranged by customers. User-defined criteria can be easily incorporated into the extraction process. We are currently developing a web-based graphical user interface for the software tool. In the following experiments, we used P/S Transformer to create appropriate data sets for analyses. These analyses are all performed offline to avoid the data communication bottleneck created in experimenting with different data sets.

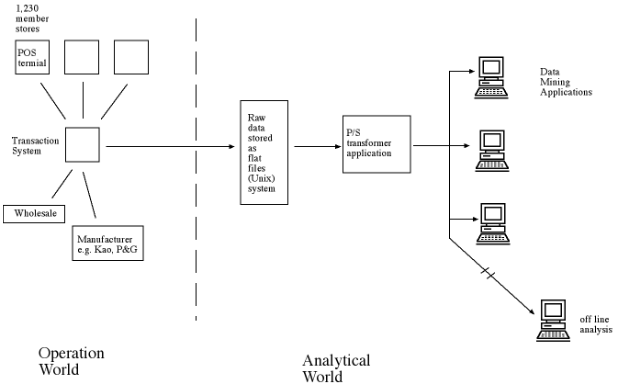

Figure 2. Architecture of the Pharma Data Mining System

Data Mining Results

A sample of 114,069 customers was used in the experiment. According to the criterion described in the Section on customer value, 12,050 (10.56%) customers were classified as "high-value"(HV). This category generated a disproportionate 52.5% of profit share and 38.4% of revenue share, exemplifying the strategic importance of the ability to manage relationship with this class of customers. The aim of the experiment is to a) construct a predictive model for identifying high value customers based on past data, b) dynamically provide the existing online system predictive values for relatively new visitors.

The dependent variable is whether or not a customer is classified as HV. This class membership variable was based on one-year data. There are two sets of independent variables used in the experiments. The first set is the logical choice of the set of the same variables used in classifying customers, namely, profitability per visit and frequency of visit. To emulate the realistic situation under which the model will be applied, we only used the first two months of data in computing the independent variables. Because of the business requirement of early identification of high-value customers, the observed period may not be sufficiently long to provide reliable data on these two variables. Therefore, in the second set of independent variables, we include other customer attributes such as customer demographics and product attributes in the hope of adding predictive power to the model. These variables include average quantity of items purchased, whether or not customer purchase a certain category of product (e.g. medicine), and number of product category being purchased. There were a total of 14 variables in this set.

To begin constructing a valid predictive model, we first divided the sample customer database into a training set that consists of a random sample of size 104,069 and a testing set of size 10,000. A data mining tool C5.0

(Quinlan 1986, Quinlan 1993) that is based on a decision tree algorithm was applied to the training data. With only two variables (profitability per visit and frequency), C5.0 provided a predictive model with a predictive accuracy of 82.9% (number of correctly classified HV / number of actual HV), and overall accuracy (number of correctly classified customers/ total customer) of 81.3%. The predictive accuracy on the testing set was respectively 81.3% (HV) and 80.1%(overall). As a baseline for comparing performance, we used the standard supervised learner logistic regression model. It was run under the Splus environment (Becker, Chambers, Wilks 1988). The logistic regression model had a 74.0% predictive accuracy for the training set and a 72.7% for the testing set, using the prior proportion of 0.1056 to 0.8944 as a cost function. The overall accuracy is 79.0% for both sets. For benchmarking purpose, we tuned the logistic regression model by lowering the probability threshold for classifying a customer as HV, so that the model reached a predictive rate of 82.7% for the training set. In this case, the predictive rate on the testing set was 82.5%. The overall accuracy rates were respectively 73.4% and 73.6%. We also compared the rates (correctly identified HV customers)/(total number of identified HV customer). The figures for C5.0 and logistic regression on the test set were respectively 31.2% and 29.1%. In summary, the decision tree learner seems to be superior in providing overall accuracy. Both supervised learning methods, however, provide reasonable predictive rates for the target variable.

To see if auxillary information could improve predictive power, we used all 14 independent variables in the second set of experiments. C5.0 tended to generate a complex tree and as a result overfitted the model. The predictive accuracy of the complex model actually was worse than the simple model when validated against a testing set. For two-month data, the predictive rate on training set was 93.0% but dropped to 73.2% on the testing set. The overall accuracy rate were respectively 86.8% and 82.6%. We also ran the experiment using logistic regression. The result showed that there were no substantial improvements. Using the prior 0.1056:0.8944, the predictive accuracy for HV and the overall accuracy for the training set was 71.1% and 79.5%, and 70.4% and 79.8% for the testing set. Results using selected subsets of variables were mixed, but none seemed to show a significant improvement over the simple model.

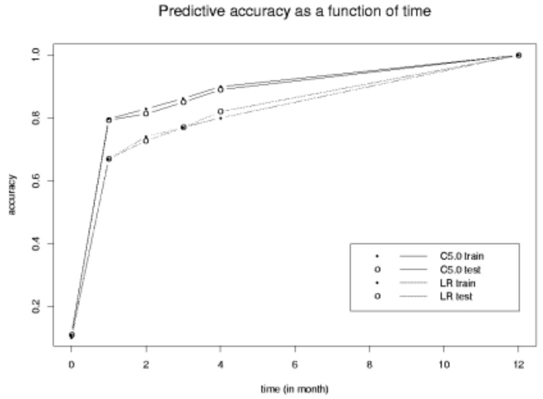

Figure 3. Predictive accuracy of C5.0 and logistic regression (LR). Logistic model is set to classify at prior distribution

To answer pressing questions from the marketing department such as "when do we know enough to start an intervention program?", we apply C5.0 and logistic regression to one-month, three-month and four-month data. We have not proceeded further than four-month because by then, the clerks would have known their customers quite well. Figure 3 summarizes the result on predictive accuracy. For predicting high-value customers, even one-month data seem to perform reasonably well, with predictive and overall accuracy of 79.3% and 71.8% on the testing set. We found that a substantial gain in total accuracy, however, is at using two-month data, with an overall accuracy of 80.1%

Summary

The study presented in the article illustrates the strategic value of data mining in managing high-value customers. We show how data mining provides a drugstore chain a better understanding of their customers and as a result leads to a higher quality of decision support to its marketing and service departments.

References

Becker, R. A., Chambers, J. M. and Wilks, A. R. *The New S Language*. Chapman and Hall, NY, 1988.

Berry, M. J, and Linoff, G. S. *Mastering Data Mining: the Art and Science of Customer Relationship Management*, Wiley, NY, 2000, Chapt. 4.

Chung, H. M., and Gray, P. " Special Section: Data Mining," *Journal of Management Information System* (16:1), Summer 1999, pp. 11-16.

Fishman, C., "This is a Marketing Revolution," *Fast Company*, May 1999, pp.206-218.

Hamuro, Y., Katoh, N. Matsuda, Y., and Yada, K. "Mining Pharmacy Data Helps to Make Profits," *Data Mining and Knowledge Discovery* (2), 1998, pp. 391-398.

Jones, T. O. and Sasser, W. E. Jr., "Why Satisfied Customers Defect," *Harvard Business Review*, November-December, 1995, pp.88-99.

Quinlan, J. R. "Induction of Decision Trees, " *Machine Learning* (1), 1986, pp. 81-106.

Quinlan, J. R. *C4.5: Programs for Machine Learning*, Morgan Kaufman, CA, 1993.

Reichheld, F. F. "Loalty-based Management," *Harvard Business Review*, March-April, 1993, pp. 64-73.

Spangler, W. E., May, J. H., and Vargas, L. G. "Choosing Data-mining Methods for Multiple Classification: Representational and Performance Measurement Implications for Decision Support," *Journal of Management Information System* (16:1), Summer 1999, pp. 37-62.

Sung T. K., Chang, N., and Lee, G. "Dynamics of Modeling in Data Mining: Interpretive Approach to Bankruptcy Prediction," *Journal of Management Information System* (16:1), Summer 1999, pp. 63-85.