INTERNATIONAL JOURNAL OF MARKETING,
FINANCIAL SERVICES & MANAGEMENT RESEARCH
Vol.1 No. 2, February 2012, ISSN 2277 3622

Online Available at indianresearchjournals.com

THE PREDICTION OF CHURN BEHAVIOUR AMONG INDIAN BANK CUSTOMERS: AN APPLICATION OF DATA MINING TECHNIQUES

S. MADHAVI
Assistant Professor, Gudlavalleru Engineering College, Gudlavalleru

ABSTRACT

The customer churn is a common measure of lost customers. By minimizing customer churn a company can maximize its profits. Companies have recognized that existing customers are most valuable assets. Customer retention is critical for a good marketing and a customer relationship management strategy. The prevention of customer churn through customer retention is a core issue of Customer Relationship Management (CRM). The paper presents churn prediction based on data mining tools in banking. In this paper, a study on modeling purchasing behavior of bank customers in Indian scenario has been attempted. A detailed scheme is worked out to convert raw customer data into meaningful and useful data that suits modeling buying behavior and in turn to convert this meaningful data into knowledge for which predictive data mining techniques are adopted. In this analysis, we have experimented with 2 classification techniques namely CART, and C 5.0. The prediction success rate of Churn class by CART is quite high but C 5.0 had shown poor results in predicting churn customers. However, the prediction success rate of Active class by C 5.0 is more effective than the other technique. But for reaping significant benefits, the models have predicted the churn behavior. Key Words: Customer Churn, Dataset, Modeling, Prediction and Active Class.

INTRODUCTION

Data mining is evolving into a strategically important dimension for many business organizations including banking sector. It is a method of analyzing the data from different viewpoints and summarizing it into valuable information. Data mining assists the banks to look for hidden pattern in a group and discover unknown relationship in the data (Fathimathabasum). Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately apparent to managers because the volume data is too large or is generated too quickly to screen by experts (Rajanish Dass). Data mining techniques help companies' particularly banking, telecommunication, insurance and retailing to build accurate customer profile based on customer behavior. Thus it is becoming a necessity in this competitive environment to analyze the data from data warehouse containing hundreds of gigabytes or terabytes of data (Fathimathabasum).

Data mining tools predict patterns, future trends and behaviors, allowing businesses to effect proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analysis of past events provided by retrospective tools typical of decision support systems.

The importance of collecting and analyzing data reflects

any business activity to achieve competitive advantage is widely recognized in today's age of information. Modeling and investigated system and discovering relations that connect variables in a database is the objective of data mining (Berson, A., et.al., 1999). Data mining uses different models for the creation of information about data which is known as discovery model. Data mining uses methodologies that can sift through the data in search of frequently occurring patterns, can detect trends, produce generalizations about the data, etc. These tools can discover these types of information with very little (or no) guidance from the user (Weiss,S.M., 1997). The main tasks such as Prediction, Classification, Detection of relations, Explicit modeling, Clustering, and Deviation Detection. Moreover, since the data mining process is systematic, it offers firms the ability to discover hidden patterns in their data-patterns that can help them understand customer behavior and market trends.

An Illustrative Data Mining Application in Banking: Churn Modeling

For forecasting the future churn, a very vigorous model should be in hand and an active model can only be built if we have a vigorous dataset in hand. Hence, data preparation is a vital step in churn prediction and it takes almost 60-70 percent of total time. In this part we will make clear the kind of data required for active system, commonly customer data available in the real life, guideline to obtain useful attributes from the obtainable data. Constructing a model for churn prediction means that we are trying to model the customer's behavior churning out. For this to be successful, the customer transaction activities should be analyzed in a specific period of time. Hence, taking a data would never be enough for the requirement. On the other hand, considering the transaction activities in a fixed time period would also not satisfy the requirement. The reason can be explained by an example. Say, for example, a model is built using data of 1000 customers of which 700 are active and 300 are known to be churned out and their 3 months activities are analyzed (say, Feb 2006 to April 2006). Here the time period is fixed and the activities done in this time period of all the 1000 customers are only analyzed. Now, out of 300 churn customers, say 50per cent of them have churned away in February. This means, the model will not be fully trained with the behavior of churn customers before churning as only one month's activity is analyzed. This problem occurred because of fixing the timeline before hand as shown in figure 1. In this paper, we consider a dynamic time period, which differs for each customer. This concept would be better explained by continuing the above example. If a customer has churned away in Feb. 2006, from that point of rime, the past 3 months activity is considered i.e., transaction activities done in Dec 2005, Jan 2006, and Feb 2006 are considered. And if another customer churns away in march, transaction activity of Jan 2006 and Feb, March 2006 should e considered. This can be seen in figure 2. This way, we adapt an active time line for every customer and hence refrain from the difficulty of not training the model accurately. This abstract idea applies to the information that constituting an account of churn that has occurred. For diligent (active) records, we can take in to account (or examine) the behavior in any 3 months portion of time. In our detailed examination we consider the behavior of most recent 3 months before last transaction date of the active customers. The total count of months of data to be considered for churn analysis is a business problem. Usually, considering the transaction activities of 3 months would (be adequate) satisfy the requirement.

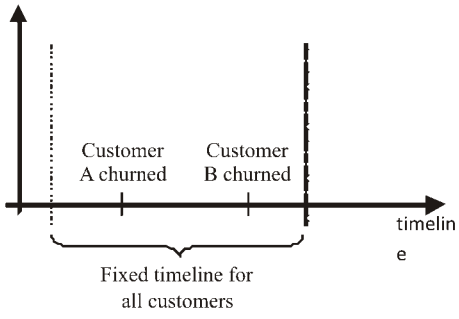

Fig 1: defined timeline for all customers

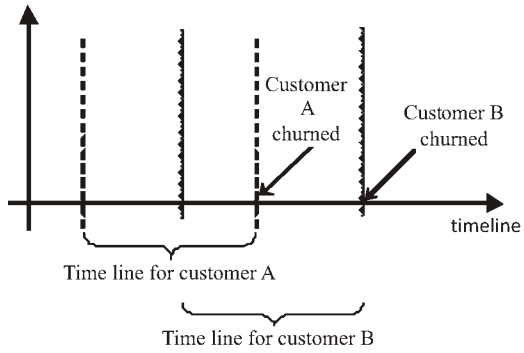

Fig 2: Dynamic time period

Next, in this part, we make clear about the actual data we used and data filtering steps to prepare an efficient dataset. We acquired the customer data from a Nationalized Indian Bank. The particulars of the data acquired are shown in Table 1.

TABLE 1
DETAILS OF CUSTOMER DATA

Table Name	Attributes	No. of Records
Customer	Custno, Name1, Name2, Address, Status, DoB (Date of Birth), Edn	40,870
General Ledger	Custno, AcNo, Descr, DOP (Date of a/c opening)	1,08,019
Dormant	Acno, Descr, Dormant	17,992
Master	Acno, Balance, Dormant flag.	31,012
Txn	Acno, Trntype, Date, Amount	31,39,010
Ttype	Ttype#, Typecode, Descr	93

Source: Bank Records

The customer table has customer details like customer number, name, address, date of birth and status. Completely there are 40,870 customers. The general ledger table holds the account numbers, account types, date of opening and description of accounts. Here, we have 1, 08,019 accounts for the above defined informed as dormant since a while. The master table holds all the account numbers and their latest balances. The Txn table holds the last 5 years transactional details of all accounts. Finally Ttype table has description for different transaction types. Completely there are 93 transaction types outlined.

The final dataset prepared on the basis of the available data shown in table 1, which contains the attributes such as: customer number, Duration (Dur), number of Credit transactions in 3 months (CRTxnx), number of Debit transactions in 3 months (DRTxns), Average credit amount in 3 months (AvgCrAmt), Average debit amount in 3 months (AvgDrAmt), total number of other accounts (NumOtherAccs), percentage of accounts closed in 3 months (PercClosedRecently), status (status).

The Duration attribute consists of the number of months the customer has transacted with the bank. When we say 3 months is an attribute, we refer to the concept of dynamic timeline explained previously. The final dataset consists of 1,484 records. Out of which 1,163 are active customers and 311 are churn customers. Extracting this dataset from the raw data available is not an easy job. Further in this section, we share our experiences in working with the above data and subsequently extracting the data for required set of attributes.

There are many discrepancies in data. In these discrepancies, the most important and crucial one is status attribute in customer able, which is explained here. The status attribute describes whether a particular customer is active or inactive or churned out. So, status attribute develops the class variable in dataset. A customer may have more than one account with the bank. Then, the difficulty here was if one of accounts of customer's is unopened, then that condition is set as churned out. But, customer is still with the bank as his/her other accounts are still active. Therefore, it was realized that, it was very difficult to prepare training dataset basing on status attribute. However, the status of each & every account is very much required to develop the class variable of the dataset. The Descr attribute in general ledger table served to verdict the status of each account. The Descr attribute consists of the description of the account type and moreover that whether it is inactive or still active. Hence by using Descr attribute in the general ledger, the status attribute of the dataset is developed. Another problem with the data was some of the attributes has missing values. Fields like DOB, DOP were partially developed. By this reason, demographic attributes such as age, gender could not be considered in the final dataset.

The data consists of different types of accounts like savings account, current account, cash credit account, loan account etc. In our analysis, the behavior of savings account customers is analyzed. There are 22,155 savings accounts out of 1, 08,019 total accounts. In these 22, 155 accounts, 6, 633 accounts are found to be inactive from dormant table. But here we are not concentrating on modeling the behavior of dormant customers so these accounts are ignored from our target customer base. There are some accounts whose duration is not so much i.e., number of months transacted with the bank was very less. These accounts gave a notion that these accounts are opened for a particular purpose and closed as soon as that purpose completed. Considering these accounts behavior may provide a poor dataset and consequently despicable predicting model. So we are ignoring the records whose duration is less than 6 months. There are some set of accounts whose duration is more than 6 months but transaction activities not much went in these accounts (but have very less transaction activities). This gave a view that these customers have just opened the savings accounts and rarely done transactions through them. Such type of data also effect in poor modeling. Hence the records whose number of transactions is less than 50 are also ignored from target customer base. After operating all these filtering steps, target customer base reduced to 1, 484 accounts. Out of 1,484 accounts 1,163 are active records and 311 are churn (inactive) records. By originating the right sort of data from the obtainable raw data, 1,484 customers transactional behavior is analyzed. As expressed before, In order to model the behavior of both active and churn customers, we have to practice (built) the model with their most recent behavior. The most recent behavior of accounts can be obtained from the Txn table which has the transactional details of all the accounts. Txn table holds Trntype field, says the kind of the

transaction involved such as credit voucher, deposit, inward cheque clearing, etc., Trntype field can take the values of credit voucher, cash deposit, inward cheque clearing etc. There are 93 different transaction types and each can be recognized as either a credit transaction or a debit transaction. After separating the credit and debit transactions done by the customers, have to calculate the number of credit transactions (CRTxn) and number of debit transactions (DBTxn) for all the 1,484 customers. The average amount of money transacted by the customers in the defined timeline may also provide support in training the model in a better manner. For this reason, two more attributes, say, the average amount included in credit transactions (AvgCRTxn) and the average amount included in debit transactions (AvgDBTxn) were intended for 1,484 customers. As we said earlier, we have taken in to account only the transaction behavior of savings accounts of the customers. As these customers have other accounts with the bank, the number of other accounts and percentage of closed accounts in the defined time line may also helps to train the model better. So two more attributes, say, number of other accounts (NumOtherAcc), percentage of other closed accounts in defined timeline of 3 months (PercClosedRecently) were intended for all the 1,484 customers.

CONSTRUCTING DATA MINING MODELS AND TRAILING OUTCOMES:

In the previous part, we have made clear about the data filtering and preparing steps in brief. After having dataset in hand, the immediate step is to prepare a predictive model by using this dataset. Usually, Data mining methods are used to prepare data models and these models consequently help for future predictions. As predicting churn is exclusively a classification problem, supervised data mining techniques are used to take away this problem. Here we used two classification tree algorithms, say, CART, C5:0 for preparing the two classification trees.

Classification tree model using CART:

Classification tree is built adopting Classification and Regression Tree (CART) model on the training dataset with the following specifications: Optimal tree cannot be discovered since CART does not use stopping rule. Thus firstly the tree is over grown and then pruned back to ensure that significant patterns are not overlooked by stopping too soon. The advantage with CART is that it performs binary splitting to make the data more sparing and to detect more patterns before too few data are left for learning. The study used Gini concentration coefficient to abridge power curves of prediction. The explanatory variables are Customer Duration, CRTxn, DRTxn, AvgCrAmt, AvgDrAmt, PerClosedRecently and target variable is Status. CART gives rules for the target variable as a function of other fields in the dataset that are previously identified as explanatory variables. 80 per cent of the dataset i.e., 1,187 samples consisting 926 active customer records and 261 churned customer records, are considered in training dataset. The remaining 20 per cent of the dataset i.e., 296 samples consisting 241 active customer records and 57 churned customer records, are considered in testing dataset. The confusion matrix and prediction success rate of training dataset and testing dataset are shown in Table 2 and Table 3 respectively.

TABLE 2

CONFUSION MATRIX AND PREDICTION SUCCESS RATE FOR TRAINING DATA

True Class	Total # samples	Predicted Active	Predicted Churn	Success percent
Active	926	797	129	85.79
Churn	261	13	248	95.01

TABLE 3
CONFUSION MATRIX AND PREDICTION SUCCESS RATE FOR TEST DATASET

True Class	Total # samples	Predicted Active	Predicted Churn	Success percent
Active	241	208	33	86.30
Churn	57	5	52	91.22

The retention rate of active customers is comparatively less because, although some of the customers status is marked as active, they have exhibited churn characteristics. It is this segment of customers on which bank has to focus upon and apply churn prevention strategies. There are 17 leaf nodes in the tree model generated using CART and hence 17 decision rules can be drawn from it. Table 4 depicts the 17 decision rules.

TABLE 4
DECISION RULES GENERATED BY CART

Rule #	Rule	Predicted Class	# Cases
1	AvgDrAmt <= 608 and AvgCrAmt <= 37.5	Churn	93
2	AvgDrAmt <= 608 and AvgCrAmt > 37.5 and AvgCrAmt <= 1655.5 and Duration > 18	Active	61
3	AvgCrAmt <= 1655.5 and AvgDrAmt > 608	Churn	102
4	AvgCrAmt > 1655.5 and Duration <= 23.5 and AvgDrAmt <= 1300.5	Active	20
5	AvgCrAmt > 1655.5 and Duration <= 23.5 and AvgDrAmt > 1300.5 and PercClosedRecently > 0.0416667	Churn	26
6	AvgCrAmt > 1655.5 and Duration > 23.5 and Duration <= 27.5	Active	615
7	Duration > 27.5 and Duration <= 68.5 and AvgDrAmt <= 3421.5 and AvgCrAmt > 1655.5 and AvgCrAmt <= 3674	Churn	16
8	Duration > 27.5 and Duration <= 68.5 and AvgDrAmt <= 3421.5 and AvgCrAmt > 3674	Active	18
9	AvgCrAmt > 1655.5 and Duration > 27.5 and Duration <= 68.5 and AvgDrAmt > 3421.5	Churn	38
10	AvgCrAmt > 1655.5 and Duration > 68.5	Active	43
11	AvgDrAmt > 1300.5 and PercClosedRecently <= 0.04 and AvgCrAmt > 1655.5 and AvgCrAmt <= 17894.5 and Duration <= 17.5	Churn	70
12	PercClosedRecently <= 0.04 and Duration > 17.5 and Duration <= 23.5 and AvgDrAmt > 1300.5 and AvgDrAmt <= 7449 and AvgCrAmt > 3527 and AvgCrAmt <= 17894.5	Active	32
13	PercClosedRecently <= 0.04 and Duration > 17.5 and Duration <= 23.5 and AvgDrAmt > 1300.5 and AvgDrAmt <= 7449 and AvgCrAmt > 1655.5 and AvgCrAmt <= 3527	Churn	10
14	AvgDrAmt <= 608 and AvgCrAmt > 37.5 and AvgCrAmt <= 1655.5 and Duration <= 18	Churn	4
15	PercClosedRecently <= 0.04 and AvgCrAmt > 1655.5 and AvgCrAmt <= 17894.5 and Duration > 17.5 and Duration <= 23.5 and AvgDrAmt > 7449	Churn	10
16	Duration <= 23.5 and PercClosedRecently <= 0.04 and AvgCrAmt > 17894.5 and AvgDrAmt > 1300.5 and AvgDrAmt <= 14238.5	Active	10
17	Duration <= 23.5 and PercClosedRecently <= 0.04 and AvgCrAmt > 17894.5 and AvgDrAmt > 14238.5	Churn	10

Among the 17 rules generated by CART, 12 rules have adequate number of cases and these rules can be adopted by the manager for predicting future churn customers.

Classification tree Model using C5.0:

Another classification algorithm that produces decision trees with variable branches per node is C 5.0. Status is taken as target variable and Customer Duration, CRTxns, DRTxns, AvgCrAmt, AvgDrAmt, PercClosedRecently are used as explanatory variables. 80 per cent of the dataset i.e., 1,187 samples consisting 926 active customer records and 261 churned customer records, are taken in training dataset. The remaining 20 per cent of the dataset i.e., 296 samples consisting 241 active customer records and 57 churned customer records, are taken in testing dataset. The confusion matrix and prediction success rate of training dataset and testing dataset are shown in Table 5 and Table 6 respectively.

TABLE 5
CONFUSION MATRIX AND PREDICTION SUCCESS RATE FOR TRAINING DATA

True Class	Total # samples	Predicted Active	Predicted Churn	Success percent
Active	926	881	45	95.14
Churn	261	60	181	69.3

TABLE 6
CONFUSION MATRIX AND PREDICTION SUCCESS RATE FOR TEST DATA

True Class	Total # samples	Predicted Active	Predicted Churn	Success percent
Active	241	232	9	96.26
Churn	57	18	39	68.4

DISCUSSION OF RESULTS

In this analysis, we have experimented with 2 classification techniques namely CART, and C 5.0 on 1,484 sample of bank customers in which 1,163 were active customers and 311 were churn customers. We used CART and C5.0 to trace out significant customer characteristics to predict churn. While CART yielded 95.01 per cent classification rate on training data and 91.22 per cent on test data, C5.0 yielded 69.3 per cent classification rate on training data and 68.9per cent on test data. The prediction success rate of Churn class by CART is quite high but C 5.0 exposed poor results in predicting churn customers. However, the prediction success rate of Active class by C 5.0 is more than the other technique. In order to have significant benefits, the model should be able to predict the churn behavior better. Thus, a model with higher prediction success rate of Churn class (i.e., CART) has to be chosen for reaping higher benefits. In all the decision tree models, all the explanatory attributes were found to be influencing the target variable, i.e., status of the customer.

CONCLUSION

In recent years, data mining has gained widespread attention and increasing popularity in the commercial world. Transforming raw customer data into information is the goal of data mining projects. But failure to turn this useful information into customer satisfaction and increased profits is the key to why many such projects often fall short of expectations. Thus, it is essential that the customers, indicated by the churn model, to become churn should be focused. If the churn prevention program is effective, the bank can look forward to reaping significant benefits from its efforts. A company that can retain 5per cent of its current customers can raise its profits by 25per cent. In

this paper, we have given a detailed guideline of converting raw customer data of a bank into useful data and then convert this data info useful information using data mining techniques. We have explained the concept of dynamic timeline that should be considered while converting raw data into useful data. We have extracted the data for chosen attributes from raw customer data for a chosen set of 1,484 customers. In these 1,484 samples, 1,163 customers have the status of active and 311 customers have the status of churn. We used CART and C5.0 to recognize significant customer characteristics to predict churn. While CART yielded 95.01per cent classification rate on training data and 91.22per cent on test data, C5.0 yielded 69.3per cent classification rate on training data and 68.4per cent on test data. The study predicts the future churn of banking customers that can be checked, by formulating intervention strategies based on churn prediction to reduce the lost revenue by increasing customer retention. It is expected that, with a better understanding of these characteristics, bank managers can develop a customized approach to customer retention activities within the context of their Customer Relationship Management efforts.

REFERENCES:

1. Athanassopoulos, A.D., 2000, "Customer satisfaction cues to support market segmentation and explain switching behavior", Journal of Business Research, 47 (3), 191 - 207.
2. Berry M.J.A. & Linoff G. (1999) Mastering Data Mining: The Art and Science of Customer Relationship Management. John Wiley and Sons, Inc. New york, pp.335-341.
3. Berson,A.,et.al., Building Data Mining Applications for CRM, McGraw-Hill,Rockefeller center, New york, 1999, 825-844.
4. Fayyad U., Grinstein G.E. & Wierse A. (2002). Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann Publishers. California, Academic Press
5. Hanjiewei, Michelin Kamber, Data mining concepts & Techniques", 2/e, Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann Publishers, Massachusetts, pp. 224-256.
6. Indiastudychannel.com, fathimathabasum: Datamining in banks, accessed on 22nd Feb2011. http://www.indiastudychannel.com/projects/1804-Data-Mining-anks.aspx.
7. J. Gehrke, V. Ganti, R. Ramakrishnan, and W.-Y. Loh, "BOAT - Optimistic decision tree construction," in Proc. ACM SIGMOD Int. Conf.Management of Data, Philadelphia, PA, 1999, pp. 169-180.
8. J. Gehrke, R. Ramakrishnan, and V. Ganti, "RainForest - A framework for fast decision tree construction of large datasets," in Proc. 24th Int.Conf. Very Large Data Bases, New York, 1998, pp. 416-427.
9. J. R. Quinlan, "Decision trees as probabilistic classifiers," in Proc. 4th Int. Workshop Machine Learning, Irvine, CA, 1987, pp. 31-37.
10. J. Shafer, R. Agrawal, and M. Mehta, "SPRINT: A scalable parallel classifier for data mining," in Proc. 22nd Int. Conf. Very Large Data Bases, Mumbai (Bombay), India, 1996. pp. 544-555.
11. Kotler, P. Keller, K.L, Marketing Management (12th Edi.) Pearson Prentice Hall, Upper Saddle River, N.J., Northwestern University, 2006. pp.177-192.
12. M. sadiq sohali , Service quality in hospitals:

more favourable than you might think, MCB UP Ltd Publishers, Vol. 13 Iss: 3,UK, pp.197 - 206.
13. Pressler Margaret, Signs of Fraud go beyond signature, Washington post, Washington, 2003. p H05.
14. Rajanish Dass, Data Mining in Banking and Finance: A note for Bankers, Indian Institute of Management, Ahmedabad,.Prentice-Hall of India,.pp. 3-9.
15. R. Rastogi and K. Shim, "PUBLIC: A decision tree classifier that integrates building and pruning," in Proc. 24th Int. Conf. Very Large Data Bases, NewYork, 1998, pp. 404-415.
16. Reichheld, FF and Sasser, WE (1990) Zero Defections: Quality Comes to Service, Harvard Business Review, Vol. 69, September-October, pp. 105- 11.
17. Smith A. "CRM and customer service: strategic asset or corporate overhead?", Handbook of Business Strategy, MA USA, Vol .7, 2006, pp.87 - 93 .
18. Turban, E., Mclean, E., Wetherbe, J., Bolloju, N., Davison, R. (2002). Information Technology for Management: Transforming Business In The Digital Economy. John Wiley and Sons, INC: New York, PP.441-458.
19. Van den Poel, D.,& Lariviere, B., "Customer attrition analysis for financial services using proportional hazard models", European Journal of Operational Research, Elsevier, vol. 157(1), pages 196-217.
20. Weiss, S.M., Predictive Data Mining: A Practical Guide, Morgan Kaufmann Publishers, Massachusetts, 1997.

❖❖❖❖❖