Customer classification in retail marketing by data mining

Narendra Kumar Jha, Manoj Kumar, Anurag Kumar, Vijay Kumar Gupta

Abstract— Capital investment in retail sector and competition in the market has changed the style of marketing. At the same time the enhancements in the field of information technology provided an upper hand to the marketer to know the exact need, preference and perches trend of the customer. By knowing the actual need, preference and purchase trend of customers the marketer can make a future business plan to increase the sale and earn more profit. This paper provides a framework to the retail marketer to find the potential customer by analyzing the previous purchase history of the customer. This task can be accomplished by the use of data mining technique. In this paper we have used k-mean clustering algorithm and Navie Byes' classifier for indentifying potential customer for a particular section of products of the retailer.

Index Terms—CRM, Data mining, Frequency, K-mean clustering, Monetary, Marketing, Naive Byes' classifier, Recency, Sales, Weighted score

1 Introduction

During recent decades the enhancement in information technology and capital investment in the field of sales and marketing changed the way of marketing. Now a day the marketer creates and manages large volumes of data on the customers. These databases contain valuable information which is hidden [1],[2]. The marketer maintains transaction as well as the customer database. The volume of database is typically very large and manual manipulation of these databases is hectic and time consuming in fact if database is very large then the task is almost impossible. Customer is the wealth of any business enterprises'. Philip Kotler has pointed out specially that customer- centered companies not only need to manage the products, but also need to manage the customers [5]. Retail industry, being the fifth largest in the world, is one of the sunrise sectors with huge growth potential and accounts for 14-15% of the country's GDP. Comprising of organized and unorganized sectors, Indian retail industry is one of the fastest growing industries in India, especially over the last few years.

According to the Global Retail Development Index 2012, India ranks fifth among the top 30 emerging markets for retail. The recent announcement by the Indian government with Foreign Direct Investment (FDI) in retail, especially allowing 100% FDI in single brands and multi-brand FDI has created positive sentiments in the retail sector.

Since revenue and the competition is increasing in the field of retail marketing therefore every marketer wishes is to increase Profits through sales, but this can't be possible without managing customers.

2 Problem Statements

Every business organization has a primary goal to increase sales and through which it earns profit. To increase sales they apply marketing and sales promotion strategies so that customers can know about their product and their promotion activities such as a discount on a particular item or an entire section. Generally for these activities organization apply mass marketing which causes decrease in intensity of effort. If they apply their effort into a particular direction then the intensity of effort will increase. The current marketing and sales promotion in retail field is almost dependent on the mass marketing. The marketer promotes the product to the mass of the customer without knowing their need of such products. Mass marketing is a market coverage strategy in which a firm decides to broadcast a message that will reach the largest number of people possible. Traditional mass marketing has focused on radio, television and newspapers as the media used to reach this broad audience. By reaching the largest audience possible, exposure to the product is maximized. There is an increasing awareness that effective customer relationship management can be done only based on an actual understanding of the needs and preferences of the customers. Under these conditions, data mining tools can help uncover the hidden information which resided already in the database.

But still there is a lack of such system which can target the customers according to their need. The existing system broadcasts the sales and promotion news to all that become more expensive and less effective. Customers are important resources for an enterprise. Therefore, it is essential for, enterprises to successfully acquire new customers and retain high value customers [6]. To achieve these aims many enterprises have gathered significant numbers of large databases, which then can be analyzed and applied to develop new business

- *Narendra Kumar Jha is currently pursuing M.Tech program in Software engineering from Babu Banarsi Das University, Lucknow, PH- 09795367355.E-mail: nkj.jha@gmail.com*
- *Manoj Kumar is currently working as an Assistant Professor in the department of Information and Technology Babu Banarasi Das National Institute of technology and management PH-09807347497. E-mail: manoj.brnwl82@gmail.com*
- *Anurag Kumar is currently pursuing M.Tech program in Software engineering from Babu Banarsi Das University, Lucknow, PH-07895361338. E-mail: anurag.kumar269@gmail.com*
- *Vijay Kumar Gupta is currently pursuing M.E program in Computer Engineering from National Institute of Techers Training and Research, Chandigarh PH- 08957680057. E-mail: vijaythesoft84@gmail.com*

strategies and opportunities. However, instead of targeting all customers equally or providing the same incentive offers to all customers, enterprises can select only those customers who meet certain profitability criteria based on their individual needs or purchasing behaviours [6]. These potential customers are the main contributor to the revenue of the company.

3 Research methodology

In this work we have provide a framework for the retail marketing promotion by analyzing their database and finding the potential customer for a particular product on the basis of which the business organization can make marketing decisions. Here main motto is to analyze the customer behavior and their purchasing activities so that a pattern can be obtained. For this purpose the data mining provides a technique for analysis and dependency analysis to discover the pattern and target the appropriate customer who can benefit the point of sales increment**.**

Data Mining Task:

Tan, Kumar and Steinbach define data mining in their book "introduction to data mining", as the process of automatically discovering useful information in large data repository [7]. Data mining techniques are deployed to scour large databases in order to find novel and useful patterns that might remain unknown. The mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have been more appropriately named "knowledge mining from data" [8]. Data mining, also known as knowledge discovery in databases is a rapidly emerging. This technology is motivated by the need of new techniques to help analyze, understand or even visualize the huge amounts of stored data gathered from business and scientific applications. It is the process of discovering interesting knowledge, such as patterns, associations, changes, anomalies and significant structures from large amounts of data stored in databases, data warehouses, or other information repositories. It can be used to help companies to make better decision to stay competitive in the marketplace. The major data mining functions that are developed in commercial and research communities include summarization, association, classification, prediction and clustering. These functions can be implemented using a variety of technologies, such as database-oriented techniques, machine learning and statistical techniques. For the intended purpose we have used clustering and classification.

Basic concept of transaction database:

The customer transaction record generally consists of following attributes. Customer mobile number, items of purchase, mode of payment, age group of the customer etc. If we denote a transaction as a set T, then its elements will be:

Mb: Mobile number of the customer.
Dp: Date of purchase.
Sp: Section of purchase.
Ap: amount of purchase.
Mp: mode of payment.
Ag: Age group of the customer.
G: gender of the customer.
Ig: income group of the customer. (This should be already known to the marketer. The income group does not depend on the single factor. The marketer should know this by several researches, such as the customer has his/her own house or not, the customer owns four- wheeler or not etc. This research is beyond scope of our discussion.). So a particular transaction can be represented as a set T.

T= {Mb, Dp, Sp, Ap, Mp, Ag, G, Ig }

These records reside in the database and contain lots of information that is hidden which can be used to enhance the promotion activities.

Transactional database is treated as universal set U of all the transactions. At the first step we derive the following information.

U = {T1,, T2, T3, ..., Tn}

$\forall i \ [T_i \in U]$ where T_i represents the ith transaction.

Using these transactions we can find three important parameters frequency, size and recency of purchases [3],[4],[6].

Frequency of purchase:

How often does the customer buy any product? By knowing this the marketer can build targeted promotion.

Size of purchase:

How much does the customer spend? This information will help to marketer to pay high attention during a promotion to target these customers.

Recency of purchase:

How long has been the customer made a purchase? The marketer may investigate the reasons a customer or a group has not purchased over a long period of time.

On the basis of these three parameters the customers can be grouped into two categories, i.e. more profitable and less profitable category. For this purpose we have used K-mean clustering algorithm.

Basic k-means algorithm:

1. Select K points as initial centroids.
2. Repeat
3. Form k clusters by assigning each point to its closest centroid.
4. Recompute the centroid of each cluster.
5. Until centroids do not change.

Since the k-means algorithm requires weightings point on the basis of that the transaction data can be clustered in the numbers of desired cluster.

Calculation of weighted score:

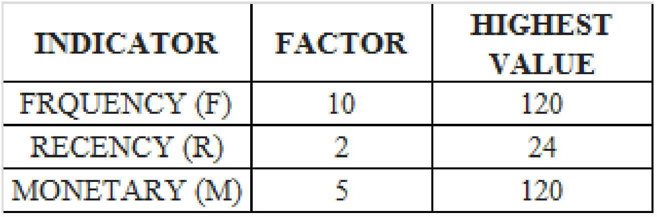
For the analysis purpose the transaction records of the customers are chosen between the specific periods of time. The time period can vary from research to research or it may depend on the condition and requirement of the marketer. Here we are taking time period of 12 months for analysing customer's transaction database. During these 12 months all transaction records from the database will be analysed and weighted value is assigned to each customer that can be distinguished by their mobile number (Mb). This will act as the primary key in the database. The total weighted score is calculated on the basis of three individual factors that is frequency, monetary and recency.

Frequency Score F:
The frequency score tells how many times the customer visits the store during the specified period of time i.e. twelve months. If a customer visited the store and purchased only once then his frequency score is 1. Each time the customer visits the store again the score is increased by one. But we have assumed constant twelve as the threshold value. This assumption can vary as per the market segments. The very high threshold value of frequency will make result biased towards Frequency score. Now we will calculate Fw (weighted frequency score) so that better clusters can be generated. To find Fw we multiply F by a constant Cf (multiplying factor). In proposed model we have taken Cf equals 10. That means if a customer purchases three times during the last three month their weighted score will be:
Weighted Frequency Score of ith customer is:
$Fw_i = 3*10 = 30$

Recency Score R:
Recency tells how recent the customer is. Least score is given to the customer who has not visited the store for the longest time. Our assumption is based on the observation that customers who have visited in the last few months will probably be visiting again in the next few months and contribute to the revenue growth of the company. Conversely, customers that have not visited for long will probably not visit for a long time and will contribute less to the company's profit. If a customer purchased within last month then his recency score will be 12. If a customer purchased in the second last but not in the last month then his recency score will be 11. If a customer purchased in the third last but not in the second or last month then his recency score will be 10 and so on. But I have assumed constant twelve as the threshold value. This assumption can vary as per the market segments. The very high threshold value of recency will make result biased towards recency score. Now we will calculate Rw (weighted recency score) so that better clusters can be generated. To find Rw we multiply R by a constant Cr (multiplying factor). In proposed model we have taken Cr equals 2. So if the customer has last purchased 3 month before then his weighted recency score will be
Weighted Frequency Score of Ith customer is:
$Rw_i = 10*2 = 20$

Monetary Score M:
The monetary score tells how much amount a customer purchased during specified period of time. For our purpose the period is taken this period as twelve months. Total purchase done by the customer during the the last twelve months is calculated. Then we calculate the average purchase dividing the total purchase by the number of times the customer makes purchase. To find the Monetary score M average purchase is divided by 1000, so that equal weightage can be assigned to all the indicators. To find weighted monetary score Mw, we multiply M by a constant Cm (multiplying factor). In proposed model we have taken Cm equal to five. Same as the Cm, Cr is taken as the multiplying facter for better clustering.
Suppose a customer has purchased 3 times and his total purchase is Rs 6000.
Then,
Average purchase Ap = 6000/3 = 2000
Monetary score M = Average purchase/1000
= 2000/1000 = 2
Weighted monetary score $Mw_i = (2000/1000)*5 = 10$

Since a higher recency, monetary or frequency score can lead to biased clusters towards one of the indicator. So Threshold limit of each indicator is given below as discussed above. As discussed above we have assumed multiplying factor so that better clustering can be done.



INDICATOR	FACTOR	HIGHEST VALUE
FRQUENCY (F)	10	120
RECENCY (R)	2	24
MONETARY (M)	5	120

Calculation of total weighted score:
The total weighted scores Tw_i for each customer Mb_i will be the sum of all the individual indicator's weighted score.
Total weighted score = weighted score of frequency
+ weighted score of recency
+ weighted score of monetary.
$Tw_i = Fw_i + Rw_i + Mw_i$
Through this the total weighted score Tw for each customer is calculated. For clustering these records k-means algorithm is applied as discussed above.
Input:
$Mb_i = \{Mb_1, Mb_2,......,Mb_n\}$ // set of all customer.
K // number of desired cluster.

K-means is an iterative clustering algorithm in which items are moved among set of cluster until the desired set is reached. Here we assume K=2 for our purpose.
In this way the customer database is devided in two clusters.
First cluster contains the customer data which has more weighted frequency or more recency or more monetary.
Second cluster contains the customer data they are either less frequency, less monetary or less recency.
The first cluster is more profitable as the point of view of marketer. Hence further during marketing and sales promotion cluster 1 is targeted and the cluster 2 is ignored.
Clustering of customers in cluster
Again we will apply k-mean clustering algorithm on the data in cluster 1 on the original value of frequency score (F), recency score (R) and monetary score (M). Since one higher value of any of the indicator (R, M and F) may lead the customer to be the part of the first cluster hence further clustering is required. Here input will be cluster 1 data and K (number of clusters) = 3 (high, medium, and low) for each score (R, M, F). The overall scenario is as shown in fig1.

Classification of target customer for promotion:

For classification problem we have used Naive Bayes Classifier which is based on Bayes Theorem.

Bayes Theorem: Let X and Y be a pair of random variables. Their joint probability, P(X=x, Y=y), refers to the probability that variable X will take on the value x and variable Y will take on the value y. A conditional probability is the probability that a random variable will take on a particular value given that the outcome for the another variable is known.

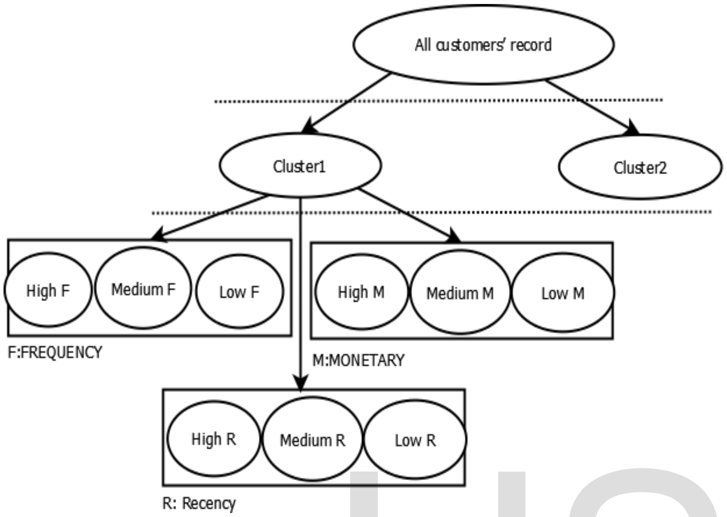


Fig 1: Scenarios

For example we have following table on the basis of above computation:

Mb	Ag	F	R	M
****1	Youth	High	High	Middle
****2	Midd.age	Low	Midd	Low
****3	Youth	Low	Midd	High
****4	Old	Midd	Midd	High
****5	Midd.age	High	Low	Low
****6	Midd.age	High	High	High
****7	Youth	Low	Low	Low
****8	Old	Midd	Low	Middle
***10	Midd.age	High	High	Low
***11	Old	Low	Low	Low
***13	Midd.age	Midd	Midd	Middle
***14	Old	Midd	Low	Low
***15	Youth	Low	High	High
***16	Youth	High	Midd	High
***17	Midd.age	Midd	Midd	Low
***18	Midd.age	Midd	High	High
***20	Youth	High	Low	Middle
***21	Youth	Low	High	Low
***22	Old	Midd	Low	Middle
***24	Youth	Low	High	Low

Table 2 (a): Customer Record

Mb	Ig	Profit / product	Buy_a product P	G
****1	High	High	Yes	Male
****2	High	High	Yes	Male
****3	Low	Medium	No	Female
****4	Low	Low	Yes	Female
****5	High	Medium	Yes	Male
****6	Low	Medium	Yes	Female
****7	Medium	Low	No	Male
****8	High	Low	No	Male
***10	High	Low	No	Male
***11	High	High	Yes	Female
***13	High	Medium	Yes	Male
***14	High	Medium	Yes	Male
***15	Low	Low	No	Female
***16	Low	High	Yes	Female
***17	Medium	High	Yes	Male
***18	Medium	Medium	No	Female

***20	High	Medium	Yes	Male
***21	Medium	Low	Yes	Male
***22	Medium	Medium	No	Male
***24	Medium	low	No	Female

Table 2 (b): Custome Record

If there is an item $P \in I_i$ (section of item) which need to promote for increasing the sale of the product P, the marketer has to find the potential customer for the product P then in the proposed model he applies Naive Bayes classifier work as follow:

1. Let the product P∈ I_{th} (section of item) has to be promoted then the customer data of section Ii is used as the training set of the classifier. Each tuple is represented as n-dimensional attribute.

$X=\{x_1, x_2, x_3, ..., x_n\}$,

where x1 = frequency, x2= monitory and so on.
2. if there are two predefined classes which is buy-the-product and not-buy-product
3. For the testing tuple X from entire customer database the classifier predict that $X \in c_i$ having the highest posterior probability.

$$P(ci/\mathrm{X}) = P(x/ci)\ P\ (ci)/P(x)$$

4. Thus P(X/ci) = $\prod_{k=1}^{n} (P(x_k / c_i)$

P (x_1/ci)* P (x_2/ci)* …..* P (x_n/ci)

This value should be maximum for the test data to belong to class ci.
5. In order to predict the class label of x, P(x/ci)* P(ci) is evaluated for each class.
6. the classifier predict that the class label of testing tuple x is the class ci if and only if

$P(x/c_i)\ P(c_i) > P(x/c_j)\ p(c_j)$ for 1<=j<=m, j≠i.

Now we wish to classify the customer having following attributes {Mb=***99, Ag=midd.age, F= High, Ig= High, Profit/product= Low, G= Male}. will have probability to buy a particular product "P". Now suppose the above table is the table of section from which the product "p" belongs.
Then P(ci) is prior probability of each attribute is computed as follow:

P(buy_the_product = yes)= 12/20 = 0.6
P(buy_the_product = No)= 8/20= 0.4

Calculation of P(X/ci)

P (age = midd.age s/ buy_the_product = yes)= 5/12= 0.416
P (age = midd.age/ buy_the_product = no) = 2/8= 0.25
P (F = high/ buy_the_product = yes) = 5/12= 0.416
P (F = high/ buy_the_product = no) = 1/8= 0.125
P (R = mid/ buy_the_product = yes) = 5/12= 0.416
P (R = mid/ buy_the_product = no) =1/8 = 0.125
P (M= high/ buy_the_product = yes) =3/12=0.25
P (M= high/ buy_the_product = no) = 3/8= 0.375
P (Ig= high/ buy_the_product = yes) = 7/12 = 0.584
P (Ig = high/ buy_the_product = no) = 3/8 = 0.375
P ((profit/product) = low/ buy_the_product = yes)
= 2/12 = 0.167
P ((profit/product) = low/ buy_the_product = no)
= 5/8 = 0.625
P (gender= male/ buy_the_product = yes) = 8/12 = 0.667
P (gender= male/ buy_the_product = no) = 4/8 = 0.5

Hence,
P (**X/ (buy_the_product = yes**)) = 0.416 * 0.416 * 0.416 * 0.25 * 0.584 * 0.167 * 0.667 = 1.1708 x 10^{-3}
P (**X/ buy_the_product = no**) = 0.25 x 0.125 x 0.125 x 0.375 x 0.375 X 0.625 x 0.5 = 1.7167 x 10^{-4}

P (X/ buy_the_product = yes) P (buy_the_product = yes)
= 1.170 x 0.6 = 7.0248 x 10^{-4}
P (X / buy_the_product = no) P (buy_the_product = no)
= 1.7167 x 10^{-4} x 0.4=6.8668 x10^{-5}

Since,
P (X/ buy_the_product = yes) P (buy_the_product = yes)
> P (X/ buy_the_product = yes) P (buy_the_product = yes)
Hence,
The customer will belongs to class buy_the_product = yes.

4 Conclusion

Data mining technique is an important technique which has verities of use in the technology field such as scientific discovery, banking, insurance, decision support and customer relationship management etc. In the current business environment the marketer focus on mass marketing for the promotion of any product. But this way of marketing will not give fruitful result until we know the actual need and preference of the customer. In the proposed model we have provided a framework for classification of profitable customer for the retail marketer. The framework analyse the customer's previous purchase history to identify his/her purchase behaviour. We have used k-mean clustering technique and Naive Bayes classifier for this purpose to identify the class of customer that is potential buyer of a particular product in the retail store. In this, marketer can perform one to one marketing instead of mass marketing. Since as per our assumption if the effort is applied in a particular direction, the sale will increase and this will result in increased profit which is the ultimate goal of the marketer.

Acknowledgments

i specially thanks to my respected guide Mr. Manoj Kumar who has encouraged me to write this paper and helped me in throughout this research work. I thank to Dr. Manuj Darwari my M-tech co-ordinater for there valuable suggestion and motivation. I also thanks Mr. Anurag Kumar my classmate for compiling this paper and giving proper help now finally I thanks my parents for there blessing .

REFERENCES

[1] Michael J. Shaw “Knowledge management and data mining for marketing” Decision Support Systems 31 (2001). 127–137, Elsevier.
[2] Syed Riaz Ahmed “Applications of Data Mining in Retail Business” Proceedings of the International Conference on Information Technology: Coding and Computing 2004 IEEE
[3] E.W.T. Ngai “Application of data mining techniques in customer relationship management: A literature review and classification” Expert Systems with Applications 36 (2009) 2592–2602 Elsevier.
[4] Pei Chao “Identifying the customer profiles for 3c-prpduct retailers: a data mining approach” International Journal of Electronic Business Management, Vol. 6, No. 4, 195-202 (2008).
[5] Marketing Management Millenium Edition, Tenth Edition, by Philip Kotler
[6] Nan-Chen Hsieh “An integrated data mining and behavioral scoring model for analyzing bank customers” Expert Systems with Applications 27 (2004) 623–633, Elsevier
[7] “Introduction to data mining” by Pang-Ning Tan, Vipin kumar, Michael Steinbach.
[8] “Data Mining: Concepts and Techniques”
Second Edition by Jiawei Han and Micheline Kamber