International Journal of Advanced Research in IT and Engineering

ISSN: 2278-6244

USING DATA MINING TECHNIQUES FOR PREDICTIVE MODELLING IN THE RETAIL CONTEXT

Dr Ruchi Mittal*

Abstract: *Based on the observance that rivalry among retailers has increased sharply in recent decades, with a growing number of stores and consequently more options for consumers, the aim of this study was to discover what attributes valued by consumers in their choices on where to shop. The specific focus is on the causal relationship between the attributes offered by grocery stores with loyalty towards these stores. The study is based on statistical data mining approach of predictive model building by regressing factor scores identified by factor analysis. The results identified positive and statistically significant relations between the attributes offered by the retailers and their influence on store loyalty.*

* Associate Professor, Maharaja Agrasen Institute of Management and Technology, Jagadhri

INTRODUCTION - DATA MINING

Data mining is an inter-disciplinary emerging field that focuses on access of information useful for high-level decisions and includes Machine Learning **[L96]**, Statistics and Probabilities **[EP96]**, On Line Analytical processing **[CD97]**, data visualization **[LOQ95]**, information science **[SG91]**, high-performance computing **[FL96]**, etc. Thus data mining is the confluence of multiple disciplines **[M01].** Data mining enables business executives to manage their data and to make relevant decisions. Simply stated, data mining refers to extracting or "mining" of knowledge from large amount of data. **[HK06]**

DATA MINING IN RETAILING

Retail is the main field of application of data mining technology. Through statistical analysis and data mining techniques to analyze relevant data, giving strong support to decision making, has reflected important thinking of data mining in retail **[LSZ07]. [B99]** classified the most effective customers in the marketing activities through data mining. **[MBO00]** also mined customer's credit card data using the RFM model to predict valuable hotel customers. **[YP03]** discussed "pattern-based" clustering approaches to group customer transactions, and presented a new technique, YACA, that generates a highly effective clustering of transactions. **[LW04]** studied RFM analysis, often used by catalog retailers and direct marketers for segmenting customers. **[LL05]** studied the implementation of clustering data mining method to customer analysis of department store to analyze customer characteristics and the relationship between customers and product categories. In a study of 301 grocery shoppers, 10 different super markets, **[M06]** identified 13 retail store attributes, relevant to super market service quality using decision trees due to its visual appeal, simplicity in setting useful rules, and efficiency in classifying rules. **[WCCK06]** identified "valuable travelers" and predicts their "next foreign destination" using data mining techniques. In a study of modeling the relationship between store image, store satisfaction, and store loyalty conducted in Switzerland on around 300 respondents, using hierarchal regression, **[BR98]** proved that a favorable perception of store image leads to store satisfaction which in turn leads to store loyalty. This is also confirmed by **[O93]** who states that customers' patronage behavior towards a particular store is dependent upon the image of a particular store. The review of literature shows that very little research has been done on retailing in general and almost negligible in the area of data-mining in retailing. The

researchers seek to add to this knowledge by researching customers' loyalty behavior in retailing in the Indian context by applying data-mining techniques.

OBJECTIVES OF THE STUDY

1. To identify customer perception of grocery store images.
2. To develop a predictive model of customer store choice for grocery shopping scenario.

RESEARCH METHODOLOGY AND ANALYSIS PATTERN

This study follows the approach followed by **[A2001]; [GK2008] and [E2008]**. They identified various dimensions of service quality & store attributes through factor analysis, and then used these factors as predictor variables to explore the impact they had on customer satisfaction and loyalty as criterion variables. A similar approach was also used by **[ERG08]** to identify service-related factors in the Indian mobile telecommunications market. In their study, data was analyzed in two stages. In the first stage the 32-variables related to service quality were factor analyzed using principal component analysis with varimax orthogonal rotation. They then used service quality factor scores as independent variables in three multiple regression analyses with customer satisfaction, repurchase intentions and recommendation of service to others as dependent variables, respectively.

SAMPLING DESIGN

1. **Universe of the study:** All adult male and female shoppers (above 20 years of age) residing in the National Capital Region (Delhi, Faridabad and Gurgaon), Haryana and Chandigarh.
2. **Survey (Target) Population:** All adult shoppers (above 20 years of age) residing in NCR, Haryana and Chandigarh who could be contacted outside identified retail outlet(s) on the days when the schedule was administered, or those who were willing to respond to the questionnaire at their residence or workplace.
3. **Sample design:** 500 respondents constituted the sample. Non-probability purposive sampling has been employed and only adult individual consumers who shop either for grocery products or for apparels were contacted. Finally, 300 complete questionnaires were received.

SURVEY INSTRUMENT DEVELOPMENT

Measures: This study has considered shopping in the context of Food & Grocery products.

The instrument was designed using scales and store attributes/ store dimensions from previous related research. The survey included the following sections:

- Questions related to respondent demographics- Gender, Age, Occupation, Education, and Monthly Household Income (MHI);
- Questions related to shopping behavior- No. of visits, Expenditure on shopping and preference towards shopping alone or with someone;
- 4 questions measuring store loyalty;
- 23 questions to evaluate grocery store attributes which comprise the store image;

Store Loyalty Scale: The scale consists of four questions covering all aspects of loyalty. The questions, alongwith their origin is stated as follows:

Q1 I think myself as a loyal customer of this store (Origin: Baumgartner and Steenkamp 1996)
Q2 I recommend this store to my family and friends (Origin: Sirohi, McLaughlin and Wittink 1998)
Q3 I make a special effort to shop at this store (Origin: Ailawadi, Neslin and Gedenk 2001).
Q4 A large majority of my grocery purchases are from this store (Origin: De Wulf, Odekerken-Schroder and Iacobucci 2001).

Store Attributes Questions: A total of 23 store attributes have been included in the questionnaire. These attributes have been taken from the "CONSUMER RETAIL STORE IMAGE SCALE" developed by Dickson & Albaum (1997). This scale called CIRS (Consumer Image of Retails Stores) encompasses attitudes towards retail prices, products, store layout and facilities, service and personnel, promotion, and "others" (Dickson & Albaum 1997). The original CIRS had a test-retest reliability of 0.91. According to Hair et al (1998) reliability is used to determine the degree of consistency of a scale. Cronbach's alpha is the most widely used measure of reliability. It measures the consistency of an entire scale. Generally, .70 is an acceptable lower limit, with **.60** being acceptable for exploratory research.

The original scale developed by **[ANG01]** reported a composite reliability of .876 and the validity was found acceptable. For data collection the respondents were asked to provide information on their personal attributes and were requested to rate the importance of the

store attributes in choosing a store. The attributes were measured on a 7-point Likert type scale of importance with 1 being extremely unimportant and 7 being extremely important.

DATA ANALYSIS

The data analysis included the following three stages:

1. Factor analysis was conducted to derive factors and factor scores
2. The reliability of the factors was tested using Cronbach's alpha
3. The factor scores were used in the regression model wherein shopping loyalty would be the criterion variable and the store attribute factors would be the predictor variables.

REGRESSION MODEL:

The theoretical model to be tested can be represented by the following equation:

$$Y = \gamma + \beta 1X1 + \beta 2X2\ldots + \beta nXn$$

Where: Y is Store Loyalty (Criterion Variable) & X1…Xn are Loyalty driving store attributes (Predictor Variables) derived from an Exploratory Factor analysis of various grocery store attributes.

For all analysis, statistical significance was set at a level of .05. The data analysis was done using Factor analysis based on the principal component analysis (PCA). Previous research suggests that store attributes produce factors. Factor analysis is used to reduce the environmental dimension scales into smaller, more manageable factors. This multivariate technique is also used to identify the underlying patterns or relationships for a large number of variables **[H88].** In the present study, Factor analysis was used to summarize the variables by examining correlations between the variables, and to create an entirely new set of variables to replace original variables. Factors were derived using component or principal components, which summarizes the original information into factors for prediction. Only factors with latent roots or eigen values greater than 1 were included. Factors were rotated using the varimax rotation method. According to **[H88],** factor loadings at ±.30 are considered minimal, ± .40 more important, ± .50 or greater practically significant. Items with loadings greater than or equal to ± .50 were retained. However, those with several high loadings on more than one factor, variables with low loadings, and those that did not load on any factor were evaluated for possible deletion. In addition to the variable loading, the communality, total amount of variance shared with other variables was evaluated before deleting the variable. Variables that did not load with communalities less than .50

were deleted. After the factors were formed, they were named according to those variables with higher factor loadings. The 23 store attributes identified were factor analyzed to get these results (Table 1). The Extraction Method used was the Principal Component Analysis (PCA). Rotation Method used was Varimax with Kaiser Normalization. Rotation converged in 7 iterations. This procedure short-lists 33 store attributes for factor analysis out of the original 40. The KMO score is above .50 (0.895) and the Bartlett's test is significant ($\chi 2$ = 6132.84, df = 253). Thus, factor analysis is suitable for this research **[M04], [H88].** The 23 store attributes are factor analyzed to produce several factors. Factors with an eigen value of more than 1 were retained. An eigen value represents the amount of variance associated with the factor. The result was that there were a total of 6 factors (dimensions), which explained for 77.986% of the total variance. The factors considered should together account for more than 60% of the total variance, so 77.986% is a very suitable statistic **[M04].**

Step 1 & 2: Factor Analysis of Grocery Store Attributes & Reliability Analysis of Factors

Factor	Factor Loading	Reliability Cronbach's alpha	Statement
F1: Store Ambience & Layout	.846	.903	This store has a well organized layout
	.645		This is a Spacious Store
	.904		This store is clean
	.697		This store has fast checkout
	.707		This store has good displays
	.855		In this store it is easy to search items
F2: Service and Loyalty Schemes	.648	.938	This store offers very good schemes & sales
	.895		This store has good service
	.921		In this store it is easy to return purchases
	.950		This store has attractive loyalty schemes
F3: Price and Quality	.826	.914	This store has a very good reputation
	.857		This store has good quality products
	.500		This store sells fresh products
	.804		This store has low prices
	.519		This store provides great value for money
F4: One Stop Shopping	.762	.912	This store stocks well known brands
	.827		This store's own products are of good quality
	.752		This store has a vast variety of products
	.827		This store offers everything under one roof
F5: Convenience	.907	.809	This store has a very convenient location
	.911		This store has good parking facilities
	.419		This store offers option to pay by credit/debit card
F6: Salesman	.852s	No estimate	This store has helpful salesmen

Table 1: Grocery Store Factors Statistics

Internal Consistency Reliability: is used to assess the reliability of a summated scale where several items are summed to form a total score. In a scale of this type, each item measures some aspect of the construct measured by the entire scale, and the items should be consistent in what they indicate about the characteristic. This measure of reliability focuses on the internal consistency of the set of items forming the scale. The measure used here is the coefficient alpha or Cronbach's alpha, which is the average of all possible split -half coefficients resulting from different ways of splitting the scale items. An alpha value of 0.6 or less indicates unsatisfactory internal consistency reliability. The Cronbach alpha value in the table --- for all individual factors are well above 0.9, thus indicating a very strong reliability of the factors. The composite reliability of all the factors together consisting of all 23 attributes is 0.930 which is a very strong indicator of reliability.

Step 3: Predictive model of customer store choice for grocery shopping scenario;

In the First objective, the data mining technique used was the principal component analysis (PCA). The reason PCA was:

Firstly, the original 23 attributes were reduced to a more manageable 6 factors/ dimensions which can further describe the consumer perception of grocery store images. Secondly, PCA enables computation of exact factor scores for each of the 300 respondents for use in the subsequent development of a predictive model of customer store choice.

Thus factor analysis has a capacity to transform the original data into a new set with a reduced dimension space which can be used to calculate composite variables (factors) for further multivariate analysis such as multiple regressions used in predictive modeling.

In this study, factor scores (defined as composite scores estimated for each respondent on the derived factors) have been computed for each respondent where a factor score for the *i*th factor may be estimated as follows:

Fi = Wi1X1+Wi2X2+Wi3X3+---------WikXK

Where Fi = estimate of ith factor, Wi = weight of factor score coefficient, K = number of variablesIn this study the factor scores have been estimated using the regression method of saving variables. The regression method is a method for estimating factor score coefficients. The scores produced have meant of 0 and a variance equal to the squared multiple correlation between the estimated factor scores and the true factor values. The scores may be correlated even when factors are orthogonal. The factor scores are now used to create a

predictive model using multiple regression analysis where the general form of the multiple regression models is as follows:

$Y= \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X4 + \ldots\ldots\ldots\ldots \beta kXk + e$ which is estimated by:

$Y= a + b1X + b2X2 + b3X3\ldots\ldots\ldots\ldots +bkXk$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.817(a)	.667	.661	.714	.667	97.994	6	293	.000

A Predictors: (Constant), REGR factor score 6 for analysis 1, REGR factor score 5 for analysis 1, REGR factor score 4 for analysis 1, REGR factor score 3 for analysis 1, REGR factor score 2 for analysis 1, REGR factor score 1 for analysis 1

B Dependent Variable: Store Loyalty.

Interpretation:

This table displays R, R squared, adjusted R squared, and the standard error.

R, the multiple correlation coefficients, is the correlation between the observed and predicted values of the dependent variable. The values of R for models produced by the regression procedure range from 0 to 1. Larger values of R indicate stronger relationships. The value of R is 0.817 which indicates strong relationship.

R squared, the coefficient of multiple determination, is the proportion of variation in the dependent variable explained by the independent variables in the regression model. Small values indicate that the model does not fit the data well. The R squared value of 0.667 indicates that the data fits the model very well.

The sample R squared tends to optimistically estimate how well the models fits the population. Adjusted R squared attempts to correct R squared to more closely reflect the goodness of fit of the model in the population. The adjusted R squared value of 0.661 indicates that 66.10 % of the variance was explained by the model.

ANOVA (b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	299.444	6	49.907	97.994	.000(a)
	Residual	149.223	293	.509		
	Total	448.667	299			

a Predictors: (Constant), REGR factor score 6 for analysis 1 , REGR factor score 5 for analysis 1 , REGR factor score 4 for analysis 1 , REGR factor score 3 for analysis 1 , REGR factor score 2 for analysis 1 , REGR factor score 1 for analysis 1

B Dependent Variable: STORELOYALTY

The ANOVA table explains the variation that is accounted for in the model. In this case the significance value of 'F' statistic is .000, which shows that the independent variables do a good job explaining the variation in the dependent variable.

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	**(Constant)**	**4.867**	.041		118.116	**.000**					
	Ambience/ Layout	**.236**	.041	.193	5.729	**.000**	.193	.317	.193	1.000	1.000
	Service & Loyalty Schemes	**.533**	.041	.435	12.914	**.000**	.435	.602	.435	1.000	1.000
	Price & Quality	**.566**	.041	.462	13.705	**.000**	.462	.625	.462	1.000	1.000
	One Stop Shopping	**.342**	.041	.279	8.289	**.000**	.279	.436	.279	1.000	1.000
	Convenience	**.449**	.041	.367	10.886	**.000**	.367	.537	.367	1.000	1.000
	Salesmen	**.151**	.041	.123	3.648	**.000**	.123	.208	.123	1.000	1.000

a Dependent Variable: STORELOYALTY

The unstandardized coefficients are the coefficients of the estimated regression model. The t statistics can help you determine the relative importance of each variable in the model. Values of 't' normally higher than 2 indicate that the variable is very important. A perusal of table--- provides information on the confidence with which the support can be estimated for each factor. Since the value of't' is more than 2 for all factors and the significance values of't' statistic are is very small (.000) it can be assumed that the beta values denoted by 'B' are true at a 95% level of confidence. The higher the value of beta the more important the factor is in predicting the dependent variable. The predictive model of store loyalty based on store image variables can now be described in light of the above tables in the following terms:

Store Loyalty = 4.867 + .236*Ambience + .533*Service + .566*Price + .342*OSS + .449 *Convenience + .151*Salesmen

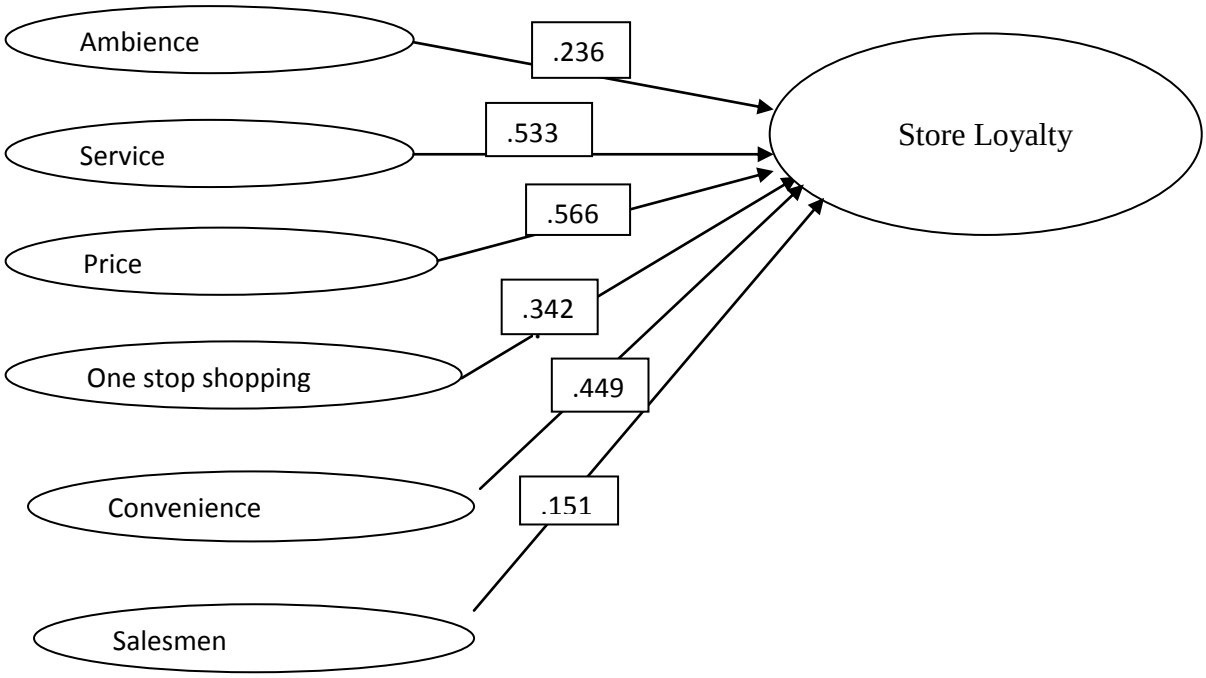

Figure 1: Grocery Store Preferences in an Evolving Market [MM08]

To estimate the reliability of the factors the Cronbach's alpha has been calculated to check the relationships between items in the scale. Cronbach's alpha is one of the most popular and reliable model of internal consistency of a scale, based on the average inter-item

correlation. The Cronbach's alpha scores have been mentioned the third column of Table no.2.

FINDINGS AND CONCLUSION

The statistical data mining techniques used in this study are factor analysis, reliability analysis followed by multiple regressions. The six factors (vide Table 1) having significant values for predicting store loyalty are Ambience, Service, Price, OSS, Convenience and Salesmen. Taking a clue from the proposed store choice model of **[SB04], [MM08]** - Framework for examining store preferences in an evolving market- the factors can be called "store loyalty drivers". To ensure primary store loyalty grocery retailers must concentrate on providing the "loyalty drivers" in a suitable combination called "Store Format". As per the model developed the most important loyalty driver, in the context of grocery retailing, is pricing followed by Service.

REFERENCES

[1] **[MM08]** Mittal and Mittal (2008), Store Choice in the Emerging Indian Apparel Retail Market: An Empirical Analysis, IBSU Scientific Journal (Georgia), 2(2), 21-46 Available at: www.ibsu.edu.ge/journal/index.php/ibsu/article/view/57/0

[2] **[ERG08]** Eshghi, A., Roy, S.K. and Ganguli, S. (2008) "Service Quality and Customer Satisfaction: An Empirical Investigation in Indian Mobile Telecommunicaions Services", Fall, pp 119-144

[3] **[SG91]** P. Smyth and R.M. Goodman. Rule induction using information theory. In G. Shapiro and J. Frawley (eds.). Knowledge discovery in databases, pages 159-176, Cambridge, MIT Press, 1991.

[4] **[LOQ95]** H.Y. Lee, H.L. Ong and L.H. Quek. Exploiting visualization in knowledge discovery. Proc. 1st Int. Conf. Knowledge discovery and data mining, AAAI, pages 198-203, 1995.

[5] **[EP96]** J.F. Elder and D. Pregibon. A statistical perspective on knowledge discovery in databases. In U.M. Fayyard et al. (eds.). Advances in knowledge discovery in data mining, pages 83-113, AAAI/ MIT Press, 1996.

[6] **[FL96]** A.A. Freitas & S.H. Lavington. Parallel data mining for very large relational databases. In H. Liddel et al. (eds.). *Proc. Int. Conf*. High performance computing and networking (HPCN'96), Brussels, Belgium, pages 158-163, Springer, 1996.

[7] **[L96]** P. Langley. Elements of Machine Learning. Morgan Kaufmann, 1996.

[8] **[CD97]** S. Chaudhuri and U. Dayal. An overview of datab warehousing and OLAP technology. ACM-SIGMOD Record, 26, pages 65-74, 1997

[9] **[B99]** S. Bhattacharya. Direct marketing performance modeling using genetic algorithms. INFORMS Journal on Computing, 11(3), pages 248-257, March 1999.

[10] **[MBO00]** A.M. Morrison, G. Bose and J.T. O'leary. Can statistical modeling help with data mining? A database marketing application for U.S. Hotels. *J.* Hospitality & Leisure Marketing, 6(4), pages 91-110, 2000.

[11] **[YP03]** Y. Yang and B. Padmanabhan. Segmenting customer transactions using a pattern-based clustering approach. In Proc. *Third IEEE Int. Conf.* Data Mining (ICDM'03), pages 411-418, November 19-22, 2003.

[12] **[LW04]** M. Levy and B. Weitz. Retailing Management, Tata Mc Graw Hill, 2004.

[13] **[LL05]** W. Liu Y. Luo. Applications of clustering data mining in customer analysis in department store. In *Proc*. IEEE *Int. Conf*. Services Systems and Services Management, Vol. 2, pages 1042-1046, June 13-15, 2005.

[14] **[M06]** H. Min. developing the profiles of supermarket customers through data mining. The Service Industries Journal, 26(7), pages 747-763, October 2006.

[15] **[M01]** N. Melab. Data Mining: A key contribution to e-business. Information & Communications Technology Law, Taylor & Francis group, pages 309-318, 10(3), October 2001

[16] **[LSZ07]** Liu, H. Su, B. Zhang and Bixi. The Applications of Association Rules in Retail Marketing Mix. In Proc IEEE Int. Conf. Automation and Logistics, pages 2514-2517, Jinan, China, August 18-21, 2007.

[17] **[HK06]** J. Han and M. Kamber. Data Mining: Concepts and Techniques, 2/e, Elsevier publisher, 2006

[18] **[WCCK06]** J.Y. Wong, H.J.Chen, P.H. Chung and N.C. Kao. Identifying valuable travelers and their next foreign destination by the application of data mining techniques. Asia Pacific *J.* of Tourism Research, Vol. 11, No. 4, pages 355-373, December 2006.

[19] **[BR98]** Bloemer, J. and Ruyter, K.(1998), A on the Relationship Between Store Image, Store Satisfaction and Store Loyalty, European Journal of Marketing, 32(5/6), 499-513.

[20] **[H98]** Hildebrandt, L. (1988), Store Inage and the Prediction of Performance in Retailing, Journal of Business Research, Elsevier, Vol. 17(1), pages 91-100.

[21] **[ERG08]** Eshghi, A., Roy, S.K. and Ganguli, S. (2008) "Service Quality and Customer Satisfaction: An Empirical Investigation in Indian Mobile Telecommunicaions Services", Fall, pp 119-144

[22] **[M04]** Malhotra, N. Marketing Research, Prentice Hall, 2004.

[23] **[SB04]** Sinha, P.K.and Banerjee, A. (2004) "Store Choice Behavior in An Evolving Market", International Journal of Retail and Distribution Management, Vol. 32(10), pp 482-494.