International Journal of Computer Applications Technology and Research
Volume 4– Issue 12, 884 - 888, 2015, ISSN: 2319–8656

Analyzing Customer Behaviour through Data Mining

Sandeep Kumar
Department of Computer Science
Krishna Engineering College
Ghaziabad, UP, India

Rakesh Kumar Arora
Department of Computer Science
Krishna Engineering College
Ghaziabad, UP, India

Abstract: Indian organized retail industry is poised for growth. In order to attract customers price discounts in retail sector are the norm rather than an exception. The most common type of monetary promotions includes discounts, coupons, and rebates. Promotions helps organizations to grow market share, increase sales volume, sell faster, cultivate loyal customers, and drown out competitor advertising. With the widespread use of sales promotions, it has become important for the managers to understand such practices and understand challenges.

This paper will assist the manager in identifying the set of customers that are attracted towards the departmental store when the discount coupons were issued. The manager will also be able to formulate new policies to attract new set of customers.

Keywords: Data Mining, Retail Sector, Decision Tree, WEKA

1. INTRODUCTION

A large number of retail chains have opened over the last decade with the objective of providing quality products at low prices to the customers. Fierce competition and narrow profit margins have pushed retailers in implementing data warehouse earlier than other industries. This has given the retailers a better opportunity to take advantage of data mining.

Large retail chains and grocery stores store vast amounts of data collected over the years that are rich in information. Data Mining helps in reducing information overload along with the improved decision-making by searching for relationships and patterns from the huge dataset collected by retailers. It enables a retail industry to focus on the most important information in the database and allows retailers to make more knowledgeable decisions by predicting future trends and behaviors. The Data Mining uses the business data as raw material using a predefined algorithm to search through the vast quantities of raw data, and group the data according to the desired criteria that can be useful for the future target marketing.[1]

This paper uses Data Mining Technique to improve the sales in the departmental store by distribution of coupons among customers visiting the departmental store such that both customers and departmental stores can gain because of increased sales volume. Data mining, the extraction of hidden predictive information from large databases is a powerful technology with great potential to help managers in the departmental stores to have larger market share and cultivate loyal customers. It discovers information within the data that queries and reports can't effectively reveal. After gathering data regarding customer profiles submitted by customers at the time of issuing of coupons and feedback form filled by the customers at the time of redemption of coupons, data mining technique need to be applied to identify set of customers that will help in increase sales volume and market share.

With the help of data mining techniques, such as clustering, decision tree or association analysis it is possible to discover the key characteristics from the details of customers and possibly use those characteristics for future prediction. This paper presents decision tree algorithm as a simple and efficient tool to analyze the customer details and distribution of coupons for higher sales and larger market share.[2]

2 . METHODOLOGY

Decision trees are a simple, but powerful form of multiple variable analysis. A decision tree is a special form of tree structure. The tree consists of internal nodes where a logical decision has to be made, and connecting branches that are chosen according to the result of this decision. The nodes and branches that are followed constitute a sequential path through a decision tree that reaches a leaf node (final decision) in the end.[3]

In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. In the case of numeric attributes, the condition refers to a range. Each leaf is assigned to one class representing the most appropriate target value. [4]

The decision tree algorithm is simple top down greedy algorithm. The major step of algorithm is to continue to divide leaves that are not homogeneous into leaves that are as homogeneous as possible until no further division is possible. The algorithmic steps for decision tree algorithm is as follows:[5]

1. Let the set of training data be S. If some of the attributes are continuous-valued, they should be discretized. Once that is done, put all of S in single tree node.
2. If all the instances in S are in same class, then stop.
3. Split the next node by selecting an attribute A from amongst the independent attributes that best divides or splits the objects in the node into subsets and create decision tree node.
4. Split the node according to the values of A
5. Stop if any of the following conditions are met, otherwise continue with step 3

Figure 1: Steps for Decision Tree Algorithm

Pruning is very important technique to be used in tree creation because of outliers. It also addresses overfitting. Datasets may contain little subsets of instances that are not well defined. To classify them correctly, pruning can be used. There are two types of pruning:

1. Post pruning (performed after creation of tree)
2. Online pruning (performed during creation of tree) [6].

The steps to extract classification rules from tree are mentioned below:

1. Represent the knowledge in the form of IF-THEN rules.
2. One rule is created for each path from the root to a leaf.
3. Each attribute-value pair along a path forms a conjunction.
4. The leaf node holds the class prediction

The analysis using decision tree is being done with the help of WEKA tool. WEKA, formally called Waikato Environment for Knowledge Learning supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. WEKA is an open source application that is freely available under the GNU general public license agreement. Originally written in C the WEKA application has been completely rewritten in Java and is compatible with almost every computing platform. It is user friendly with a graphical interface that allows for quick set up and operation. WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values. The WEKA application allows novice users a tool to identify hidden information from database and file systems with simple to use options and visual interfaces. [7]

3. ANALYSIS

The study was carried out on the profiles of the customers who have visited the departmental store from July 2014 to Dec 2014. The attributes considered for analysis of customers along with their description are reflected in Table 1.

Parameters	**Description**
Sex	Male (M) /Female (F)
Age_Group	Less than and equal to 20(Y) / More than 20 and less than equal to 30(L) / More than 30 and less than equal to 40(M) / More than 40 and less than equal to 50(N) / Above 50(O)
Professsion	Salaried (S) / Businessman (B)
Qualification	Under graduate (U) / Graduate (G) / Post graduate (P)
Income	Less than and equal to 25,000/- (Low) / More than 25,000 and less than 50,000/- (Medium) / More than 50,000/- (High)
Coupon Utilized	Yes / No

Table 1: Parameters used for analysis

.

	A	B	C	D	E	F
1	Sex	Age_Group	Profession	Qualification	Income(Rs)	Coupon Utilized
2	M	Y	S	U	Low	Y
3	F	L	S	G	Low	Y
4	F	M	B	P	Low	Y
5	M	L	S	U	Low	Y
6	M	O	B	G	Low	N
7	F	O	S	P	Low	N
8	F	M	B	U	Low	N
9	M	L	S	G	Medium	Y
10	F	M	B	P	Medium	N
11	M	Y	S	U	Low	Y
12	M	M	B	G	Medium	N
13	M	L	S	P	Low	Y
14	F	L	B	U	Low	N
15	F	L	S	G	Low	N
16	F	M	B	P	Low	Y
17	M	M	S	U	Low	Y
18	M	N	B	G	Medium	N
19	M	N	S	P	Low	Y
20	M	Y	B	U	Low	Y
21	F	O	S	G	Low	Y
22	F	O	B	P	Low	N
23	F	Y	S	U	Low	N

Figure 2: Sample view of Dataset

The data file normally used by WEKA is in ARFF (Attribute-Relation File Format) file format, which consist of special tags to indicate different things in the data file. Figure 2 shows the sample view of dataset and Figure 3 shows the ARFF format of desired dataset. To convert an Excel format into ARFF format an Excel to ARFF convertor is being used. The ARFF format dataset is represented in Figure 3

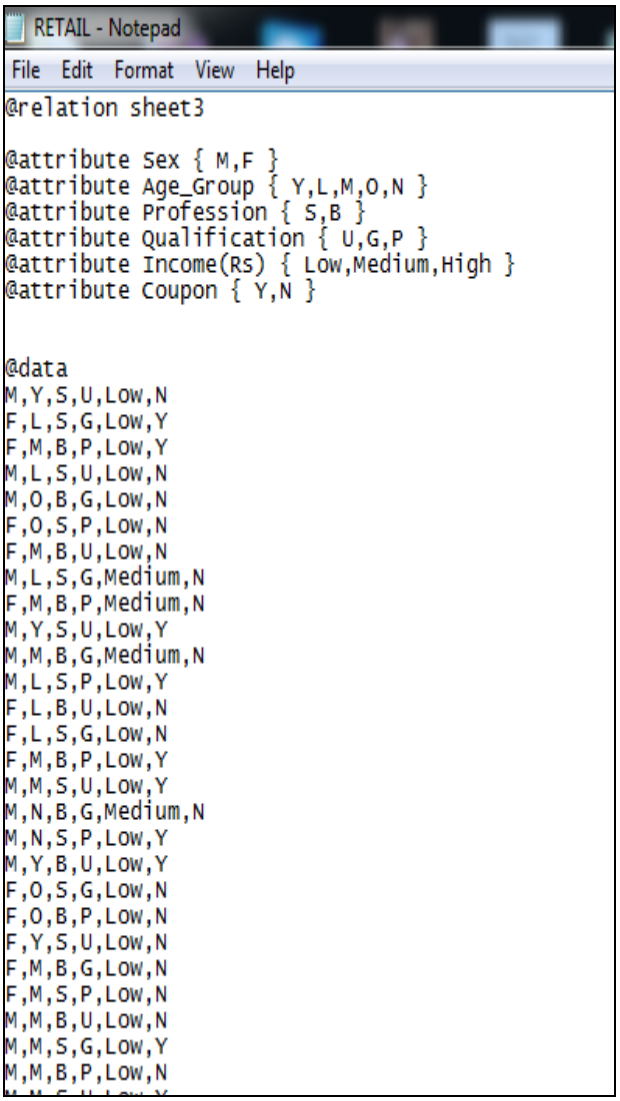

Figure 3: ARFF Format of Sample Dataset

After collecting and cleaning the data, the classification of data is done using J48. J48 is an open source Java implementation of the C4.5 algorithm in the WEKA data mining tool. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The output generated is displayed in Figure 4.

Figure 4: Output

The accuracy is around 58%. The kappa statistic measures the agreement of prediction with the true class where value 1.0 signifies complete agreement. The confusion matrix or contingency table in this example has thirteen classes, and therefore a 13x13 confusion matrix is being displayed. The number of correctly classified instances is the sum of diagonals in the matrix; all others are incorrectly classified. The True Positive (TP) rate is the proportion of examples which were classified as class x, among all examples which truly have class x, i.e. how much part of the class was captured. It is equivalent to Recall. The False Positive (FP) rate is the proportion of examples which were classified as class x, but belong to a different class, among all examples which are not of class x. The Precision is the proportion of the examples which truly have class x among all those which were classified as class x. The F-Measure is simply 2 * Precision * Recall / (Precision + Recall), a combined measure for precision and recall.[8]

As per J48 Algorithm, parameters that reflect noise or outliers need to be removed, hence only those targeted node are shown by tree which have some value of precision and recall. The tree generated is represented in Figure 5.

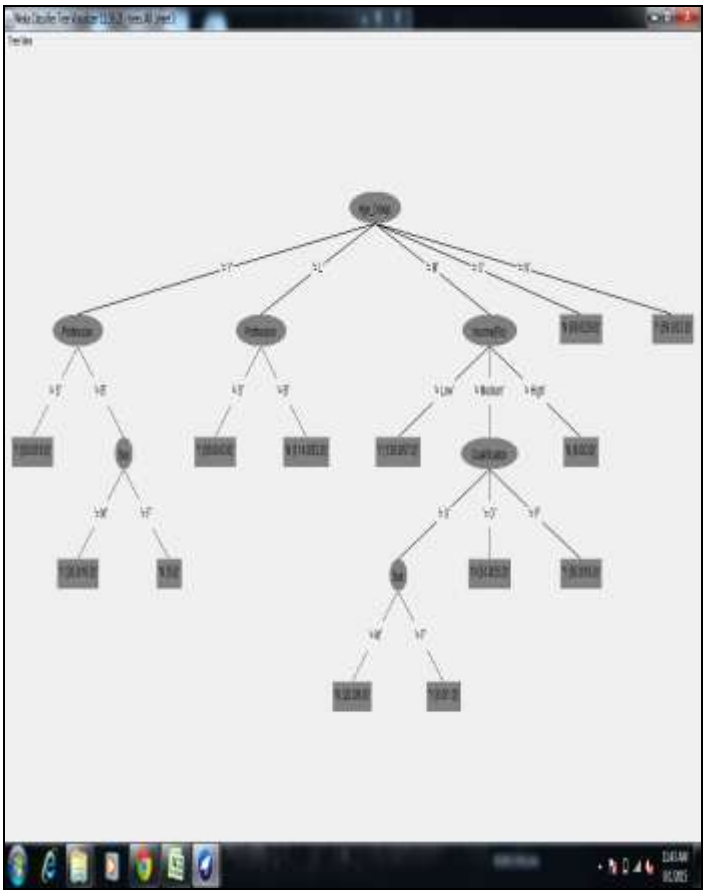

Figure 5: Decision Tree

5. RESULTS

The classification rules extracted from tree when the coupon is utilized are:

1) If Age_group is less than 20 and Profession is Service than coupon is utilized.
2) If Age_group is less than 20 ,Profession is Buisnessman and Sex is Male than coupon is utilized.
3) If Age_group is between 20 and 30 and Profession is Service than coupon is utilized.
4) If Age_group is between 30 and 40 , Income is than equal to Rs 25,000/- than coupon is utilized.
5) If Age_group is between 30 and 40 , Income is more than Rs 25,000/- and less than Rs 50,000/- , Qualification is under graduate and Sex is Female than coupon is utilized.
6) If Age_group is between 30 and 40 , Income is more than Rs 25,000/- and less than Rs 50,000/- and Qualification is post graduate than coupon is utilized.

7) If Age_group is between 40 and 50 than coupon is utilized.

Disadvantages of J48 algorithm: The run-time complexity of the algorithm matches to the tree depth, which cannot be greater than the number of attributes. Tree depth is linked to tree size, and thereby to the number of examples. So, the size of C4.5 trees increases linearly with the number of examples. C4.5 rules slow for large and noisy datasets Space complexity is very large as we have to store the values repeatedly in arrays [9].

6. CONCLUSION

In this paper, a simple methodology based on decision tree algorithm is being used to analyze the customer details for distribution of coupons. This methodology will assist the Managers in the store to identify set of customers that are likely to purchase from store. The manager will also be able to devise the new schemes in order to attract new set of customers. This model will play important role in understanding demographics of customers by clearly differentiating between the customers that need to retained and that need to be targeted . This will have significant effect in improving sales and hereby achieving targets of departmental store.

7. REFERENCES

[1] Ahmed, S. R. (2004), 'Applications of Data Mining in Retail Business', Proceedings of the International Conference on Information Technology: Coding and Computing, Vol.2, pp. 455- 459 IEEE

[2] Arora K. Rakesh, Badal Dharmendra," Admission Management using Data Mining using WEKA", IJARCSSE Vol. 3, Issue 10, October 2013

[3] [Online]http://www.estard.com/decisiontree/decision_trees_definition.asp

[4] [Online]http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf

[5] Gupta K G. "Introduction to Data Mining with case studies" , PHI

[6] Moertini, Veronica S. "Towards the use of C4.5 algorithm for classifying banking dataset." Vol. 8 No. 2, October 2003 (2003): 12. Web. 24 Jan. 2013

[7] [Online]http://www.gtbit.org/downloads/dwdmsem6/dwdmsem6lman.pdf

[8] [Online] http://weka.wikispaces.com/Primer

[9] Juneja, Deepti, et al. "A novel approach to construct decision tree using quick C4.5 algorithm." Oriental Journal of Computer Science & Technology Vol. 3(2), 305-310 (2010) (2010): 6. Web. 18 Feb. 2013.

[10] Arora K. Rakesh, Badal Dharmendra," Subject Distribution using Data Mining ", IJRET Vol. 2, Issue 12, December 2013

[11] Arora K. Rakesh, Badal Dharmendra," Placement Prediction through Data Mining ", IJARCSSE Vol. 4, Issue 7, July 2014.

[12] [Online]http://www.gallimoreinc.com/learn/how-coupons-help-sell-more-product.php