CUSTOMER PUZZLED BEHAVIORAL ANALYSIS - A STEP TOWARDS VALUING CUSTOMER’S INTERESTS

T.N. Natesh  
Department of Commerce,  
Dayananda Sagar University, Bengaluru, India

N. Shobha Rani  
Department of Computer Science,  
Amrita School of Arts and Sciences, Mysuru,  
Amrita Vishwa Vidyapeetham, India

ABSTRACT

In this paper, an attempt is made towards interpretation of customer’s buying patterns which is most widely useful for improving their satisfaction during their shopping in large shopping malls or stores. One of the significant aspects focused in this work is to value the customer needs and improve the level of customer satisfaction through analyzing their varied purchase patterns from time to time. The prediction of correlation between the type of items the customer’s opting in terms of various factors like brand, model, cost, their budget, demographics attributes, frequently shopped products and regular/consistent brands etc, the organization can make better attempts on designing the strategies for alignment of products/items in their shelves so that customer fulfill their needs with minimal risk. Also, the organization can improve the productivity of items/products which are at high demand along with modeling of new marketing strategies. The prediction of frequently sold products of certain brands can be used as basis for design of attractive offers assisting the sales of non frequently used brands so that the demand can be improved. Datasets on whole customer information is employed for experimentation comprised of 8 attributes including region, channel and various other food categories and about 440 instances. The visualization of datasets is carried out using Weka tool and overall summarization of data is carried out using attribute wise with respect to various food item categories.

Key words: Customer data analysis, visualization, data summarization , valuing customers and behavioral analysis.

http://www.iaeme.com/IJMET/issues.asp?JType=IJMET&VType=9&IType=7
1. INTRODUCTION

Most of the large scale organizations today are unable to spend enough amount of time on valuing the customer needs and their interests due to the pace at which contemporaries are moving in the business world. It is obvious that organizations spend most of their time in analysis, planning and implementation of various marketing campaigns, response statistics and various other aspects relating to improving the productions and demand of less preferred brands etc. Although there exists the usage of smart technologies at each and every point of business process, some aspects relating to the customer value additions are short sighted by organizations. The impact and influence of this significant aspect may lead to reduction of customer satisfaction and that in turn may drop the customer’s interest towards a particular brand or organizations. Therefore, it is very much essential to focus on valuing the needs of customers by analyzing the various past requirements and purchase history of the customers.

Usually, the most common factors addressed in connection with this predictive task include customer value, propensity to spend, behavior/attitude and other demographics [1]. Emphasizing on the various customers values assists in pre determination of statistics like marketing budget such as the instances on which the money needs to be spent [2] and where it is not, setting rankings based on their frequency and magnitude of purchases [3], determining more interested or prospective customers based on their wallet analysis [4][5] etc. Lifetime Value framework [6] aims to predict profits obtained through a particular customer based on their interests towards a particular brand. The other which closely associates with lifetime value framework is the propensity model [7] which helps in figuring out the ways to prioritize the customers by investing the organization resources on the high demand brands in campaign; Also, the model reduces the cost spent in domestic needs like post or couriers etc. further the model produces all the required statistic sassed on the analysis of past purchasing history, demographics, life style of the customers. The other methods that focus on valuing the customers is mail communications [8] on the new releases of a particular brand and compelling them to grab the offers. In the past, numerous investigations are carried out on analysis of customer interests and valuing customers, details of some of the important contributions are as summarized below.

Marko et al [9] had investigated on various analytical techniques for marketing, sales and after sales challenges in the automotive industry through grouping of past purchase history of data using clustering technique. Lawrence et al [10] had proposed a personalized recommender system designed to suggest new products to supermarket customers which are apt for working in a computing environment through the use of smart devices to prepare and submit their orders. These systems are most eliminates the needs of directly visiting the stores for purchases/shopping which on the other hand is a disadvantage for offline shopping stores. Desarbo et al [11] had developed a statistical approach for customer value analysis through recursive simultaneous equation model comprising buyer heterogeneity through which relative effects and integration rules of perceived value drivers can be estimated at the market segment level. Punj et al [12] had intensively surveyed the types of clustering techniques that should be applied on the customer and marketing research data and also provided the suggestions for employing clustering on various applications of marketing and academic research data. Berne et al [13] had analyzed the negative role of variety seeking on customer retention for services through testing the hypothesis with structural equation modeling and also study is carried out on food service at three Universities. The results obtained indicate variety seeking negatively affects customer retention and reduces the impact of the management efforts to improve service quality and customer satisfaction. Burez et al [14] had investigated performance increase of both random and advanced under-sampling along with
gradient boosting and weighted random forests modeling techniques in prediction of customer churn. The metrics like AUC and lift are employed for evaluation which had produced better overall prediction accuracy.

Min et al [15] had proposed a methodology for detecting a user's time-variant pattern to improve the performance of collaborative filtering recommendations comprising of three phases profiling, detecting changes, and recommendations. The prediction of changes in customer behavior at different periods of time is carried out which had resulted in improved performance in recommendations. Srinivasan et al [16] had suggested a new technique for measuring, analyzing, and predicting a brand’s equity in a product market by considering sources of brand equity, brand awareness, attribute perception biases, and non-attribute preference as essential parameters. The results indicate good face validity and convergent validity, with brand awareness playing the role, followed by non-attribute preference. Park et al [17] had developed a customer profile model for improving quality of the recommendation system using individual and group behavior information such as clicks, basket insertions, purchases, and interest fields. The evaluation of recommendation performance had provided better recommendation performance than existing models. Schmittlein et al [18] had performed customer base analysis based on past purchase history to identify number of active customers, change in number over time, customers who are most likely still active, the duration the customer likely to remain an active customer, and expected purchases from each in future. The work is validated by applying stochastic models to industrial purchase processes and industrial marketing decisions which indicate effectiveness in predicting purchase patterns and in generating insights into how key customer groups differ.

Most of the works in the literature focus on capturing the customer interests and churning the predictions in various areas of marketing. Few of the works focused on enhancement of functionalities in the recommendation systems for customers where as few works are directed towards addressing the issues of managerial authorities in understanding the customer interests and behavior during shopping. The strategies developed towards valuing customer interests by companies are restricted to classification and clustering tasks and also analyzing correlations between various customers purchase history through statistical and stochastic methods. In the proposed work, we focus on analyzing the behavior of customer buying patterns by understanding correlations between different types of items or products, types of brands, their demographics, thresholds of one time shopping, shopping cycles/frequencies etc through the techniques of association rule mining and rule based learners.

The subsequent sections focus on depicting the details of various data mining approaches used in the proposed work for prediction of customers puzzled behavior during shopping.
2. APPROACHES EMPLOYED FOR CUSTOMER BEHAVIOR PREDICTION

The prediction of customer behavior with special reference to the data available in purchase history of large shopping malls is carried out in two different ways. One is collection of data from customers which comprises of attributes like demographic attributes, behaviors, preferences, their budget, buying habits, frequently used items etc and analyzing correlations between them. The correlations between the various attributes is usually carried out by one of the widely used analysis technique which is association rule mining [21], on the other hand the correlations between the attributes can also be analyzed using the Chi-square test [22], Karl pearson correlation coefficient [23] etc. The primary purpose of analyzing these data is find some interesting and meaning purchase patterns from the purchase history of a customer. Further, these patterns will be helpful in designing marketing strategies which values customer interests greatly. Individual attention on each customers buying habits can be interpreted clearly leading to better productivity of the brands/products in demand. Few instances of interesting patterns include purchase frequencies, sequences, spending, number of items and inter-order times etc.

The derivation of all the above patterns is accomplished in relation with advanced data analytics tools and techniques. Some of the prominent techniques which has significant role in this process is as discussed subsequently.

2.1. Data Mining

Data Mining is the process of analyzing the huge volumes of data in order to predict the interesting data patterns which can help in various managerial decision making activities [24]. The process of mining for interesting patterns comprises the sequence of stages as depicted in figure 1.

The variety of mid level operations in this process will integrate the predictive capabilities from the known data which is to be helpful in deducing new data. The predicted patterns are further given a descriptive form suitable for easy analysis of results obtained. Some of the essential techniques used for prediction of unknown data from known data includes, regression, association rule mining, classification and clustering.

2.1.1. Association Rule Discovery

Association rule is a combination of two counterparts, which depicts the probabilistic relationship of first counterpart with second one. Mostly the two counterparts indicates the frequent combination in the database of the concerned domain. The process of discovery is an intelligent analysis of the determining rules which involves identifying various frequently occurring combinations of items in the database.
The strength of the rules discovered is analyzed using two measures, support and confidence respectively. Support represents the frequency of occurrence of the items in a specific rule where as the confidence is the number of times the item 2 is taken away along with item 1.

Let $P_1$ and $P_2$ are the items in the database, if $N(T)$ is the number of instances containing both $P_1$ and $P_2$, $N(P_1)$ is the instances in the database containing only items $P$ and $N$ is the total number of transactions, then $S$ and $C$ is the support and confidence of the items which are as given by equations (1) and (2).

\[
\text{Support } S = \frac{N(T)}{N} \quad (1)
\]

\[
\text{Confidence } C = \frac{N(T)}{N(P_1)} \quad (2)
\]

### 2.1.2. Classification

Classification focuses on the finding the similarities between the data and modeling them into classes so that the data objects within a particular class will possess high similarity. Basically, it is a process of supervised learning which involves assumption prior knowledge on the number of class types and characteristic properties of each class type. The decision of assigning each data object to particular class involves the definition of logical predicates with respect to each class. The logical predicate is usually a set of rules that suffices the decision making process of assigning of data objects to a class.

### 2.1.3. Clustering

Though the sole purpose of classification and clustering is to categorize the objects based on their similar characteristics, the clustering differs in its approach initially. Unlike classification, clustering does not require any prior assumptions or inputs on the number of class types and its characteristics. Clustering is generally applied on the huge volume of data with extremely variant characteristics from one data object to the other, therefore the clustering is termed as unsupervised learning approach [25]. These approaches shall not require any initial inputs or assumptions and performs the grouping of data based on interpretation of properties of each individual object and further converting into specific groups with respect to the characteristics the object possess.

Basically, the classification or clustering is performed in two important considerations; one is the training set which is used to build the knowledge model, while the other one is the test set used to validate it. The performance of the model built is determined on basis of validation conducted on the test set.

### 2.1.4. Regression

Regression is the most straightforward, simple, version of what we call “predictive power.” When we use a regression analysis we want to predict the value of a given (continuous) feature based on the values of other features in the data, assuming a linear or nonlinear model of dependency [29].

Regression techniques are very useful in data science, and the term “logistic regression” will appear almost in every aspect of the field. This is especially the case due to the usefulness and strength of neural networks that use a regression-based technique to create complex functions that imitate the functionality of our brain.
3. DATASETS

3.1. Wholesale customers Data Set

The data set comprises the clients of a wholesale distributor [26] [27][28] which includes the annual spending in monetary units (m.u.) on diverse product categories. The attribute information is as follows.

**Attribute Information:**
- **FRESH** annual spending (m.u.) on fresh products (Continuous)
- **MILK** annual spending (m.u.) on milk products (Continuous)
- **GROCERY** annual spending (m.u.) on grocery products (Continuous)
- **FROZEN** annual spending (m.u.) on frozen products (Continuous)
- **DETERGENTS_PAPER** annual spending (m.u.) on detergents and paper products (Continuous)
- **DELICATESSEN** annual spending (m.u.) on delicatessen products (Continuous)
- **CHANNEL** Customers channel - Horeca (Hotel/Restaurant) or Retail channel (Nominal)
- **REGION** Customers region (Lisbon, Oporto or Other) (Nominal)

The central tendency quantities of the datasets are as tabulated in table 1.

**Table 1** Descriptive statistics of whole sale customers dataset

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fresh</td>
<td>3</td>
<td>112151</td>
<td>12000.30</td>
<td>12647.329</td>
</tr>
<tr>
<td>Milk</td>
<td>55</td>
<td>734498</td>
<td>5796.27</td>
<td>7380.377</td>
</tr>
<tr>
<td>Grocery</td>
<td>3</td>
<td>92780</td>
<td>7951.28</td>
<td>9503.163</td>
</tr>
<tr>
<td>Frozen</td>
<td>25</td>
<td>60869</td>
<td>3071.93</td>
<td>4854.673</td>
</tr>
<tr>
<td>Detergents_paper</td>
<td>3</td>
<td>40827</td>
<td>2881.49</td>
<td>4767.854</td>
</tr>
<tr>
<td>Delicatessen</td>
<td>3</td>
<td>47943</td>
<td>1524.87</td>
<td>2820.106</td>
</tr>
</tbody>
</table>

The frequency of the various items as specified in table 1 for the various regions for around le 440 instances is as depicted in table 2.

**Table 2** Region wise statistics of items in table 1.

<table>
<thead>
<tr>
<th>Region</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lisbon</td>
<td>77</td>
</tr>
<tr>
<td>Oporto</td>
<td>47</td>
</tr>
<tr>
<td>Other region</td>
<td>316</td>
</tr>
</tbody>
</table>

**Table 3** Channel wise statistics of items in table 1

<table>
<thead>
<tr>
<th>Channel</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horeca</td>
<td>298</td>
</tr>
<tr>
<td>Retail</td>
<td>142</td>
</tr>
</tbody>
</table>
Performing data analytics on the dataset may result in various benefits such as identification of repeated customers so that the decision can be taken on what good is spending money to acquire new customers if they only buy once and do not return? Based on a customer's frequency to purchase, it is not only important to predict likelihood to buy for first-time buyers, but it is equally important to predict likelihood to buy for repeat buyers. The ultimate objective is to retain customers coming back time to time again. It is happy and loyal customers who have a large lifetime value, and many customers with a large lifetime value make for large revenues and profits for the company. The prediction of frequent patterns from the dataset can be determined using association rule mining using apriori algorithm.

The dataset is also further subject to classification trying to predict the next purchase by customers based on the customer’s past behaviour. Classification retrieves the x last purchases as feature vector in order to make the predictions.

4. EXPERIMENTAL ANALYSIS:

The dataset is analyzed with the decision tree classification technique and further the visualization of dataset analyzed is as depicted in the figure 2 through figure. The distribution of customers based on their interests with respect to various regions and channels are visualized. It is observed that the region 1 and region 3 customers are more in number compared to region 2.

Figure 2 Customers distribution in datasets- Region wise

Figure 3 Region versus preference towards number of fresh products
Figure 4 Region versus preference towards number of milk products

Figure 5 Region versus preference towards number of Grocery items

Figure 6 Region versus preference towards number of frozen foods

Figure 7 Region versus preference towards number of Detergents
Figure 8 depicts the outcome of the decision tree classification.

![Decision tree classification](http://www.iaeme.com/IJMET/index.asp)

### Figure 8 Decision tree classification

#### 5. CONCLUSIONS

In this work, the whole sale customer data is analyzed for the visualization of the customers' interests with respect to the three regions as depicted in the results. The decision tree classification had provided the effective visualization. Statistical techniques are employed for computing the mean distribution of data and the data is classified using decision tree classification.

### REFERENCES


[27] Jean-Patrice Baudry, Margarida Cardoso, Gilles Celeux, Maria José Amorim, Ana Sousa Ferreira (2012). Enhancing the selection of a model-based clustering with external qualitative variables. RESEARCH REPORT NÂ° 8124, October 2012, Project-Team SELECT. INRIA Saclay - Î©le-de-France, Projet select, UniversitÃ© Paris-Sud 11
