Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 09, September-2020

Sales Prediction of Market using Machine Learning

Mr. Soham Patangia
B.E Student, Electronics and Telecommunication
Rajiv Gandhi Institute Of Technology,
Mumbai, India

Ms. Rachana Mohite
B.E Student, Electronics and Telecommunication
Rajiv Gandhi Institute Of Technology,
Mumbai, India

Mr. Kevin Shah
B.E Student, Electronics and Telecommunication
Rajiv Gandhi Institute Of Technology,
Mumbai, India

Mr. Gaurav Kolhe
B.E Student, Electronics and Telecommunication
Rajiv Gandhi Institute Of Technology,
Mumbai, India

Mrs. Madhura Mokashi
Professor, Electronics and Telecommunication
Rajiv Gandhi Institute Of Technology,
Mumbai, India

Mrs. Prajakta Rokade
Professor, Electronics and Telecommunication
Rajiv Gandhi Institute Of Technology,
Mumbai, India

Abstract:- **Connected devices, sensors, and mobile apps make the retail sector a relevant testbed for big data tools and applications. We investigate how big data is, and can be used in retail operations. Based on our state-of-the-art literature review, we identify four themes for big data applications in retail logistics: availability, assortment, pricing, and layout planning. Our semi-structured interviews with retailers and academics suggest that historical sales data and loyalty schemes can be used to obtain customer insights for operational planning, but granular sales data can also benefit availability and assortment decisions. External data such as competitors' prices and weather conditions can be used for demand forecasting and pricing. However, the path to exploiting big data is not a bed of roses. Challenges include shortages of people with the right set of skills, the lack of support from suppliers, issues in IT integration, managerial concerns including information sharing and process integration, and physical capability of the supply chain to respond to real-time changes captured by big data. We propose a data maturity profile for retail businesses and highlight future research directions. Association Rules is one of the data mining techniques which is used for identifying the relation between one item to another. Creating the rule to generate the new knowledge is a must to determine the frequency of the appearance of the data on the item set so that it is easier to recognize the value of the percentage from each of the datum by using certain algorithms, for example apriori. This research discussed the comparison between market basket analysis by using apriori algorithm and market basket analysis without using algorithm in creating rule to generate the new knowledge. The indicator of comparison included concept, the process of creating the rule, and the achieved rule. The comparison revealed that both methods have the same concept, the different process of creating the rule, but the rule itself remains the same.**

Key Words: Big data; retail operations; maturity; availability; assortment; replenishment; pricing; layout; logistics.

I. INTRODUCTION

One of the challenges for companies that have invested a lot in consumer data collection is how to mine important information from their vast customer databases and product feature databases, in order to gain economical advantage. Several aspects of market basket analysis have been studied in academic literature, such as using customer interest profile and interests on particular products for one to one marketing, purchasing patterns in a multi-store environment to improve the sales. Market basket analysis has been intensively used in many companies as a means to discover product associations and base a retailer's promotion strategy on them.

A retailer must know the needs of customers and adapt to them. Market basket analysis is one achievable way to find out which items can be place together. Market basket analyses gives retailer good information about related sales on group of goods basis Customers who buys bread often also buy several products related to bread like milk, butter or jam. It makes sense that these groups are placed side by side in a retail centre so that consumers can access them promptly. Such related groups of goods also must be located side-by-side in order to remind customers of related items and to lead them through the centre in a logical manner. Market basket analysis determines which products are bought together and to design the supermarket arrangement, and also to design promotional campaigns. Therefore, the Market consumer behaviours need to be analysed which can be done during dissimilar data mining techniques. Well-versed decision can be made easily about product placement, pricing, endorsement, profitability and also finds out if there are any successful products that have no significant related elements. Similar products can be found so those can beplaced near each other or it can be cross- sold.

Association rules can be mined and this process of mining the association rules is one of the most important and powerful aspect of data mining. One of the main criteria of ARM is to find the relationship among various items in a database. An association rule is of the form A→B where A is the antecedent and B is the Consequent. and here A and B are item sets and the underlying rule says us purchased by the customers who purchase A are likely to purchase B with a probability percentage factor as %C where C is known as confidence such a rule is as follows: "seventy per cent of people who purchase beer will also like to purchase diapers" This helps the shop managers to study the behaviour or buying habits of the customers to increase the sales . based on this study items that are regularly purchased by the customers are put under closed

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 09, September-2020

proximity. For example, persons who purchase milk will also likely to purchase Bread.
The interestingness measures like support and confidence also plays a vital role in the association analysis. The support is defined as percentage of transactions that contained in the rule and is given by ***Support = (# of transactions involving A and B) / (total number of transactions).***

The other factor is confidence it is the percentage of transactions that contain B if they contain A

Confidence = Probability (B if A) = P(B/A) Confidence = (# of transactions involving A and B) / (total number of transactions that have A)

Consider the following example:

TABLE I. EXAMPLE OF PREDICTIVE ANALYSIS

Customer	Item Purchased	Item Purchased
1	Burger	Coke
2	Puff	Mineral Water
3	Burger	Mineral Water
4	Puff	Tea

If A is "purchased Burger " and B is "purchased mineral water" then

Support=P(A and B)=1/4

Confidence=P(B/A)=1/2

Item sets that satisfy minimum support and minimum confidence are called strong association rules.
Predictive analytics is composed of two words predict & analysis, but it works in reverse viz. first analyse then predict. It is human nature to want to know and
predict what the future holds. Predictive analytics deals with the prediction of future events based on previously observed historical data by applying sophisticated methods like machine learning. The historical data is collected and transformed by using various techniques like filtering, correlating the data, and so on. Prediction process can be divided into four steps:

- Collect and pre-process raw data;
- Transform pre-processed data into a form that can be easily handled by the selected machine learning method;
- Create the learning model (training) using the transformed data;

Report predictions to the user using the previously created learning model.
An essential goal in data mining is to create and enhance the precision of predictive models, and a basic challenge toward this end lies in the discovery of new features, inputs or predictors. This paper illustrates how rules generated from Market Basket Analysis (MBA) may be utilized to improve predictive models. The goal of data mining process is the extraction of information from large data sets, transform such information into some understandable structure for future application. It is a process of using different techniques to find useful patterns or models from data. This process is use to select, explore and model large amount of data.

II. BIG DATA

Big data is an "imprecise description of a rich and complicated set of characteristics, practices, techniques, ethical issues, and outcomes all associated with data". A more technical and complementary definition is "datasets that could not be perceived, acquired, managed, and processed by traditional [infor-mation technology] and software/hard-ware tools within a tolerable time" For companies, big data is a cornucopia of digitalized content about consumers' cognitions, emotions, behaviors, and reactions critical to the ongoing data-driven industrial revolution.
'Big data' has different definitions that focus on size, range, and speed. While some incorporate a dynamic definition as beyond the ability of typical software, some use a static number as 100 terabytes or have a growth rate that is higher than 60% annually. Big data is also defined as using the population in the analysis rather than the sample, as this can change the mind-set of data analysis in three ways:

Accept errors as they can be smoothed by sheer quantity, accept correlation rather than causality, and discover value of data besides its original purpose. Big data is also defined based on volume, variety, velocity, and veracity.
Volume represents to the amount of data, which can be affected by the scope, and is the primary attribute.
Variety refers to the data included in the analysis which comes from various sources and are in many mediums. Velocity includes the frequency of both data generation and deliveries. Finally, veracity is to do with the uncertainty of the data. In this paper, the definition used is: all population data that can be accessed timely. This definition implicitly involves volume, variety, velocity, and veracity features of big data.

III. RETAIL CHALLENGES AND POTENTIAL BIG DATA APPLICATIONS

Although retailers were distributing products passively in the past, with the information about the end customer demand, they are becoming more proactive in the supply chain. The retail supply chain contains four major activities: assorting goods, breaking goods into smaller packages, holding inventory near the customer, and providing value-adding services such as gift wrapping and warranties. Agrawal and Smith mapped the supply chain of two major home furnishing retailers in the US and proposed a comprehensive set of planning processes: product design (private label merchandise) and assortment planning, sourcing and vendor selection, logistics planning, distribution planning and inventory management, clearance and markdown optimization, and cross-channel optimization. Major activities of retailing are to decide what products to carry (assortment, product design, procurement), how to sell products to customers (marketing, including pricing), and how to complete these efficiently (supportive functions, such as logistic planning). Out of stock (OOS) situations and poor on shelf availability (OSA) lead to customer dissatisfaction and consequently loss of market share . Assortment, pricing, and store layout are challenging processes in retail operations that are expected to benefit from big data analysis. Assortment and pricing involve a large amount of granular decisions and can be affected by various factors, such as customer preferences, store location, and demand elasticity. Without effective quantitative

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 09, September-2020

analysis, making these decisions can be resource-consuming and ineffective. Store layout is also worth exploring as it can affect customers' purchasing decisions.

IV. ASSORTMENT, PRICING AND STORE LAYOUT

Assortment is one of the most difficult tasks in the retail supply chain and has received high priority due to its impact on sales. Big data is used to micro-segment customers and optimize assortment. Common practices include analysis of the correlations between items purchased as well as time- and location- dependent purchase patterns. The output helps retailers understand customer preferences and compositions, and improve forecast processes.

Pricing is one of the most difficult issues facing retailers due to the large amounts of SKUs whose price can vary in different locations and over time according to local demand and competition. In the case of markdowns and promotions, pricing is even more difficult because of the increased level of uncertainty and the lack of historical data with the same promotion conditions. As the promotions are a function of what else is also on the market, it is almost impossible to replicate a promotion, although some similarities can be captured for forecasting purposes. With extra computing and analysis capacity, big data analysis enables promotion decisions to be taken more effectively and efficiently. Big data can also help retailers evaluate sources of sales lifts and plan future promotions more effectively.

Store layout is worth exploring in terms of how it can benefit from big data analysis as it has an impact on purchasing decisions. In the past, customer in-store behavior could be analyzed by observing sampled customers and this information could help retailers optimize store layout and shelf design. However, this exercise requires a significant level of effort and reports have suggested that using videos, mobile services, WI-FI, and RFID tags attached to shopping carts to track customer movement in store can provide high volumes of information at low cost. Although in-store traffic alone cannot provide information on customer behavior, customers' in-store location data still provide useful correlations with sales.

V. MINING ASSOCIATION RULES

Till now, we have looked at the Apriori algorithm with respect to frequent itemset generation. There is another task for which we can use this algorithm, i.e., finding association rules efficiently. For finding association rules, we need to find all rules having support greater than the threshold support and confidence greater than the threshold confidence.

But, how do we find these? One possible way is brute force, i.e., to list all the possible association rules and calculate the support and confidence for each rule. Then eliminate the rules that fail the threshold support and confidence. But it is computationally very heavy and prohibitive as the number of all the possible association rules increase exponentially with the number of items.

Given there are n items in the set I, the total number of possible association rules is 3n–2n+1+1.

We can also use another way, which is called the two-step approach, to find the efficient association rules. The two-step approach is:

Step 1: Frequent itemset generation: Find all itemsets for which the support is greater than the threshold support following the process we have already seen earlier in this article.

Step 2: Rule generation: Create rules from each frequent itemset using the binary partition of frequent itemsets and look for the ones with high confidence. These rules are called candidate rules.

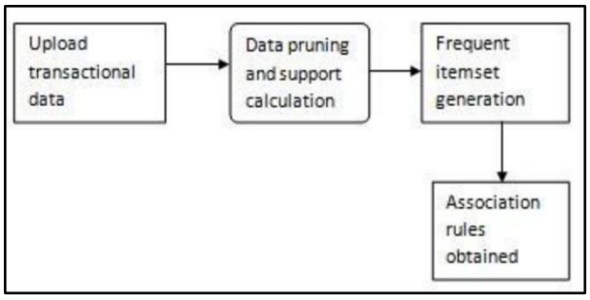

Fig.1. Proposed Model

Algorithm for sales growth using Apriori.

Step1: Start.

Step2: Select the database of the market.

Step3: Apply minimum support to find all the frequent sets with k items in a database.

Step4: Use the self-join rule to find the frequent sets with k+1 items with the help of frequent k-itemsets. Repeat this process from k=1 to the point when we are unable to apply the self-join rule.

Step5: RawData P 7−−> D F 7−−> X d RFCn0 7−−> ˆy*TopN 7−−> ˆ y//
RawData: Exploratory Data Analysis
P: Partition RawData into Dtrain, Dtest, ...
F: Feature Design builds design matrix X from d RFCn0: Random Forest Classifier makes probabilistic prediction ˆ y*
X n0: Hyperparameters found using OOB classifier
TopN: TopN Variants make binary predictions ˆ y from ˆ y* ˆ
y: Prediction Explorer inspects binary predictions ˆy//

Step6: RawData P −> DDSets
RawData 7→{Ds}s∈DSets //DSets = {train,test,kaggle},
P is the unique partition defined by a partition of the set of users, U, ordered by, say, user id, into {Us}s∈DSets Utrain: 80% of 131,209 users with available ultimate orders. Utest: 20% of 131,209 users with available ultimate orders.
Ukaggle: 75,000 users whose ultimate orders are withheld by Kaggle.
{Ds}s∈DSets is the image of RawData under P.//[5]

Step7: : F : D → Mn×m(R); D 7→ X //F = (fj)m 1 is an m-tuple of functions fj : D →Rn a feature fj(D) is a column of X fj are compositions of aggregations, filtrations, arithmetic ... compute some fj via an unsupervised learning technique, Latent Dirichlet Allocation (LDA), introduced in [BNJ03]. Prod(u) are all products purchased by u ∈ U 10

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 09, September-2020

Rows of X are user-product pairs ((u,p) | u ∈ U & p ∈ Prod(u))//

Step8: RFCn0 : Mn×m(R) → [0,1]n; X → ˆ y*

Step9: TopN : [0,1]n →{0,1}n ˆ y* 7→ ˆy //TopN: maps make binary predictions from probability
TopNthreshold: chooses a classification threshold p0; best for optimizing threshold metrics (F1-score); not best for product applications TopNu uses user's mean basket size as N; set basket size by user reduces its variance; higher precision versions to autopopulate carts?
TopNN: uses a constant value N for each user; worse metrics but best for displaying on fixed-width web page (zero basket size variance).//

Step10: Stop

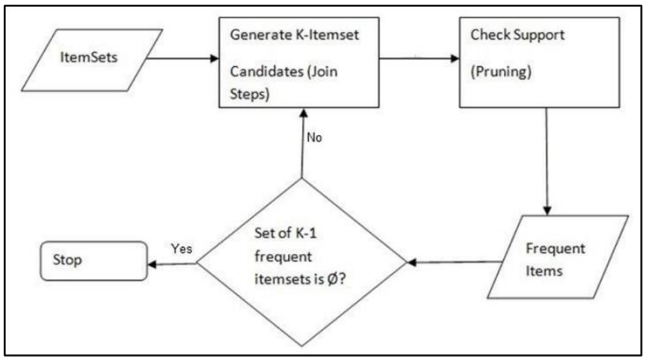

Fig.2. Sales Growth Algorithm

VI. SYSTEM MAINTENANCE AND EVALUATION

Types of Maintenance:
- Corrective Maintenance
- Adaptive Maintenance
- Perfective maintenance
- Preventive maintenance

Corrective Maintenance :-
Even with the best quality assurance activities, it is likely that the customer will uncover defects in the software. Corrective maintenance changes the software to correct defects.

Adaptive Maintenance :-
Over time, the original environment (e.g., CPU, operating system, business rules, external product characteristics) for which the software was developed is likely to change. Adaptive maintenance results in modification to the software to accommodate changes to its external environment.

Perfective maintenance :-
As software is used, the customer/user will recognize additional functions that will provide benefit. Perfective maintenance extends the software beyond its original functional requirements.

Preventive maintenance :-
Computer software deteriorates due to change, and because of this, preventive maintenance, often called software reengineering, must be conducted to enable the software to serve the needs of its end users. In essence, preventive maintenance makes changes to computer programs so that they can be more easily corrected, adapted, and enhanced.

Evaluation:
System evaluation provides framework for classification scheme to identifying sets of similar systems. The framework integrates previous studies on software evaluation, productivity models, software quality factors, and total quality models. It also classifies information about software systems from the perspective of the project, the system, and the environment.

VII. RESULTS

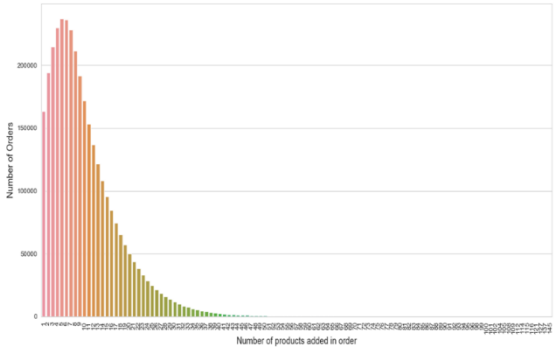

Fig.3. Graph of Number of products added in a cart at a time

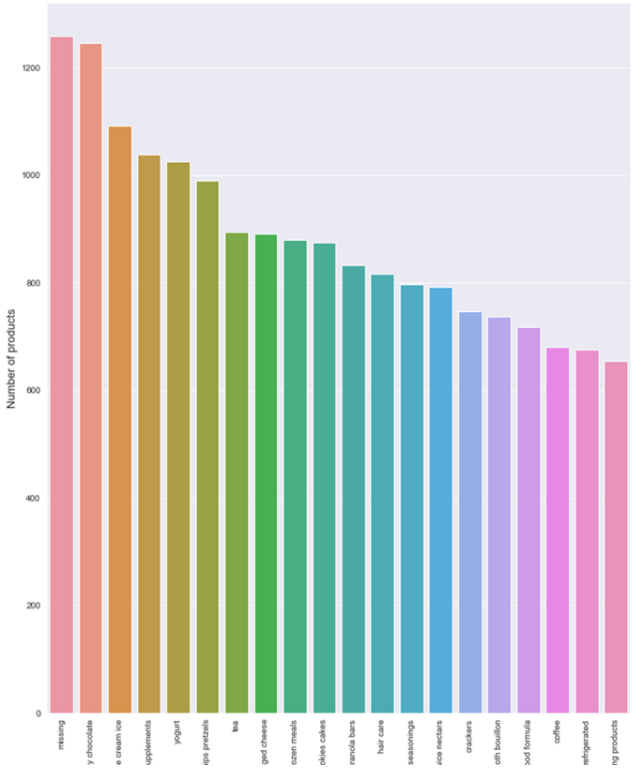

Fig.4. Graph of total products as per aisle

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 09, September-2020

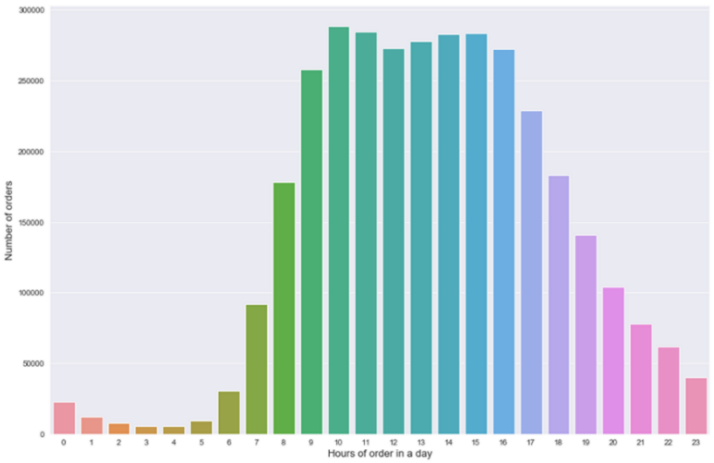

Fig.5. Graph of sales as per hour of the day.

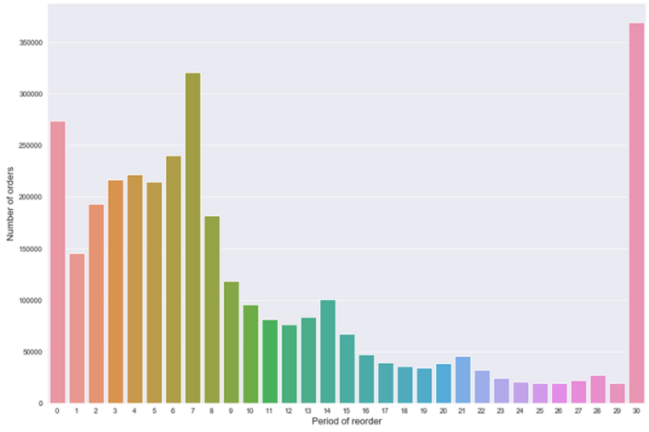

Fig.6. Graph of period of reorder since last order

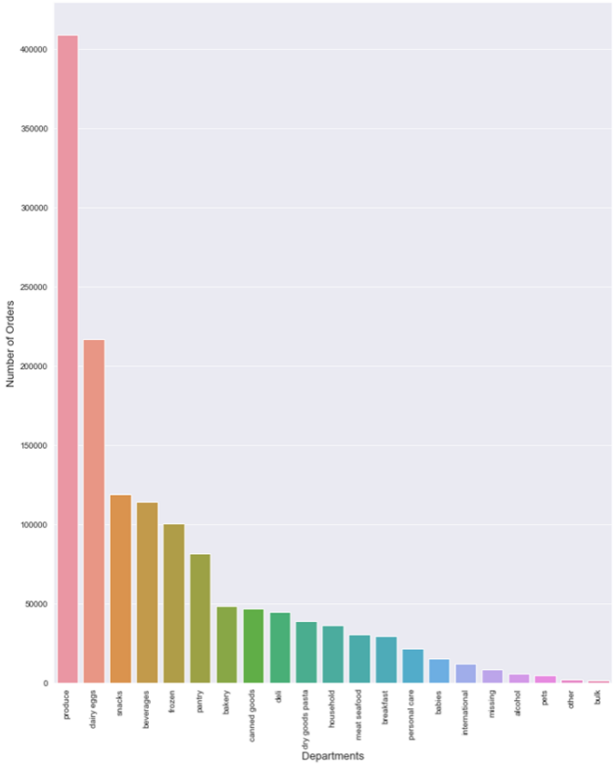

Fig.7. Graph of bestselling as per departments

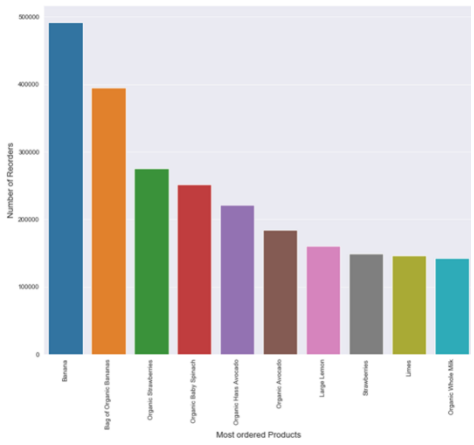

Fig.8. Graph of product reorder

VIII. CONCLUSION

Market basket analysis generates the frequent item set i.e. association rules can easily tell the customer buying behaviour and the retailer with the help of these concepts can easily setup his retail shop and can develop the business in future. The main algorithm used in market basket analysis is the Apriori algorithm. It can be a very powerful tool for analysing the purchasing patterns of consumers. The three statistical measures in market basket analysis are support, confidence. Support measures the frequency an item appears in a given transactional data set, confidence measures the algorithm's predictive power or accuracy. In our example, we examined the transactional patterns of grocery purchases and discovered both obvious and not-so-obvious patterns in certain transactions. Association rules and the existing data mining algorithms usage for market basket analysis , also it clearly mentioned about the existing algorithm and its implementation clearly and also about its problems and solutions. Predictive modelling offers the potential for firms to be proactive instead of receptive. Predictive modelling using transactional data create particular challenges which need to be carefully addressed to develop valuable models. With MBA, leading retailers can drive more profitable advertising and promotions, attract more customers, increases the value of market basket and much more. Consumers, planners, merchandisers and store administrators have started to recognize how this new era of easy-to-use market basket analysis tools helps to work more intelligent and compete more successfully. Our future work would be to design and develop intelligent prediction models to generate the association rules that can be adopted on recommendation system to make the functionally more operational. Better and effective rule mining techniques can be used for better performance of the recommendation system. In future the same algorithm can be modified and it can be extended in the future work which also decreases the time complexity.

IX. FUTURE SCOPE

Since this system provides solutions to small supermarkets and is compatible to limited size of data there can be a wide range of scope possible

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 09, September-2020

- With stronger device storage and proper ram management devices it can hold data of larger supermarkets and can hold data for a longer duration

- Expanding the limitations will allow the user to directly access the data from the cash counter as the product is scanned from the barcode

- A system can be designed in which Using the history of the previous bills of the customer a new shopping list and list of recommendations can be provided for shopping next time, even the customer can manually edit the shopping list.

Supermarket shopping can be digitalized and product review can be shown just by scanning the barcode , and same time the same product will pop in the recommendation for the user, this will help the market boost the economy from the medium of digital marketing.

REFERENCES

[1] Trnka., "Market Basket Analysis with Data Mining Methods", International Conference on Networking and Information Technology (ICNIT) ,2010.
[2] W Yanthy, T. Sekiya, K. Yamaguchi , "Mining Interesting Rules by association and Classification Algorithms", FCST 09.
[3] Chiu, K.S.Y., Luk, R.W.P, Chan, K.C.C., and Chung, K.F.L, "Market-basket Analysis with Principal Component Analysis: An Exploration", IEEE International Conference on Systems, Man and Cybernetics, Vol.3, 2002.
[4] Cunningham , S.J. and Frank, E., "Market Basket Analysis of Library Circulation Data", International Conference on Neural Information Processing, Vol.2. 1999.
[5] Vo,B.and Le,B.,"Mining traditional association rules using frequent itemsets lattice",International Conference on Computers & Industrial Engineering. 2009.
[6] Rastogi, R.. and kyuseok Shim, "Mining optimised association rules with Categorical and numerical attributes", IEEE transactions on Knowledge and Data Engineering, vol.14, 2002.