Judicious Analysis of Market Basket using K-Apriori Algorithm

[1]Siddhita Ashok Bagwe, [2]Soumya Anil Shenoy, [3]Kunjal Kiran Shirwalkar,[4] Gauravi Hemant Tembulkar,[5]Manya Gidwani

[1]Student, [2]Student, [3]Student,[4]Student,[5]Professor
Department of Information Technology
Shah & Anchor Kutchhi Engineering College, Mumbai-400088, India

***Abstract*—** Market Basket Analysis is a useful method of discovering customer purchasing patterns by extracting association or co-occurrences from store's transactional databases. The information obtained from the analysis can be used in forming marketing, sales, services and operation strategies. The existing system however fails to discover important purchasing patterns from the customer point of view in the current multi-store situations. Providing personalized services to the customers is the main challenge for the supermarket chains these days. To provide this personalization, it is of utmost importance for the retailers to understand the frequency, commonness and understandable patterns of the buys made by frequent customers. The personalization thus provided will be of the form where each revisiting customer is provided with the products he has been purchasing for multiple of his previous buys. Along with the predicted cart, comprising of the previously bought items, the user shall also be provided with the items which were popularly bought by other customers from the aisle similar to the user's most purchased from aisle.

Keywords— *Cart Prediction, Classification and Regression Tree Algorithm, K- Apriori Algorithm, Market Basket Analysis*

I. Introduction

Nowadays, a lively challenge for supermarket chains is to offer personalized services to their customers. Detecting the purchase habits of customers and their evolution in time is a crucial challenge for effective marketing policies and engagement strategies. In fact, purchasing patterns of individuals evolve in time and can experience changes due to both environmental reasons, like seasonality of products or retail policies, and personal reasons, like diet changes or shift in personal preferences. Thus, a satisfactory solution to next basket prediction must be adaptive to the evolution of a customer's behavior, the recurrence of their purchase patterns and their periodic changes. In such context one of the most promising facilities retail markets can offer to their customers is market basket prediction, i.e., the automated forecasting of the next basket that a customer will purchase. An effective basket recommender can act as a shopping list reminder suggesting the items that the customer could probably need. A successful realization of this application requires an in-depth knowledge of an individual's general and recent behavior. Market basket prediction, i.e., supplying the customer a shopping list for the next purchase according to her current needs, is one of these services. Current approaches are not capable of capturing at the same time the different factors influencing the customer's decision process: co-occurrence, sequentially, periodicity and recurrences of the purchased items.

Market Basket Analysis is a modeling technique based upon the theory that if you buy a certain group of items, you are likely to buy the same group of items during your next purchase. It is one of the most common and useful types of data analysis for marketing and retailing. To put it another way, it allows retailers to identify relationships between the items that people buy. The purpose of market basket analysis is to determine what products customers purchase together. It takes its name from the idea of customers throwing all their purchases into a shopping cart (a "market basket") during grocery shopping. Knowing what products people purchase as a group can be very helpful to a retailer or to any other company. A store could use this information to place products frequently sold together into the same area, while a catalogue or World Wide Web merchant could see it to determine the layout of their catalogue and order form. Direct marketers could use the basket analysis results to determine what new products to offer their prior customers.

II. Literature Review

Initially, to identify our system scope and requirements we referred papers suggesting the use of Association Rule Mining, K-Apriori and CART algorithms. Association Rule Mining provided relationships between two items popularly purchased together. This helped to find items that could be sold together as a combination, since they were retailing at a faster rate when sold together. However it didn't serve our purpose. Jake Morgan's "Classification and Regression Tree Analysis", [2] CART algorithm assisted us in the visualization of the dataset and hence determining the factors to be considered for mapping of the classification tree. Mesut Gumus, Mustafa S. Kiran's, "Crude oil forecasting using XGBoost", [3] provided us with a gradient library which is efficient and can solve problems in fast and accurate ways. Nitin Kumar Mishra, Vimal Mishra, Saumya Chaturvedi, "Solving Cold Start Problem using Market Basket Analysis", [4] using which we were able to eradicate the problem using the attributes from the Kaggle dataset. It provides the new user with a predicted cart based on day_of_the_week and hour_of_the_day attribute. Dr. Neeraj Bhargava, Renuka Purohit, Sakshi Sharma, Abhishek Kumar's, "Prediction of Arthritis using Classification and Regression Tree algorithm" , [5] M Balamurugan, S Kannan's, "Performance Analysis Of CART and C5.0 using Sampling Techniques"[6] and Jatinder Kaur, Jasmeet Singh Gurm's, "Description of Genetic and CART Algorithm using Data Mining Tool" [7] were used for visualization and determining the patterns. K-Apriori is a combination of K-means which sort the similar and dissimilar data into respective groups and Apriori used for frequent item mining and associative rule mining. It is an algorithm which is said to have improved the performance of the original Apriori algorithm.

III. METHODOLOGY

The data that we've considered for this evaluation is taken from a static source (Kaggle.com).[1] The dataset is a comparative set of files. These files show the customers' ordering patterns and the orders over time. The objective of carrying out this procedure is to forecast the products in the users next purchasing cart. In the data available, each user profile is provided with the information ranging from the user's individual purchases, the order in which the product is purchased, sequence of products purchased in order. Data regarding the week and hour of day when the order was placed, and a relative measure of time between the two orders placed for a particular user is also provided. All of the entities be it the customer name, product, order, aisle, etc. each of them has an associated unique id.

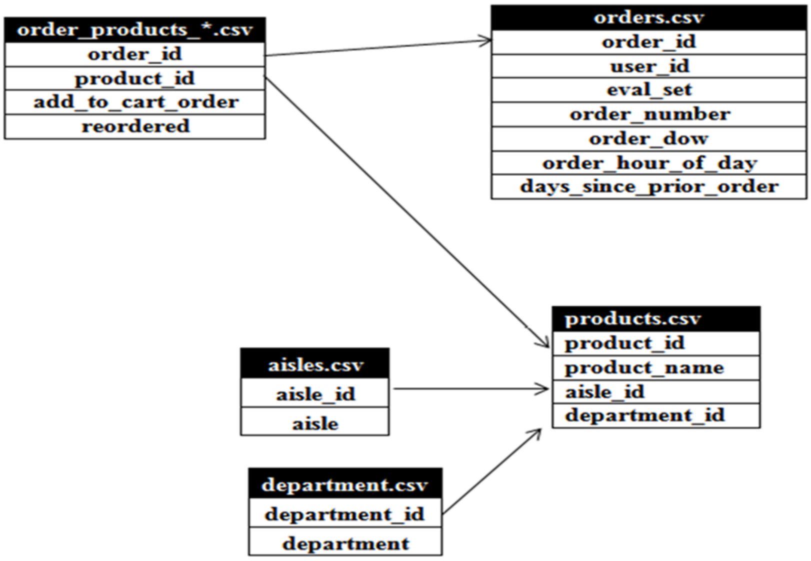


Figure 1. Data Set Description

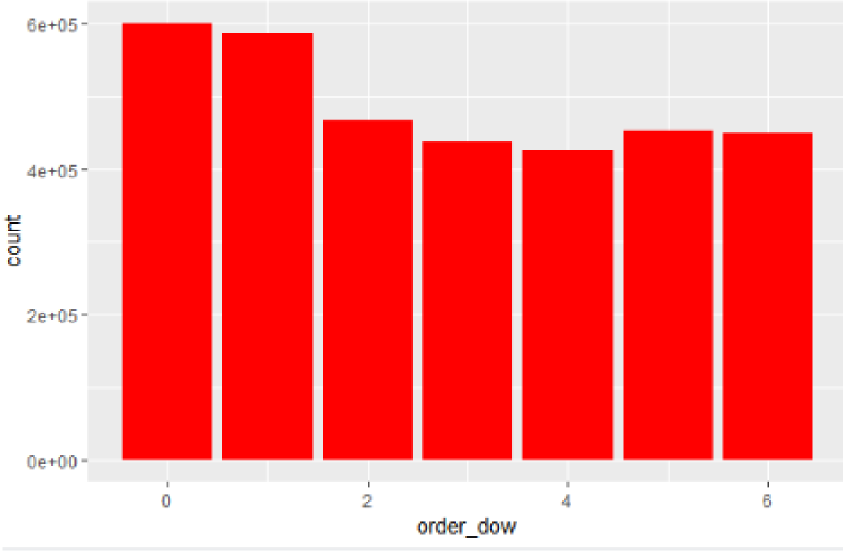


Figure 2. Histogram for Day_of_week

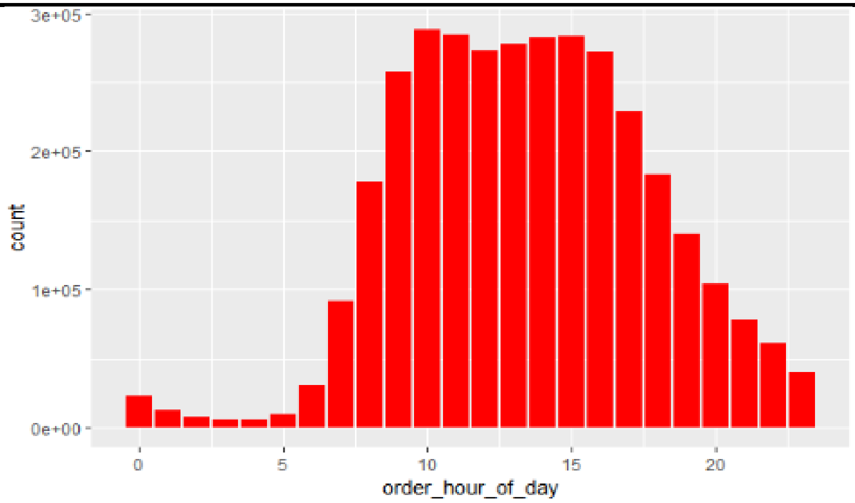


Figure 3. Histogram for order_hour_of_day

IV. ALGORITHM

We begin with Reading and Merging of the Dataset. After Reading the dataset (provided from Kaggle), merging of user_id, order_id and product_id table is performed, based on which the average cart product count is obtained. The product_ids bought till now for each user for each product_id are being counted initially. Later sorting is performed based on Most Bought Item for each user. Consider a temporary variable Temp1 which indicates the Total products bought by each user and Temp2 which contains the Total times user went to supermarket. Using Temp1 and Temp2 we calculate the average number of items bought in each visit by the user. For predicting the products within the cart of each user we are using K Apriori Algorithm. In this, we select top x (which is calculated earlier) product_ids based on most bought item. The output table obtained is merged with the Products dataset where we will get actual names of the product user is going to buy. For suggested items, the cart is predicted based on the aisle the user has visited. The no. of aisle_id is considered from the items bought from that aisle, thus suggesting the items from the aisle that is most visited by the user. The most popularly bought item is suggested to the user from that particular aisle. We are considering top 4 aisle_ids and top 2 product_ids from each aisle for the suggested cart. Later, the repeated products are removed which are already present in the cart. Finally, we have the predicted and suggested cart ready for the existing user.

1. The dataset is first loaded into the software (here python)
2. The tables with primary key user_id, product_id and order_id is merged.
3. For each user the most purchased products are listed in the descending order
4. A temporary variable x stores the products with the number of times they were purchased
5. Another variable y stores the number of visits made by user to the website/supermarket
6. Average of the variables x and y gives us average number of items bought in each visit by each user (say z)
7. Now using k-apriori algorithm select top z product ids from x.
8. Merging the output of step 7 with product table we will get our predicted cart

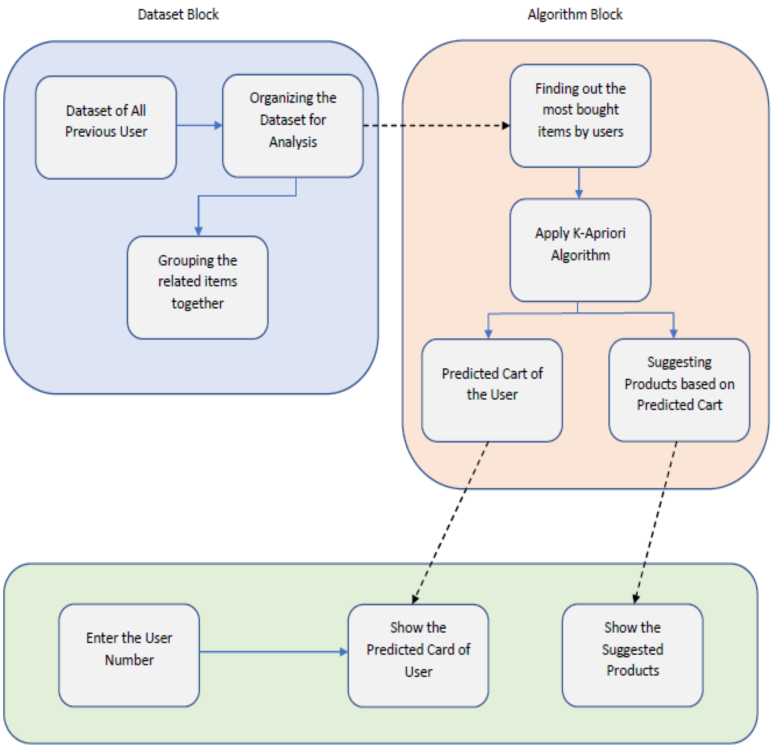
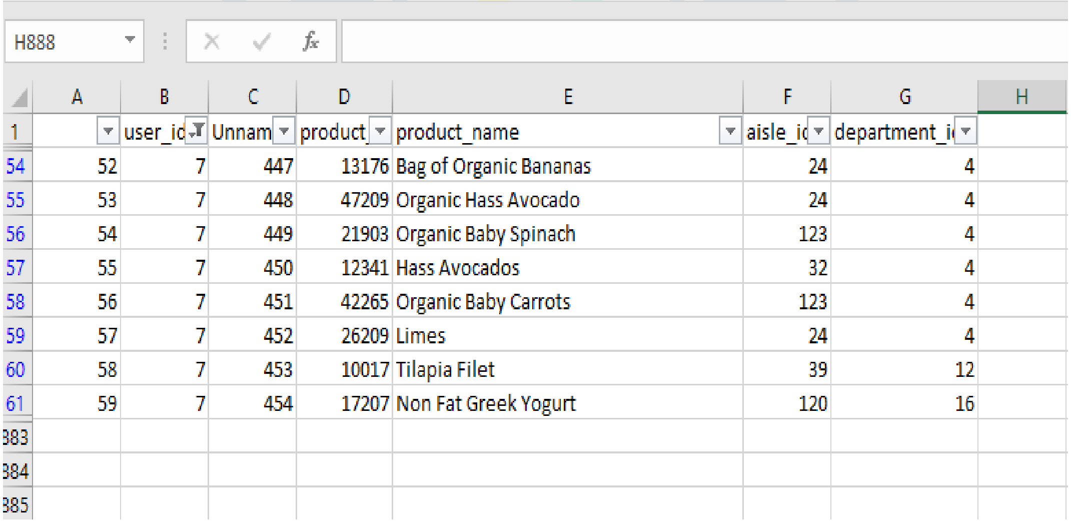

Dataset Block
Algorithm Block
Dataset of All Previous User
Organizing the Dataset for Analysis
Grouping the related items together
Finding out the most bought items by users
Apply K-Apriori Algorithm
Predicted Cart of the User
Suggesting Products based on Predicted Cart
Enter the User Number
Show the Predicted Card of User
Show the Suggested Products

Figure 3. Working

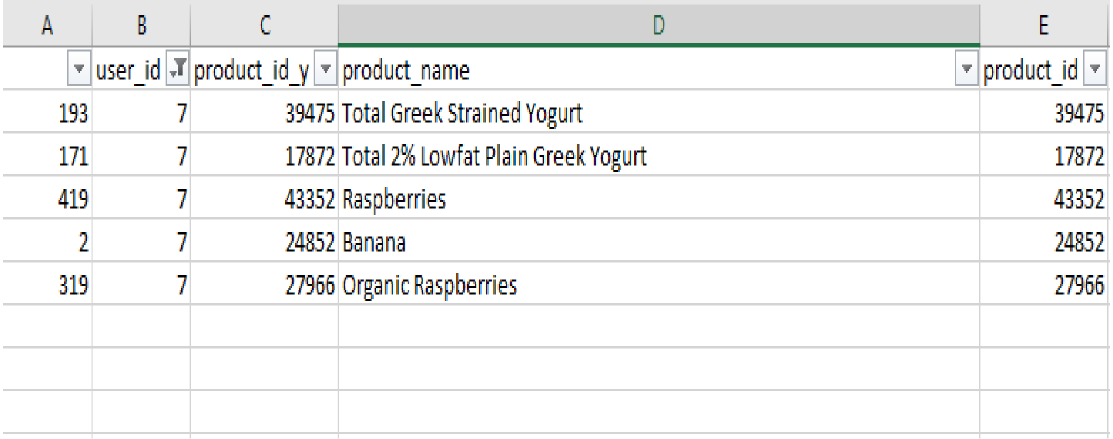
V. RESULT

We have successfully premeditated and devised an algorithm that identifies the purchasing patterns of the buyers. For the users who join for a repeated buying session, the expectation of having predicted items already in his cart is successfully achieved. The items predicted are on the basis of the items he previously purchased; the order in which the products were purchased in the preceding buy is also important. The recommended items in the suggested item cart are predicted on the basis of the frequency of the items purchased on regular intervals from a particular aisle. The frequency thus calculated after each purchase made from an aisle, is used to predict an item for the user during his next visit.



	A	B	C	D	E	F	G	H
1		user_id	Unnam	product_	product_name	aisle_ic	department_i	
54	52	7	447	13176	Bag of Organic Bananas	24	4	
55	53	7	448	47209	Organic Hass Avocado	24	4	
56	54	7	449	21903	Organic Baby Spinach	123	4	
57	55	7	450	12341	Hass Avocados	32	4	
58	56	7	451	42265	Organic Baby Carrots	123	4	
59	57	7	452	26209	Limes	24	4	
60	58	7	453	10017	Tilapia Filet	39	12	
61	59	7	454	17207	Non Fat Greek Yogurt	120	16	
383								
384								
385								

Table 1. Predicted Cart for customer user_id 7.



A	B	C	D	E
	user_id	product_id_y	product_name	product_id
193	7	39475	Total Greek Strained Yogurt	39475
171	7	17872	Total 2% Lowfat Plain Greek Yogurt	17872
419	7	43352	Raspberries	43352
2	7	24852	Banana	24852
319	7	27966	Organic Raspberries	27966

Table 2. Suggested Cart for customer user_id 7

VI. CONCLUSION

With this project we aim to proffer an auto generated shopping cart. For this we obtained a pre made data set from a closed Kaggle competition. The data obtained is thus unbiased, to the best of our knowledge. The Classification and Regression Tree algorithm provides output in tree format, which then helps us in taking accurate decisions. Market Basket Analysis is one of the key techniques used by large retailers to uncover associations between items. It works by looking for combinations of items that occur together frequently in transactions. To put it another way, it allows retailers to identify relationships between the items that people buy. This technique will be profitable to not only the customers but also the retailers and shop owners who when gets a prior idea of the goods the customers will buy, can pre order the stocks accordingly. This in turn will help them to initiate a prompt delivery, accounting as positive review for the retailer.

VII. ACKNOWLEDGMENT

The authors would like to thank the review committee for their invaluable suggestions and feedback without whom our work would have been very difficult. We also take this opportunity to express our profound gratitude and deep regards to our guide Ms. Manya Gidwani for her exemplary guidance, monitoring and constant encouragement throughout the course of this project. The blessing, help and guidance given by her time to time shall carry us a long way in the journey of life on which we are about to embark. No project is ever complete without the guidance of these experts who have already established a mark on this path before and have become master of it. So, we would like to take this opportunity to thank all those who have helped us in implementing this project.

VIII. REFERENCES

[1] Kaggle, Instacart Dataset https://www.kaggle.com/c/instacart-market-basket-analysis/data

[2] Jake Morgan, "Classification and Regression Tree Analysis", Boston University School of Public Health, May 2014

[3] Mesut Gumus, Mustafa S. Kiran, "Crude oil forecasting using XGBoost", International Journal of Advanced Technology & Engineering Research, May 2018, pg. 1100-1103.

[4] Nitin Kumar Mishra, Vimal Mishra, Saumya Chaturvedi, "Solving Cold Start Problem using Market Basket Analysis", IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), June 2018, pg. 1598-1601.

[5] Dr. Neeraj Bhargava, Renuka Purohit, Sakshi Sharma, Abhishek Kumar, "Prediction of Arthritis using Classification and Regression Tree algorithm", February 2017, pg. 606-610.

[6] M Balamurugan, S Kannan, "Performance Analysis of CART and C5.0 using Sampling Techniques", IEEE International Conference on Advances in Computer Applications, March 2016, pg. 72-75.

[7] Jatinder Kaur, Jasmeet Singh Gurm, "Description of Genetic and CART Algorithm using Data Mining Tool", International Journal of Advanced Research in Computer Science and Software Engineering, June 2015, pg. 948-952