Market Basket Analysis on Instacart Data

Shradha Atmaram Yewale
San Jose State University
shradhaatmaram.yewale@sjsu.edu

Dhruv Dinesh Soni
San Jose State University
dhruvdinesh.soni@sjsu.edu

Vineet Batthina
San Jose State University
vineet.batthina@sjsu.edu

Abstract - Instacart is a grocery ordering and delivery app. Market Basket Analysis on Instacart data is a technique used by businesses to better understand how their customers shop their product items. This technique will help us determine the association between different product items. It looks for the combination of items that occur together frequently in transactions. The goal of this project is touse market basket analysis to establish associations between products and customer recommendation system to improve overall customer experience. Association rule mining will be considered for modeling. We can also predict customers product purchasing patterns based on their previously purchased product and determine what can be the customers next purchase using Random Forest classifier to improve the user recommendation system. We will beusing the data of Instacart. The data will include the product purchase history of a customer along with product items that were bought together by other customers.

I. INTRODUCTION

The retail and e-commerce businesses are changing dramatically as more people prefer to shop online. Among the most prevalent business issues faced by online merchants are how to adopt a more data-driven client retention approach and how to determine consumer needs by evaluating who they are and what they purchased frequently. In addition, these online businesses are collecting millions of transaction data in real time. As a result, these organizations may be able to extract useful insights from large data, which might help them achieve a competitive advantage and attract more and more customers to buy their products online.

To provide best customer service, sellers should understand and accommodate customer's needs as well as their choices. To understand the customer's needs we can use association rule mining to find patterns in the data. Association rule learning is the process of finding the rules to govern associations between the sets of product items in huge datasets. This technique is extensively used and designed to analyze transaction data, to detect strong rules revealed in transaction data using metrics of interest based on the strong rule notion. Market basket analysis is a popular application of Association Rules.

One method of determining whether related items can be combined is to do a market basket analysis. It provides retailers with useful data on connected sales by group of product items. Customers that purchase biscuits frequently also purchase other related items such as milk, coffee, or tea. These product item groupings should be arranged side by side in a retail center or should display in recommended products to buy section, so that shoppers may quickly reach them. Such linked sets of products must also be placed sequentially to remind buyers of related things and guide them to logically search product items through the Instacart.

Instacart is a company that provides grocery delivery service. Market basket analysis has been used in many companies to discover product associations and base a retailer's product's promotion strategy on them. [1] Retailers may quickly make informed decisions about product placement, pricing, and marketing using market basket analysis to understand customer's needs, and related products can be located and placed near one other.

There are two main algorithms used in market basket analysis. First one is apriori algorithm which computes the frequent items in the dataset through several iterations. Second one is FP-Growth which is a more advanced version of Apriori which finds frequent items without candidate generation. It uses a tree data structure and scans the database twice instead of frequent iterations which improves the performance significantly. We will be using both the approaches and comparing its results.

II. BACKGROUND AND LITERATURE REVIEW

The significant amount of research and analysis has been conducted in an attempt to better understand consumer needs and behavior on e-commerce platforms. With more individuals purchasing on e-commerce websites on a daily basis, the industry is continuously evolving to make the customer experience simpler and smoother. Market basket research will assist these e-commerce websites, in this case Instacart, in better understanding each user's needs and behavior, analyzing frequent orders by customers and associated product items with those orders, and using this information to personalize the user experience, manage resources efficiently, and boost business.

A. Market Basket Analysis

Market basket analysis establishes trends and find connections between customer's purchases. A conditional algorithm is used to model the relationship:

IF {X, Y} THEN {Z}

This notation indicates that "the items in the right are likely to be ordered with items on the left".

{Ai} -> {Ci}

The antecedent of the rule is the set of product items on the left(cookies, sandwich) whereas the consequent is the set of items on the right (coffee, tea, milk). The probability that client purchasing a cookies and sandwich will occur is the rule's support.

The probability that a customer will purchase a particular item based on the purchasing another item is referred to as the confidence of the rule. It can be used for price changing strategy, item placement, and accordingly increasing the overall profitability. Placing high margin product items near associated high confidence items can increase the overall margin on the purchases of product items. [2]

The lift of the rule is the ratio of the support of the right-hand side of the rule co-occurring with the left-hand side of the rule, divided by the probability that the right-hand side and left-hand side co-occur if the two are independent. In the project we have considered two algorithms which are usedto find association between the products:

1. Apriori Algorithm
2. FP-Growth Algorithm. [3]

B. Association Rule Mining

Association Rule Mining is a process for searching and discovering relationships between product items in a dataset. The goal of rules of association is to locate the information about products that alternate as the rules are in the form of rules. [4]

C. Apriori Algorithm

The Apriori Algorithm is a basic algorithm proposed by Agrawal and Srikant in 1994 for the determination of the frequent product item set for Boolean association rules. It is divided into many stages called narratives. It consists of (1) The establishment of candidate items, (2) Calculation of support of each k-itemset candidate, (3) Set the high frequency pattern and (4) If no new high frequency pattern is obtained the whole process is stopped. If not, then k plus one and return part 1 [5].

D. FP-Growth

The FP-Growth algorithm is a development of the Apriori algorithm. FP-Growth algorithm improves the Apriori algorithm's flaws [6]. The generating candidate is necessary in Apriori to obtain frequent product item sets. However, this technique is not implemented since FP-Growth searches for frequent product item sets using the concept of tree development. Because of this, the FP-Growth algorithm is faster than the Apriori algorithm.

E. Instacart Dataset

Instacart is a company that operates a grocery delivery and pick-up service in the United States and Canada. It allows customers to order groceries from participating retailers with the shopping being done by a personal shopper. The dataset is a relational set of files describing customer's orders over time. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 300,000 Instacart users. For each user, there are approximately 4 to 100 orders, with the sequence of products purchased in each order. The week and hour of day the order was placed, and a relative measure of time between orders is also given.

III. EXPERIMENTAL SETUP AND IMPLEMENTATION

A. Setup

Language used	Python
Development Environment	Jupyter notebook
Version control	Git

B. Libraries

- Pandas - Pandas library was used for data manipulation and analysis.
- Sklearn - Sklearn provides a selection of efficient tools for machine learning and statistical modeling
- Seaborn - This library was used for data visualization.
- FPgrowth - Frequent pattern growth algorithm implemented in Python.
- Numpy - Numpy was used for carrying out mathematical operations on the data efficiently.
- Apyori - A simple implementation of the Apriori algorithm in Python.

.

C. Dataset

The datasets were provided by Instacart Technology Company and were taken from Kaggle to perform analysis.

D. Dataset description

- **Aisles** - This dataset provides information on the aisles such as aisle ID and aisle names, through which the products were organized.
- **Orders** - This dataset has information about the customer orders like order ID, order number, week day of the order, hour of theorder, days since prior order, and user ID.
- **Order_Products_Prior** - This dataset givesinformation on the orders, products, and reordered products.
- **Order_Products_train** - This dataset is the same as order_products_prior and it is a trained dataset.
- **Products** - This dataset gives information on the products such as product name, product ID, aisle and departments.
- **Department** - This dataset provides information on the departments such as department names and department Id.

E. Implementation flow

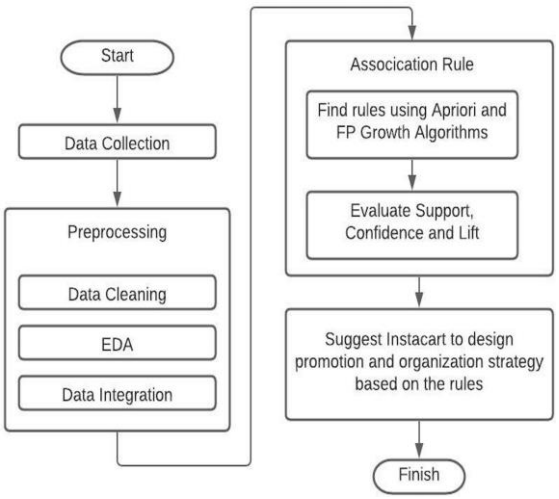

Fig. 1 Implementation Flow

F. Algorithm matrices

Random Forest Classifier - Random forest was implemented using parameters max features as "log2", max depth as 11 and n estimators as 24. The data had 8 million records in it of which 33% of data was used to test the model.

Apriori Algorithm –

- Support - It's the default popularity of an item. That means, the support of item A is the ratio of transactions involving A to the total number of transactions.
- Confidence - Likelihood of the customer who bought both A and B products. It is the ratio of the number of transactions involving both A and B and the number of transactions involving B.

  Confidence (A => B) = Support (A, B)/Support(A)
- Lift - Increase sale of A when product B is sold.

  Lift*(A=>B) =Confidence (A, B)/ Support(B).*

If the lift value is 1 that means there is no correlation within the product item set. If it is > 1 then there is a positive correlation which means that products in the set are more likely to be bought together. If it is < 1 then the products in the set are unlikely to be bought together.

IV. RESULT

A. EDA

First, we plot the product items count by department to find out which department has the greatest number of orders. In Fig.1 we can see that most of the products that are ordered are from the personal care and snack department. The count for bulk and other is least.

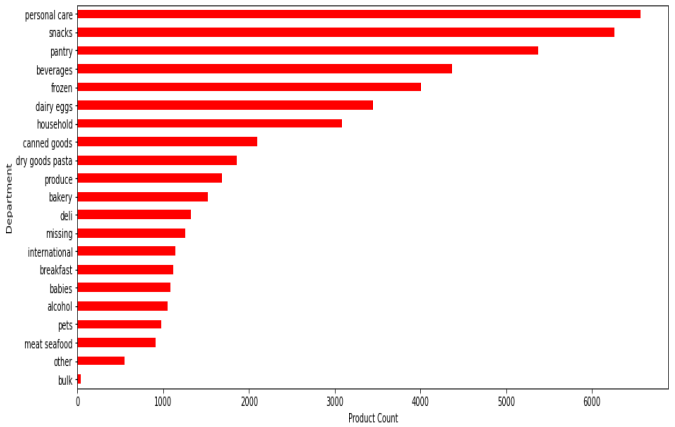

Fig. 2 Product Frequency plot according to department

To find out which products are more frequently bought by customers, we plot the word cloud from the product by department table. Word clouds are used for simple text analysis. In a retail market business appealing words and packaging also plays an important key role in sales. Fig. 2 shows the result of word cloud plot. From the below word cloud, few of the most frequent words seem to be Organic, Original, Gluten Free and natural.

We also check how many products are reordered and how many products are not reordered using bar graph. Fig.4 shows that over 2 million products are reordered and over1.25 million products are not reordered.

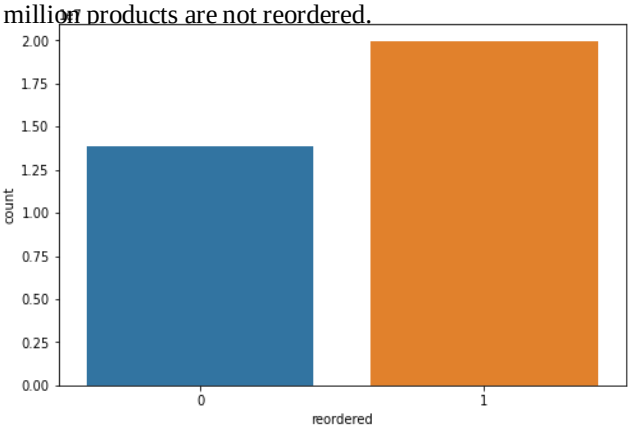

Fig. 4 Number of products reordered

Fig 5 is a bar graph showing that approximately 3.5 million customers reorder in 7 days. Following table shows most reordered and least reordered products. From the table we can see that milk, sparkling water, fresh fruit and eggs are most reordered.

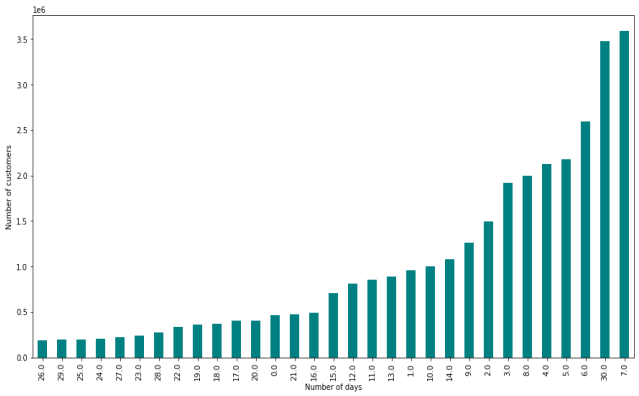

Fig. 5 Number of days after customer reorders

B. Reorder Prediction

To predict which product will be reordered next we used random forest classifier which gives us 90.45% accuracy.

```
Accuracy using randomforest classifier:  0.9045400568325 05
CPU times: user 2min 39s, sys: 2.38 s, total: 2min 42s
Wall time: 2min 43s
```

Fig. 6 Random forest classifier accuracy

Below is the feature importance graph for random forest classifier. The top five features are - reordered_latest, reordered_sum, reordered_count, Uxp_total_bought, p_reorder_ratio.

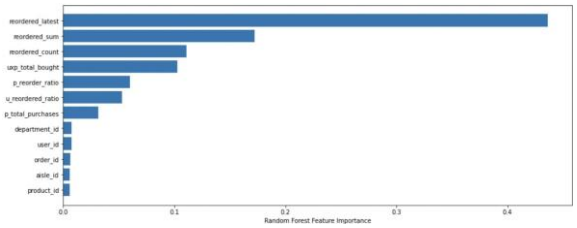

Fig. 7 Feature importance graph

	aisle	department	reordered
0	milk	dairy eggs	0.781812
1	water seltzer sparkling water	beverages	0.72993
2	fresh fruits	produce	0.718823
3	eggs	dairy eggs	0.706359
4	soy lactosefree	dairy eggs	0.692361
5	...	...	...
6	beauty	personal care	0.212858
7	first aid	personal care	0.195864
8	kitchen supplies	household	0.194802
9	baking supplies decor	pantry	0.167595
10	spices seasonings	pantry	0.152933

Table 1. Reordered Products

C. Association rule mining

Following table shows the confidence and lift scores between product A and product B using apriori algorithm. From the table we can see that bananas are bought more frequently with other Organic food items such as Fruits, Vegetables, Greens.

This shows a pattern that there is a high probability that healthy and natural products are brought together.

	A	B	support_A	support_B	support_AB	confidence	lift
0	Organic Raspberries	Organic Strawberries	0.042123	0.082295	0.010395	0.246766	2.998576
1	Organic Fuji Apple	Banana	0.027774	0.147230	0.010576	0.380801	2.586433
2	Organic Raspberries	Bag of Organic Bananas	0.042123	0.118016	0.012375	0.293772	2.489255
3	Organic Hass Avocado	Bag of Organic Bananas	0.066126	0.118016	0.019117	0.289098	2.449649
4	Cucumber Kirby	Banana	0.030195	0.147230	0.010044	0.332637	2.259300
5	Organic Avocado	Banana	0.054675	0.147230	0.016426	0.300424	2.040508
6	Organic Strawberries	Bag of Organic Bananas	0.082295	0.118016	0.019175	0.233004	1.974342
7	Strawberries	Banana	0.044267	0.147230	0.012854	0.290370	1.972220
8	Large Lemon	Banana	0.047652	0.147230	0.012637	0.265192	1.801204
9	Organic Baby Spinach	Bag of Organic Bananas	0.075295	0.118016	0.015664	0.208033	1.762751
10	Limes	Banana	0.043796	0.147230	0.010016	0.228705	1.553382
11	Organic Baby Spinach	Banana	0.075295	0.147230	0.016083	0.213594	1.450750
12	Organic Strawberries	Banana	0.082295	0.147230	0.017477	0.212377	1.442480

Table 2. Apriori algorithm result

Below table shows the confidence and lift score between two products using the FP-Growth algorithm. We can observe a similar pattern which we observed using apriori algorithm. We can see that banana and organic products are brought together.

	antecedent	consequent	confidence	lift	support
0	[Bag of Organic Bananas]	[Organic Strawberries]	0.162478	1.974342	0.019175
1	[Bag of Organic Bananas]	[Organic Baby Spinach]	0.132726	1.762751	0.015664
2	[Bag of Organic Bananas]	[Organic Hass Avocado]	0.161986	2.449649	0.019117
3	[Bag of Organic Bananas]	[Organic Raspberries]	0.104856	2.489255	0.012375
4	[Banana]	[Organic Strawberries]	0.118708	1.442480	0.017477
5	[Banana]	[Organic Baby Spinach]	0.109234	1.450750	0.016083
6	[Banana]	[Organic Avocado]	0.111564	2.040508	0.016426
7	[Banana]	[Large Lemon]	0.085832	1.801204	0.012637
8	[Banana]	[Strawberries]	0.087305	1.972220	0.012854
9	[Banana]	[Limes]	0.068032	1.553382	0.010016
10	[Banana]	[Cucumber Kirby]	0.068220	2.259300	0.010044
11	[Banana]	[Organic Fuji Apple]	0.071835	2.586433	0.010576
12	[Cucumber Kirby]	[Banana]	0.332637	2.259300	0.010044
13	[Organic Avocado]	[Banana]	0.300424	2.040508	0.016426
14	[Large Lemon]	[Banana]	0.265192	1.801204	0.012637
15	[Organic Strawberries]	[Bag of Organic Bananas]	0.233004	1.974342	0.019175
16	[Organic Strawberries]	[Banana]	0.212377	1.442480	0.017477
17	[Organic Strawberries]	[Organic Baby Spinach]	0.144915	1.924631	0.011926
18	[Organic Strawberries]	[Organic Hass Avocado]	0.155520	2.351868	0.012798
19	[Organic Strawberries]	[Organic Raspberries]	0.126310	2.998576	0.010395

Table 3. FP-Growth algorithm result

Following table shows another association rule example which indicates that consumers take more than one flavored yogurt and water.

	A	B
0	Non Fat Raspberry Yogurt	Icelandic Style Skyr Blueberry Non-fat Yogurt
1	Icelandic Style Skyr Blueberry Non-fat Yogurt	Non Fat Raspberry Yogurt
2	Icelandic Style Skyr Blueberry Non-fat Yogurt	Vanilla Skyr Nonfat Yogurt
3	Vanilla Skyr Nonfat Yogurt	Icelandic Style Skyr Blueberry Non-fat Yogurt
4	Total 2% Lowfat Greek Strained Yogurt With Blueberry	Total 2% with Strawberry Lowfat Greek Strained Yogurt
5	Total 2% with Strawberry Lowfat Greek Strained Yogurt	Total 2% Lowfat Greek Strained Yogurt With Blueberry
6	Total 2% with Strawberry Lowfat Greek Strained Yogurt	Total 2% Lowfat Greek Strained Yogurt with Peach
7	Total 2% Lowfat Greek Strained Yogurt with Peach	Total 2% with Strawberry Lowfat Greek Strained Yogurt

Table 4 Association rule example

V. CONCLUSION

From the results we can conclude that to predict which product will be reordered next we can use random forest classifier which gives us 90.45% accuracy. Using the apriori algorithm and FP-Growth algorithm we can find out which product will be brought together by the customer. The model thus formed can be used to push targeted advertisements to customers. Physical stores can use this information to organise items effectively. Both the algorithms gave similar results.

VI. REFERENCES

[1] Annie, L. C. M., & Kumar, A. D. (2012). Market basket analysis for a supermarket based on frequent itemset mining. International Journal of Computer Science Issues (IJCSI), 9(5), 257.

[2] P. Pravallika and K. Narendra, "Analysis on Medical Data sets using Apriori Algorithm Based on Association Rules," IJSRSET, vol. 4, no. 1, pp. 717–722, 2018.

[3] R. Gangurde, D. B. Kumar, and D. S. D. Gore, "Building Prediction Model using Market Basket Analysis," Int. J. Innov. Res. Comput. Commun. Eng., vol. 5, no. 2, pp. 1302–1309, 2017.

[4] M. Kaur and S. Kang, "Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining," Procedia Comput. Sci., vol. 85, no. Cms, pp. 78–85, 2016.

[5] R. Kanapaka, "Association Rule Mining using Apriori algorithm For food dataset," Int. J. Comput. Appl., vol. 112, no. 4, p. 8887, 2015

[6] D. Hunyadi, "Performance comparison of Apriori and FP-Growth algorithms in generating association rules," Proc. Eur. Comput. Conf., pp. 376–381, 2011.