Market Basket Analysis

-TANUSH SINGH

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. One common technique used for market basket analysis is Association Rule Mining. The main idea is to understand purchase pattern in a transaction and get frequently bought together itemsets. E-commerce retailers uses these results to gain maximum profit by recommending items to their customers. In a store, all vegetables are placed in the same aisle, all dairy items are placed together and cosmetics form another set of such groups. Investing time and resources on deliberate product placements like this not only reduces a customer's shopping time, but also reminds the customer of what relevant items he might be interested in buying, thus helping stores cross-sell in the process. Our approach will be using Apriori and FP-Growth algorithms to analyze the dataset and get appropriate results. We will then analyze the results obtain from both the algorithms to decide their efficiency on a given dataset. According to our understanding we are expecting FP growth to perform well when compared with Apriori when it comes to Execution time.

Keywords • Market Basket Analysis, Association Rules, Apriori Algorithm, FP Growth, Pattern Mining, Frequent Itemsets.

1. INTRODUCTION

How come E-commerce websites are almost always able to make appropriate suggestions when you add something to the cart? How is it possible that every time you visit a supermarket to buy something, you come back with a bunch of other useful stuff that you did not plan on buying? These are some of the questions we have always wondered about. The answer to this ? Market Basket Analysis.

Market Basket Analysis is a technique to find out the affinities between the collection of data. It is used by retailers to increase their revenue by understanding customer purchase pattern. It analyzes large transactional data to uncover products groupings. It uses association rule learning where it will look at all the combinations of products that occur together frequently. Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable. This will help the retailers to identify the relationships between the items that people buy.

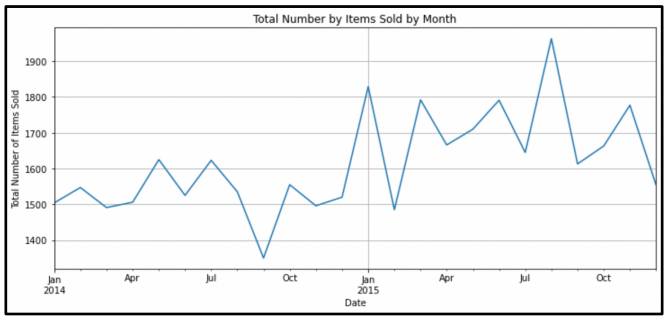
We have chosen a groceries dataset for a retail brand. The dataset has 38,765 rows and 3 columns of the purchase orders of people from the grocery stores namely Member_number, Date, itemDescription. Our main aim in this project is to find Customer Shopping patterns to gain insights about the products customers prefer to buy together by applying Association Rule Mining algorithms. Apriori algorithm and FP-Growth algorithms are widely used to get frequent itemsets. The algorithms will mine frequent sets of items that are bought together by the customer. We will be comparing the associations formed using these algorithms and obtain some useful information. It is possible that FP-growth algorithm, which is believed to be an improved version of apriori could provide better association rules than apriori for the same dataset. The strength of obtained association rules will be measured on three metrics namely support, confidence and lift. The algorithms will identify rules which has a support value greater than a predefined threshold. Similarly, confidence is calculated for all the transactions and will keep only the itemsets which has confidence greater than a predefined threshold. The other metric we will be comparing the algorithms on is their computational performance by observing how much time each one of them takes to complete.

2. MODELS AND TECHNIQUES:

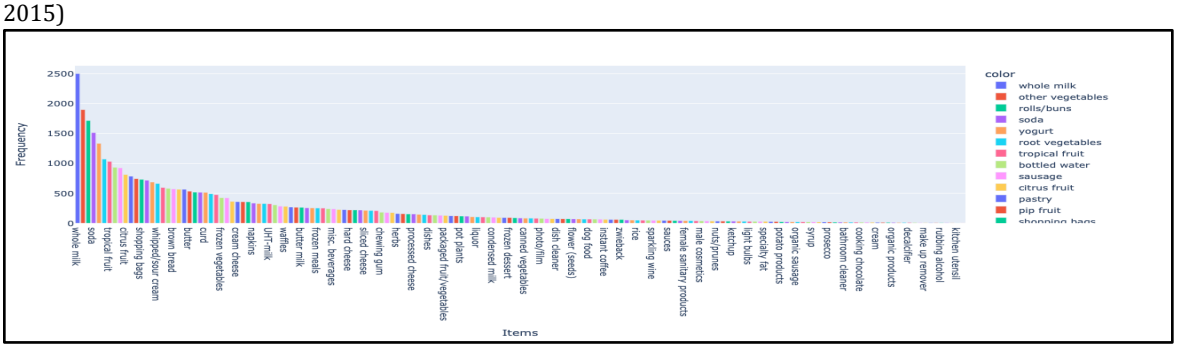
2.1. Data Visualization

```
In total, 167 unique items have been purchased by customers
The total number of transactions in 38765 days are 728
With an average daily sale of 53 items
```

We used Libraries like plotly and matplotlib for visualization of data. Below images gives us some idea about the sales trends and popular products that have been purchased by customers in our dataset.



The above plot shows number of items sold each month over the span of dates in our dataset (Jan 2014 - Dec 2015)



The above visualizations clearly shows which products are purchased very frequently by the customers and could definitely considered as frequent itemsets.

2.2. Data Preprocessing

In our data, every row consists of an item, the customer who purchased it and the date it was purchased. For our requirements, we have combined all items purchased by one customer in a day as a single transaction using TransactionEncoder.

2.3. Applying Apriori and FP Growth Algorithm

The goal of these algorithms is to find combinations of products that are often bought together, which we call frequent itemsets. The technical term for the domain is Frequent Itemset Mining.

Frequent Itemset Mining basically gives us three metrics to interpret:

- Support (the number of times, or percentage, that the products co-occur).
- Confidence (the number of times that a rule occurs, also the conditional probability of the right-hand side given the left-hand side).
- Lift (the strength of association).

After implementation, we noticed that Apriori and FP Growth give us similar list of frequent itemsets

Processing 28 combinations | Sampling itemset size 4e 3
Frequent itemsets using Apriori Algorithm

	support	itemsets
0	0.004010	(Instant food products)
1	0.021386	(UHT-milk)
2	0.001470	(abrasive cleaner)
3	0.001938	(artif. sweetener)
4	0.008087	(baking powder)
...	...	...
745	0.001136	(whole milk, sausage, rolls/buns)
746	0.001002	(soda, whole milk, rolls/buns)
747	0.001337	(yogurt, whole milk, rolls/buns)
748	0.001069	(soda, sausage, whole milk)
749	0.001470	(yogurt, sausage, whole milk)

750 rows × 2 columns

Frequent itemsets using FP Growth Algorithm

	support	itemsets
0	0.157923	(whole milk)
1	0.051728	(pastry)
2	0.018780	(salty snack)
3	0.085879	(yogurt)
4	0.060349	(sausage)
...	...	...
745	0.001403	(yogurt, chewing gum)
746	0.001069	(other vegetables, chewing gum)
747	0.001002	(chewing gum, soda)
748	0.001069	(pasta, whole milk)
749	0.001002	(seasonal products, rolls/buns)

750 rows × 2 columns

3. DISCUSSION AND RESULTS

The frequent itemsets obtained using Apriori and FP Growth algorithms can be used to obtain useful association rules. Initially, we derived these association rules based on a certain threshold for confidence, which is basically the conditional probability of buying a product given some other product has been already bought. However, confidence is not the best metric to obtain association rules, since some items can have a high probability of being purchased regardless of the other item and thus doesn't really give us a good association. For example, in our dataset, milk is a very frequently purchased product and therefore if we try to associate milk with let's say, toothpaste, it will still give us a good confidence because a lot of people will buy milk but it is obviously not a very god relation.

Thus we switched to lift as a better metric of association instead. Below, we can observe how using lift as a metric instead of

Association Rules based on confidence

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(UHT-milk)	(bottled water)	0.021386	0.060683	0.001069	0.050000	0.823954	-0.000228	0.988755
1	(bottled water)	(UHT-milk)	0.060683	0.021386	0.001069	0.017621	0.823954	-0.000228	0.996168
2	(other vegetables)	(UHT-milk)	0.122101	0.021386	0.002139	0.017515	0.818993	-0.000473	0.996060
3	(UHT-milk)	(other vegetables)	0.021386	0.122101	0.002139	0.100000	0.818993	-0.000473	0.975443
4	(UHT-milk)	(rolls/buns)	0.021386	0.110005	0.001804	0.084375	0.767013	-0.000548	0.972009
...	...	...	...	...	...	...	...	...	...
1195	(yogurt, sausage)	(whole milk)	0.005748	0.157923	0.001470	0.255814	1.619866	0.000563	1.131541
1196	(yogurt, whole milk)	(sausage)	0.011161	0.060349	0.001470	0.131737	2.182917	0.000797	1.082219
1197	(sausage, whole milk)	(yogurt)	0.008955	0.085879	0.001470	0.164179	1.911760	0.000701	1.093681
1198	(yogurt)	(sausage, whole milk)	0.085879	0.008955	0.001470	0.017121	1.911760	0.000701	1.008307
1199	(sausage)	(yogurt, whole milk)	0.060349	0.011161	0.001470	0.024363	2.182917	0.000797	1.013532

1200 rows × 9 columns

confidence gives us fewer but more reliable association rules.

Association rules based on confidence

Association Rules using Apriori Algorithm

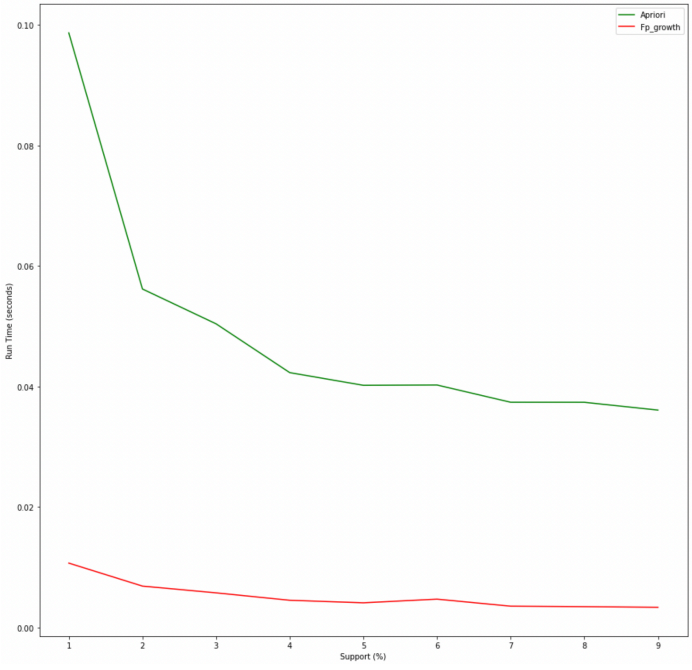
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(tropical fruit)	(UHT-milk)	0.067767	0.021386	0.001537	0.022682	1.060617	8.785064e-05	1.001326
1	(UHT-milk)	(tropical fruit)	0.021386	0.067767	0.001537	0.071875	1.060617	8.785064e-05	1.004426
2	(beef)	(brown bread)	0.033950	0.037626	0.001537	0.045276	1.203301	2.597018e-04	1.008012
3	(brown bread)	(beef)	0.037626	0.033950	0.001537	0.040853	1.203301	2.597018e-04	1.007196
4	(beef)	(citrus fruit)	0.033950	0.053131	0.001804	0.053150	1.000349	6.297697e-07	1.000020
...	...	...	...	...	...	...	...	...	...
235	(yogurt, whole milk)	(sausage)	0.011161	0.060349	0.001470	0.131737	2.182917	7.967480e-04	1.082219
236	(sausage, whole milk)	(yogurt)	0.008955	0.085879	0.001470	0.164179	1.911760	7.012151e-04	1.093681
237	(yogurt)	(sausage, whole milk)	0.085879	0.008955	0.001470	0.017121	1.911760	7.012151e-04	1.008307
238	(sausage)	(yogurt, whole milk)	0.060349	0.011161	0.001470	0.024363	2.182917	7.967480e-04	1.013532
239	(whole milk)	(yogurt, sausage)	0.157923	0.005748	0.001470	0.009310	1.619866	5.626300e-04	1.003596

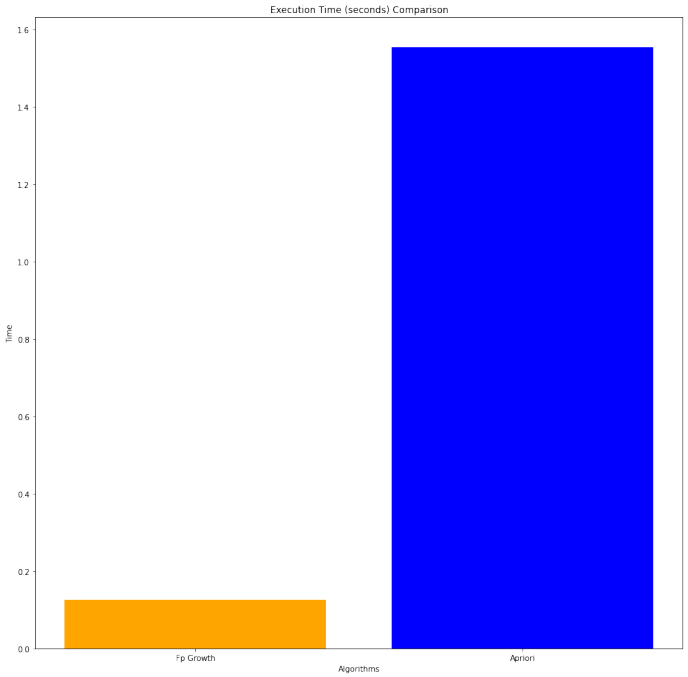
240 rows × 9 columns

Association rules based on lift

As stated above, both Apriori and FP Growth algorithms provide us similar results for frequent item sets and hence we can say that the will provide same association rules.

However, the internal working of both these algorithms vary significantly. Apriori Algorithm scans the dataset to find the items that occur more than the minimum support threshold, then repeat the process for pair of two items and so on. Thus it needs to scan the datasets multiple times to collect frequent item sets. On the other hand, FP Growth algorithm just scans the dataset once and maintains all the items in nodes of FP trees. It links the items that are associated together in the data based on minimum threshold of support. This difference in working of these two algorithms is evident in the runtime comparison below.





Another thing worth noting in the second image is that as mentioned above, for support of 1%, Apriori has to scan the dataset very high number of times and thus performs very poorly whereas the performance of FP Growth doesn't change too much as it only scans the dataset once regardless of the threshold and then links the frequent items.

4. CONCLUSION

Performing market basket analysis gave us a better understanding of how frequent itemset generation algorithms work in real life, and how useful association rules can be obtained from those. How metrics are useful to obtain and assess the reliability of association rules. And how based on different requirements different algorithms can be used, for example Apriori can be implemented when time is not a big constraint but there is a limited availability of space and FP Growth can be implemented when water execution time is required however space is not a constraint.

Another point worth noting is that these rules can prove out to be very useful in gaining insights and make recommendations, however, they cannot guarantee customer behavior completely.

6. CITING RELATED WORK

http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/ACKNOWLEDGMENTS

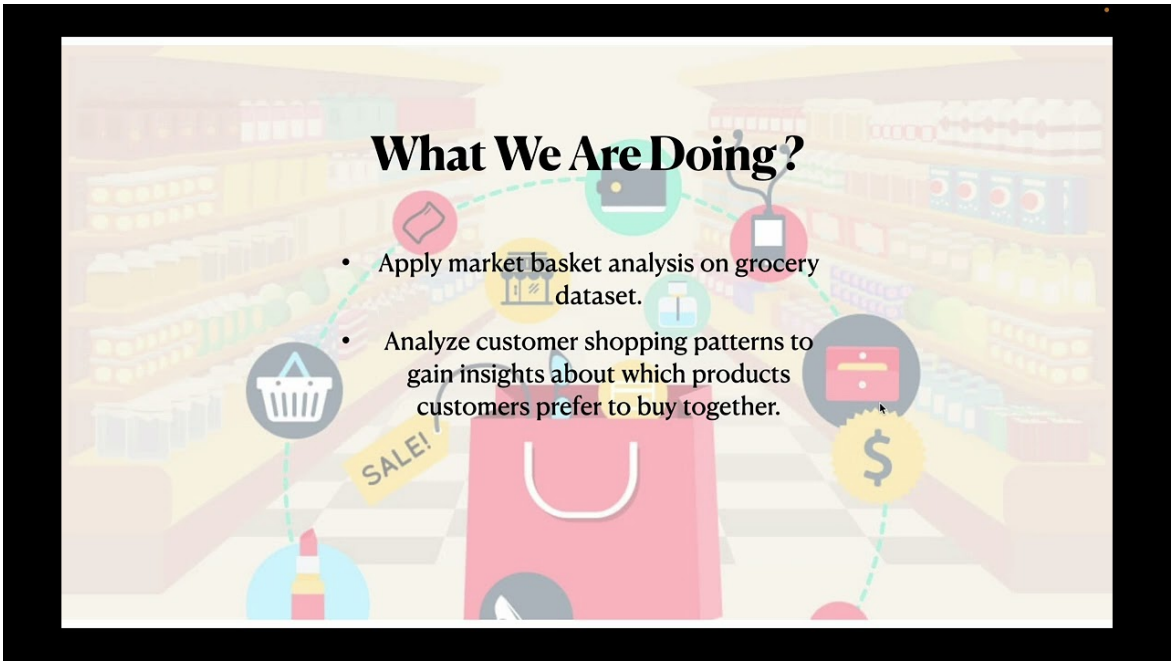
https://www.ibm.com/docs/en/db2/10.5?topic=SSEPGG_10.5.0/com.ibm.im.model.doc/c_lift_in_an_association_rule.html

https://analyticsindiamag.com/apriori-vs-fp-growth-in-market-basket-analysis-a-comparative-guide/

https://towardsdatascience.com/association-rules-2-aa9a77241654

Dataset can be found here: https://www.kaggle.com/datasets/heeraldedhia/groceries-dataset?datasetId=877335&sortBy=voteCount

For video description of the project or complete list of association rules, please refer the link below:



https://drive.google.com/drive/folders/1hO4wC-FaeeRnD4u50c2yZ0OHxY7UaImi?usp=sharing