

MARKET BASKET ANALYSIS OF RETAIL STORE

SOLO- Anirudha Patil

Step 1: Prototype Selection

Abstract:

Developing a Market Analysis for a new business or new product provides an entrepreneur a better understanding of the size and specific segments of a market, as well as an evaluation to determine if the target market will support the business' growth plans. Market Analysis is a critical part of any business plan created whether to inform the business or to communicate to potential investors the size of the opportunity.

Industry Type: Retail

- **Industry Category:** All the small scale retail stores, grocery stores, marts etc. that sell day to day needs can be helped using associate rules that is market basket analysis. We can help these stores to align the products, which are bought frequently and help increase their sales.

Problem Statement:

To understand the shopping pattern of the customer and help the retail vendor increase their sales accordingly.

Business Assessment:

The approach is based on the theory that customers who buy a certain item (or group of items) are more likely to buy another specific item (or group of items). For example: while at a quick-serve restaurant (QSR), if someone buys a sandwich and cookies, they are more likely to buy a drink than someone who did not buy a sandwich. This correlation becomes more valuable if it is shown to be stronger than that between the sandwich and drink without the cookies.

More and more organizations are discovering ways of using market basket analysis to gain useful insights into associations and hidden relationships. As industry leaders continue to explore the technique's value, a predictive version of market basket analysis is making in-roads across many sectors in an effort to identify sequential purchases

Target Specifications and Characterization

- **Gaining market share:** Once a business reaches its peak growth, finding new ways to do so might be difficult. Market basket analysis may be used to integrate gentrification and demographic data to locate the sites of new businesses or geo-targeted marketing.
- **Campaigns and promotions:** MBA is used to identify the goods that work well together as well as the products that serve as the cornerstones of their product range.
- **Behavior analysis:** A fundamental tenet of marketing is comprehending consumer behavior patterns. MBA may be used for anything, including UI/UX and basic catalog designs.

- **Optimization of in-store activities:** MBA is useful in deciding what goes on the shelves as well as at the back of the shop. Because geographic patterns are a major factor in determining the strength or popularity of particular products, MBA is increasingly used to manage inventory for each store or warehouse.

Information and Data Analysis:



Association Rules (Antecedent and Consequent)

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It identifies frequent if-

then associations called association rules which consists of an antecedent (if) and a consequent (then) Association Rules is one of the very important concepts of machine learning being used in market basket analysis. Let's now see what an association rule exactly looks like. It consists of an antecedent (if) and a consequent (then), both of which are a list of items.

For example: “__If tea and milk, then sugar__” (“If tea and milk are purchased, then sugar would also be bought by the customer”)

Antecedent: Tea and Milk

Consequent: Sugar.

What is Association Rule Learning?

Association Rule Learning is rule-based learning for identifying the association between different variables in a database. One of the best and most popular examples of Association Rule Learning is the Market Basket Analysis. The problem analyses the association between various items that has the highest probability of being bought together by a customer.

For example, the association rule, {onions, chicken masala} => {chicken} says that a person who has got both onions and chicken masala in his or her basket has a high probability of buying chicken also.

Apriori Algorithm:

The algorithm was first proposed in 1994 by Rakesh Agrawal and Ramakrishnan Srikant. Apriori algorithm finds the most frequent item sets or elements in a transaction database and identifies association rules between the items.

```
importing Libraries

:
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
```

Loading Data:

```
data.head()
```

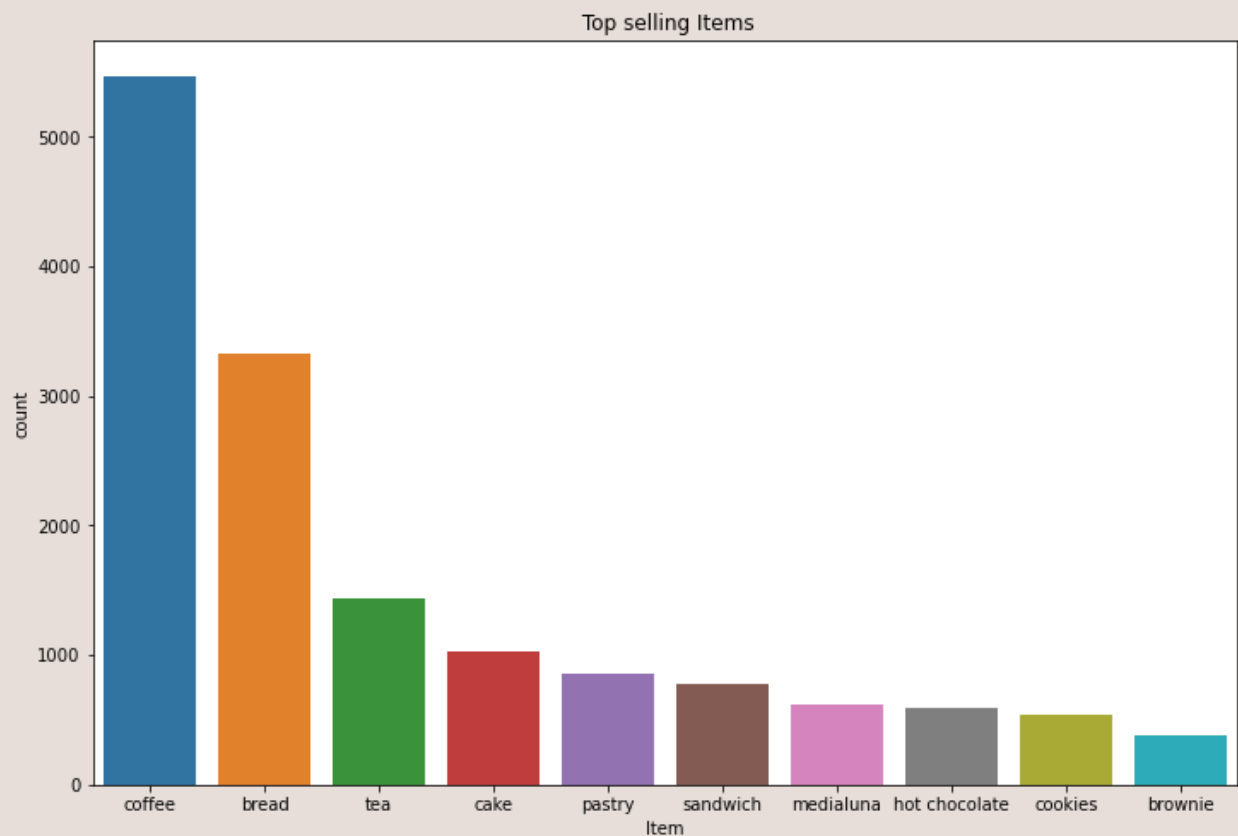
	Date	Time	Transaction	Item
0	2016-10-30	09:58:11	1	Bread
1	2016-10-30	10:05:34	2	Scandinavian
2	2016-10-30	10:05:34	2	Scandinavian
3	2016-10-30	10:07:57	3	Hot chocolate
4	2016-10-30	10:07:57	3	Jam

Understanding Data

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 21293 entries, 0 to 21292  
Data columns (total 4 columns):  
#   Column      Non-Null Count  Dtype  
---  -  
0   Date        21293 non-null  object  
1   Time        21293 non-null  object  
2   Transaction 21293 non-null  int64  
3   Item        21293 non-null  object  
dtypes: int64(1), object(3)  
memory usage: 665.5+ KB
```

Getting Count of Items:



Frequent Items Bought:

```
freq_item=apriori(a,min_support=0.01,use_colnames=True)
freq_item.sort_values(by='support',ascending=False)
```

	support	itemsets
6	0.478394	(coffee)
2	0.327205	(bread)
26	0.142631	(tea)
4	0.103856	(cake)
34	0.090016	(bread, coffee)
19	0.086107	(pastry)
21	0.071844	(sandwich)
16	0.061807	(medialuna)
12	0.058320	(hot chocolate)
42	0.054728	(cake, coffee)
8	0.054411	(cookies)
55	0.049868	(tea, coffee)
50	0.047544	(coffee, pastry)
3	0.040042	(brownie)
9	0.039197	(farm house)
15	0.038563	(juice)
18	0.038457	(muffin)
51	0.038246	(coffee, sandwich)

Association Rules:

A collection of items purchased by a customer is an itemset. The set of items on the left-hand side (sandwich, cookies in the example above) is the antecedent of the rule, while the one to the right (drink) is the consequent. The probability that the antecedent event will occur, i.e., a customer will buy a sandwich and cookies, is the support of the rule. That simply refers to the relative frequency that an itemset appears in transactions. In a QSR, the support of an item or item combination helps to identify keystone products. Hence, if a sandwich and cookies have high support, then they can be priced to attract people to the store.

- A lift greater than 1 suggests that the presence of the antecedent increases the chances that the consequent will occur in a given transaction
- Lift below 1 indicates that purchasing the antecedent reduces the chances of purchasing the consequent in the same transaction. Note: This could indicate that the items are seen by customers as alternatives to each other
- When the lift is 1, then purchasing the antecedent makes no difference on the chances of purchasing the consequent
























```
rules=association_rules(freq_item,metric='lift',min_threshold=1)
rules
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(coffee)	(alfajores)	0.478394	0.036344	0.019651	0.041078	1.130235	0.002264	1.004936
1	(alfajores)	(coffee)	0.036344	0.478394	0.019651	0.540698	1.130235	0.002264	1.135648
2	(bread)	(pastry)	0.327205	0.086107	0.029160	0.089119	1.034977	0.000985	1.003306
3	(pastry)	(bread)	0.086107	0.327205	0.029160	0.338650	1.034977	0.000985	1.017305
4	(coffee)	(brownie)	0.478394	0.040042	0.019651	0.041078	1.025860	0.000495	1.001080
5	(brownie)	(coffee)	0.040042	0.478394	0.019651	0.490765	1.025860	0.000495	1.024293
6	(cake)	(coffee)	0.103856	0.478394	0.054728	0.526958	1.101515	0.005044	1.102664
7	(coffee)	(cake)	0.478394	0.103856	0.054728	0.114399	1.101515	0.005044	1.011905
8	(hot chocolate)	(cake)	0.058320	0.103856	0.011410	0.195652	1.883874	0.005354	1.114125
9	(cake)	(hot chocolate)	0.103856	0.058320	0.011410	0.109868	1.883874	0.005354	1.057910

Support:

The support of item I is defined as the ratio between the number of transactions containing the item I by the total number of transactions expressed as :

Support indicates how popular an itemset is, as measured by the proportion of transactions in which an itemset appears. In Table 1 below, the support of {apple} is 4 out of 8, or 50%. Itemsets can also contain multiple items. For instance, the support of {apple, beer, rice} is 2 out of 8, or 25%.

Transaction 1	   
Transaction 2	  
Transaction 3	 
Transaction 4	 
Transaction 5	   
Transaction 6	  
Transaction 7	 
Transaction 8	 

If you discover that sales of items beyond a certain proportion tend to have a significant impact on your profits, you might consider using that proportion as your support threshold. You may then identify itemsets with support values above this threshold as significant itemsets.

Confidence:

This is measured by the proportion of transactions with item I1, in which item I2 also appears. The confidence between two items I1 and I2, in a transaction is defined as the total number of transactions containing both items I1 and I2 divided by the total number of transactions containing I1. (Assume I1 as X , I2 as Y)

$$\text{confidence}(X \rightarrow Y) = \frac{\text{Number of transactions containing } X \text{ and } Y}{\text{Number of transactions containing } X}$$

Confidence says how likely item Y is purchased when item X is purchased, expressed as $\{X \rightarrow Y\}$. This is measured by the proportion of transactions with item X, in which item Y also appears. In Table 1, the confidence of $\{\text{apple} \rightarrow \text{beer}\}$ is 3 out of 4, or 75%.

One drawback of the confidence measure is that it might misrepresent the importance of an association. This is because it only accounts for how popular apples are, but not beers. If beers are also very popular in general, there will be a higher chance that a transaction containing apples will also contain beers, thus inflating the confidence measure. To account for the base popularity of both constituent items, we use a third measure called lift.

Lift:

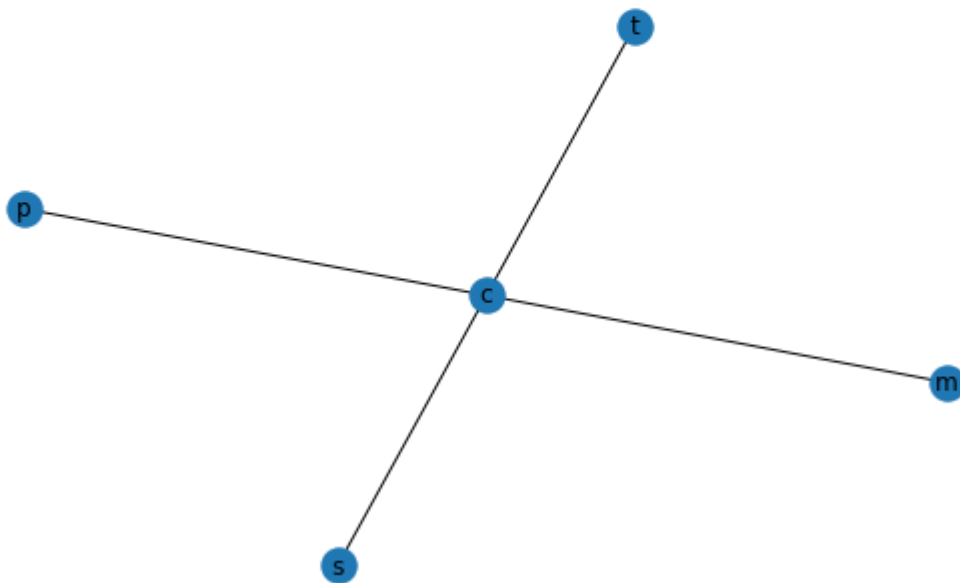
Lift is the ratio between the confidence and support.

$$\text{Lift} \{ \text{🍏} \rightarrow \text{🍺} \} = \frac{\text{Support} \{ \text{🍏}, \text{🍺} \}}{\text{Support} \{ \text{🍏} \} \times \text{Support} \{ \text{🍺} \}}$$

Lift says how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is. In Table 1, the lift of $\{\text{apple} \rightarrow \text{beer}\}$ is 1, which implies no association between items. A lift value greater than 1 means that item Y is likely to be bought if item X is bought, while a value less than 1 means that item Y is unlikely to be bought if item X is bought. (here X represents apple and Y represents beer)

Most Bought Together:

```
import networkx as nx
rules.antecedents=rules.antecedents.apply(lambda x: next(iter(x)))
rules.consequents=rules.consequents.apply(lambda x: next(iter(x)))
fig,ax=plt.subplots(figsize=(10,6))
ga=nx.from_pandas_edgelist(rules,source='antecedents',target='consequents')
nx.draw(ga,with_labels=True)
```



In addition its popularity as a retailer's technique, MBA is applicable in many other areas:

- Manufacturing: predictive analysis of equipment failure
- Pharmaceutical/Bioinformatics: discovery of co-occurrence relationships among diagnosis and pharmaceutical active ingredients prescribed to different patient groups
- Financial/Criminology: fraud detection based on credit card usage data

- Customer Behavior: associating purchases with demographic and socio-economic data
- More and more organizations are discovering ways of using market basket analysis to gain useful insights into associations and hidden relationships. As industry leaders continue to explore the technique's value, a predictive version of market basket analysis is making in-roads across many sectors in an effort to identify sequential purchases.

A) Feasibility

This project can be developed and deployed within a few years as SaaS(Software as a Service) for anyone to use.

B) Viability

As the retail industry grows in India and the world, there will always be small businesses existing which can use this service to improvise on their sales and data warehousing techniques. So, it is viable to survive in the long-term future as well but improvements are necessary as new technologies emerge.

C) Monetization

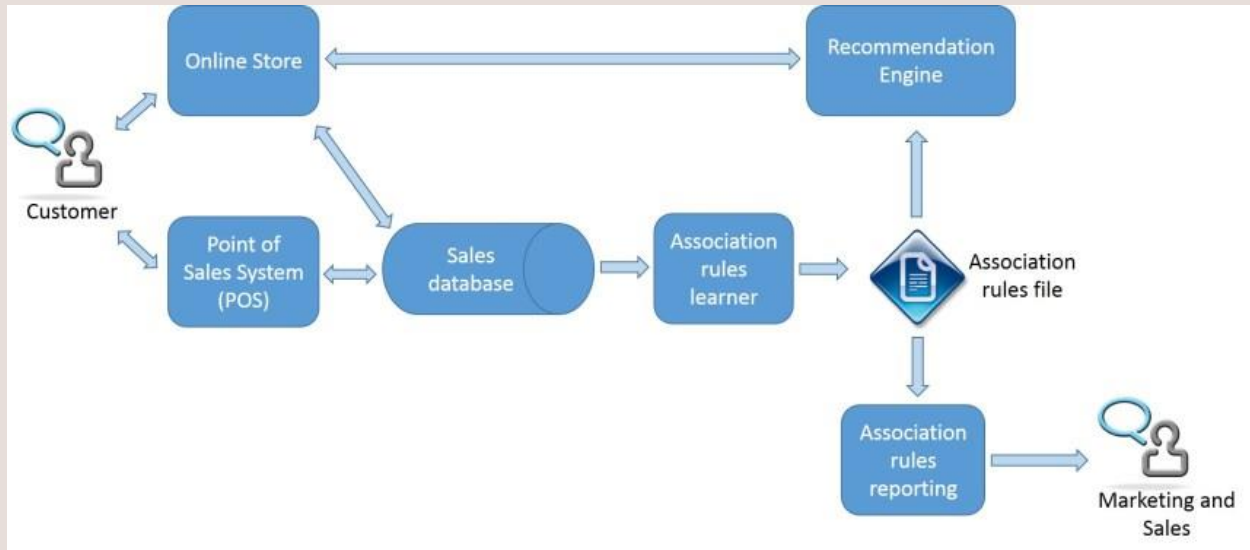
This service is directly monetizable as it can be directly released as a service on completion which can be used by businesses.

Step 2: Prototype Development

Git Link:

https://github.com/patilanirudh/ML_Unsupervised_Learning/blob/main/Apriori_Algorithm_Market_Basket_Analysis.ipynb

Step 3: Business Modeling



- First we will get the data of past customer behavior
- Then we will analyse and study their buying behavior
- Apply Association rules
- Provide the recommendations
- Sales will increase gradually
- Modify Accordingly

Conclusion:

At present many data mining algorithms have been developed and applied on variety of practical problems. However periodic mining is a new approach in data mining which has gained its significance these days. This field is evolving due to needs in different applications and limitations of data mining. This would enhance the power of existing data mining techniques.

Finding out the patterns due to changes in data is in itself an interesting area to be explored. It may helpful in Find out interesting patterns from large amount of data. Automatically track the changes in facts from previous data; due to this feature it may be helpful in fraud detection. Predicting future association rules as well as gives us right methodology to find out outliers.

In addition association rules will help to analyse the right product so as to increase the customers and their sales.