Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 24 Issue 1 Version 1.0 Year 2024
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Market Basket Analysis using Machine Learning

By Atul Sharma, Dr. Mohammad Salim Hamidi & Yousuf Hotak

Jahan University

Abstract- Market Basket Analysis is a technique used to analyse the items which are most likely to be purchased together mostly in the retail and economic sector. This technique is especially beneficial for optimization purpose. In this research paper we have used the market basket dataset from the Kaggle repository. This database has been analysed for very well know topic Market Basket Analysis using Python language.

Keywords: *market basket analysis, machine learning, python.*

GJCST-C Classification: *DDC Code: 006.31*

Strictly as per the compliance and regulations of:

Market Basket Analysis using Machine Learning

Atul Sharma [α], Dr. Mohammad Salim Hamidi [σ] & Yousuf Hotak [ρ]

Abstract- **Market Basket Analysis is a technique used to analyse the items which are most likely to be purchased together mostly in the retail and economic sector. This technique is especially beneficial for optimization purpose. In this research paper we have used the market basket dataset from the Kaggle repository. This database has been analysed for very well know topic Market Basket Analysis using Python language.**

Keywords: market basket analysis, machine learning, python.

I. Introduction

Market Basket Analysis is a valuable tool for businesses seeking to optimize their product offerings, increase cross-selling opportunities, and improve marketing strategies. Market basket analysis can be used to enhance the profitability of any business. Machine Learning is rewarding the retail industry in a unique way. It supports the retail sector in all areas, from predicting sales success to locating customers. Market basket analysis (MBA) is one such top retail application of machine learning. It helps retailers know what products people are purchasing together so that the store/website layout can be designed in the same manner.[1]

We have followed the below mentioned process for the task of Market Basket Analysis research project

- Gather transactional data, including purchase history, shopping carts, or invoices.
- Analyse product sales and trends.
- Use algorithms like Apriori or FP-growth to discover frequent item sets and generate association rules.
- Interpret the discovered association rules to gain actionable insights.
- Develop strategies based on the insights gained from the analysis.

II. Review of Literature

(Chaudhary, S. (2022, February 11) has talked about the importance of Market Basket Analysis in his research; (Stevens, S. (2023, September 7) has talked critically about the Data Analysis implication using Machine Learning; (Simplilearn. (2022, November 22) has discussed about the key components of the Market Basket Analysis; (McColl, L. (2022, March 1) has discussed about the Market Basket Analysis using Python; (*How to use market basket analysis for retail and marketing*. (2023, December 19) talks about the analysis of Market Basket analysis for retail sector; *Overview of market basket analysis*. (n.d.) discusses about the overview related to the Market basket analysis; Predoiu, O. (2024, April 2) talks about customer behavior analysis; Elnahla, N. (2021) discusses about Retail lance and its Marketing Implications with reference to Market Basket Analysis.

III. Research Methodology

We have worked on the Quantitative research. The past (historical) research data has been downloaded from the Kaggle repository for analysis. Now this data has been analyzed very effectively using Python language. According to Dawson (2019), a research methodology is the primary principle that will guide your research. It becomes the general approach in conducting research on your topic and determines what research method you will use. A research methodology is different from a research method because research methods are the tools you use to gather your data (Dawson, 2019). You must consider several issues when it comes to selecting the most appropriate methodology for your topic. Issues might include research limitations and ethical dilemmas that might impact the quality of your research.[2]

IV. Data Analysis & Interpretation

Even with years of professional experience working with data, the term "data analysis" still sets off a panic button in my soul. And yes, when it comes to serious data analysis for your business, you'll eventually want data scientists on your side. But if you're just getting started, no panic attacks are required.[3]

Author α: Asstt. Professor. e-mail: atul.sharma@jpemgzb.ac.in
Author σ: Vice Chancellor. Jahan University, Kabul, Afghanistan.
e-mail: avc@jahan.edu.af
Author ρ: Dean Jahan University, Kabul, Afghanistan.
e-mail: dean_cs@jahan.edu.af

```
Python 3.12.0 (tags/v3.12.0:0fb18b0, Oct  2 2023, 13:03:3
Type "help", "copyright", "credits" or "license" for more
>>> import pandas as pd
>>> import plotly.express as px
>>> import plotly.io as pio
>>> import plotly.graph_objects as go
>>> pio.templates.default = "plotly_white"
>>> data = pd.read_csv("E:/market_basket_dataset.csv")
>>> print(data.head())
   BillNo  Itemname  Quantity  Price  CustomerID
0    1000    Apples         5   8.30       52299
1    1000    Butter         4   6.06       11752
2    1000      Eggs         4   2.66       16415
3    1000  Potatoes         4   8.10       22889
4    1004   Oranges         2   7.26       52255
>>> _
```

Figure 1: Importing Utilities & Reading Dataset

Figure 1 above shows us steps to import utilities in Python which would be required for our Data analysis.

```
>>> print(data.isnull().sum())
BillNo        0
Itemname      0
Quantity      0
Price         0
CustomerID    0
dtype: int64
```

Figure 2: Verification of the Consistency of Data

Figure 2 above shows that we do not have any null data in our dataset.

Further we go ahead to check for Summary Statistics of the dataset as shown below (Figure 3).

```
>>> print(data.describe())
            BillNo    Quantity       Price    CustomerID
count   500.000000  500.000000  500.000000    500.000000
mean   1247.442000    2.978000    5.617660  54229.800000
std     144.483097    1.426038    2.572919  25672.122585
min    1000.000000    1.000000    1.040000  10504.000000
25%    1120.000000    2.000000    3.570000  32823.500000
50%    1246.500000    3.000000    5.430000  53506.500000
75%    1370.000000    4.000000    7.920000  76644.250000
max    1497.000000    5.000000    9.940000  99162.000000
```

Figure 3: Statistics for the Dataset

Figure 3 above shows the Statistical results of dataset.

Now let us look at the pictorial representation Sales Distribution of the items as.

```
>>> print(rules[['antecedents', 'consequents', 'support',
  antecedents consequents   support  confidence      lift
0    (Apples)     (Bread)  0.045752    0.280000  1.862609
1     (Bread)    (Apples)  0.045752    0.304348  1.862609
2    (Apples)    (Butter)  0.026144    0.160000  0.979200
3    (Butter)    (Apples)  0.026144    0.160000  0.979200
4    (Apples)    (Cereal)  0.019608    0.120000  0.592258
5    (Cereal)    (Apples)  0.019608    0.096774  0.592258
6    (Apples)    (Cheese)  0.039216    0.240000  1.311429
7    (Cheese)    (Apples)  0.039216    0.214286  1.311429
8    (Apples)   (Chicken)  0.032680    0.200000  1.530000
9   (Chicken)    (Apples)  0.032680    0.250000  1.530000
```

Figure 4: Sales Distribution

- *Antecedents:* These are the items that are considered as the starting point or "if" part of the association rule. Here is our case we have Bread, Butter, Cheese, and Chicken as the antecedents in our analysis. The entities or "itemsets" produced from the data are called antecedents. To put it another way, it's the IF element on the left. In the situation before, bread serves as the antecedent.[4]
- *Consequent:* These are the items that tend to be purchased along with the antecedents or the "then" part of the association rule. The term "consequent" refers to an item or group of items that are encountered along with the antecedent. The THEN part of the sentence is displayed on the right-hand side. The result in the aforementioned case is butter.[5]
- *Support:* Support measures how frequently a particular combination of items (both antecedents and consequents) appears in the dataset. It refers to the proportion of transactions in which the items are expected to be bought together. For example, the first rule indicates that Bread and Apples are bought together in approximately 4.58% of all transactions. Support refers to the frequency or occurrence of a specific combination of items in the dataset. Thus indicates frequency of item set appearing in the transactions being analyzed.[6]
- *Confidence:* Confidence quantifies the likelihood of the consequent item being purchased when the antecedent item is already in the basket. Alternately it shows the probability of buying the subsequent item wherein the antecedent item is already in the basket. Figure above shows that there is a 30.43% chance of buying Apples when Bread is already kept in the basket after purchase. The probability that a transaction that contains the items on the left hand side of the rule (in our example, pencil and paper) also contains the item on the right hand side (a rubber). The higher the confidence, the greater the likelihood that the item on the right hand side will be purchased or, in other words, the greater the return rate you can expect for a given rule.[7]
- *Lift:* Lift measures the degree of association between the antecedent and consequent items, while considering the baseline purchase probability of the consequent item. If we find a lift with a value greater than 1 then this would indicate a positive association between the antecedent and the consequent item then it would indicate that the items are most likely to be bought together rather than independently. If we obtain a value which is less than 1 then it would indicate a negative association between the two. We can find a lift of 1.86 suggests a positive association between Bread and Apples. Lift is the measure of the effect of purchasing item A on purchasing item B. It is used to determine whether the combination of items has practical value. In other words, it is used to see if the combination of items is purchased more frequently than the individual items. If the value is greater than 1, it means that the combination is effective, while if it is less than 1, it means that it is ineffective.[8]

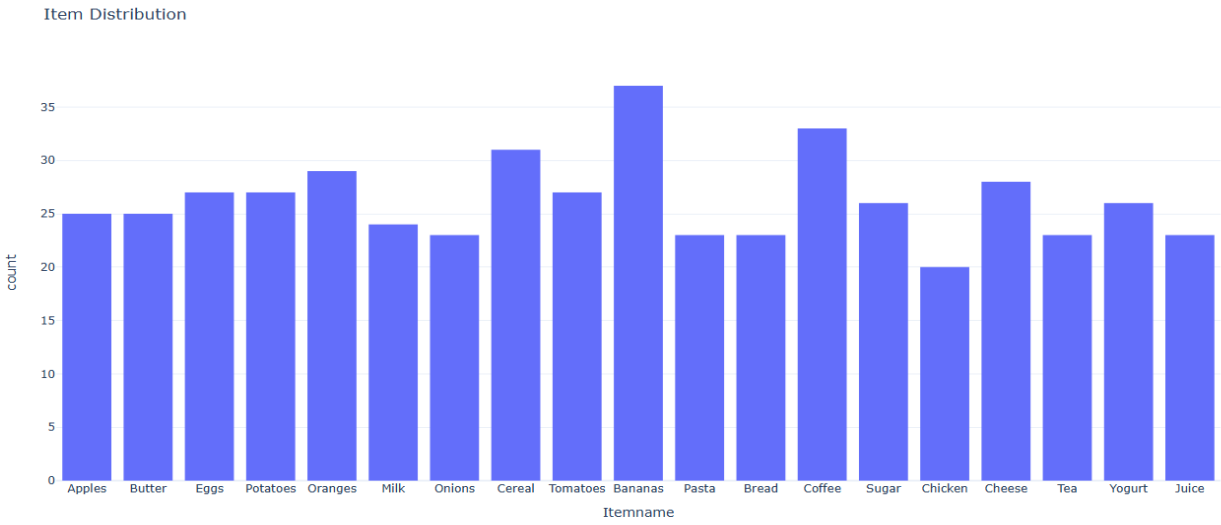

Figure 5: Item Distribution

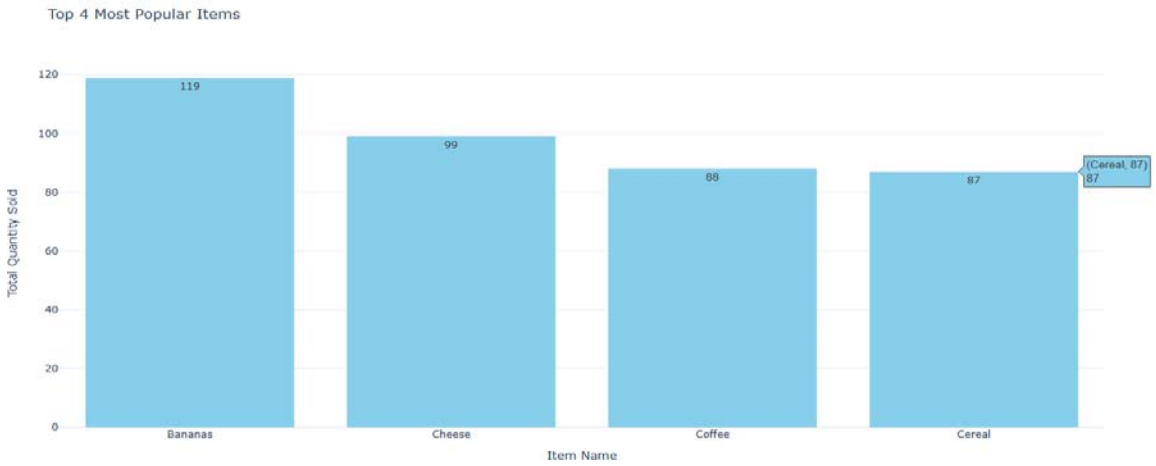

Figure 6: Top four Most Popular Items

It is observed that bananas are the most popular item sold in the store.

Understanding Customer Behavior.

Figure 7: Understanding Customer Behavior

By the term customer behavior, we understand the trends in the buying habits and factors which influence the decision to buy something else along with previous item. Here in Figure 7 above we explore customer behavior by comparing average quantity and total spending. Customer Behavior Analysis represents the study of how people make buying decisions concerning a product, service, and /or organization.[9]

V. Conclusion, Implications, and Scope For Future Research

Henceforth it may be concluded that the historic data can be analyzed very effectively using Python language which is highly flexible and simple. This data analysis would be highly beneficial to end users in terms of decision-making in the future. They can very easily plan out their investment based upon the results that have been obtained with the help of this application. It would help them to have a better decision-making which would result in generating more profits. Since Market Basket Analysis is a highly productive tool to optimize the selling opportunities hence this project becomes utmost important. In the near future we would design a model wherein the predictions can be made beforehand. Artificial intelligence has revolutionized market basket analysis by automating the process of data analysis and rule discovery.[10]

Acknowledgement

We would like to express our deepest gratitude to my adviser, Professor Mamta Bansal, for her invaluable guidance and support throughout this research. Her expertise and dedication have been a source of inspiration and motivation.

References Références Referencias

1. Chaudhary, S. (2022, February 11). Understanding market basket analysis in data mining. Hire the World's Most Deeply Vetted Developers & Teams Turing. https://www.turing.com/kb/market-basket-analysis.
2. Stevens, S. (2023, September 7). What is data analysis? Examples and how to get started. https://zapier.com/blog/data-analysisexam- ple/
3. Simplilearn. (2022, November 22). *What is market basket analysis? Overview, uses, types, and examples* Simplilearn.com. https://www.simplilearn.com/what-is-market-basket-analysis-article.
4. Simplilearn. (2022, November 22). *What is market basket analysis? Overview, uses, types, and examples*.Simplilearn.com. https://www.simplilearn.com/what-is-market-basket-analysis-article.
5. *How to use market basket analysis for retail and marketing*. (2023, December 19). Thought Spot. https://www.thoughtspot.com/data-rends/analytics/market-basket-analysis.
6. McColl, L. (2022, March 1). *Market basket analysis: Understanding customer behaviour*. Select Statistical Consultants. https://select-statistics.co.uk/blog/market-basket-analysisunderstanding-customer-behaviour/
7. *Overview of market basket analysis*. (n.d.). Modern Big Data Analytics & BI Software - Fine BI. https://intl.finebi.com/blog/market-basket-analysis

8. Predoiu, O. (2024, April 2). Customer Behavior analysis. Omniconvert Ecommerce Growth Blog. https://www.omniconvert.com/blog/customer-behavior-analysis/
9. Elnahla, N. (2021). Your Retailer Needs You: Retail lance and its Marketing Implications. https://doi.org/10.22215/etd/2021-14633