Application of Apriori Algorithm on the given MBA Dataset

Jehezkel Rosh T. Rodriguez
College of Computer Studies
Silliman University
jehezkeltrodriguez@su.edu.ph

I. Abstract

Market Basket Analysis (MBA) serves as a fundamental tool for uncovering significant patterns within consumer purchasing behavior. In this study, The researcher used the Apriori algorithm to analyze the dataset, seeking to extract meaningful associations among items frequently bought together. The Apriori algorithm, known for its effectiveness in identifying frequent itemsets, is employed to reveal latent relationships and dependencies within the dataset.

The method in this study involves dataset preprocessing, application of the Apriori algorithm to uncover frequent itemsets, and the subsequent generation of association rules. By setting a minimum support threshold, noisy data are filtered out. The resulting association rules are evaluated based on confidence, providing a comprehensive understanding of product relationships.

The selected dataset, representing authentic retail transactions, forms the basis in the analysis. This research aims to contribute to the field of market basket analysis by showcasing the practical application of the Apriori algorithm and its ability to uncover valuable insights from transactional data.

The researcher then found out after the implementation of the algorithm, out of 14,964 transactions, there are 125 frequent itemsets and in which 19 generated association rules are derived with the minimum support of 0.005 and minimum confidence level of 0.1. This sheds a light to the purchasing habits of customers which then can be harnessed to optimize various components in business operations.

II. Introduction

The analysis of grocery data and association rule mining has become increasingly crucial in the world of retail and marketing. With the vast amounts of transactional data generated in grocery stores, understanding consumer purchasing patterns has emerged as a key strategy for businesses. Association rule mining, particularly through algorithms like Apriori, offers a powerful approach to uncovering hidden relationships among products. By identifying frequently co-occurring items in transactions, businesses can gain insights into consumer behavior, optimize inventory management, and enhance the overall shopping experience.

Statement of the Problem

The researcher aims to address the following research question: **What are the association rules generated from the Apriori algorithm when applied to the given MBA datasets?**

III. Review of Related Literature

Market Basket Analysis has been widely researched and applied in various industries, particularly in retail. The foundational work by Agrawal, Imielinski, and Swami [1] introduced the Apriori algorithm, which has since become a cornerstone in association rule mining. The technique involves identifying frequent itemsets and generating association rules, providing valuable insights into customer purchasing patterns. Several studies [2] have explored variations of the Apriori algorithm and its applications in different domains.

The Apriori algorithm, proposed by Agrawal et al. [1], remains a key algorithm in association rule mining. The algorithm efficiently discovers frequent itemsets and generates association rules based on a priori knowledge. Numerous studies have extended and optimized the Apriori algorithm, such as the FP-Growth algorithm proposed by Han et al.[2], which addresses some of the limitations of the original Apriori algorithm.

Python has become a popular programming language for data science and machine learning due to its simplicity and versatility. The 'pandas' library is a fundamental tool for data manipulation and analysis. Wes McKinney's book [3], "Python for Data Analysis," provides a comprehensive guide to using 'pandas' for data-related tasks.

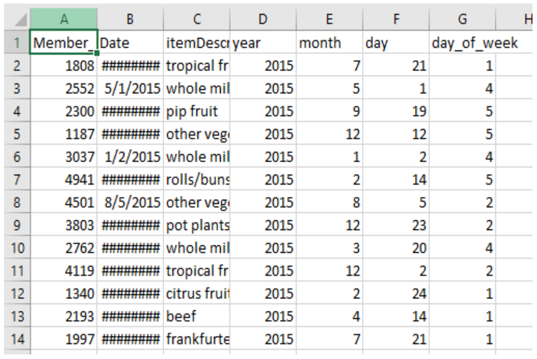
[4] The 'mlxtend' library is an extension of 'scikit-learn' and provides tools for exploratory data

analysis, feature engineering, and association rule mining. Raschka [5] provides insights into 'mlxtend' and its applications in machine learning in his book "Python Machine Learning."

IV. Methodology

Dataset

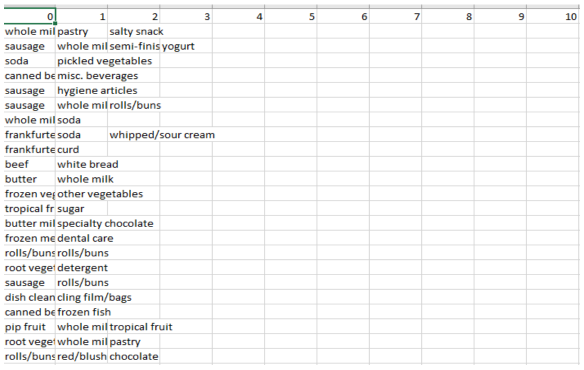
The [6] grocery dataset used in this study is provided to the researcher. The dataset can also be accessed through kaggle, providing transparency and reproducibility. The dataset contains the details member number, date purchased, item description, year, month, day and day of the week.



	A	B	C	D	E	F	G	H
1	Member_	Date	itemDescı	year	month	day	day_of_week	
2	1808	########	tropical fr	2015	7	21	1	
3	2552	5/1/2015	whole mil	2015	5	1	4	
4	2300	########	pip fruit	2015	9	19	5	
5	1187	########	other veg	2015	12	12	5	
6	3037	1/2/2015	whole mil	2015	1	2	4	
7	4941	########	rolls/buns	2015	2	14	5	
8	4501	8/5/2015	other veg	2015	8	5	2	
9	3803	########	pot plants	2015	12	23	2	
10	2762	########	whole mil	2015	3	20	4	
11	4119	########	tropical fr	2015	12	2	2	
12	1340	########	citrus fruit	2015	2	24	1	
13	2193	########	beef	2015	4	14	1	
14	1997	########	frankfurte	2015	7	21	1	

Figure 1: A snippet of the uncleaned MBA dataset.

Prior to applying the Apriori algorithm, the dataset underwent preprocessing to ensure data quality and relevance. Steps included handling missing values, removing duplicates, and encoding categorical variables. Additionally, outliers were identified and addressed to mitigate their impact on the association rule mining process. The cleaned dataset was then formatted into a suitable structure for input into the Apriori algorithm.

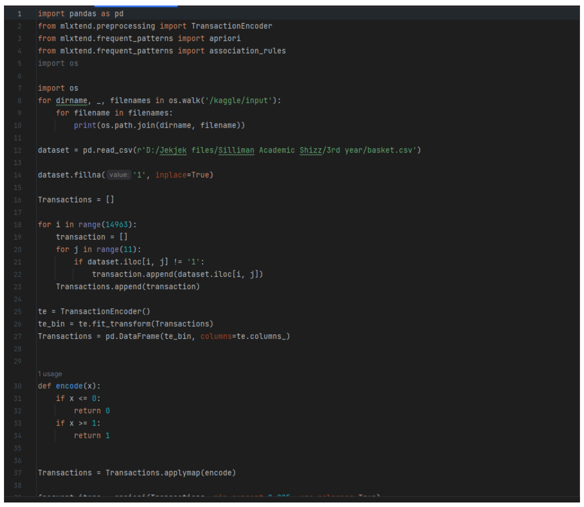


0	1	2	3	4	5	6	7	8	9	10
whole mil	pastry	salty snack								
sausage	whole mil	semi-finis	yogurt							
soda	pickled vegetables									
canned be	misc. beverages									
sausage	hygiene articles									
sausage	whole mil	rolls/buns								
whole mil	soda									
frankfurte	soda	whipped/sour cream								
frankfurte	curd									
beef	white bread									
butter	whole milk									
frozen veg	other vegetables									
tropical fr	sugar									
butter mil	specialty chocolate									
frozen me	dental care									
rolls/buns	rolls/buns									
root veget	detergent									
sausage	rolls/buns									
dish clean	cling film/bags									
canned be	frozen fish									
pip fruit	whole mil	tropical fruit								
root veget	whole mil	pastry								
rolls/buns	red/blush	chocolate								

Figure 2: A snippet of the cleaned MBA dataset.

The Apriori algorithm was implemented using the 'mlxtend' library in Python. The algorithm follows a two-step process: (1) identifying frequent itemsets, and (2) generating association rules. The 'mlxtend' implementation efficiently navigates the dataset, employing the Apriori principle to iteratively discover frequent itemsets of varying lengths. Subsequently, association rules were generated based on these frequent itemsets, considering the specified minimum support and confidence thresholds.

The choice of support and confidence thresholds is a critical aspect of association rule mining. In this study, the support threshold was set to 0.005, reflecting the minimum proportion of transactions that must contain a specific itemset for it to be considered frequent. A confidence threshold of 0.1 was chosen to ensure that only rules with a strong association between antecedent and consequent items were retained. These parameter settings were determined through an iterative process, balancing the need for meaningful associations with computational efficiency.



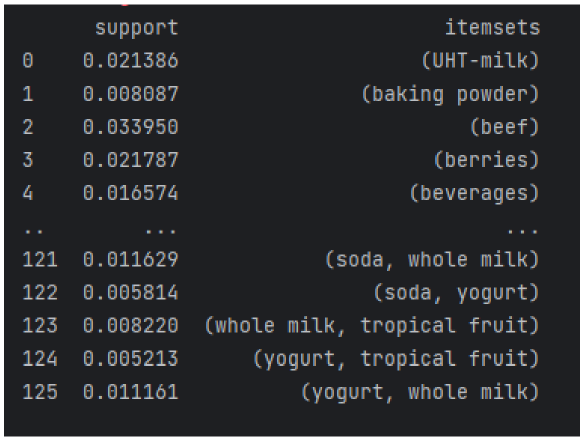
```
import pandas as pd
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules
import os

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

dataset = pd.read_csv(r'D:/Jekjek files/Silliman Academic Shizz/3rd year/basket.csv')

dataset.fillna( value: '1', inplace=True)

Transactions = []

for i in range(14963):
    transaction = []
    for j in range(11):
        if dataset.iloc[i, j] != '1':
            transaction.append(dataset.iloc[i, j])
    Transactions.append(transaction)

te = TransactionEncoder()
te_bin = te.fit_transform(Transactions)
Transactions = pd.DataFrame(te_bin, columns=te.columns_)


1 usage
def encode(x):
    if x <= 0:
        return 0
    if x >= 1:
        return 1


Transactions = Transactions.applymap(encode)

```

Figure 3: A code snippet of the python program showing 'pandas' and 'mlxtend' being used in the algorithm.

V. Results

Frequent Itemsets

The application of the Apriori algorithm on the grocery dataset resulted in the discovery of 126 frequent itemsets. These itemsets represent combinations of products that are frequently purchased together by customers. The support values, indicating the proportion of transactions containing each itemset, provide insights into the popularity of these combinations.



```
     support                      itemsets
0    0.021386                    (UHT-milk)
1    0.008087               (baking powder)
2    0.033950                        (beef)
3    0.021787                     (berries)
4    0.016574                   (beverages)
..        ...                           ...
121  0.011629            (soda, whole milk)
122  0.005814                (soda, yogurt)
123  0.008220  (whole milk, tropical fruit)
124  0.005213      (yogurt, tropical fruit)
125  0.011161          (yogurt, whole milk)
```

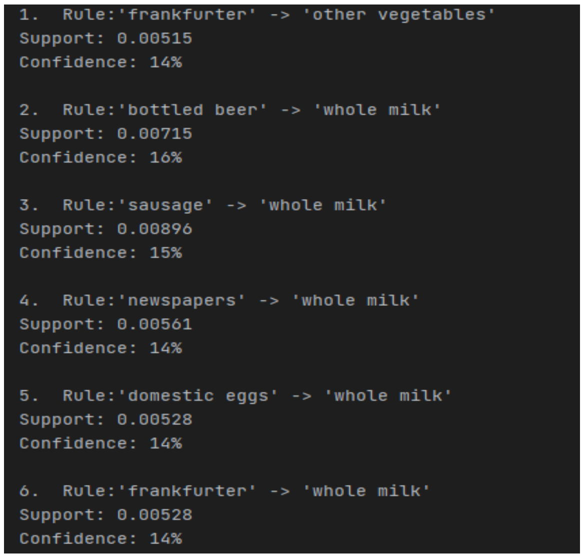
Figure 4: Results in python showing the first and the last 5 frequent itemsets.

Association Rules

There are 19 association rules derived from the frequent itemsets that contribute to the understanding of customer behavior. Each rule consists of an antecedent and a consequent, along with confidence and support as its metrics. The association rules are as follows:

Rule	Support	Confidence
Frankfurter -> other vegetables	0.00515	14%
Bottled beer -> whole milk	0.00715	16%
Sausage -> whole milk	0.00896	15%
Newspaper -> whole milk	0.00561	14%
Domestic eggs -> whole milk	0.00528	14%
Frankfurter -> whole milk	0.00528	14%
Pork -> whole milk	0.00501	14%
Pip fruit -> whole milk	0.00662	13%
Citrus fruit -> whole milk	0.00715	13%
Shopping bags -> whole milk	0.00635	13%
Yogurt -> whole milk	0.01116	13%
Canned beer -> whole milk	0.00601	13%
roll/buns -> whole milk	0.01397	13%
Pastry -> whole milk	0.00648	13%
Other vegetables -> whole milk	0.01484	12%
Tropical fruit -> whole milk	0.00822	12%
Soda -> whole milk	0.01163	12%
Bottled water -> whole milk	0.00715	12%
Root vegetables -> whole milk	0.00755	11%

Table 1: Association rules with its support and confidence



```
1.  Rule:'frankfurter' -> 'other vegetables'
Support: 0.00515
Confidence: 14%

2.  Rule:'bottled beer' -> 'whole milk'
Support: 0.00715
Confidence: 16%

3.  Rule:'sausage' -> 'whole milk'
Support: 0.00896
Confidence: 15%

4.  Rule:'newspapers' -> 'whole milk'
Support: 0.00561
Confidence: 14%

5.  Rule:'domestic eggs' -> 'whole milk'
Support: 0.00528
Confidence: 14%

6.  Rule:'frankfurter' -> 'whole milk'
Support: 0.00528
Confidence: 14%
```

Figure 5: Results in python showing the first 6 out of 19 association rules derived from the frequent itemsets.

Analysis

The association rules derived from the analysis of the dataset reveal several patterns in customer purchasing behavior. Remarkable items like 'bottled beer,' 'sausage,' 'newspapers,' and 'domestic eggs' are frequently associated with the purchase of 'whole milk,' each exhibiting a confidence ranging from 14% to 16%. This indicates a consistent preference for combining these products in shopping transactions. Additionally, common grocery items such as 'rolls/buns,' 'pastry,' 'other vegetables,' and 'tropical fruit' also exhibit a significant association with 'whole milk,' ranging from 11% to 13%. This suggests that customers tend to purchase these items together.

Conclusion

The association rule analysis provides valuable information into the purchasing patterns of the customers in the dataset. The frequent associations between items like ‘bottled beer,’ ’sausage’ and ‘newspaper’ with ‘whole milk’ shows a consistent consumer preference offering the retailers actionable information for product placement and targeted marketing. With the confidence of each rule consistently staying above the minimum threshold, this shows that the reliability of each rule is dependable. These findings contribute to the data-driven understanding of consumer behavior in the grocery retail area, emphasizing the importance of association rule mining.

IV. References

[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. *ACM SIGMOD Record* 22, 2 (1993), 207-216. DOI:https://doi.org/10.1145/170036.170072

[2] Jiawei Han, Jian Pei, and Yiwen Yin. 2000. Mining frequent patterns without candidate generation. *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (2000). DOI:https://doi.org/10.1145/342009.335372

[3] McKinney, W. (2017). Python for Data Analysis. *O'Reilly Media*.

[4]Raschka, S. (2018). Python Machine Learning. *Packt Publishing*.

[5]Raschka, S. (2021). mlxtend: Extensions and Helper Functions for Python's Data Analysis Library. *Journal of Open Source Software.*

[6]Rashik Rahman. 2020. Groceries dataset for Market Basket Analysis(MBA). *kaggle*. Retrieved November 19, 2023 from https://www.kaggle.com/datasets/rashikrahmanpritom/groceries-dataset-for-market-basket-analysismba/