Advanced Market Basket Analysis

Mostafa Dorrah[1], Ahmed Ashraf[2], Abdallah ElSaadany[3], Yousif Khalil[4], Mohamed Adel[5], Omar Tarek[6]

ITCS-AI Department, Nile University
AIS331

Under Supervision of:
Dr. Mona Arafa

***Abstract*— Making sense of the data and devising marketing strategies has become more important in the modern world as more people and transactions take place. Marketing tactics are greatly influenced by the data's hidden patterns, which can be revealed in order to improve performance and maximise profit in the face of fierce market competition and build value-driven, long-term connections with customers. For learning association rules, various algorithms can be applied. such as the FP-Growth algorithm and the Apriori algorithm. In this project, the "Apriori Algorithm" and "FP-Growth" will be used to handle product association analysis, and the sales data of an e-commerce company will be used to make the best product offers for the customer who is in the sales process.**

I. Introduction

Offering the appropriate product to the right consumer at the right time is the foundation of cross-selling and loyalty programmes within the context of customer retention and growing lifetime value; the development of rule-based strategies is no longer conceivable in the era of big data. Therefore, it has become essential for businesses to use these patterns of association while offering their products and to create efficient marketing plans. One of the uses of association rules is market basket analysis. By creating a pattern from prior consumer behaviour and habits, it enables us to forecast the goods that customers are likely to purchase in the future. In this project, the "Apriori Algorithm" and "FP-Growth" will be used to handle product association analysis. And the dataset used for this project is the online retail II dataset, which includes the sales data of the UK-based online sales store with more than 500,000 data-point, the sales data was between 2010 to 2011. To suggest products to users at the basket stage using both Apriori Algorithm and FP-Growth Algorithm we will have to consider the work in 5 main steps: Data Pre-processing and feature engineering, analyzing both algorithms and then apply it to the data, checking and analyzing the result, trying different threshold and min. support, and evaluating and visualizing the final results with analysis.

II. Literature review

The Apriori method is a well-known and popular data mining technique for extracting association rules from huge datasets. The technique has been the subject of multiple research investigations since it was first introduced by Agrawal et al. in 1994 with the goal of enhancing its effectiveness, scalability, and applicability to other fields.The creation of effective data structures is a big improvement to the Apriori method. Due to memory and computational constraints brought on by the enormous number of itemsets produced during the candidate generation phase, traditional systems frequently had these problems. Researchers have suggested novel data structures to address this problem, including hash-based techniques, tree-based structures, and vertical data formats. Pruning tactics are a further area of development. The so-called "combinatorial explosion" problem, where the search space grew exponentially with the number of items and itemsets, plagued early versions of the Apriori algorithm. Pruning approaches have been developed by researchers to lessen this issue's impact and get rid of useless itemsets. Anti-monotonicity, subset testing, and multi-level filtering are some of these techniques that effectively prune the search space depending on the lowest support criterion. These pruning methods greatly increase the algorithm's runtime effectiveness by lowering the number of candidates itemsets.

The Han et al.-developed FP-Growth algorithm is a well-liked technique for effective and scalable mining of frequent itemsets in huge datasets. It uses the FP-tree, a small data structure that speeds up the development of frequent itemsets without creating candidate itemsets as the Apriori method does. The creation of the FP-tree, the production of frequent itemsets by mining conditional patterns, and the post-processing steps to extract the entire set of frequent itemsets are all covered in this section's description of the essential processes of the FP-Growth method. The novel data structure called the FP-tree, which is part of the FP-growth algorithm, is a noteworthy development. The FP-growth approach creates a compact and condensed data structure that captures the dataset's frequent patterns, as opposed to the Apriori algorithm, which generates a vast number of candidate itemsets. The FP-tree eliminates the need for pricey candidate development and several database scans, enabling effective and quick mining of frequent itemsets. The efficiency of the programme is substantially improved by this specific data structure, especially for datasets with high dimensionality and lots of transactions. The mining procedure used by the FP-growth algorithm has also been optimized. Other solutions have been suggested to enhance the algorithm's functionality and scalability. These include parallel and distributed implementations that use numerous processors and distributed computing frameworks to speed up the processing and handling of large datasets. In order to increase the algorithm's effectiveness even further, optimizations such as pruning tactics, post-processing methods, and memory management strategies have been created.

A data mining algorithm called H-mine [1] (memory-based hyperstructure mining of frequent patterns) is used for frequent itemset mining, which is the process of identifying patterns that regularly appear in huge transactional datasets.
H-mine performs better in terms of time and space complexity compared to the Apriori and FP-Growth algorithms. The H-struct data structure and a more effective search space traversal technique are used to accomplish this.
With the use of a brand-new data structure called the H-struct, the H-mine algorithm enhances the FP-Growth algorithm. It is more effective for frequent itemset mining since the H-struct is a hybrid data structure that combines the advantages of both horizontal and vertical data layouts.

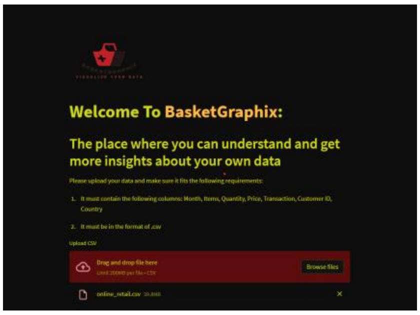
The advantages of both horizontal and vertical data layouts are combined in the hybrid data structure known as the H-struct (Hyper-structure). For particular kinds of datasets, it tries to maximise the effectiveness of data storage and retrieval

In a horizontal data architecture, the attributes or fields are stored sequentially within each record, which takes up a contiguous block of memory. This design is advantageous for tasks that require simultaneous access to all of a record's properties.
In a vertical data architecture, all of the records' relevant values are stored together but each attribute or field is recorded separately. This design is advantageous for procedures that require swiftly accessing a particular attribute across several records.
The H-struct strategically combines both layouts to benefit from them. It divides the dataset into segments, each of which has a subset of the dataset's records and properties. The ability to store and access each segment separately makes it possible to efficiently retrieve particular properties for a subset of data.

A new and promising data mining method called the H-Mine algorithm focuses on extracting high-dimensional, high-quality patterns from large datasets. Despite being new, it has garnered considerable interest from experts in the field. The H-Mine algorithm tackles the problems brought on by high-dimensional data, where conventional data mining algorithms frequently fail because of the dimensionality curse. This algorithm seeks to extract patterns from high-dimensional spaces that are highly relevant, significant, and intriguing in order to provide more precise and insightful data. The H-Mine algorithm's capacity to manage the combinatorial expansion of itemsets in high-dimensional datasets is a crucial feature. To limit the search space and increase computing efficiency, it makes use of sophisticated pruning techniques and effective data structures. Advanced statistical and quality measures are also incorporated into the H-Mine algorithm to guarantee that the patterns found are both statistically significant and have high-quality attributes. Researchers and practitioners can concentrate on patterns that really do represent important linkages and interactions within the data by using these measurements to help filter out noise and irrelevant patterns. The algorithm improves the accuracy and value of the found patterns by taking into account both statistical significance and quality measures.

III. METHODOLOGY

In our methodology we are going to implement an application with Stream lit and python. The application is designed to work on any data with some constrains.



Welcome To BasketGraphix:

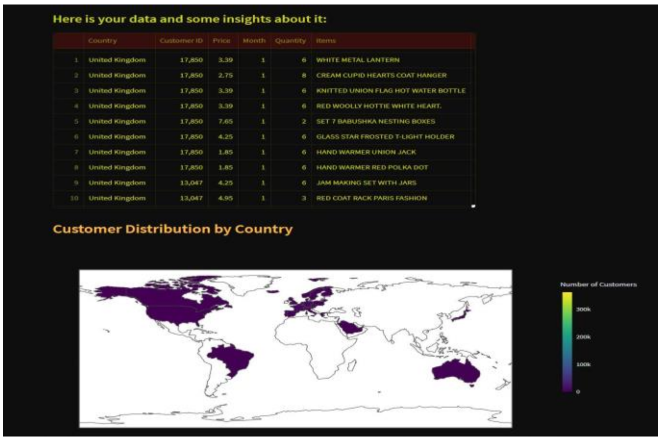
The place where you can understand and get more insights about your own data

Please upload your data and make sure it fits the following requirements:

1. It must contain the following columns: Month, Items, Quantity, Price, Transaction, Customer ID, Country
2. It must be in the format of .csv

Upload CSV

In the start we take your data and make sure it meets the requirements.

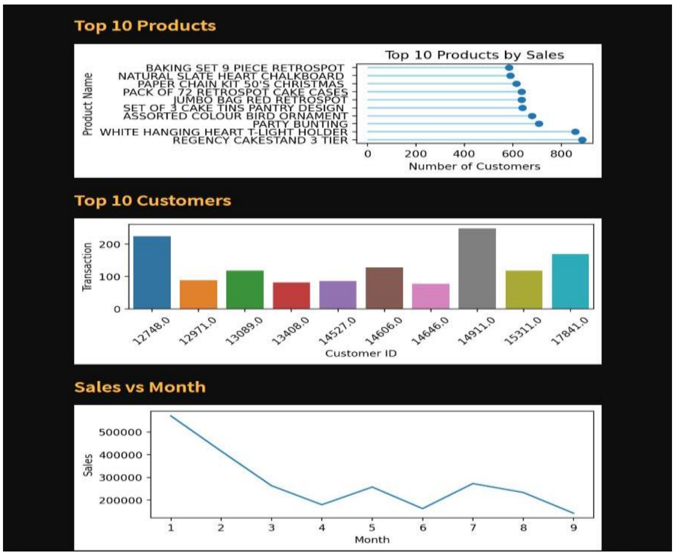


Here is your data and some insights about it:

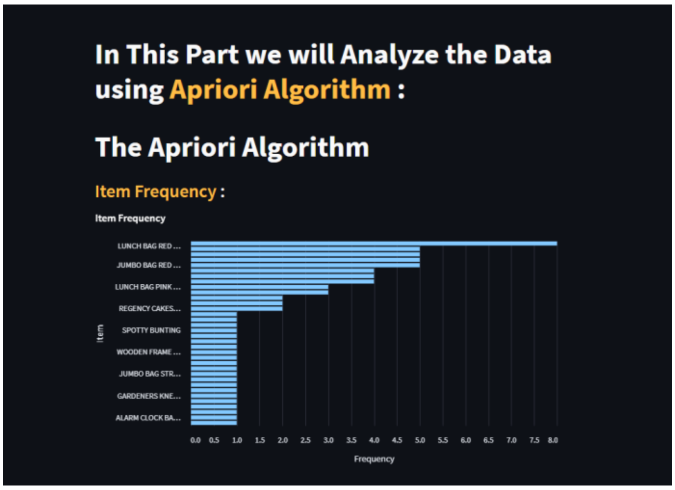
	Country	Customer ID	Price	Month	Quantity	Items
1	United Kingdom	17,850	3.39	1	6	WHITE METAL LANTERN
2	United Kingdom	17,850	2.75	1	8	CREAM CUPID HEARTS COAT HANGER
3	United Kingdom	17,850	3.39	1	6	KNITTED UNION FLAG HOT WATER BOTTLE
4	United Kingdom	17,850	3.39	1	6	RED WOOLLY HOTTIE WHITE HEART.
5	United Kingdom	17,850	7.65	1	2	SET 7 BABUSHKA NESTING BOXES
6	United Kingdom	17,850	4.25	1	6	GLASS STAR FROSTED T-LIGHT HOLDER
7	United Kingdom	17,850	1.85	1	6	HAND WARMER UNION JACK
8	United Kingdom	17,850	1.85	1	6	HAND WARMER RED POLKA DOT
9	United Kingdom	13,047	4.25	1	6	JAM MAKING SET WITH JARS
10	United Kingdom	13,047	4.95	1	3	RED COAT RACK PARIS FASHION

Customer Distribution by Country

Then we clean the data and view it. Then we start the graphs section, First we graph the Customers vs the countries to view where are the most sales.



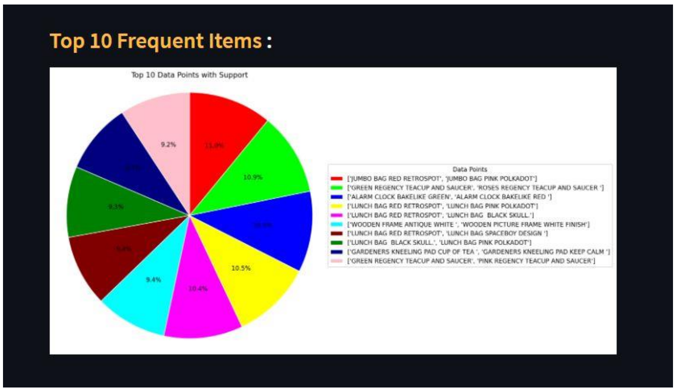
The other 3 graphs are we have Products vs Customers, most frequent Customers and the total sales in each month.


In This Part we will Analyze the Data using Apriori Algorithm :

The Apriori Algorithm

Item Frequency :

Then we applied the apriori , FPgrowth, and Hmine on the data to get more information out of it.


Top 10 Frequent Items :
Top 10 Data Points with Support
9.2%
10.9%
9.3%
9.4%
10.4%
10.5%
Data Points
['JUMBO BAG RED RETROSPOT', 'JUMBO BAG PINK POLKADOT']
['GREEN REGENCY TEACUP AND SAUCER', 'ROSES REGENCY TEACUP AND SAUCER ']
['ALARM CLOCK BAKELIKE GREEN', 'ALARM CLOCK BAKELIKE RED ']
['LUNCH BAG RED RETROSPOT', 'LUNCH BAG PINK POLKADOT']
['LUNCH BAG RED RETROSPOT', 'LUNCH BAG BLACK SKULL.']
['WOODEN FRAME ANTIQUE WHITE ', 'WOODEN PICTURE FRAME WHITE FINISH']
['LUNCH BAG RED RETROSPOT', 'LUNCH BAG SPACEBOY DESIGN ']
['LUNCH BAG BLACK SKULL.', 'LUNCH BAG PINK POLKADOT']
['GARDENERS KNEELING PAD CUP OF TEA ', 'GARDENERS KNEELING PAD KEEP CALM ']
['GREEN REGENCY TEACUP AND SAUCER', 'PINK REGENCY TEACUP AND SAUCER']

IV. Conclusion

In conclusion, we found that on the data we used to test our application that all three algorithms have close output. But it all depends on the data provided that is why we tried making our application general as possible. In the future we want to try and make a combination between the 3 algorithms for better results.

References

[1] "Example: Mining Frequent Itemsets Using the HMine Algorithm (SPMF - Java)." Www.philippe-Fournier-Viger.com, www.philippe-fournier-viger.com/spmf/HMine.php. Accessed 27 May 2023.

[2] GeeksforGeeks. "Apriori Algorithm - GeeksforGeeks." GeeksforGeeks, 4 Sept. 2018, www.geeksforgeeks.org/apriori-algorithm/..

[3] "Mlxtend.frequent Patterns - Mlxtend." Rasbt.github.io, rasbt.github.io/mlxtend/api_subpackages/mlxtend.frequent_patterns/. Accessed 27 May 2023.

[4] Zeng, Yi, et al. "Research of Improved FP-Growth Algorithm in Association Rules Mining." Scientific Programming, vol. 2015, 2015, pp. 1–6, https://doi.org/10.1155/2015/910281. Accessed 9 Jan. 2023.

[5] "Hmine-Mlxtend." *Rasbt.github.io*, rasbt.github.io/mlxtend/user_guide/frequent_patterns/hmine/. Accessed 27 May 2023.

[6] Science, International Journal of Scientific Research in, and Engineering and Technology Ijsrset. "Using Hyper-Structure Mining to Ascertain Recurrent Patterns in Large Dataset." *Www.academia.edu*, www.academia.edu/25501152/Using_Hyper_structure_mining_to_Ascertain_Recurrent_Patterns_in_Large_Dataset.