[Market Basket Analysis Using Association Rules and PySpark]

**Submission date: *[15-04-2024]***

Contents

[**1. ABSTRACT** 1](#_Toc164039209)

[**2. INTRODUCTION** 2](#_Toc164039210)

[**3**. **DATASET** 4](#_Toc164039211)

[**4. METHODS** 5](#_Toc164039212)

[**5. VISUALIZATION** 7](#_Toc164039213)

[**6. MODEL BUILDING** 11](#_Toc164039214)

[**7. MODEL EVALUATION** 13](#_Toc164039215)

[**8. DISCUSSION** 14](#_Toc164039216)

[**9. CONCLUSION** 15](#_Toc164039217)

[**10. FUTURE WORK** 15](#_Toc164039218)

[**11. REFERENCES** 17](#_Toc164039219)

# **1. ABSTRACT**

Market basket analysis is a widely used data mining technique to identify patterns of co-occurrence among products that customers frequently purchase together. In this project, both the Apriori and FP-Growth algorithms were employed for market basket analysis. The Apriori algorithm, a classical method, was utilized for generating frequent itemsets and association rules. Conversely, FP-Growth, known for its efficiency, was favored, especially for large datasets. Key concepts, methodologies, and practical applications of both algorithms in market basket analysis were discussed. Furthermore, insights were provided on how these algorithms can facilitate the identification of relevant product combinations and optimization of product placement and promotion strategies in the retail industry.

# **2. INTRODUCTION**

In this digital era, online shopping has become exceedingly popular, with a vast majority of individuals having been exposed to it in some form. To illustrate the significance of the concepts explored in this paper, let's envision a scenario in a traditional grocery store. Imagine a situation where the store notices a significant increase in sales of chuck steak. Recognizing this trend, the store decides to leverage market basket analysis to further capitalize on this opportunity. By analyzing customer purchase data, the store identifies a strong association between chuck steak, jam, and marinade. Armed with this insight, the store strategically lowers the prices of jam and marinade, effectively upselling these complementary items to customers who purchase chuck steak. This scenario underscores the importance of market basket analysis in understanding customer behavior and optimizing sales strategies.

Market basket analysis (MBA) serves as a fundamental data mining technique for unraveling purchase patterns in various retail domains. It encompasses a set of statistical computations aimed at identifying patterns that aid business leaders in better understanding and serving their customers. At its core, MBA seeks to uncover the most prevalent product combinations in transactions, thereby shedding light on consumer preferences and tendencies. Simply put, MBA allows store owners to dissect product associations, discern related items, and pinpoint frequently co-purchased products, thereby laying the groundwork for strategic decision-making.

Market basket analysis emerges as a cornerstone in this endeavor, offering retailers a powerful tool to decode intricate purchasing patterns and unlock hidden opportunities for growth. Within the context of our exploration, we delve deeper into the methodologies and applications of market basket analysis, shedding light on its pivotal role in driving business success.

# **3**. **DATASET**

The dataset which we used here consists of 7 attributes helpful which help to analyze customer buying patterns. The dataset contains 522065 rows. The 7 features are as follows:

BillNo: 6-digit number assigned to each transaction. Nominal.

Item name: Product name. Nominal.

Quantity: The quantities of each product per transaction. Numeric.

Date: The day and time when each transaction was generated. Numeric.

Price: Product price. Numeric.

CustomerID: 5-digit number assigned to each customer. Nominal.

Country: Name of the country where each customer resides. Nominal.

# **4.** **METHODS**

**Dataset Selection and Exploration**

Our journey commenced with the meticulous selection of a dataset that encapsulated the essence of customer transactions within a retail environment. Leveraging our expertise and research acumen, we scoured online repositories to unearth a dataset rich in attributes conducive to market basket analysis. Upon discovery, we meticulously examined the dataset's structure and content, ensuring its suitability for our analytical pursuits.

**Implementation of Market Basket Analysis Algorithms**

Central to our methodology was the implementation of two prominent market basket analysis algorithms: Apriori and FP-Growth. These algorithms, renowned for their efficacy in extracting frequent itemsets and association rules, formed the bedrock of our analytical framework. Through meticulous experimentation and fine-tuning, we harnessed the capabilities of these algorithms to unearth meaningful insights from our dataset.

**Agile Data Preprocessing and Cleaning**

Upon loading the dataset, our initial focus shifted towards meticulous data cleaning and preprocessing to ensure the integrity and reliability of our analysis. Here's an overview of the agile methodologies employed:

Null Value Handling: We commenced our data cleaning journey by scrutinizing the dataset for null values. Fortunately, our thorough inspection revealed the absence of any null entries, alleviating the need for extensive imputation strategies.

**Zero Quantity Removal**: Recognizing the importance of accurate quantity data, we proceeded to scrutinize the 'Quantity' column for zero values. Given that zero quantities hold no significance in our analysis, we swiftly removed such entries from our dataset to maintain data coherence.

**Total Price Calculation**: To facilitate comprehensive analysis, we embarked on computing the total price of items by multiplying the quantity with the respective item price. This resulted in the creation of a new column containing the total price values, thereby enhancing the granularity of our dataset.

**Outlier Detection using IQR Method**: Leveraging the robustness of the Interquartile Range (IQR) method, we undertook outlier detection to identify aberrant data points within our dataset. Calculating the upper and lower quartiles (Q3 and Q1, respectively), we determined the IQR as the difference between these quartiles.

**Flooring and Capping Technique**: Armed with insights gleaned from the IQR method, we employed the flooring and capping technique to address outliers effectively. By setting threshold values based on the 75th and 25th percentiles, we defined upper and lower bounds for outlier detection.

**Outlier Removal**: With the upper and lower bounds delineated, we proceeded to remove outliers from our dataset. Entries falling below the lower bound were replaced with the lower bound value, while those exceeding the upper bound were substituted with the upper bound value, ensuring data integrity and coherence.

**Visualization for Enhanced Understanding**: To augment our understanding of outlier distribution and efficacy of outlier removal, we harnessed the power of data visualization. Employing box plots, we visualized the distribution of total price values, providing insights into the impact of outlier removal on data coherence and analytical robustness.

# **5. VISUALIZATION**

Upon completing the data cleaning phase, we delved into visualization techniques to gain deeper insights into our dataset. Leveraging the Plotly and Seaborn libraries, we crafted interactive and informative visualizations to elucidate various aspects of customer purchasing behaviors.

**Plotly**:

Plotly stands as a robust open-source data visualization library proficient in generating interactive, web-based visualizations across multiple platforms. Below are the key features and visualizations crafted using Plotly:

1. **Country-wise Total Price of Quantity**: Fig 1 illustrates the distribution of total price values of quantities across different countries. This visualization aids in discerning variations in purchasing patterns and expenditure among diverse geographical regions.

A graph of the number of countries/regions

Description automatically generated

Observation: With a high level view the graph depicts to have a descending trend from highest revenue generated in United Kingdom to the least in Saudi Arabia.

1. **Quantity of Items Sold vs. Total Gain of Each Item**: Fig 2 juxtaposes the quantity of items sold (represented by bars) against the total gain garnered from each item. This visualization unveils intriguing insights into the correlation between sales volume and revenue generation, highlighting items with disproportionate gains relative to their sales volume.

A graph with red lines

Description automatically generated

Observation: From the given graph the bars represents Quantities of Items sold and the line represents Prices of items respectively. It can be depicted that the majority of represented items have equally low volume with respective prices except for a few items such as - PAPER CRAFT, LITTLE BIRDIE with high price, high quatity and REGENCY CAKESTAND 3 TIER with high price but low quatity.

1. **Distribution of Total Price Data**: Fig 3 portrays the distribution of total price data, providing a comprehensive view of the skewness within our dataset. Utilizing a bell graph, this visualization facilitates the assessment of data symmetry and identifies potential outliers or anomalous data points.

A blue and red graph

Description automatically generated

Observation: From the graph above, we can infer with 2 peeks in log scale the graph is symmetrical on both ends.

**Skewness**:

Skewness, a pivotal statistical measure, elucidates the asymmetry present within a distribution. Fig 3, complemented by textual insights, expounds on the calculation and interpretation of skewness within our dataset, enriching our understanding of data distribution characteristics.

# **6. MODEL BUILDING**

Our model building endeavors revolved around implementing both the Apriori and FP-Growth algorithms to extract meaningful association rules from our dataset. Below, we outline the methodologies employed and insights derived:

1. **Apriori Algorithm**:

The Apriori algorithm, a cornerstone of data mining, facilitated the extraction of frequent itemsets and association rules from our transactional data. We meticulously followed a step-by-step approach, beginning with setting minimum support thresholds and culminating in the generation of robust association rules.

A screenshot of a computer

Description automatically generated

1. **FP-Growth Algorithm**:

In addition to the Apriori algorithm, we leveraged the FP-Growth algorithm to efficiently mine frequent itemsets from our dataset. By employing a tree-based data structure, FP-Growth offered enhanced scalability and performance, making it well-suited for large transactional datasets.

A screenshot of a computer

Description automatically generated

# **7. MODEL EVALUATION**

Following algorithm implementation, rigorous evaluation ensued to assess the efficacy and relevance of the extracted association rules. By scrutinizing metrics such as lift, confidence, and support, we gauged the strength and reliability of each association rule. Noteworthy observations, including country-specific insights and the identification of strong association rules, emerged from our evaluation endeavors.

•**Lift** : Association appears to be higher in FP growth than Apriori

•**Confidence** : appears to be more favourable in Apriorithan FP growth

•**Support** : appears to be dominant in Apriorithan in FP growth

# **8. DISCUSSION**

The preprocessing steps effectively cleaned the dataset, ensuring its suitability for subsequent analysis and modeling.

Outlier detection and handling techniques were essential for maintaining data integrity and reliability.

The analysis revealed valuable insights into customer purchasing behaviors, top-selling products, and revenue generation across different countries.

Both the Apriori and FP-Growth algorithms were employed to extract association rules, uncovering meaningful relationships between items in transactions.

Visualizations aided in the interpretation of results and provided intuitive representations of key findings.

# **9. CONCLUSION**

In conclusion, our agile approach to data preprocessing, coupled with insightful visualizations and robust model building using both the Apriori and FP-Growth algorithms, unveiled valuable insights into customer purchasing behaviors. Armed with these insights, businesses can optimize inventory management, devise targeted marketing strategies, and foster enhanced customer engagement, ultimately fostering sustainable growth and profitability.

# **10. FUTURE WORK**

**Visualization and Interpretation**: Develop interactive and visually appealing dashboards to present the results of the market basket analysis, allowing stakeholders to explore the association rules interactively and gain insights more intuitively.

**Integration with Recommendation Systems**: Integrate the association rules generated from market basket analysis into recommendation systems to provide personalized product recommendations to customers based on their purchasing history and preferences. Experiment with collaborative filtering techniques and hybrid recommendation approaches to enhance the accuracy and relevance of product recommendations.

**Real-Time Analysis and Deployment**: Explore real-time market basket analysis by integrating streaming data sources and deploying the analysis pipeline in a production environment. Implement automated model updating and retraining mechanisms to adapt to changes in customer behavior and market dynamics over time.

# **11. REFERENCES**

1. Apriori Algorithm in Data Mining: Implementation With Examples. (2023, February 17). Software Testing Help. <https://www.softwaretestinghelp.com/apriori-algorithm/>
2. Ganiyu, I. S. (2023, January 18). Market Basket Analysis in Data Mining Simplified 101 - Learn | Hevo. Learn | Hevo. <https://hevodata.com/learn/market-basket-analysis-in-data-mining/>
3. Chen, J. (2019). Learn About Skewness. Investopedia. [https://www.investopedia.com/terms/s/skewness.asp](https://www.investopedia.com/terms/s/skewness.asp%20%20%20%20%20%20%20%20%20%20%20%20%20%20%20)
4. Apriori Algorithm in Data Mining: Implementation with Examples. (2024, March 10). Softwaretestinghelp.com.<https://www.softwaretestinghelp.com/apriori-algorithm/>
5. Apriori Algorithm in Data Mining: Implementation with Examples. 2024, March 7. <https://www.softwaretestinghelp.com/fp-growth-algorithm-data-mining/>
6. Market\_Basket\_Analysis\_Pyspark\_Apriori. (n.d.). Kaggle.com. Retrieved April 14, 2023, from <https://www.kaggle.com/code/gabrieleottaviani/market-basket-analysis-pyspark-apriori/input>