January 2024

Market Basket Analysis

A machine learning project

Prepared by: Mekki Meriem
Bouazza Ayat
Seddik Dounia

IASD G2

Abstract

This research project explores market basket analysis using the Apriori algorithm to uncover hidden patterns within transactional data. The study aims to optimize business strategies by identifying frequent itemsets and relevant association rules. Utilizing a retail dataset, we preprocess transactional data and apply the Apriori algorithm to extract meaningful co-occurrence patterns among products.

The methodology involves setting appropriate support and confidence thresholds, allowing the Apriori algorithm to reveal frequent itemsets. The generated association rules provide actionable insights for targeted marketing, inventory management, and overall business optimization. We investigate the impact of varying support and confidence levels, offering a nuanced understanding of algorithm sensitivity.

Results highlight associations among products, unveiling customer preferences and cross-selling opportunities. The project contributes to market basket analysis by demonstrating the Apriori algorithm's effectiveness in extracting meaningful patterns from transactional data, with practical implications for businesses seeking data-driven strategies.

This research enhances our understanding of consumer behavior, laying the groundwork for more informed decision-making and improved business outcomes in the dynamic retail and commerce landscape.

Table of contents

Introduction

Market basket analysis is a pivotal tool in the retail and commerce landscape, offering profound insights into consumer behavior. Understanding the intricacies of customer purchasing patterns holds significant implications for businesses seeking to tailor their strategies effectively. In this context, this project explores the application of the Apriori algorithm to transactional data, aiming to reveal hidden patterns that can inform strategic decision-making in the retail sector.

B**ackground:**

Market basket analysis, a fundamental technique in the field of retail analytics, involves the examination of customer transactional data to uncover patterns and associations among purchased items. In the dynamic and highly competitive realm of retail and commerce, understanding customer behavior is paramount for businesses aiming to enhance customer satisfaction, optimize inventory, and tailor marketing strategies effectively.

The importance of market basket analysis lies in its ability to unveil hidden correlations between products, enabling businesses to identify frequently co-occurring items in transactions. This not only aids in the creation of targeted marketing campaigns but also empowers retailers to optimize product placements and inventory management.However, this analytical approach is not without its challenges. As the volume and complexity of transactional data continue to grow, businesses face the dual challenge of extracting meaningful insights from this data deluge and adapting swiftly to evolving consumer preferences.

The retail sector, characterized by its diverse product offerings and constantly changing consumer trends, provides a fertile ground for the application of market basket analysis. Businesses must navigate through a multitude of choices and consumer behaviors, making it imperative to leverage analytical tools like the Apriori algorithm to uncover actionable insights.

This project delves into this intricate landscape, motivated by the need to address the complexities of modern retail. By employing the Apriori algorithm in market basket analysis, we seek to not only understand customer preferences but also provide businesses with practical strategies to navigate the intricate interplay of products in their inventory. In doing so, we aim to contribute valuable insights to the broader field of retail analytics, providing businesses with a competitive edge in today's fast-paced market.

Motivation:

The motivation behind this research lies in the recognition of the immense value derived from uncovering hidden patterns within transactional data. In the highly competitive business environment, the ability to decipher customer behaviors and preferences provides a strategic advantage, enabling businesses to adapt and thrive.

Objective of the Study:

This project's primary objective is to apply the Apriori algorithm to conduct market basket analysis. We aim to identify frequent itemsets and derive meaningful association rules, contributing to a deeper understanding of customer purchasing behavior in a retail contex**t.**

.

Scope of the Study:

The scope of this study is defined by the boundaries and limitations set for our research. We focus on a specific dataset within the retail sector, emphasizing a particular type of transactional data to ensure a targeted and insightful analysis.

Significance of the Research:

This study contributes to the existing knowledge in market basket analysis by providing practical insights into the application of the Apriori algorithm. The potential impact of our findings extends to improved decision-making processes for businesses, particularly in areas such as targeted marketing and inventory management.

Structure of the Paper:

This paper strategically unfolds with a focused exploration of market basket analysis. Commencing with a concise "Introduction," we highlight the significance of understanding customer purchasing behavior. The subsequent "Literature Review" delves into existing research, providing a theoretical framework that informs our study. This foundation sets the stage for the "Methodology," where we detail our systematic application of the Apriori algorithm to transactional data.
Following the methodology, the paper seamlessly transitions to the "Results" section, presenting key findings from the analysis. This structured approach ensures a logical progression, moving from the theoretical context established in the literature review to the practical application detailed in the methodology and results. Such a organization aims to offer a clear and coherent narrative throughout our exploration of market basket analysis.

Literature Review

An In-Depth Look at Market Basket Analysis through Machine Learning:

Definition:

Market Basket Analysis (MBA) is a data mining technique used in retail and e-commerce to identify associations between products that customers tend to buy together. It analyzes customer purchasing patterns to discover relationships and dependencies among different items in a transaction, aiming to understand what products are frequently bought together.

How Market Basket Analysis Works:

Market Basket Analysis works by examining transactional data, typically in the form of sales receipts or online shopping carts. The process involves identifying patterns or associations between items that co-occur in transactions. This analysis often employs association rules, which are logical statements that describe the relationships between items in a dataset.

Types of Market Basket Analysis:

Association Rule Mining: Focuses on discovering relationships between items in a dataset.
Sequential Pattern Mining: Examines the order in which items are purchased over time.
Affinity Analysis: Explores relationships between products and customer demographics.

Applications of Market Basket Analysis:

Inventory Management: Helps optimize stock levels by identifying complementary products.
Product Placement: Aids in strategic placement of products in stores or online platforms.
Cross-Selling and Upselling: Enables targeted recommendations to increase sales.
Promotion Planning: Assists in designing effective promotional campaigns.
Customer Segmentation: Facilitates the grouping of customers based on their purchasing behavior.

Association Rule for Market Basket Analysis:

Association rules consist of two parts: antecedent (if) and consequent (then). They are written in the form "if A, then B," where A is the antecedent item or set of items, and B is the consequent item or set of items. The rules are quantified by metrics like support, confidence, and lift.

Algorithms Used in Market Basket Analysis:

Apriori Algorithm: Identifies frequent itemsets and generates association rules.

AIS Algorithm (Association Rule Index Structure): Optimizes the process of rule generation by creating an index structure.

SETM Algorithm (Sequential Evaluation of Temporal Patterns): Focuses on sequential patterns in transactional data over time.

FP Growth (Frequent Pattern Growth): Builds a compact data structure to discover frequent itemsets efficiently.

Advantages of Market Basket Analysis:

Increased Revenue: Boosts sales through targeted promotions and cross-selling.
Improved Customer Experience: Enhances the shopping experience by offering relevant product suggestions.

Optimized Inventory: Helps reduce excess inventory and avoid stockouts.

Market Basket Analysis From the Customers' Perspective:

Personalized Recommendations: Customers receive personalized suggestions based on their preferences.
Time-saving: Efficiently find related products, saving time during shopping.
Discovery of New Products: Customers may discover complementary items they hadn't considered.

The algorithm chosen: Apriori Algorithm

Introduction:

Market Basket Analysis (MBA) plays a pivotal role in retail analytics, providing insights into customer purchasing behavior. Among the various algorithms employed for MBA, the Apriori algorithm stands out as a foundational technique. Originally proposed by Rakesh Agrawal and Ramakrishnan Srikant in 1994, Apriori has been extensively studied and applied in diverse retail settings. This literature review delves into the theoretical underpinnings, operational mechanisms, and parameter considerations associated with the Apriori algorithm.

Theoretical Underpinnings:

At its core, the Apriori algorithm operates based on the "apriori property," a fundamental principle in association rule mining. This property posits that if an itemset is frequent, then all of its subsets must also be frequent. By leveraging this property, Apriori efficiently identifies frequent itemsets from transactional data.

The algorithm follows a two-phase approach:

Candidate Generation: Starting with individual items, the algorithm generates candidate itemsets of increasing size by combining frequent itemsets from the previous step.

Support Counting and Pruning: The algorithm scans the transactional data to count the support of each candidate itemset. Subsequently, candidate itemsets below a predefined support threshold are pruned, focusing computational efforts on the most relevant patterns.

Operational Mechanisms:

Apriori's operational flow involves a breadth-first search strategy, systematically exploring the space of potential itemsets. This systematic approach ensures that no frequent itemset is overlooked. The iterative process continues until no new frequent itemsets can be generated, providing a comprehensive view of associations within the dataset.

The algorithm's efficiency stems from its ability to avoid redundant candidate generation and support counting by leveraging the monotonicity of the support measure. By adhering to the apriori property, Apriori minimizes the number of candidate itemsets to be considered, making it suitable for large-scale transactional datasets.

Parameter Considerations:

The Apriori algorithm introduces tunable parameters crucial for its performance and the patterns it uncovers:

Support Threshold: This parameter determines the minimum occurrence frequency for an itemset to be considered "frequent." A higher support threshold yields more concise and potentially more meaningful results, focusing on stronger associations. However, setting it too high may miss subtle but relevant patterns. Conversely, a lower threshold captures a broader range of patterns but may introduce noise into the analysis.

Support Threshold Calculation:

- Calculate the support of individual items and itemsets in the dataset.
- Set a minimum support threshold, often as a percentage of total transactions.
- Only itemsets with support above this threshold are considered frequent.

Confidence Threshold: In the generation of association rules, confidence represents the conditional probability of the consequent item(s) given the antecedent item(s). The confidence threshold filters rules based on their strength, ensuring that only highly confident rules are considered. Similar to support, striking the right balance is crucial to obtaining meaningful insights.

Confidence Threshold Calculation:

. Calculate the confidence of association rules.
. Set a minimum confidence threshold to filter out weaker rules.
. Confidence is computed as the support of the combined itemset divided by the support of the antecedent itemset.

Lift Threshold: Lift measures the strength of an association rule by comparing the observed support of the rule with what would be expected if the items were independent. Setting a lift threshold helps filter out spurious rules, emphasizing those with a significant impact.

Lift Threshold Calculation:

Calculate the lift for association rules.
Set a minimum lift threshold to emphasize rules with a significant impact.
Lift is computed as the ratio of the observed support to the expected support under independence

Algorithm Steps:

1- **Initialize:**
Identify individual items in the dataset.
Set minimum support and confidence thresholds.

2- **Generate Candidate Itemsets:**
Create candidate itemsets of size k by joining frequent itemsets of size k-1.
Prune candidate itemsets that do not satisfy the apriori property.

3- **Count Support:**
Scan the transactional data to count the support of each candidate itemset.

4- **Prune:**
Eliminate candidate itemsets with support below the specified threshold.

5- **Repeat:**
Repeat steps 2-4 until no new frequent itemsets can be generated.

6- **Generate Association Rules:**
For each frequent itemset, generate association rules with confidence above the specified threshold.

7- **Evaluate Lift:**
Calculate lift for each rule and filter based on the lift threshold.

Apriori in comparison with other algorithms

The Apriori algorithm, known for its generality and interpretability, efficiently identifies frequent itemsets but can be computationally intensive. FP-Growth, offering improved efficiency and reduced memory usage, is advantageous for large datasets but may be less interpretable. Eclat, similar to FP-Growth, excels in efficiency and reduced memory requirements but has limitations in applicability to binary transaction datasets. Shared strengths include scalability and flexibility, but all algorithms are sensitive to parameter settings and potential noise in data. The choice depends on specific dataset characteristics, computational resources, and interpretability needs. Apriori remains accessible, while FP-Growth and Eclat provide efficiency benefits, each with its trade-offs.

Methodology

Choosing the dataset

In our research methodology, the choice of the dataset was a deliberate process aimed at ensuring the authenticity and real-world relevance of the market basket analysis. The decision to source the dataset from Kaggle was based on its reputation as a platform hosting diverse and high-quality datasets. Our primary criterion was to obtain a dataset containing real transactions, thereby avoiding the need to generate artificial or synthetic transactions that might not accurately reflect genuine consumer behavior.

The selected dataset, found on Kaggle, aligns with our research objectives by providing a rich source of transactional data. Real transactions inherently capture the complexity and nuances of actual customer interactions, allowing for a more accurate and insightful market basket analysis. This approach not only enhances the credibility of our findings but also ensures the applicability of the results to real-world retail scenarios.

As we delve into the subsequent stages of our research, the authenticity of the chosen dataset serves as a cornerstone, laying the groundwork for meaningful and actionable insights into market basket dynamics. The transparency of our dataset selection process contributes to the reliability and validity of our study, fostering confidence in the outcomes of our market basket analysis. It is important to note that datasets containing authentic transactions, particularly of this nature, are relatively rare online due to the private nature of transactional data.

This rarity underscores the ethical considerations surrounding the handling of customer information and reinforces the value of our meticulous dataset selection process, ensuring both authenticity and privacy compliance in our market basket analysis.

Online retail dataset

A real online retail transaction data set of two years

kaggle.com

Data preprocessing

1- Loading the dataset:

In Jupyter Notebook using Python, we loaded the dataset and inspected its size, revealing (541909, 8) rows and columns.

```
df.head()
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
26	536370	22728	ALARM CLOCK BAKELIKE PINK	24	2010-12-01 08:45:00	3.75	12583.0	France
27	536370	22727	ALARM CLOCK BAKELIKE RED	24	2010-12-01 08:45:00	3.75	12583.0	France
28	536370	22726	ALARM CLOCK BAKELIKE GREEN	12	2010-12-01 08:45:00	3.75	12583.0	France
29	536370	21724	PANDA AND BUNNIES STICKER SHEET	12	2010-12-01 08:45:00	0.85	12583.0	France
30	536370	21883	STARS GIFT TAPE	24	2010-12-01 08:45:00	0.65	12583.0	France

2- Variable Identification:

here we identified the data type and category of the variables**.**

```
df.dtypes
```

```
InvoiceNo              object
StockCode              object
Description            object
Quantity                int64
InvoiceDate    datetime64[ns]
UnitPrice             float64
CustomerID            float64
Country                object
dtype: object
```

3- Univariate Analysis

At this stage, we explore variables one by one. Method to perform uni-variate analysis will depend on whether the variable type is categorical or continuous

Categorical values:

for these variables we calculated the frequency of each category.

Continuous values:

for these one's we inspected their stats.

```
df.describe()
```

	Quantity	InvoiceDate	UnitPrice	CustomerID
count	8557.000000	8557	8557.000000	8491.000000
mean	12.911067	2011-07-13 01:19:17.424331008	5.028864	12677.995996
min	-250.000000	2010-12-01 08:45:00	0.000000	12413.000000
25%	5.000000	2011-04-07 13:07:00	1.250000	12571.000000
50%	10.000000	2011-08-17 08:50:00	1.790000	12674.000000
75%	12.000000	2011-10-19 13:49:00	3.750000	12689.000000
max	912.000000	2011-12-09 12:50:00	4161.060000	14277.000000
std	21.425031	NaN	79.909126	276.742088

4- Variable transformation

In this phase, we engaged in comprehensive variable transformation to optimize the dataset for compatibility with the Apriori algorithm. This involved meticulous handling of inherent irregularities and discrepancies within the variables

During the transformation process, we:

Removed Spaces in 'Description': Stripped spaces within the 'Description' variable for enhanced consistency and interpretability.

Excluded Negative Quantity Samples: Implemented a filtering step to eliminate samples where the 'Quantity' variable exhibited negative values, ensuring a refined and coherent dataset for effective Apriori-based market basket analysis.

5- Basket formation:

In this phase, a specialized basket representation is formulated for transactions conducted . The process involves organizing the data into a matrix-like structure where each row signifies a unique invoice, and each column corresponds to a distinct product. The matrix cells denote the aggregated quantity of each product within a specific invoice. To enhance the suitability for subsequent analysis, any instances of missing values are replaced with zeros. Additionally, a custom hot encoding function is introduced to convert quantity values into binary form, simplifying the representation to indicate the presence or absence of items in each transaction. This refined dataset sets the foundation for further exploration using association rule mining techniques in market basket analysis

6- Exclusion of Postage Items:

In this step, postage items, deemed as ubiquitous and common across almost all transactions, are intentionally excluded from the basket representation. The decision to remove 'POSTAGE' items is motivated by their consistent presence, which may not contribute significantly to diverse association rule patterns. The exclusion is implemented by dropping the 'POSTAGE' column from the basket, ensuring a more focused and meaningful dataset for subsequent market basket analysis.

Implementing the algorithm

Frequent Itemset Generation:

In this phase, frequent itemsets are generated using the Apriori algorithm on the basket dataset. The apriori function is employed with a specified minimum support threshold of 0.06, indicating that itemsets with a support value equal to or exceeding 6% of total transactions are considered frequent. The resulting frequent itemsets capture patterns of co-occurring products that occur frequently together in transactions.

Association Rule Mining:

In this step, association rules are mined from the frequent itemsets using the association_rules function from the MLxtend library. The mining process employs the lift metric with a minimum threshold set at 1, indicating that only rules with a lift value greater than 1 are considered. The resulting set of association rules provides insights into the relationships and dependencies between different products in customer transactions.

These rules highlight the likelihood of one product being purchased when another is present, contributing to a deeper understanding of customer behavior and informing strategic decision-making in retail settings.

Results:

The output represents the discovered association rules based on the Apriori algorithm applied to the basket dataset. Here's a general description of the key components:

Antecedents and Consequents:

The "Antecedents" are the items that precede or are found in transactions, while the "Consequents" are the items that follow or are associated with the antecedents.

Support:

Support indicates the proportion of transactions in the dataset that contain the antecedent and consequent items. Higher support values suggest a more frequent occurrence of the rule.

Confidence:

Confidence represents the likelihood that the presence of the antecedent will lead to the presence of the consequent in a transaction.

Lift:

Lift measures how much more likely the consequent is to be purchased when the antecedent is present compared to when it is not.

Leverage and Conviction:

Leverage and Conviction are additional metrics providing insights into the dependency and significance of the association rules.

Zhang's Metric:

Zhang's metric is another measure of the strength of association between antecedents and consequents in the rules.

The association rule results obtained through the Apriori algorithm reveal compelling insights into co-occurrence patterns among different products in customer transactions. Each rule consists of antecedent items, representing those present or purchased first, and consequent items, denoting those associated or subsequently purchased. The support metric indicates the prevalence of each rule in the dataset, offering a measure of how frequently the associated products appear together. Confidence highlights the reliability of the rules, showcasing the probability that the presence of the antecedent leads to the consequent in a transaction. Lift provides a crucial measure, indicating the extent to which the presence of one item influences the likelihood of another. Leveraging additional metrics like Zhang's metric, these association rules provide a nuanced understanding of customer purchasing behavior, offering valuable strategic insights for product placement, marketing, and potential bundling strategies in the retail domain.

[25]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE GREEN)	0.102041	0.096939	0.073980	0.725000	7.478947	0.064088	3.283859	0.964734
1	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE PINK)	0.096939	0.102041	0.073980	0.763158	7.478947	0.064088	3.791383	0.959283
2	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.094388	0.096939	0.079082	0.837838	8.642959	0.069932	5.568878	0.976465
3	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.096939	0.094388	0.079082	0.815789	8.642959	0.069932	4.916181	0.979224
4	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE PINK)	0.094388	0.102041	0.073980	0.783784	7.681081	0.064348	4.153061	0.960466
5	(ALARM CLOCK BAKELIKE	(ALARM CLOCK BAKELIKE								

Deployment

In the deployment of market basket analysis insights, a web application was meticulously developed for a retail store. The backend of the application seamlessly incorporates the dataset's products and the association rules derived from the Apriori algorithm. Leveraging Django API's robust capabilities, the products and rules are efficiently stored in a database, creating a dynamic foundation for real-time analysis. The backend logic is ingeniously engineered to handle the recommendation system, orchestrating the display of associated or recommended products when a customer explores a particular item on the web platform. This integration not only enhances the customer experience by suggesting complementary products but also facilitates data-driven decision-making for the retail store.

Complementing the powerful backend, the frontend of the web app is crafted using React, a versatile JavaScript library for building user interfaces. The React framework provides a responsive and interactive interface for customers to navigate and explore products effortlessly. The seamless integration with the Django backend ensures a cohesive and efficient user experience. Through the web app's frontend, customers can intuitively discover recommended products based on market basket analysis rules, creating an engaging and personalized shopping journey. This frontend-backend synergy enhances the overall functionality of the web app, delivering a tailored and data-driven shopping experience to users.

Conclusion

In conclusion, our market basket analysis journey has unveiled valuable insights into customer purchasing behavior, offering a deeper understanding of product associations within the retail store dataset. Leveraging the Apriori algorithm, we efficiently identified frequent itemsets and association rules, shedding light on co-occurring product patterns. The subsequent deployment of these insights into a web application, orchestrated through Django API's robust backend and React's interactive frontend, exemplifies the practical applications of data-driven decision-making in the retail domain. The seamless integration of recommended products based on association rules not only enhances the customer experience but also empowers the retail store with actionable strategies for merchandising, marketing, and inventory management. As we traverse the realms of market basket analysis, this comprehensive report serves as a testament to the significance of data-driven approaches in shaping informed business strategies, fostering customer engagement, and ultimately contributing to the success of the retail enterprise.