Market Basket Analysis using FP-Growth in PySpark

Mothi Gowtham Ashok Kumar

Fall 2023 – INFO-H516: CLOUD COMPUTING FOR DATA SCIENCE

Faculty – Sunandan Chakraborty, *Ph.D*.

Luddy School of Informatics, Computing, and Engineering, Indianapolis

INTRODUCTION

Market Basket Analysis (MBA) in e-commerce, especially for grocery shopping, stands as a critical tool in understanding and leveraging customer purchasing patterns in the digital marketplace. This project focuses on employing the FP-Growth algorithm within PySpark to identify and predict product combinations that customers frequently buy together. Such insights are pivotal for refining marketing strategies, optimizing product placement, and tailoring promotional activities. MBA not only aids in inventory management and campaign customization but also enhances the shopping experience by offering personalized recommendations, like suggesting jelly when a customer buys peanut butter. With the surge in e-commerce data, the necessity for robust algorithms and infrastructure capable of handling large datasets becomes paramount. Utilizing Apache Spark MLlib FP-growth, this initiative aims to efficiently process extensive data for market basket analysis, enabling businesses to better understand and respond to their customer base, ultimately leading to more personalized and efficient shopping experiences. This approach marks a significant stride in both technical and strategic realms, reshaping how businesses interact with and comprehend their customers' needs and preferences.

PROBLEM DESCRIPTION

The challenge addressed by this project is rooted in the complexity and scale of the dataset, which comprises over 3 million grocery orders from more than 200,000 users. This dataset [1], sourced from Kaggle, presents a rich tapestry of consumer behaviors, preferences, and purchasing patterns. It includes detailed information on each order, such as the sequence of products purchased and order history. The primary obstacle is the extraction of actionable insights from this vast, multi-dimensional dataset. The project aims to parse through this extensive data to uncover patterns of product co-purchases, translating these findings into practical recommendations for grocery shopping. The successful analysis of such a dataset can provide a blueprint for understanding customer preferences on a granular level, which is crucial for businesses looking to tailor their offerings to the evolving needs of their consumer.

DESCRIPTION OF DATA

The dataset is taken from "Instacart Online Grocery Shopping Dataset 2017" [1] - Data Source. The dataset is a relational set of files describing customers' orders over time. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 users. For each user, the data provides between 4 and 100 of their orders, with the sequence of products purchased in each order.

METHODOLOGY

The methodology of this project is a comprehensive approach that encompasses several key phases: data ingestion, pre-processing, Exploratory Data Analysis (EDA), data manipulation, and machine learning modeling. Initially, the data ingestion phase involves the acquisition and integration of the extensive Kaggle dataset into a workable format. Following this, the pre-processing stage aims to clean and structure the data, ensuring its suitability for analysis. EDA is then conducted to gain preliminary insights and identify

potential patterns within the data. The crucial phase of data manipulation involves organizing the dataset into 'shopping baskets' conducive to the FP Growth model. The model is then applied to mine frequent item sets and establish association rules. This process is critical for uncovering the hidden relationships between different products and understanding the purchasing behavior of customers. Finally, the project moves towards making predictions based on these associations and evaluating the results to gauge the effectiveness of the model and refine it further.

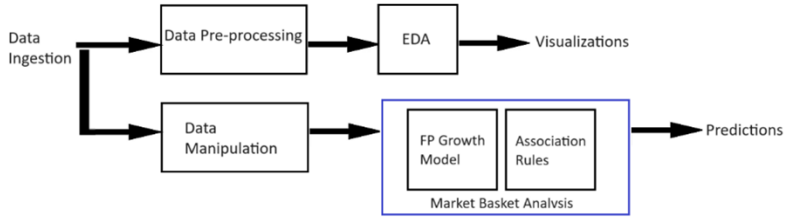
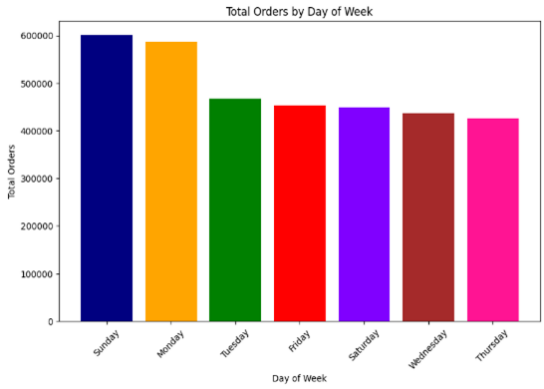


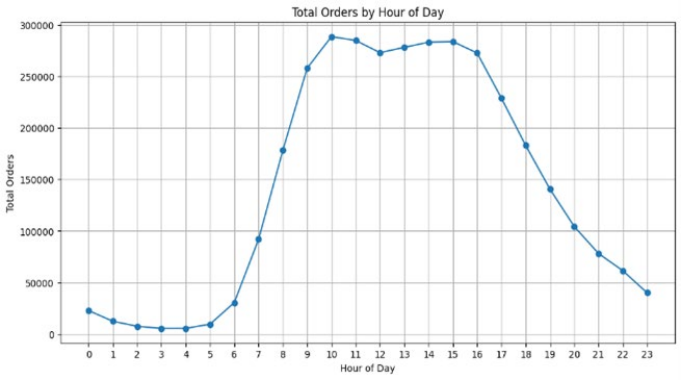
Fig 4: Strategic Plan of Action

DATA INGESTION

We have uploaded the dataset files to the personal account's Google Drive. We have mounted the drive into the Google Collaboratory environment where we have our codebase. We read the dataset files using PySpark [4] and create DataFrames using spark.read.csv.

EXPLORATORY DATA ANALYSIS (EDA)





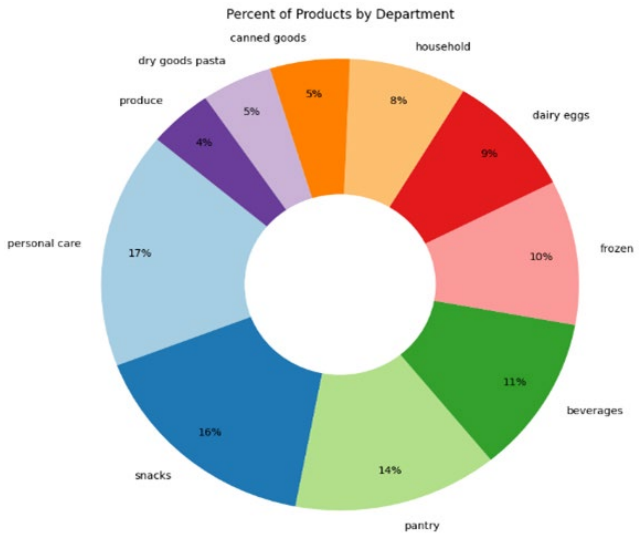


Fig 1: Bar Chart for "Total Orders by day of Week", Line Graph for "Total Orders by Hour of Day" & Donut Chart for "Percentage of Products by Department"

The bar chart indicates the highest number of orders occurs on Sunday and Monday, with the least on Thursday. The line graph depicts total orders by hour of the day, peaking during the midday hours from 10 AM to 4 PM, and shows fewest orders during the early morning hours. Finally, the donut chart illustrates the percentage of products by department, with the personal care department having the highest percentage of products, followed by snacks and pantry, indicating a diverse product range with a focus on non-food items.

FP-GROWTH MODEL

The Frequent Pattern Growth (FP-Growth) un-supervised learning algorithm [2][3] represents a significant advancement in the field of data mining, particularly for the task of identifying frequent item sets in large datasets. Its key strength lies in its efficient handling of database transactions through a compact tree structure known as the FP-Tree. Unlike the Apriori algorithm, which relies heavily on multiple database scans and candidate generation, FP Growth adopts a divide-and-conquer strategy, significantly reducing computational overheads. This algorithm operates in a two-pass process: the first pass identifies the frequencies of items, and the second pass constructs the FP-Tree. The efficiency of FP Growth not only lies in its speed but also in its reduced memory requirements, making it particularly well-suited for large datasets like the one used in this project. The application of FP Growth in this context demonstrates its practical utility in extracting meaningful patterns from complex, large-scale data, providing valuable insights into customer purchasing behaviors.

Fig 2: Top five most frequent item-sets

ASSOCIATION RULES

Association rules [5] are a fundamental component, particularly in the context of retail basket data. These rules aim to uncover patterns within large sets of transaction data, revealing relationships between different items purchased together. At their core, association rules are about finding correlations between items in a dataset, often expressed in the form of "antecedent-consequent" (if-then) statements. For instance, an association rule might state that if a customer buys peanut butter (the antecedent), there is a certain probability or confidence that they will also buy jelly (the consequent).

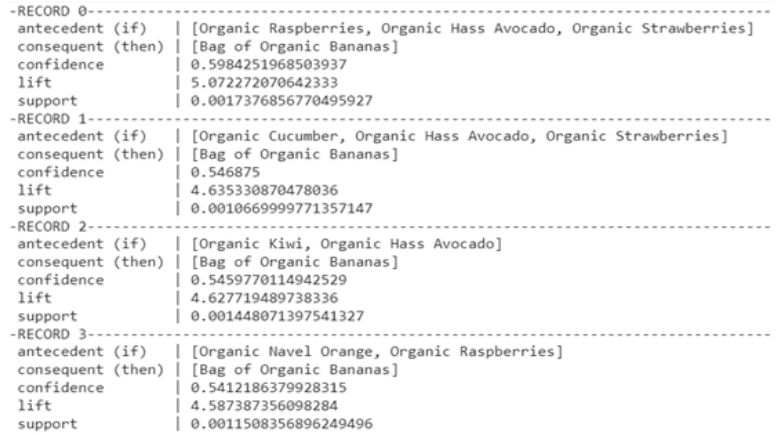


Fig 3: Top four most frequent item-sets

As can be seen in the above table, there is relatively strong confidence that if a shopper has organic raspberries, organic avocados, and organic strawberries in their basket, then it may make sense to recommend organic bananas as well. Interestingly, the top 10 (based on descending confidence) association rules - i.e. purchase recommendations - are associated with organic bananas or bananas.

PREDICTIONS

Basket-1 Items: [Organic Hass Avocado, Organic Carrot Bunch, Organic Strawberries]

Top 25 Recommended Items: [Organic Large Green Asparagus, Organic Baby Rainbow Carrots, Organic Basil, Organic Large Brown Grade AA Cage Free Eggs, 100% Raw Coconut Water, Green Beans, Red Raspberries, Organic Banana, Organic Large Extra Fancy Fuji Apple, Jalapeno Peppers, Organic Whole String Cheese, Blackberries, Limes, Organic Peeled Whole Baby Carrots, Raspberries, Hass Avocado, Bartlett Pears, Organic Broccoli Florets, Uncured Genoa Salami, Spring Water, Michigan Organic Kale, Yellow Onions, Baked Aged White Cheddar Rice and Corn Puffs, Bag of Organic Bananas, Banana]

Basket-2 Items: [Organic Fuji Apple, Organic Hass Avocado, Organic D 'Anjou Pears, Organic Large Brown Grade AA Cage Free Eggs]

Top 25 Recommended Items: [Organic Large Green Asparagus, Organic Basil, 100% Raw Coconut Water, Red Raspberries, Organic Banana, Organic Large Extra Fancy Fuji Apple, Jalapeno Peppers, Organic Whole String Cheese, Limes, Organic Peeled Whole Baby Carrots, Raspberries, Organic Broccoli Florets, Uncured Genoa Salami, Spring Water, Michigan Organic Kale, Yellow Onions, Organic Yellow Onion, Organic Red Radish, Bunch, Cucumber Kirby, Lime Sparkling Water, Unsweetened Almond milk, Organic Garnet Sweet Potato (Yam), Honeycrisp Apple, Organic Lacinato (Dinosaur) Kale, Seedless Red Grapes]

EVALUATION

Evaluating the performance of the FP Growth algorithm involves several key metrics, each providing unique insights into the effectiveness of the association rules generated. Confidence assesses the reliability of the rules generated, indicating the likelihood of purchasing a consequent item when an antecedent item is bought. Support measures the frequency with which item sets appear in the dataset, offering a basic yet crucial understanding of how common certain combinations of items are. Lift is another critical metric, adjusting for the base popularity of items and assessing the strength of association between item sets. In the context of this project, these metrics are instrumental in determining the quality and relevance of the product recommendations generated by the algorithm.

Fig 4: Evaluation metrics for Basket-1 & Basket-2

CONCLUSION

The project showcases the implementation of the FP Growth algorithm in market basket analysis within the PySpark framework, highlighting the utility of data mining in the e-commerce sector. It draws valuable insights that could enhance the customer experience and inform business strategy and marketing initiatives. Despite some association between recommended items and the initial basket, as indicated by lower confidence and support values, the more favorable lift values suggest a positive relationship. For practical business applications, aiming for higher confidence and support values would likely yield more solid and actionable recommendations. This exploration sets a foundational methodology for future endeavors, suggesting that subsequent research could focus on refining the algorithm's precision,

extending its applicability to various datasets and sectors, and combining it with other analytical tools for improved data-driven strategies in retail.

REFERENCES

[1] Instacart Online Grocery Shopping Dataset 2017:

https://www.kaggle.com/competitions/instacart-market-basket-analysis/data

[2] FP-Growth Advantages: https://www.scaler.com/topics/data-mining-tutorial/fp-growth-in-data-mining/

[3] FP-Growth: https://medium.com/@shruti.dhumne/fp-growth-d0bf62384292#:~:text=FP%2DGrowth%20(Frequent%20Pattern%20Growth,the%20need%20for%20multiple%20scans.

[4] MBA PySpark: https://towardsdatascience.com/market-basket-analysis-using-pysparks-fpgrowth-55c37ebd95c0

[5] Association Rules: https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce

[6]: Colab Project Codebase: https://colab.research.google.com/drive/1KM2UmeC3pGFUM5zt2HqxtMiJyuwxcpt3?usp=sharing

APPENDIX

Contributions from each member

Tasks	**Team Members**
Dataset selection	Vishal, Shashikant, Mothi
EDA	Vishal, Shashikant, Mothi
Modeling	Vishal, Shashikant, Mothi