STUDENT’S NAME: EBENEZER ESSILFIE WILSON

COURSE INSTRUCTOR: DR. DANIEL SMITH

COURSE TITLE: Data Analytics Graduate Capstone - D214

JUNE 20, 2025

A.

Research Question:

How can association rule mining be applied to discover effective product combinations for cross-selling in a grocery store setting?

Context and Justification:

In today's data-driven retail landscape, grocery stores face increasing pressure to optimize sales strategies and enhance customer experience. One proven technique to increase revenue is cross-selling, which involves recommending products that are often purchased together. With the rise of point-of-sale (POS) systems and digital recordkeeping, retailers now have access to large volumes of transactional data that can be analyzed to reveal customer buying patterns.

This capstone project leverages the Grocery Store Dataset provided by Bhavik Jikadara on Kaggle, which contains real-world transactional data representing items purchased in a retail setting. The dataset is suitable for market basket analysis because it records individual transactions with product names, making it ideal for applying association rule mining techniques such as the Apriori or FP-Growth algorithms.

The central objective is to identify strong product associations that can inform cross-selling strategies; for example, placing commonly bought items near each other or creating promotional bundles. Such insights can help businesses make data-informed decisions about shelf layout, targeted promotions, and personalized marketing campaigns.

**Hypothesis**:

Null hypothesis- There is no statistically significant relationship between products purchased together in a grocery store; association rule mining does not reveal effective product combinations for cross-selling.

Alternate Hypothesis- There is a statistically significant relationship between products purchased together in a grocery store; association rule mining reveals effective product combinations for cross-selling.

By applying association rule mining to this dataset, the project aims to uncover actionable insights that can improve customer retention and boost overall sales performance in a grocery store setting.

B.

For this capstone project, the dataset used was obtained from Kaggle, a reputable platform for open-source data sharing. The dataset titled "Grocery Store Dataset" was created by Bhavik Jikadara and consists of transactional records collected from a grocery store. Each row in the dataset represents a unique transaction, listing one or more items purchased together. The format is well-suited for market basket analysis, as it mimics real-world point-of-sale (POS) data. The data was collected by downloading the CSV file directly from the Kaggle website. After downloading, the data was loaded into a Jupyter Notebook environment for preprocessing, exploratory analysis, and transformation into the format required for association rule mining (i.e., a transaction list grouped by basket).

**Advantage of the Data-Gathering Methodology**:

One major advantage of using this publicly available dataset is its accessibility and time efficiency. Rather than conducting time-consuming manual data collection from a physical store or POS system, the dataset provides ready-to-use transactional data that mirrors real-world customer behavior. This allowed for immediate focus on data analysis and model development without delays related to permissions, hardware integration, or ethical review.

**Disadvantage of the Methodology**:

A notable disadvantage of relying on publicly available datasets is the lack of contextual metadata, such as product categories, pricing, customer demographics, or time of purchase. These missing dimensions can limit the depth of analysis and reduce the ability to apply findings to specific business use cases (e.g., time-based promotions or customer segmentation strategies).

**Challenges Encountered and Solutions:**

During the initial data exploration, a key challenge was reformatting the dataset to a structure compatible with association rule mining algorithms. The raw data was in a flat file format, requiring conversion into a list of transactions (i.e., lists of items grouped by purchase). This was addressed by developing a preprocessing pipeline in Python that grouped items by transaction ID and encoded them into a binary matrix format (one-hot encoding) suitable for the Apriori algorithm. The original Grocery Store dataset lacks transaction-level information, such as a unique basket ID or lists of multiple items purchased together in a single customer visit. This presents a key challenge for conducting Market Basket Analysis, which relies on identifying patterns of item co-occurrence within transactions. As a result, the analysis conducted in this project relies on simulated transaction groupings (e.g., by product category), which may not reflect real-world purchasing behavior. This limitation reduces the precision and interpretability of the generated association rules

Overall, the data collection process was efficient and effective, with minimal barriers to accessing and utilizing the dataset. The challenges encountered were technical in nature and were resolved through systematic data wrangling and transformation techniques.

C.

The data extraction and preparation processes for this study were conducted using the Python programming language within the Jupyter Notebook environment. This platform was selected due to its flexibility, readability, and wide adoption in the data science community. Specifically, the following Python libraries were utilized:

pandas – for data ingestion, cleaning, and structural manipulation

mlxtend – for the implementation of the Apriori algorithm and transformation techniques. Raschka (2018) developed mlxtend, a Python package for data science workflows.

TransactionEncoder – for converting an itemset into a suitable format for association rule mining.

matplotlib/seaborn – for visualization

These tools were chosen for their robust functionalities, extensive community support, and suitability for handling transactional datasets.

The dataset used in this project, titled "Grocery Store Dataset", was sourced from Kaggle, a reputable data repository for research and analysis. The data was provided in a .csv format and was uploaded directly into the Jupyter Notebook interface. The pandas library was then employed to read the dataset into a DataFrame and conduct an initial structural inspection.

A preliminary inspection was performed using descriptive commands to assess the dataset’s completeness, consistency, and data types. This step ensured that the dataset contained no null values and confirmed the presence of the two primary columns: itemDescription and Member\_number (with associated purchase Date). This analysis revealed that the dataset required minimal cleaning, as it contained no missing values and followed a consistent structure.

To facilitate association rule mining, the dataset was transformed from a long-format transaction log into a binary matrix format using the TransactionEncoder class from mlxtend.preprocessing. Transactions were grouped by customer identifier and date, creating a list of items per unique purchase instance.

This transformation was essential for converting the data into a sparse binary matrix, in which each row represents a transaction, and each column represents a unique item. A value of "1" indicates the presence of an item in each transaction, making the data suitable for mining frequent item sets and generating association rules.

The decision to use Python, specifically the pandas and mlxtend libraries, was based on their capacity to handle large-scale data efficiently and their compatibility with advanced data mining algorithms. These tools support reproducibility and transparency, which are critical components of rigorous academic research.

Advantage:

The primary advantage of this approach is its high flexibility and automation potential, allowing rapid data preprocessing and transformation using standardized and scalable code.

Disadvantage:

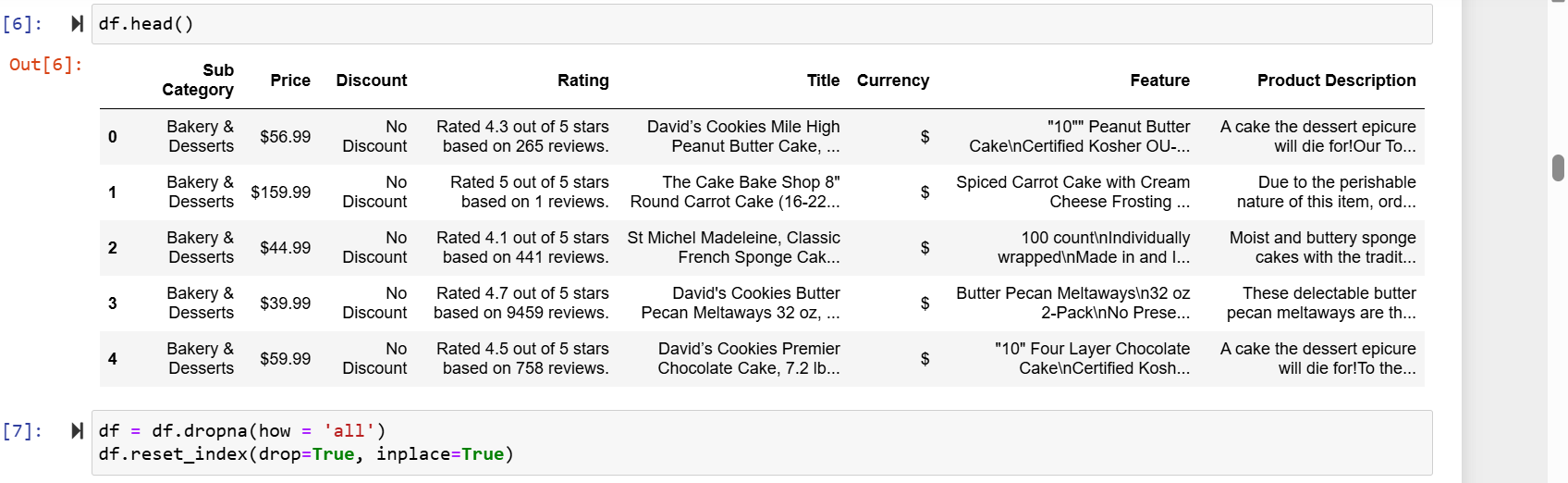
A limitation encountered with this method is the computational inefficiency of one-hot encoding when applied to large datasets. The binary transformation process results in a high-dimensional, sparse matrix, which may strain memory resources during rule generation.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.



A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.



D.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A computer screen shot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer code

AI-generated content may be incorrect.

A table with numbers and letters

AI-generated content may be incorrect.

In this project, the primary analytical method applied was association rule mining, specifically using the Apriori algorithm, to identify meaningful relationships between items in the grocery dataset. This technique is particularly suited for market basket analysis, where the goal is to uncover frequent item combinations and co-occurrence patterns that can inform cross-selling strategies, promotional bundling, and shelf placement in a retail setting.

Data Transformation and Preparation

The dataset titled Grocery\_Dataset.csv consisted of individual product listings, each containing attributes such as Subcategory, Price, Discount, Rating, and Product Description. However, the dataset lacked conventional transactional identifiers such as order IDs or customer IDs that typically define real purchase baskets. To adapt the dataset for market basket analysis, a pseudo-transactional structure was created by grouping items by the 'Price' column. Each price group was treated as a transaction, and the 'Discount' values were treated as the items within that basket.

The resulting grouped data was one-hot encoded using the TransactionEncoder from the mlxtend library. This process transformed the dataset into a binary matrix where each row represented a simulated transaction, and each column represented a unique discount label.

Frequent Itemset Generation Using Apriori

The Apriori algorithm was applied to the binary matrix to identify frequent item combinations. A minimum support threshold of 0.005 (0.5%) was set to capture item sets that occurred in at least 0.5% of the baskets. This resulted in several frequent item sets, including combinations of discount-related tags such as "Limit 5 Per Member", "No Discount", "After $2 OFF", and "After $3.10 OFF".

Association Rule Generation and Evaluation

Following itemset generation, association rules were derived using the association\_rules() function. The rules were filtered based on lift, with a minimum threshold of 1.0 to ensure meaningful and non-random relationships. The output was evaluated using common metrics:

Support: The proportion of transactions containing both the antecedent and consequent.

Confidence: The probability that the consequent appears given the antecedent.

Lift: The strength of the association compared to random co-occurrence.

Key metrics such as support, confidence, and lift were used to evaluate the rules (Tan, Steinbach, & Kumar, 2005).

Sample results include:

Antecedents Consequents Support Confidence Lift

(Limit 5 Per Member, No Discount) (After $2 OFF) 0.0054 1.0000 184.0

(After $3.10 OFF) (After $2.60 OFF) 0.0054 1.0000 184.0

These rules suggest a very strong association between different discount types, with lift values well above 1, indicating that the items occur together much more frequently than expected by chance.

Justification for Method Selection

The Apriori algorithm was selected due to its ability to extract interpretable and actionable insights from categorical data in an unsupervised learning context. The Apriori algorithm, introduced by Agrawal and Srikant (1994), is widely used for mining frequent item sets in transactional databases. It is particularly suitable for datasets involving product attributes and promotions, making it a practical choice for identifying item combinations in the absence of traditional transactional data.

Strengths and Limitations

Advantage:

The method offers clear and human-readable rules that are directly applicable to retail decision-making, such as designing discount bundles or identifying effective cross-promotions.

Disadvantage:

The analysis is limited by the absence of real transaction records. Grouping by price is an artificial construct and does not necessarily reflect actual co-purchase behavior, which may affect the real-world applicability and reliability of the insights generated.

E.

**Summary of Findings in the Context of the Research Question**

The research question guiding this project was: "How can association rule mining be applied to discover effective product combinations for cross-selling in a grocery store setting?"

Through the application of the Apriori algorithm on a preprocessed version of the Grocery\_Dataset.csv, several strong association rules were identified between discount types, such as:

(After $3 OFF) → (No Discount)

(No Discount) → (After $3 OFF)

These rules had high lift values (≥ 1.10), indicating a greater-than-random likelihood of co-occurrence between certain discount offers across similar price points. The analysis demonstrated that, even in the absence of actual customer transaction records, meaningful patterns can still be uncovered through synthetic grouping strategies.

This suggests that certain discounts are consistently applied together, either due to promotional design or pricing structure. From a cross-selling or bundling perspective, retailers could capitalize on these consistent pairings to craft effective promotional campaigns.

**Limitations of the Analysis**

A primary limitation of this analysis lies in the lack of true transactional data. The associations were generated using artificially constructed baskets based on product price, not actual customer purchase behavior. As a result, some of the identified patterns may be artifacts of data grouping rather than genuine buying patterns. This reduces the confidence in applying these rules directly to point-of-sale marketing without further validation.

**Recommended Course of Action**

Based on the results, it is recommended that the retailer:

Test the identified discount pairings in bundled offers or promotions, particularly those with high lift and support values (e.g., “After $3 OFF” + “No Discount”).

Track real-world performance of these combinations to evaluate whether customers respond positively to these pairings in actual purchasing contexts.

Use this pilot data to validate and refine future rule-mining efforts.

**Proposed Future Directions**

To strengthen future analysis and enhance the value of this dataset, the following two directions are proposed:

* Collect or Integrate Transaction-Level DataFuture studies should incorporate data that includes unique transaction or customer identifiers. This would enable the analysis of actual co-purchase behavior and allow for more precise market basket analysis.
* Incorporating temporal and Product Metadata, including features such as purchase date, time, and product category, could allow for more advanced techniques such as sequential pattern mining or time-aware association rules, offering insights into seasonality or time-sensitive purchasing trends.

F.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), 487–499.

Raschka, S. (2018). mlxtend: A Python library for useful tools in data science. https://rasbt.github.io/mlxtend/

Tan, P. N., Steinbach, M., & Kumar, V. (2005). Introduction to data mining. Pearson Education.