DATA MINING AND KNOWLEDGE DISCOVERY PROJECT

Anna Abrahamyan

ABSTRACT

In the highly competitive retail industry, understanding customer behavior and purchasing patterns is essential for improving business strategies and increasing revenue. This paper aims to leverage data mining techniques, specifically market basket analysis to analyze transaction data from a bakery and uncover valuable insights. By identifying frequent itemsets and association rules using algorithms such as Apriori, I aim to reveal which products are commonly purchased together and the strength of their relationships. This information can be used to optimize product placement, plan menus, develop new products, and segment customers effectively. Through this project, bakery owners and managers can make data-driven decisions to enhance customer satisfaction and drive business growth.

Keywords Data Mining · Association Rules · Sequence Analysis · Transaction Data · Data Analysis

1 Introduction

The retail industry is characterized by intense competition and evolving consumer preferences. To stay ahead in this competitive landscape, bakery owners and managers need to constantly innovate and adapt their strategies to meet customer demands. One key aspect of this is understanding customer behavior and purchasing patterns.

In this project, I'm appling market basket analysis to transaction data from a bakery using data mining techniques, implementing sequence analysis and linear regression. Market basket analysis finds out customers' purchasing patterns by discovering important associations among the products which they place in their shopping baskets. It not only assists in decision making process but also increases sales in many business organizations. Apriori and FP Growth are the most common algorithms for mining frequent itemsets.[**?**]

Market basket analysis(MBA) aims to understand customer behavior and preferences by identifying which products are often purchased together. There are three major types of MBAs - Association Rule Mining, Sequence Analysis and Cluster Analysis. Association Rule Mining identifies frequent item sets and generating association rules that express the likelihood of one item being purchased with the purchase of another item. Sequence Analysis identifies frequent item sequences and generates sequential association rules describing the likelihood of one item sequence being followed by another. Cluster Analysis groups similar items or transactions into clusters or segments based on their attributes to identify customer segments with similar purchasing behaviors.

In the paper, I will explore association rule mining, sequence analysis and linear regression, using algorithms such as Apriori, to discover frequent itemsets. These rules will provide valuable insights into which bakery products are commonly bought together and can be used to inform product placement, menu planning, product development, optimized product offerings and customer segmentation strategies. For example, you may find that customers who buy bread are also likely to buy butter, or customers who buy coffee are likely to buy pastries. This information can be used for various purposes, such as product recommendations, store layout optimization, and targeted marketing campaigns. By leveraging data-driven insights, bakery owners and managers can make informed decisions to enhance the customer experience, increase sales, and ultimately achieve business success in a competitive market environment. Analysis of association rules can help optimize the layout of the bakery by placing related items closer together. For instance, if coffee and pastries are often purchased together, placing them near each other can encourage customers to buy both items, increasing sales. It can help the managers to improve menu planning and product development. By identifying popular item combinations, the bakery can introduce new products that complement existing offerings or create meal deals that include popular combinations. By identifying groups of customers who exhibit similar purchasing behavior, the bakery can tailor marketing campaigns and promotions to specific customer segments.

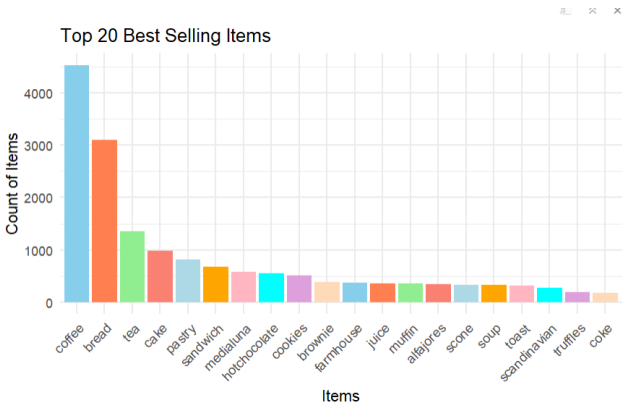

Figure 1: Top 20 Items

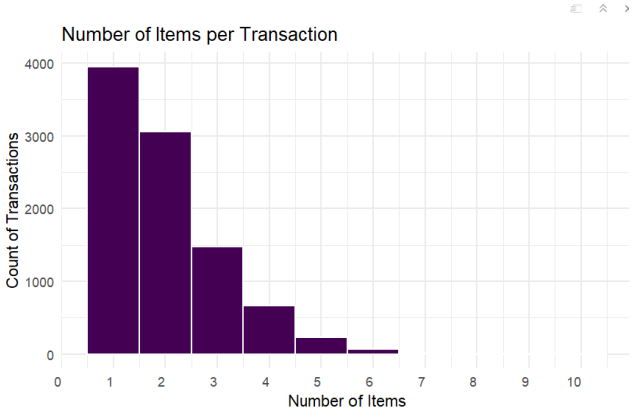

Figure 2: Number of Items per Transaction

2 Dataset

The dataset belongs to "The Bread Basket" bakery, which is located in Edinburgh is sourced from Kaggle. This dataset contains transactional data from a bakery, providing insights into customer purchasing behavior. Each observation in the dataset represents a transaction, with information including the items purchased and the date and time of the transaction. It includes 20507 entries and 4 columns, recording over 9000 transactions of customers who ordered different items from this bakery online from 2016 October to 2017 April.

The dataset contains 20,507 observations (rows). There are 5 variables (columns) in the dataset: *Transaction*, *Item*, *date_time*, *period_day*, and *weekday_weekend*.

- The '*Transaction*' variable is numerical, representing unique transaction IDs.
- The '*Item*' variable is character type, indicating the items purchased in each transaction.
- The '*date_time*' variable is character type, indicating the date and time of each transaction.
- The '*period_day*' variable is character type, representing categorical information about the time of day (e.g., morning, afternoon, night)
- The '*weekday_weekend*' is also character variable and represents categorical whether it's a weekday or weekend.

2.1 Data Preprocessing

Prior to conducting market basket analysis, I implemented dataset preprocessing to ensure data integrity and reliability. Firstly, missing values were taken into consideration, and it was determined that the dataset didn't have any. Next, I identified duplicate entries to avoid skewing analysis outcomes. Consequently, there are no extreme values that are significantly higher or lower than the typical range of values. This suggests that the distribution of the variables does not contain any noticeable outliers, or imbalanced class.

Next, I implemented some data manipulation techniques. The '*date_time*' column was converted into the right format for easier extracting. Moreover, I extracted Date, Hour, Month and Weekday columns from it to be able to perform thorough data analysis. Then, I implemented some normalization changes to the textual features. I changed the item names to lowercase and removed any spaces. After cleaning we get 18887 rows of transactions. The total number of unique transactions was 9465. And the total number of unique items was 94.

2.2 Exploratory Data Analysis

In Figure 1, we can wee the Top 10 best selling items after the cleaned and manipulated dataset. Coffee is by far the most popular item sold, accounting for 26.7% of total number of items sold. Bread ranks second at 16,2%. Over 80% of total items have low frequency of below 1% of total quantity sold. Figure 2 shows that over 38% of transactions consist of only one item. Most of transactions (approx. 95%) have equal or less than 5 items.

Figure 3, provides a summary of the transaction counts during different periods of the day, allowing us to understand the variability and central tendency of transaction volumes. And we can clearly see, the transactions mostly happen

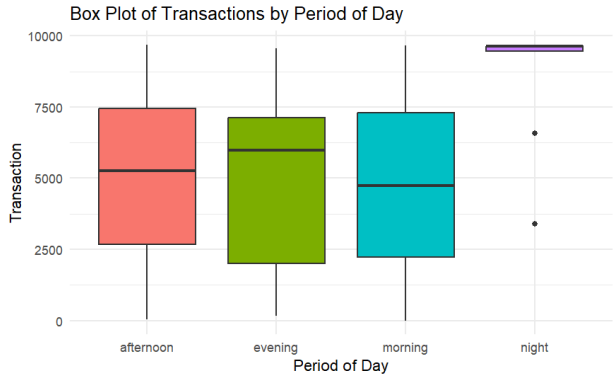

Figure 3: Transactions by period of the day

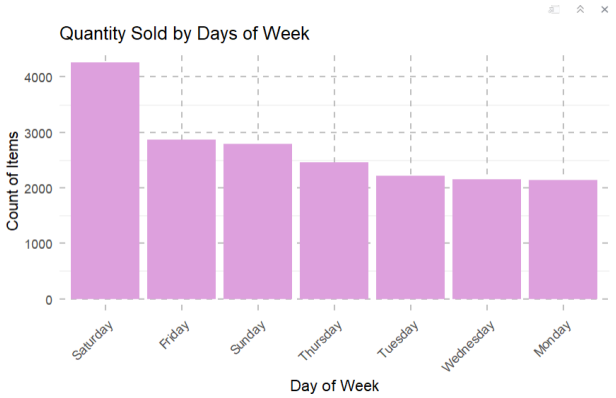

Figure 4: Number of Sales by the Days of week

from morning to evening and are comparably very low during night. From Figure 4, it's quite obvious that we sold more items on weekends, especially Saturday, with 20% more than average. Monday has lowest sales quantity with 30% less than average.

3 Models

In this project, I use Apriori Algorithm in R using alures package [**?**], TraMineR package for sequence analyis and linear regression model.

3.1 Association Rules

For Apriori algorithm I created a a transactions matrix in a sparse format with 9465 rows (transactions) and 94 columns (items), which contains observations consisting of a set of items and a transaction identifier (a market basket). For generating association rules through the Apriori algorithm, we need to define a minimum Support (supp) and Confidence (conf) for the set the rules to be analysed. Based on the most frequent items plot, it will be reasonable to apply a low support to get a relevant number of rules. The model is implemented with these parameters: minimum support of 0.001, minimum confidence of 0.30, and minimum length of items per rule of 2. And the model gave us a set of 178 rules.

Table 1: Rule Length Distribution

Length	Count	
2	3	4
42	133	3

Table 1 helps us to identify more thorough information about the rules. The majority of rules have a length of 3, followed by rules with a length of 2 and then 4. This indicates that most association rules involve three items, suggesting that customers tend to purchase sets of three items together.

	labels <chr>	support <dbl>	confidence <dbl>
41	{cake} => {coffee}	0.0547	0.527
42	{tea} => {coffee}	0.0499	0.350
40	{pastry} => {coffee}	0.0475	0.552
38	{sandwich} => {coffee}	0.0382	0.532
36	{medialuna} => {coffee}	0.0352	0.569
37	{hotchocolate} => {coffee}	0.0296	0.507

Figure 5: Association rules

As there were many association rules(178), I decided to take a look at the distribution of their support, lift and confidence to be able to do reasonable filtering (Figure 7). Support indicates how frequently the item set appears in the data set. Confidence measures the percentage of times that item B is purchased, given item A was purchased. Lift is a measure, which represents the strength of the association between two items, taking into account the frequency of both items

in the dataset. A lift value of 1 indicates that the two items are independent. So a value greater than 1 or less than 1 indicates a positive association and negative association, respectfully.

```
     lhs                rhs                support
 Length:178         Length:178         Min.   :0.0011
 Class :character   Class :character   1st Qu.:0.0013
 Mode  :character   Mode  :character   Median :0.0020
                                       Mean   :0.0050
                                       3rd Qu.:0.0039
                                       Max.   :0.0547
   confidence        coverage             lift            count
 Min.   :0.302   Min.   :0.0013   Min.   : 0.6   Min.   : 10
 1st Qu.:0.400   1st Qu.:0.0025   1st Qu.: 1.0   1st Qu.: 12
 Median :0.518   Median :0.0044   Median : 1.1   Median : 18
 Mean   :0.513   Mean   :0.0103   Mean   : 1.8   Mean   : 47
 3rd Qu.:0.600   3rd Qu.:0.0085   3rd Qu.: 1.4   3rd Qu.: 37
 Max.   :0.875   Max.   :0.1426   Max.   :43.2   Max.   :518
```

Figure 6: Summary of association rules

For this project, the minimum support threshold that an itemset must meet to be considered frequent is identified at 0.1% in total number of transactions and minimum confidence of 0.30. 0.1% might seem very low, but with this we capture more detailed and specific patterns, which will be better and more accurate for suggestions regarding business problems, so I decided to move with that. However, an important aspect to take into consideration is lift measure when assessing the associations between items, as it takes into account the support of each item to measure the strength of the association between them. For example, the item "coffee" has a high support value, but that does not mean there is a strong association between coffee and any other particular item. In this project, I will filter the rules with minimum threshold of lift as 1.1 and minimum threshold of confidence as 0.5. This will return the association rules that occur at least 10% more often than we would expect if they were random.

3.2 Sequence Analaysis

Moreover, for implementing sequence analyis I transformed the transaction data into sequences of events. These sequences represent the occurrence of items purchased during different periods of the day.

	Freq <dbl>	Percent <dbl>
6560/1-Coffee/1-afternoon/1	4	0.0195
6850/1-Coffee/1-morning/1	4	0.0195
6887/1-Coffee/1-afternoon/1	4	0.0195
104/1-Coffee/1-morning/1	3	0.0146
247/1-Coffee/1-afternoon/1	3	0.0146
346/1-Coffee/1-morning/1	3	0.0146
635/1-Coffee/1-afternoon/1	3	0.0146
2156/1-Coffee/1-afternoon/1	3	0.0146
2196/1-Coffee/1-afternoon/1	3	0.0146
2552/1-Sandwich/1-afternoon/1	3	0.0146

Figure 7: Sequence Analysis

Figure 5, represents the frequency table of the most common sequences. For example, the sequence "coffee-morning" and "coffee-afternoon" have a frequency of 4, indicating that 4 transaction contains these sequences. (As we can see from the picture there are more versions of this sequences with frequency of 3, this only indicates that the overall frequency of the sequence is higher than 4) This suggests that coffee is purchased more in the afternoon and morning. Similarly, the sequence "sandwich-afternoon" has a frequency of 3, indicating that 3 transaction contains the sequence "sandwich" purchased in the afternoon, which is logical as many people take lunch during afternoon.

3.3 Regression

Next, I applied linear regression model to analyze the relationship between the transaction amount (dependent variable) and various factors such as the type of item purchased, the time of day, the day of the week, and the month (independent variables). Linear regression will be useful for market basket analysis because it helps quantify the relationship between the transaction amount and other factors. By analyzing this relationship, businesses can gain insights into how different factors influence customer spending behavior. Linear regression can help us reveal which items have a significant

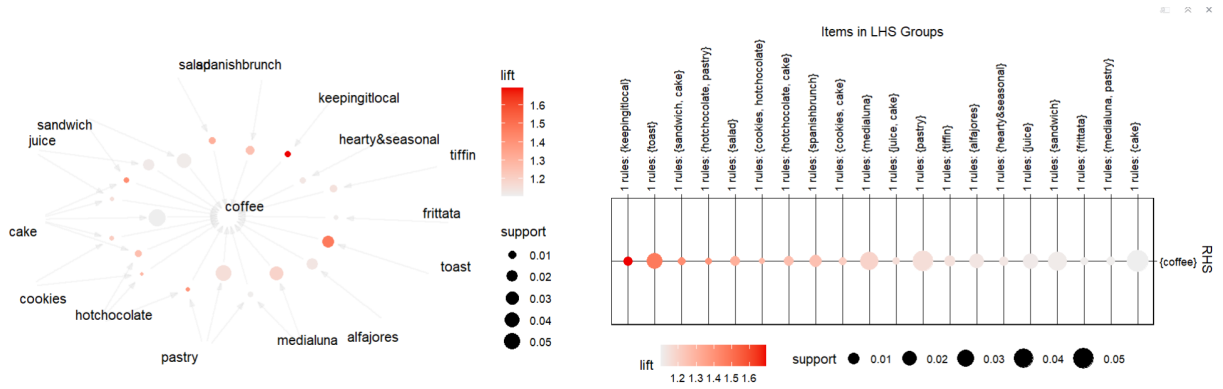

Figure 8: Top 20 Association rules

impact on transaction amounts. It can also identify trends such as whether transaction amounts tend to increase during certain times of the day, days of the week, or months of the year. This information can be valuable for optimizing inventory management, marketing strategies, and overall business operations.

4 Results

First plot in Figure 8 helps us to visualizes the association rules in a clear way, for easier interpretation of the relationships within the dataset. We can see that in the center is coffee which is in most of the association rules. In second plot, the association rules with coffee are presented more thoroughly. We can see that the bigger and darker the dots are it means that they are more common and frequent. From the plots we can notice that coffee has very high support and it appears in approximately half of total orders. Among top 20 association rules presented in Figure 9, coffee appears in all of the rules. Although, we can see from the whole rules that bread is the second most sold items, however it's not in the filtered version of top rules.

Regarding the linear regression, the coefficient estimate for the weekday_weekend variable is 37.25. This suggests that, on average, transactions increase by approximately 37.25 units on weekends compared to weekdays, holding other factors constant. The p-values associated with each coefficient estimate indicate whether the relationship between the independent variable and the dependent variable is statistically significant. For instance, the p-value for the weekday_weekend variable is less than 0.05, indicating that it has a statistically significant effect on transactions. The R-squared value (0.99) indicates the proportion of variance in the dependent variable (Transaction) explained by the independent variables included in the model. In this case, the high R-squared value suggests that the model fits the data well, explaining approximately 99% of the variance in transaction values. Next, we pay attention to residuals and find RMSE of 63.24412, which suggests that the model's predictions may have a moderate level of error.

Some important insights from the plots for analysis are as follows:

- 5.47% of transactions include both cake and coffee. When customers order coffee (support=0.05%), 50% of them will also buy cake.
- For pastry (support=0.047%), 55.2% of customers will buy coffee as well.
- 2.37% of transactions include both toast and coffee. Among transactions that include toast, 70.4% also include coffee. The lift value of 1.47 indicates that the likelihood of buying toast with coffee is 1.47 times more than random.
- The most common sequences with highest freuqencies are "coffee in the morning" and "coffee in the afternoon"
- Transactions increase by approximately 37.25 units on weekends compared to weekdays, holding other factors constant

5 Conclusion

In conclusion, this project utilized market basket analysis techniques, including association rule mining, sequence analysis, and regression, to gain insights from transaction data collected from a bakery. The association rule mining revealed the most popular items sold at the bakery, such as coffee, bread, and pastry. Analysis of association rules

highlighted potential opportunities for promoting additional purchases through targeted promotions and product bundling strategies. Suggestions for improving sales included offering promotions in the evening and on weekdays, as well as combining frequent items with less frequent ones to boost overall sales. Sequence analysis provided insights into the temporal patterns of item purchases, revealing that certain items were more commonly purchased during specific periods of the day. For example, coffee was found to be more popular in the afternoon and morning, while sandwiches were more common during the afternoon, coinciding with lunchtime. Regression analysis further investigated the factors influencing transaction values, identifying significant effects of weekends, specific items, hours of the day, and months on transaction amounts. These insights can inform pricing strategies, marketing campaigns, and operational management decisions for the bakery. Overall, the combination of these analyses provides valuable insights into customer purchasing behavior, allowing bakeries to optimize their offerings, promotions, and business strategies to enhance sales and customer satisfaction. Some interesting suggestion would be to offer Promotions/Discounts/Loyalty Cards for orders taken on weekdays or evenings to increase sales, especially on Monday. Another interesting example, would be to combine more frequent items with less frequent ones to boost overall sales: cake & juice, muffin & tea, cookies & smoothies, sandwich & coke, etc. For future work, exploring interaction effects between variables could further refine the understanding of customer behavior and inform more targeted business strategies.

References

[1] **ADITYA MITTAL:** (2020) "The Bread Basket Dataset". $doi : https : //www.kaggle.com/datasets/mittalvasu95/the-bread-basket/data$

[2] **Maliha Hossain; A H M Sarowar Sattar; Mahit Kumar Paul** (2019) "Market Basket Analysis Using Apriori and FP Growth Algorithm". $doi : https : //ieeexplore.ieee.org/abstract/document/9038197$

[3] **apriori: Mining Associations with Apriori** () "apriori: Mining Associations with Apriori ". $doi : https : //www.rdocumentation.org/packages/arules/versions/1.6-2/topics/apriori$