Group 7 - Customer Segmentation and Market Basket Analysis

Team members: Shrey Shah, Gaurav Bhandari, Venkatasubramanian Narasimman

Abstract:
In this project, we analyze an online retail dataset covering transactions. We perform customer segmentation based on purchase behavior using clustering methods. We analyze co-purchased items using association rule mining to enhance cross-selling opportunities.

Introduction and Motivation:
The project uses data mining on e-commerce customer data for customer clustering and item association discovery, aiding targeted marketing and revenue growth. This understanding of customer behavior is vital in a competitive business landscape, improving customer relationships, sales, and inventory management in retail and e-commerce.

Related Work:
According to (Marisa et al., 2019), a typical customer segmentation model is based on the variables recency, frequency, and monetary. Fashion bags were used as the product in the development of a novel methodology (Zhouzhou et al., 2021) for real-time customer segmentation. (Gomes et al. 2023) provide a comprehensive overview of methods such as manual feature selection, RFM analysis, and k-means clustering for customer representation and segmentation. (Griva et al. 2018) uses K-means clustering with an emphasis on understanding shopping behavior at the level of individual visits, as opposed to generalized customer behavior. The Apriori algorithm is examined in the paper by (Mujianto et al., 2019). (Arivazhagan et al., 2022) have studied the FP-Growth algorithm for performing association rule mining that has great scaling capability. Apriori, FP-Growth, and Eclat are three association rule algorithms compared by (M.R. Narasingha Rao et al., 2018). According to the findings, the Apriori algorithm requires more scans to generate item sets, whereas the FP-Growth algorithm requires fewer scans and Eclat can be used efficiently for small datasets.

Methodology:

Exploratory data analysis: The dataset contains features like ‘InvoiceNo,’ ‘StockCode,’ ‘Description,’ ‘Quantity,’ ‘InvoiceDate,’ ‘UnitPrice,’ ‘CustomerID,’ and ‘Country.’ It includes various data types, with some negative values in 'Quantity' and 'UnitPrice,' possibly indicating returns. Summary statistics show that 75% of ‘Quantity’ values fall between 0 and 5 units, and 75% of ‘UnitPrice’ data is under £4.13. Outliers, with 80995 units and £38970 prices, may require further investigation. Despite only 241 entries with prices exceeding £50, it is crucial to retain high-value entries for clustering analysis, as they contribute significantly to revenue in retail data.

Preprocessing: We removed 2.15% of the records with negative ‘Quantity’ or ‘UnitPrice’ values. Additionally, we dropped 135080 rows with null ‘CustomerID,’ which accounted for 24.6% of the dataset. This step is necessary for our RFM model as we try to compute the recency, frequency, and monetary value of the customers and groups based on their IDs. Null ‘Descriptions’ were found alongside null ‘CustomerID.’ As seen in Fig 1, we finally have 397884 rows in our dataset. We changed the ‘CustomerID’ data type to an integer. Inconsistencies exist in ‘Description,’ with multiple variations for the same item, maybe due to spelling mistakes. We replaced erroneous descriptions with the mode value to ensure consistency.

Adding new features: Since the dataset is limited to the sales records we implemented an RFM (Recency, Frequency, Monetary Value) model to evaluate customer value as pointed out in (Marisa et al., 2019). These variables exhibit a similar distribution with positive skew and high kurtosis, suggesting outliers as observed in Fig 2. We resolved it by applying log transformation to ensure uniform input to the algorithm.

	types	counts	distincts	nulls	missing_ratio
CustomerID	float64	406829	4373	135080	24.926694
Description	object	540455	4224	1454	0.268311
Country	object	541909	38	0	0.000000
InvoiceDate	datetime64[ns]	541909	23260	0	0.000000
InvoiceNo	object	541909	25900	0	0.000000
Quantity	int64	541909	722	0	0.000000
StockCode	object	541909	4070	0	0.000000
UnitPrice	float64	541909	1630	0	0.000000

	types	counts	distincts	nulls	missing_ratio
Amount	float64	397884	2939	0	0.0
Country	object	397884	37	0	0.0
CustomerID	int64	397884	4338	0	0.0
Description	object	397884	3647	0	0.0
Internal	object	397884	2	0	0.0
InvoiceDate	datetime64[ns]	397884	17282	0	0.0
InvoiceNo	object	397884	18532	0	0.0
Quantity	int64	397884	301	0	0.0
StockCode	object	397884	3665	0	0.0
UnitPrice	float64	397884	440	0	0.0
days_since_last_purchase	float64	397884	360	0	0.0

Fig 1. Summary statistics of the dataset before and after preprocessing with RFM features

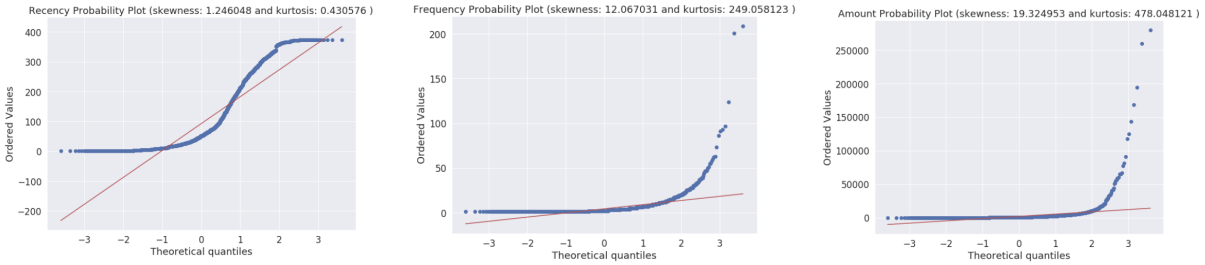

Fig 2. QQ plot for Recency, Frequency and Monetary Value (Amount) features

Customer Segmentation: We employed K-means++ clustering to segment customers using RFM features. We aim to maximize inter-cluster distance and minimize intra-cluster variance. To determine the optimal number of clusters (k), we used the elbow method, calculating the 'within-cluster sum of squares' for values of k from 2 to n, where 'k' is a hyperparameter. For assessing clustering quality, we use the Silhouette score, Dunn Index, and Davies-Bouldin Index. Higher Silhouette scores, Dunn Index, and lower Davies-Bouldin Index suggest a better quality of clusters.

Association Rule Mining: For this, we use the original dataset with just rows removed where the Description was missing. We have included the rows where the CustomerID was null (24.6% of the data), as it includes critical transactional data which might be useful for uncovering patterns. We applied the Apriori algorithm. We used support to identify frequent itemsets, confidence, and lift to assess association rule quality, and Zhang's index to measure deviations from expected item independence.

Results:

Cluster analysis:
Using the elbow method as seen in Fig 3 (a), we explored k=2 to 7 and found that k=2 is optimal. The scores for this range are as follows:
Silhouette scores: [0.43, 0.34, 0.33, 0.30, 0.31, 0.31], Dunn Index: [1.472, 1.099, 1.012, 0.888, 0.903, 0.867] and Davies-Bouldin Index: [0.887, 1.043, 1.0162, 1.047, 1.013, 0.982]

k=2 had the highest Silhouette score and Dunn Index, and the lowest Davies-Bouldin Index, indicating better cluster quality. Plots for 2 cluster groups (k=2) show distinct customer segments based on the log (amount spent) vs. log (recency) and log (amount) vs. log (frequency) are plotted as seen in Fig 3 (b). Cluster 0 comprises new, low-engagement customers with high recency and low spending, ideal for new promotions or discounts. In contrast, Cluster 1 represents loyal, high-spending customers with low recency, deserving of loyalty rewards and retention efforts.

Fig 3 (a). Elbow plot for optimal k (b) Silhouette analysis and visualization of 2 clusters

Association Rule Mining:

Association rule mining was performed on the transactions with support thresholds of 0.01 and 0.03, and a confidence level of 0.6. From these settings, we derived 461 frequent item sets, 164 rules, and 98 frequent item sets, 11 rules respectively. A data pruning approach was applied to selectively retain essential transactional data based on transaction length of 5 items and total sales percentage. This ensured a meaningful exploration of significant associations. Some of the frequent item sets observed include 'Jumbo Bag Red Retrospot', 'White Hanging Heart T-light Holder', 'Lunch Bag Red Retrospot', 'Regency Cakestand 3 Tier.'

Evaluation of the rules was done using support, confidence, lift, and Zhang's method. The rules ranked using Zhang's method are as seen in Fig 4. The ideal confidence of rules as high as 0.7 means reliable rules. A positive value of lift more than 1 indicates a strong positive association. Zhang's method ranks rules based on their interestingness and significance, measures deviations from independence in itemset occurrences. Zhang's index lies between -1 and 1. Higher the value indicates stronger negative or positive association. In our case, the rules have a high positive association as the value is closer towards 1.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	zhangs_metric
(LUNCH BAG BLACK SKULL., LUNCH BAG RED RETROS...	(PACK OF 72 RETROSPOT CAKE CASES)	0.014056	0.193684	0.011501	0.818182	4.224317	0.774157
(LUNCH BAG RED RETROSPOT, REGENCY CAKESTAND 3 ...	(JAM MAKING SET WITH JARS)	0.014239	0.167032	0.010040	0.705128	4.221522	0.774141
(REGENCY CAKESTAND 3 TIER, PACK OF 72 RETROSPO...	(JAM MAKING SET WITH JARS)	0.016429	0.167032	0.011501	0.700000	4.190820	0.774101
(HEART OF WICKER SMALL, PACK OF 72 RETROSPOT C...	(JAM MAKING SET WITH JARS)	0.015152	0.167032	0.010588	0.698795	4.183607	0.772679
(LUNCH BAG BLACK SKULL., REGENCY CAKESTAND 3 ...	(PACK OF 72 RETROSPOT CAKE CASES)	0.020993	0.193684	0.016612	0.791304	4.085547	0.771429

Fig 4. Top 5 association rules with support (0.01), confidence (0.6), lift, ranked using Zhang's metric

Plan of work:

Customer segmentation using K-Means++ had challenges like determining the optimal cluster count, sensitivity to outliers, and the assumption of uniform clusters. We plan to explore hierarchical clustering and DBSCAN for better results. For Apriori, it efficiently finds frequent itemsets and association rules but can be computationally expensive for large datasets. We will explore tree-based algorithms like FP-Growth and depth-first search-based Eclat.

Conclusion:

In summary, we employed the RFM model with K-means++ for customer clustering, gaining insights into buying behaviors. Additionally, we used the Apriori algorithm to discover itemsets and association rules, helping us understand purchasing patterns and product affinities. This combined approach informs tailored marketing and improves customer satisfaction, empowering businesses to make smart decisions.

References:

Arivazhagan, B., Pandikumar, S., Sethupandian, S. B., & Subramanian, R. S. (2022). Pattern discovery and analysis of customer buying behavior using association rules mining algorithm in e-commerce. *2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*. https://doi.org/10.1109/iceeict53079.2022.9768473

Gomes, Miguel Alves, and Tobias Meisen. "A Review on Customer Segmentation Methods for Personalized Customer Targeting in E-Commerce Use Cases - Information Systems and e-Business Management." *SpringerLink*, Springer Berlin Heidelberg, 9 June 2023, link.springer.com/article/10.1007/s10257-023-00640-4

Griva, A., Bardaki, C., Pramatari, K., & Papakiriakopoulos, D. (2018). Retail business analytics: Customer visit segmentation using Market Basket Data. *Expert Systems with Applications*, *100*, 1–16. https://doi.org/10.1016/j.eswa.2018.01.029

Marisa, F., Syed Ahmad, S. S., Mohd Yusof, Z. I., Fachrudin, F., & Akhriza Aziz, T. M. (2019a). Segmentation model of customer lifetime value in small and medium enterprise (smes) using K-means clustering and LRFM model. *International Journal of Integrated Engineering*, *11*(3). https://doi.org/10.30880/ijie.2019.11.03.018

M.R. Narasingha Rao, D., V.L Sita Ratnam, K., D.S. Prasanth, M., & Lakshmi Bhavani, P. (2018). A survey on analysis of online consumer behaviour using association rules. *International Journal of Engineering & Technology*, *7*(2.32), 206. https://doi.org/10.14419/ijet.v7i2.32.15568

Mujianto, A. H., Mashuri, C., Andriani, A., & Jayanti, F. D. (2019). Consumer Customs analysis using the Association rule and Apriori Algorithm for determining sales strategies in retail central. *E3S Web of Conferences*, *125*, 23003. https://doi.org/10.1051/e3sconf/201912523003

Yan, Zhouzhou and Zhao, Yang, "Customer Segmentation Using Real Transactional Data in E-Commerce Platform: A Case of Online Fashion Bags Shop" (2021). ICEB 2021 Proceedings (Nanjing, China). 12.
https://aisel.aisnet.org/iceb2021/12