**E-Commerce Customer Behaviors Prediction**

Choo-Kim Tan1, 2, a) and Jing-Ting Soon1, b)

1*Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450, Bukit Beruang, Malaysia.*

2*Centre for Intelligent Cloud Computing, CoE for Advanced Cloud, Multimedia University, Jalan Ayer Keroh Lama, 75450, Bukit Beruang, Malaysia*

a) Corresponding author: cktan@mmu.edu.my

b) erikasoon520@gmail.com

**Abstract.** Individuals nowadays choose to utilize and buy products from online marketplaces like Amazon, TaoBao, Shopee, and Lazada. Businesses now need to gain insight into customer buying habits to remain attractive with the explosive expansion of e-commerce. Data mining can be useful in identifying trends, and demographic data can be utilized to determine important factors that indicate purchase intent based on behaviors. One mining approach for determining the relationship between items customers frequently purchased is market basket analysis. Market Basket Analysis is a statistical technique that uses a specific kind of algorithm known as Association Rules. It uses a variety of techniques and algorithms, including the ECLAT Algorithm and Apriori. The goal of this project is to compare association rules and Machine Learning techniques to predict how customers make choices regarding their purchasing behavior for different products. The research design uses quantitative methods to study patterns of purchased items through association rules. The e-commerce platform InstaCart maintains a dataset of 3.4 million user purchase items. The data preprocessing process has been executed to achieve data analysis accuracy. The researchers utilized Python for the execution of association rule and Machine Learning on their dataset.

# INTRODUCTION

These days, individuals choose to utilize and purchase products from online marketplaces like Amazon, Shopee and Lazada. To make better judgments, people can access multiple online mediums to evaluate costs, quality, and reviews. Various sites have offered various approaches, including cashback, free delivery, and discounts. Since it is more accessible and user-friendly, e-commerce has emerged as a significant aspect of most households. Businesses now need to gain insight into customer buying habits to remain attractive with the explosive expansion of e-commerce. Personalized suggestions, cross-selling, and client retention techniques may all benefit greatly from anticipating what consumers are likely to buy next.

One of the data mining approaches for determining the relationship between items customers frequently select is market basket analysis. Market Basket Analysis is a statistical technique that uses a specific kind of algorithm known as Association Rules. It uses a variety of techniques and algorithms, including the Apriori, and ECLAT algorithms. Due to the massive amount of datasets, researchers nowadays use Machine Learning (ML) like Random Forest, Linear Regression, XGBoost, etc. to make predictions. ML methods can help the analysis become more efficient and easier. This research’s objective is to apply Market Basket Analysis and Machine Learning techniques like XGBoost, Linear Regression, and Random Forest, to identify which techniques are more efficient and reliable in predicting consumer behavior and determining future purchasing patterns. By solving and anticipating potential challenges, this research can assist the company in increasing its competitiveness in marketing, improving decision-making, and collecting essential information to improve its strategy for marketing.

Since customer retention is crucial in the digital world now, Lazada and other e-commerce platforms need to create new marketing ideas and ensure that customers are satisfied [1]. Customers’ actions play a role in their loyalty and satisfying to an e-commerce platform [1]. Understanding spending patterns can show which activities are related to customers deciding to stay or leave [1]. Researching customers' spending behaviors allows e-commerce platforms to strengthen their customer relationships and make effective marketing decisions for a better experience [1].

Data mining functions as an analytical process that handles big data to discover relationships and interrelationships and identify multiple patterns [2]. The method has found applications across economic sectors and e-commerce and bioinformatics and other fields [3]. Data mining contains standard techniques, which consist of clustering and classification and association rule mining, and predictive mining [4]. These approaches enable the definition of extensive, complicated information sets alongside the detection of regular patterns within them. Organizations use data mining techniques to create valuable business decisions through information extraction while optimizing their marketing plans for objective achievement [4].

Market Basket Analysis (MBA) operates as a data mining method for analyzing data to study customer purchase patterns. Market basket analysis finds its most widespread applications across retail stores and online businesses as well as e-commerce platforms. MBA exists to identify and understand the relationships that exist between items inside shopping carts [2]. Measuring MBA enables businesses to analyze consumption patterns through the evaluation of historical transactional data [5]. Businesses use customer transaction records to predict market trends for the future through analysis of factors including client attributes and site experience hnique that utilizes association rule algorithms to uncover patterns in customer purchasing behavior [6] [7]. According to [2], association rules as a data mining approach focused on identifying the frequency of itemset occurrences. The standard structure of an association rule is framed as an "IF-THEN" statement: the "IF" part, also known as the antecedent, represents the item(s) selected by a customer in their shopping cart, while the "THEN" part, or consequent, refers to the item(s) that are likely to be purchased alongside the antecedent items. Based on the research from [8] [9], the confidence, support, and lift were employed to determine the itemsets' frequency and validity. Support shows, as a percentage, how frequently the item has appeared in past transactions. Confidence is used to determine whether the frequency within the rule is valid. It is a metric used for calculating the likelihood of one item occurring in a transaction where another exists. Lift is employed to determine if two items are independent or interdependent. It is used to determine if the rule is accurate or not. If the value is larger than 1, it indicates that the items are dependent on each other. There are some well-known algorithms, such as Apriori, FR-Tree Algorithm and ECLAT Algorithm (Equivalence Class Transformation).

The Apriori Algorithm serves as a basic method to discover purchasing patterns in datasets [8]. The Apriori Algorithm stands out because it excels in processing extensive itemsets, along with its parallelism features and user-friendly implementation properties [10] [11]. By using the Apriori Algorithm, large-scale databases can be analyzed more quickly and accurately, however, it requires more time [9] [12]. FP-Growth is highly regarded as an effective association rule for frequent pattern management in transaction databases [2] [7]. Among all mining methodsFP-Growth works the fastest and is excellent for frequent item set analysis. The algorithm splits the grouped database into FP trees before finding out all the frequently-occurring items [7] [13]. The Equivalence Class Transformation Algorithm (ECLAT) represents a method that identifies frequent items within transactional databases [5].The ECLAT Algorithm is an easy method to detect frequently purchased items [14] [15].

As the amount of data has increased, combining Deep Learning, Machine Learning or similar methods with MBA helps researchers discover how customers behave, in terms of personalization, possible churn, segmentation and forecasting of sales [6]. According to [16], there are several useful techniques to analyze and predict customer behavior, like Random Forest, K-Nearest Neighbours (KNN), K-Means Clustering, and Extreme Gradient Boosting (XGBoost). K-means is a one of a well-known clustering techniques and easier unsupervised clustering algorithms used to partition data into K clusters or groups [17]. KNN can classify customers based on their shopping habits, making the analysis simpler and efficient to identify the behavior prediction [16]. Random Forest enhances prediction accuracy and combines multiple decision tree predictions and considers the average of all the decision tree predictions [16] [18]. XGBoost can build model step-by-step, each one correcting the mistakes of the previous one, which helps in making accurate predictions even for challenging cases [16]. Logistic Regression is also a well-known technique to predict customer churn and use to predict outcomes in 0 or 1 based on features [18] . Precision, accuracy, F1-score, and recall were employed to evaluate the model’s performance [19]. According to [19], precision was to evaluate the accuracy of a classifier and indicate how accurately the model made predictions, while accuracy quantified the proportion of correct predictions out of all predictions made by model. The recall measured the ability of a model of making accurate predictions, and F1-score was used as a performance metric that combined precision and recall into a single harmonic mean. A higher F1-score indicated better model performance by balancing both precision and recall.

According to [20], both Machine Learning and MBA techniques are useful for identifying the customer purchase behavior, but there are some limitations to applying MBA in prediction. Based on the research from [05], the MBA system mainly focuses on product co-occurrence but may not fully capture the impact of specific products attributes such as size and color on purchasing patterns. There are also lack of cross-channel integration and MBA may not adapt quickly to rapidly changing customer trends or short-term fluctuations in purchasing behavior.

# RESEARCH METHOD

This project will implement the quantitative research method which uses secondary data, InstaCart’s historical transactional data to do the analysis. The research will utilize the MBA and Association Rules to identify customer purchasing behavior patterns. The research will be designed in six different stages. This research has a sequence as shown in Figure 1.

A diagram of a machine learning process

AI-generated content may be incorrect.

**FIGURE 1.** Research methodology

The data is collected from InstaCart Online Grocery [21], which is an online grocery ordering and delivery app. This data has six tables including 3.4 million users purchase data. The key components are included with:

* + Orders: The transactions or purchase instance.
  + Products: The items available in-store for purchase.
  + Customers: Anonymized information about the users.
  + Aisles and Departments: Hierarchical categorization of products.
  + Order Sequences: Details about repeat purchases, timings, and patterns.

Before analysis, the dataset was cleaned to remove missing values, duplicates, and inconsistencies. Product and department details were merged for meaningful insights. Due to computational limits, a representative sample of 10,000 transactions was selected. The study began with Exploratory Data Analysis (EDA) using MBA techniques, including Apriori and ECLAT algorithms, to identify frequent itemsets and association rules. A minimum support threshold of 0.01 was applied, with lift and confidence used to uncover department-level co-purchase patterns.

In addition to identifying patterns, XGBoost, Logistic Regression, and Random Forest were used in supervised machine learning models. The data was made into an interaction matrix between users and products, and 80% was used for training while 20% was used for testing. In order to understand individuals’ shopping habits, models were used to predict when a product will be ordered and tested using accuracy, precision, and AUC, fostering a deeper and valuable insight into customer purchase behavior.

## RESULTS AND DISCUSSION

In this study, both the techniques of MBA and ML have been used on the Instacart dataset to study and forecast customer behavior. In the beginning, the Apriori and ECLAT Algorithms are used to discover the frequency of co-purchase patterns of the customers. In addition, machine learning models, including Random Forest, XGBoost, and Logistic Regression, are used to do predictive analysis to identify the customer’s future purchase behavior. Each method is checked against the others to see which is best suited for retail environments.

## Association Rules Results

Table 1 presents the top 5 association rules from the Apriori algorithm, highlighting frequent department-level itemset combinations. High lift values, which are above 4.4 reveal strong associations between departments such as frozen goods, pantry, bakery, and dry goods pasta, indicating common bundling behaviors. Supporting this, Figure 2 shows a co-purchase heatmap, where darker cells reflect stronger department co-occurrence, consistent with Table 1 patterns such as frozen & snacks, pantry & dry goods pasta. Figure 3 displays department-wise support derived from ECLAT algorithm. "Produce" leads with 75.14% support, followed by "dairy eggs" and the "dairy eggs, produce" combination. These trends emphasize the popularity of fresh goods. Other frequent pairings, like "produce & snacks" and "beverages & dairy eggs", align with the association rules, reinforcing the reliability of the MBA results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TABLE 1. Association rules results | | | | |
| Antecedents | **Consequents** | **Support** | **Confidence** | **Lift** |
| frozen, canned goods, dairy eggs, bakery | snacks, pantry, dry goods pasta | 0.0104 | 0.24946 | 4.497211 |
| snacks, pantry, dry goods pasta | frozen, canned goods, dairy eggs, bakery | 0.0104 | 0.187489 | 4.497211 |
| frozen, canned goods, bakery | snacks, pantry, dry goods pasta, dairy eggs | 0.0104 | 0.224719 | 4.456944 |
| snacks, pantry, dry goods pasta, dairy eggs | frozen, canned goods, bakery | 0.0104 | 0.206267 | 4.456944 |
| snacks, pantry, dry goods pasta, produce | frozen, canned goods, bakery | 0.01034 | 0.204996 | 4.429474 |

A diagram of a company

AI-generated content may be incorrect.

**FIGURE 2.** Department co-purchase heatmap

## Machine Learning Results

Table 2 summarizes the performance of three machine learning models in predicting customer reorders. XGBoost achieved the highest AUC score of 0.8053, indicating strong overall predictive capability. While Random Forest had the highest recall and F1-score, Logistic Regression offered better precision. Overall, XGBoost provides the best balance between accuracy and interpretability for forecasting purchase behavior.

Table 3 displays the top 5 association rules predicted using XGBoost, highlighting strong co-purchase patterns among product categories. Each rule shows a high lift value, which is above 5.2, indicating that the occurrence of the consequent is significantly more likely when the antecedent is present. For example, customers who purchase frozen, produce, pantry, dairy eggs, and snacks are likely to also buy dry goods pasta, breakfast items, and beverages. These results confirm that XGBoost effectively captures meaningful itemset relationships for predictive modeling.

A graph of food prices

AI-generated content may be incorrect.

**FIGURE 3.** Top 10 Most Frequently Purchased Items

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TABLE 2. Machine Learning Model Performance | | | | |
| Model | Precision | Recall | F1-Score | AUC |
| Logistic Regression | 0.5734 | 0.0764 | 0.1348 | 0.7774 |
| Random Forest | 0.5206 | 0.1684 | 0.2545 | 0.7786 |
| XGBoost | 0.6065 | 0.1425 | 0.2308 | 0.8053 |

To better visualize the strength and reliability of these predictions, Figure 4 presents a bubble chart plotting confidence against lift for the top 5 rules. The size and color of each bubble represent the lift value, while the position indicates rule confidence. This chart clearly shows that rules with both high confidence and high lift, such as Rule 3 and Rule 2, are especially valuable for prediction. The clustering of points with moderate confidence but high lift also highlights the model’s ability to detect strong but less frequent associations.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TABLE 3. XGBoost’s Prediction Results | | | | |
| Antecedents | **Consequents** | **Support** | **Confidence** | **Lift** |
| frozen, produce, pantry, dairy eggs, snacks | dry goods pasta, breakfast, beverages | 0.01 | 0.137362637 | 5.303577 |
| dry goods pasta, breakfast, beverages | frozen, produce, pantry, dairy eggs, snacks | 0.01 | 0.386100386 | 5.303577 |
| dry goods pasta, produce, breakfast, beverages | dairy eggs, frozen, snacks, pantry | 0.01 | 0.427350427 | 5.262936 |
| dairy eggs, frozen, snacks, pantry | dry goods pasta, produce, breakfast, beverages | 0.01 | 0.123152709 | 5.262936 |
| dry goods pasta, snacks, breakfast | frozen, produce, beverages, pantry, dairy eggs | 0.01 | 0.328947368 | 5.204863 |

## Comparison between MBA and ML

According to the analysis, this study explains the main differences between MBA and ML in analyzing customer buying patterns. Using association rules, MBA uncovers which products are often bought together from previous transactions. They can be understood by examples, with support, confidence, and lift being used to display what items customers commonly bought together before. However, MBA is not well-designed for making predictions. XGBoost, one of the Ml techniques, can estimate the probability that a user will reorder the item again in the future. For these reasons, ML is better for making recommendations just for users and planning future business approaches.

**A graph with a number of dots

AI-generated content may be incorrect.FIGURE 4.** XGBoost’s Prediction Results

## CONCLUSION

MBA assists in discovering frequently bought product combinations and how they are linked when processed through transactional data. With 100,000 random samples from the Instacart dataset, it was discovered that produce and dairy eggs, being key everyday items, were the items consumers commonly selected at once. As a result, businesses can improve how their products are displayed, bundled and promoted to meet what their customers are looking for. However, MBA can only offer insights into what has happened with sales so far, but it cannot predict future purchases. Hence, this study also relied on Machine Learning (ML) models, specifically on XGBoost, to estimate the probability of customers ordering the same products again. By looking at a user’s old transactions, ML algorithms were better at estimating what they could buy in the future.

Both approaches are strong in different areas. MBA is good for understanding top purchasing trends and relationships, and ML allows for creating predictive models that support immediate decisions. All of these approaches offer a comprehensive analysis of how customers behave. The use of MBA supports strategic planning, while the use of ML helps create customized suggestions and automated tools. Using both strategies lets organizations improve customer satisfaction, improve business operations, and boost their growth by making decisions with data.

# Acknowledgments

The author wants to express sincere gratitude to everyone who supported and actively contributed to this project's achievement. The valuable help provided by these supporters has played a crucial role in achieving research success.

# References

1. M. A. A. Hakim and T. Terttiaavini, “Predictive Buyer Behavior Model as Customer Retention Optimization Strategy in E-commerce,” *Journal of Intelligent System and Computation*, vol. 6, no. 1, pp. 32–38, Apr. 2024, doi: 10.52985/insyst.v6i1.379.
2. A. Sharma and H. Babbar, “Analysis of Data Mining Algorithms in Market Basket Analysis,” in *2023 International Conference on Advancement in Computation and Computer Technologies, InCACCT 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 275–280. doi: 10.1109/InCACCT57535.2023.10141816.
3. B. Arivazhagan, S. Pandikumar, S. B. Sethupandian, and R. S. Subramanian, “Pattern Discovery and Analysis of Customer Buying Behavior Using Association Rules Mining Algorithm in E-Commerce,” in *2022 1st International Conference on Electrical, Electronics, Information and Communication Technologies, ICEEICT 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICEEICT53079.2022.9768473.
4. H. Zhang, “Precision Marketing System for Cross-Border E-commerce Based on Data Mining Technology,” in *2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data, ICEIB 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 578–582. doi: 10.1109/ICEIB61477.2024.10602599.
5. D. Rajan and V. L. Helen Josephine, “Data Mining Techniques to Enhance Customer Segmentation and Targeted Marketing Strategies,” in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, Jun. 2024, pp. 1–5. doi: 10.1109/ICCCNT61001.2024.10725770.
6. M. L. Arnumukti, S. Sudianto, and U. Athiyah, “Product Layout Recommendations based on Customer Behavior and Data Mining,” in *Proceeding - COMNETSAT 2023: IEEE International Conference on Communication, Networks and Satellite*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 330–334. doi: 10.1109/COMNETSAT59769.2023.10420563.
7. V. L. Helen Josephine and D. Rajan, “Enhancing Customer Experience and Sales Performance in a Retail Store Using Association Rule Mining and Market Basket Analysis,” in *2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICCCNT56998.2023.10307411.
8. U. G. ÇİÇEKLİ and İ. KABASAKAL, “Market Basket Analysis of Basket Data with Demographics: A Case Study in E-Retailing,” *Alphanumeric Journal*, vol. 9, no. 1, pp. 1–12, Jun. 2021, doi: 10.17093/alphanumeric.752505.
9. S. S. Khedkar and S. Kumari, “Market Basket Analysis using A-Priori Algorithm and FP-Tree Algorithm,” in *Proceedings - 2021 1st IEEE International Conference on Artificial Intelligence and Machine Vision, AIMV 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/AIMV53313.2021.9670981.
10. D. Durgadevi, “An Association Rule Mining’s Algorithm for E-Commerce Technology.” [Online]. Available: https://www.researchgate.net/publication/380434315
11. N. A. H. Mohd Rosli and N. H. Ibrahim Teo, “MARKET BASKET ANALYSIS USING APRIORI ALGORITHM: GROCERY ITEMS RECOMMENDATION,” *Advanced International Journal of Business, Entrepreneurship and SMEs*, vol. 4, no. 14, pp. 01–09, Dec. 2022, doi: 10.35631/aijbes.414001.
12. C. Ignatius Hermina, “MARKET BASKET ANALYSIS FOR A SUPERMARKET.” [Online]. Available: https://www.researchgate.net/publication/365489098
13. S. O. Fageeri, M. A. Kausar, and A. Soosaimanickam, “MBA: Market Basket Analysis Using Frequent Pattern Mining Techniques,” *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, pp. 15–21, May 2023, doi: 10.17762/ijritcc.v11i5s.6591.
14. M. Gito Resmi, T. I. Hermanto, and M. Al Ghozali, “Product Recommendation System Application Using Web-Based Equivalence Class Transformation (Eclat) algorithm,” *SinkrOn*, vol. 7, no. 3, pp. 957–961, Jul. 2022, doi: 10.33395/sinkron.v7i3.11454.
15. A. Gopakumar, A. Shine, A. Ajim, and T. Anjali, “Consumer Behaviour Analysis for Customer Segmentation and Purchase Prediction,” in *10th International Conference on Advanced Computing and Communication Systems, ICACCS 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 1127–1131. doi: 10.1109/ICACCS60874.2024.10717226.
16. L. Kumari, K. Bhattacharjee, N. Sharma, S. Kumar, and A. Kumari, “Machine Learning Models in Customer Behaviour Prediction: A Comparative Analysis,” in *Proceedings of International Conference on Contemporary Computing and Informatics, IC3I 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 957–959. doi: 10.1109/IC3I61595.2024.10828637.
17. R. F. Sabahi, F. Razavi, and A. Asadi, “Developing Personalized Marketing Strategies Based on Customer Behavior Analysis Using Clustering and Machine Learning,” in *2025 11th International Conference on Web Research (ICWR)*, IEEE, Apr. 2025, pp. 391–402. doi: 10.1109/ICWR65219.2025.11006252.
18. S. Lalwani, “Implementation of Machine Learning Algorithms to Predicting Customer Churn for HRMS Software Vendors,” 2024. [Online]. Available: https://www.jisem-journal.com/
19. S. Anisah and M. Wasesa, “Improving Café Reputation: Machine Learning Analytics for Predicting Customer Engagement on Google Maps,” *Journal of Information Systems Engineering and Business Intelligence*, vol. 11, no. 1, pp. 91–102, Feb. 2025, doi: 10.20473/jisebi.11.1.91-102.
20. Z. Farheen and A. Dharani, “Prediction of Customer Purchasing Patterns for Retail Optimization Using Market Basket Techniques,” in *8th IEEE International Conference on Computational System and Information Technology for Sustainable Solutions, CSITSS 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/CSITSS64042.2024.10816740.
21. M Yasser H, “InstaCart Online Grocery Basket Analysis Dataset,” Kaggle. Accessed: Apr. 24, 2025. [Online]. Available: https://www.kaggle.com/datasets/yasserh/instacart-online-grocery-basket-analysis-dataset