Asia Pacific
Academy of Science Pte. Ltd.

Business and Management Theory and Practice 2025, 2(2), 3216.
https://doi.org/10.54517/bmtp3216

Article

Exploring consumer purchasing behavior: Business insights for precision marketing

Kai-Hsun Wang[1,2,†], Yi-Hsien Tai[1,†], Ben-Chang Shia[1,2], Mingchih Chen[1,2,*]

[1] Graduate Institute of Business Administration, College of Management, Fu Jen Catholic University, New Taipei City 242062, Taiwan
[2] Artificial Intelligence Development Center, Fu Jen Catholic University, New Taipei City 242062, Taiwan
*** Corresponding author:** Mingchih Chen, 081438@mail.fju.edu.tw
[†] These authors contributed equally to this work.

CITATION

Wang KH, Tai YH, Shia BC, Chen M. Exploring consumer purchasing behavior: Business insights for precision marketing. Business and Management Theory and Practice. 2025; 2(2): 3216. https://doi.org/10.54517/bmtp3216

ARTICLE INFO

Received: 13 January 2025
Accepted: 10 April 2025
Available online: 6 May 2025

Abstract: Understanding consumer purchasing behavior is crucial for businesses aiming to enhance customer engagement and optimize marketing strategies. In today's digital economy, traditional marketing approaches are becoming less effective due to evolving consumer behaviors, the rise of online communities, and the widespread use of ad-blocking software. To remain competitive, businesses must adopt data-driven strategies to analyze consumer preferences and tailor their marketing efforts accordingly. Machine learning provides a powerful tool for predicting consumer purchasing behavior, enabling businesses to anticipate customer needs and implement targeted marketing campaigns. Previous studies have demonstrated the effectiveness of machine learning in consumer analysis, particularly in customer segmentation and purchase prediction. However, while much research focuses on technical model optimization, relatively few studies have applied machine learning specifically for marketing prediction and strategic decision-making. This study addresses that gap by leveraging machine learning to analyze consumer purchasing behavior and generate practical insights for marketing strategies and business applications. Using a dataset of 4680 transactions, we employ Generalized Linear Models (GLM), Logistic Regression, Random Forest, and XGBoost to predict repurchase behavior within a specified timeframe. Our objective is to provide practical implications for businesses, such as improving targeted promotions, refining customer segmentation, and enhancing demand forecasting.

Keywords: customer behavior; precision marketing; machine learning

1. Introduction

Major convenience store chains have increasingly partnered with well-known foreign brands. With the widespread use of online communities, the way consumers receive information has changed, impacting the marketing models of these retail channels. Promotional activities in major convenience store chains are often disseminated through online platforms to attract customers to physical stores. According to the Taiwan Network Information Center (TWNIC) 2018 Taiwan Internet Report [1], the number of internet users in the country has reached 18.66 million, with an overall internet penetration rate of 79.2%. The survey shows that content media and social media have the highest usage rates among netizens, at 88.5% and 80.6%, respectively—both exceeding eighty percent. Among these, Facebook and Instagram are the two most widely used social media platforms, with usage rates of 98.5% and 38.8%, respectively.

With the rise of the anti-advertising movement, nearly $50 billion worth of global advertisements are being blocked. One-third of adult internet users employ ad-blocking software, and 48% of users actively avoid website ads. The effectiveness of advertising reach has been declining annually, with the average reach rate now below 5%.

In recent years, the rapid development of data science and machine learning has led to the widespread application of these advanced technologies in various commercial scenarios, including the prediction of consumer purchasing behavior. Consumer buying behavior is a key factor in business decisions, significantly affecting a company’s sales and marketing strategies [2]. As consumer preferences are diverse and ever-changing, predicting their purchasing behavior has become particularly important. In today’s business environment, characterized by intensified market competition and diversified consumer behavior, companies face increasing challenges. One major challenge is accurately predicting consumer purchasing behavior, especially in fast-moving consumer goods sectors like coffee. The motivation for this study stems from the urgent need to use advanced technology to address this challenge. Machine learning, as a powerful data analysis tool, offers a new approach to understanding and predicting consumer behavior, which is crucial for enhancing a company’s market strategies and operational efficiency. In the coffee market, consumer purchasing behavior may be influenced by various factors, including price, product quality, brand image, and consumers’ lifestyles and taste preferences [3]. Therefore, predicting consumer purchasing behavior requires consideration of these factors’ impact.

Companies should implement effective marketing strategies to reduce customer churn. Discovering customers’ purchasing behaviors and needs is an important method for enhancing competitiveness [4]. However, many companies’ marketing activities often treat all customers as a single entity, applying the same marketing tactics to different customers, which contradicts the concept of database marketing [5]. The RFM (Recency, Frequency, Monetary) model in direct marketing helps companies understand customers based on their past purchasing behaviors. A key aspect of this model is customer segmentation, which allows companies to decide whether to further engage with specific customers [6].

Traditional prediction methods usually rely on statistical models and consumer surveys, but these approaches can be inefficient in handling large volumes of complex data and may struggle to capture the nonlinearities and interactions in consumer behavior [7]. In contrast, machine learning methods, such as random forests, can process a large number of input variables and effectively handle nonlinearities and interaction effects [8].

Studies have used machine learning techniques, such as decision trees, cluster analysis, and Bayesian algorithms, to analyze customer characteristics and attributes based on historical purchase records. Furthermore, these methods help identify key factors influencing potential customer purchasing behavior by selecting models with higher promotion rates. The results indicate that the most important factors influencing customer purchasing behavior are age and the number of cars owned by the household [2].

For example, research combining cluster analysis with the RFM model explored the purchasing behavior of customers at a furniture company to develop precise marketing strategies [9]. Another study combined K-means and Random Forest methods, using the K-means algorithm for preliminary clustering to obtain labels for user data. Subsequently, Random Forest was used to select labeled user data to obtain feature importance rankings, applying these rankings as weight parameters for user characteristics [10].

In today's data-driven business environment, machine learning technology plays an increasingly crucial role in predicting consumer purchasing behavior. The study by Zuo et al. [11] is one of the pioneering works in this field, demonstrating how machine learning techniques can be used to analyze and predict the purchasing behavior of grocery store consumers. This research not only provides a deep understanding of consumer behavior patterns but also offers valuable insights for retailers to optimize inventory management and marketing strategies.

The study focused on the online shopping domain, exploring the application of deep learning technologies in predicting consumer online shopping behavior. Their research not only proves the effectiveness of machine learning technologies in analyzing large volumes of online shopping data but also highlights how these analytical results can enhance customer experience and increase sales efficiency [12].

Yao explored the role of machine learning in predicting and analyzing consumer purchasing behavior in e-commerce. This study emphasizes the importance of machine learning in understanding consumer purchasing decisions and highlights how these insights can be used to design more effective personalized marketing strategies [13]. A review study by Bangyal et al. synthesized multiple applications of machine learning in online shopping consumer behavior research. Their research demonstrates the diverse applications of machine learning technologies in predicting consumer behavior and underscores the potential of these technologies in enhancing business competitiveness and customer satisfaction [14].

This study aims to bridge that gap by applying machine learning to analyze consumer purchasing behavior and derive business insights. Using a dataset of 4680 transactions, we employ Generalized Linear Models (GLM), Logistic Regression, Random Forest, and XGBoost to predict repurchase behavior. Our goal is to provide practical implications for businesses, such as improving targeted promotions, refining customer segmentation, and enhancing demand forecasting.

This paper is structured as follows: Section 1 introduces the research problem, highlights the significance of consumer behavior analysis, and presents an overview of previous studies. Section 2 details the dataset, data preprocessing techniques and the machine learning models applied in this study. Section 3 presents the experimental results, evaluates model performance, and provides a performance of different algorithms. Section 4 discusses key findings, interprets the advantages and limitations of this study, and explores the implications for businesses. Finally, Section 5 concludes the study by summarizing its key contributions, highlighting managerial and research implications, and suggesting potential directions for future research.

2. Materials and methods

H Company's app allows users to accumulate steps through walking to earn rewards and redeem various discounts and gifts. This not only promotes health among users but also enhances their motivation to visit stores and strengthens their connection to the brand. Additionally, users receive personalized recommendations and discounts. The products have already attracted millions of users globally, generating a substantial amount of transaction data daily. This study follows a structured methodology to predict consumer purchasing behavior using machine learning models. The process begins with data collection, where a dataset comprising 4680 cases was gathered from 1 January 2020, to 31 August 2020. This dataset serves as the foundation for analyzing consumer purchasing patterns.

Following data collection, the next step involves data preprocessing to ensure data quality. This includes data cleaning, where missing values and inconsistencies are addressed, ensuring the dataset is ready for analysis. Key variables used in the study include age, gender, steps taken, and the number of previous purchases, which serve as input features for the predictive models.

The study then defines two types of predictive variables. The first is a binary classification task, where the models predict whether a consumer will make a purchase within 7 days or 14 days. The second is a regression task, which predicts the number of days until the next purchase as a continuous variable. This dual approach allows for a comprehensive understanding of consumer purchasing behavior.

To develop predictive models, four machine learning techniques are employed: Logistic Regression (LG), Generalized Linear Models (GLM), Random Forest (RF), and XGBoost. These models are trained using the dataset and tested to determine their effectiveness in making accurate predictions.

Finally, the models undergo evaluation using multiple performance metrics, including accuracy, specificity, precision, recall, and F1-score. These metrics provide a detailed assessment of each model’s strengths and weaknesses, ensuring a robust comparison of their predictive capabilities.

By following this structured workflow, the study aims to offer valuable insights into consumer purchasing behavior, enabling businesses to develop more effective marketing strategies and improve customer engagement. The Data Processing Flowchart is illustrated in **Figure 1**.

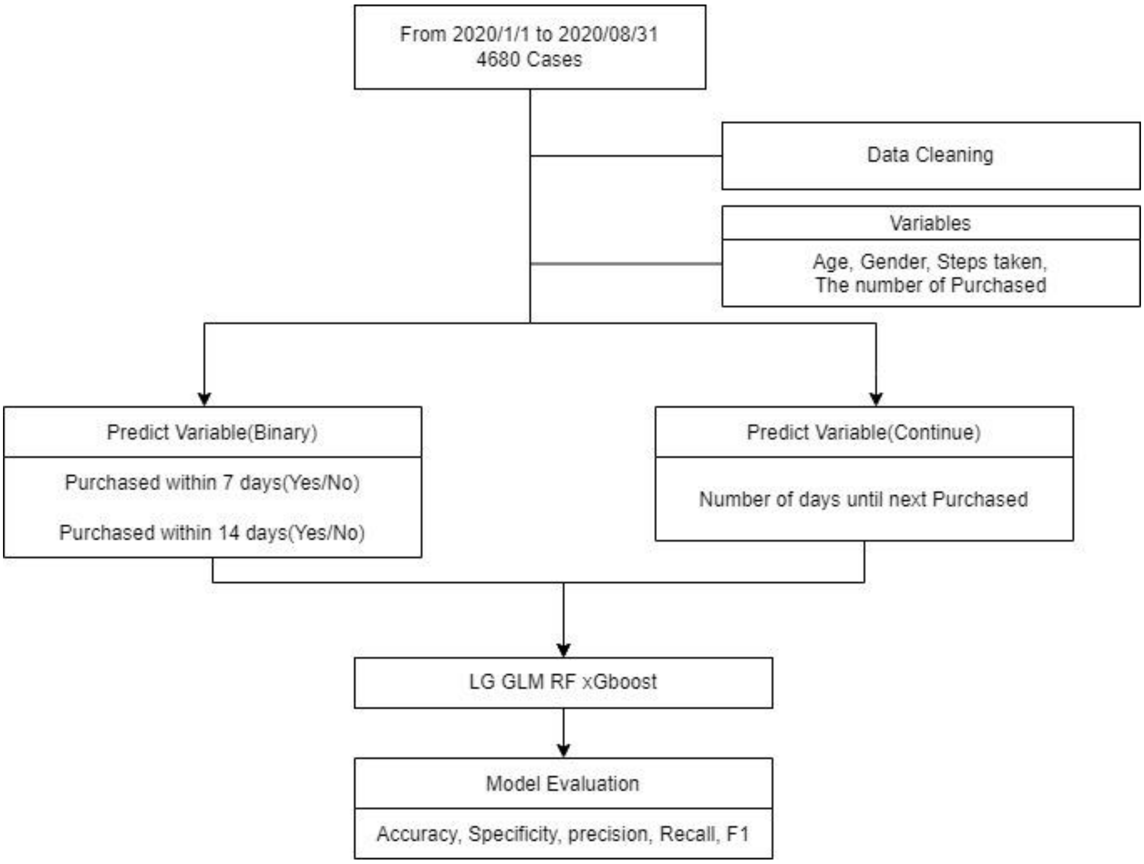

From 2020/1/1 to 2020/08/31
4680 Cases
Data Cleaning
Variables
Age, Gender, Steps taken, The number of Purchased
Predict Variable(Binary)
Purchased within 7 days(Yes/No)
Purchased within 14 days(Yes/No)
Predict Variable(Continue)
Number of days until next Purchased
LG GLM RF xGboost
Model Evaluation
Accuracy, Specificity, precision, Recall, F1

Figure 1. Data processing flowchart.

2.1. Data source

The data for this study come from the product sales records of H Company. The products combine walking steps, lifestyle perks, and social interaction, where accumulated steps can be exchanged for various rewards and gifts. This not only brings health benefits to users but also strengthens their motivation and connection to visit stores. It offers personalized recommendations and discounts. These products have accumulated millions of users globally and generate a large volume of transaction data daily.

2.2. Variable description

The independent variables used in this study are age, gender, steps taken, coffee purchase, the number of coffee purchases, the number of dessert purchases, the number of beverage purchases, the number of food purchases, the number of luxury life product purchases, total purchase frequency, discount price, and total purchase amount. The independent variables used in this study are the number of days between the last and the penultimate coffee purchase, the number of days between the penultimate and antepenultimate.

2.3. Statistical methods

We used four machine learning methods to predict consumers' coffee purchasing behavior: GLM, Logistic Regression, Random Forest, and XGBoost. To validate the predictive effectiveness of our models, we employed 5-fold cross-validation. This method divides the dataset into five equal parts, using one part as the test set and the remaining four as the training set. This process is repeated five times, each time with a different part as the test set, and the model's average predictive performance is calculated. This approach allows for a more accurate assessment of the model's performance and avoids variations in results due to different data-splitting methods.

2.3.1. Generalized Linear Models (GLM)

Generalized Linear Models (GLM) are statistical models that allow us to predict a dependent variable based on one or more explanatory variables. GLM extends the linear regression model by accommodating irregular error distributions in the dependent variable. A key feature of GLM is its combination of linear regression and link functions, enabling the model to handle various types of dependent variables like binomial and Poisson distributions [15,16]. GLM is widely used in insurance data as it can effectively handle large datasets and varying exposure times [17]. Additionally, GLM is used to predict PM10 concentration in urban outdoor environments based on concentrations of air pollutants in the atmosphere and meteorological variables [18].

2.3.2. XGBoost

XGBoost is a gradient-boosting machine learning method used to construct a strong predictive model, which is a linear combination of weak predictive models (like decision trees) [19]. In this study, we used XGBoost to predict when consumers will repurchase coffee. XGBoost builds a strong predictive model by iteratively adding new weak predictive models to improve prediction errors. One major advantage of XGBoost is its efficient handling of a large number of input variables and its capability to manage nonlinearities and interaction effects.

2.3.3. Logistic Regression

Logistic Regression is a statistical learning method used to predict the probability of a binary response variable [20]. In our study, we used Logistic Regression to predict whether consumers will purchase coffee at the next time point. Logistic Regression establishes a predictive model by converting a linear combination of input variables into a probability between 0 and 1. A major advantage of Logistic Regression is its provision of estimates for the influence of each input variable on the predicted probability.

2.3.4. Random Forest

Random Forest is an ensemble learning method that combines the predictions of multiple decision trees to improve the accuracy and stability of predictions [21]. In our study, we used Random Forest to predict when consumers will repurchase coffee. Random Forest increases model diversity by introducing randomness in the construction process of each decision tree. A key advantage of Random Forest is its effective handling of a large number of input variables and its capability to manage nonlinearities and interaction effects.

We used the R software to implement these four machine learning methods. Our dataset includes consumers' purchase history, demographic information, and other factors that might influence purchasing behavior. We used this data to train our models and employed cross-validation to evaluate the performance of the models.

3. Results

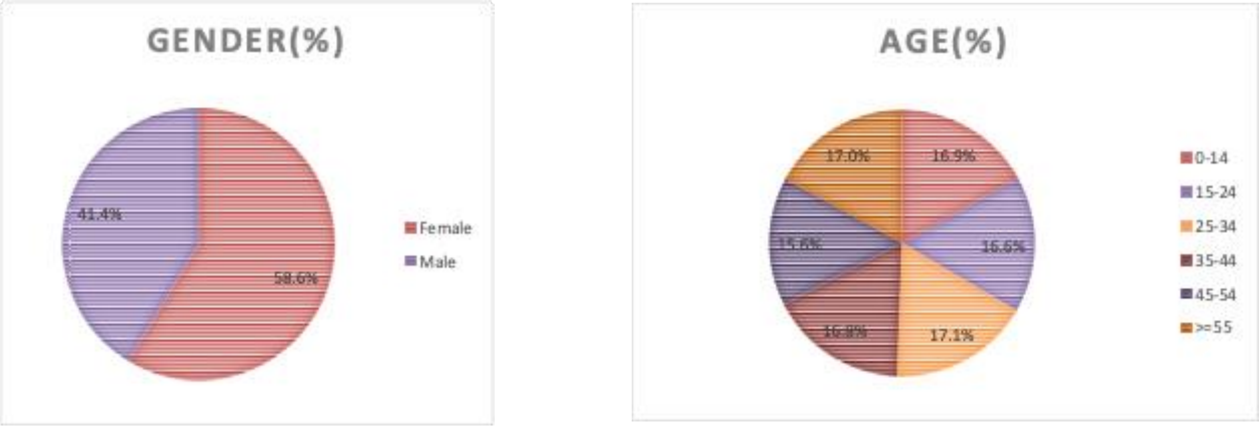
As shown in **Table 1**, the sample size of this study is 4680 individuals. In terms of gender, there are 2744 females, accounting for 58.6% of the total, and 1936 males, making up 41.4% (**Figure 2A**). The age distribution is relatively balanced across different age groups. There are 791 individuals aged 0–14 years (16.9%), 777 individuals aged 15–24 years (16.6%), 801 individuals aged 25–34 years (17.1%), 786 individuals aged 35–44 years (16.8%), 728 individuals aged 45–54 years (15.6%), and 797 individuals aged 55 years and above (17%) (**Figure 2B**).

On average, consumers take 6,492.6 steps, make 11.6 purchases, and spend 402.8 TWD in total. The number of steps taken ranges from 121 to 23,997, while purchase frequency varies from 2 to 321 times. Total spending ranges from 60 to 9636 TWD. The standard deviation of total spending is 779.2 TWD, the standard deviations for steps and purchase frequency are 3458.3 and 24.2(**Table 1**).

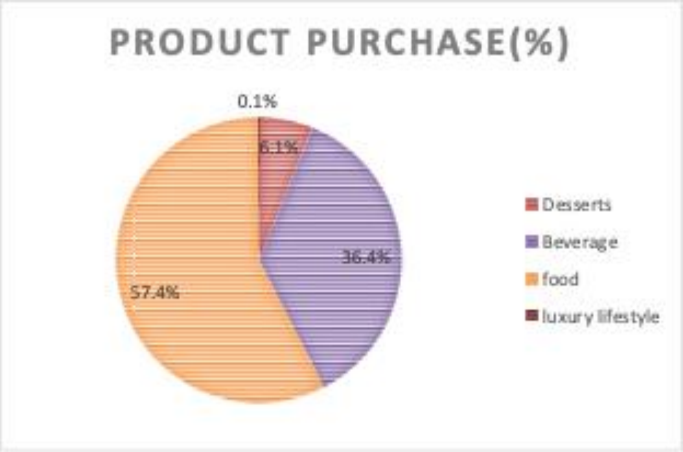
Regarding the frequency of purchasing items, desserts were purchased 400 times (6.1%), beverages 2376 times (36.4%), food 3744 times (57.4%), and luxury lifestyle products only 8 times (0.1%) (**Figure 2C**).

As indicated in **Table 2**, the average amount spent on coffee purchases is 423.25 TWD for the 0–14 age group, 399.46 TWD for 15–24 years, 394.81 TWD for 25–34 years, and 381.79 TWD for 35–44 years. Observing this data, we find that the younger age group (0–14 years) spends more on average for coffee, and the average spending decreases with age. The 35–44 age group seems to spend the least on average for coffee purchases, we observe that the frequency of coffee purchases across different age groups is relatively similar. However, the 0–14 age group seems to purchase slightly more, with 9694 coffee purchases, accounting for 17.79% of the total.

Table 3 shows the average amount spent on coffee purchases by different genders. The average for females is 369.88 TWD, while for males, it is significantly higher at 449.63 TWD. These data indicate that males, on average, spend more on coffee purchases than females. Females purchased coffee 28,248 times, representing 51.8% of the total purchases, while males made 26,240 coffee purchases, accounting for 48.2%. In total, there were 54,488 coffee purchases. Although the number of purchases by females and males is close, females purchased slightly more. This might suggest that, in this sample, females buy coffee more frequently than males.



(A) **(B)**



(C)

Figure 2. Data distribution of **(A)** gender; **(B)** age; **(C)** product purchase.

Table 1. Descriptive statistical distribution.

	Cases(n)	Percentage (%)	
Total	4680	100%	
Gender			
Female	2744	58.6	
Male	1936	41.4	
Age			
0–14	791	16.9	
15–24	777	16.6	
25–34	801	17.1	
35–44	786	16.8	
45–54	728	15.6	
>= 55	797	17	
Number of product purchases			
Desserts	400	6.1	
Beverage	2376	36.4	
food	3744	57.4	
luxury lifestyle	8	0.1	
	Number of coffee purchases per person	**Discount amount spent on coffee per person**	**Number of walking steps taken for coffee purchases**
Mean	11.6	402.8	6492.6
SD	24.2	779.2	3458.3
Max	320.0	9636.0	23997.0
Min	2.0	60.0	121.0

Table 2. Cross-table of age ranges and number of coffee purchases and average amount spent on coffee purchases.

Age Ranges	Average Purchase Amount	Number of Purchases	Percentage (%)
0–14	423.25	9694	17.79%
15–24	399.46	8943	16.41%
25–34	394.81	9161	16.81%
35–44	381.79	8650	15.88%
45–54	397.39	8300	15.23%
>= 55	419.87	9740	17.88%

Table 3. Cross-table of gender and number of coffee purchases and average amount spent on coffee purchases.

Gender	Average Purchase Amount	Number of Purchases	Percentage (%)
Female	369.88	28248	51.8%
Male	449.63	26240	48.2%

Table 4 primarily predicts whether coffee will be purchased within seven days. According to the data, regardless of whether the number of coffee purchases is included as a variable, the XGBoost model has the highest accuracy, about 0.807. The accuracy of the GLM model ranges between 0.7127 and 0.7149, while that of the Random Forest model is between 0.7785 and 0.7851.

Table 5 predicts the number of days within which coffee will be purchased. Three evaluation metrics were used: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The LM model, without including the number of coffee purchases, has an MSE of 80.0754, which significantly increases to 792,437.8083 when this variable is included. Conversely, the MSE values for XGBoost and Random Forest models are relatively close in both scenarios, ranging between 55 and 58.

Overall, the XGBoost and Random Forest models perform more consistently across these three prediction models, while the GLM and LM models show more variability, especially when considering the number of coffee purchases as a variable (see also Appendix A.).

Table 4. Prediction of coffee purchase s within seven days whether the number of coffee purchases is included as an independent variable.

	GLM		XGBoost		Random Forest	
	No	**Yes**	**No**	**Yes**	**No**	**Yes**
Accuracy	0.7127	0.7149	0.8070	0.8070	0.7851	0.7785
Specificity	0.9758	0.9789	0.9851	0.9851	0.9522	0.9403
Precision	0.2000	0.2222	0.8837	0.8837	0.7091	0.6667
Recall	0.016	0.016	0.3140	0.3140	0.3223	0.3306
F1	0.0296	0.0299	0.4634	0.4634	0.4432	0.4420

Table 5. Prediction of the number of days until the next coffee purchase whether the number of coffee purchases is included as an independent variable.

	LM		**XGBoost**		**Random Forest**	
	No	**Yes**	**No**	**Yes**	**No**	**Yes**
MSE	80.0754	792437.8083	55.0468	55.1244	56.8027	57.2339
RMSE	8.9482	532.6803	7.4193	7.4245	7.5367	7.5653
MAPE	205.1352	2135.5098	145.4540	146.4057	139.9343	141.6914

4. Discussion

As businesses increasingly rely on data-driven strategies, understanding consumer purchasing behavior through machine learning can provide valuable insights for marketing optimization. This study aimed to explore how machine learning models can enhance sales predictions and support strategic decision-making in precision marketing. By analyzing consumer transactions, we identified key factors influencing repurchase behavior and evaluated the effectiveness of different predictive models. The following discussion highlights the practical implications, model performance, feature impact, and challenges associated with implementing machine learning in marketing applications.

4.1. Model selection and business implications

Understanding consumer purchasing behavior is essential for businesses seeking to enhance marketing strategies. Our findings confirm that tree-based models, particularly XGBoost, outperform traditional statistical models in predicting consumer repurchase behavior. The adaptability of tree-based models to changing consumer trends and market conditions makes them highly suitable for dynamic marketing environments. In contrast, Generalized Linear Models (GLM) and Linear Models (LM) often struggle with nonlinear patterns, reinforcing the need for advanced machine learning methods in consumer behavior analysis [22].

By leveraging machine learning models, businesses can improve customer targeting, optimize promotions, and refine product recommendations based on historical purchasing behavior. These insights contribute to more precise and efficient marketing strategies, helping companies enhance customer engagement and retention [23].

4.2. Feature impact and marketing insights

Feature selection plays a crucial role in consumer behavior prediction, as different factors influence purchasing decisions. Our analysis found that incorporating the number of previous coffee purchases as a predictor did not significantly improve accuracy, suggesting that historical purchases alone may not fully capture repurchase intent. However, the increase in MSE in the LM model after adding this feature indicates potential collinearity or outliers, both of which can reduce model performance [24].

This finding aligns with previous research showing that consumer behavior is influenced by multiple interrelated factors, such as demographics, pricing, and promotional incentives (see also Appendix B). Businesses should consider a holistic approach to consumer analysis, integrating multiple variables to improve prediction accuracy and develop more effective marketing campaigns [25].

4.3. Impact of prediction period on business decisions

Predictive accuracy tends to decline as the forecasting period extends, primarily due to evolving consumer behavior and external market influences [26]. While tree-based models demonstrate better resilience to longer prediction periods, they still face limitations in capturing long-term shifts in purchasing trends.

For businesses, this underscores the need for frequent model updates and recalibration to maintain accuracy in marketing forecasts. Companies should adopt adaptive learning techniques and real-time data analysis to ensure that marketing strategies remain responsive to changing consumer preferences and external factors such as seasonal trends and economic shifts [26].

4.4. Choice of evaluation metrics for business applications

Selecting appropriate evaluation metrics is critical for ensuring that predictive models align with business objectives. While MSE, RMSE, and MAPE are commonly used for model evaluation, each metric has different business implications. MSE and RMSE are more sensitive to extreme values, making them useful for identifying high-impact purchasing behaviors, whereas MAPE provides intuitive interpretability, making it more suitable for non-technical business stakeholders [25].

For practical marketing applications, businesses may benefit from customized performance metrics that align with revenue impact and customer engagement goals. Future studies could explore business-driven evaluation metrics that optimize not only prediction accuracy but also real-world marketing effectiveness.

4.5. Future model development for business optimization

Advancing consumer behavior prediction models requires deeper data exploration and refined feature engineering. Understanding key purchasing drivers—such as personalized promotions, loyalty program participation, and seasonal demand fluctuations—can help businesses develop more targeted marketing strategies [23].

Additionally, advanced data analysis techniques such as real-time machine learning and AI-driven segmentation could enhance the effectiveness of predictive models, allowing businesses to anticipate consumer needs and adapt marketing efforts accordingly [25].

4.6. Limitations and challenges

Despite the promising results, this study presents several limitations and challenges that must be considered. First, data limitations restrict the scope of our analysis, as the dataset used in this study is confined to a specific time period and consumer segment. This may limit the generalizability of our findings to broader markets or different consumer behavior patterns. Future studies should explore additional datasets from various time frames and geographic locations to enhance model robustness.

Second, external factors such as economic conditions, market trends, and seasonal variations may significantly impact consumer purchasing behavior. Since our model is trained on historical transaction data, it may not fully account for these dynamic influences. Integrating external data sources, such as macroeconomic indicators or sentiment analysis from social media, could improve predictive accuracy.

Third, while XGBoost demonstrated superior predictive performance, model interpretability remains a challenge. Decision-tree-based ensemble models are often considered "black-box" techniques, making it difficult for businesses to fully understand the reasoning behind specific predictions. Future research should focus on incorporating explainable AI (XAI) techniques, such as Shapley Additive explanations (SHAP) values or feature importance analysis, to enhance transparency and trust in the model's predictions.

Additionally, the study did not include external validation with unseen data due to dataset availability constraints. While five-fold cross-validation was applied to mitigate overfitting and assess model reliability, testing on an independent dataset would further confirm the model's generalization ability. Future work should consider applying the model to new datasets to evaluate its adaptability in real-world scenarios.

Lastly, implementing machine learning models in business operations poses managerial challenges. Companies must determine how to effectively translate predictive insights into actionable strategies, such as customer retention initiatives, personalized promotions, or inventory adjustments. Practical deployment considerations, such as computational efficiency, data privacy concerns, and integration with existing marketing systems, should also be addressed.

5. Conclusion

In today's business environment, predicting consumer purchasing behavior has become a core element of competitive advantage. Understanding when consumers are likely to repurchase products is crucial for inventory management, sales strategy, and marketing campaign planning. From a managerial perspective, this predictive ability offers several practical benefits. First, businesses can more accurately forecast sales volumes, optimizing inventory management and reducing overstock or stockout situations. Second, based on predictive results, businesses can adjust marketing strategies, such as sending coupons or launching promotional activities before the expected repurchase time. Additionally, this predictive model enables businesses to better understand consumer needs and preferences, allowing them to offer more personalized products and services.

Our findings indicate that XGBoost consistently outperforms other machine learning models in predicting consumer repurchase behavior. This superiority stems from its ability to handle nonlinear relationships, capture complex feature interactions, and effectively manage missing values and outliers. Compared to traditional models such as Generalized Linear Models (GLM) and Logistic Regression, which assume linear relationships, XGBoost leverages gradient boosting and decision trees, allowing it to make more accurate predictions by iteratively refining its learning process. Furthermore, its regularization techniques help prevent overfitting, enhancing generalization to unseen data.

Beyond accuracy, our evaluation demonstrates that XGBoost not only achieves higher classification accuracy but also performs better across other key metrics. While traditional models primarily rely on Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) for regression tasks, our classification-based evaluation incorporates precision, recall, and F1-score to provide a more comprehensive assessment. The results reveal that XGBoost exhibits higher precision and recall, particularly in identifying customers likely to repurchase coffee within the predicted timeframe. This suggests that XGBoost is not only highly accurate but also minimizes false positives and false negatives—an essential factor in developing effective targeted marketing strategies.

Author contributions: Conceptualization, MC; methodology, KHW and YHT; software, KHW and YHT; validation, BCS and MC; formal analysis, KHW and YHT; investigation, KHW and YHT; resources, BCS; data curation, KHW; writing—original draft preparation, KHW; writing—review and editing, KHW and MC; visualization, KHW; supervision, MC; project administration, MC; funding acquisition, YHT. All authors have read and agreed to the published version of the manuscript.

Institutional review board statement: Not applicable.

Informed consent statement: Not applicable.

Conflict of interest: The authors declare no conflicts of interest.

References

1. Taiwan Network Information Center. 2018 Taiwan Internet Report. Taiwan Network Information Center; 2019.
2. Li J, Pan S, Huang L, Zhu X. A Machine Learning Based Method for Customer Behavior Prediction. Tehnički vjesnik. 2019; 26(6): 1774–1781. doi: 10.17559/tv-20190603165825
3. Oster E. Diabetes and Diet: Purchasing Behavior Change in Response to Health Information. American Economic Journal: Applied Economics. 2018; 10(4): 308–348. doi: 10.1257/APP.20160232
4. Malthouse EC, Blattberg RC. Can we predict customer lifetime value. Journal of Interactive Marketing. 2005; 19(1): 2–16.
5. Kahan R. Using database marketing techniques to enhance your one-to-one marketing initiatives. Journal of Consumer Marketing. 1998; 15(5): 491–493.
6. Miglautsch JR. Thoughts on RFM scoring. Journal of Database Marketing & Customer Strategy Management. 2000; 8: 67–72.
7. Pfeiffer J, Pfeiffer T, Meiβner M, et al. Eye-Tracking-Based Classification of Information Search Behavior Using Machine Learning: Evidence from Experiments in Physical Shops and Virtual Reality Shopping Environments. Information Systems Research. 2020; 31(3): 675–691. doi: 10.1287/ISRE.2019.0907
8. Kharfan M, Chan VWK, Firdolas Efendigil T. A data-driven forecasting approach for newly launched seasonal products by leveraging machine-learning approaches. Annals of Operations Research. 2020; 303: 159–174. doi: 10.1007/s10479-020-03666-w
9. Purnomo MRA, Azzam A, Khasanah AU. Effective Marketing Strategy Determination Based on Customers Clustering Using Machine Learning Technique. In: Proceedings of the 1st Bukittinggi International Conference on Education; 17–18 October 2019; West Sumatera, Indonesia.
10. Lim T. K-Means Clustering-Based Market Basket Analysis: UK Online E-Commerce Retailer. In: Proceedings of the 2021 International Conference on Information Technology (ICIT); 14–15 July 2021; Amman, Jordan. pp. 126–131.

11. Zuo Y, Yada K, Ali ABMS. Prediction of Consumer Purchasing in a Grocery Store Using Machine Learning Techniques. In: Proceedings of the 2016 Asia-Pacific World Congress on Computer Science and Engineering; 5–6 December 2016; Nadi, Fiji.
12. Nisha, Singh AS. Customer Behavior Prediction using Deep Learning Techniques for Online Purchasing. In: Proceedings of the 2023 International Conference on Innovative and Online Learning; 3–5 March 2023; Bangalore, India.
13. Yao S. Method and Research of E-commerce Consumers' Purchasing Behavior Forecast and Analysis Based on Machine Learning. In: Proceedings of the 2021 ACM International Conference on Information Technology for Social Good; 23–25 October 2021; Manchester, United Kingdom.
14. Bangyal W, Ashraf A, Shakir R, et al. A Review on Consumer Behavior Towards Online Shopping using Machine Learning. International Journal of Emerging Multidisciplinary Sciences and Artificial Intelligence. 2022; 1(1). doi: 10.54938/ijemdcsai.2022.01.1.84
15. Yang C, Chandler RE, Isham VS, et al. Spatial-temporal rainfall simulation using generalized linear models. Water Resources Research. 2005; 41(11).
16. Czado C, Raftery AE. Choosing the link function and accounting for link uncertainty in generalized linear models using Bayes factors. Statistical Papers. 2006; 47(3): 419–442.
17. Jong PD, Heller GZ. Generalized Linear Models for Insurance Data. Cambridge University Press; 2008.
18. Garcia JM, Teodoro F, Cerdeira R, et al. Developing a methodology to predict PM10 concentrations in urban areas using generalized linear models. Environmental Technology. 2016; 37(18): 2316–2325.
19. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 13–17 August 2016; San Francisco, CA, USA. pp. 785–794.
20. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. John Wiley & Sons; 2013.
21. Breiman L. Random forests. Machine learning. 2001; 45(1): 5–32.
22. Hayati SNY, Dhiaa A, Ain N. Strategies for Accurate Arabica Coffee Export Forecasts. Klover Multidisciplinary Journal of Engineering. 2023; 8(2): 134–143.
23. von Loeben SC, Gornott C, Abigaba D, et al. Climate risk analysis for adaptation planning in Uganda's agricultural sector: An assessment of maize and coffee value chains. Available online: https://publications.pik-potsdam.de/pubman/faces/ViewItemFullPage.jsp?itemId=item_28791_2&view=ACTIONS (accessed on 5 January 2025).
24. Hasyati B. Customer Churn Prediction Model Design Using Predictive Analytics for Modern Coffee Shop [Master's thesis]. IPB University; 2023.
25. Dias CG, Martins FB, Martins MA. Climate risks and vulnerabilities of the Arabica coffee in Brazil under current and future climates considering new CMIP6 models. Science of the Total Environment. 2024; 907: 167753.
26. Nguyen NT, Phan VT, Duong TAN, et al. A Model for Alliance Partner Selection Based on GM (1, 1) and DEA Frameworks—Case of Vietnamese Coffee Industry. In: Proceedings of the 12th Conference on Information Technology and Its Applications; 28–29 July 2023; Da Nang City, Vietnam.

Appendix A. ROC Curves for GLM, XGBoost, and Random Forest models

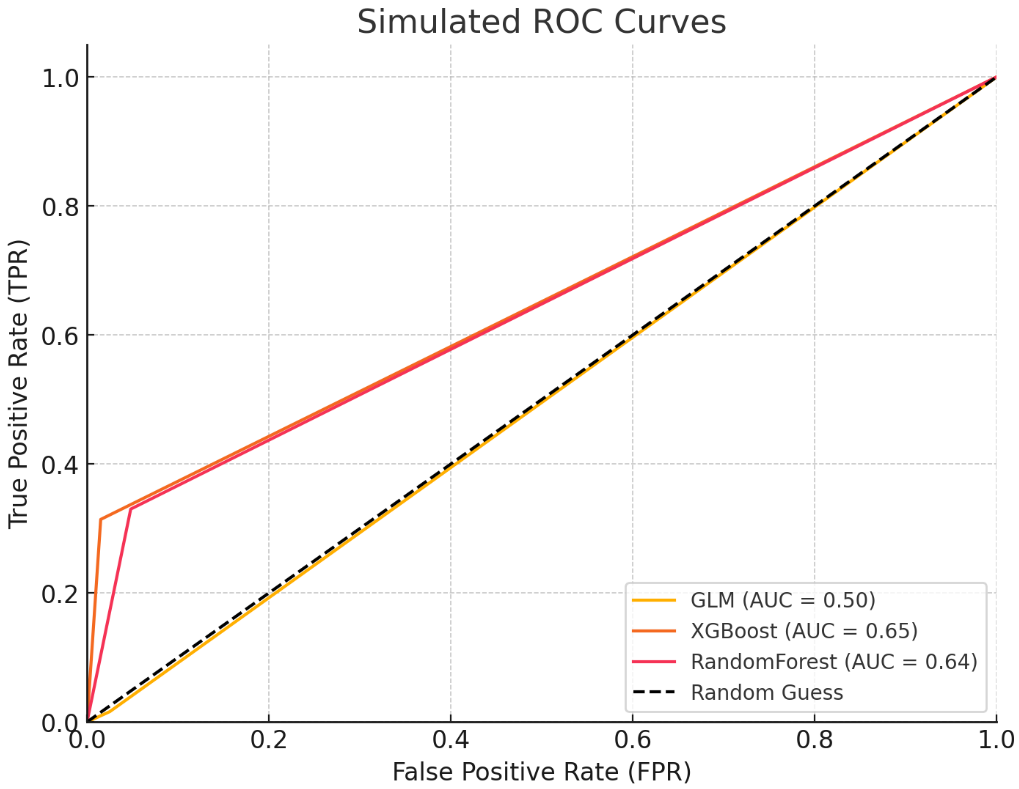


Figure A1. AUC-ROC Curves.

Appendix B. Correlation Coefficient Matrix and Plot

Table B1. Correlation Coefficient Matrix.

Variable	Steps Taken	Coffee Purchase Frequency	Total Coffee Purchase Amount
Steps Taken	1.0000	0.0004	−0.0050
Coffee Purchase Frequency	0.0004	1.0000	0.9778
Total Coffee Purchase Amount	−0.0050	0.9778	1.0000

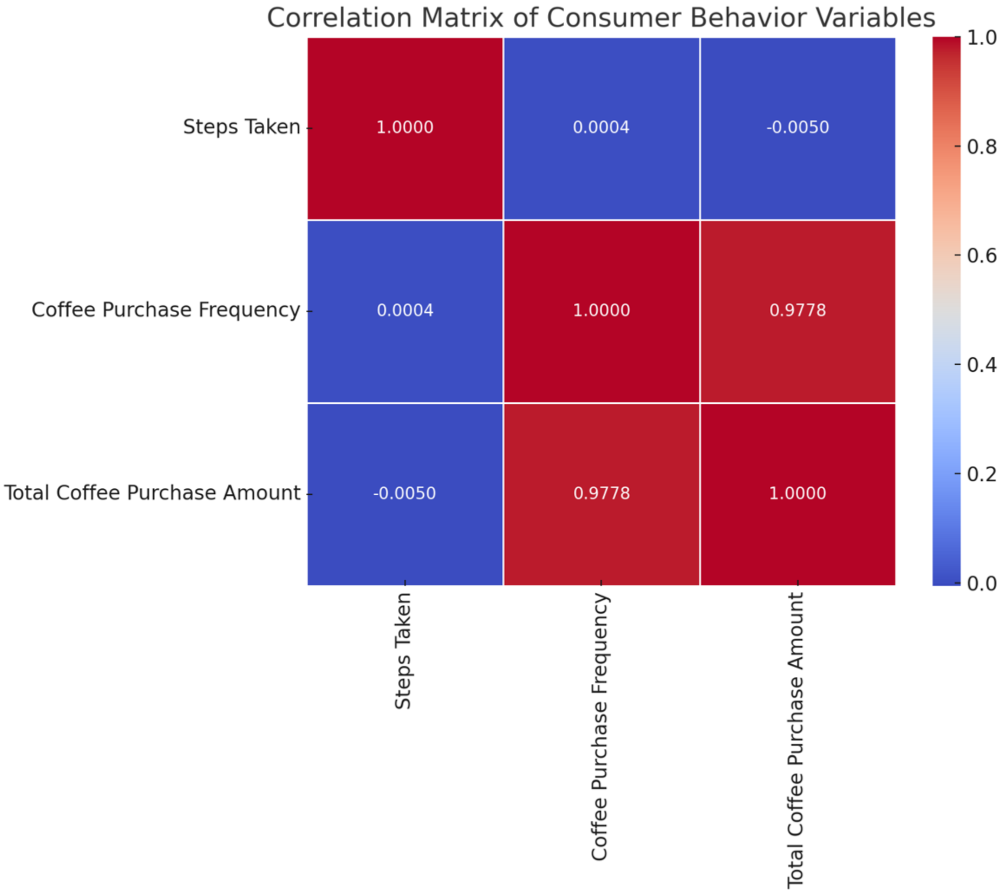


Figure B1. Correlation Coefficient Matrix Plot.