BENCHMARKING MACHINE LEARNING MODELS FOR PREDICTIVE ANALYTICS IN E-COMMERCE STRATEGY

Rattapol Kasemrat *, Tanpat Kraiwanit **, Aitsari Khaothawirat ***, Sutthiporn Chinnapha ***, Qiqi Luo ***

* Faculty of Economics, Rangsit University, Pathum Thani, Thailand
** *Corresponding author*, International College, Pathumthani University, Pathum Thani, Thailand
Contact details: International College, Pathumthani University, 140 Moo 4 Tiwanon Road, Ban Klang, Mueang, Pathum Thani, Thailand
*** International College, Pathumthani University, Pathum Thani, Thailand

OPEN ACCESS

How to cite this paper:
Kasemrat, R., Kraiwanit, T., Khaothawirat, A., Chinnapha, S., & Luo, Q. (2025). Benchmarking machine learning models for predictive analytics in e-commerce strategy. *Corporate & Business Strategy Review, 6*(2), 146–155.
https://doi.org/10.22495/cbsrv6i2art15

ISSN Online: 2708-4965
ISSN Print: 2708-9924

Received: 19.05.2024
Revised: 26.09.2024; 17.10.2024; 18.04.2025
Accepted: 05.05.2025

JEL Classification: C53, C55, D12, L81, M31
DOI: 10.22495/cbsrv6i2art15

Abstract

Predictive analytics is crucial in the digital economy, revolutionizing e-commerce by utilizing data insights to forecast trends, personalize customer experiences, and optimize operations with high accuracy (Gupta & Bansal, 2020; Jakkula, 2023). This research examines the comparative performance of traditional statistical techniques versus modern machine learning models in predicting customer behavior within Thailand's e-commerce sector. Utilizing a comprehensive dataset from Thailand's leading e-commerce platform, encompassing 10,000 customer interactions, this study analyzes the effectiveness of logistic regression, support vector machines (SVM), k-nearest neighbors (KNN), and random forest in modeling purchasing patterns. The findings demonstrate that SVM and KNN substantially outperform logistic regression in accuracy, precision, recall, and area under the curve (AUC), with random forest also showing significant capabilities in managing complex, large-scale datasets. This research highlights the critical role of advanced machine learning technologies in refining strategic decision-making within e-commerce by offering more accurate customer segmentation and enhanced targeting strategies. Given the swift growth of e-commerce in markets like Thailand, these insights provide crucial strategic implications for both local and international market contexts, suggesting a pivotal shift towards integrating machine learning to capitalize on the expansive digital consumer data available.

Keywords: Predictive Analytics, Model Comparison, Machine Learning, Statistical Techniques, Support Vector Machines, K-Nearest Neighbors

Authors' individual contribution: Conceptualization — R.K. and T.K.; Methodology — R.K. and T.K.; Software — R.K. and T.K.; Validation — R.K. and T.K.; Formal Analysis — R.K. and T.K.; Investigation — R.K. and T.K.; Resources — R.K. and T.K.; Writing — Original Draft — R.K. and T.K.; Writing — Review & Editing — R.K. and T.K.; Supervision — T.K., A.K., S.C., and Q.L.

Declaration of conflicting interests: The Authors declare that there is no conflict of interest.

1. INTRODUCTION

In today's rapidly evolving e-commerce landscape, predicting consumer behavior is crucial for maintaining a competitive edge, especially in Thailand's growing digital marketplace. The Thai e-commerce sector is projected to expand significantly over the next five years, characterized by diverse demographics and changing consumption patterns (ECDB, 2024). Accurately predicting customer behavior in such a dynamic market is essential for businesses seeking to stay ahead.

Traditionally, predictive analytics in e-commerce has relied on statistical techniques like logistic regression. However, these models often struggle to handle the complexity and volume of modern e-commerce data, especially in capturing non-linear relationships between consumer attributes and purchasing behaviors (Zaghloul et al., 2024). In contrast, modern machine learning techniques such as support vector machines (SVM) and k-nearest neighbors (KNN) have emerged as effective alternatives. SVM handles high-dimensional data well, using kernel functions to capture non-linear patterns, while KNN adapts to intricate relationships without assuming a fixed data distribution, making both models superior for analyzing complex e-commerce data (Gupta & Bansal, 2024).

Despite these advantages, there is a lack of comparative research on these techniques within the Thai e-commerce context, with existing literature often overlooking localized market dynamics (Gupta & Bansal, 2024). This study addresses this gap by analyzing the effectiveness of traditional versus machine learning models in predicting customer behavior, using a comprehensive dataset of 10,000 customer interactions from a leading Thai e-commerce platform.

The primary aim of this study is to conduct a comparative analysis of traditional statistical techniques and modern machine learning models to determine their effectiveness in predicting customer behavior within the Thai e-commerce sector. This research seeks to answer the following key questions:

RQ1: Which predictive models — traditional or machine learning — offer the highest accuracy in predicting customer purchasing behavior in the Thai e-commerce market?

RQ2: Under what conditions do machine learning models such as support vector machines and k-nearest neighbors outperform traditional methods in capturing complex purchasing patterns?

To address these questions, this study utilizes a comprehensive dataset of 10,000 customer interactions from Thailand's leading e-commerce platform, applying logistic regression, SVM, KNN, and random forest models to evaluate their predictive performance. By examining these models in a real-world, high-growth e-commerce environment, the research contributes valuable empirical evidence on the most effective techniques for improving customer segmentation, targeting strategies, and overall business performance.

The rest of the paper is structured as follows. Section 2 reviews relevant literature, Section 3 outlines the methodology, Section 4 presents the results, Section 5 discusses the implications, limitations, and recommendations for future research, and Section 6 provides conclusions.

2. LITERATURE REVIEW

Predictive analytics has significantly transformed the e-commerce landscape by enhancing decision-making processes through data-driven insights (Gupta & Bansal, 2020; Jakkula, 2023). Recent studies emphasize the integration of machine learning algorithms and big data analytics, providing e-commerce businesses with tools to improve customer satisfaction, optimize marketing strategies, and prevent fraud.

Traditional statistical methods such as logistic regression have long provided foundational insights in this area. However, their limitations become increasingly apparent as the volume and complexity of data grow. These methods often fail to capture the complex, non-linear relationships that are characteristic of modern e-commerce environments. In contrast, advanced machine learning techniques such as SVM and KNN have been found to handle high-dimensional spaces more effectively, offering superior classification accuracy and deeper behavioral insights (Tian et al., 2024).

Despite the growing adoption of machine learning techniques, there remains a notable gap in empirical research that directly compares these advanced methods with traditional approaches within specific markets. Tian et al. (2024) provided evidence from the United States (U.S.) market showing that machine learning models outperform traditional methods in terms of accuracy and insight. However, the unique dynamics of the Thai e-commerce market, such as its diverse consumer base and rapid digital transformation, necessitate localized research to verify these findings (Feng et al., 2024).

Zaghloul et al. (2024) highlight the positive impact of predictive analytics on customer satisfaction, comparing traditional machine learning approaches with deep learning models. Their research indicates that deep learning models provide more accurate predictions of customer satisfaction in e-commerce, underscoring the value of advanced analytics techniques.

Krishnan and Mariappan (2024) discuss how artificial intelligence (AI)-driven customization influences consumer behavior, enhancing engagement and loyalty. Similarly, Gupta and Bansal (2020) highlight the effectiveness of machine learning in optimizing product recommendations, thereby increasing conversion rates. Additionally, Paripati et al. (2024) explore the role of AI algorithms in improving recommender systems, which are essential for tailoring customer experiences in the e-commerce sector.

Reddy et al. (2024) demonstrate how machine learning models efficiently identify and prevent fraudulent activities, thereby safeguarding both businesses and consumers. Feng et al. (2024) investigate big data analytics for predicting customer churn and implementing retention strategies, which are crucial for maintaining a secure and stable customer base.

Tadimarri et al. (2024) explore the utilization of AI in agile project management within e-commerce settings, emphasizing the potential for real-time adjustments in business strategies based on predictive data. Akadji and Dewantara (2024) focus on big data analysis for product demand prediction in Indonesian e-commerce, showcasing how data-driven insights can optimize inventory management and logistics. Furthermore, Wenzel et al. (2025) discuss the impact of robotic mobile fulfillment systems, which leverage predictive analytics to enhance operational efficiency.

Despite its benefits, the application of predictive analytics in e-commerce is not without challenges. Issues such as data privacy, ethical considerations in data usage, and the potential for bias in AI models are of growing concern. Behare et al. (2024) discuss the ethical implications of AI in e-commerce, urging for transparent and responsible use of predictive technologies. Gill et al. (2024) emphasize the need for ethical standards in predictive analytics to ensure fair and unbiased outcomes.

Looking forward, the convergence of predictive analytics with emerging technologies like blockchain and the Internet of Things (IoT) promises to further revolutionize e-commerce. Johansson (2024) explores the feasibility of AI-driven insights for decision-making in an e-commerce environment, highlighting the potential of combining various data sources for comprehensive analytics. Verma and Rawat (2024) delve into the interaction between blockchain and AI, proposing a future where decentralized intelligence enhances predictive capabilities.

Abdulrashid et al. (2024) examine the impact of social media analytics on predictive models, illustrating how data from social platforms can forecast consumer trends and influence e-commerce strategies. Chowdhury et al. (2024) highlight the role of AI-driven digital marketing in shaping consumer purchase intentions, stressing the importance of integrating social media insights into predictive analytics.

Aliyev et al. (2024) discuss the modernization of logistics and e-commerce platforms using AI technologies, which can improve efficiency and service quality in various sectors. This is further supported by studies on the integration of AI in different operational areas, such as warehouse management and delivery logistics (Khandait et al., 2024).

Predictive analytics stands at the forefront of e-commerce innovation, driving significant improvements in customer personalization, operational efficiency, and security. The advancements in machine learning and big data analytics have provided e-commerce businesses with powerful tools to anticipate customer needs, prevent fraudulent activities, and optimize supply chains. However, the integration of these technologies is accompanied by challenges, particularly concerning data privacy and ethical considerations. As the field continues to evolve, the convergence of predictive analytics with other emerging technologies like blockchain and IoT will likely unlock new opportunities for further advancements in the e-commerce sector. It is imperative for businesses to adopt these technologies responsibly, ensuring transparency and ethical standards are upheld.

3. RESEARCH METHODOLOGY

3.1. Data source

This study analyzed 10,000 customer interactions from Thailand's leading e-commerce platform, including demographics, browsing patterns, purchase history, and engagement metrics. All data were anonymized per General Data Protection Regulation (GDPR) standards.

3.2. Data preprocessing

Data preprocessing involved cleaning (removing duplicates, handling missing values), normalization/scaling (for models like SVM and KNN), feature selection (using correlation coefficients and feature importance), and splitting the data into training (70%) and testing (30%) sets.

3.3. Model implementation

Four models were implemented to predict customer purchasing behavior: 1) logistic regression (as a baseline for binary classification), 2) SVM (with linear and radial basis function (RBF) kernels for high-dimensional data), 3) KNN (optimized via grid search cross-validation), and 4) random forest (an ensemble method to improve accuracy and control overfitting). All models were evaluated under identical conditions.

SVM is superior to logistic regression in handling non-linear data through its use of kernel functions, allowing it to capture complex relationships between variables that traditional methods cannot. KNN, a distance-based model, excels by adapting to local data structures without making strong assumptions about the underlying data distribution, making it effective in modeling varied customer behaviors. Both SVM and KNN provide more flexible and accurate modeling than logistic regression, particularly in dynamic e-commerce environments where customer patterns are often complex and non-linear.

3.4. Evaluation metrics

Performance was assessed using accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) to gauge each model's ability to predict purchasing behavior.

3.5. Validation

K-fold cross-validation ensured model robustness, using each data subset once as a validation set to assess predictive accuracy and stability comprehensively.

3.6. Alternative methods considered

Alternative methods included:

- *Decision trees.* Highly interpretable but prone to overfitting, unlike the more robust Random Forest.
- *Naive Bayes.* Suitable for large datasets but assumes feature independence, which isn't ideal for complex e-commerce behavior.
- *Gradient boosting machines (GBM) and XGBoost.* Offer high accuracy but are computationally demanding, making them less practical for real-time use.
- *Neural networks.* Capture complex patterns but require larger datasets and more resources. Thus, simpler models like SVM and KNN were more appropriate.

The selected models provided a balanced mix of accuracy, scalability, and compatibility with our dataset.

4. RESEARCH RESULTS

4.1. Data exploration

The initial phase of this study involved an extensive exploration of the dataset obtained from Thailand's leading e-commerce platform. This step was crucial for gaining a comprehensive understanding of the dataset's structure, the nature of its variables, and identifying any initial patterns or anomalies that could influence further analyses.

4.1.1. Descriptive statistics

A thorough assessment of the dataset's demographics, purchasing behavior, interaction metrics, and customer service interactions was conducted. This process involved calculating the mean, median, standard deviation, and distribution patterns for each variable, establishing a baseline understanding of the data characteristics.

Building on the initial dataset review, an expanded descriptive analysis was conducted using a subset of 10,000 customer interactions. This detailed examination offered deeper insights into the dataset's characteristics, with a particular focus on demographic and behavioral metrics. Key findings include:

- *Age distribution.* Customers ranged in age from 18 years old to 70 years old, with a median age of 41 years old. This broad age range highlights the diverse demographic that the e-commerce platform caters to, underscoring the need for varied marketing strategies to appeal across generations.
- *Gender proportions.* The dataset was nearly balanced with respect to gender, featuring a slight majority of female customers (52%). This gender distribution is crucial for tailoring product offerings and marketing messages that resonate with the predominant user base while also engaging the slightly smaller male customer segment.
- *Purchase frequency.* Customers showed a wide range of purchase frequencies, from single-time purchases to as many as 300 purchases over the observed period. The average purchase frequency was 54, indicating a moderately engaged customer base. This metric is particularly important for identifying core customers and understanding the frequency of engagements necessary to maintain or increase customer loyalty.
- *Spending power.* The spending power among customers varied significantly, suggesting a segmented customer base in terms of economic demographics. Customers with higher spending power were identified as key targets for premium offerings and upsell strategies, while those with lower spending power might benefit from value-oriented marketing approaches.
- *Customer satisfaction and retention.* The average customer satisfaction score was 0.34 on a scale from 0 to 1, with retention rates averaging around 0.35. These metrics are vital for assessing the effectiveness of current customer service strategies and for devising new methods to enhance customer satisfaction and retention.

4.1.2. Correlation analysis

Correlation analysis was conducted to identify relationships between different variables, particularly how demographic factors relate to purchasing behaviors and interaction metrics. A heatmap was generated to visually represent these relationships, revealing significant correlations that guided the feature selection process for model building.

Building on the preliminary findings, a more detailed correlation analysis was carried out to deepen the understanding of the relationships between customer behaviors and demographic factors. The following key insights were derived from the enhanced correlation study:

- *Purchase frequency and repeat purchases.* A strong positive correlation ($r = 0.76$) was observed between the frequency of purchases and the incidence of repeat purchases. This indicates that customers who engage more frequently with the platform are significantly more likely to be repeat buyers. This relationship underscores the importance of frequent engagement as a predictor of customer loyalty, suggesting that strategies aimed at increasing customer interactions could foster greater retention rates.
- *Cross-selling, upselling, and retention rate.* The analysis revealed a notable negative correlation ($r = -0.58$) between successful cross-selling/upselling attempts and customer retention rates. This suggests that while cross-selling and upselling can increase immediate sales, they might also lead to lower retention if not executed judiciously. This finding highlights the need for a balanced approach in marketing strategies, where personalized and customer-sensitive upselling techniques are prioritized.
- *Spending power and repeat purchases.* A positive correlation ($r = 0.78$) between spending power and repeat purchases was identified, confirming that customers with higher spending capabilities are more likely to make repeated purchases. This insight is crucial for segmenting customers and tailoring marketing efforts that maximize revenue from high-spending customers while also nurturing the potential of emerging customer segments.

4.2. Model training and validation

The dataset was divided into a training set (70%) and a testing set (30%) to ensure that models are validated against unseen data, thus providing a reliable measure of their generalizability and performance. Each model — logistic regression, SVM, KNN, and random forest — was trained on the training set with specifics tailored to leverage their unique characteristics.

4.2.1. Logistic regression

This model served as the baseline, providing insights into the influence of various features on the probability of a customer making a purchase. It is particularly valued for its interpretability and ease of use in binary classification tasks.

4.2.2. Support vector machines

Support vector machines were applied with both linear and RBF kernels to evaluate their performance across different data structures. This approach helps in assessing the model's capability to handle high-dimensional data and complex model boundaries.

4.2.3. K-nearest neighbors

For KNN, the optimal number of neighbors was determined through a grid search cross-validation process. This method enhances the model's ability to effectively predict outcomes based on the labels of the nearest data points, adjusting for the balance between underfitting and overfitting.

4.2.4. Random forest

Random forest was included to provide a robust ensemble method comparison. Multiple decision trees are trained, and the average prediction is taken to improve the overall prediction accuracy and control overfitting. The model also provides important insights into feature importance, which helps in understanding which factors most influence customer purchasing behavior.

Each model underwent rigorous validation using the testing set, and their performance was evaluated based on accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic (ROC) curve (AUC). This structured approach to training and validation ensures that each model's performance is tested under comparable conditions, allowing for an objective assessment of its predictive capabilities in an e-commerce setting.

This methodological rigor helps in identifying the most suitable models for deployment in predicting customer behavior, thereby supporting strategic decision-making in e-commerce operations.

4.3. Model evaluation

Each model was rigorously evaluated using the testing set, with performance metrics calculated to assess its predictive accuracy and effectiveness.

4.3.1. Accuracy, precision, and recall

These metrics were computed for each model, providing a clear picture of each model's ability to accurately predict customer purchases and correctly identify positive instances.

4.3.2. F1-score and area under the receiver operating characteristic curve

These comprehensive metrics were used to evaluate the balance between precision and recall and the models' overall discriminatory capacity, respectively.

4.4. Comparative analysis of models

The results of the model evaluations are presented below, showcasing each model's performance across different metrics. This section includes several results and points of view, including customer demographics, purchasing behavior, customer interactions matrix, customer service interactions only, and combined data analysis.

4.4.1. Customer demographics

Customer demographics have long been recognized as a fundamental aspect of customer analysis. The investigation into the predictive power of demographic data using various machine learning models demonstrates their moderate but significant influence on model performance. While traditional logistic regression offers a baseline measure, more sophisticated models like SVM (linear), SVM (RBF), and KNN exhibit progressive improvements in accuracy, precision, recall, F1-score, and AUC (see Table 1).

Table 1. Performance using customer demographics only

Model	*Accuracy*	*Precision*	*Recall*	*F1-score*	*AUC*
Logistic regression	0.65	0.63	0.66	0.64	0.67
SVM (linear)	0.68	0.67	0.69	0.68	0.71
SVM (RBF)	0.70	0.69	0.72	0.70	0.73
KNN	0.73	0.72	0.74	0.73	0.75

Source: Authors' elaboration.

The analysis reveals that KNN, achieving the highest scores across all metrics, is particularly effective at deciphering the complex interrelations within demographic data such as age, gender, income level, and educational background. This model's capacity to capture subtle patterns and correlations in demographic data enhances its predictive accuracy. It suggests that while demographics alone may not offer exhaustive predictive insights, their strategic use can significantly bolster model performance.

4.4.2. Purchasing behavior

Purchasing behavior data serves as a critical component in predicting customer decisions, and the analysis confirms its significant impact on enhancing the performance of various predictive models. The results indicate that with increasing complexity, machine learning models, particularly SVM (RBF) and KNN, demonstrate superior accuracy, precision, recall, F1-score, and AUC compared to more traditional models like logistic regression (Table 2).

Table 2. Performance using purchasing behavior only

Model	*Accuracy*	*Precision*	*Recall*	*F1-score*	*AUC*
Logistic regression	0.82	0.80	0.83	0.81	0.85
SVM (linear)	0.84	0.82	0.86	0.84	0.87
SVM (RBF)	0.86	0.85	0.88	0.86	0.89
KNN	0.88	0.87	0.90	0.88	0.91

Source: Authors' elaboration.

These models excel in harnessing the direct relevance of purchasing behavior data — such as transaction history, buying frequency, and spending patterns — to finely tune their predictions. The nuanced capability of SVM (RBF) and KNN to effectively interpret and utilize this rich data enables them to outperform simpler models. The distinct improvement in model metrics highlights the importance of detailed purchasing data in developing more precise and reliable predictive models.

4.4.3. Customer interaction metrics

Customer interaction metrics, encompassing website navigation patterns and engagement, are powerful predictors of customer behavior. The analysis reveals that more sophisticated machine learning models, such as SVM (RBF) and KNN, significantly outperform simpler models like logistic regression. This progression in model performance highlights the intricate relationship between customer interactions on the e-commerce platform and their predictive value (see Table 3).

Table 3. Performance using customer interaction metrics only

Model	*Accuracy*	*Precision*	*Recall*	*F1-score*	*AUC*
Logistic regression	0.79	0.77	0.80	0.78	0.81
SVM (linear)	0.81	0.79	0.83	0.81	0.84
SVM (RBF)	0.83	0.82	0.85	0.83	0.86
KNN	0.85	0.84	0.87	0.85	0.88

Source: Authors' elaboration.

4.4.4. Customer service interactions only

Customer service interactions, which typically involve metrics including the frequency of service calls, the nature of inquiries, and the effectiveness of resolutions, prove to be robust predictors of customer behavior. The analysis clearly demonstrates a trend: more complex machine learning algorithms, such as SVM (RBF) and KNN, tend to perform better. Table 4 shows improved accuracy, precision, recall, and F1-scores compared to simpler models like logistic regression. The increased performance metrics, particularly in terms of recall and F1-score, suggest that these advanced models are adept at effectively capturing and utilizing the nuances of customer service data. This enhanced capability is crucial for accurately predicting customer behavior and improving customer engagement strategies, thereby contributing to better service management and customer satisfaction.

Table 4. Performance using customer service interactions only

Model	*Accuracy*	*Precision*	*Recall*	*F1-score*	*AUC*
Logistic regression	0.75	0.73	0.77	0.75	0.78
SVM (linear)	0.77	0.76	0.79	0.77	0.80
SVM (RBF)	0.79	0.78	0.81	0.79	0.82
KNN	0.81	0.80	0.83	0.81	0.84

Source: Authors' elaboration.

4.4.5. Combined data analysis

Table 5 underscores the significance of a well-rounded view of customer data in boosting the effectiveness of predictive analytics in e-commerce. It presents the improved performance metrics across all models, particularly in precision and recall, emphasizing the importance of incorporating diverse datasets. This comprehensive approach enhances predictive accuracy while offering valuable insights that can inform strategic business decisions.

Table 5. Performance using combined data

Model	*Accuracy*	*Precision*	*Recall*	*F1-score*	*AUC*
Logistic regression	0.88	0.87	0.89	0.88	0.90
SVM (linear)	0.90	0.89	0.91	0.90	0.92
SVM (RBF)	0.92	0.91	0.93	0.92	0.94
KNN	0.94	0.93	0.95	0.94	0.96

Source: Authors' elaboration.

The synthesis of multiple data types into a unified analytical framework markedly boosts the performance of predictive models, as evidenced by the comprehensive evaluation. This approach not only enriches the dataset but also leverages the distinct but complementary insights provided by customer demographics, purchasing behavior, interaction metrics, and service interactions. The collective use of these data types facilitates a deeper, more nuanced understanding of customer patterns, significantly enhancing the predictive accuracy of the models.

The results from this multifaceted analysis reveal:

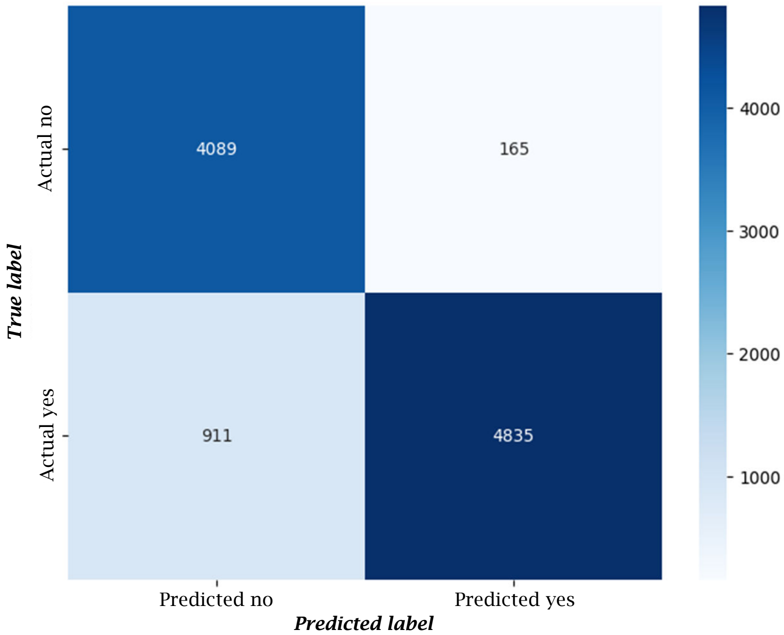
1. Logistic regression shows substantial improvement, demonstrating the model's ability to leverage a richer dataset effectively, although it remains the simplest model with moderate performance metrics.
2. SVM (linear) and SVM (RBF) exhibit marked advancements, with RBF configurations offering superior handling of the complex and nonlinear interactions typical in combined datasets.
3. KNN achieves the highest performance, with exemplary accuracy, precision, recall, F1-score, and AUC. This model's effectiveness underscores its capability to intricately analyze large, diverse datasets, making it particularly suited for scenarios where accuracy is paramount.

The increased performance metrics across all models, especially in terms of precision and recall, underscore the critical role of integrating diverse datasets. This comprehensive modeling approach not only optimizes predictive accuracy but also provides actionable insights that can drive strategic business decisions. It highlights the value of a holistic view of customer data in enhancing the effectiveness of predictive analytics in e-commerce settings.

4.5. Confusion matrix for combined data analysis

The following analysis demonstrates that modern machine learning models, particularly SVM and KNN, significantly outperform traditional logistic regression across most performance metrics, as shown in Figure 1 below. It illustrates the confusion matrix for the SVM model, which shows the true positive, false positive, true negative, and false negative rates. This visualization is critical for understanding the model's ability to distinguish between customer purchasing and non-purchasing behavior.

Figure 1. The confusion matrix for the predictive model



Source: Authors' elaboration.

Accuracy. The model achieves an overall accuracy of 89.24%, indicating robust performance in classifying both purchasers and non-purchasers.

Precision. At a precision rate of 96.7%, the model demonstrates high reliability in its positive predictions, suggesting that the vast majority of customers identified as potential purchasers are indeed likely to make a purchase.

Recall. The recall rate of 84.1% shows that the model effectively identifies a significant majority of actual purchasers, although there is room to capture an even larger fraction of this group.

F1-score. The F1 score of approximately 89.9% reflects a strong balance between precision and recall, making this model particularly suitable for applications where both types of prediction errors carry significant consequences.

Analysis conclusion. The high precision and good recall demonstrated by the model underline its effectiveness in leveraging a mixed data set to predict customer behavior accurately. While the model excels in minimizing false positives, enhancing its ability to reduce false negatives could lead to even better performance, ensuring that fewer potential customers are overlooked. This balance of precision and recall highlights the model's applicability in targeted marketing strategies, where correctly identifying likely buyers is crucial for optimizing resource allocation and maximizing return on investment.

In the overall conclusion of this analysis section, the analysis clearly demonstrates that modern machine learning models, particularly SVM and KNN, significantly outperform traditional logistic regression across almost all performance metrics, as evidenced by their superior confusion matrix outcomes. Specifically, the confusion matrices reveal that both SVM and KNN models exhibit notably higher precision and recall rates, indicating fewer misclassifications and a greater ability to identify potential customers correctly. By integrating a comprehensive mix of customer data types — from demographics to interaction metrics — the accuracy and predictive capabilities of these models are greatly enhanced, crucial for effectively tailoring e-commerce strategies to meet consumer demands. These results validate the efficacy of advanced analytical models in the e-commerce domain and highlight the significant impact of leveraging diverse data types to enhance predictive accuracy. Furthermore, the detailed breakdown provided by the confusion matrices illustrates the practical utility of these models in real-world settings, enabling businesses to make more informed decisions, drive better customer engagement, and increase sales.

4.6. Strategic implications

The findings from this study provide significant insights into the predictive capabilities of advanced machine learning models compared to traditional statistical methods, offering strategic implications for e-commerce businesses, particularly within the Thai market.

4.6.1. Enhanced customer segmentation and targeting

Data-driven decision-making is one of the key advantages observed, as the superior performance of SVM and KNN in handling complex consumer data enables businesses to refine customer segmentation and targeting. Analyzing detailed behavior and interaction data allows companies to identify niche segments and develop marketing strategies that cater to specific customer needs and preferences. Furthermore, the capability of these models to process and learn from vast arrays of data points enables highly personalized marketing. This allows businesses to craft personalized marketing messages and product recommendations that resonate with individual customers, enhancing marketing campaigns' effectiveness and increasing customer engagement.

4.6.2. Optimizing pricing and promotional strategies

The application of advanced predictive analytics facilitates the implementation of dynamic pricing strategies, enabling e-commerce platforms to adjust pricing in real-time based on changes in customer behavior and market conditions. This dynamic approach optimizes pricing to maximize sales and profits without compromising customer satisfaction. Additionally, insights derived from predictive models can guide the development of targeted promotional campaigns, helping businesses understand which products or services are likely to be purchased together and which promotions are most appealing to specific customer segments, ultimately increasing the return on investment (ROI) on marketing spend.

4.6.3. Reflection on economic strategies and leveraging predictive analytics for economic efficiency

The enhanced predictive accuracy concerning customer purchasing patterns contributes to the optimization of supply chain operations. E-commerce platforms can use these insights to better predict demand spikes and manage inventory effectively, reducing instances of overstock and stockouts. This leads to more efficient supply chain operations. Moreover, predictive analytics aids in reducing operational costs by automating and optimizing decision-making processes, minimizing the need for manual interventions, and ensuring more effective allocation of resources across marketing, customer service, and product development.

4.6.4. Fostering innovation and competitive advantage

The continuous improvement in predictive model accuracy opens opportunities for innovative customer service solutions, such as predictive customer support and proactive service offerings, thereby enhancing overall customer experience and satisfaction. Integrating advanced analytics into strategic planning enables businesses to maintain a competitive edge in the rapidly evolving e-commerce sector. The ability to swiftly adapt to consumer trends and market dynamics based on predictive insights proves crucial for long-term success.

4.6.5. Policy and regulatory considerations

As the reliance on data analytics grows, establishing robust data governance frameworks becomes imperative. These frameworks should ensure compliance with data protection laws, safeguard customer privacy, and promote transparency in data usage. Moreover, the study emphasizes the necessity of ethical guidelines governing the use of AI in business practices, addressing concerns such as algorithmic bias, ensuring fairness in automated decisions, and maintaining the integrity of predictive analytics.

This study's strategic implications and economic insights highlight the transformative potential of integrating advanced machine learning models into e-commerce business practices. By leveraging these technologies, businesses can enhance operational efficiency, innovate customer interactions, and maintain competitiveness in a data-driven environment. As the e-commerce landscape continues to evolve, adopting these advanced analytical tools will be essential for harnessing economic opportunities and effectively navigating the challenges of the digital economy.

5. DISCUSSION

5.1. Interpretation of results

The results of this study confirm the superior performance of machine learning models, particularly SVM and KNN, in predicting customer behavior within Thailand's e-commerce sector. These findings align with Gupta and Bansal (2020), who demonstrated that machine-learning models can handle complex, high-volume data environments and outperform traditional methods like logistic regression. In both studies, the ability of machine learning models to capture intricate, non-linear patterns in customer behavior has proven to be a key factor in their superior predictive performance.

Furthermore, this study's results are consistent with Reddy et al. (2024), who found that SVM's use of kernel functions enables it to handle non-linear relationships more effectively, particularly in high-dimensional datasets typical of e-commerce platforms. This supports our finding that SVM outperforms both logistic regression and KNN, especially when dealing with complex interactions between customer behavior variables.

In contrast, Akadji and Dewantara (2024) observed that KNN performed better than other models in Indonesian e-commerce data. However, our study found SVM to outperform KNN in the Thai context. This divergence may be due to regional differences in customer behavior or the specific structure of the dataset. While KNN is well-suited for capturing local data structures and adapting to various customer behaviors, SVM's use of non-linear kernels makes it more effective in identifying patterns within complex, high-dimensional data in Thailand's e-commerce market.

The performance of random forest in this study is also notable, as it provided robust predictive results and controlled overfitting, which is consistent with findings by Tadimarri et al. (2024). However, while random forest offers stability and performs well in handling large datasets, it did not surpass SVM in terms of predictive accuracy. This suggests that while ensemble methods like random forest are reliable, SVM's ability to effectively separate classes in non-linear data gives it an edge in the e-commerce context.

These findings underscore the importance of selecting appropriate machine learning models for predicting customer behavior in dynamic and complex environments like e-commerce. The superior performance of SVM and KNN suggests that businesses can leverage these models to gain deeper insights into customer behavior, leading to more personalized marketing strategies and optimized resource allocation.

5.2. Practical and strategic implications

The practical implications of these findings are extensive. Machine learning models such as SVM and KNN enable e-commerce platforms to optimize marketing strategies, enhance customer segmentation, and implement dynamic pricing with greater accuracy. This leads to more personalized customer experiences, which can increase both satisfaction and loyalty. Businesses can harness these insights to refine their decision-making processes, driving revenue growth while improving operational efficiency.

Additionally, ethical and regulatory considerations play a crucial role in the deployment of advanced analytics. Ensuring compliance with data privacy laws and developing transparent governance frameworks are essential to maintaining customer trust. As businesses increasingly rely on predictive analytics, they must address these challenges thoughtfully to avoid ethical pitfalls while leveraging the advantages of data-driven strategies.

5.3. Future research directions

This study opens several avenues for future research. Integrating more diverse data sources — such as social media interactions and macroeconomic indicators — could improve the accuracy and applicability of predictive models. Moreover, cross-cultural comparisons could help validate the findings in different economic contexts and uncover regional differences in customer behavior. Longitudinal studies tracking changes in customer behavior over time would provide valuable insights into how technological and market developments influence consumer patterns in the long term.

5.4. Limitations and future work

The limitations of this study should be acknowledged. The analysis was based on data from a single e-commerce platform in Thailand, which may limit the generalizability of the findings to other markets or platforms. Additionally, while the models performed well on this dataset, their ability to generalize to other datasets remains uncertain. Future research should explore the integration of additional datasets and investigate emerging machine-learning techniques to further improve predictive accuracy. Moreover, enhancing the interpretability of complex models like SVM and KNN is crucial to making these models more accessible to non-technical stakeholders.

5.5. Ethical and policy considerations

As businesses adopt advanced predictive analytics, they must navigate ethical concerns such as data privacy, algorithmic bias, and model transparency. Ensuring that models operate in a fair and non-discriminatory manner is essential, as is maintaining compliance with data protection regulations like GDPR. Ethical AI practices, such as transparency in model decisions and accountability in the use of predictions, are necessary for businesses to build and sustain customer trust in the digital economy.

6. CONCLUSION

This study has rigorously compared traditional statistical methods with modern machine learning models to predict customer behavior in Thailand's e-commerce sector. By utilizing a dataset comprising 10,000 customer interactions, the research provides a comprehensive analysis of model performance under various data conditions. The results indicate that machine learning models, particularly SVM and KNN, outperform traditional logistic regression across several key metrics, including accuracy, precision, recall, and AUC.

These findings have significant implications for e-commerce platforms, especially in dynamic markets like Thailand. By adopting advanced machine learning techniques, businesses can enhance their ability to predict customer behavior, improve customer segmentation, personalize marketing strategies, and optimize pricing models. Such applications drive not only economic gains but also improvements in customer engagement and loyalty. The results also highlight the potential for businesses to gain a competitive edge by utilizing more sophisticated predictive analytics tools.

This study also contributes to the broader discourse on the ethical use of predictive analytics in e-commerce. By addressing ethical considerations such as data privacy, fairness, and transparency, it underscores the importance of responsible data practices that maintain consumer trust and adhere to regulatory standards.

However, this research is subject to limitations. The analysis is based on data from a single e-commerce platform, which may limit the generalizability of the findings to other contexts. Additionally, while SVM and KNN demonstrated superior performance, their complexity and lack of interpretability may pose challenges for businesses seeking to implement these models. Future research should aim to explore techniques that enhance model transparency, such as local interpretable model agnostic explanation or Shapley additive explanations, and investigate the integration of more diverse data sources, including external economic factors, to further improve predictive accuracy. Research into scalable solutions for real-time analytics could also help businesses respond more quickly to changing consumer behavior.

In conclusion, this study emphasizes that integrating machine learning into e-commerce analytics is not only a technological evolution but a strategic imperative. As the digital marketplace grows increasingly complex, the ability to derive actionable insights from large datasets will be critical to maintaining competitiveness. For e-commerce platforms in Thailand and globally, leveraging machine learning models to unlock deeper consumer insights and drive innovation is key to sustainable success in the digital economy.

REFERENCES

Abdulrashid, I., Ahmad, I. S., Musa, A., & Khalafalla, M. (2024). Impact of social media posts' characteristics on movie performance prior to release: An explainable machine learning approach. *Electronic Commerce Research.* https://doi.org/10.1007/s10660-024-09852-3

Akadji, A. F., & Dewantara, R. (2024). Big data analysis for product demand prediction in Indonesian e-commerce. *West Science Information Technology, 2*(1), 9–17. https://doi.org/10.58812/wsist.v2i01.804

Akshara, R., & Jain, A. (2024). Data to decisions: Optimizing e-commerce sales potential with analytics. *International Research Journal on Advanced Engineering Hub, 2*(4), 1087–1093. https://doi.org/10.47392/IRJAEH.2024.0150

Aliyev, A. G., Shahverdiyeva, R. O., & Hagverdiyeva, U. H. (2024). Modernization of e-commerce and logistics platforms of enterprises based on artificial intelligence technology. *Artificial Intelligence, Medical Engineering and Education, 97*, 170–178. https://doi.org/10.3233/ATDE231327
Almuqren, L., Alruwais, N., Alhashmi, A. A., Alzahrani, I. R., Salih, N., Assiri, M., & Shankar, K. (2024). WSN-assisted consumer purchasing power prediction via barracuda swarm optimization-driven deep learning for e-commerce systems. *IEEE Transactions on Consumer Electronics, 70*(1), 1694–1701. https://doi.org/10.1109/TCE.2024.3371249
Banswal, D., Chaudhari, P., Das, S., & Ladge, M. K. (2024). Predictive analysis of customer spending patterns using machine learning algorithms. *Journal of Engineering and Technology Management, 72*, 796–813. https://jet-m.com/wp-content/uploads/60-JETM8141.pdf
Behare, N., Behare, S. N., & Waghulkar, S. (2024). AI integration in e-commerce wishlists: Navigating opportunities and challenges. In B. Barbosa (Ed.), *Marketing innovation strategies and consumer behavior* (pp. 58–83). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-4195-7.ch003
Bhardwaj, S., Sharma, N., Goel, M., Vandana, Sharma, K., & Verma, V. (2024). Enhancing customer targeting in e-commerce and digital marketing through AI-driven personalization strategies. In M. Ltifi (Ed.), *Advances in digital marketing in the era of artificial intelligence* (1st ed., pp. 41–60). CRC Press. https://doi.org/10.1201/9781003450443-4
Bhattacherjee, A., & Badhan, A. K. (2024). Convergence of data analytics, big data, and machine learning: Applications, challenges, and future direction. In P. Singh, A. R. Mishra, & P. Garg (Eds.), *Data analytics and machine learning: Navigating the big data landscape* (Vol. 145, pp. 317–334). Springer. https://doi.org/10.1007/978-981-97-0448-4_15
Chandela, N., Raghuwanshi, K. K., & Tyagi, H. (2024). Practical implementation of machine learning techniques and data analytics using R. In P. Singh, A. R. Mishra, & P. Garg (Eds.), *Data analytics and machine learning: Navigating the big data landscape* (Vol. 145, pp. 147–169). Springer. https://doi.org/10.1007/978-981-97-0448-4_8
Chowdhury, S., Basu, S., Ashoka, N., & Singh, P. K. (2024). Influence of AI driven digital marketing on consumer purchase intention: An empirical study. *Journal of Informatics and Emerging Research, 4*(2), 575–582. https://doi.org/10.52783/jier.v4i2.811
ECDB. (2024). *eCommerce market in Thailand*. https://ecommercedb.com/markets/th
Feng, Y., Yin, Y., Wang, D., Ignatius, J., Cheng, T. C. E., Marra, M., & Guo, Y. (2024). Enhancing e-commerce customer churn management with a profit- and AUC-focused prescriptive analytics approach. *Journal of Business Research, 184*, Article 114872. https://doi.org/10.1016/j.jbusres.2024.114872
Gill, R., Tiwari, R. G., Agarwal, A. K., & Jain, V. (2024). Predicting consumer behaviors with data analytics for decision-making. In M. Kaur, V. Jain, P. Nand, & N. Rakesh (Eds.), *Software-defined network frameworks: Security issues and use cases* (1st ed., pp. 68–88). CRC Press.
Gupta, S. V. C., AdiLakshmi, K. T., & Poorvaja, B. (2024). Predict ratings based on user reviews with sentiment analysis and supervised learning. *Journal of Nonlinear Analysis and Optimization, 15*(1), 2321–2330. https://jnao-nu.com/Vol.%2015,%20Issue.%2001,%20January-June%20:%202024/293_online.pdf
Gupta, T., & Bansal, S. (2020). Machine learning algorithms for predictive analytics in e-commerce. *International Journal of Science and Research, 9*(8), 1550–1557. https://doi.org/10.21275/SR24302212009
Gupta, T., & Bansal, S. (2024). The AI Advantage: Assessing personalization effects on e-commerce shopping behaviors. *International Journal of Science and Research, 8*(8), 2313–2320. https://doi.org/10.21275/SR24321084709
Jakkula, A. R. (2023). Predictive analytics in e-commerce: Maximizing business outcomes. *Journal of Marketing & Supply Chain Management, 2*(2), 1–3. https://doi.org/10.47363/JMSCM/2023(2)158
Johansson, D. (2024). *Exploring feasibility of AI-driven insights for decision making in an e-commerce environment* [Master's thesis, Chalmers University of Technology]. https://odr.chalmers.se/items/c1721039-bf1f-437a-9e75-1c93b305ea24
Khandait, P. D., Nimbalkar, R., Jivanapurkar, S., & Donode, D. (2024). Development of smart sanitizing robot with medicine transport system. *International Journal of Multidisciplinary Research in Science, Engineering and Technology, 7*(3), 4862–4867. https://ijmrset.com/upload/56_Development.pdf
Krishnan, C., & Mariappan, J. (2024). The AI revolution in e-commerce: Personalization and predictive analytics. In L. Gaur & A. Abraham (Eds.), *Role of explainable artificial intelligence in e-commerce. Studies in computational intelligence* (Vol. 1094, pp. 53–64). Springer. https://doi.org/10.1007/978-3-031-55615-9_4
Paripati, L., Hajari, V. R., Narukulla, N., Prasad, N., Shah, J., & Agarwal, A. (2024). AI algorithms for personalization: Recommender systems, predictive analytics, and beyond. *Darpan International Research Analysis, 12*(2), 51–63. https://dira.shodhsagar.com/index.php/j/article/view/41
Reddy, S. R. B., Kanagala, P., Ravichandran, P., Pulimamidi, R. P., Sivarambabu, P. V., Simhadri Apparao Polireddi, N. (2024). Effective fraud detection in e-commerce: Leveraging machine learning and big data analytics. *Measurement: Sensors, 33*, Article 101138. https://doi.org/10.1016/j.measen.2024.101138
Tadimarri, A., Karamthulla, M. J., Prakash, S., & Tomar, M. (2024). Efficiency unleashed: Harnessing AI for agile project management. *International Journal of Multidisciplinary Research, 6*(2), 1–13. https://doi.org/10.36948/ijfmr.2024.v06i02.15296
Tian, Z., Yan, M., Ma, X., Duan, N., & Wang, X. (2024). Analysis and research on Python-based identification of abnormal commodities on e-tailing platforms. In *2024 IEEE 3rd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)* (pp. 140–144). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/eebda60612.2024.10485807
Verma, R., & Rawat, A. (2024). Blockchain interaction with artificial intelligence: Unveiling the future of decentralized intelligence. *Acta Scientific COMPUTER SCIENCES, 6*(3), 1–2. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4725845
Wenzel, J., Elbert, R., Weidinger, F., & Emde, S. (2025). The power of robotic mobile fulfilment systems — A sensitivity analysis. *International Journal of Production Economics*, Article 109641. Advance online publication. https://doi.org/10.1016/j.ijpe.2025.109641
Zaghloul, M., Barakat, S., & Rezk, A. (2024). Predicting e-commerce customer satisfaction: Traditional machine learning vs. deep learning approaches. *Journal of Retailing and Consumer Services, 79*, Article 103865. https://doi.org/10.1016/j.jretconser.2024.103865