ADBA
Information Technology in Economics and Business
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8943
RESEARCH ARTICLE
Vol.1 / No.1 / 2024 / pp.1-6
https://doi.org/10.69882/adba.iteb.2024071

Predicting Customer Purchase Behavior Using Machine Learning Models

Emre Deniz *,[1] **and Semanur Çökekoğlu Bülbül** *,[2]
*Department of Computer Engineering, Hitit University, 19030, Corum, Turkiye.

ABSTRACT In this study, we aim to predict customer purchase behavior using various machine learning models to better understand customer tendencies and enhance marketing strategies. We use a dataset containing demographic and behavioral data, including age, gender, annual income, number of purchases, product category, time spent on the website, loyalty program membership, and discounts availed. Our analysis involves data preprocessing, exploratory data analysis (EDA), and feature engineering. We then train and evaluate six different machine learning models: Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost. The models are assessed using metrics such as accuracy, precision, recall, F1-score, and ROC AUC. Results indicate that ensemble models, specifically Random Forest and Gradient Boosting, outperform the other models in terms of accuracy and ROC AUC. The study concludes that ensemble models are highly effective for predicting customer purchase behavior, providing valuable insights for businesses to tailor their marketing efforts. Future research could explore additional features, more advanced models, and real-time prediction capabilities.

KEYWORDS
Predictive analytics
Customer purchase behavior
Machine learning
Business intelligence

INTRODUCTION

Predicting customer purchase behavior is a critical aspect of modern business strategies, enabling companies to optimize their marketing efforts, enhance customer satisfaction, and ultimately increase profitability. With the rapid advancement of technology and the proliferation of data, businesses now have access to a wealth of information about their customers. This information, when analyzed effectively, can provide deep insights into customer preferences and behaviors, allowing for more personalized and targeted marketing approaches.

In this study, we leverage various machine learning techniques to predict whether a customer will make a purchase based on their demographic and behavioral data. The dataset used in this study includes attributes such as age, gender, annual income, number of purchases, product category, time spent on the website, loyalty program membership, and discounts availed. These features provide a comprehensive view of customer profiles and their interaction with the business.

Machine learning has emerged as a powerful tool in predictive analytics, offering the ability to identify complex patterns and relationships within data that traditional statistical methods may overlook. By applying machine learning models to our dataset, we aim to develop a robust predictive framework that can accurately forecast purchase behavior.

We employ six different machine learning models in this study: Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost. These models are chosen for their diverse approaches and strengths in handling various types of data and relationships. Logistic Regression is known for its simplicity and interpretability, while ensemble methods like Random Forest and Gradient Boosting are renowned for their high performance and ability to capture complex interactions. SVM and KNN offer robust alternatives for classification tasks, and XGBoost is a highly efficient implementation of gradient boosting that has gained popularity for its speed and performance.

The methodology of this study involves several key steps: data preprocessing, exploratory data analysis (EDA), feature engineering, model training, and evaluation. Data preprocessing ensures that the data is clean and suitable for analysis, while EDA helps in understanding the underlying patterns and relationships within the data. Feature engineering is crucial for enhancing the predictive power of the models by creating relevant features. The models are then trained on the processed data and evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC AUC to determine their effectiveness.

Manuscript received: 14 July 2024,
Revised: 27 July 2024,
Accepted: 28 July 2024.

[1] emredeniz@hitit.edu.tr (**Corresponding author**)
[2] semanurcokekoglu@hitit.edu.tr

By comparing the performance of these models, we aim to identify the most suitable approach for predicting customer purchase behavior. The insights gained from this analysis can help businesses make data-driven decisions, improve customer targeting, and design more effective marketing strategies. Additionally, this study contributes to the growing body of research on the application of machine learning in customer analytics, providing a reference for future studies and practical implementations.

Recent advancements in customer behavior prediction have demonstrated the effectiveness of multi-objective evolutionary algorithms (MOEAs). By leveraging techniques such as Word2Vec for feature extraction and boosted ant colony optimization (BACO) for feature selection, the MOEA approach significantly enhances prediction accuracy and efficiency compared to traditional machine learning methods (Krishnamoorthy *et al.* 2023). This innovative approach highlights the growing trend of combining advanced algorithms and feature engineering techniques to improve predictive models.

Studies on customer purchase intentions have shown significant progress with the use of clickstream data. The MBT-POP model, which incorporates multi-behavioral trendiness and product popularity, has achieved notable improvements in predictive performance, as evidenced by an impressive F1 score of 0.9031 (Rana *et al.* 2024). This model exemplifies the utility of integrating behavioral data to refine prediction accuracy in customer purchase behaviors.

In the banking sector, machine learning models such as Logistic Regression (LR) and Naive Bayes (NB) have been effectively utilized to predict customer churn. These models analyze various customer data points like age, location, gender, and credit card information to identify customers most likely to leave. Findings indicate that the Naive Bayes model surpasses Logistic Regression in predictive accuracy (Wen *et al.* 2023). This demonstrates the continued relevance of traditional machine learning models in specific domains like banking.

The integration of RFID technology with machine learning models has also proven effective in analyzing customer shopping behavior in physical stores. By leveraging received signal strength (RSS) data from RFID tags, time-domain features were extracted and used for classification, significantly enhancing the prediction of customer activities such as browsing and product interaction. This approach demonstrated high accuracy, precision, recall, and F1-score, providing valuable insights for product placement and customer recommendations (Alfian *et al.* 2023).

Federated learning has emerged as a promising technique for predicting customer behavior while preserving data privacy. Utilizing differential privacy and homomorphic encryption, federated learning ensures that data privacy is maintained without compromising the accuracy of the predictive models. This method enables multiple entities to collaboratively train models without sharing their data, addressing privacy concerns effectively (Thabet *et al.* 2023). This approach is particularly relevant in today's data-driven world where privacy concerns are paramount.

Advanced machine learning techniques have significantly enhanced the understanding of customer purchase patterns in e-commerce. Models such as Random Forest, Decision Tree, K-Nearest Neighbors, Neural Networks, and Support Vector Machine are employed to analyze vast amounts of transaction data. These models improve customer experiences, streamline inventory management, and optimize marketing strategies, as validated by recent research (Kumar *et al.* 2023). This highlights the multifaceted applications of machine learning in e-commerce.

Similarly, machine learning models have proven highly effective in predicting customer purchase intent. Various models including Random Forest, XGBoost, Decision Trees, K-Nearest Neighbors (KNN), and Logistic Regression have been applied to marketing data, significantly enhancing prediction accuracy and optimizing marketing efforts. This approach provides valuable insights into customer behavior, improving the overall efficiency of marketing strategies (Krishna *et al.* 2023).

In the context of online grocery shopping, machine learning models such as Artificial Neural Networks (ANN), Decision Trees (DT), Recurrent Neural Networks (RNN), and Naive Bayes (NB) have shown significant promise. These models estimate the kind and timing of client transactions with high accuracy rates. For instance, ANNs identified intricate patterns with an accuracy of 97.6%, while Decision Trees achieved precision and accuracy rates of 97.3% and 97.8%, respectively (Chaudhary *et al.* 2024). These insights enable businesses to better understand customer behavior and optimize targeted marketing efforts.

Big data analytics and machine learning have significantly improved the analysis of customer behavior for digital marketing. Employing various machine learning algorithms and the ML pipeline, businesses can forecast customer churn, identify high-propensity prospects, determine optimal communication channels, and enhance customer experiences through sentiment analysis. These techniques effectively analyze large datasets, providing real-time insights and enabling data-driven decisions that enhance customer engagement and satisfaction (Deniz *et al.* 2022).

Lastly, analyzing e-commerce customer reviews through multi-label classification provides in-depth insights into customer opinions beyond simple sentiment analysis. Techniques such as Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, GloVe, and Bidirectional Encoder Representations from Transformers (BERT) have been employed to extract meaningful features from text. Using algorithms like Binary Relevance, Random Forest, and XGBoost, researchers have achieved high accuracy in classifying multi-label customer reviews, highlighting the diverse opinions customers hold about products (Agarwal *et al.* 2022). This approach underscores the importance of nuanced text analysis in understanding customer feedback.

In conclusion, the integration of advanced machine learning techniques, big data analytics, and privacy-preserving methods has significantly advanced the field of customer behavior prediction. These approaches not only improve predictive accuracy but also provide actionable insights that enhance customer experiences and optimize business strategies. The continuous evolution of these technologies promises further advancements in understanding and predicting customer behavior.

MATERIALS AND METHODS

The dataset used in this study is a comprehensive collection of customer demographic and behavioral information, which provides a robust foundation for predicting purchase behavior. The dataset contains the following key attributes:

Age: This attribute represents the age of the customer. Age is a significant demographic factor that can influence purchasing behavior, as different age groups often exhibit distinct preferences and spending patterns.

Gender: Gender is another crucial demographic variable, coded as 1 for male and 0 for female. Understanding gender-specific preferences can help tailor marketing strategies and product offerings.

Annual Income: This attribute captures the annual income of the customer. Income level is a vital indicator of purchasing power

and can significantly affect purchasing decisions and frequency.

Number of Purchases: This attribute indicates the total number of purchases made by the customer. It is a direct measure of customer engagement and loyalty, reflecting how often a customer interacts with the business through purchases.

Product Category: This categorical variable represents the category of the product purchased. Different product categories can have varying levels of appeal to different customer segments, influencing purchase behavior.

Time Spent on Website: This attribute measures the total time a customer spends on the business's website. Time spent can indicate the level of interest and engagement a customer has with the online platform, potentially correlating with the likelihood of making a purchase.

Loyalty Program: This binary attribute indicates whether the customer is a member of the loyalty program (1 for yes, 0 for no). Loyalty programs are designed to enhance customer retention and encourage repeat purchases by offering rewards and incentives.

Discounts Availed: This attribute captures the number of discounts the customer has utilized. Discounts can be a significant motivator for purchases, and customers who frequently avail discounts may exhibit different purchasing behaviors compared to those who do not.

Purchase Status: This is the target variable in our analysis, indicating whether the customer made a purchase (1 for yes, 0 for no). This binary variable is what our machine learning models aim to predict.

The dataset is rich in both demographic and behavioral data, allowing for a multifaceted analysis of customer purchase behavior. Each attribute provides valuable insights into different aspects of customer interactions and preferences.

Data Preprocessing

Before feeding the data into the machine learning models, several preprocessing steps are undertaken to ensure data quality and compatibility with the models. These steps include:

Handling Missing Values: Ensuring that there are no missing values in the dataset, as missing data can lead to inaccuracies in the model training process.

Encoding Categorical Variables: Converting categorical variables, such as Gender and Product Category, into numerical representations using techniques like one-hot encoding. This step is crucial for enabling the models to process these variables effectively.

Feature Scaling: Normalizing numerical features such as Age, Annual Income, Number of Purchases, and Time Spent on Website to a standard scale. This is essential to prevent features with larger numerical ranges from disproportionately influencing the model.

Exploratory Data Analysis (EDA) EDA is performed to understand the underlying structure and distribution of the data. This involves generating summary statistics, visualizing data distributions through histograms and box plots, and identifying correlations between features using a correlation heatmap. EDA helps in uncovering patterns and relationships that inform feature engineering and model selection.

Figure 1 shows the histograms and box plots for the numerical features in the dataset. These visualizations help us understand the distribution and potential outliers in the data.

Figure 2 displays the count plots for the categorical features, illustrating the distribution of gender, product category, loyalty program membership, and purchase status among customers.

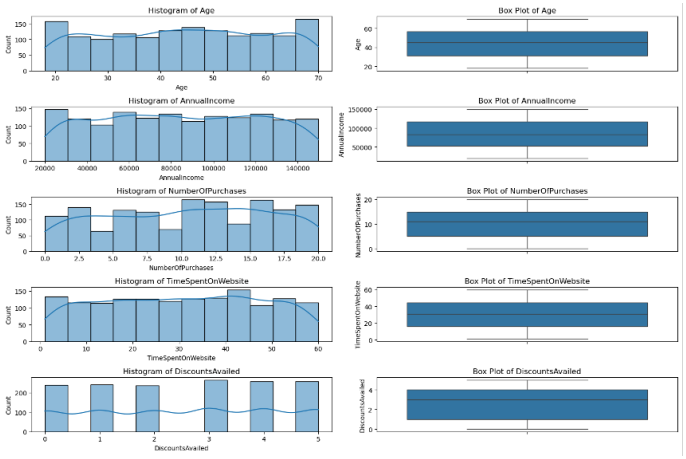

Figure 1 Histograms and boxplots of numerical features

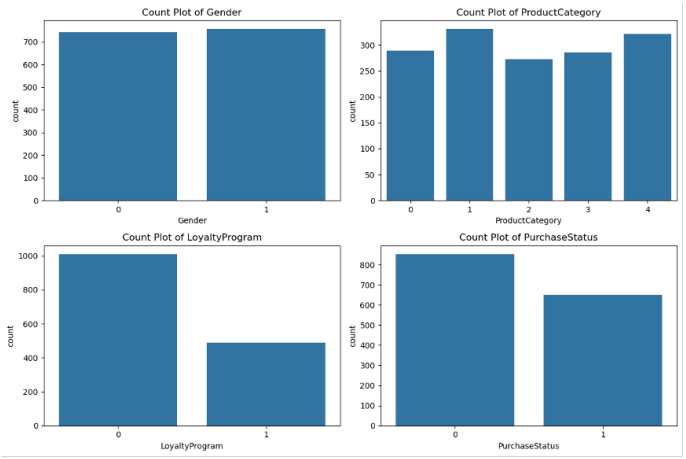

Figure 2 Histograms and boxplots of numerical features

The correlation heatmap in Figure 3 highlights the relationships between different features in the dataset. This visualization helps identify which features are strongly correlated and can provide insights into feature engineering.

Feature Engineering Feature engineering involves creating new features or transforming existing ones to enhance the predictive power of the models. In this study, we derive two additional features: 1. Spender Segment: Categorizing customers into 'High Spender', 'Medium Spender', and 'Low Spender' based on the number of purchases. This segmentation helps in understanding different spending behaviors (Figure 4).

2. Age Group: Grouping customers into age brackets such as '18-30', '31-45', '46-60', and '61-70'. Age grouping provides insights into age-specific purchasing trends (Figure 5).

Model Training and Evaluation

We employed six different machine learning models for predicting customer purchase behavior:

Logistic Regression Random Forest Gradient Boosting Support Vector Machine (SVM) K-Nearest Neighbors (KNN) XGBoost

Each model was trained using the preprocessed dataset, and their performance was evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC AUC.

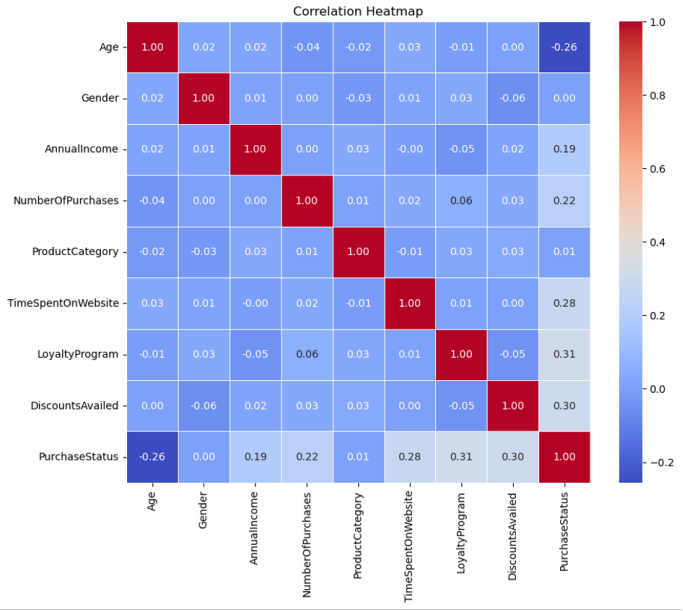

Figure 3 Correlation Heatmap

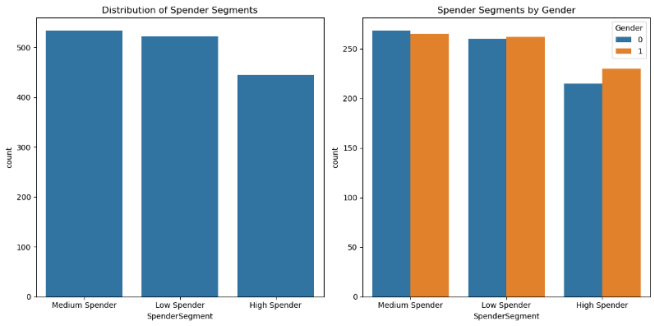

Figure 4 Distribution of Spender Segments and Spender Segments by Gender

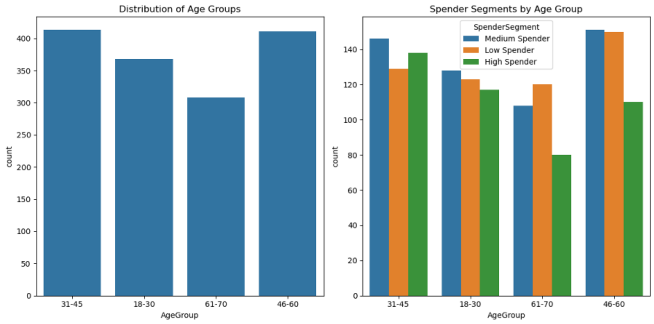

Figure 5 Distribution of Age Groups and Spender Segments by Age Group

Model Training The models were trained on 70% of the data (training set) and tested on the remaining 30% (test set). Feature scaling was applied to ensure that numerical features were on a comparable scale. For categorical variables, one-hot encoding was used.

Model Evaluation The performance of each model was evaluated based on:

Accuracy: The proportion of correctly predicted instances out of the total instances. Precision: The proportion of true positive predictions out of the total positive predictions. Recall: The proportion of true positive predictions out of the actual positive instances. F1-score: The harmonic mean of precision and recall, providing a single measure of a model's performance. ROC AUC: The area under the receiver operating characteristic curve, indicating the model's ability to distinguish between classes.

RESULTS

The performance metrics for each model are summarized in Table 1. The table presents the performance metrics of six different machine learning models applied to predict customer purchase behavior. The models evaluated include Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost. The performance metrics used to assess these models are Accuracy, Precision, Recall, F1-Score, and ROC AUC.

Logistic Regression achieved an accuracy of 82.2%, with a precision of 0.84, a recall of 0.73, an F1-score of 0.78, and an ROC AUC of 0.81. This model, known for its simplicity and interpretability, shows moderate performance across the metrics. Random Forest, an ensemble learning method, shows superior performance with an accuracy of 94.4%, a precision of 0.97, a recall of 0.90, an F1-score of 0.93, and an ROC AUC of 0.94. These results indicate that Random Forest can effectively capture complex interactions within the data, leading to high predictive accuracy.

Gradient Boosting also demonstrates high performance, with an accuracy of 94.2%, a precision of 0.96, a recall of 0.90, an F1-score of 0.93, and an ROC AUC of 0.94. This ensemble technique boosts the performance by combining weak learners to create a strong predictive model, closely matching the performance of Random Forest. Support Vector Machine (SVM) achieved an accuracy of 92.0%, a precision of 0.95, a recall of 0.85, an F1-score of 0.90, and an ROC AUC of 0.92. SVM is effective in high-dimensional spaces and shows robust performance in this study.

K-Nearest Neighbors (KNN) reported an accuracy of 89.0%, a precision of 0.89, a recall of 0.88, an F1-score of 0.88, and an ROC AUC of 0.89. While KNN is simple and easy to interpret, it shows slightly lower performance metrics compared to the ensemble models. XGBoost, another powerful ensemble method, outperformed all other models with an accuracy of 94.5%, a precision of 0.97, a recall of 0.91, an F1-score of 0.94, and an ROC AUC of 0.94. XGBoost's efficiency and scalability make it an excellent choice for predictive modeling in large datasets.

In summary, the ensemble models, particularly Random Forest, Gradient Boosting, and XGBoost, exhibit superior performance in predicting customer purchase behavior, as reflected in their high accuracy, precision, recall, F1-score, and ROC AUC values. These results highlight the effectiveness of ensemble learning methods in capturing complex patterns and interactions within customer data, thereby providing valuable insights for enhancing marketing strategies.

TThe ROC AUC curves for all models are shown in Figure 6. The Random Forest and XGBoost models achieve the highest ROC AUC scores, indicating their superior performance in distinguishing between customers who will and will not make a purchase.

Table 1 Results of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	0.82	0.84	0.73	0.78	0.81
Random Forest	0.94	0.97	0.90	0.93	0.94
Gradient Boosting	0.94	0.96	0.90	0.93	0.94
Support Vector Machine	0.92	0.95	0.85	0.90	0.92
K-Nearest Neighbors	0.89	0.89	0.88	0.88	0.89
XGBoost	0.94	0.97	0.91	0.94	0.94

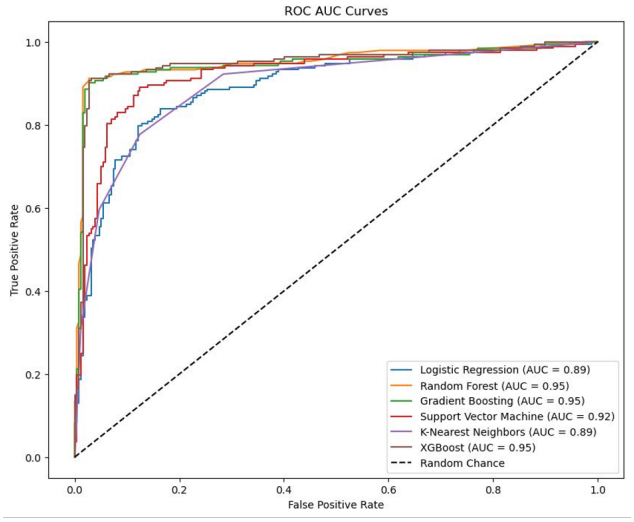

Figure 6 Distribution of Age Groups and Spender Segments by Age Group

DISCUSSION

Our analysis demonstrates that ensemble models such as Random Forest and Gradient Boosting are highly effective in predicting customer purchase behavior. Logistic Regression, while simpler and more interpretable, provides lower accuracy and ROC AUC scores compared to ensemble models. Support Vector Machine and K-Nearest Neighbors offer robust alternatives, with KNN being particularly effective for datasets with a clear neighborhood structure. XGBoost, known for its efficiency and performance, also delivers excellent results.

The insights gained from this study can help businesses make data-driven decisions, improve customer targeting, and design more effective marketing strategies. Additionally, this study contributes to the growing body of research on the application of machine learning in customer analytics, providing a reference for future studies and practical implementations.

CONCLUSION

Predicting customer purchase behavior using machine learning can significantly enhance marketing strategies and customer understanding. Our study shows that ensemble models, particularly Random Forest and Gradient Boosting, provide the best performance. Future work could explore additional features, more advanced models, and real-time prediction capabilities.

Availability of data and material

Not applicable.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

LITERATURE CITED

Agarwal, V., S. Taware, S. A. Yadav, D. Gangodkar, A. L. N. Rao, *et al.*, 2022 Customer - churn prediction using machine learning. In *2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, pp. 893–899, IEEE.

Alfian, G., M. Q. H. Octava, F. M. Hilmy, R. A. Nurhaliza, Y. M. Saputra, *et al.*, 2023 Customer shopping behavior analysis using rfid and machine learning models. Information **14**: 551.

Chaudhary, R., S. Chaudhary, A. Singh, A. Bhanu, S. Chandela, *et al.*, 2024 Artificial intelligence-based digital marketing for discovering shopping possibilities and enhancing customer experience. In *2024 2nd International Conference on Disruptive Technologies (ICDT)*, pp. 413–417, IEEE.

Deniz, E., H. Erbay, and M. Coşar, 2022 Multi-label classification of e-commerce customer reviews via machine learning. Axioms **11**: 436.

Krishna, R. D., M. Mahadev, S. Hariprasad, S. Abhishek, and T. Anjali, 2023 Cultivating customer purchase intent: Leveraging machine learning for precise predictions. In *2023 12th International Conference on System Modeling & Advancement in Research Trends (SMART)*, pp. 374–379, IEEE.

Krishnamoorthy, R., K. P. Kaliyamurthie, B. S. H. S. Ahamed, N. Harathi, and R. S. Selvan, 2023 Multi objective evaluator model development for analyzing customer behavior. In *2023*

3rd International Conference on Advancement in Electronics & Communication Engineering (AECE), pp. 640–645, IEEE.

Kumar, M. M., A. S. Venkat, M. V. N. Balaji, C. H. N. Kumar, S. Srithar, *et al.*, 2023 Driving e-commerce success with advanced machine learning: Customer purchase pattern insights. In *2023 International Conference on Sustainable Communication Networks and Application (ICSCNA)*, pp. 1196–1203, IEEE.

Rana, M. N. U., S. A. Al Shiam, S. A. Shochona, M. R. Islam, M. Asrafuzzaman, *et al.*, 2024 Revolutionizing banking decision-making: A deep learning approach to predicting customer behavior. Journal of Business and Management Studies **6**: 21–27.

Thabet, M., M. Messaadia, and M. Kumar, 2023 Distributed machine learning for predicting customer behavior while preserving privacy. In *2023 International Conference on Decision Aid Sciences and Applications (DASA)*, pp. 613–617, IEEE.

Wen, Z., W. Lin, and H. Liu, 2023 Machine-learning-based approach for anonymous online customer purchase intentions using clickstream data. Systems **11**: 255.

How to cite this article: Deniz, E. and Bülbül, S. Ç. Predicting Customer Purchase Behavior Using Machine Learning Models. *Information Technology in Economics and Business*, 1(1), 1-6, 2024.