Int. J. Communications, Network and System Sciences, 2024, 17, 81-103
https://www.scirp.org/journal/ijcns
ISSN Online: 1913-3723
ISSN Print: 1913-3715

Predicting Purchasing Behavior on E-Commerce Platforms: A Regression Model Approach for Understanding User Features that Lead to Purchasing

Abraham Jallah Balyemah[1], Sonkarlay J. Y. Weamie[1], Jiang Bin[1], Karmue Vasco Jarnda[2], Felix Jwakdak Joshua[1]

[1]College of Computer Science and Engineering, Hunan University, Changsha, China
[2]Department of Health Inspection and Quarantine, Xiangya School of Public Health, Central South University, Changsha, China
Email: skweamie@hnu.edu.cn

How to cite this paper: Balyemah, A.J., Weamie, S.J.Y., Bin, J., Jarnda, K.V. and Joshua, F.J. (2024) Predicting Purchasing Behavior on E-Commerce Platforms: A Regression Model Approach for Understanding User Features that Lead to Purchasing. *Int. J. Communications, Network and System Sciences*, **17**, 81-103.
https://doi.org/10.4236/ijcns.2024.176006

Received: March 25, 2024
Accepted: June 25, 2024
Published: June 28, 2024

Abstract

This research introduces a novel approach to improve and optimize the predictive capacity of consumer purchase behaviors on e-commerce platforms. This study presented an introduction to the fundamental concepts of the logistic regression algorithm. In addition, it analyzed user data obtained from an e-commerce platform. The original data were preprocessed, and a consumer purchase prediction model was developed for the e-commerce platform using the logistic regression method. The comparison study used the classic random forest approach, further enhanced by including the K-fold cross-validation method. Evaluation of the accuracy of the model's classification was conducted using performance indicators that included the accuracy rate, the precision rate, the recall rate, and the F1 score. A visual examination determined the significance of the findings. The findings suggest that employing the logistic regression algorithm to forecast customer purchase behaviors on e-commerce platforms can improve the efficacy of the approach and yield more accurate predictions. This study serves as a valuable resource for improving the precision of forecasting customers' purchase behaviors on e-commerce platforms. It has significant practical implications for optimizing the operational efficiency of e-commerce platforms.

Keywords

E-Commerce Platform, Purchasing Behavior Prediction, Logistic Regression Algorithm

1. Introduction

Consumer behaviors in the realm of e-Commerce are shaped by a multitude of factors, ranging from personal preferences and demographic attributes to larger-scale economic patterns. With the emergence of big data and sophisticated analytics, organizations now have an unparalleled opportunity to extract valuable information from extensive databases, allowing them to get deep insights into customer behavior [1]. The regression model technique is a leading strategy in this research, providing a systematic and statistically reliable way to find crucial elements that impact purchase decisions. The use of regression models to forecast e-Commerce purchasing patterns is based on the notion of comprehending the associations between different independent factors (also known as "features") and the probability of making a purchase (the dependent variable) [2]. The features encompass a wide variety of information, including user-specific data such as age, gender, and browsing history, as well as product-related elements such as price, reviews, and availability. Through the analysis of these characteristics, regression models can offer important forecasts regarding which clients are more inclined to make a purchase and the specific circumstances under which this is likely to occur.

Understanding customer behavior is crucial in the highly competitive e-Commerce sector, as it directly impacts successful marketing, inventory management, and overall business strategy. By discerning the qualities that are most closely correlated with purchase choices, firms can optimize their marketing strategies, customize customer experiences, and refine their product offers to better align with consumer needs. Furthermore, the use of the regression model methodology provides a higher degree of detail and precision that may be absent in alternative methodologies. It enables the examination of how alterations in certain characteristics can influence the likelihood of a purchase, offering insights that are both practical and measurable. This feature is essential in an online setting because even slight modifications to a website or product listing can have a substantial impact on client behavior. Implementing regression models in e-Commerce presents some difficulties. The large amount and diverse range of data, the requirement for advanced model selection and validation methodologies, and the fast rate of change in online consumer behavior present substantial challenges. Moreover, comprehending the findings of the model requires a sophisticated grasp of both statistical techniques and the business environment [3]. However, given the continuous growth and evolution of e-Commerce, it is imperative to underscore the need to develop robust, data-driven strategies to understand and predict customer purchasing behaviors. This research has the power not only to increase company performance in the digital marketplace but also to provide insight into the larger trends and patterns of consumer behavior in an increasingly online world.

2. Related Work

Digital commerce has revolutionized customer behavior, allowing for better

analysis of buying trends. Paulo and Tiago (2019) [4], investigate the impact of e-service quality on customer satisfaction, trust, and behavior in online buying. The study focuses on repurchase intention, word of mouth, and site re-visit. This study has limitations in forecasting consumer buying patterns and suggests that the efficiency of the e-service quality model may vary depending on product categories or industrial sectors. Jawaid *et al.* (2021) [5] examined the impact of social media campaigns, television advertising, electronic newspaper advertisements, and word of mouth marketing on customer purchasing behavior. The results showed that e-paper ads and word-of-mouth marketing significantly influenced purchasing behavior, while social media campaigns and television commercials did not, which emphasizes the need for e-commerce enterprises to adapt strategies to evolving customer behaviors during the epidemic.

However, it did not incorporate theoretical discoveries and included secondary data, which could have provided a more comprehensive understanding of consumer purchasing behavior. Yang Zhao, Lin Wang *et al.* (2020) [6] explores the impact of information quality, social psychological distance, and trust on consumer buying intentions. Data was collected from Xiaohongshu users through a questionnaire survey. The results show that information quality positively influences social psychological distance and trust, which in turn affect purchase intentions. Meanwhile, the study's nonuniform sample distribution, primarily female, results in an imbalanced representation. This is due to differing perceptions of information reliability in electronic word-of-mouth (eWOM). Neha Chaudhuri *et al.* conducted a study in 2021 [7] that compared deep learning methods to conventional machine learning techniques and analyzed a data set of over 50,000 online visits. It was focused on variables such as platform participation and client attributes to forecast online sales.

The data was exclusive of retail sales on a specific e-Commerce platform, focusing on European online shopping and a specific product category. The lack of scenarios and real-time prediction accuracy, as in real-time analysis, provides inaccurate predictions from the model. In 2020 [8], Mariya Hendriksen, Pim Nauts, and her team conducted a study on forecasting purchase intention in e-Commerce. They analyzed more than 95 million sessions on a prominent European e-commerce platform, examining variations in consumer behavior between anonymous and identifiable individuals, factors like session duration, device category, channel type, and search queries to distinguish between purchase and non-purchasing behaviors were analyzed. However, the limited data sampling of four weeks may present a sample bias, making the findings susceptible to unknown temporal or seasonal trends. This could affect the precision and relevance of the results across multiple seasons or various market scenarios.

The researcher primarily analyzes the on-to-offline (O2O) commerce paradigm but does not fully explore the developing off-to-offline commerce paradigm. While variables such as pricing and competitive positioning influence customer purchasing decisions, the methodology does not consider these factors.

Instead, it focuses on positive online content, ignoring the potential impact of negative information on consumer purchasing choices. This highlights the need for more research to better understand and predict consumer behavior in different O2O commerce scenarios [9]. The researcher examines the impact of price sensitivity and promotional methods on customer purchasing behavior.

However, the lack of these elements hinders a comprehensive explanation. Meanwhile, further research is lacking on consumer behavior disparities across product variety, services offered, and offline and online retailers. Furthermore, analyzing customer relationship behavior and the effects of trust, commitment, and information satisfaction on purchasing behavior in relational settings could be beneficial [10]. This study reveals key factors influencing online purchases and provides valuable insights for platform designers. However, it suggests the need for further research on how past purchases, indicating consumer loyalty, influence future purchases. This is crucial for the development of efficient e-Commerce systems and improving prediction models. The lack of research underscores the complexity of consumer behavior and the need for continuous investigation to understand the dynamics of online shopping [11].

In this paper, the limitations of Deep Item-Based Collaborative Filtering for Top-N Recommendation in accurately forecasting purchase behaviors, particularly the impact of customer loyalty on future purchases, identify factors that can predict online purchases, and provide insights for platform designers. However, it suggests the need for further studies to understand the impact of customer loyalty on future purchasing choices. It also emphasizes the importance of considering competition and alternative channels when creating online platforms. Prioritizing features that facilitate faster purchases may yield better results in situations of intense competition [12]. Neha Chaudhuri *et al.* [13] suggest that future research should explore the impact of customer loyalty on subsequent purchasing patterns.

Understanding customer loyalty can provide insight into consumer behavior. The study suggests that optimizing online platforms for faster transactions is beneficial in intense competition, but not always resulting in favorable purchasing decisions. Cheng-Ju Liu and colleagues [14] [15] have developed a machine learning model for the repurchase behavior in e-commerce, aiming to improve conventional methods for predicting online buying behavior. Methods like linear logistic regression and decision tree-based XGBoost reveal that nonlinear models are more effective in enhancing predictions.

However, the limitations suggest that its findings should be cautiously applied to wider e-Commerce contexts and require further research across other platforms and over longer durations. Zhenzhou Wu and his team use bidirectional LSTM Recurrent Neural Networks to simulate purchasing behavior in e-Commerce, based on consumers' clicking behaviors. This approach minimizes the need for significant feature engineering, a common requirement in machine learning models, by directly representing the sequence of click events [16].

3. Methodology

The study was carried out in Liberia, and participants were selected as respondents based on certain criteria to get the required information for the investigation. Participants were carefully selected to ensure that they had a clear recollection of encounters with the website of an online retailer. The selection criteria for respondents consisted of Liberian individuals who were actively or nonactivity Internet users and had engaged with online merchants by visiting, purchasing from, or utilizing their services at least once within a month. The research included all adults in Liberia who were 18 years of age or older, regardless of gender. To evaluate the suggested model, a questionnaire was created and answers were used. Data collection was carried out using distributed questionnaires. Participants were instructed to go to a certain location to submit their questionnaires. The participants were directed to answer according to the content of the questions. In accordance with the PII regulations, the personal information of the respondents was neither retained nor documented.

The study was carried out in Liberia and the respondents were selected based on specific criteria to gather the information required for the research. Participants were carefully chosen to ensure that they reflected various demographics and characteristics typical of Liberia's broader e-commerce consumer base.

The selection criteria for respondents consisted of Liberian individuals who were actively or non-active Internet users and had engaged with online merchants by visiting, purchasing from, or utilizing their services at least once within a month. Furthermore, efforts were made to include participants from various regions of Liberia to capture regional differences in the behavior of e-commerce.

Implementing this approach minimized the likelihood of excluding specific demographic segments and ensured that the findings were applied to Liberia's broader population of e-commerce users. This stratification helped ensure that the sample reflected the diversity of the Liberian population and the e-commerce consumer base.

Additionally, random sampling methods were used within each stratum to mitigate selection bias in the selection of participants. This method ensured that the results could be applied to the larger population of Liberian E-Commerce customers while reducing the possibility of leaving out specific demographic groups.

The research included all adults in Liberia who were 18 years or older, regardless of sex. A questionnaire was prepared and distributed to gather information to assess the proposed model. Data was collected using distributed questionnaires, and participants were randomly selected based on stratification criteria.

Participants were instructed to go to a specific location to submit their questionnaires. Clear instructions were provided to ensure consistency in the data collection process. Participants were asked to respond according to the content of the questions, with an emphasis on providing accurate and honest responses.

In compliance with Personally Identifiable Information (PII) regulations, the

personal information was not retained nor documented, ensuring the anonymity and confidentiality of the participants.

3.1. Data Collection and Preparation

The initial step involves the comprehensive collection of data from consumers of e-Commerce platforms. Data often encompasses several user attributes, such as gender, age, educational attainment, employment, monthly income, demographic characteristics, online purchasing behavior, preferred payment methods, and purchase intent, among other relevant factors. The data was clean and preprocessed, ensuring all categorical variables were encoded appropriately and missing values were handled. After this process, we initialize the Logistic Regression model and prepare it for training with the user attribute data. The model was trained using the preprocessed user data, where the target variable is the user's purchasing decision (purchased or not purchased). The trained model was then used to predict purchasing probabilities for new or existing users based on their attributes. Then the model Interpret output as the likelihood of purchase, with higher scores indicating a greater probability of purchasing behavior. The data set was subsequently partitioned into separate training and testing sets, often following a conventional split of 80% for training and 20% for testing. We ensure that adherence to data privacy and ethical norms is of utmost importance, particularly in the management of confidential user data.

The initial step involves collecting data from consumers on various e-commerce platforms. The data encompass a wide range of user attributes, including but not limited to gender, age, educational attainment, employment status, monthly income, demographic characteristics, online purchasing behavior, preferred payment methods, and purchasing intent.

It is critical to recognize that sample strategies, platform-specific user demographics, or data collection procedures may have introduced biases in the data. Pre-processing and data cleaning were carefully considered to reduce these biases. Categorical variables were appropriately encoded and missing values were handled using established techniques, such as imputation or removal. In addition, steps were taken to identify and address any systematic biases in the data set during preprocessing.

Following data pre-processing, the logistic regression model was initialized and prepared for training with the user attribute data. The model was trained using the preprocessed user data, with the target variable being the user's purchasing decision (i.e. whether they purchased or not). This training process involved optimizing model parameters to maximize predictive performance while minimizing the potential for overfitting.

Subsequently, the trained model was used to predict the purchasing probabilities for both new and existing users based on their attributes. The model output was interpreted as the likelihood of purchase, with higher scores indicating a higher probability of purchasing behavior.

The data set was divided into two halves, the training set and the testing set,

following the standard split of 80% training and 20% testing, to assess the model's accuracy and generalizability. This partitioning strategy allows for robust model evaluation while ensuring that the model's performance is assessed on unseen data.

It is essential to emphasize that the adherence to data privacy and ethical norms is paramount throughout the entire data handling process, particularly when handling confidential user data. Strict protocols were followed to safeguard user privacy and confidentiality, and all data handling procedures were conducted according to relevant regulations and ethical guidelines.

By meticulously addressing potential biases in data collection and preparation and ensuring adherence to ethical standards, the developed model aims to be applicable across various e-Commerce platforms, providing valuable insights into user purchasing behavior while maintaining the integrity and trustworthiness of the analysis.

3.2. Data Preprocessing

The initial step involves collecting data from consumers on various e-commerce platforms. Data cover many user attributes, including gender, age, educational attainment, employment status, monthly income, demographic characteristics, online purchasing behavior, preferred payment methods, and purchasing intent.

As a precaution against bias and to ensure the model's generalizability, rigorous data pretreatment processes were carried out:

1) Handling Missing Values: Missing values in the data set were identified and appropriately handled. Depending on the nature and extent of missingness, techniques such as mean or median imputation, mode imputation for categorical variables, or advanced imputation methods such as K-nearest neighbors (KNN) were utilized. By addressing missing values effectively, we minimize the impact of incomplete data on the model performance and ensure robustness across diverse datasets.

2) Encoding Categorical Variables: Categorical variables within the dataset were encoded to numerical representations suitable for model training. Techniques such as hot encoding or label encoding were used based on the nature of categorical variables and the requirements of the chosen machine learning algorithm. This step ensures that categorical attributes are adequately represented in a format that the model can interpret accurately.

3) Handling Imbalanced Data: Imbalanced data, where one class may significantly outnumber the other, poses challenges for model training and evaluation. Techniques such as oversampling (e.g. Synthetic Minority Over-sampling Technique—SMOTE) or undersampling were applied to balance the distribution of classes within the dataset. We need to fix the class imbalance to make the model more generalizable and less biased towards the dominant class.

4) Feature scaling: Continuous features in the data set were scaled to a standard range to prevent attributes with larger scales from dominating the learning process. Normalization, which sets the averages and standard deviations of the

characteristics to zero and one, respectively, and Min-Max scaling, which scales features to a range between 0 and 1, are popular scaling techniques. Feature scaling ensures that all features contribute equally to the model's learning process and enhances its ability to generalize across different feature magnitudes.

5) Dimensionality reduction: Dimensionality reduction strategies, such as principal component analysis (PCA) or feature selection approaches, were used to lower the dataset's complexity while preserving pertinent information in situations with high-dimensionality or multicollinearity around features. Dimensionality reduction improves model interpretability, reduces computational complexity, and improves generalizability by focusing on the most informative features.

By meticulously applying these data preprocessing steps, we enhance the quality and robustness of the dataset, thereby improving the model's generalizability across diverse e-commerce platforms. These steps ensure that the model can learn effectively from the data, capture underlying patterns, and make accurate predictions on unseen instances, ultimately facilitating its broader applicability and utility in real-world scenarios.

3.3. Model Designed

Figure 1 diagram illustrates the typical workflow of a model from raw data to predictions, structured into three main stages. Initially, preprocessing involves cleaning, normalizing, extracting features, and other necessary transformations of the raw dataset to prepare it for analysis, followed by partitioning the data into separate training and testing sets to prevent overfitting and accurately assessing the model on unseen data. Subsequently, the processed data is divided into

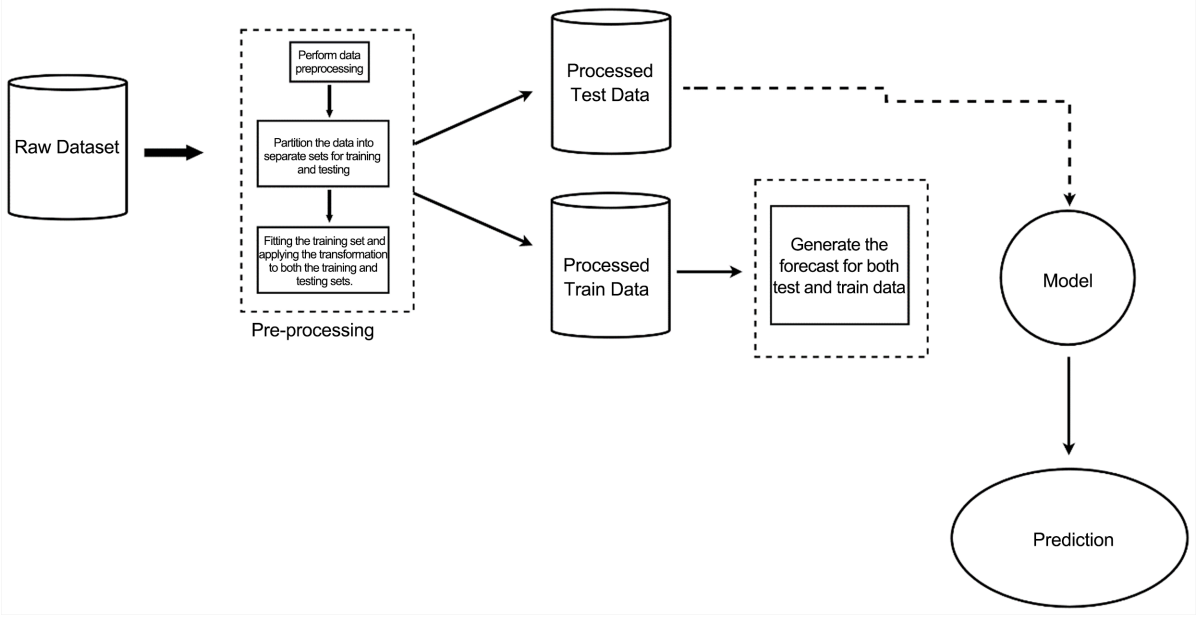

Figure 1. Model diagram.

training data, which is used to train the machine learning model, and testing data, used to evaluate the model's predictive performance. In the final stage, model training and prediction, the trained model applies its algorithms to both datasets to generate forecasts, culminating in predictions that reflect learned patterns from the training data. The diagram emphasizes the systematic approach taken to maintain the integrity, robustness, and accuracy of the model during the training and validation phases, ensuring it performs well on new, unseen data.

3.3.1. Algorithm

Algorithm: Predicting willingness to purchase

Input: A collection of user attributes

Output: Prediction to purchase

Steps:

1. $i \leftarrow []$
2. a = user_data.drop (u_1 , axis=1)
3. b= user_data [u_1]
4. $a[\mathrm{M}]$, $a[\mathrm{O}]$, $b[\ddot{\varphi}]$, $b[\dot{\varphi}]$= (G) (a, b, test_size=0.2, random_state = 42)
5. pipeline.fit (a [M], $b[\dot{\varphi}]$)
6. i = 0
7. p= []
8. while i < len (a [O]):
9. instance = a [O].iloc[I].values.reshape(1, -1)
10. u= pipeline. U(q) [0][1]
11. p.append(u)
12. i += 1

3.3.2. Algorithm Flowchart

The algorithm depicted in the flow chart begins by initializing an empty list. It then proceeds to segregate the user attributes from the target variable, subsequently extracting the target variable. Afterward, the data is partitioned into several sets for training and testing purposes. After training on the training set, a predictive model is used to iterate over the testing set. During each iteration, the program takes the attributes of the current sample, employs the model to forecast the likelihood of purchase intention, and adds this probability to a list. Iteration persists until all samples in the testing set have been processed, resulting in the retrieval of the list containing all the predicted probabilities. This is vividly shown in **Figure 2** below.

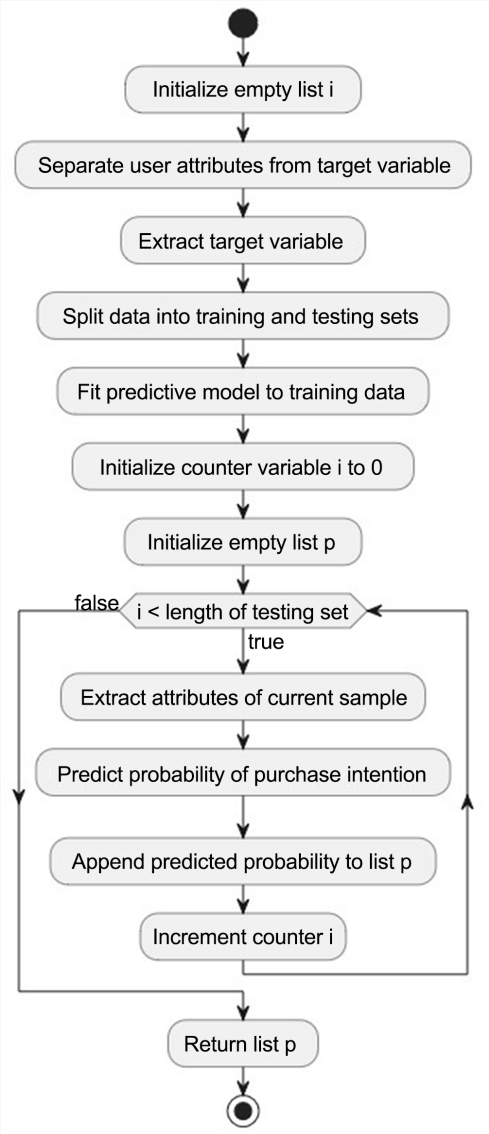

Figure 2. Algorithm flow chat.

3.4. Model

3.4.1. Logistic Regression

Logistic regression is widely recognized as a key technique in the field of binary classification due to its capacity to predict binary outcomes, such as (Yes/No or 1/0, utilizing many predictor variables. Fundamentally, logistic regression employs the logistic function, often known as the sigmoid function, $\sigma(z)=\frac{1}{1+e-z}$ to estimate the likelihood that a given input is associated with a specific category. The function as mentioned above is designed to assign a value between (0 and 1) to any given input z, which is often a linear combination of characteristics denoted as $z=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n$. This assigned value represents the probability that the dependent variable will be equal to 1. Estimation of model coefficients (β_0, $\beta_1,\cdots,\beta_n$) is performed by techniques such as maximum like-

lihood estimation, which aims to maximize the probability of witnessing the given sample data. The relevance of logistic regression is mainly ascribed to its interpretability and simplicity, making it a preferred methodology in diverse fields such as business, medicine, and finance. The probabilistic aspect of this phenomenon is highly respected due to its ability to provide valuable insights into the possibility of various outcomes. This attribute is particularly significant in the context of risk assessment and decision-making procedures. The capacity to modify the classification threshold, often established at 0.5, affords adaptability in various situations, enabling the optimization of the trade-off between false positives and false negatives according to the particular environment. Logistic regression has a notable level of precision, particularly in cases where the association between predictors and the outcome follows a linear pattern, despite its straightforwardness. The model has a lower susceptibility to overfitting compared to more intricate models. Furthermore, its interpretability is a notable benefit, since the coefficients establish direct associations with the odds ratios, offering lucid insights into the influence of each predictor variable.

3.4.2. Random Forest Regression

Random forest is often regarded as a highly integrated technological solution because of its simplicity, user-friendly nature, and low computational requirements. The Random Forest algorithm has demonstrated strong performance in addressing a wide range of practical challenges encountered in real-world scenarios. The random forest technique uses random attribute selection to train decision trees through bagging integration. Decision tree-based learners commonly employ this strategy. In an optimal scenario, it would be possible to partition the existing sample set into two equal subsets and employ the resultant decision tree model to effectively predict future events. In practical application, however, circumstances are never this straightforward. When the level of division is too exact, the algorithm experiences an overflow of data and overfits, while a substantial disparity between the anticipated and actual values poses challenges in accurately completing the prediction task. Scholars have devised novel concepts such as the Gini index and knowledge gain in order to evaluate the need for increased segregation.

Definition 1: (information entropy) Assume that the proportion of class k samples in the current sample set C is $p_k\left(k=1,\ 2,\cdots,\left|\gamma\right|\right)$, Then the information entropy of C is defined as

$$Ent\left(c\right)=-\sum_{k=1}^{\left|Y\right|}Pk\log 2Pk$$

Definition 2: (Information gain) Assume that the discrete attribute a has U possible values $\left\{a^1,a^2,\cdots,a^u\right\}$.

C^u represents the samples in sample set C whose value is a^u when divided according to discrete attribute a, then the information gain obtained after partitioning the sample set C with attribute a is,

$$Gain(c,a) = Ent(c)\sum_{U=1}^{U}\frac{|c^u|}{|c|}Ent(c^U)$$

Typically, the "purity boost" that is achieved by partitioning with attribute a grows in direct correlation with the information gain. So, the information gained may be utilized to decide if this characteristic is split by the following step when the decision tree is being formed.

3.5. Rationale for Tuning Parameters for Both Models

3.5.1. Regularization Technique: L1 Regularization

L1 regularization was chosen for the logistic regression model due to its ability to induce sparsity in coefficients, automating feature selection, and enhancing interpretability. This approach was reinforced by empirical experimentation, aligning with Occam's Razor principle of preferring simpler models. The liblinear solver was selected for optimization, given its compatibility with L1 regularization and efficiency on small to medium-sized datasets. Its fast convergence on sparse datasets complements L1 regularization's feature selection effect, ensuring a balance between complexity and performance for interpretable and accurate predictions in predictive modeling.

3.5.2. Number of Trees

In developing our random forest model, we conducted experiments to find the optimal number of trees balancing predictive performance and computational efficiency. We tested a range of values from 10 to 200 trees and found that model accuracy improved significantly up to 100 trees, beyond which gains plateaued. We chose 100 trees for our final model as it provided robust performance while managing computational time effectively. This decision reflects a pragmatic approach guided by empirical evidence to achieve an accurate and efficient model. By selecting 100 trees, our random forest model delivers high-quality predictions promptly, making it suitable for practical applications in e-commerce platforms where accuracy and efficiency are crucial.

3.5.3. Maximum Depth of Trees

The maximum depth of trees in a random forest model regulates model complexity, balancing between capturing intricate patterns and avoiding overfitting. We determined the maximum depth by allowing trees to grow until nodes contained a minimum number of samples before splitting. This approach ensures sufficient detail to capture underlying data patterns while preventing unnecessary complexity. The chosen depth strikes a balance between model complexity and generalizability, validated empirically to prevent overfitting while retaining essential data patterns.

3.5.4. Feature Sampling Strategy

Random Forest models employ feature sampling to enhance generalization by reducing correlation between trees. In our approach, features were randomly

sampled, divided by the square root of the total number of characteristics, ensuring diversity in feature selection across trees. This strategy aimed to increase robustness to noise and reduce overfitting. Empirical testing, including cross-validation, demonstrated improved model performance, affirming the successful balance between tree diversity and predictive power.

3.6. Cross-Validation

Cross-validation is crucial for assessing machine learning models' predictive performance and generalizability. In our study, we employed 10-fold cross-validation to validate both logistic regression and random forest models. This method partitions the data into 10 subsets, using 9 for training and 1 for testing, repeating the process 10 times. The choice of 10-fold cross-validation balances computational efficiency with reliable performance estimates, ensuring each observation is used for training and testing exactly once. Results from folds are averaged to provide a stable performance metric, offering an accurate assessment of the models' ability to generalize to unseen data.

3.7. Feature Importance and Selection

Feature importance and selection play vital roles in enhancing interpretability and efficiency in machine learning models. In our study, for the logistic regression model, L1 regularization facilitated feature selection by penalizing irrelevant features, making the model simpler and interpretable. Conversely, in the random forest model, feature importance was assessed using the reduction in impurity (Gini index) across all trees, identifying the most influential features in predicting outcomes. This approach allowed us to reduce model complexity while maintaining predictive accuracy, leading to increased computing efficiency and a deeper understanding of the connections between the target variable and characteristics in e-Commerce platforms.

3.8. Model Evaluation Metrics

Selecting appropriate evaluation metrics is vital for accurately assessing model performance and ensuring alignment with study objectives. In our analysis, we utilized accuracy, precision, recall, and F1 score as primary metrics for evaluating both models. Accuracy measures overall correctness, while precision evaluates the model's ability to identify relevant instances accurately. Recall assesses the model's ability to capture all relevant instances, and the F1 score balances precision and recall to provide a comprehensive performance metric. These metrics were chosen to address the correctness of predictions and handle class imbalances effectively, ensuring the models' accuracy and practicality for real-world applications were identifying positive instances and minimizing false positives are crucial.

Incorporating these general enhancements into the logistic regression and random forest models has contributed significantly to their robustness, inter-

pretability, and alignment with the objectives of the study. Cross-validation ensured that our performance estimates were reliable and indicative of the model's generalization ability. The importance and selection processes of the models' efficiency and interpretability of the models, allow a deeper understanding of the underlying factors that influence purchasing behavior. Finally, the thoughtful selection of evaluation metrics ensured a comprehensive assessment of the models' performance, highlighting their practical utility in predicting customer purchasing behavior on e-commerce platforms.

4. Results and Discussion

4.1. Confusion Matrix

Table 1 indicates the confusion matrix results illustrating the performance metrics of two models, Logistic Regression and Random Forest Regression, on a binary classification task involving Class 0 and Class 1. The Logistic Regression model shows a precision that indicates no false positives, and a recall, while the Random Forest model achieves perfect scores.

The classification model exhibits a notable level of performance, achieving an overall accuracy rate of 91.94%. This high accuracy suggests that the model is useful in accurately predicting outcomes in the majority of instances. The analysis of the confusion matrix indicates that the model has a high level of accuracy in predicting the positive class (class 1). With 141 true positives and no false negatives, it successfully identifies every instance of class 1 in the test set, achieving a recall rate of 100%. In contrast, the accuracy for the negative class (class 0) is perfect at 100%; however, the recall is somewhat lower at 83%. This indicates that the model fails to identify 17% of the genuine class 0 occurrences, resulting in 22 false positives. The observed difference in performance between the two classes is also evident in the F1 scores, which are reported as 91% for class 0 and slightly higher at 93% for class 1. The macro- and weighted average F1 scores, both at 92%, indicate a well-balanced performance across classes, taking into account the relative proportions of each class within the dataset.

The classification model demonstrates a commendable level of performance, boasting an impressive overall accuracy rate of 91.94%. This high precision underscores the effectiveness of the model in predicting outcomes in most instances.

Table 1. Confusion matrix of the prediction.

Confusion matrix results between Logistic Regression and Random Forest Regression							
	Class 0			Class 1			
Logistic Regression	Precision	Recall	F1-score	Precision	Recall	F1-score	Accuracy
	1.00	0.83	0.91	0.87	1.00	0.93	92%
Random Forest Regression	1.00	1.00	1.00	1.00	1.00	1.00	100%

Upon closer examination of the confusion matrix, it becomes evident that the model predicts the positive class (class 1). In particular, with 141 true positives and no false negatives, the model achieves a perfect recall rate of 100% for class 1. This signifies the model's exceptional ability to identify every instance of class 1 in the test set, which is crucial in scenarios where correctly identifying positive cases is paramount.

On the contrary, while the negative class (class 0) accuracy is also perfect at 100%, the recall is marginally lower at 83%. This discrepancy implies that the model overlooks approximately 17% of genuine Class 0 occurrences, leading to 22 false positives. Although the model achieves a high accuracy rate for class 0, the relatively lower recall highlights its limitation in capturing all instances of class 0, which may have practical implications in specific applications.

The F1 scores, which serve as a harmonic mean of precision and recall, provide further insight into the performance of the model. For class 0, the F1 score is 91%, indicating a strong balance between precision and recall. In comparison, the F1 score for class 1 is slightly higher at 93%, reflecting the model's superior performance in correctly identifying positive cases.

The macro and weighted average F1 scores, at 92%, signify well-balanced performance between classes, accounting for the varying proportions of each class within the data set. This holistic evaluation underscores the model's ability to maintain consistent predictive accuracy across different class distributions, which is crucial for its generalizability and applicability in diverse real-world scenarios.

In summary, while the classification model overall demonstrates impressive accuracy, there is room for further improvement, particularly in enhancing the recall rate for class 0. Future research could explore refined feature engineering techniques or alternative model architectures to address this limitation and improve the model's performance in identifying negative cases. Moreover, comparative analysis with existing literature on similar classification tasks could offer valuable insights into the model's strengths and areas for refinement, thereby contributing to the advancement of predictive modeling methodologies in the field.

4.1.1. The Situation of the Training and Testing Sets Prediction

Figure 3 and **Figure 4** depict the actual versus predicted probabilities for a binary classification model, with **Figure 3** focusing on the training set and **Figure 4** on the testing set. The plot for the training set shows a dense collection of predicted probabilities with many samples having high probabilities, which indicates a model that is confidently fitted to the training data. The testing set plot demonstrates more variation in predicted probabilities, suggesting that the model is less certain about its predictions on unseen data.

Figure 3 and **Figure 4** represent the actual vs. predicted probabilities for a binary classification model. **Figure 3** shows the training set, and **Figure 4** shows the testing set. The training plot shows a dense collection of predicted probabili-

ties, with many samples having high probabilities, indicative of a confident model fit to the training data. The testing plot shows more variation in the predicted probabilities, suggesting the model is less certain about its predictions on unseen data. Overall, these plots are used to evaluate the calibration of a classification model – how well the predicted probabilities match the actual outcomes. They can also help identify whether the model is overfitting, as an overfitted model would typically show very high confidence on the training set but perform less confidently on the testing set.

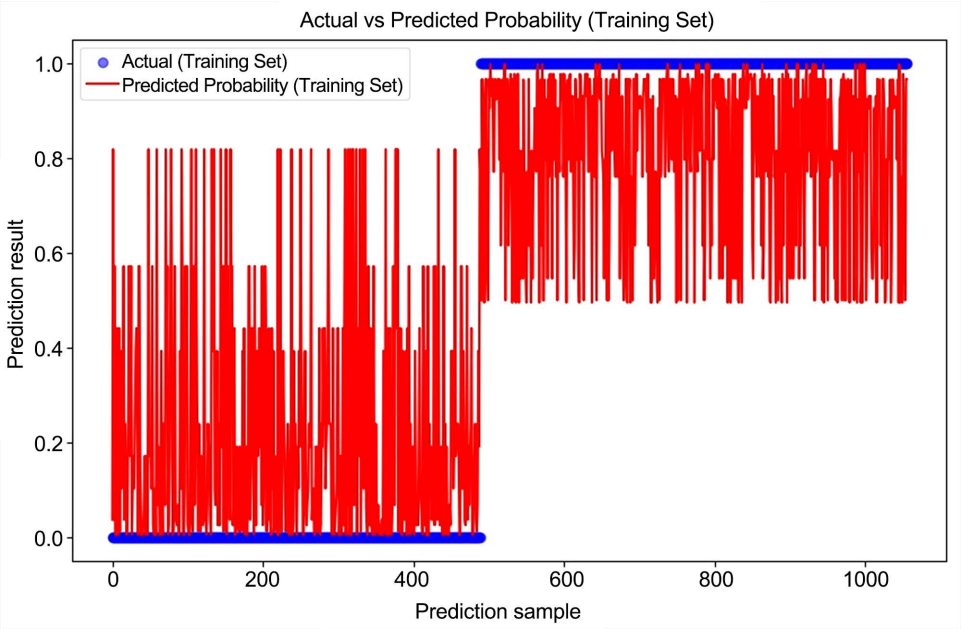

Figure 3. The situation of the training set prediction.

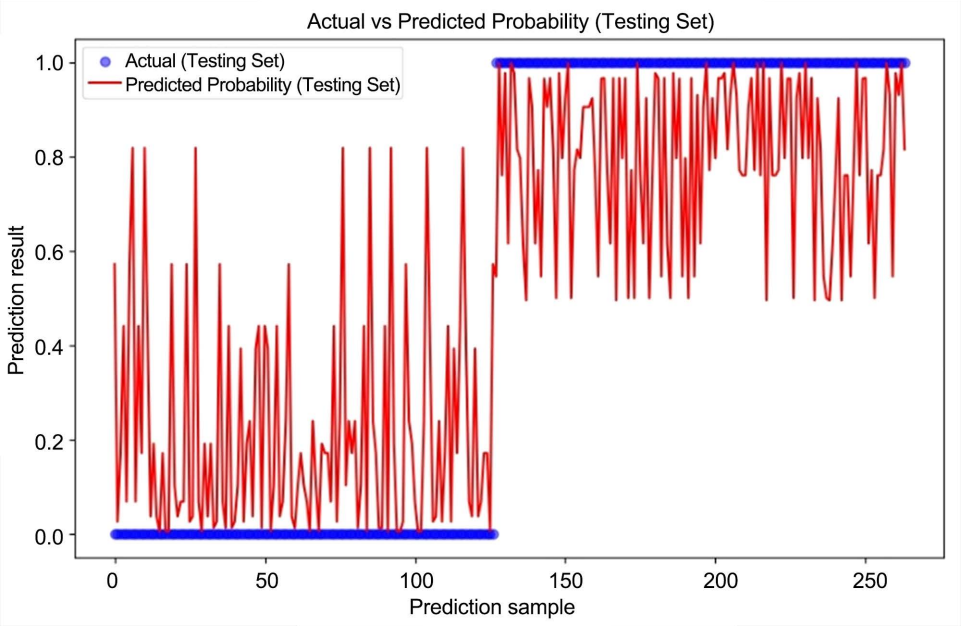

Figure 4. The situation of the testing set prediction.

4.1.2. Confusion Matrix for both Training and Testing

Figure 5 shows two confusion matrices: one for the training data on the left and one for the testing data on the right, illustrating the performance of a binary classification model.

Figure 5 below presents two confusion matrices: one for the training data on the left and one for the testing data on the right, illustrating the performance of a binary classification model. For the training data, the matrix shows 466 true negatives (TN), indicating that 466 instances were correctly identified as class 0, and 69 false positives (FP), where 69 instances were incorrectly labeled as class 1. The testing data matrix reveals 450 true negatives, confirming accurate predictions for class 0 in 450 cases, but shows an increase in false positives to 106, reflecting more instances misclassified as class 1. Additionally, it records 19 true positives (TP), where 19 instances were correctly predicted as class 1. These matrices are crucial for evaluating the accuracy of the classifier, indicating how well the model performs on both the data it was trained on and new, unseen data. The results suggest that the model tends to be more conservative in predicting class 1 on new data, leading to a higher number of false positives during testing, which can be an area of focus for improving the model's generalization capabilities.

4.1.3. The Feature Importance

Figure 6 shows the detailed importance of various features in a predictive model. The features range across demographic data, experiences with data protection, behavioral attributes, and awareness levels. At the top of the importance scale, "Age" stands out as the most significant feature. As we move from left to right in the chart, the importance of the features gradually decreases, with "Monthly Income" ranked as the least significant in predicting the target variable.

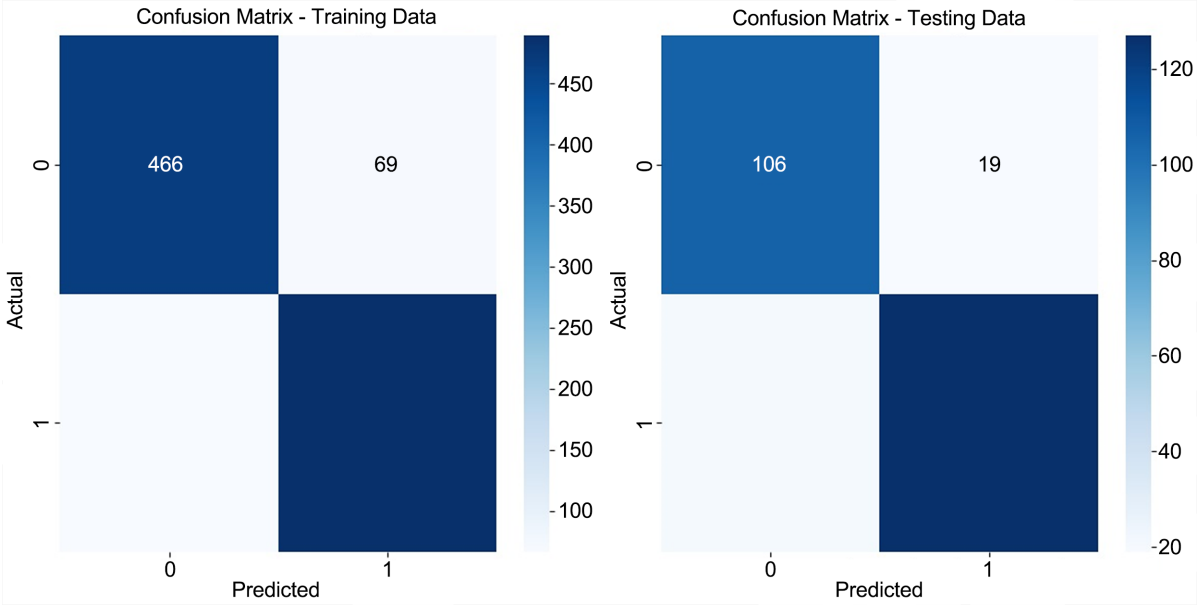

Figure 5. Confusion matrix for both training and testing.

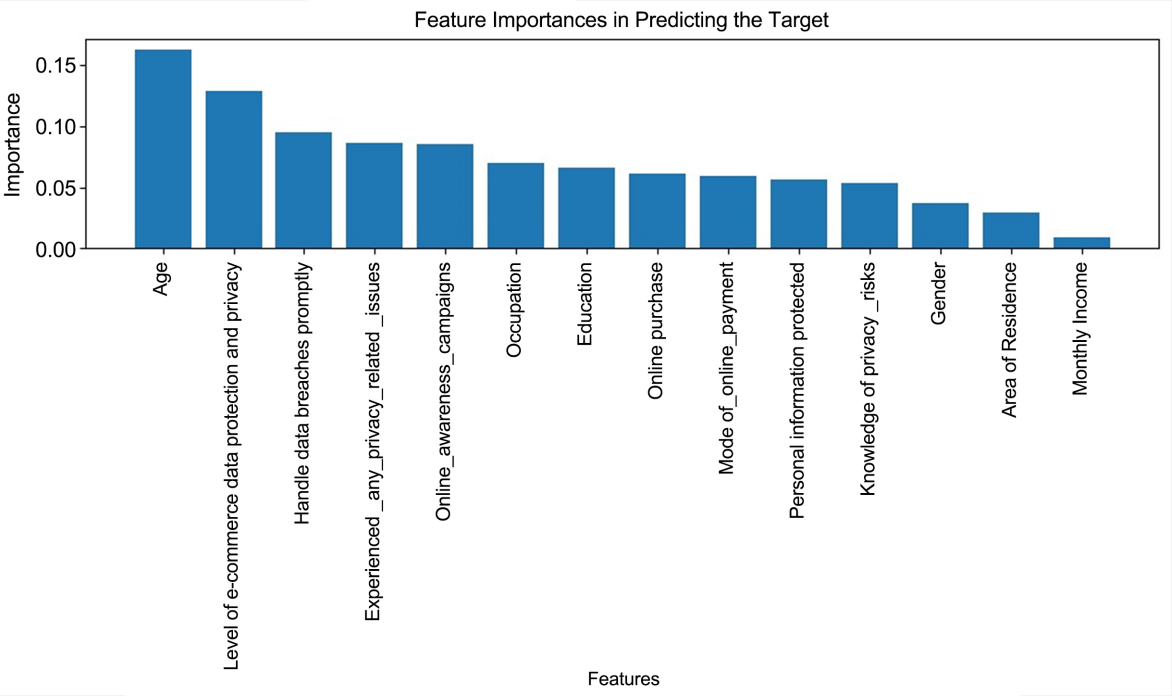

Figure 6. The feature importance.

The feature exhibits a hierarchical arrangement of characteristics that exert an effect on purchase decisions made on an e-commerce platform. These factors have been established through the use of a predictive model. The variable "Age" is identified as the most prominent factor, implying that customer age has a substantial impact on online purchase behaviors, indicating distinct preferences or purchasing capabilities among various age groups. The subsequent section is titled "Level of e-Commerce data protection and privacy awareness", emphasizing the crucial significance of trust in ensuring data security during e-commerce transactions. The features related to proactive management of data breach management, the experience of privacy issues, and the influence of online awareness efforts constitute the middle level of importance. This highlights the intricate manner in which a user's worries and experiences with privacy influence their purchase decisions. In the context of online buying behaviors, the variables "Occupation" and "Education" indicate a potential association with socioeconomic aspects. The factors "Knowledge of privacy hazards", "Gender", "Area of residence", and "Monthly Income", although ranked lower in terms of effect separately, have the potential to provide more comprehensive behavioral insights when considered together. This forecast highlights the complex and diverse aspects of online consumer behavior, where demographic characteristics and personal privacy experiences play a crucial role in shaping purchase decisions.

4.2. Business Impact and Recommendations

The feature importance graph provides valuable information that can be used by

e-commerce enterprises and stakeholders. It highlights the importance of understanding and utilizing customer demographics and trust factors, which can result in notable competitive advantages. The significant impact of age on consumer purchasing behavior suggests that the implementation of customized marketing strategies and product offers for different age cohorts has the potential to generate considerable improvements in conversion rates. The significant significance of safeguarding data and preserving privacy underscores the notion that implementing strong security protocols and maintaining transparent privacy rules are not just matters of regulatory adherence, but are also crucial for fostering customer confidence and loyalty. This highlights a significant opportunity for companies to distinguish themselves in the market by highlighting their commitment to data protection, particularly following any possible security breaches. Furthermore, the model places significant importance on the influence of online awareness campaigns, indicating that the provision of information to customers on the security and advantages of online shopping has the potential to bolster their trust and thus increase their inclination to participate in e-Commerce transactions. The relatively modest influence of career and education on purchase decisions presents opportunities for nuanced marketing approaches. This implies that tailor-made techniques that take into account customer professional and educational profiles have the potential to enhance engagement and effectiveness. In the meantime, although criteria such as awareness of privacy hazards, gender, geographical location, and monthly income have a less significant influence, they nevertheless have value in facilitating comprehensive customer segmentation, guiding personalized content strategies, and enhancing user satisfaction. These collective insights have the potential to guide e-commerce platforms in enhancing their strategy in several operational areas, including marketing, customer service, product management, and cybersecurity. By strategically adapting their strategies to correspond to these essential consumer characteristics, companies have the potential to improve customer happiness and establish credibility, ultimately leading to sustainable growth within the highly competitive online marketplace.

4.3. Model Accounts for Rapidly Changing E-Commerce Behaviors and Trends over Time

In addressing the challenges posed by rapidly changing e-commerce behaviors and trends over time, the researcher implemented adaptive strategies within the model. First, regular model updates are crucial to keeping up with evolving patterns in e-commerce dynamics. It can capture emerging trends and consumer behaviors by periodically retraining models with the latest data, ensuring their relevance and effectiveness in predicting outcomes. Additionally, feature engineering plays a pivotal role in reflecting evolving market dynamics. The model may adjust to changing customer preferences and market situations by continuously improving the collection of predictor variables to include new elements, such as social media engagement metrics and product popularity trends. Fur-

thermore, regression's dynamic threshold modification enables real-time adaptation to changing e-commerce patterns.

The model can optimize the balance between false positives and false negatives by flexibly modifying the classification threshold based on evolving dynamics, thus enhancing their predictive accuracy. It can also extract insights from textual data sources, such as social media comments and customer reviews because it incorporates sophisticated natural language processing techniques such as sentiment analysis and text mining.

This allows for a deeper understanding of the changing sentiments and preferences. Furthermore, adopting adaptive learning algorithms allows for dynamic adjustments to the model parameters in response to changing data patterns, ensuring continuous adaptation to evolving e-commerce dynamics. Finally, robust cross-validation techniques enable systematic assessment of model performance and generalizability across diverse datasets, including historical and recent data, thus validating the models' efficacy in capturing evolving trends and behaviors in the e-commerce landscape. By implementing these adaptive strategies, our models can effectively account for the dynamic nature of e-commerce environments, ensuring their relevance and applicability over time.

4.4. Limitation and Future Work

Although the predictive model offers useful information on the purchase behavior of e-commerce, it is not exempt from some restrictions. The model has a high level of accuracy in recognizing class 1 while displaying a more cautious approach in classifying instances as class 0. This suggests a possible need for improvement, particularly in situations where the failure to forecast class 0 accurately could have significant consequences. The existing scope may not adequately consider the fast evolution of consumer trends and technological breakthroughs, which have the potential to substantially modify purchase behaviors. Moreover, the dependence on historical data might potentially create biases or fail to account for developing patterns that have not yet been included in the data set. The effectiveness may also be limited by the level of detail in the data that is accessible, which might result in an oversimplification of intricate consumer behaviors into visible characteristics. Another constraint exists in the comprehensibility of the model, particularly when employing sophisticated machine learning algorithms that function as "black boxes", rendering it challenging to ascertain the underlying reasoning behind certain predictions. In future research endeavors, it is recommended to augment the model by integrating real-time data analysis techniques to capture emerging patterns with greater precision. The use of unstructured data, such as social media-derived consumer sentiment research, has the potential to enhance the predictive capabilities of the model. Progress made in the field of artificial intelligence has the potential to provide a more intricate understanding of consumer behavior. This may be achieved through the utilization of techniques such as deep learning, which possesses the capability to discern patterns beyond the capabilities of conventional models.

The continuous process of refining and validating the model against up-to-date data will be vital to maintaining its relevance and accuracy. Furthermore, it is recommended that future versions of the system prioritize the enhancement of openness and explainability. This will serve to strengthen user confidence and provide more comprehensible information for people responsible for making informed decisions.

5. Conclusions

Examining e-commerce purchasing behavior using regression modeling provides valuable information to organizations seeking to enhance the performance of their online platforms. The investigation underscores the importance of age, data protection, and privacy concerns as crucial determinants in shaping consumer buying behavior. By giving priority to these areas, e-commerce companies can customize user experiences to meet consumer expectations. This, in turn, leads to improvement in satisfaction and trust, both of which play a crucial role in driving online sales. The model's precision, as evidenced by its accuracy and the feature significance graph, implies a robust basis for forecasting consumer behavior. However, it is essential to acknowledge the limits of the model, such as its reliance on historical data and the possibility of inherent biases. These factors underscore the need for continuous adaptation and development. Future advancements should prioritize the integration of real-time data, the acknowledgment of the intricacies of consumer behavior through more sophisticated analytics, and the preservation of model transparency to ensure ongoing relevance in the ever-expanding digital marketplace. In summary, although the existing model provides significant predictions that can drive specific initiatives, its development must ensure its continued effectiveness in properly forecasting e-commerce purchase behavior.

To practically implement the insights derived from the findings, e-Commerce platforms can undertake several strategic initiatives. First, by taking advantage of the importance of age on consumer purchasing behavior, platforms can tailor their marketing strategies to different age cohorts, crafting personalized campaigns and product offerings to improve conversion rates. Second, platforms prioritizing data protection measures can invest in robust cybersecurity systems and transparent privacy protocols to foster customer confidence and loyalty, especially after security breaches. Third, by conducting online awareness campaigns highlighting the security and advantages of online shopping, platforms can bolster consumer trust and participation in e-commerce transactions.

In addition, e-commerce platforms can adopt nuanced marketing approaches based on customers' professional and educational profiles to enhance engagement and effectiveness. Therefore, by using comprehensive customer segmentation strategies and tailoring content and services to various consumer characteristics, e-commerce platforms can improve user satisfaction and retention. By integrating these insights into various operational areas, including marketing, customer service, product management, and cybersecurity, e-commerce plat-

forms can optimize their operations, improve customer experiences, and establish credibility in the highly competitive online marketplace, ultimately driving sustainable growth.

Acknowledgments

We are grateful to our colleagues and peers for their significant contributions, inspirational discussions, persistent encouragement, and collaborative efforts throughout this project. Their different perspectives and insights have proven quite beneficial.

We appreciate that this project would have been much more difficult without the collaborative commitment and support of all the authors involved. We greatly appreciate everyone's dedication and contributions that have made this project come to fruition.

Authors' Contributions

AB Jallah Balyemah: Conceptualization, Methodology, Validation, Implementation, Visualization, Analysis, Data Collection, Writing: Original Draft, Revised Manuscript.

Sonkarlay J.Y. Weamie: Validation, Methodology, Editing, Visualization, Implementation, Analysis, Writing: Original Draft, Revised Manuscript.

Jiang Bin: Supervision, Editing, Validation, Implementation, Writing: Revised Manuscript.

Felix Jwakdak Joshua: Editing, Implementation, Analysis, Writing: Revised Manuscript.

Karmue Vasco Janda: Implementation, Analysis, Writing: Revised Manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interest about the publication of this research.

References

[1] Sivarajah, U., Kamal, M.M., Irani, Z. and Weerakkody, V. (2017) Critical Analysis of Big Data Challenges and Analytical Methods. *Journal of Business Research*, **70**, 263-286. https://doi.org/10.1016/j.jbusres.2016.08.001

[2] Kumar, S.S., Margala, M., Shankar, S. and Chakrabarti, P. (2023) A Novel Weight-Optimized LSTM for Dynamic Pricing Solutions in e-Commerce Platforms Based on Customer Buying Behaviour. *Soft Computing*, 1-13. https://doi.org/10.1007/s00500-023-08729-1

[3] Janiesch, C., Zschech, P. and Heinrich, K. (2021) Machine Learning and Deep Learning. *Electronic Markets*, **31**, 685-695. https://doi.org/10.1007/s12525-021-00475-2

[4] Rita, P., Oliveira, T. and Farisa, A. (2019) The Impact of E-Service Quality and Customer Satisfaction on Customer Behavior in Online Shopping. *Heliyon*, **5**, e02690. https://doi.org/10.1016/j.heliyon.2019.e02690

[5] Idrees, M., Khan, M. and Khan, A. (2020) Factors Affecting Consumer Buying Behavior for Electronic Notebook. *European Journal of Business and Management Research*, **5**, No. 3. https://doi.org/10.24018/ejbmr.2020.5.3.339

[6] Zhao, Y., Wang, L., Tang, H. and Zhang, Y. (2020) Electronic Word-of-Mouth and Consumer Purchase Intentions in Social E-Commerce. *Electronic Commerce Research and Applications*, **41**, Article ID: 100980. https://doi.org/10.1016/j.elerap.2020.100980

[7] Chaudhuri, N., Gupta, G., Vamsi, V. and Bose, I. (2021) On the Platform But Will They Buy? Predicting Customers' Purchase Behavior Using Deep Learning. *Decision Support Systems*, **149**, Article ID: 113622. https://doi.org/10.1016/j.dss.2021.113622

[8] Astuti, R. and Pulungan, D. (2022) Analysis of Factors Affecting E-Commerce Customer Purchase Decisions. *Morfai Journal*, **2**, 1-20. https://doi.org/10.54443/morfai.v2i1.190

[9] Delina, R., Gróf, M.and Dráb, R. (2021) Understanding the Determinants and Specifics of Pre-Commercial Procurement. *Journal of Theoretical and Applied Electronic Commerce Research*, **16**, 104-124. https://doi.org/10.4067/S0718-18762021000200107

[10] Park, C.H. and Kim, Y.G. (2003) Identifying Key Factors Affecting Consumer Purchase Behavior in an Online Shopping Context. *International Journal of Retail & Distribution Management*, **31**, 16-29. https://doi.org/10.1108/09590550310457818

[11] Xue, F., He, X., *et al*., (2019) Deep Item-Based Collaborative Filtering for Top-N Recommendation. *ACM Transactions on Information Systems*, **37**, 1-25. https://doi.org/10.1145/3314578

[12] Jain, A., Nagar, S., *et al.* (2020) EMUCF: Enhanced Multistage User-Based Collaborative Filtering through Non-Linear Similarity for Recommendation Systems. *Expert Systems with Applications*, **161**, Article ID: 113724. https://doi.org/10.1016/j.eswa.2020.113724

[13] Song, P. (2020) An XGBoost Algorithm for Predicting Purchasing Behaviour on E-Commerce Platforms. *Technical Gazette*, **27**, No. 5. https://doi.org/10.17559/TV-20200808113807

[14] Liu, C.J., *et al.* (2020) Machine Learning-Based E-Commerce Platform Repurchase Customer Prediction Model. *PLOS ONE*, **15**, e0243105. https://doi.org/10.1371/journal.pone.0243105

[15] Zhou, Y., Mishra, S., *et al.* (2019) Understanding Consumer Journey Using Attention Based Recurrent Neural Networks. *Proceedings of the* 25*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, July 2019, 3102-3111. https://doi.org/10.1145/3292500.3330753

[16] Wu, Z., Tan, B.H., *et al.* (2015) Neural Modeling of Buying Behaviour for E-Commerce from Clicking Patterns. *Proceedings of the* 2015 *International ACM Recommender Systems Challenge*, September 2015, 1-4. https://doi.org/10.1145/2813448.2813521