MESR
منار الشرق للدراسات والأبحاث
Manar Elsharq for Studies & Research
Manar Alsharq Journal
for Education and Instructional Technology
Homepage:
http://meijournals.com/ojs/index.php/majeit/index
ISSN: 2790-6698
مجلة منار الشرق
للتربية و تكنولوجيا
التعليم

Customer Classification in E-Commerce Using Random Forest Algorithm

Noor Shobar Ali

Department computer software

Azad University, South Tehran

استلام البحث: 15/02/2026 **مراجعة البحث:** 17/03/2026 **قبول البحث:**10/04/2026

الملخص:

تدرس هذه الدراسة مدى كفاءة خوارزميات التعلم الآلي لتصنيف الغابات العشوائية في التنبؤ بمستويات رضا العملاء عند التسوق عبر الإنترنت. تم تجميع مجموعة بيانات اصطناعية تضم 350 سجلًا للعملاء، تحتوي على بيانات ديموغرافية وسلوكية وبيانات معاملات، بهدف تصنيف العملاء إلى ثلاث فئات من الرضا (محايد/راضٍ/غير راضٍ). قُسّمت جميع البيانات إلى مجموعتي بيانات تدريب واختبار باستخدام تقسيم طبقي بنسبة 20/80، وخضعت جميع البيانات للمعالجة المسبقة والتحويل إلى رموز رقمية. استُخدمت مقاييس تقييم مختلفة لتقييم الأداء، تشمل الدقة، والضبط، والاستدعاء، ومقياس F1، ومصفوفة الارتباك، والتحقق المتبادل، ومساحة منحنى ROC متعدد الفئات. تفوقت خوارزمية الغابات العشوائية على الخوارزميات الأخرى بمعدل دقة إجمالي بلغ 98.57%، ومعدل دقة متوسط للتحقق المتبادل بلغ 98.86%، ومتوسط مساحة منحنى ROC بلغ 0.9963. أظهر تحليل أهمية الخصائص أن عدد الأيام منذ آخر عملية شراء وإجمالي الإنفاق هما العاملان الأكثر تأثيرًا على رضا العملاء. وقد أثبتت خوارزمية الغابة العشوائية أداءً متميزًا مع مجموعات بيانات التجارة الإلكترونية الكبيرة والمعقدة، مما يُساعد تجار التجزئة على تحسين استراتيجياتهم للاحتفاظ بالعملاء وإنشاء برامج تسويقية مُخصصة. تُعد هذه الدراسة جزءًا من مجموعة متنامية من المعارف التي تهدف إلى استخدام البيانات لتحسين جودة التنبؤ برضا العملاء لدى تجار التجزئة عبر الإنترنت.

الكلمات المفتاحية: سلوك العملاء، رضا العملاء، التجارة الإلكترونية، التعلم الآلي، الغابة العشوائية

Abstract

This study investigates how well Random Forest Classification machine learning algorithms predict customer satisfaction levels when shopping online. A synthetic dataset of 350 customer records containing demographics, behavior and transaction data was assembled with the purpose of segmenting customers into three categories of satisfaction (Neutral/Satisfied/Unsatisfied). All data was divided into training and test datasets using an 80/20 stratified split, and all data was preprocessed and transformed into numerical codes. Different evaluation metrics were used for evaluating performance; these include Accuracy, Precision, Recall, F1 Score, Confusion Matrix, Cross Validation, and Multiclass ROC-AUC. The Random Forest algorithm outperformed the other algorithms with a total accuracy rate of 98.57%, a mean Cross Validation accuracy rate of 98.86%, and a mean ROC-AUC score of 0.9963. The feature importance analysis indicated that Days Since Last Purchase and Total Spend are the most important factors influencing customer satisfaction. The Random Forest algorithm has shown superior performance with large and complex e-commerce datasets, which can help retailers improve how they retain customers and create customized marketing programs. This study is part of an increasing body of knowledge aimed at using data to improve the quality of customer satisfaction forecasting for online retailers.

Keywords: Customer Behavior, Customer Satisfaction, E-commerce, Machine Learning, Random Forest

I. INTRODUCTION

Context and Background

The exponential expansion of e-commerce has changed the way businesses interface with customers and execute transactions entirely. During the last decade, online shopping has become a fixture in consumers' daily habits, allowing shoppers to purchase nearly any product, or service, to be delivered to their homes or mobile devices, with great ease. (Vanderveld, Pandey, Han, & Parekh, 2016) This digital transformation is also granting businesses access to more customer data than ever. (Bernat, Koning, & Fok, 2019)Businesses now have access to extensive customer data - demographic, purchasing behavior, or behavioral engagement data and trends. Such data represents a tremendous opportunity for businesses to understand their customers, predict their future actions, and ultimately help design more targeted engagement and marketing approaches that will drive customer satisfaction. (Chamberlain, Cardoso, Liu, Pagliari, & Deisenroth, 2017) Customer satisfaction has become essential for the success of e-commerce businesses. (Chen, 2018) Satisfied customers will increase a business' revenue through repeat purchases, brag on platforms for positive experiences, and recommend a platform to friends, which enhances loyalty to a platform, all adding revenue potential. Unsatisfied customers will have the opposite impact. (Data Science Group, Amperity Inc., 2019)As the online marketplace is fast-paced and overwhelmingly competitive, it is detrimental to a business' profitability and reputation. Therefore, e-commerce companies must prioritize understanding Customer Satisfaction to retain and ultimately grow their customer base. (Farzanfar & Delafrooz, 2016) Even with a wealth of customer data at hand, predicting customer behavior and satisfaction is still an enormous challenge. E-retailers typically produce multiple complex, heterogeneous datasets, containing both numerical features (e.g., total spend, quantity of items purchased, average rating of products purchased) and categorical features (e.g., gender, city, membership type). (Jangid, Kothari, Spear, & Wadsworth, 2014) Additionally, customer behavior is a function of an array of factors (e.g., personal preference, frequency of shopping, promotional discounts, overall state of the economy), which interact in non-linear and often unpredictable ways (Karlssson, 2016). (Jasek, Vrana, Sperkova, Smutny, & Kobulsky, 2019) Because of all this complexity, it is challenging for standard analytical techniques to accurately model and forecast customer satisfaction. (Glady, Baesens, & Croux, 2008)Hence, businesses require refined analytical techniques that can handle large, heterogeneous datasets and identify meaningful patterns to facilitate better, data-driven decision-making. (Rathi, 2011)

Problem Statement

As a result of analyzing various aspects of the customer's behavior and the complexity that results from the way that this customer will behave and interact with an organization. Thus how do we determine the satisfaction levels of our customers through their interaction? Through the process of accurately classifying customers based on three broad classes: "satisfied", "neutral", or "not satisfied"; organisations, by correctly identifying which customers fall into what classification will be able to better position themselves to identify those customers that are most at risk, to improve their ability to better market to their customers and to enhance the customer experience overall. Simple techniques such as traditional descriptive statistics and correlation, while informative, do not provide adequate insight into the specific nuances of customer satisfaction based upon the interaction of multiple behaviours or transactions. For that reason, machine learning is needed, which can take advantage of rich data sources associated with e-commerce to accurately predict customer satisfaction. Satisfaction is ultimately a subjective measurement, and there are many variables that can influence customer satisfaction that are different from one customer to the next. Some customers could be concerned primarily with shipping times, and pricing, while other customers may only be focused on brand attributes like quality, reputation, and personal service. For any

predictive model to have any predictive capacity, it must be including more than one variable at a time, and it should also strengthen predictive accuracy with multiple dimension variables. One challenge in predictive modeling is to accommodate an imbalanced dataset in which some categories of satisfaction may contain fewer instances, causing biased predictions based on the absence of representations in frequency counts, while others categories contain a lot of cases (frequency) to draw predictive value from. Creating a predictive model that diminishes the challenges and weakness that have been outlined is the first step in helping businesses establish good, informed decision-making processes that enable them to successfully promote customer loyalty and enhance profits through today and into the future.

Research Objective

This study aims to develop a predictive model for classifying customer satisfaction using the Random Forest algorithm, which is a strong ensemble learning algorithm. Random Forest fits “many” decision trees during training time and “averages” their outputs to improve predictive performance and reduce overfitting to training data. Random Forest is a good option for analyzing e-commerce data because it fits both numerical and categorical data, handles missing data, and can fit complex, non-linear relationships. By fitting the Random Forest model to the E-commerce customer dataset collected, this study will examine the salient factors identifying customer satisfaction, assess model performance using accuracy and cross-validation statistics, and explore the importance of various features to predict customer satisfaction. There are a few components that go into the research method. First, there will ultimately be pre-processing of the data to train the model that will then go through feature importance analysis for determining what has the highest impact on customer satisfaction. Next, to gather even a classification process, the categorical variable will be encoded. Furthermore, a training set and a test set of data will need to be created from the dataset. The decision tree that will be used for this research is a Random Forest method that will be trained alongside evaluating the model through each of the metrics. Finally, visualizations such as confusion matrices and ROC curves will be provided as interpretations of the model predictive power. The purpose of this not only to have an accurate classification level of high, but to inform a business direction that is available to optimize their business practices and customer experience.

Significance of Study

This study's results are also important for e-commerce businesses since they provide a data-driven model to help explain and predict customer satisfaction. By discriminating customers based on their satisfaction levels, businesses can create tailored marketing campaigns to customer segments, like loyal, neutral, or even at-risk customers. (Munna, Rifat, & Badrudduza, 2020) Businesses can boost client involvement through promotional activities and customized suggestions which target neutral or dissatisfied customers. The identification of key satisfaction elements allows organizations to direct their operational changes and resource distribution effectively. (Diwakar, Kumar, Gour, & Khan, 2019)The model indicates Days Since Last Purchase and Total Spend as essential predictors which directs companies to create retention strategies and loyalty programs and personalized incentives for sustaining customer satisfaction. (Hossain, Rahman, Uddin, & Hossain, 2022)The analysis of feature importance enables product development and customer service improvement and user experience optimization. (Noor & Islam, 2019) (Singh & Sarraf, 2020) From a more high-level lens, this paper contributes to a growing trend of applications of fused machine learning in e-commerce analytics. The paper demonstrates the use of Random Forest for customer classification and emphasizes the tangible advantages of using powerful analysis of advanced concepts to solve knowledge adoption complexities in a business context. (Yi & Liu, 2020) The study provides a framework for other organizations that either are or are trying to develop predictive modeling for customer satisfaction experience, and the contributing findings can be translated for viable machine learning applications in marketing, sales, and operations. The study also has implications for customer relationship management (CRM) systems, which are increasingly relying on advanced predictive analytics

for decision-making. (Karn, Karna, Kondamudi, Bagale, Pustokhin, Pustokhina, & Sengan, 2022) Customer satisfaction predictive modeling will lead to the ability to make proactive decisions like sending communications before customers have problems, sending them promotions based on their interests or preferences, or helping them prior to their request escalating to customer service. All of these experience provide long-term customer retention and profit opportunities through proactive decision making. E-commerce has been developing and maintaining huge amounts of structured and unstructured customer data, providing e-commerce companies the ability to use machine-learning algorithms to their advantage by providing improved responses to new customer needs and by responding to changing market conditions. The e-commerce industry faces both challenges and opportunities while leveraging the data generated by its customers to enhance satisfaction and retention levels . Large datasets enable organizations to understand customer preferences and actions through their extensive data collection yet customer behavior remains complex which requires sophisticated analytical techniques to discover valuable patterns in customer data. The project addresses the challenge of classifying customer satisfaction from both transaction and behavioral data, and proposes one potential solution to classification, the Random Forest algorithm. (Shrirame, Sabade, Soneta, & Vijayalakshmi, 2020)The study's proposed research will develop an effective predictive model that classifies customer satisfaction by identifying key drivers of satisfaction for marketing, retention, and personalization purposes. As a result, the study's research will help to contribute to academic understanding as well as business application in the e-commerce analytic domain by providing a useful tool for improving customer experience and fostering sustainable growth in e-commerce through satisfaction suggestions.

II. LITERATURE REVIEW

In recent years, machine learning applications are critical to understanding e-commerce customer behavior and predicting if customers will be satisfied, segment customers, or trigger optimal marketing. Eshra (2021) examined customer segmentation and classification using two approaches including product purchased history and Recency, Frequency, and Monetary (RFM) metrics. Eshra (2021) used K-Means clustering along with 8 classification models (Random Forest, Logistic Regression, Gradient Boosting, etc.) and achieved over 90% accuracy with predicting Customer Class. This illustrates an importance of both transactional patterns and previous purchase history to understanding customer behavior.

Also, Than Than Win & Khin Sundee Bo (2020) studied Customer Lifetime Value to predict Customer Class as well. The research and implementation used Random Forest models with hyperparameter tuning to prove that CLV predictive models were a useful tool to allocate resources in Customer Relationship Management. The research was primarily based on monetary features and did not directly address if customers were satisfied.

Alghazzawi et al. (2023) introduced a technique that utilized an ensemble of Random Forest and XGBoost (ERF-XGB) in the task of classifying sentiments in product reviews. The results on data such as IMDB and ChnSentiCorp produced accuracies above 98%. While a strength of the study was demonstrating the predictive capacity of ensemble-based machine learning methods for textual inputs and recognizing sentiment polarity, it does not specifically conduct customer satisfaction classifications using behavioral and transactional data.

Wong and Marikannan (2020), however, began to specifically examine e-commerce customer satisfaction by comparing different techniques, including decision trees, random forest, artificial neural networks, and support vector machines across three years of historical data from retailer transactions. The research noted that delivery time and order fulfillment influenced satisfaction illustrated by the Random Forest technique with the best accuracy. Although the authors provide useful insights, they identified the presence of

challenges that included imbalanced and highly skewed datasets, and early on, a portion of their study suggested limiting interpretability of the features and distributions provided inadequate evaluations.

Ghosh and Banerjee (2020) utilized a modified Random Forest algorithm to investigate customer purchasing behavior in cloud services. Their model incorporates multiple variables, such as prior purchases, advertisement exposure, and location and achieved a high level of accuracy in prediction. As an example of the Random Forest's versatility, the study is once again context-specific and does not make claims about generalizability to online purchasing behavior in broader e-commerce contexts.

Weiwei Wei (2024) used a Random Forest recommendation algorithm to capture user behavior across the major Chinese e-commerce platforms. With an accuracy of 87.5% in prediction, the experiment showed the feasibility of using this model to recognize user discoverer intention. However, the dataset has performance deficiencies compared to more sophisticated ensemble approaches and developing predictive models involving a combination of behavioral and purchase transactional data appears to capture the purchase processes more efficiently when privacy is managed appropriately.

Finally, Lilhore et al. (2021) presented a hybrid weighted Random Forest model that utilized tree-level weights in conjunction with the C4.5 algorithm to predict online purchases. The hybrid model achieved increased accuracy over the standard Random Forest model; however, the additional complexity of the hybrid model required more sophisticated computational capacity, which hindered interpretability into actionable business processes.

Literature Matrix:

No.	Study	Objective	Dataset / Domain	Methodology / Algorithm	Key Findings / Accuracy	Limitations / Gaps
1	Eshra, 2021	Classify online retail customers based on purchase history and RFM (Recency, Frequency, Monetary)	Retail data	K-Means Clustering + 8 classification models (Logistic Regression, Gradient Boosting, Random Forest)	Predicted customer classes with >90% accuracy	Focused on segmentation; limited exploration of feature importance and model interpretability
2	Than Than Win & Khin Sundee Bo, 2020	Predict customer class using Customer Lifetime Value (CLV)	Super Store Retail dataset (~10,000 transactions)	Random Forest + Random Search tuning	Improved predictive accuracy compared to AdaBoost	Limited to CLV-based features; did not consider satisfaction explicitly
3	Alghazzawi et al., 2023	Sentiment analysis of e-commerce product reviews	IMDB and ChnSentiCorp datasets	Ensemble RF-XGBoost (ERF-XGB)	Accuracy: 98.7% (ChnSentiCorp), 98.2% (IMDB)	Focused on sentiment polarity; not directly on customer

						satisfaction classification
4	Wong & Marikannan, 2020	Identify factors influencing e-commerce customer satisfaction	E-commerce retailer data (3-year historical data)	Decision Tree, Random Forest, ANN, SVM	Random Forest performed best; top factors: delivery time, order fulfillment	Dataset challenges: imbalance, skewness; limited exploration of behavioral features
5	Ghosh & Banerjee, 2020	Predict customer purchase behavior in cloud services	Advertisement log dataset	Modified Random Forest	High accuracy in predicting purchase behavior	Domain-specific (cloud services); less focus on general e-commerce context
6	Weiwei Wei, 2024	Analyze user behavior on e-commerce platforms	Chinese e-commerce platforms (Taobao, Jingdong, Tiktok)	Random Forest Recommendation Algorithm	Accuracy: 87.5%; feasible for user intention recognition	Accuracy lower than other ensemble approaches; limited to user behavior metrics
7	Lilhore et al., 2021	Predict online purchasing behavior	Large-scale online shopping data	Weighted Random Forest (Hybrid with C4.5)	Improved prediction compared to standard RF	Complexity increased due to weighting; limited interpretability for business insights

Literature Gaps:

Although there is empirical evidence of the usefulness of machine learning, particularly Random Forest and ensemble-based models in the forecasting of customer behavior and satisfaction, many gaps remain in the literature. For example, most studies either look at segmentation, purchase prediction, or sentiment analysis of reviews, but only a few have considered a complete set of behavioral, transactional, and demographic features to provide a holistic classification of customer satisfaction. Relatedly, the fields of feature importance and interpretability have only minimally been explored, constraining potential practical insights to guide data-informed decision-making in an e-commerce context. Additionally, some studies considered imbalanced datasets and multi-class customer satisfaction classification with a strong evaluation/validation protocol with respect to cross-validation and ROC-AUC metrics, and limited studies examining an efficacious evaluation and validation protocol is needed to achieve reliable predictions in varied contexts. Lastly, some studies are domain specific (i.e., cloud services or product reviews), limiting

the extent of findings and implications that could be made for the broader e-commerce literature space. Hence, it is necessary to develop a predictive model that can reliably classify customer satisfaction, but which additionally is trained on relevant features that identify potential actions to inform future marketing zones and engaged strategies for customer retention to enhance revenue growth and optimization.

III. DATASET DESCRIPTION

This research utilizes a synthetic e-commerce dataset obtained from Kaggle, which is structured as 350 records indicating 11 features: Customer ID, Gender, Age, City, Membership Type, Total Spend, Items Purchased, Average Rating, Discount Applied, Days Since Last Purchase, and Satisfaction Level. The prediction target variable is Satisfaction Level, which has three classes: Neutral, Satisfied, and Unsatisfied. During the data preprocessing stage, it was determined that there were 2 records that had missing values specifically in the Satisfaction Level column and were determined to utilize mode imputation to handle the missing values, thus making the dataset whole for modeling, without bias and data loss.

IV. METHODOLOGY

The methodology employed in this research aims to methodically develop a discretionary and highly predictive forthcoming model for the classification of customer satisfaction level in the context of e-commerce. The first step toward building the forthcoming model is preprocessing the dataset, which is necessary to ensure the data is cleaned, consistent, and organized for the machine learning analysis. In this study, the dataset consisted of 350 records, eleven features, and two missing values in the variable, Satisfaction Level. Mode imputation was performed for the missing values, as it replaces missing entries with the most common class while preserving the original distribution in the term of the target variable to reduce bias. All categorical variables (e.g., Gender, City, Membership Type, Satisfaction Level) were transformed into integer values because of machine learning algorithms' (e.g., Random Forest) requirement for numeric data to be processed effectively.

The process continued after data cleaning and encoding with feature selection as the following step. The entire dataset features proved useful for forecasting customer satisfaction but Customer ID proved useless because it functions solely as an identification number. The model concentrates on features which have direct impact on customer behavior through Total Spend, Items Purchased, Average Rating, Discount Applied and Days Since Last Purchase.

The research team chose Random Forest Classifier as their model selection tool because this method provides reliable classification results through its ensemble learning approach. Random Forest (RF) was chosen as an approach to build an ML model because RF is capable of effectively modeling complex datasets through the use of a group of trees (200 trees) working together, as opposed to a single decision tree; RF trees are able to detect and capture nonlinear relations between features while also providing an indication of the relevance of features to model predictions. To train and evaluate the Random Forest model's performance, the data was split into training and testing datasets using an 80/20 ratio, maintaining a stratified approach to ensure balanced representation of the three categories of satisfaction (Neutral, Satisfied, and Unsatisfied) in both datasets. Preserving the original distribution of class labels allows for the model to avoid having a bias toward the majority class. To comprehensively assess the Random Forest model's classification performance, various performance measures were utilized, including: Overall Accuracy - how well the model performed overall; Precision - the number of correct True Positives divided by the total number of predicted positives; Recall - the number of correct True Positives divided by the total number of actual positives; and F1-score - the harmonic mean of Precision and Recall which accounts for both false positives and false negatives.

The model's prediction performance became visible through the confusion matrix which showed both accurate and inaccurate predictions for each class to reveal the model's specific error patterns. The model underwent stability and generalizability testing through 5-fold cross-validation which divides the dataset into five equal parts for training and validation during five separate iterations that use different subsets for validation. The method shows how well the model performs when trained on different data segments.
The evaluation of multiclass ROC-AUC analysis helped determine how well the model distinguishes between the three satisfaction levels. The ROC-AUC (Receiver Operating Characteristic - Area Under Curve) metric evaluates how well a classifier can order positive instances above negative ones which creates a solid evaluation method for models with multiple classes. The methodology achieves an accurate customer satisfaction prediction model for e-commerce platforms through the integration of preprocessing with feature selection and model training and evaluation metrics and cross-validation and ROC-AUC analysis.

V. RESULTS

The dataset used for this research included 350 records which contained 11 features that represented demographic details and behavioral patterns and transactional data of e-commerce customers. The dataset included Customer ID and Gender and Age and City and Membership Type and Total Spend and Items Purchased and Average Rating and Discount Applied and Days Since Last Purchase and Satisfaction Level (Neutral, Satisfied, Unsatisfied) as the target variable. The Satisfaction Level field in the dataset contained two missing values out of 350 records which required mode imputation to maintain the target variable distribution. The dataset became prepared for model training after categorical features received numeric value encoding. Table 1 presents an overview of all dataset features together with their corresponding missing value counts.

Table 1: Dataset Summary

Feature	Data Type	Non-Null Count	Missing Values
Customer ID	int64	350	0
Gender	object	350	0
Age	int64	350	0
City	object	350	0
Membership Type	object	350	0
Total Spend	float64	350	0
Items Purchased	int64	350	0
Average Rating	float64	350	0
Discount Applied	bool	350	0
Days Since Last Purchase	int64	350	0
Satisfaction Level	object	348	2

The dataset underwent preprocessing and encoding before it got divided into two parts: a training set containing 280 records and a testing set containing 70 records through an 80/20 stratified split which preserved the original distribution of satisfaction classes.

Random Forest Model Performance

A Random Forest model was built with 200 estimators and the maximum depth of 10 levels. The model achieved an overall accuracy of 98.57% on the testing set. Table 2 presents the complete classification metrics that apply to every class.

Table 2: Classification Report

Class	Precision	Recall	F1-Score	Support

Neutral	0.96	1.00	0.98	22
Satisfied	1.00	0.96	0.98	25
Unsatisfied	1.00	1.00	1.00	23
Accuracy			0.99	70
Macro Average	0.99	0.99	0.99	70
Weighted Avg	0.99	0.99	0.99	70

The confusion matrix (Figure 1) confirms the high performance of the model, with only minor misclassifications for the Satisfied class.

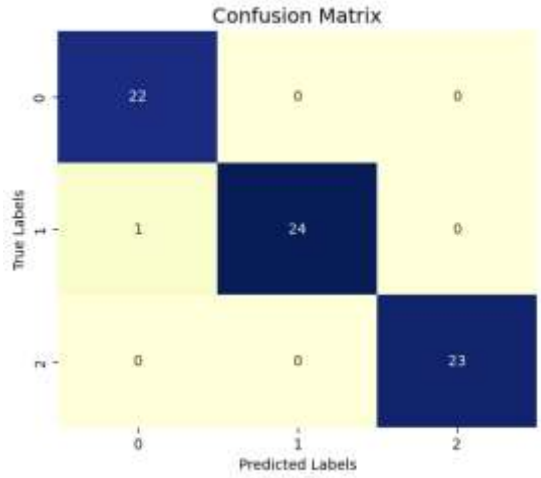

Figure 1 Confusion Matrix

Feature Importance

The Random Forest model provides a ranking of feature importance, helping identify which attributes most influence customer satisfaction predictions. Table 3 lists the top predictive features.

Table 3: Top Predictive Features

Feature	Importance
Days Since Last Purchase	0.2456
Total Spend	0.1998
Discount Applied	0.1322
Items Purchased	0.1293
Average Rating	0.1205

The analysis shows that Days Since Last Purchase is the most important feature, while demographic features such as Gender had the least influence on predictions.

Cross-Validation and ROC-AUC

To assess model robustness, 5-fold cross-validation was conducted. The model showed perfect accuracy during the first fold but all remaining folds achieved 98.57% accuracy which resulted in an overall cross-validated accuracy of 98.86%.

Furthermore, the mean ROC-AUC score for the multiclass classification was 0.9963, indicating an excellent ability of the model to distinguish between Neutral, Satisfied, and Unsatisfied classes.

Table 4: Cross-Validation and ROC-AUC Metrics

Metric	Value
Mean Cross-Validation Accuracy	98.86%
Mean ROC-AUC Score	0.9963

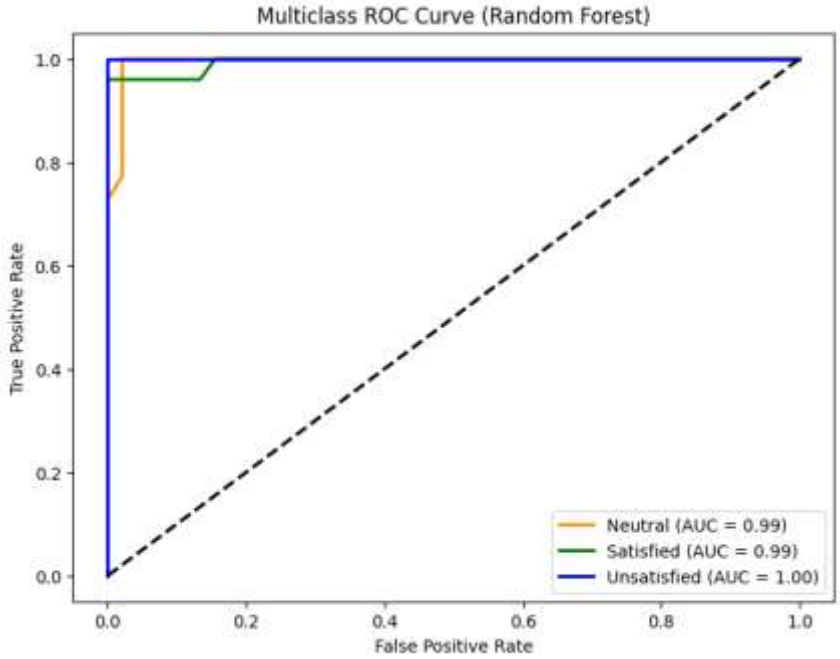

Figure 2 Multiclass ROC Curve (Random Forest)

VI. DISCUSSION

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

The research findings demonstrate that machine learning applications including the Random Forest classifier deliver strong performance in predicting customer satisfaction within e-commerce operations. The model demonstrated outstanding testing dataset performance with 98.57% accuracy and achieved 98.86% accuracy through cross-validation which shows its ability to perform consistently across various data segments. The model shows outstanding performance through a mean ROC-AUC score of 0.9963 which proves its ability to separate Neutral, Satisfied and Unsatisfied satisfaction levels and Random Forest method delivers reliable multi-class classification results for this task. The model achieves outstanding predictive performance through its ability to reduce classification mistakes which is shown by its high precision and recall scores and F1 scores across all classes. The comprehensive analysis of

feature importance indicated that behavioral and transactional features, such as Days Since Last Purchase, and Total Spend had the highest effect on predicting customer satisfaction, while demographic features like Gender had the least effect. This conclusion is in accordance with prior research findings in the domain, which suggests that it is the customer's behavior and the dynamics of their engagement that are more predictive of satisfaction than any static demographic features. To put it another way, customers who shop more often and/or are fluent in their engagement with returning purchases are more likely to be satisfied. Therefore, the e-commerce company has shown that tracking these behaviors with the Random Forest model is useful for targeting customers who may be at risk of being unhappy or dissatisfied with their purchases. Instead of relying just on demographic segmentations based on basic characteristics to reach out to customers, e-commerce businesses need to leverage this information to identify customer engagement opportunities. Using the Random Forest algorithm also helped the company through the challenges that e-commerce companies often face with class imbalances, skewed distributions, and noisy data. In addition to providing improved stability and accuracy, Random Forest also reduces overfitting, allowing it to learn and capture intricate relationships between variables.

The Random Forest model is particularly appropriate for e-commerce because it handles high-dimensional data very well, especially when it includes large transaction datasets with multiple attributes associated with customer interactions and product interactions along with metrics for customer engagement. Utilization of stratified sampling to separate training and validation data enabled more reliable evaluation metrics and provided e-commerce companies with additional methods for mitigating bias in the assessment of their models. The Random Forest model's high performance level demonstrates both its capability to produce valid predictions and the opportunity for its use to create market-specific marketing, promotion and retention strategies for e-commerce retailers. For example, knowing about probable dissatisfaction can facilitate outreach for retention by considering loyalty programs, sending a timely offer, or communicating something about service delivery to enhance the customer experience and improve the possibility of retaining the individual. In addition to this, knowing the primary sources of satisfaction, which are part of the feature importance analysis, allows companies to spend their resources wisely and simply focus on the prioritized components of customer perception; for example, delivery times or recent purchases, or incentives to spend additional capital are influences on customer satisfaction. When comparing these findings with prior studies, it strengthens the validity of machine learning approaches for customer satisfaction prediction. Mohamed Abdellatif Eshra's (2021) research identified the effectiveness of Random Forest and other ensemble models in segmenting and predicting customers classes based on, for example, purchase history and RFM (Recency, Frequency, Monetary) information. Relatedly, Than Win and Khin Sundee Bo (2020) and Wong and Poolan Marikannan (2020) studies also explained both the effectiveness of Random Forest models and their superior performance to other classification algorithms, especially in high-dimensional complex and imbalanced e-commerce datasets. The current research supports these research findings, and extends their application to synthetic data representing a modern e-commerce transaction, thus supporting the generalizability and use of Random Forest in an e-commerce business context. The findings have provided insight into how behavioral and transaction data generate predictive indicators for customer satisfaction, as well as how machine learning ensemble techniques (Random Forests) are used to develop e-commerce analytics. The predictions produced by the developed algorithm showed high accuracy and excellent cross-validation, which indicates that businesses can use the results of their predictive modelling to enhance customer experiences and marketing strategies and make decisions informed by data. The use of essential features for prediction will enable businesses to obtain usable insights, which will enable them to retain customers and create loyal customers, enabling them to be successful in the rapidly evolving e-commerce sector. Therefore, this research supports the

increasing amount of evidence for the use of machine learning techniques within customer relationship management practices and offers practical insight into how to connect data-based approaches to both customer satisfaction and sustainable business value.

I. CONCLUSION

This study illustrates how machine learning may be used to predict customer satisfaction through the Random Forest Classifier, and how this use of machine learning could benefit e-commerce businesses (retailers). Using a synthetic dataset of 350 customers with demographic, transactional and behavioural information, the researchers produced a strong predictive model that accurately classifies customers into Neutral (N), Satisfied (S) and Unsatisfied (U). The predictive model has an overall accuracy of 98.57% with validation accuracy of 98.86%. The average ROC-AUC score of 0.9963 across all classes represents excellent discriminative and classification ability of the predictive model and supports the potential for the predictive model to be used on more expansive and diverse geographical areas in e-commerce as an evolving field of business. The predictors identified as important for determining customer satisfaction were behavioral-based, specifically the variable "Days Since Last Purchase" and the variable "Total Spend." Whereas gender was determined to be a lesser predictive variable. The study identified that e-commerce businesses could use the findings of the predictive model as a means to develop customer retention and marketing strategies to encourage continued engagement based on customer activity, Days Since Last Purchase engagement metrics, and transaction attributes. Random Forest is especially good at handling some of the biggest challenges that e-commerce datasets face, including issues with class imbalance, an abundance of dimensionality, and potential noise, due to Random Forest's ensemble structure and its ability to withstand overfitting. As a result, incorporating stratified testing on train-test splits, using cross-validation, and measuring performance using different metrics allow for a credible evaluation of the model performance and the availability of the model across multiple datasets.The research ultimately confirms that machine learning models can be utilized to enhance the customer experience, assist with business decision-making, and facilitate data-driven decision-making in online retail.The research also supports the growing body of literature about the application of behavioral data in customer satisfaction and provides actionable strategies for businesses to establish long-term loyalty from current customers, maintain retention and improve their competitive position in the rapidly evolving e-commerce landscape.

REFERENCES

[1] Bernat, J.R., Koning, A.J., & Fok, D., 2019. Modelling customer lifetime value in a continuous, noncontractual Time Setting. Netherlands.

[2] Chamberlain, B.P., Cardoso, A., Liu, C.H.B., Pagliari, R., & Deisenroth, M.P., 2017. Customer lifetime value prediction using embeddings. 23rd ACM SIGKDD International Conference, Canada.

[3] Chen, S., 2018. Estimating customer lifetime value using machine learning techniques. London.

[4] Data Science Group, Amperity Inc., 2019. Predicting Customer Lifetime Value with Unified Customer Data.

[5] Farzanfar, E., & Delafrooz, N., 2016. Determining the Customer Lifetime Value based on the Benefit Clustering in the Insurance Industry. Indian Journal of Science and Technology, India.

[6] Jangid, C., Kothari, T., Spear, J., & Wadsworth, E., 2014. Custoval: estimating customer lifetime value using machine learning techniques. Dept. of CIS-Senior Design.

[7] Jasek, P., Vrana, L., Sperkova, L., Smutny, Z., & Kobulsky, M., 2019. Modeling and application of customer lifetime value in online retail. Informatics 5, India.

[8] Karlssson, M., 2016. Predicting customer lifetime value using machine learning algorithms.

[9] Glady, N., Baesens, B., & Croux, C., 2008. Modeling churn using customer lifetime value. KU Leuven KBI Working Paper, Belgium.

[10] Rathi, T., 2011. Customer lifetime value measurement using machine learning techniques. IGI Global, USA.

[11] Vanderveld, A., Pandey, A., Han, A., & Parekh, R., 2016. An engagement-based customer lifetime value system for e-commerce. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York.

[12] Munna, M.H., Rifat, M.R.I. & Badrudduza, A.S.M., 2020. Sentiment analysis and product review classification in e-commerce platform. In: Proceedings of the 2020 23rd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 19–21 December 2020. IEEE, Piscataway, NJ, USA, pp.1–6.

[13] Diwakar, D., Kumar, R., Gour, B. & Khan, A.U., 2019. Proposed machine learning classifier algorithm for sentiment analysis. In: Proceedings of the 2019 Sixteenth International Conference on Wireless and Optical Communication Networks (WOCN), Bhopal, India, 19–21 December 2019. IEEE, Piscataway, NJ, USA, pp.1–6.

[14] Noor, A. & Islam, M., 2019. Sentiment Analysis for Women's E-commerce Reviews using Machine Learning Algorithms. In: Proceedings of the 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 6–8 July 2019. IEEE, Piscataway, NJ, USA, pp.1–6.

[15] Singh, S.N. & Sarraf, T., 2020. Sentiment analysis of a product based on user reviews using random forests algorithm. In: Proceedings of the 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 29–31 January 2020. IEEE, Piscataway, NJ, USA, pp.112–116.

[16] Yi, S. & Liu, X., 2020. Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review. Complex & Intelligent Systems, 6, pp.621–634. DOI: [CrossRef]

[17] Hossain, M.S., Rahman, M.F., Uddin, M.K. & Hossain, M.K., 2022. Customer sentiment analysis and prediction of halal restaurants using machine learning approaches. Journal of Islamic Marketing, ahead-of-print. DOI: [CrossRef]

[18] Karn, A.L., Karna, R.K., Kondamudi, B.R., Bagale, G., Pustokhin, D.A., Pustokhina, I.V. & Sengan, S., 2022. Customer centric hybrid recommendation system for E-Commerce applications by integrating hybrid sentiment analysis. Electronic Commerce Research, 23, pp.279–314. DOI: [CrossRef]

[19] Shrirame, V., Sabade, J., Soneta, H. & Vijayalakshmi, M., 2020. Consumer Behavior Analytics using Machine Learning Algorithms. In: Proceedings of the 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 2–4 July 2020. IEEE, Piscataway, NJ, USA, pp.1–6.

[20] Eshra, M.A., 2021. Classification of Online Retail Customers using Machine Learning Techniques. Master's thesis, International University of Valencia, Madrid.

[21] Than, T.W. & Bo, K.S., 2020. Predicting Customer Class using Customer Lifetime Value with Random Forest Algorithm. 2020 International Conference on Advanced Information Technologies (ICAIT), 4–5 November 2020, Yangon, Myanmar. IEEE. DOI: 10.1109/ICAIT51105.2020.9261792

[22] Alghazzawi, D.M., Alquraishee, A.G.A., Badri, S.K. & Hasan, S.H., 2023. ERF-XGB: Ensemble Random Forest-Based XG Boost for Accurate Prediction and Classification of E-Commerce Product Review. Sustainability, 15(9), p.7076. DOI: 10.3390/su15097076

[23] Wong, A.-N. & Marikannan, B.P., 2020. Optimising e-commerce customer satisfaction with machine learning. Journal of Physics: Conference Series, 1712, 012044. DOI: 10.1088/1742-6596/1712/1/012044

[24] Ghosh, S. & Banerjee, C., 2020. A Predictive Analysis Model of Customer Purchase Behavior using Modified Random Forest Algorithm in Cloud Environment. 2020 IEEE 1st International Conference for Convergence in Engineering (ICCE), 5–6 September 2020, Kolkata, India. IEEE. DOI: 10.1109/ICCE50343.2020.9290700

[25] Wei, W., 2024. User Behavior Analysis of E-commerce Platforms Under Random Forest Recommendation Algorithm. 2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT), 15–16 March 2024, Bengaluru, India. IEEE. DOI: 10.1109/ICDCOT61034.2024.10516196

[26] Lilhore, U.K., Simaiya, S., Prasad, D. & Verma, D.K., 2021. Hybrid Weighted Random Forests Method for Prediction & Classification of Online Buying Customers. Journal of Information Technology Management, [online] DOI: 10.22059/jitm.2021.310062.2607

https://www.kaggle.com/datasets/uom190346a/e-commerce-customer-behavior-dataset

```
# ============================================
# Import Libraries
# ============================================
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix, roc_curve, auc
from sklearn.preprocessing import label_binarize
from sklearn.multiclass import OneVsRestClassifier

# ============================================
# Load Dataset
# ============================================
df = pd.read_csv("E-commerce Customer Behavior.csv")

print("Dataset Overview:")
print(df.info())
print("\nDataset Shape:", df.shape)
print("\nMissing Values:\n", df.isnull().sum())

# ============================================
# Data Cleaning
# ============================================
# Fill missing values for 'Satisfaction Level'
df['Satisfaction Level'] = df['Satisfaction Level'].fillna(df['Satisfaction Level'].mode()[0])

# ============================================
# Encoding Categorical Columns
# ============================================
categorical_cols = ['Gender', 'City', 'Membership Type', 'Satisfaction Level']
for col in categorical_cols:
    df[col] = df[col].astype('category').cat.codes

print("\nData Encoding Complete!")
print(df.head())

# ============================================
# Feature and Target Split
# ============================================
X = df.drop(columns=['Customer ID', 'Satisfaction Level'])
y = df['Satisfaction Level']
```

```
# Train-test split
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)

print("\nTraining Set Shape:", X_train.shape)
print("Testing Set Shape:", X_test.shape)

# ==========================================
# Random Forest Model Training
# ==========================================
model = RandomForestClassifier(
    n_estimators=200,
    max_depth=10,
    random_state=42,
    class_weight='balanced'
)
model.fit(X_train, y_train)
print("\nRandom Forest Model Trained Successfully!")

# ==========================================
# Model Evaluation
# ==========================================
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred) * 100

print(f"\nModel Accuracy: {accuracy:.2f}%")
print("\nClassification Report:")
print(classification_report(y_test, y_pred, target_names=['Neutral', 'Satisfied', 'Unsatisfied']))

# ==========================================
# Confusion Matrix Visualization
# ==========================================
plt.figure(figsize=(6,5))
sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt='d', cmap='YlGnBu', cbar=False)
plt.title("Confusion Matrix", fontsize=14)
plt.xlabel("Predicted Labels")
plt.ylabel("True Labels")
plt.show()

# ==========================================
# Feature Importance Visualization
# ==========================================
importances = model.feature_importances_
feature_importance = pd.DataFrame({
    'Feature': X.columns,
    'Importance': importances
}).sort_values(by='Importance', ascending=False)
```

```
plt.figure(figsize=(8,5))
sns.barplot(data=feature_importance, x='Importance', y='Feature', hue='Feature', legend=False, palette='viridis')
plt.title("Feature Importance (Random Forest)", fontsize=14)
plt.xlabel("Importance Score")
plt.ylabel("Feature")
plt.show()

print("\nTop Predictive Features:")
print(feature_importance.head())

# ==========================================
# Cross-Validation for Robustness
# ==========================================
cv_scores = cross_val_score(model, X, y, cv=5, scoring='accuracy')
print(f"\nCross-Validation Accuracy Scores: {cv_scores}")
print(f"Mean CV Accuracy: {cv_scores.mean() * 100:.2f}%")

# ==========================================
# ROC Curve and AUC (Multiclass)
# ==========================================
# Binarize the output for multiclass ROC
y_test_bin = label_binarize(y_test, classes=[0, 1, 2])
n_classes = y_test_bin.shape[1]

# One-vs-Rest classifier for ROC
rf_ovr = OneVsRestClassifier(RandomForestClassifier(random_state=42, n_estimators=200, max_depth=10))
rf_ovr.fit(X_train, label_binarize(y_train, classes=[0, 1, 2]))
y_score = rf_ovr.predict_proba(X_test)

# Compute ROC curve and AUC for each class
fpr, tpr, roc_auc = dict(), dict(), dict()
for i in range(n_classes):
    fpr[i], tpr[i], _ = roc_curve(y_test_bin[:, i], y_score[:, i])
    roc_auc[i] = auc(fpr[i], tpr[i])

# Plot ROC Curves
plt.figure(figsize=(8,6))
colors = ['darkorange', 'green', 'blue']
class_names = ['Neutral', 'Satisfied', 'Unsatisfied']

for i, color in zip(range(n_classes), colors):
    plt.plot(fpr[i], tpr[i], color=color, lw=2, label=f"{class_names[i]} (AUC = {roc_auc[i]:.2f})")

plt.plot([0,1],[0,1],'k--', lw=2)
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Multiclass ROC Curve (Random Forest)')
```

```
plt.legend(loc="lower right")
plt.show()

# Mean ROC-AUC
mean_auc = np.mean(list(roc_auc.values()))
print(f"\nMean ROC-AUC Score: {mean_auc:.4f}")

# ============================================
# Final Summary
# ============================================
print("\nSummary:")
print(f'Final Model Accuracy: {accuracy:.2f}%")
print(f'Cross-Validated Accuracy: {cv_scores.mean() * 100:.2f}%")
print(f'Mean ROC-AUC Score: {mean_auc:.4f}")
print(f'Most Important Feature: {feature_importance.iloc[0, 0]}")
print(f'Least Important Feature: {feature_importance.iloc[-1, 0]}")
```