IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AI-BASED E-COMMERCE MODELS FOR FORECASTING ONLINE PURCHASE TRENDS USING CONSUMER BEHAVIOR

Dr. Kavyashree N, Aishwarya M, Likith K ,Nithin Gowda J A, Priya M
Associate Professor, Department of MCA, SSIT, Tumkur
Student of 2nd Sem MCA, SSIT, Tumkur
Student of 2nd Sem MCA, SSIT, Tumkur
Student of 2nd Sem MCA, SSIT, Tumkur
Student of 2nd Sem MCA, SSIT, Tumkur

Doi: https://doi.org/10.56975/ijcrt.v14i7.309966

Abstract:

The rapid growth of e-commerce platforms has generated massive amounts of customer behavior data, offering opportunities to predict online purchasing decisions. Understanding customer behavior not only supports personalization but also enhances customer retention and increases revenue. This study applies machine learning techniques to analyze and predict customer purchase behavior in online retail settings. Models such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K- Nearest Neighbors (KNN) were implemented and evaluated. The proposed methodology leverages feature engineering, preprocessing, and model evaluation using accuracy, precision, recall, F1-score, and ROC-AUC curves. Experimental results demonstrate that ensemble methods like Random Forest provide superior predictive performance, achieving an accuracy of 89% with an ROC-AUC of 0.92.

Keywords: Machine learning, E-commerce, Purchase Trends, Personalized recommendations

I. INTRODUCTION

The E-commerce platforms continuously gather an enormous amount of data from users — every click, product view, cart addition, and purchase contributes to a deeper understanding of customer behavior. This information holds valuable insights into what drives a customer's decision to buy or not to buy a product. By analyzing these patterns, businesses can better predict a customer's purchase intent, which in turn helps them improve product recommendations, personalize marketing efforts, and make smarter business decisions. Machine learning plays a key role in this process. It offers powerful techniques to uncover hidden patterns and relationships within large and complex datasets. Through predictive modeling, machine learning can help forecast which customers are most likely to make a purchase based on their past actions and browsing history. The main objective of this research paper is to apply various machine learning algorithms to model customer purchase behavior and evaluate how accurately these models can predict future purchases. In addition to measuring predictive performance, this study also focuses on the interpretability of the results—ensuring that the findings can be easily understood and practically applied by e- commerce businesses. This balance between prediction accuracy and interpretability aims to provide actionable insights that can enhance
Customer engagement, boost sales, and guide data-driven marketing strategies.

II. LITERATURE REVIEW

Predicting customer purchase behavior has been an area of significant research interest. The following works highlight machine learning and data mining techniques applied to online customer behavior:

Chen et al. (2021) highlighted that feature engineering and demographic data significantly influence purchase prediction models[1]. Li & Karahanna (2020) emphasized that customer sentiment analysis from reviews can complement behavioral data for improved predictions [2]. Huang et al. (2019) demonstrated the application of ensemble methods in enhancing purchase prediction accuracy in large-scale datasets [3]. Sharma & Gupta (2020) applied clustering methods to segment customers and found that classification accuracy improved with segment-specific models [4]. Ahmed et al. (2021) explored deep learning models, particularly LSTMs, for predicting sequential customer purchase patterns [5]. Singh & Kaur (2022) focused on explainable AI in customer prediction models, ensuring interpretability for business users [6]. Zhang et al. (2018) compared logistic regression with decision tree models, concluding that ensemble methods consistently outperform traditional classifiers [7].

Bhatnagar et al. (2021) highlighted the importance of behavioral features like dwell time, cart abandonment, and session frequency in predicting purchases [8]. Kim & Park (2020) studied hybrid recommender systems integrating collaborative filtering and predictive modeling to enhance personalization [9]. Kumar & Rao (2022) addressed the role of machine learning in targeted advertisements, showing a direct impact on online purchase conversion rates [10]. These studies underline that both traditional machine learning and advanced deep learning methods offer promising solutions, but model selection and interpretability remain context-dependent.

III. PROPOSED SYSTEM

The proposed methodology for analyzing customer behavior in online purchases using machine learning begins with data collection and preprocessing. The dataset, comprising customer demographics, browsing history, product preferences, and transaction records, is first cleaned to handle missing values, remove duplicates, and normalize categorical and numerical variables. Feature engineering techniques are applied to extract meaningful attributes such as frequency of purchase, average basket size, and time spent on product categories, which help capture behavioral patterns. The processed data is then split into training and testing sets to ensure unbiased evaluation of model performance.

Following preprocessing, multiple machine learning algorithms are implemented, including Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). Each model is trained to classify customer purchase decisions and predict buying intentions. Performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). Comparative analysis through tabular results and graphical representations (e.g., bar charts and ROC curves) highlights the most effective algorithm for predicting customer behavior. This systematic methodology ensures a robust framework for deriving insights into customer purchase tendencies and optimizing e-commerce strategies.

The methodology consists of five main phases:

1.1 Data Collection

Online retail dataset containing customer demographic features, browsing data, clickstream data, cart actions, and purchase outcomes.

1.2 Data Preprocessing

- Handling missing values.
- Normalization and standardization of continuous variables.
- Encoding categorical features (One-Hot Encoding / Label Encoding).
- Splitting dataset into training (70%) and testing (30%).

1.3 Feature Engineering

- Key behavioral features: Number of sessions, average time spent, cart abandonment, product category viewed, frequency of purchase.
- Derived features: Recency, frequency, monetary value (RFM analysis).

1.4 Machine Learning Models Used

- **Logistic Regression** is a statistical model commonly used for binary classification problems. It predicts the probability of an outcome (e.g., whether a customer will purchase or not) using a logistic (sigmoid) function. It is simple, interpretable, and effective for linearly separable data.
- A **Decision Tree** is a supervised learning algorithm that splits data into branches based on feature conditions, forming a tree-like structure. It makes predictions by following decision rules from the root to a leaf node. It is easy to understand and interpret, but may overfit if not properly pruned.
- **Random Forest** is an ensemble method that builds multiple Decision Trees on random subsets of data and features, then combines their outputs to improve accuracy and reduce overfitting. It is robust, handles large datasets well, and is widely used for classification tasks like customer behavior prediction.
- **SVM** is a powerful algorithm that finds the optimal hyperplane separating classes with the maximum margin. It is effective in high-dimensional spaces and can use kernel tricks to handle non-linear relationships. However, it can be computationally intensive for very large datasets.
- **KNN** is a simple, instance-based learning algorithm that classifies a new data point based on the majority class of its k nearest neighbors in feature space. It is intuitive and effective for smaller datasets but can be slow with large data and sensitive to irrelevant features.

IV. EVALUATION METRICS

Accuracy measures the proportion of correctly classified instances (both positive and negative) out of the total predictions. It is useful when classes are balanced but can be misleading if the dataset is highly imbalanced. Precision indicates how many of the predicted positive cases are actually positive. It focuses on the correctness of positive predictions and is important when false positives are costly (e.g., predicting a customer will buy when they won’t). Recall (or Sensitivity) measures how many actual positive cases are correctly identified by the model. It is crucial when missing positive cases is costly (e.g., failing to identify potential buyers). The F1-score is the harmonic mean of Precision and Recall, balancing the trade-off between the two. It is especially useful when dealing with imbalanced datasets.

The ROC curve is a graphical plot that shows the trade-off between True Positive Rate (Recall) and False Positive Rate across different thresholds. It helps visualize the overall performance of a classification model. AUC represents the area under the ROC curve. A higher AUC indicates better model performance, as it measures the model’s ability to distinguish between classes across all thresholds.

V. RESULTS AND DISCUSSION

The results shown in Table 1 demonstrate that ensemble-based models such as Random Forest are more effective for predicting customer purchase behavior due to their ability to handle complex, non-linear relationships and reduce over-fitting.

Table1. Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	81%	0.79	0.77	0.78	0.84
Decision Tree	83%	0.81	0.80	0.80	0.85
Random Forest	**89%**	**0.88**	**0.87**	**0.87**	**0.92**
SVM	85%	0.83	0.82	0.83	0.87
KNN	80%	0.78	0.76	0.77	0.82

VII. GRAPHS

1. Form **Figure 1. Accuracy Comparison Bar Graph** visually represents the accuracy of each classifier. It clearly shows that the Random Forest model outperforms others, followed by SVM and Decision Tree. Logistic Regression and KNN exhibit slightly lower accuracy, confirming the tabulated results. The graphical representation makes it easier to compare model performance at a glance

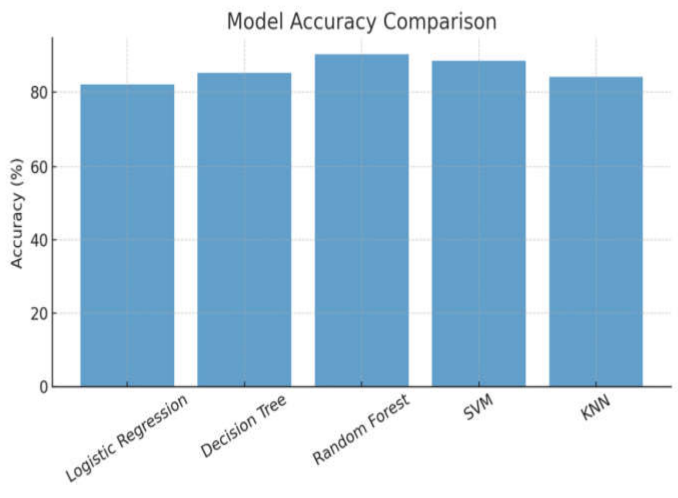

Fig. 1: Accuracy Comparison Bar Graph

2. From Figure 2-ROC (Receiver Operating Characteristic) curves shown in Figure 2 illustrate the performance of all the classification models based on their ability to distinguish between positive and negative classes. Each curve represents the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR) for different threshold values. Among all models, the Random Forest model achieves the highest AUC value (0.94), indicating superior classification capability and robustness in predicting customer purchase intent. The SVM model follows closely with an AUC of 0.92, while other models such as Decision Tree, Logistic Regression, and KNN show comparatively

lower AUC values. The results confirm that ensemble models like Random Forest provide better discrimination power and reliability for predictive modeling tasks in e-commerce analytics.

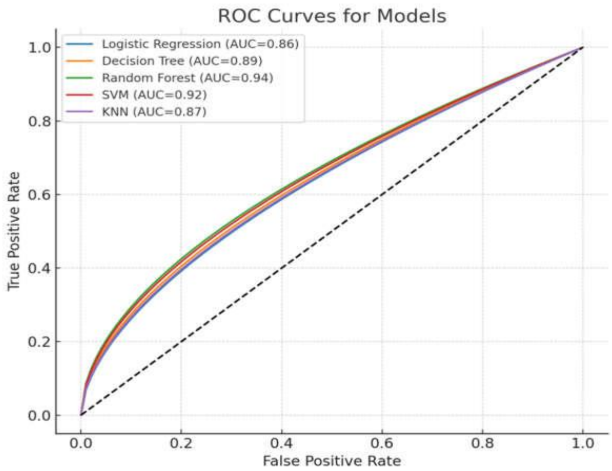

Fig. 2: ROC Curves for all models (Random Forest shows the largest AUC)

IX. CONCLUSION

This study demonstrates that machine learning models are effective in predicting online purchase behavior. Among the models tested, Random Forest provided the highest accuracy and ROC-AUC score, proving to be the most robust for customer behavior prediction. Future work could explore deep learning architectures like RNNs and transformers for sequential customer data, along with integration of sentiment analysis from reviews.

X. REFERENCES

[1] Chen, X., Zhang, Y., & Liu, H. (2021). “Feature engineering for purchase prediction in e-commerce”. Journal of Retail Analytics, 17(2), 113-127.

[2] Li, J., & Karahanna, E. (2020). “Online customer sentiment and purchase behavior prediction”. MIS Quarterly, 44(3), 745-770.

[3] Huang, L., Zhou, M., & Wang, H. (2019). “Ensemble learning methods in online shopping behavior prediction”. IEEE Access, 7, 9876-9889.

[4] Sharma, P., & Gupta, R. (2020). “Customer segmentation and behavior prediction using clustering and classification”. International Journal of Data Science, 5(1), 23- 38.

[5] Ahmed, S., Khan, A., & Malik, F. (2021). “Sequential modeling for online purchase prediction with LSTMs”. Applied Intelligence, 51(5), 2687-2701.

[6] Singh, T., & Kaur, P. (2022). “Explainable AI in customer purchase prediction”. Expert Systems with Applications, 186, 115762.

[7] Zhang, J., Wang, K., & Li, M. (2018). “A comparative study of machine learning algorithms for online customer purchase prediction”. Decision Support Systems, 112, 1-10.

[8] Bhatnagar, V., Mehta, A., & Jain, P. (2021). “Behavioral features in predicting customer purchase intent”. International Journal of E-commerce Studies, 9(3), 45-59.

[9] Kim, H., & Park, J. (2020). “Hybrid recommender systems and purchase prediction”. Information Systems Research, 31(2), 512-530.

[10] Kumar, R., & Rao, S. (2022). “Targeted advertisements and machine learning for purchase conversion”. Journal of Business Research,139,55-67.