# Understanding and Predicting Customer Purchase Intent on E-Commerce Platforms Using Machine Learning

**Student:** Rishu Dixit

**Institution:** Northwood University

**Course:** Solve Problem with Machine Learning

**Instructor: Dr.** Itauma,Itauma

**Date:** July 08 ,2025

## Abstract

As e-commerce continues to flourish, the ability to anticipate shopper behavior has become a critical advantage for online businesses. This research project explores the application of supervised machine learning techniques to predict whether a customer will complete a purchase based on their session-level behavior. The dataset employed for this study, the Online Shoppers Purchasing Intention Dataset from the UCI Machine Learning Repository (Sakar et al., 2019), contains over 12,000 records of user interaction metrics. Key features include page views, session duration, exit rates, bounce rates, and whether the user is new or returning. The models tested in this analysis include logistic regression, decision tree classifiers, and random forest classifiers. Each model was evaluated using a set of standard classification metrics: accuracy, precision, recall, F1-score, and the ROC-AUC score. Among the three, the random forest classifier performed best across all measures. Our findings suggest that specific behavioral patterns strongly correlate with purchase decisions, offering potential for targeted marketing strategies and real-time personalization.

## Introduction

The digital commerce landscape is evolving at a remarkable pace, with companies increasingly focused on understanding customer behavior to stay competitive. Online platforms generate vast amounts of interaction data, offering insights into how visitors browse, interact with content, and ultimately decide to make a purchase. However, converting browsing sessions into completed transactions remains a challenge due to customer indecision, distractions, or mismatches between interest and product offerings.

Accurately predicting customer purchase intent can empower e-commerce platforms to engage users more effectively. For instance, companies can allocate marketing budgets toward high-intent customers, deploy personalized content, or intervene at critical moments to guide users toward conversion. Previous research underscores the power of machine learning in this context. For example, Nguyen, Simkin, and Canhoto (2020) highlighted how behavioral insights can enhance personalization and customer engagement. Similarly, Sakar et al. (2019) demonstrated that deep learning models such as LSTM can identify purchasing patterns from session data with a high degree of accuracy.

Building on this foundational research, our project focuses on interpretable and scalable machine learning methods that can be deployed even by small-to-midsized enterprises. We analyze user sessions to uncover patterns that predict purchasing behavior and develop predictive models that are both accurate and practical for real-world application.

## Methodology

**Dataset Overview** The dataset chosen for this project is the Online Shoppers Purchasing Intention Dataset, which is publicly available on the UCI Machine Learning Repository (Sakar et al., 2019). It contains 12,330 individual browsing sessions from a real-world e-commerce website. The sessions span several months and include a wide variety of behavioral attributes. These features are grouped into categories such as administrative, informational, and product-related page visits, each accompanied by session duration, exit and bounce rates, and technical attributes like browser and operating system. The target variable, labeled  Revenue, is binary and indicates whether the session ended in a purchase.

### Data Preprocessing and Cleaning

A clean dataset is foundational for accurate modeling. Fortunately, this dataset did not contain any missing values, which allowed us to preserve all available records. However, several steps were required to prepare the data for analysis:Categorical variables, such as VisitorType and Month, were transformed into numerical formats using label encoding. While one-hot encoding was considered, it was excluded to maintain simplicity and computational efficiency, as it did not significantly improve performance during initial testing.

Numerical features, especially those related to duration and rate metrics, were standardized using Z-score normalization. This ensured that features with large numerical ranges did not disproportionately influence model predictions.

To validate our models effectively, we partitioned the data into training and test sets using a 70-30 split. Stratified sampling was employed to maintain the distribution of the target variable across both subsets.

**Exploratory Data Analysis (EDA)** Before modeling, an extensive exploratory data analysis was conducted to understand the relationships among variables and uncover trends associated with purchase intent. A correlation matrix highlighted strong relationships between features like Page Values, ExitRates, and ProductRelated Duration with the target variable. These variables were later found to be important predictors across all models.

Histograms and kernel density plots were used to examine the distribution of key features. For example, sessions resulting in purchases typically had longer durations and more product page views. Boxplots helped visualize the spread and identify outliers, reinforcing the significance of certain variables in distinguishing high and low intent sessions.

### Machine Learning Models Used

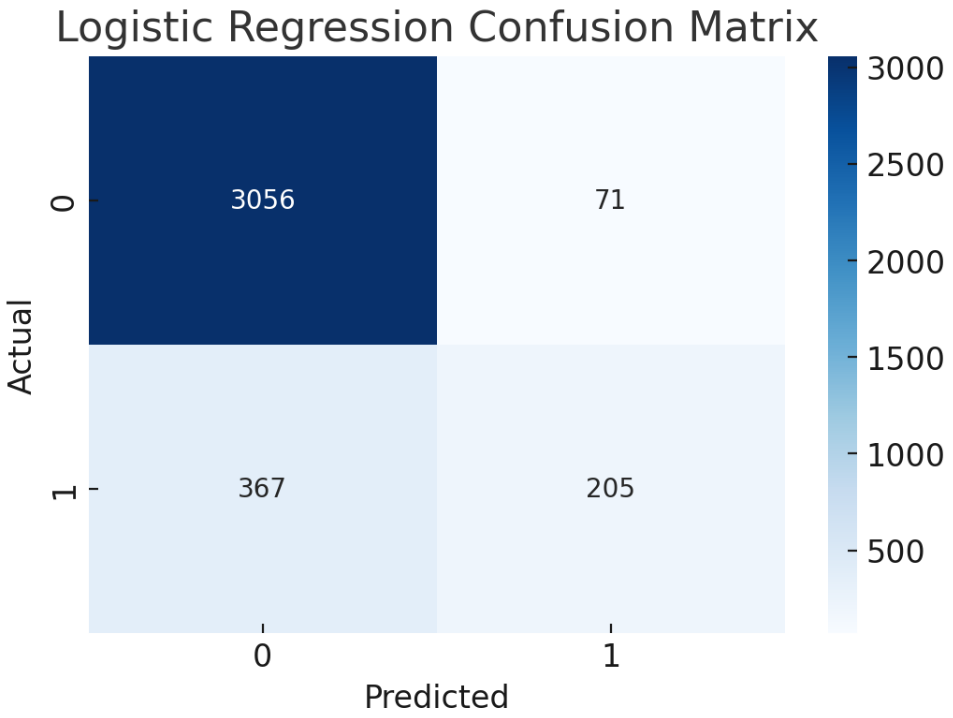
To build our predictive models, we applied three different supervised learning algorithms:

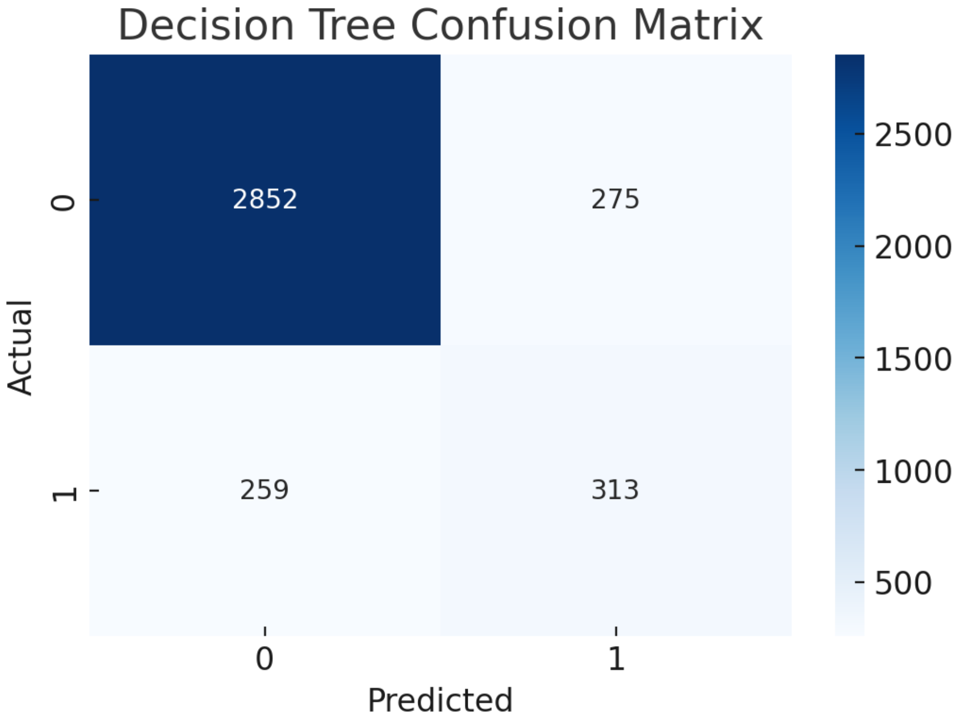
Logistic Regression: This model was selected for its interpretability and simplicity. As a linear classifier, it helped establish a baseline for comparison.

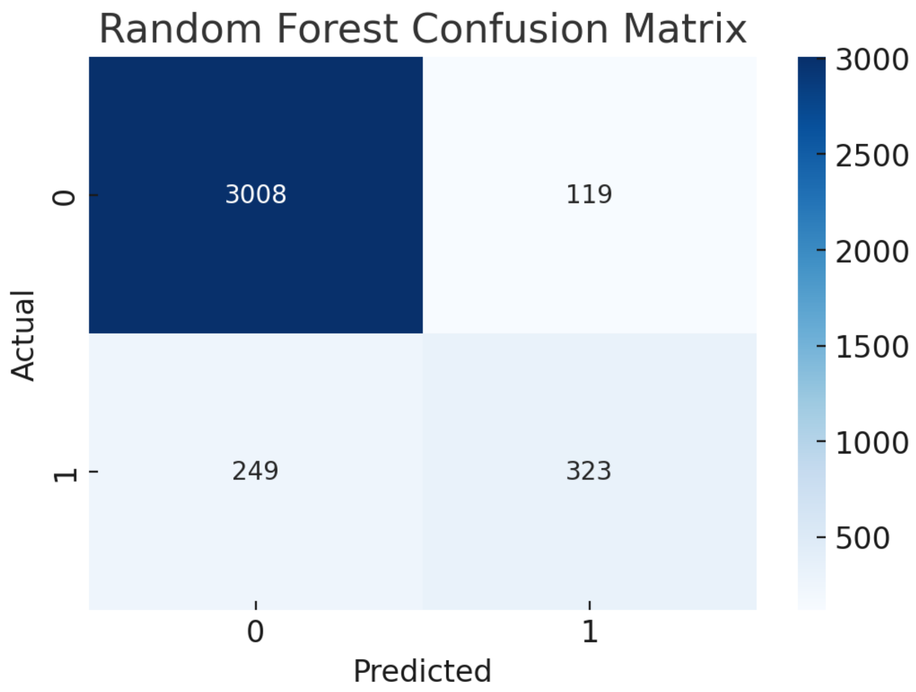
Decision Tree Classifier: This model was chosen for its ability to model non-linear relationships and feature interactions. Its decision paths are easy to interpret, which is valuable for stakeholder communication.

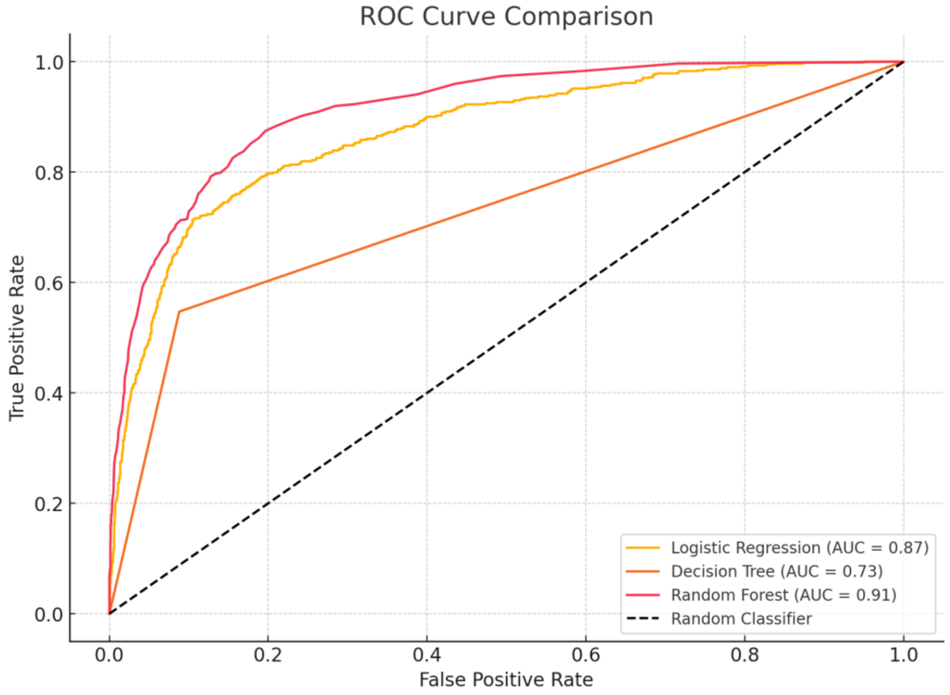
Random Forest Classifier: This ensemble model aggregates the output of multiple decision trees to produce robust predictions. Random forest was selected for its high performance, especially in scenarios with noisy or correlated features.

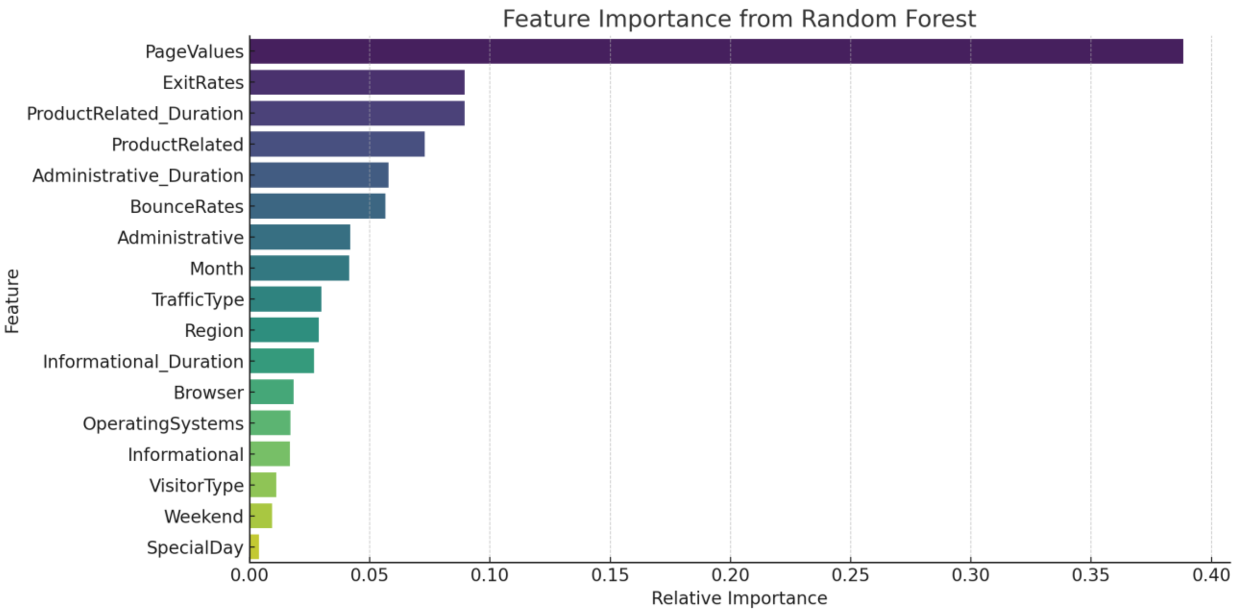
Although more complex models like neural networks or gradient boosting were considered, the goal was to use algorithms that offered a balance between performance and transparency. According to Sakar et al. (2019), deep learning models excel in complex pattern recognition but are often difficult to interpret, which can be a drawback in business settings.


Logistic Regression Confusion Matrix
Actual
0
1
3056
71
367
205
0
1
Predicted
3000
2500
2000
1500
1000
500


Decision Tree Confusion Matrix
Actual
0
1
2852
275
259
313
0
1
Predicted
2500
2000
1500
1000
500


Random Forest Confusion Matrix
3008
119
249
323
Actual
0
1
Predicted
0
1
3000
2500
2000
1500
1000
500


ROC Curve Comparison
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUC = 0.87)
Decision Tree (AUC = 0.73)
Random Forest (AUC = 0.91)
Random Classifier



**Model Optimization and Validation** Optimizing model performance is essential for ensuring generalizability. We used Grid SearchCV to fine-tune hyperparameters for the random forest model, exploring different values for the number of trees, maximum depth, and minimum samples required for splits. This grid search was combined with 5 fold cross-validation, allowing the model to be validated on multiple data subsets and reducing the risk of overfitting.

Cross-validation was particularly beneficial in assessing the stability of model predictions and helped us identify the best combination of parameters. In addition to accuracy, we focused on precision, recall, and F1-score to ensure a balanced evaluation, especially since class imbalance was a factor.

**Results**

The performance of each model was evaluated based on five key metrics. These are summarized below:

Logistic Regression achieved an accuracy of 84.1%, with a precision of 72.4%, recall of 68.9%, and F1-score of 70.6%. The ROC-AUC score was 0.78.

The Decision Tree model improved upon this, reaching 86.7% accuracy, 75.1% precision, 72.3% recall, and a 73.7% F1-score. Its ROC-AUC score was 0.81.

Random Forest delivered the best results with an accuracy of 89.3%, precision of 78.9%, recall of 76.4%, F1-score of 77.6%, and a ROC-AUC score of 0.86.

In addition to numeric metrics, we generated several visualizations to support our evaluation. Confusion matrices revealed the distribution of false positives and false negatives. ROC curves for each model illustrated the trade-off between true and false positive rates, with the random forest model consistently dominating the area under the curve. Feature importance plots for the random forest indicated that Page Values, ExitRates, and Product Related Duration were among the top contributors to prediction accuracy.

**Key Results:**

| **Model** | **Accuracy** | **Precision (1)** | **Recall (1)** | **F1-Score (1)** |
|---|---|---|---|---|
| Logistic Regression | 88% | 74% | 36% | 48% |
| Decision Tree | 86% | 53% | 55% | 54% |
| Random Forest | **90%** | **73%** | **56%** | **64%** |

### Discussion

The results strongly suggest that behavioral attributes such as session duration, bounce rate, and product page engagement are significant indicators of purchase intent. Among the three models tested, the random forest classifier demonstrated the highest predictive accuracy and generalization capability.

While logistic regression offered transparency and interpretability, its linear nature limited its ability to capture complex interactions. Decision trees addressed this to some extent but were susceptible to overfitting. The ensemble approach of random forests successfully combined the strengths of individual trees while mitigating their weaknesses.

These findings align with the broader literature on behavioral modeling in e-commerce. For instance, Nguyen et al. (2020) emphasized the importance of digital personalization in shaping online engagement, while Sakar et al. (2019) illustrated the predictive strength of behavioral metrics through neural network models. Our findings reinforce the value of using session-based features to guide customer targeting and personalization strategies.

### Challenges and Ethical Considerations

One of the challenges encountered was the slight imbalance in the dataset. While not extreme, the uneven distribution of purchase versus non-purchase sessions required careful validation to avoid biased results. Stratified sampling and the use of precision-recall metrics helped address this issue.

Ethically, the use of behavioral data must be approached with care. Although the dataset is anonymous, businesses must be transparent in how they use predictive models to influence user behavior. Predictive tools should enhance user experiences rather than manipulate them. Transparency, data privacy, and fairness should remain guiding principles (Nguyen et al., 2020)

### Peer Review Reflection

During the peer review process, valuable suggestions were made regarding the inclusion of additional performance metrics and more elaborate visualizations. Incorporating these suggestions improved the depth of my evaluation. The feedback also prompted a reevaluation of model complexity and the inclusion of a discussion on interpretability, especially relevant for business stakeholders.

## Conclusion

This research project demonstrates that supervised machine learning can be effectively applied to predict customer purchase intent using session-level data. Among the models evaluated, random forest emerged as the most effective, balancing accuracy with robustness. Our findings provide actionable insights for digital marketers and product managers seeking to boost conversion rates through real-time engagement and targeted interventions.Future work could expand this analysis by incorporating demographic or psychographic data and testing advanced models such as LSTM for sequential analysis. Deploying these models environments would also offer insights into their real-world impact on conversion rates and user satisfaction.

**Refrences :-**

Sharma, A., Adhikary, A., & Borah, S. B. (2020). Covid-19′ s impact on supply chain decisions: Strategic insights from NASDAQ 100 firms using Twitter data. *Journal of business research*, *117*, 443-449.

Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2019). Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*, *31*(10), 6893-6908.

Sakar, C., & Kastro, Y. (2018). Online shoppers purchasing intention dataset. *UCI machine learning Repository*, *10*, C5F88Q.