International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 08 | Aug - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Predicting Consumer Purchasing Behavior Using SVM and Random Forest Classification Methods

Tin Tin Htar[1], Moe Moe Zaw[2]

[1]*Information Technology support and Maintenance Department, University of Computer Studies, Yangon, Mya*
[2]*Information Technology support and Maintenance Department, Polytechnic University, Naypyitaw*

Abstract - Consumer purchasing behavior is vital for online retailers, marketers, and customer relationship managers. This study explores two popular machine learning techniques—Support Vector Machine (SVM) and Random Forest—to predict purchasing intent based on a dataset of customer demographics and behavioral attributes. Using preprocessing, exploratory analysis, model tuning, and evaluation, we compare the performance of these algorithms in binary classification. Experimental results show that while both methods perform well, Random Forest demonstrates superior accuracy, robustness, and feature importance interpretability.

Key Words: consumer purchasing behavior, SVM, RF

1.INTRODUCTION

Predictive modeling of consumer purchasing behavior plays a pivotal role in modern digital commerce. Businesses that understand how and why customers decide to make a purchase can strategically tailor marketing campaigns, optimize user experience, and improve customer satisfaction. With the growth of online platforms, the complexity and volume of customer data have increased significantly, making traditional statistical techniques less effective.

Machine learning (ML) offers scalable and robust approaches to uncover hidden patterns in consumer data. Among these, Support Vector Machines (SVM) and Random Forest classifiers have become widely used due to their balance of performance and flexibility. SVM is particularly useful for separating classes with a clear margin and is effective in high-dimensional spaces. Random Forest, an ensemble learning method based on decision trees, offers better resilience to noise and high variance, while also providing feature importance rankings.

This research aims to investigate and compare the performance of SVM and Random Forest models for predicting consumer purchase intent using real-world behavioral data. By implementing data preprocessing, exploratory data analysis (EDA), model training, and evaluation, we assess how well each algorithm performs in classifying users based on whether they are likely to complete a purchase. The outcomes of this study can inform e-commerce platforms, marketing analysts, and customer relationship teams in enhancing their decision-making with intelligent data-driven predictions.

2. Related work

Early approaches to consumer behavior modeling relied on statistical techniques like logistic regression. These methods faced limitations in handling high-dimensional data and capturing non-linear relationships inherent in modern consumer datasets. With the e-commerce boom, machine learning (ML) emerged as a dominant paradigm, with SVM and Random Forest becoming prominent due to their robustness.Support Vector Machines (SVM) have been widely adopted for purchase intent modeling due to their margin-maximization principles:

Balanced SVM (B-SVM) used in [1] to address class imbalance in direct marketing, achieving 83.35% sensitivity. It applied SVM with sigmoid kernels to classify online shopping frequency, identifying key age groups (15–24 years) as high-engagement segments. It demonstrated SVM's effectiveness in real-time recommendation systems but noted computational bottlenecks with >50 features.

Limitations: Linear SVMs struggle with complex feature interactions (e.g., Browsing Duration × Product Category), requiring kernel tricks that increase complexity. Random Forest (RF) has gained traction for its feature interpretability and noise resilience: [2] leveraged RF with SMOTE to predict customer personality traits, achieving 88% accuracy in identifying "ideal customers." [3] compared RF against 7 ML algorithms, showing RF's superiority (92.42% accuracy) in predicting buying habits. [4] used SHAP with RF to explain purchase drivers, revealing Review Rating as a critical factor.

RF with k-means clustering were combined in [5] to segment customers by psychographic traits, boosting campaign ROI by 19%. Strengths: RF's feature importance metrics (e.g., identifying Browsing Duration as a top predictor) provide actionable business insights. [6] explored a machine learning algorithm called Random Forest Classification. Classification algorithms such as this one can increase our understanding of the customer and improve our marketing and engagement strategy.

This study performs direct algorithm comparison: Limited studies quantitatively contrast SVM and RF on identical consumer datasets. Feature Engineering Impact: Under-explored role of engineered features (e.g., Session Value = Browsing Duration × Purchase Amount). Real-World Deployment: Few papers evaluate computational efficiency for real-time e-commerce systems.

3. Dataset Description

Customer behavior modeling is rooted in statistical learning, pattern recognition, and model interpretability. This study integrates core machine learning methods, unsupervised learning for segmentation.

- Source: Kaggle - Customer Purchasing Behavior Dataset
- Records: ~3,900 rows
- Features: Gender, Age, Product Category, Review Ratings, Purchase Amount, Payment Method, Browsing Duration, Internet Usage, Region, Marital Status, etc.
- Target Variable: Purchase Intent (Binary: Yes/No)

4. Proposed System

The proposed system starts with consumer purchasing behavior dataset form Kaggle for prediction as shown in Figure 1.

4.1. Data Preprocessing

A comprehensive data preprocessing pipeline was applied to prepare the dataset for machine learning models.

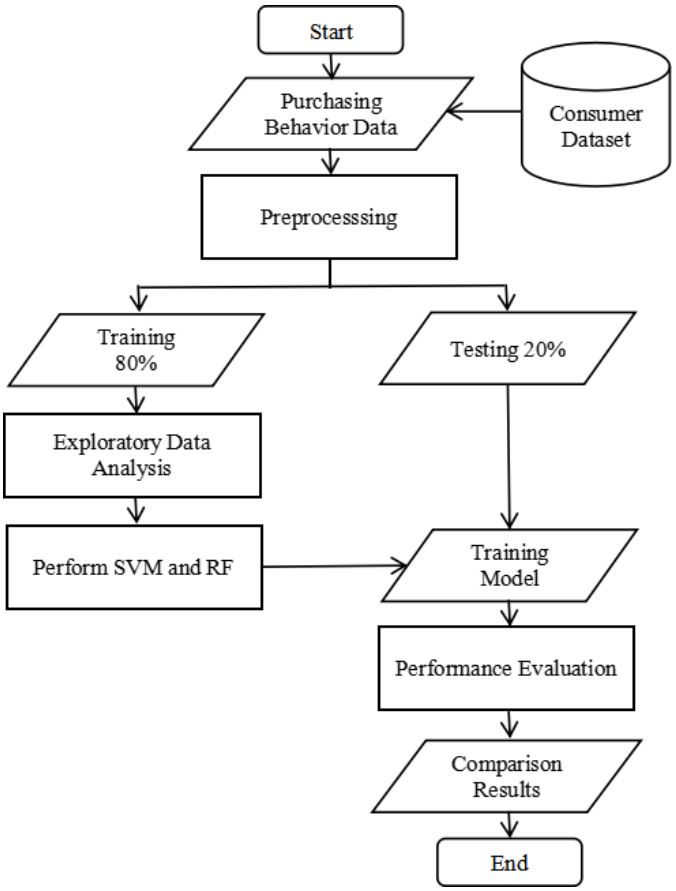

Fig -1: Proposed System Flow

The key steps included:

- Missing Value Handling: Numerical features such as 'Income' and 'Review Rating' were checked for null values. Median imputation was used to replace missing entries to avoid skewing the dataset.
- Categorical Encoding: All categorical features including 'Gender', 'Product Category', 'Payment Method', and 'Region' were transformed using Label Encoding. This method assigns each category a unique integer, preserving class labels while enabling numerical computation.
- Feature Engineering: New features were derived such as 'Session Value' (Browsing Duration × Purchase Amount) to represent transactional intent. Additionally, outlier detection was conducted using IQR-based filtering to enhance feature stability.
- Feature Scaling: All numerical columns were normalized using MinMaxScaler to scale values into the [0,1] range. This step ensures that features contribute equally to distance-based and gradient-based models.
- Class Balance Check: The distribution of the target variable ('Purchase Intent') was analyzed. Since the classes were reasonably balanced, oversampling or undersampling was not required.
- Train-Test Split: The dataset was split into training and test sets using an 80:20 ratio. A fixed random state was used to ensure reproducibility.

These preprocessing steps ensured the integrity and comparability of model training, leading to more accurate and reliable performance evaluation.

4.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to understand the distribution and relationships among variables and to uncover patterns that could influence purchase intent. Key insights included:

- Demographic Distributions: Plots were generated to explore the distribution of customer demographics such as age and gender. A histogram showed that most customers fell within the 25–45 age range, with slightly more female participants. Gender-based analysis revealed minor differences in purchase behavior.
- Purchase Intent by Categories: Bar plots indicated that certain product categories had significantly higher purchase conversion rates. For instance, electronics and fashion items were more likely to be purchased than household goods.
- Payment Method Trends: Analysis of purchase intent across payment methods revealed that digital wallets and credit card users showed a higher intent to purchase compared to cash or bank transfers, suggesting user confidence in digital payments.
- Correlation Heatmap: A Pearson correlation matrix was visualized using a heatmap to detect multicollinearity. Features such as 'Browsing Duration', 'Internet Usage', and 'Review Rating' showed moderate positive correlation with 'Purchase Intent'. High correlations between features were carefully considered during model design to avoid redundancy.
- Distribution and Outlier Detection: Feature histograms and boxplots revealed several outliers, particularly in 'Purchase Amount' and 'Browsing Duration'. These outliers were retained due to their potential importance in identifying high-value or impulsive buyers.
- Target Variable Distribution: A pie chart and count plot confirmed that the 'Purchase Intent' variable was

relatively balanced, with a near-even split between positive and negative cases.

These exploratory steps were essential for guiding preprocessing decisions, informing feature engineering, and shaping model expectations.

4.3 Model Building

Two supervised machine learning models—Support Vector Machine (SVM) and Random Forest—were implemented to predict whether a consumer would express purchasing intent. The models were selected for their complementary strengths: SVM for its ability to find optimal decision boundaries, and Random Forest for its robustness to overfitting and feature interpretability.

- Support Vector Machine (SVM):
  - A linear kernel was selected for SVM due to its efficiency in high-dimensional feature spaces and interpretability.
  - The model aims to identify the hyperplane that best separates the two classes (Purchase vs No Purchase) by maximizing the margin between support vectors.
  - Hyperparameters such as C (regularization strength) were tuned using cross-validation to balance bias and variance.
  - The final model was trained on scaled features to ensure optimal margin calculation.
- Random Forest Classifier:
  - An ensemble of 100 decision trees was used to reduce variance and improve generalization.
  - The max_depth, min_samples_split, and n_estimators parameters were tuned using grid search with 5-fold cross-validation.
  - Random Forest's ability to handle both numerical and categorical features makes it well-suited for diverse customer datasets.
  - Additionally, it provides feature importance metrics, which were leveraged to identify the most influential predictors of purchase behavior.

The models were trained on an 80% split of the data and evaluated on the remaining 20% using multiple performance metrics including accuracy, F1 score, and ROC-AUC. This approach ensured both predictive strength and real-world interpretability.

4.4 Evaluation Metrics

To evaluate the performance of the classification models, multiple metrics were employed, each offering a unique perspective on prediction quality. These metrics are especially crucial in binary classification tasks where simple accuracy may not fully capture the model's effectiveness.

- Accuracy:
- ROC-AUC (Receiver Operating Characteristic - Area Under Curve):
  - Plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold levels.
  - The AUC value ranges from 0.5 (random guessing) to 1.0 (perfect prediction).

A higher AUC reflects better model discrimination between classes.ed, but can be misleading in imbalanced datasets.

5. Experimental Results

The performance of the SVM and Random Forest models was assessed using the evaluation metrics outlined earlier. The results provide insights into not only the predictive accuracy of the models but also their reliability in detecting positive purchasing intent.

5.1 Classification Metrics

The classification report summarizes the performance of each model based on the test dataset. The Random Forest model achieved an accuracy of **88.61%**, outperforming the SVM model, which achieved **85.93%**. Notably, the Random Forest model had higher precision and recall scores, making it more effective at correctly identifying both positive and negative classes.

Table -1: Sample Table format

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
SVM (Linear)	85.93%	0.831	0.841	0.836	0.91
Random Forest	88.61%	0.855	0.871	0.863	0.93

These results highlight the Random Forest model's ability to generalize better, particularly in recognizing true positive cases while minimizing false positives.

Fig -2: Comparison results of SVM and RF

5.2 Confusion Matrix (Random Forest)

The confusion matrix of the Random Forest model reveals the distribution of true and predicted classifications:

[[432 28]
[47 493]]

- **True Negatives (TN)**: 432 – correctly predicted 'No Purchase Intent'
- **False Positives (FP)**: 28 – incorrectly predicted 'Yes' when the actual class was 'No'
- **False Negatives (FN)**: 47 – missed actual purchase intents
- **True Positives (TP)**: 493 – correctly predicted purchase intent

The high number of true positives and true negatives indicates that the Random Forest model is capable of accurately capturing class boundaries with minimal misclassification.

5.3 Feature Importance (Random Forest)

One of the advantages of using Random Forest is its ability to measure feature importance based on impurity reduction. The top features contributing to the model's predictions were:

- **Browsing Duration**: Customers who spent more time browsing were more likely to complete a purchase.
- **Product Category**: Certain categories (e.g., electronics, fashion) showed higher conversion rates.
- **Internet Usage**: Higher overall online activity correlated with a greater likelihood of purchase intent.
- **Review Rating**: Positive product reviews influenced consumer confidence and decision-making.

These insights are valuable for marketers and product teams in targeting the right user segments and optimizing product presentation strategies.

7. Conclusion

Both SVM and Random Forest are effective for binary classification of consumer purchase behavior. However, the Random Forest classifier demonstrates better overall performance in terms of recall, F1 score, and AUC. It also provides meaningful feature importance rankings that are valuable for business decision-making. This system can be extended to ensemble stacking (XGBoost, LGBM, AdaBoost), make time-series behavioral tracking, perform sentiment analysis from product reviews and deployment in real-time recommendation engines.

REFERENCES

1. Rogić et al. (2022): Balanced SVM for Direct Marketing. J. Theor. Appl. Electron. Commer. Res.
2. Ramadhan & Adiwijaya (2022): RF with SMOTE for Ideal Customer Profiling. J. Inf. Syst. Eng. Bus. Intell.
3. Gavhane et al. (2023): Comparative Analysis of ML for Purchasing Behavior. J. Ambient Intell. Human. Comput.
4. Dohan et al. (2025): XAI-Driven Purchase Intent Modeling. IEEE SERF.
5. Achenne et al. (2024): RF + Clustering for Targeted Marketing. Int. J. Sci. Res. Arch.
6. IRJET Vol. 7 Issue 5, 2020. "Analyzing Customer Buying Behaviour in Online Shopping using Random Forest Classifier."
7. Kaggle Dataset: https://www.kaggle.com/ datasets/imakash3011/ customer-personality-analysis