Using Random Forest and Support Vector Machine Algorithms to Predict Online Shopper Purchase Intention from E-Commerce Session Data

Reza Alamsyah[1,*], Sri Wahyuni[2]

[1]*Department of Information Systems, STMIK Methodist Binjai, Binjai, Indonesia*

[2]*Information Systems, Faculty of Engineering and Computer Science, Dharmawangsa University, Indonesia*

(Received: December 20, 2023; Revised: March 25, 2024; Accepted: April 29, 2024; Available online: June 2, 2024)

Abstract

This study explores the use of machine learning algorithms to predict online shopper purchase intention, aiming to provide e-commerce businesses with actionable insights into consumer behavior. The Online Shoppers Purchasing Intention dataset, containing 12,330 session records from an e-commerce site, was analyzed using two classification models: Random Forest and Support Vector Machine (SVM). The models were evaluated based on key performance metrics including accuracy, precision, recall, F1-score, and ROC AUC. Results showed that the Random Forest model outperformed the SVM model, achieving an accuracy of 90.43% and a ROC AUC score of 0.94, indicating strong predictive capability. PageValues and ProductRelated_Duration were identified as the most important features influencing purchasing behavior, with higher values of these features being strongly associated with successful purchases. The study provides valuable insights into the behaviors that drive purchasing decisions in e-commerce, showing that longer engagement with product-related content and higher monetary value pages significantly increase the likelihood of conversion. While the study contributes to understanding online shopper behavior through machine learning, it is limited by the class imbalance in the dataset and the absence of more granular customer data. Future research could address these limitations by incorporating additional features and exploring deep learning models for more accurate predictions. Practical implications of the study suggest that e-commerce businesses can improve conversion rates by optimizing product-related pages and focusing on key user behaviors that are predictive of purchases.

Keywords: Online Shopper Behavior, Machine Learning, Random Forest, Purchase Intention, E-commerce

1. Introduction

The rise of e-commerce has fundamentally transformed the global business landscape, driven by technological advancements and changing consumer behaviors. E-commerce, characterized by the buying and selling of goods and services through the internet, has become an essential component for businesses across various sectors. This transformation has been influenced largely by increasing internet penetration, enhanced mobile connectivity, and the growing trust in online transactions, which collectively foster an environment conducive to e-commerce growth [1], [2].

One significant factor contributing to the growth of e-commerce is the improvement in digital infrastructure, including the expansion of broadband and mobile networks, as well as the development of various payment methods. Enhanced technological capabilities have enabled businesses, especially Small and Medium Enterprises (SMEs), to reach broader markets that were previously inaccessible, thus amplifying their competitiveness and innovation [3], [4]. E-commerce allows these businesses to not only reduce costs associated with traditional retail but also to expand their product offerings and customer bases significantly [5].

Furthermore, e-commerce has become a crucial driver of economic growth, particularly evident during the COVID-19 pandemic when physical interactions were limited. The surge in online shopping during this period highlighted the resilience and adaptability of the e-commerce sector. Global e-commerce sales reached approximately $4.28 trillion in 2020, and it is projected to rise to $5.4 trillion by 2024. This growth trend has been particularly pronounced in developing countries, such as China, which has quickly become the largest e-commerce market globally [6], [7].

***Corresponding author: Sri Wahyuni (sriwahyuni15jun@dharmawangsa.ac.id)**
DOI: https://doi.org/10.47738/ijaim.v4i2.81

The implications of e-commerce extend beyond mere economic metrics; it has led to a substantial shift in consumer behavior and expectations. With the ease of accessing various products and services, consumers have become accustomed to lower search costs and a wider array of choices than what conventional retail can provide [5]. Additionally, e-commerce platforms have created efficiencies that benefit consumers while facilitating better inventory management and supply chain logistics for businesses. The reliance on e-commerce has established a new normal in business operations, necessitating continuous adaptation to remain competitive in a rapidly evolving marketplace.

The challenge of predicting purchase intent in online shopping sessions stems from a multitude of factors that significantly influence consumer behavior. Consumer attitudes play a critical role in shaping online purchase intentions, making their understanding paramount for e-commerce platforms [8]. For instance, the interplay between hedonic motivations—such as pleasure and satisfaction derived from shopping—and trust affects consumers' willingness to engage in online transactions. The complexity of these dynamics often makes it difficult for businesses to accurately forecast consumer behaviors.

Moreover, information overload poses a considerable challenge. Consumers today are inundated with vast amounts of information, including online reviews and product details. Such overload can lead to confusion, affecting the decision-making process and, consequently, the purchase intent [9]. However, some research suggests that electronic word-of-mouth (E-WOM) can impact purchase decisions, but its effectiveness may be diminished by excessive or contradictory information.

Additionally, perceived risk remains a significant barrier to predicting purchase intent. Consumers' anxiety about product quality, transaction security, and the overall reliability of online retailers can deter them from completing purchases. For example, in markets specializing in non-standardized products, mitigating perceived risks through quality assurances and enhanced payment security is essential for increasing purchase intent [10]. This relationship underscores the importance of trust and safety in shaping online consumer behavior, which tends to be unpredictable and varies significantly across different shopping contexts [11].

Furthermore, psychological factors also play a vital role in shaping purchase intentions. Behavioral studies indicate that internal traits, such as a consumer's propensity to be a reflexive buyer, can significantly influence purchasing decisions. Such internal factors, coupled with external influences like social norms, highlight the multi-faceted nature of consumer behavior [12]. Situational ethics also play an important role; consumers’ norms and attitudes can drive their decisions in specific contexts, particularly during events that fundamentally alter shopping behaviors, such as the COVID-19 pandemic [13].

The objective of this study is to predict online shoppers' purchase intention using machine learning algorithms. By leveraging various session-based features such as the number of pages visited, time spent on each page, and session bounce rates, the goal is to develop a predictive model that can accurately classify whether a given online shopping session will result in a purchase. Specifically, the research will utilize two popular machine learning algorithms—Random Forest and Support Vector Machine (SVM)—to build and evaluate classification models. These models will be trained to predict the "Revenue" target variable, which indicates whether the session resulted in a purchase (True) or not (False).

The significance of this study lies in its potential to improve e-commerce strategies and customer targeting. As online shopping continues to grow in scale and complexity, understanding and predicting purchasing behavior can significantly enhance personalized marketing, inventory management, and sales forecasting. By accurately predicting whether a session will lead to a purchase, e-commerce businesses can optimize their resources and tailor customer engagement efforts in real-time. This research could also provide insights into which features or behaviors are most predictive of a purchase, allowing businesses to refine their strategies and increase conversion rates.

2. Literature Review

2.1. Online Shopper Behavior

Research on factors influencing online purchase decisions has identified several key determinants that shape consumer behavior in the digital marketplace. Central to many studies is the integration of psychological theories, such as the

Theory of Planned Behavior (TPB) and the Technology Acceptance Model (TAM), which help to elucidate the reasoning behind consumer decisions. For instance, Nguyen et al. underscore the significant influence of the e-commerce platform's reputation and credibility on online purchase intention, noting that a high e-commerce exchange image directly correlates with consumers' willingness to buy [14].

Another critical factor is the role of online customer reviews and influencer marketing, highlighted by Fachmi and Sinau. Their study indicates that these variables jointly contribute to approximately 34.7% of the purchase decision-making process. The influence of social media, online brand loyalty, and the effectiveness of influencers plays a substantial role, although other external factors, such as product pricing and marketplace competition, also significantly affect purchase decisions [15].

The quality of e-services and the ease of transaction are also pivotal elements contributing to consumers' purchasing decisions. Hartono et al. emphasize that enhanced e-service quality can lead to more effective and appealing online transactions, thereby positively influencing purchasing behavior [16]. Additionally, studies focused on specific product categories, such as food products, have shown that consumers place considerable value on factors like food hygiene, price, and convenience when selecting online offerings. Similarly, Rahman and Sultana identify brand name, pricing, and product quality as substantial drivers of mobile phone purchasing behavior, further asserting the subjective nature of these purchase intentions within specific market [17].

2.2. Machine Learning in E-Commerce

The application of machine learning algorithms, particularly Random Forest (RF) and Support Vector Machine (SVM), in predicting purchase behavior has garnered significant attention in recent research. These algorithms leverage large datasets to uncover patterns and trends in consumer behavior, aiding businesses in making informed marketing and operational decisions.

Random Forest is a versatile ensemble learning method known for its effectiveness in classification tasks. In the context of e-commerce, RF has been utilized to predict customer purchase behavior by analyzing various input features derived from user interactions, preferences, and transaction histories. A study by Sunarya et al. demonstrated the capability of Random Forest to outperform logistic regression in terms of predictive accuracy, showcasing its robustness in handling complex datasets common in e-commerce environments [18]. This algorithm’s inherent ability to manage high-dimensional data while mitigating overfitting makes it particularly suited for applications in online shopping behavior prediction.

Support Vector Machine, on the other hand, is favored for its effectiveness in high-dimensional spaces and its flexibility in choosing various kernel functions. Its application has been discussed in the context of financial market predictions, underlining its capacity to discern patterns in data that could also apply to user behavior in e-commerce [19]. Additionally, RF and SVM have been compared in terms of their predictive power for e-commerce customer behavior, with some studies revealing that RF may offer advantages in scenarios characterized by diverse and non-linear relationships between variables

In terms of performance comparison, recent findings indicate that while both algorithms possess their strengths, Random Forest frequently demonstrates a superior ability to generalize across multiple datasets typically encountered in e-commerce environments [20]. Furthermore, hybrid models that optimize parameter tuning processes, such as Bayesian-optimized Random Forest, have emerged, enhancing predictive capabilities further in complex environments like international e-commerce [21].

2.3. Feature Engineering Analysis

Feature selection is a critical process in machine learning, particularly for classification tasks, as it involves identifying and selecting a subset of relevant features from a larger set of input variables. Effective feature selection can significantly enhance model performance by improving accuracy, reducing overfitting, and decreasing computation time. Various feature selection methods have been developed, each with its strengths and weaknesses, making their evaluation essential for optimal performance in classification tasks [22].

One widely used approach is filter methods, which assess the relevance of features by evaluating their statistical properties relative to the target variable. For instance, the use of mutual information to gauge the dependency between features and the outcome can aid in selecting the most informative variables [23]. Another major category is wrapper methods, which involve training and evaluating a model using a subset of features and assessing the model's performance through techniques such as cross-validation. While often leading to high accuracy, these methods can be computationally intensive, particularly with large datasets, due to the iterative nature of training multiple models [24]. Embedded methods represent a hybrid that integrates feature selection within the model training process. Techniques like LASSO (Least Absolute Shrinkage and Selection Operator) apply a penalization strategy to feature coefficients, effectively reducing their impact and enabling automatic feature selection during the model fitting process [25]. These methods tend to strike a balance between the robustness of wrapper methods and the efficiency of filters, as they consider the interaction between features when building the model.

3. Method

3.1. Dataset Overview and Data Preprocessing

The dataset used for this study is the Online Shoppers Purchasing Intention dataset, which consists of 12,330 records representing distinct web sessions from an e-commerce site over the course of one year. Each record includes several features that describe the characteristics of the shopping session, such as the number of pages visited, time spent on different types of pages, and user-related information like browser type and traffic source. The goal of this study is to predict whether a session will result in a purchase, using machine learning algorithms.

The first step in the methodology was to load and inspect the dataset using the Pandas library. A quick analysis was conducted to understand the structure of the data, checking for any missing values or anomalies. The initial exploration revealed the need for preprocessing, particularly handling missing data and preparing categorical variables for model input. Missing values in the numerical features were handled by filling them with the mean value of the respective columns, which is a common approach when the missing data is assumed to be missing at random. For categorical features, missing values were replaced with the mode (the most frequent value), as this would maintain the most common category without introducing bias. After handling missing values, the dataset was ready for further transformation, where the numerical features were scaled, and categorical features were encoded. This ensured that all variables were appropriately prepared for machine learning algorithms, which require numerical input and benefit from standardized scales.

3.2. Feature Encoding and Scaling

One of the critical steps in preparing the dataset for machine learning was the encoding of categorical variables. The dataset contained several categorical columns, such as Month, OperatingSystems, Browser, Region, TrafficType, VisitorType, and Weekend. These categorical variables were transformed into numerical representations using LabelEncoder from Scikit-Learn. Label encoding converts each category into a unique integer, enabling the machine learning models to process them efficiently.

In addition to encoding categorical variables, the numerical features needed to be standardized. The numerical features in the dataset include various counts and durations, such as the number of pages viewed in different categories (e.g., Administrative, ProductRelated) and the corresponding time spent on those pages. To ensure that each feature contributes equally to the model, especially for algorithms sensitive to feature scaling (such as Support Vector Machines), the StandardScaler was applied to normalize the numerical features. This scaled the features to have a mean of zero and a standard deviation of one, which helped improve the performance and convergence of the models. Scaling is particularly important when using algorithms like SVM, where unscaled features can lead to suboptimal performance. After encoding and scaling, the dataset was ready for model training.

3.3. Train-Test Split

After preprocessing, the dataset was split into training and testing sets using the train_test_split function from Scikit-Learn. This function randomly divides the data, ensuring that the model is trained on one portion of the data and tested on a separate, unseen portion to evaluate its generalization capability. For this study, 80% of the data was used for

training, and 20% was reserved for testing. The random state was set to ensure that the data split was reproducible. The training set was used to train the machine learning models, while the testing set was used to evaluate their performance. The split ensures that the model's performance metrics are reliable and that the evaluation reflects how well the model will perform on unseen data. The shapes of the training and testing sets were printed to confirm that the split was performed correctly, with training data consisting of 80% of the records and test data consisting of the remaining 20%.

3.4. Model Training and Evaluation

Two machine learning models were trained to predict online shoppers' purchase intention: Random Forest and Support Vector Machine (SVM). The Random Forest Classifier is an ensemble learning method that builds multiple decision trees and combines their outputs to improve prediction accuracy and prevent overfitting. The Random Forest model was trained using the default parameters from Scikit-Learn, and it was then used to predict whether a session would result in a purchase (1) or not (0).

The second model, Support Vector Machine (SVM), was trained using the SVC class from Scikit-Learn, which supports binary classification. SVM works by finding the hyperplane that best separates the classes in the feature space. For this study, the probability=True option was set, which allows the SVM model to output probabilities for the class predictions. This is particularly useful for evaluating the ROC AUC score, as it requires probability estimates for the positive class.

Both models were trained on the training set, and their predictions were made on the testing set. The model performance was evaluated using several metrics, including accuracy, precision, recall, F1-score, and ROC AUC. These metrics provide a comprehensive view of the model's ability to correctly classify sessions that result in a purchase and avoid false positives or false negatives.

3.5. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was performed to gain a deeper understanding of the data and uncover insights that might inform model development. The distribution of the target variable (i.e., whether a session resulted in a purchase) was visualized using a countplot. This visualization provided insight into the class imbalance in the dataset, with a larger number of sessions resulting in no purchase compared to those resulting in a purchase.

A correlation matrix was also generated to explore the relationships between numerical features. This helped identify which features were most strongly correlated with each other and whether any multicollinearity issues existed. A heatmap of the correlation matrix was plotted, with higher correlations highlighted in darker colors. This is crucial for understanding how features interact and whether some features might be redundant.

Additionally, a pairplot was created to visualize relationships between the top five most important features, as identified by the Random Forest model. This allowed for a deeper exploration of how certain features, such as ProductRelated_Duration, SpecialDay, and BounceRates, influence the likelihood of a purchase, with each pair showing the distribution of features for both purchase and non-purchase sessions.

3.6. Model Evaluation with ROC Curve

To assess the performance of both models, ROC curves were plotted. The ROC curve illustrates the trade-offs between sensitivity and specificity, with the True Positive Rate (TPR) plotted against the False Positive Rate (FPR). The AUC (Area Under the Curve) is a valuable metric, as it provides a single value that summarizes the model's ability to discriminate between the positive and negative classes. Higher AUC values indicate better performance. The ROC curves for both models were plotted on the same graph to compare their performance visually. For each model, the FPR and TPR were calculated using the roc_curve function, and the AUC score was computed using roc_auc_score. This allowed for a direct comparison of the Random Forest and SVM models in terms of their ability to predict purchase intention across different decision thresholds.

3.7. Feature Importance and Insights

Feature importance was evaluated for the Random Forest model, which is capable of providing insights into which features have the most influence on the model's predictions. Feature importance scores were extracted from the trained Random Forest model and sorted in descending order. These scores highlight which features, such as

ProductRelated_Duration, SpecialDay, and PageValues, play a significant role in predicting whether a session results in a purchase. A bar plot was generated to visualize the relative importance of each feature. This helps in understanding the factors that drive purchasing behavior in online shopping sessions. By examining the feature importance, businesses can gain insights into which aspects of a shopping session to focus on in order to improve user engagement and increase conversion rates.

This detailed methodology outlines the process of developing machine learning models to predict online shoppers' purchase intention using Random Forest and SVM. By combining data preprocessing, feature engineering, and exploratory data analysis, this study aims to provide actionable insights for e-commerce businesses to enhance their customer targeting and improve conversion rates.

4. Results and Discussion

4.1. Data Overview Analysis

The initial dataset consists of 12,330 records and 18 columns, with all features being non-null. A quick examination of the data revealed a range of numerical, categorical, and Boolean variables that describe different aspects of online shopper sessions. The dataset includes features such as the number of pages visited (e.g., Administrative, ProductRelated), the total duration spent on those pages (e.g., Administrative_Duration, ProductRelated_Duration), as well as session-level metrics such as BounceRates, ExitRates, and PageValues. Additionally, there are categorical variables representing the Month, OperatingSystems, Browser, Region, TrafficType, and VisitorType, along with the Revenue column, which indicates whether a session resulted in a purchase.

The initial data exploration showed no missing values, as all columns were fully populated, confirming that the dataset was complete and ready for analysis. Descriptive statistics revealed the distribution of key features, such as the high variance in ProductRelated_Duration (ranging from 0 to over 63,000 seconds) and a relatively low average PageValues. This variability in session behaviors was crucial for modeling and understanding purchasing patterns.

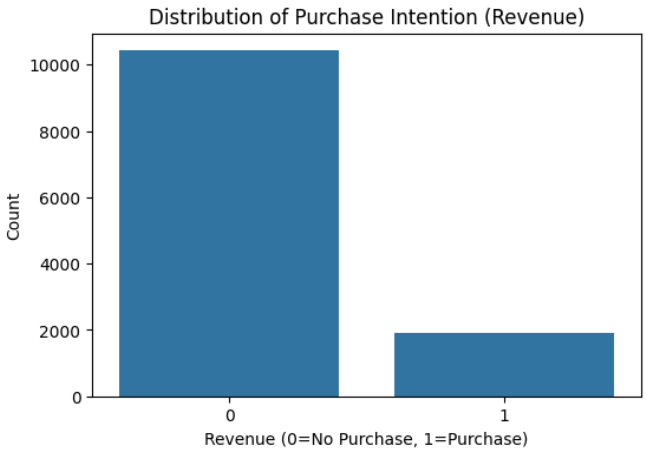

Figure 1. Distribution of Purchase Intention (Revenue)

Figure 1 displays the distribution of the Revenue variable, which indicates whether a session resulted in a purchase (1) or not (0). The chart clearly shows an imbalance in the dataset, with a significantly higher number of sessions that did not result in a purchase (0) compared to those that did (1). Specifically, there are around 10,422 sessions where no purchase was made, while only 1,908 sessions resulted in a purchase. This class imbalance is common in e-commerce data and needs to be addressed during model training, as it may influence the model's ability to correctly predict minority class outcomes (purchases).

 ISSN 2776-8007

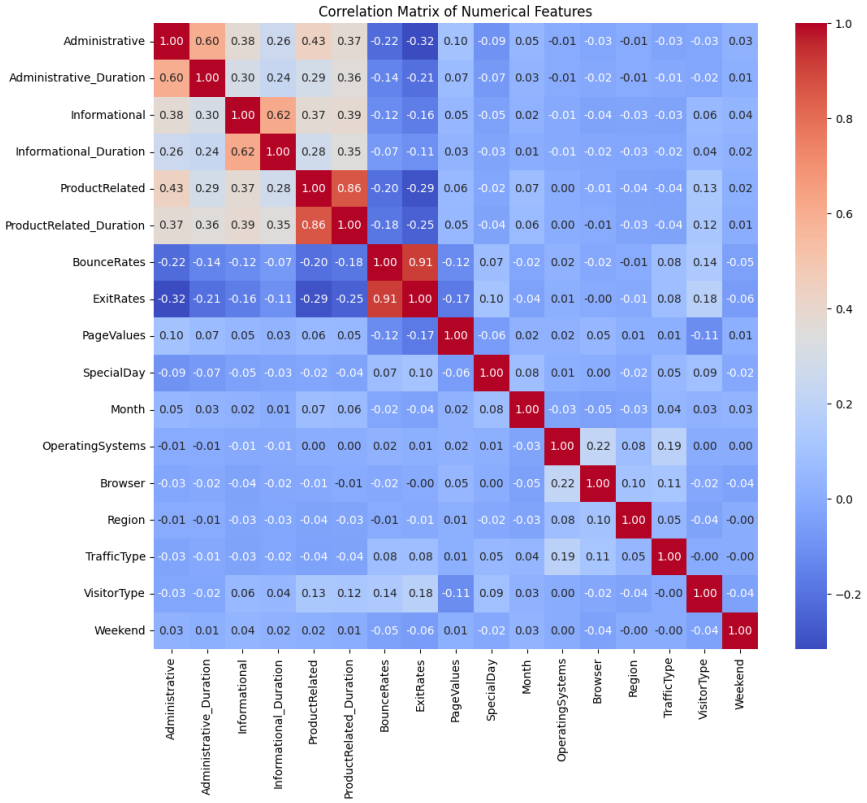

Figure 2. Correlation Matrix of Numerical Features

Figure 2 provides a correlation matrix for the numerical features in the dataset. The correlation matrix visually represents the relationships between various numerical features, with each cell showing the Pearson correlation coefficient between two features. A darker red color indicates a stronger positive correlation, while blue shades reflect a negative correlation. Several key correlations are noteworthy. ProductRelated_Duration is highly correlated with ProductRelated (0.86), suggesting that the amount of time spent on product-related pages is strongly linked to the number of product pages viewed. This aligns with the intuitive assumption that users who engage more with product-related content are more likely to convert into purchasers. BounceRates and ExitRates show a strong positive correlation (0.91), meaning that sessions with high bounce rates (users leaving the site quickly) tend to have higher exit rates (users leaving after visiting a few pages). This indicates that users who leave early are less likely to proceed with a purchase. PageValues shows a moderate correlation with ExitRates (-0.12), indicating that higher-value pages tend to have a slight negative relationship with exit rates, suggesting that valuable pages may keep users engaged longer. SpecialDay has a weak correlation with other features, suggesting that proximity to special shopping days, like holidays, has minimal direct impact on other user behavior metrics, although it could still be relevant for predicting purchasing intent during special occasions. These correlations provide valuable insights into how different user behaviors and session metrics are related and can inform which features might be most predictive when building a classification model. For example, ProductRelated_Duration and PageValues are likely to be important features for predicting whether a session leads to a purchase.

 ISSN 2776-8007

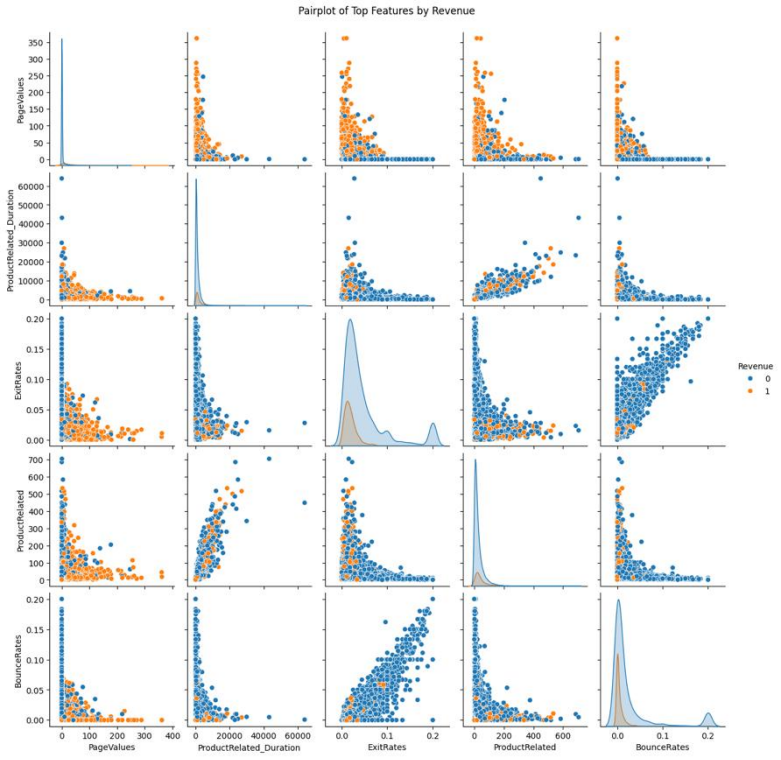

Figure 3. Pairplot of Top Features by Revenue

Figure 3 displayed above visualizes the relationships between the top five most important features—PageValues, ProductRelated_Duration, ExitRates, ProductRelated, and BounceRates—with respect to the target variable Revenue (purchase outcome). In the plot, the blue dots represent sessions where no purchase occurred (Revenue = 0), while the orange dots represent sessions that resulted in a purchase (Revenue = 1). The diagonal histograms show the distributions of each feature, and the scatter plots below the diagonal show pairwise relationships between the features. PageValues (the monetary value of pages viewed) shows a strong concentration at low values, with a significant number of sessions having a value of 0. This suggests that most users do not engage with high-value pages. There is a noticeable trend where sessions with higher PageValues are more likely to result in a purchase (orange dots), and these sessions tend to correlate with longer durations spent on product-related pages (ProductRelated_Duration). This indicates that users who engage with higher-value pages and spend more time browsing product-related content are more likely to convert into purchasers. There is an inverse relationship between ProductRelated_Duration (the time spent on product-related pages) and ExitRates (the likelihood of leaving the session). Sessions with higher time spent on product-related pages tend to have lower exit rates, suggesting that users who browse more product content are less likely to abandon the session. The scatter plot also shows that users who leave quickly (high ExitRates) typically do not engage with many product-related pages (low ProductRelated_Duration).

ProductRelated (the number of product-related pages viewed) is positively correlated with ProductRelated_Duration. This makes sense because users who view more product pages tend to spend more time on them. However, the data shows that many sessions with few product-related pages viewed (ProductRelated = 0 or 1) have zero or very short durations (ProductRelated_Duration). This could suggest that users who do not engage deeply with product content are less likely to convert into a purchase. ExitRates and BounceRates show a positive correlation, which is evident in the plot. Both metrics capture the likelihood that a user will leave the session quickly, with BounceRates representing a single-page exit and ExitRates capturing the overall session exit behavior. Interestingly, users who have high BounceRates (those who leave after viewing only one page) also tend to have higher ExitRates, indicating that a quick exit from the site usually leads to an overall session exit. These sessions (blue dots) are less likely to convert to a purchase.

The pairplot visually reinforces the finding from earlier that purchasing sessions (orange dots) tend to show higher values of PageValues, ProductRelated_Duration, and ProductRelated, while non-purchasing sessions (blue dots) are

concentrated around lower values in these features. The BounceRates and ExitRates for sessions that result in purchases (orange) are relatively lower, indicating that users who make a purchase tend to engage more with the content and exit less frequently. In summary, the pairplot provides a clear visual representation of how specific features, such as PageValues, ProductRelated_Duration, and ProductRelated, influence the likelihood of a purchase, reinforcing the importance of user engagement and content interaction in driving conversions. The scatter plots reveal the relationships between these features and suggest strategies for improving website engagement, particularly focusing on product-related content to increase the chances of conversion.

4.2. Model Evaluation of Random Forest and SVM

Two machine learning models, Random Forest and Support Vector Machine (SVM), were trained to predict whether a session would result in a purchase (the target variable Revenue). The performance of both models was evaluated using various classification metrics, including accuracy, precision, recall, F1-score, and ROC AUC. The Random Forest Model achieved an accuracy of 90.43%, indicating that it correctly predicted the outcome in the majority of cases. The precision (0.7439) was relatively high, meaning the model successfully identified positive purchases among the predicted purchases. However, the recall was lower (0.5653), reflecting that the model missed some actual purchase sessions. The F1-score (0.6424) combined precision and recall, giving a balanced evaluation of the model's performance. Most notably, the ROC AUC for Random Forest was 0.9363, suggesting that it had a strong ability to differentiate between purchase and non-purchase sessions. In contrast, the Support Vector Machine (SVM) model performed slightly worse, with an accuracy of 89.25%. While the precision was comparable to the Random Forest model (0.7477), the recall was much lower (0.4427), suggesting that SVM had more difficulty identifying purchase sessions. The F1-score for SVM was 0.5561, lower than Random Forest, and the ROC AUC was also lower at 0.8610. These results indicate that while SVM performs well overall, it has limitations in distinguishing between sessions that lead to purchases and those that do not, especially when compared to Random Forest.

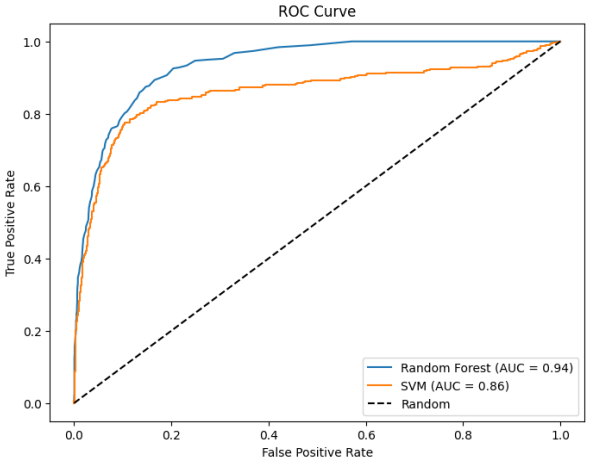

Figure 4. ROC Curve

Figure 4 displays the ROC curve for both the Random Forest and SVM models. The True Positive Rate (TPR) is plotted against the False Positive Rate (FPR) for each model. The ROC curve shows the trade-off between sensitivity (True Positive Rate) and specificity (1 - False Positive Rate). The Random Forest model has a strong ROC curve, indicating its superior performance in distinguishing between sessions that result in a purchase and those that do not. The model's AUC score of 0.94 shows that it has a high ability to correctly identify both positive and negative instances (purchases and non-purchases). The curve approaches the top-left corner, suggesting that the model achieves a good balance between false positives and true positives. The SVM model, on the other hand, performs somewhat less effectively, with an AUC score of 0.86. While it still provides a reasonable ability to distinguish between the two classes, the curve is slightly lower than that of Random Forest, showing that it has a higher rate of false positives or false negatives. The dashed diagonal line represents a random classifier, where the model is unable to make any meaningful distinctions between classes. The fact that both models significantly outperform the random classifier confirms their utility for predicting purchase behavior.

Feature importance analysis in Random Forest highlighted PageValues as the most influential feature, contributing approximately 37.33% to the model's predictions, followed by ProductRelated_Duration and ExitRates. This indicates that the monetary value of pages viewed and the time spent on product-related pages are the most significant predictors of purchase intent.

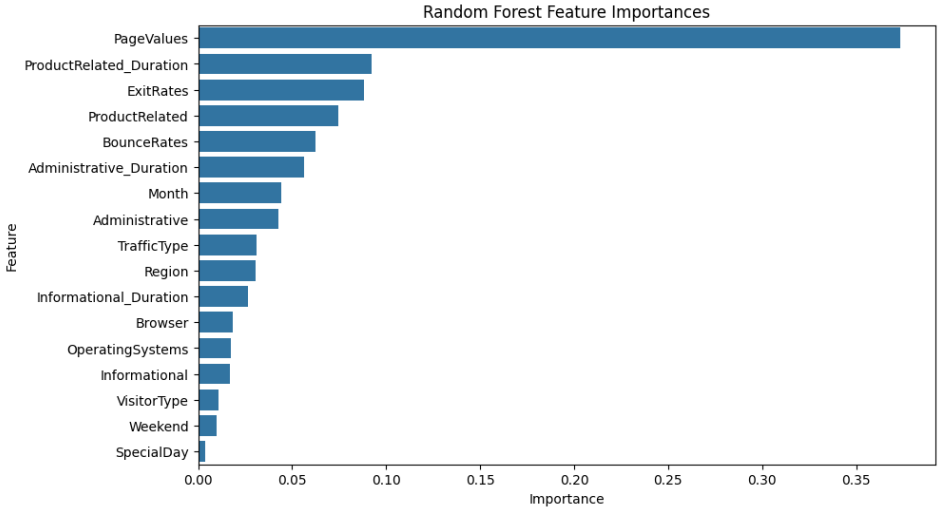

Figure 5. Random Forest Feature Importances

Figure 5 shows the feature importances for the Random Forest model. Feature importance scores are derived from the trained model and indicate how much each feature contributes to the model's decision-making process. The importance scores are normalized, such that they sum up to 1. PageValues emerges as the most influential feature, contributing approximately 37.33% to the model's predictions. This suggests that the monetary value of pages viewed is a crucial predictor of whether a session results in a purchase. Users who engage with higher-value pages are more likely to make a purchase. ProductRelated_Duration comes second with an importance score of 9.25%, which aligns with the intuition that users who spend more time on product-related pages are more likely to convert into buyers. ExitRates and ProductRelated also play significant roles in predicting purchase behavior, with importance scores of 8.81% and 7.46%, respectively. This indicates that the likelihood of exiting the session and the number of product-related pages viewed both contribute meaningfully to the prediction of a purchase. Other features, such as BounceRates, Administrative_Duration, and Month, have lower importance scores, but they still provide useful information for the model's predictions. This feature importance analysis highlights the key factors that drive purchasing behavior in online shoppers. By focusing on PageValues and ProductRelated_Duration, e-commerce businesses can better target users who are more likely to make a purchase.

4.3. Feature Importance and Insights

The Random Forest model provided valuable insights into the most influential features driving purchasing behavior. As mentioned earlier, PageValues emerged as the most important feature, suggesting that the financial value of pages viewed plays a significant role in determining whether a session will convert to a purchase. The high importance of ProductRelated_Duration also supports the idea that the more time users spend browsing product pages, the more likely they are to make a purchase. Other important features included ExitRates and BounceRates, indicating that visitors who engage with more pages and spend more time per page are more likely to purchase. These insights can guide e-commerce businesses in refining their customer engagement strategies. For instance, optimizing product page content to increase time spent on those pages or improving the flow of the website to reduce exit rates could enhance conversion rates. Moreover, understanding the importance of PageValues suggests that targeted promotions or upselling strategies on higher-value pages could increase the likelihood of purchase. In summary, the results demonstrate that machine learning models, particularly Random Forest, can be effectively used to predict online shopper purchasing intention. The feature importance analysis provides actionable insights that e-commerce businesses can use to refine their strategies and improve customer conversion.

5. Conclusion

The results of this study reveal that Random Forest outperforms Support Vector Machine (SVM) in predicting online shopper purchase intentions. The Random Forest model achieved an accuracy of 90.43%, with an impressive ROC AUC score of 0.94, indicating its strong ability to distinguish between sessions that lead to purchases and those that do not. On the other hand, the SVM model performed slightly less effectively, with an accuracy of 89.25% and a ROC AUC score of 0.86. Feature importance analysis showed that PageValues and ProductRelated_Duration were the most significant predictors of purchase behavior. These findings highlight the key factors influencing purchasing decisions, with longer engagement on product-related pages and higher-value page interactions being the most telling signs of purchase intent. This study contributes to the growing body of research on machine learning applications in e-commerce by demonstrating how Random Forest and SVM can effectively predict online shopper behavior. By utilizing commonly available session-based features like PageValues, ProductRelated_Duration, and ExitRates, the study provides valuable insights into the factors that drive conversions. The analysis of feature importance further deepens our understanding of the specific session behaviors that indicate a higher likelihood of purchase. This research paves the way for more targeted e-commerce strategies, where businesses can focus on key user behaviors to optimize marketing efforts and improve sales conversion rates.

While the study provides valuable insights, there are certain limitations to consider. One limitation is the class imbalance in the dataset, where the majority of sessions do not result in a purchase, which could skew the model's ability to predict minority class instances accurately. Additionally, the study relies on a predefined set of features, and other potentially influential factors, such as user demographics or prior purchase history, were not included. Future research could focus on addressing the class imbalance using techniques like SMOTE or exploring advanced deep learning models like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), which might capture more complex patterns in the data. Additionally, incorporating more granular customer data, such as user preferences or purchase history, could further enhance the prediction models and improve their applicability in real-world e-commerce settings. For businesses, the findings of this study offer actionable insights that can be directly applied to optimize e-commerce strategies. By focusing on the most predictive features, such as PageValues and ProductRelated_Duration, businesses can design targeted marketing campaigns that engage users who are more likely to convert. For instance, improving the quality and visibility of product pages with higher monetary value or optimizing the website flow to encourage users to spend more time on product-related content can increase the likelihood of a purchase. Additionally, understanding the importance of ExitRates and BounceRates can help businesses reduce session abandonment by improving user experience and engagement. By leveraging these insights, e-commerce companies can fine-tune their marketing efforts, increase conversions, and enhance overall user experience.

6. Declarations

6.1. Author Contributions

Conceptualization: R.A., S.W.; Methodology: S.W.; Software: R.A.; Validation: R.A., S.W.; Formal Analysis: R.A., S.W.; Investigation: R.A.; Resources: S.W.; Data Curation: S.W.; Writing—Original Draft Preparation: R.A., S.W.; Writing—Review and Editing: S.W., R.A.; Visualization: R.A. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] U. Erdoğan, "A Systematic Review on the Use of Artificial Intelligence in E-Commerce," *Toplum Ekon. Ve Önetim Derg.*, 2023, doi: 10.58702/teyd.1357551.

[2] C. Lei, S. Ji, and Z. Li, "TiSSA: A Time Slice Self-Attention Approach for Modeling Sequential User Behaviors," *in Proc. of the World Wide Web Conference*, 2019. doi: 10.1145/3308558.3313495.

[3] G. Șoavă, A. Mehedințu, and M. Sterpu, "Analysis and Forecast of the Use of E-Commerce in Enterprises of the European Union States," *Sustainability*, 2022, doi: 10.3390/su14148943.

[4] J. Tan and S. Ludwig, "Regional Adoption of Business-to-Business Electronic Commerce in China," *Int. J. Electron. Commer.*, 2016, doi: 10.1080/10864415.2016.1122438.

[5] X. Chen and C. Wang, "Online Arbitration of E-Commerce Disputes in the People's Republic of China: Due Process Concerns," *China Wto Rev.*, 2022, doi: 10.14330/cwr.2022.8.2.01.

[6] M. W. Hasanat, A. Hoque, and A. B. Abdul Hamid, "E-Commerce Optimization With the Implementation of Social Media and SEO Techniques to Boost Sales in Retail Business," *J. Mark. Inf. Syst.*, 2020, doi: 10.31580/jmis.v3i1.1193.

[7] S. Revinova, "E-Commerce Effects for the Sustainable Development Goals," *SHS Web Conf.*, 2021, doi: 10.1051/shsconf/202111401013.

[8] S. N. Rasyida, "The Impact of Indonesian Muslim Consumers Hedonic and Trust on the Online Purchase Intention With Attitude as Intervening Variable (Case Study at Shopee Marketplace)," *Iqtishodia J. Ekon. Syariah*, 2021, doi: 10.35897/iqtishodia.v6i1.520.

[9] G. C. Premananto, T. Basuki, S. Hartini, M. Kurniawati, and A. A. Rashid, "The Effects of E-Wom, Information Overload, Attitude Towards Online Purchase, and Consumer Psychological Condition on the Intention Towards Online Purchase of Laptop Product," *Majcafe*, 2023, doi: 10.60016/majcafe.v31.30.

[10] L. Chen, Md. S. Rashidin, F. Song, W. Yi, S. Javed, and J. Wang, "Determinants of Consumer's Purchase Intention on Fresh E-Commerce Platform: Perspective of UTAUT Model," *Sage Open*, 2021, doi: 10.1177/21582440211027875.

[11] L. T. Hai Ha, L. M. Hien, and P. T. Thuy Van, "The Factors Impact on Online Purchase Intention: Evidence From Intermediate Role of Customers' Trust," *Int. J. Manag. Econ. Invent.*, 2023, doi: 10.47191/ijmei/v9i11.01.

[12] S. Gu, B. Ślusarczyk, S. Hajizada, I. N. Kovalyova, and A. Sakhbieva, "Impact of the COVID-19 Pandemic on Online Consumer Purchasing Behavior," *J. Theor. Appl. Electron. Commer. Res.*, 2021, doi: 10.3390/jtaer16060125.

[13] E. Sumarliah, S. U. Khan, and I. U. Khan, "Online Hijab Purchase Intention: The Influence of the Coronavirus Outbreak," *J. Islam. Mark.*, 2021, doi: 10.1108/jima-09-2020-0302.

[14] T. T. Nguyen, H. T. Thu Truong, and T. Le-Anh, "Online Purchase Intention Under the Integration of Theory of Planned Behavior and Technology Acceptance Model," *Sage Open*, 2023, doi: 10.1177/21582440231218814.

[15] M. Fachmi and H. Sinau, "The Effect of Online Costumer Reviews and Influencer Marketing on Shopee Purchasing Decisions," *Terbuka J. Econ. Bus.*, 2022, doi: 10.33830/tjeb.v3i2.4206.

[16] C. Hartono, Y. B. Rahayu Silintowe, and A. D. Huruta, "The Ease of Transaction and E-Service Quality of E-Commerce Platform on Online Purchasing Decision," *Bisma Bisnis Dan Manaj.*, 2021, doi: 10.26740/bisma.v13n2.p81-93.

[17] Md. B. Rahman and S. Sultana, "Factors Influencing Purchasing Behavior of Mobile Phone Consumers: Evidence From Bangladesh," *Open J. Soc. Sci.*, 2022, doi: 10.4236/jss.2022.107001.

[18] V. L. Miguéis and R. Teixeira, "Predicting Market Basket Additions as a Way to Enhance Customer Service Levels," in *Proceedings of the International Conference on Exploring Services Science*, vol. 11706, pp. 121–134, 2020. doi: 10.1007/978-3-030-38724-2_9.

[19] E. V. Sinitsyn and A. Tolmachev, "Model of the Decision Support System on the Financial Markets for Enterprises Based on Probability Analysis and Machine Learning," *Bull. Ural Fed. Univ. Ser. Econ. Manag.*, 2019, doi: 10.15826/vestnik.2018.17.3.019.

[20] W. Wang *et al.*, "A User Purchase Behavior Prediction Method Based on XGBoost," *Electronics*, 2023, doi: 10.3390/electronics12092047.

[21] S. Rhouas, A. E. Attaoui, and N. E. Hami, "Optimization of the Prediction Performance in the Future Exchange Rate," *in Proc. 2023 9th Int. Conf. on Optimization and Applications (ICOA)*, pp. 1–6, 2023. doi: 10.1109/ICOA58279.2023.10308858.

[22] W. Mostert, K. M. Malan, and A. P. Engelbrecht, "A Feature Selection Algorithm Performance Metric for Comparative Analysis," *Algorithms*, 2021, doi: 10.3390/a14030100.

[23] D. L. Dan Liu *et al.*, "Research on Mutual Information Feature Selection Algorithm Based on Genetic Algorithm," 電腦學刊, 2022, doi: 10.53106/199115992022123306011.

[24] N. Hoque, M. Singh, and D. K. Bhattacharyya, "EFS-MI: An Ensemble Feature Selection Method for Classification," *Complex Intell. Syst.*, 2017, doi: 10.1007/s40747-017-0060-x.

[25] Y. Akhiat, M. Chahhou, and A. Zinedine, "Ensemble Feature Selection Algorithm," *Int. J. Intell. Syst. Appl.*, 2019, doi: 10.5815/ijisa.2019.01.03.