E-Commerce Buyer Decision Prediction Model

EDA & Supervised Learning Model of Online Shoppers Purchasing Intention Dataset

Table of Contents

2. Name of the Dataset 3
3. Problem Statement 3
4. Objectives 3
Exploratory Data Analysis (EDA) 3
Classification Using Supervised Learning 3
Model Evaluation and Selection 3
5. Data Description 4
6. Exploratory Data Analysis of Online Shoppers Intention Dataset 5
Univariate Analysis 5
Output of Histograms 6
Interpretation of Histograms 6
Boxplot Output: 8
Interpretation of Boxplots 8
7. Bivariate Analysis – Scatter Diagrams 9
8. Bivariate Analysis – Heatmap 10
9. Classifying the 'Online Shoppers Purchasing Intention' Dataset Using Different Machine Learning Models 11
10. Logistic Regression 11
Interpretation- Logistic Regression Results 12
11. Decision Tree 13
Interpretation – DT results 13
12. Random Forest 15
Interpretation – Random Forest Results 15
Comments on the Result: 16
13. Gradient Boosting Machine (GBM) 16
Interpretation - GBM 16
14. k-Nearest Neighbor (kNN) 17
Interpretation – kNN Results 17
15. Support Vector Machine (SVM) 18
Interpretation – SVM Results 18
16. Model Evaluation and Selection 19
Comparison of the ML Models 19
Performance Analysis of the models 20
Selecting the Best Model 20
17. Conclusion 20

1. Name of the Dataset

Online Shoppers Purchasing Intention Dataset

2. Problem Statement

Predict whether a visitor to an e-commerce website will generate revenue by making a purchase during their session based on features such as the number of pages visited, session duration, type of visitor, operating system, browser, and other relevant attributes.

3. Objectives

The main objective of this study is to explore the online shoppers purchasing intention dataset, conduct exploratory data analysis, develop and assess classification models, and determine the most effective model for predicting buyers' purchasing intention. Thus, objectives are the following:

Exploratory Data Analysis (EDA)

- Understand the structure and characteristics of the "Online Shoppers Purchasing Intention Dataset".
- Explore the distribution of various features such as number of pages visited, session duration, type of visitor, operating system, browser, etc.
- Identify any patterns, trends, or anomalies in the data.
- Visualize relationships between different features and the target variable (revenue generation).

Classification Using Supervised Learning

- Preprocess the dataset by handling missing values, encoding categorical variables, and scaling numerical features if necessary.
- Split the dataset into training and testing sets.
- Select appropriate supervised learning algorithms for classification, such as logistic regression, decision trees, random forests, support vector machines, etc.
- Train multiple models using different algorithms and hyperparameters.
- Evaluate the performance of each model using appropriate evaluation metrics such as accuracy, precision, recall, and ROC AUC.
- Fine-tune the hyperparameters of the best-performing models using techniques like grid search or random search.

Model Evaluation and Selection

- Compare the performance of different classification models based on evaluation metrics.
- Select the best-performing model based on the evaluation results.
- Validate the selected model using cross-validation techniques to ensure its robustness.
- Interpret the results and provide insights into which features are most influential in predicting revenue generation.
- Optionally, deploy the selected model for real-time predictions if applicable.

The study will provide interpretations of the implications of the selected model's performance, considering its effectiveness in predicting revenue generation.

4. Data Description

The Online Shoppers Purchasing Intention Dataset contains information about various attributes of white wines, with a focus on factors that might influence their quality. The dataset consists of a total of 12330 instances and 18 columns.

Variable Name	Variable Description	Value Level	Measurements	Level of Appropriate Measures
Administrative	Number of administrative pages visited during the session	Integer	Discrete	Ordinal/Nominal
Administrative_ Duration	Total duration of administrative page visits during the session	Integer	Continuous	Interval/Ratio
Informational	Number of informational pages visited during the session	Integer	Discrete	Ordinal/Nominal
Informational_Duration	Total duration of informational page visits during the session	Integer	Continuous	Interval/Ratio
Product Related	Number of product-related pages visited during the session	Integer	Discrete	Ordinal/Nominal
Product Related_ Duration	Total duration of product-related page visits during the session	Continuous	Continuous	Interval/Ratio
Bounce Rates	Bounce rate of the website, i.e., the percentage of visits in which the visitor leaves the website without browsing further	Continuous	Continuous	Interval/Ratio
Exit Rates	Exit rate of the website, i.e., the percentage of exits from the website that occurred on a particular page	Continuous	Continuous	Interval/Ratio

Variable Name	Variable Description	Value Level	Measurements	Level of Appropriate Measures
Page Values	Average value of the web page visited by the user before completing an e-commerce transaction	Integer	Continuous	Interval/Ratio
Special Day	Indicates if the visit occurred on a special day (e.g., Mother's Day, Valentine's Day)	Integer	Discrete	Ordinal/Nominal
Month	Month of the visit	Categorical	Nominal	Nominal
Operating Systems	Operating system of the visitor	Integer	Discrete	Ordinal/Nominal
Browser	Browser used by the visitor	Integer	Discrete	Ordinal/Nominal
Region	Geographic region of the visitor	Integer	Discrete	Ordinal/Nominal
Traffic Type	Type of traffic source	Integer	Discrete	Ordinal/Nominal
Visitor Type	Type of visitor (e.g., returning visitor, new visitor)	Categorical	Nominal	Nominal
Weekend	Indicates if the visit occurred on a weekend (1 for yes, 0 for no)	Binary	Nominal	Nominal
Revenue	Indicates if a purchase was made during the session (1 for yes, 0 for no)	Binary	Nominal	Nominal

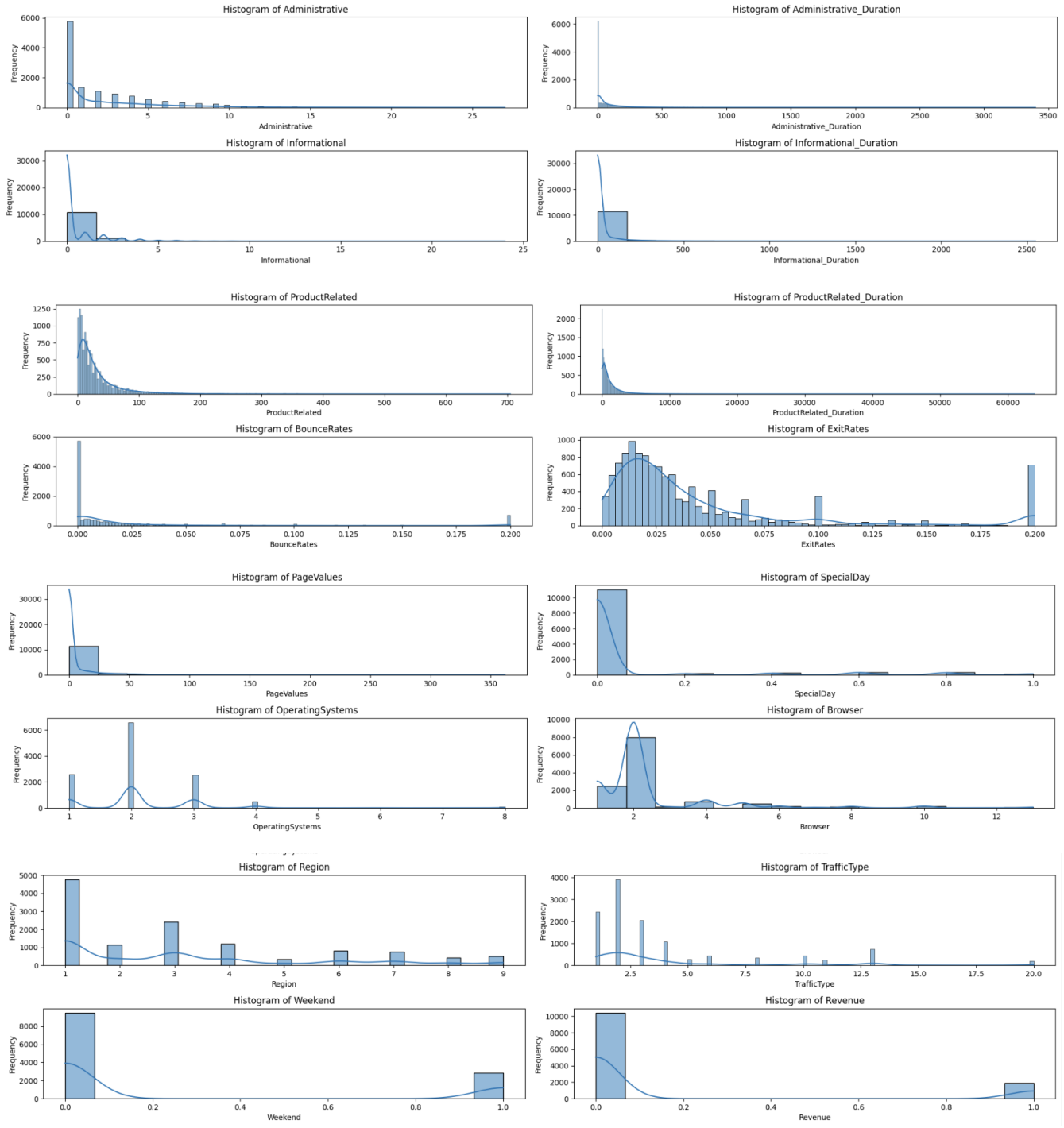
Table 1: Variable's Summary Information of Online Shoppers Purchasing Decision Dataset

5. Exploratory Data Analysis of Online Shoppers Intention Dataset

Univariate Analysis

The univariate analysis helps in understanding the characteristics and distribution of individual variables in the dataset. It provides initial insights into the data structure, potential outliers, and data quality issues that may need further investigation.

Output of Histograms



Interpretation of Histograms

1. Histogram of Administrative: The histogram of Administrative displays the distribution of administrative page visits during online sessions. Most sessions involve a lower number of administrative page visits, with a peak around 1 to 3 visits.

The range extends from 0 to approximately 27 visits, showcasing variability in the number of administrative interactions across sessions.

2. Histogram of Administrative_Duration: The histogram of Administrative_Duration illustrates the distribution of durations for administrative page visits. Duration values are predominantly shorter, with a peak around 0 to 100 seconds. The range spans from 0 to over 3,600 seconds (1 hour), indicating occasional longer durations for administrative interactions.

3. Histogram of Informational: The histogram of Informational represents the distribution of informational page visits per session. Most sessions involve fewer informational interactions, concentrated around 0 to 2 visits. Some sessions show higher informational engagement, with the range extending up to approximately 24 visits.

4. Informational_Duration: The histogram of Informational_Duration visualizes the distribution of durations for informational page visits. Duration values are typically short, with a peak around 0 to 100 seconds. The range varies from 0 to over 2,000 seconds, suggesting occasional longer durations for informational content.

5. ProductRelated: The histogram of ProductRelated depicts the distribution of product-related page visits during sessions. Majority of sessions involve a moderate number of product-related interactions, with a peak around 0 to 200 visits. Some sessions exhibit higher engagement with product-related content, with the range extending up to over 2,000 visits.

6. ProductRelated_Duration: The histogram of ProductRelated_Duration showcases the distribution of durations for product-related page visits. Duration values are generally short to moderate, with a peak around 0 to 2,000 seconds. The range spans from 0 to over 150,000 seconds (over 41 hours), indicating varying durations for product-related interactions.

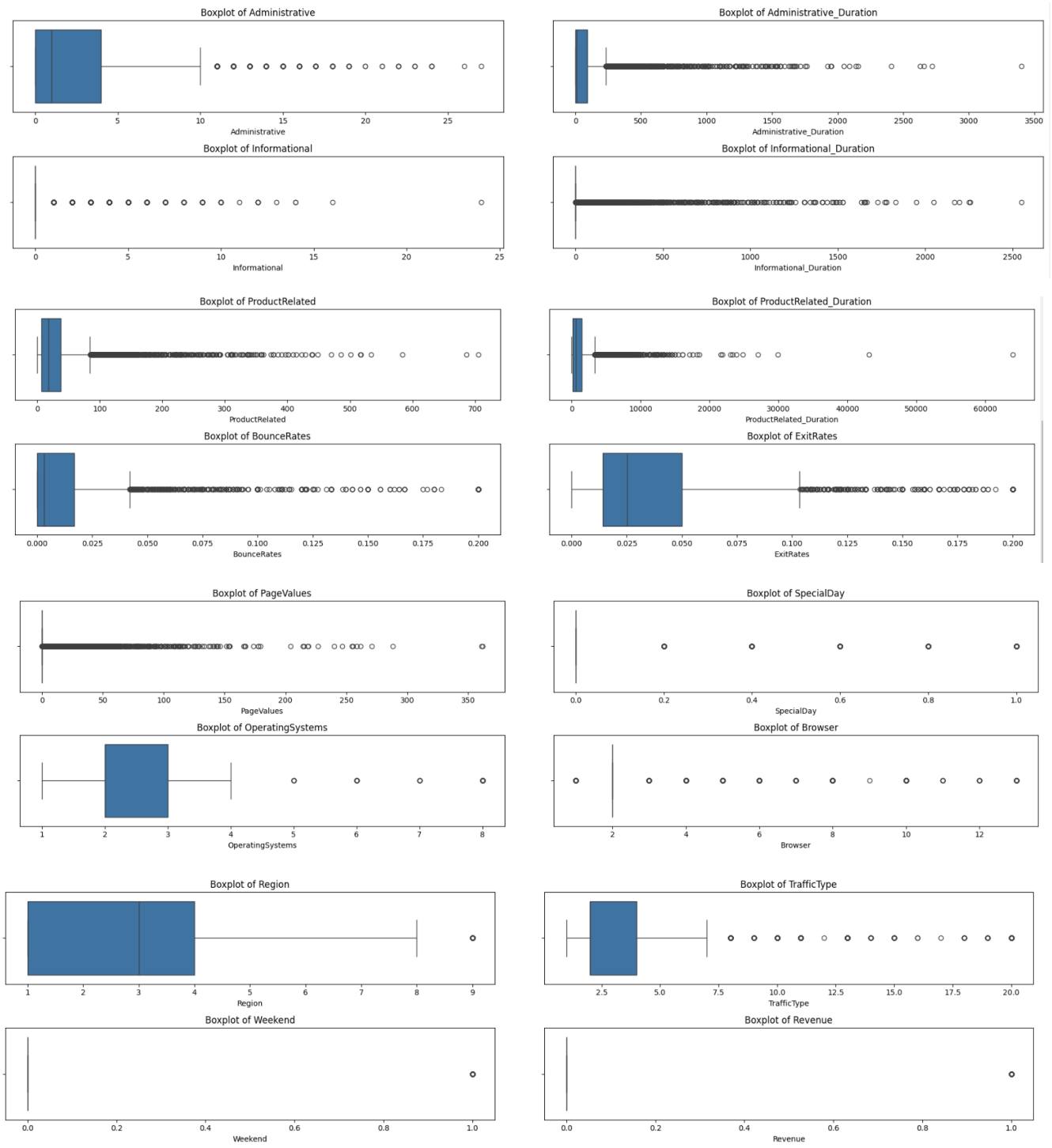
7. BounceRates: Interpretation: The histogram of BounceRates reveals the distribution of bounce rates across sessions. Most sessions have lower bounce rates, concentrated around 0% to 20%. Outliers with higher bounce rates are also observed, extending the range beyond 100%.

8. ExitRates: The histogram of ExitRates displays the distribution of exit rates for page views. Majority of sessions exhibit lower exit rates, with a peak around 0% to 20%. Some sessions show higher exit rates, with outliers extending the range beyond 100%.

9. PageValues: The histogram of PageValues illustrates the distribution of page values perceived by users. Page values are typically lower, concentrated around 0 to 50. Some pages exhibit higher perceived values, with the range extending beyond 250.

10. SpecialDay: The histogram of SpecialDay represents the distribution of special day indicators. Distinct values are observed corresponding to specific events or occasions impacting user behavior. The range showcases variations in special day occurrences, providing insights into seasonal trends or promotions.

Boxplot Output:



Interpretation of Boxplots

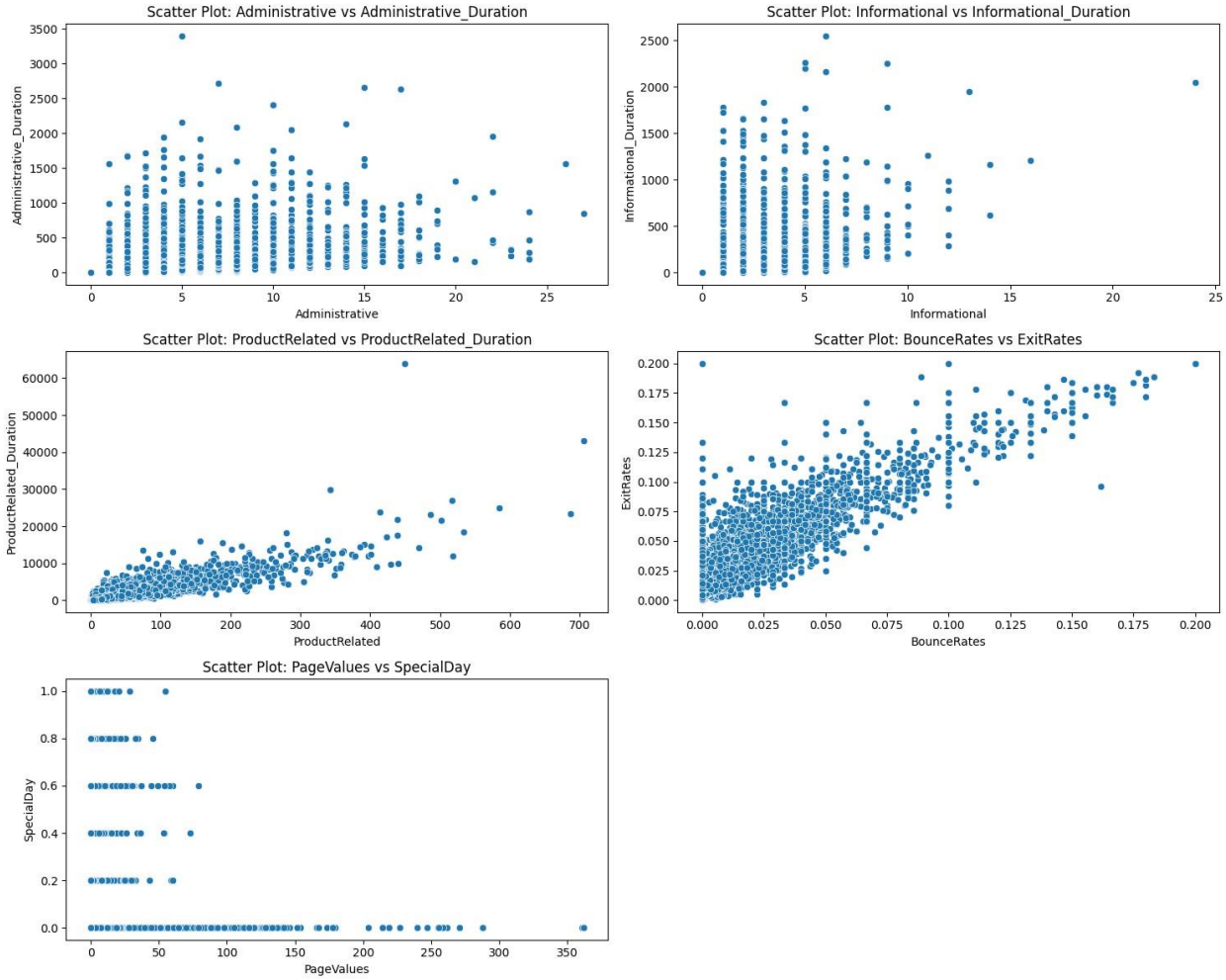
The median (middle line inside the box) represents the typical number per session.

Interquartile Range (IQR): The IQR spans from Q1 (25th percentile) to Q3 (75th percentile), showing the spread of values where most sessions fall.

Outliers: Points beyond the whiskers indicate sessions with exceptionally high numbers of administrative visits.

6. Bivariate Analysis – Scatter Diagrams

Output



Interpretation:

1. Administrative vs Administrative_Duration: This scatter plot shows the relationship between the number of administrative page visits (Administrative) and their respective durations (Administrative_Duration). There appears to be a positive correlation, as sessions with more administrative page visits tend to have longer durations. However, the correlation is not very strong, indicating variability in session durations even with similar numbers of administrative visits.

2. Informational vs Informational_Duration: This scatter plot illustrates the relationship between the number of informational page visits (Informational) and their durations

(Informational_Duration). There is a weak positive correlation, suggesting that sessions with more informational interactions may have slightly longer durations. However, the spread of data points indicates variability in session durations regardless of the number of informational visits.

3. ProductRelated vs ProductRelated_Duration: This scatter plot depicts the relationship between the number of product-related page visits (ProductRelated) and their durations (ProductRelated_Duration). There appears to be a positive correlation, indicating that sessions with more product-related interactions tend to have longer durations. The correlation is relatively stronger compared to other pairs, suggesting a more consistent relationship between product-related visits and durations.

4.BounceRates vs ExitRates: This scatter plot shows the relationship between bounce rates (BounceRates) and exit rates (ExitRates). There seems to be a positive correlation, as higher bounce rates are associated with higher exit rates. The data points exhibit a linear trend, indicating that pages with higher bounce rates also tend to have higher exit rates.

5. PageValues vs SpecialDay: This scatter plot illustrates the relationship between page values (PageValues) and special days (SpecialDay). There doesn't appear to be a clear correlation between page values and special days. Data points are scattered across the plot, suggesting that page values are influenced by factors other than special days.

7. Bivariate Analysis – Heatmap

High Positive Correlation (Dark Red): Cells with dark red color (values close to 1) represent strong positive correlations.

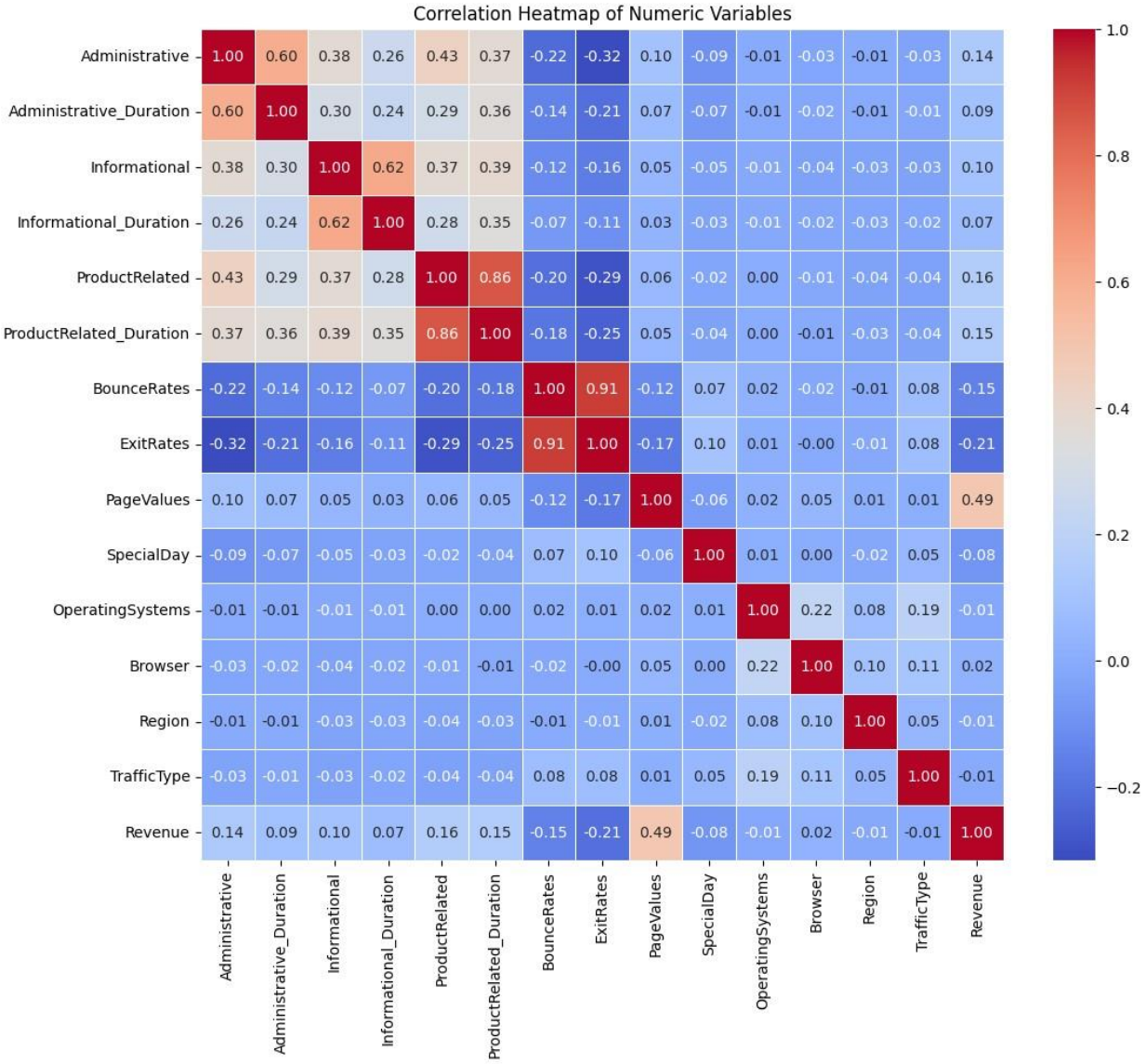
High Negative Correlation (Dark Blue): Cells with dark blue color (values close to -1) represent strong negative correlations.

Weak or No Correlation (Lighter Colors): Cells with lighter colors (values close to 0) indicate weak or no correlation.

Feature Selection:

Using the the heatmap to identify highly correlated variables. Redundant or highly correlated variables may be candidates for feature selection to avoid multicollinearity in predictive modeling.

Here, for the revenue (target variable) we can interpret that comparing with the other variables, PageValues variable has highest correlation with the revenue.



8. Classifying the 'Online Shoppers Purchasing Intention' Dataset Using Different Machine Learning Models

9. Logistic Regression

Code Output - Confusion Matrix:

[[2533 61]
[323 166]]

Accuracy: 0.8754
Precision (Positive Predictive Value): 0.7313
Recall (Sensitivity): 0.3395, Specificity: 0.9765, AUC-ROC: 0.8798

Interpretation- Logistic Regression Results

The confusion matrix reveals important insights into the model's performance:

True Negative (TN): 2533 False Positive (FP): 61

False Negative (FN): 323 True Positive (TP): 166

This breakdown indicates that the model correctly predicted 2533 non-revenue sessions (TN) and 166 revenue sessions (TP). However, it misclassified 61 non-revenue sessions as revenue (FP) and missed 323 revenue sessions (FN).

Accuracy: The achieved accuracy of approximately 87.54% signifies the overall correctness of the model's predictions across both revenue and non-revenue sessions. This metric is calculated as the ratio of correct predictions (TN + TP) to the total number of predictions.

Precision (Positive Predictive Value): The precision score of around 73.13% represents the model's ability to accurately identify revenue-generating sessions among those predicted as revenue. A higher precision score indicates that when the model predicts a session as revenue, it tends to be correct more often.

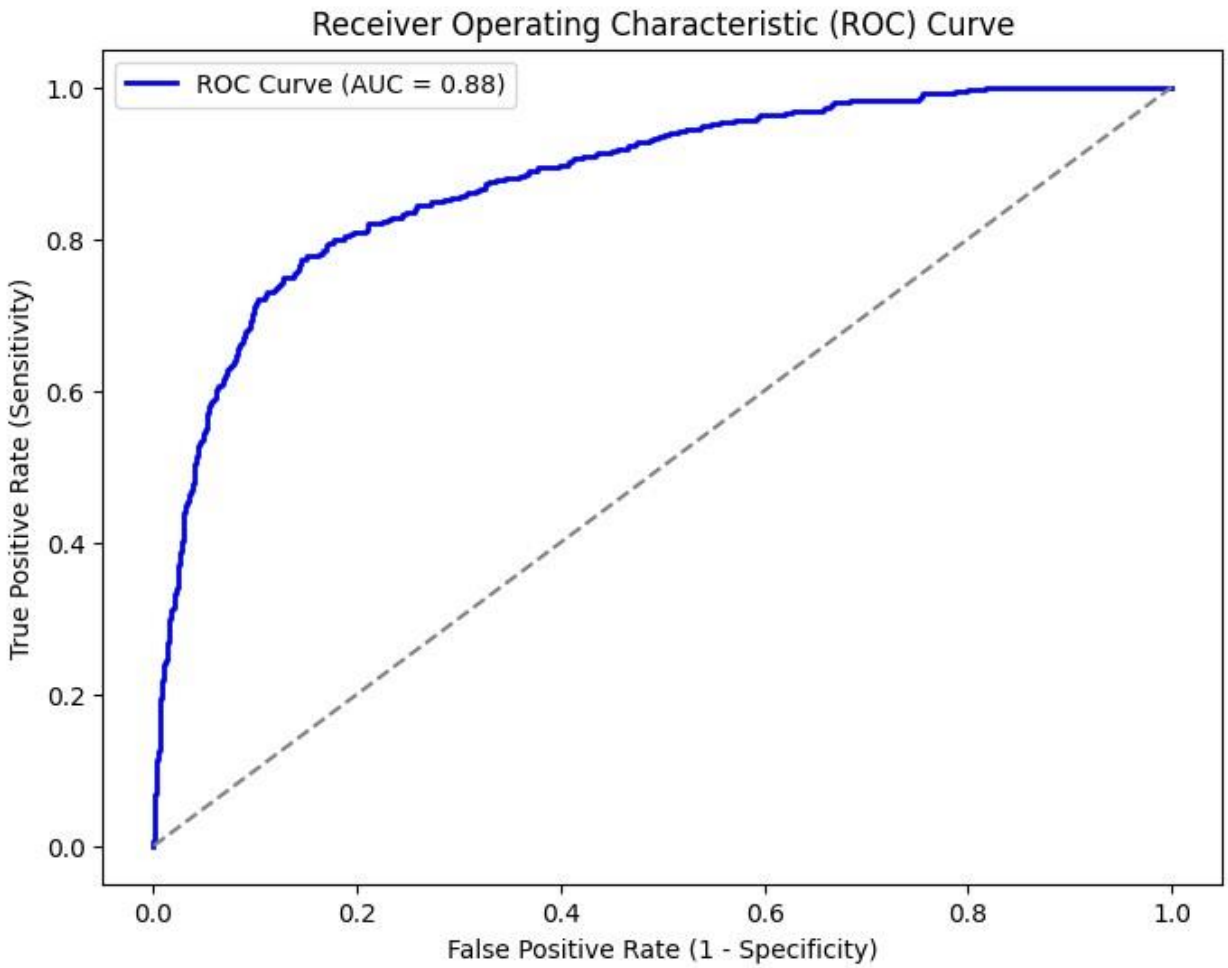
Recall (Sensitivity): The recall score of approximately 33.95% reflects the model's proficiency in capturing actual revenue sessions. This metric highlights the proportion of true revenue sessions correctly predicted by the model out of all actual revenue sessions.

Specificity: With a specificity score of about 97.65%, the model's capability to correctly classify non-revenue sessions is emphasized. A higher specificity score suggests that the model is effective at identifying sessions that truly do not result in revenue.

In summary, while the Logistic Regression model shows promising accuracy and specificity, there is room for improvement in capturing a higher proportion of revenue-generating sessions (improving recall) without compromising precision. This evaluation provides valuable insights for refining the model and optimizing its performance for revenue prediction in e-commerce scenarios.

AUC ROC & ROC Curve:

In summary, an AUC-ROC value of 0.8798 suggests that the Logistic Regression model performs well in distinguishing between revenue and non-revenue sessions, with a strong balance between sensitivity and specificity. The ROC curve visually represents this trade-off and provides insights into the model's performance across different decision thresholds.



10. Decision Tree

Output:

Confusion Matrix:
[[2364 230]
[224 265]]

Accuracy: 0.8527; Precision (Positive Predictive Value): 0.5354
Recall (Sensitivity): 0.5419; Specificity: 0.9113

AUC-ROC: 0.7266

Interpretation – DT results

Accuracy: The achieved accuracy of approximately 85.27% indicates the overall correctness of the model's predictions across both revenue and non-revenue sessions.

Precision (Positive Predictive Value): The precision score of around 53.54% signifies the model's ability to accurately identify revenue-generating sessions among those predicted as revenue. A higher precision score indicates that when the model predicts a session as revenue, it tends to be correct more often.

Recall (Sensitivity): The recall score of approximately 54.19% reflects the model's proficiency in capturing actual revenue sessions. This metric highlights the proportion of true revenue sessions correctly predicted by the model out of all actual revenue sessions.

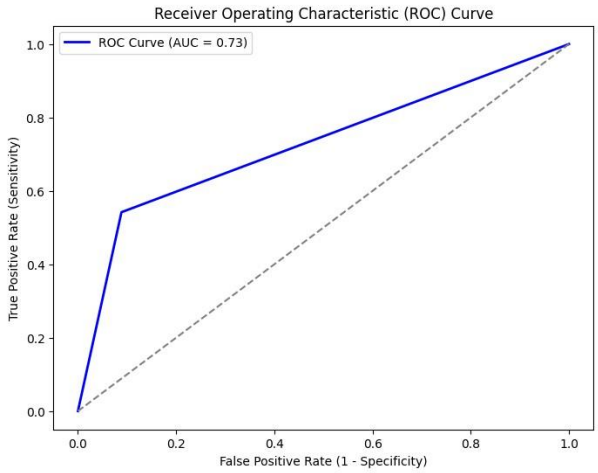
Specificity: With a specificity score of about 91.13%, the model's capability to correctly classify non-revenue sessions is highlighted. A higher specificity score implies that the model is effective at identifying sessions that truly do not result in revenue.

AUC-ROC (Area Under the ROC Curve): The AUC-ROC value of 0.7266 quantifies the overall performance of the model in distinguishing between revenue and non-revenue sessions. This metric measures the area under the Receiver Operating Characteristic (ROC) curve, with a higher value indicating better discrimination ability.

Comments on Result:

- The Decision Tree model demonstrates strong accuracy and specificity, indicating its effectiveness in predicting non-revenue sessions.
- However, the model's precision and recall for revenue sessions are relatively moderate, suggesting room for improvement in correctly identifying revenue-generating sessions without increasing false positives.
- The AUC-ROC value indicates a fair discrimination ability of the model, with potential for further optimization to enhance its performance in revenue prediction on e-commerce platforms.

ROC Curve



11. Random Forest

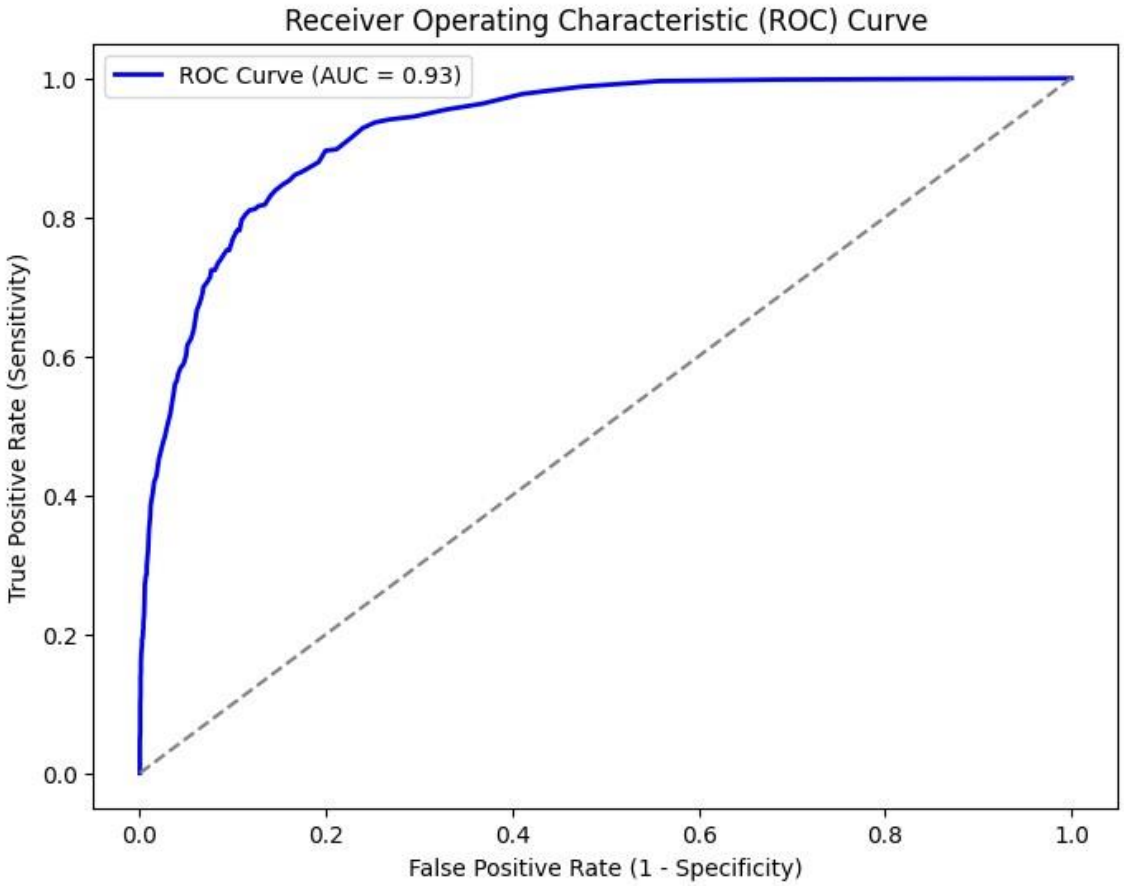
Output

Confusion Matrix:
[[2491 103]
[213 276]]

Accuracy: 0.8975; Precision (Positive Predictive Value): 0.7282
Recall (Sensitivity): 0.5644; Specificity: 0.9603

AUC-ROC: 0.9275

ROC Curve:



Interpretation – Random Forest Results

The Random Forest model achieved a high accuracy of 89.75%, indicating its overall correctness in predicting both positive and negative instances (revenue and non-revenue sessions) from the dataset.

Precision (Positive Predictive Value): The precision score of 72.82% suggests that when the model predicts a session to result in revenue, it is correct approximately 72.82% of the time.

Recall (Sensitivity): The recall score of 56.44% reflects the model's ability to correctly identify revenue-generating sessions out of all actual revenue sessions in the dataset.

Specificity: With a specificity score of 96.03%, the model is highly effective in correctly classifying non-revenue sessions.

AUC-ROC (Area Under the ROC Curve): The AUC-ROC value of 92.75% indicates strong discriminatory performance of the model in distinguishing between revenue and non-revenue sessions.

Comments on the Result:

In summary, the Random Forest model demonstrates good overall accuracy and specificity, although there is room for improvement in recall (sensitivity) to capture more true revenue sessions among the predicted positive instances.

12. Gradient Boosting Machine (GBM)

Output - Confusion Matrix:

[[2481 113]
[207 282]]
Accuracy: 0.8962; Precision (Positive Predictive Value): 0.7139
Recall (Sensitivity): 0.5767; Specificity: 0.9564

AUC-ROC: 0.9296

Interpretation - GBM

Accuracy: The model demonstrated an overall accuracy of 89.62%, indicating a high level of correctness in predicting both revenue and non-revenue sessions.

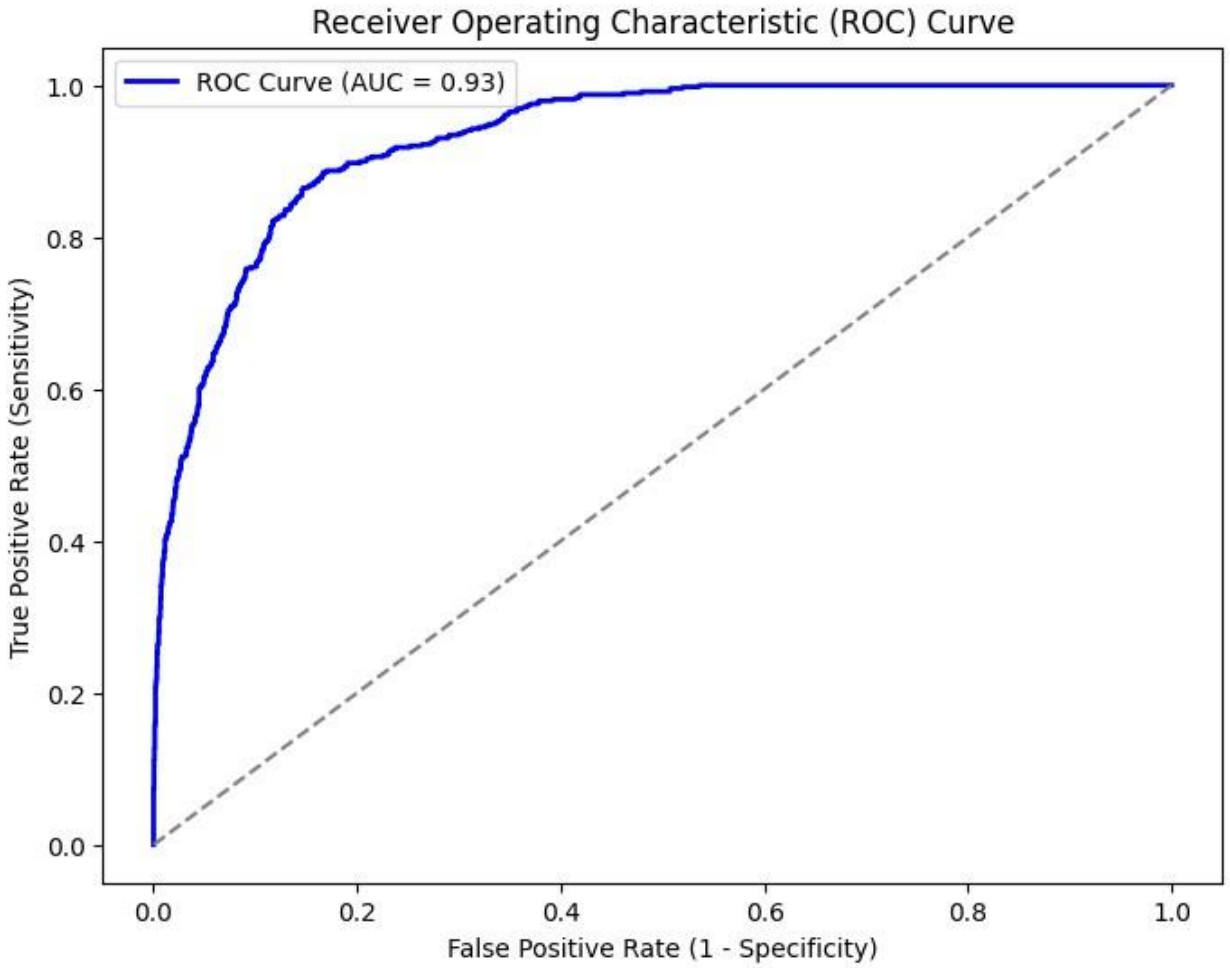
Precision (Positive Predictive Value): The precision score of 71.39% suggests that when the model predicts a session to result in revenue, it is correct approximately 71.39% of the time.

Recall (Sensitivity): The recall score of 57.67% reflects the model's ability to capture a significant portion of actual revenue-generating sessions among all true revenue instances.

Specificity: With a specificity score of 95.64%, the model effectively identifies non-revenue sessions, showcasing its ability to avoid false positive predictions.

AUC-ROC (Area Under the ROC Curve): The AUC-ROC value of 92.96% highlights the strong discriminatory power of the model in distinguishing between revenue and non-revenue sessions.

ROC Curve:



13. k-Nearest Neighbor (kNN)

Output- Confusion Matrix:

[[2510 84]
[309 180]]
Accuracy: 0.8725; Precision (Positive Predictive Value): 0.6818
Recall (Sensitivity): 0.3681; Specificity: 0.9676

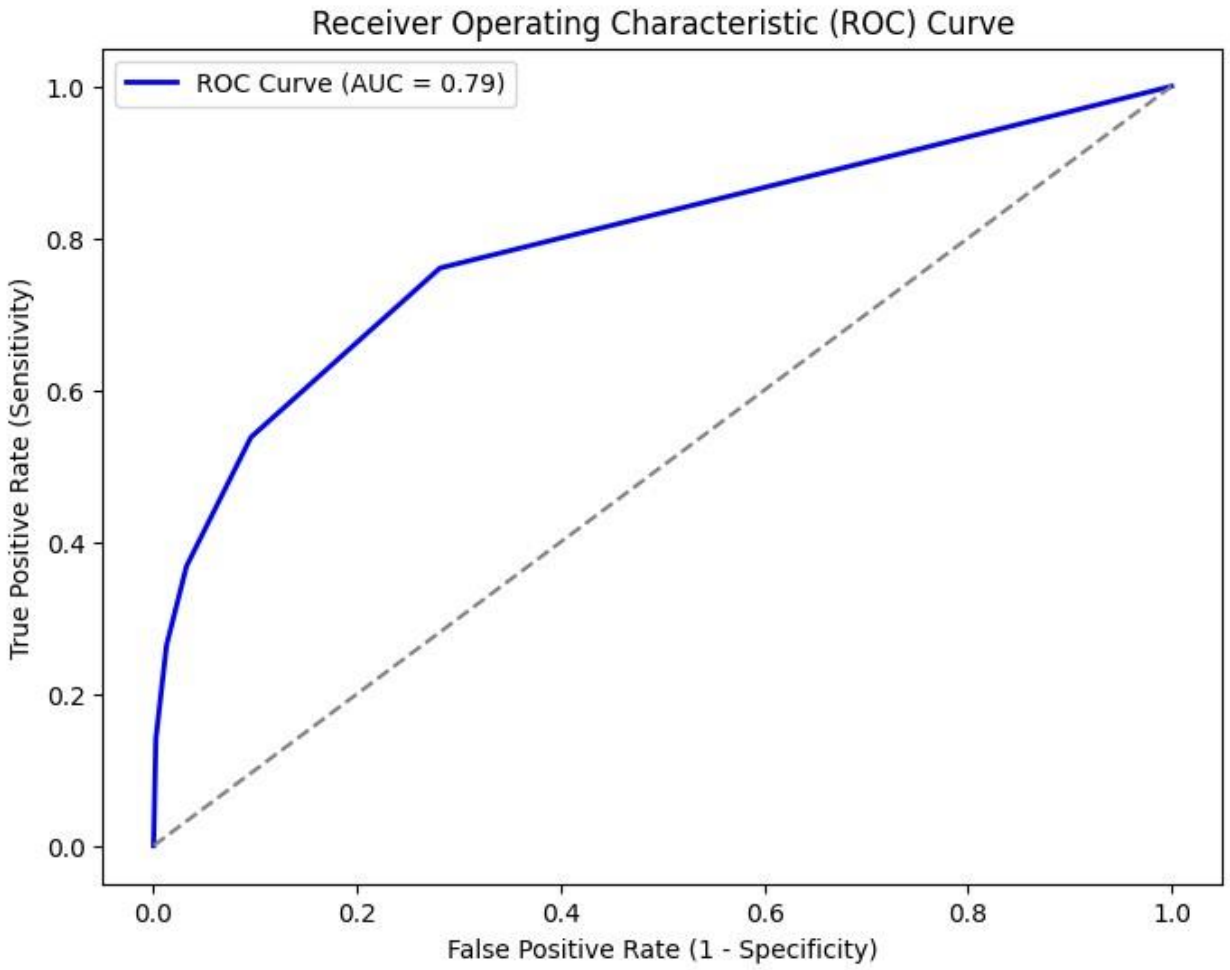
AUC-ROC: 0.7904

Interpretation – kNN Results

The model achieved an accuracy of 87.25%, indicating a relatively high overall correctness in predicting both revenue and non-revenue sessions. The precision score of 68.18% suggests that when the model predicts a session to result in revenue, it is correct approximately 68.18% of the time. The recall score of 36.81% reflects the model's ability to capture a moderate portion of actual revenue-generating sessions among all true revenue instances. The AUC-ROC value of 79.04%

indicates the model's moderate discriminatory power in distinguishing between revenue and non-revenue sessions.

ROC Curve:



14. Support Vector Machine (SVM)

Confusion Matrix:

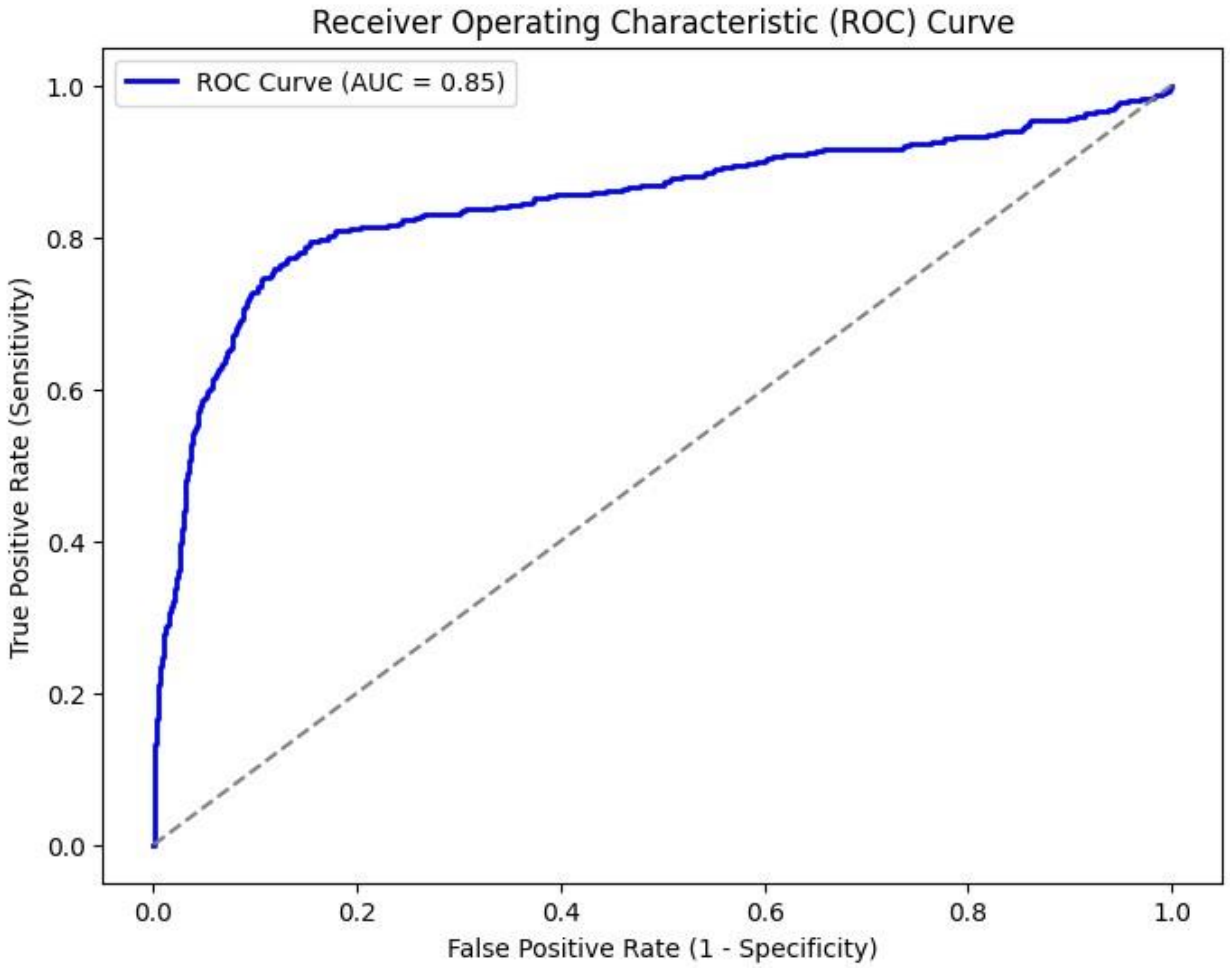
[[2516 78]
[275 214]]
Accuracy: 0.8855; Precision (Positive Predictive Value): 0.7329
Recall (Sensitivity): 0.4376; Specificity: 0.9699
AUC-ROC: 0.8472

Interpretation – SVM Results

The model demonstrated an accuracy of 88.55%, indicating a relatively high overall correctness in predicting both revenue and non-revenue sessions. The precision score of 73.29% suggests that

when the model predicts a session to result in revenue, it is correct approximately 73.29% of the time. The recall score of 43.76% reflects the model's ability to capture a moderate portion of actual revenue-generating sessions among all true revenue instances. With a specificity score of 96.99%, the model effectively identifies non-revenue sessions, showcasing its ability to avoid false positive predictions. The AUC-ROC value of 84.72% indicates the model's moderate discriminatory power in distinguishing between revenue and non-revenue sessions.

ROC Curve



15. Model Evaluation and Selection

Comparison of the ML Models

Model	Accuracy	Precision	Recall (Sensitivity)	Specificity	AUC-ROC
Logistic Regression	0.8754	0.7313	0.3395	0.9765	0.8798
Decision Tree	0.8527	0.5354	0.5419	0.9113	0.7266
Random Forest	0.8975	0.7282	0.5644	0.9603	0.9275
Gradient Boosting	0.8962	0.7139	0.5767	0.9564	0.9296

Model	Accuracy	Precision	Recall (Sensitivity)	Specificity	AUC-ROC
k-Nearest Neighbor	0.8725	0.6818	0.3681	0.9676	0.7904
Support Vector Machine	0.8855	0.7329	0.4376	0.9699	0.8472

Performance Analysis of the models

Accuracy: Random Forest achieved the highest accuracy of 89.75%, closely followed by Gradient Boosting (89.62%) and Logistic Regression (87.54%).

Precision (Positive Predictive Value): Random Forest exhibited the highest precision at 72.82%, indicating its effectiveness in predicting revenue sessions among those predicted as revenue. Logistic Regression also showed strong precision at 73.13%.

Recall (Sensitivity): Gradient Boosting demonstrated the highest recall rate (57.67%), indicating its ability to capture a substantial proportion of actual revenue sessions among all true revenue instances.

Specificity: The k-Nearest Neighbor (kNN) model achieved the highest specificity (96.76%), indicating its capability to correctly identify non-revenue sessions.

AUC-ROC: Gradient Boosting and Random Forest exhibited the highest AUC-ROC values (92.96% and 92.75%, respectively), suggesting superior discriminatory power in distinguishing between revenue and non-revenue sessions.

Selecting the Best Model

- Random Forest and Gradient Boosting emerged as the top-performing models overall, showcasing strong accuracy, precision, recall, specificity, and AUC-ROC.
- Logistic Regression, while achieving decent performance in accuracy and precision, demonstrated relatively lower recall compared to ensemble methods like Random Forest and Gradient Boosting.
- Decision Tree and k-Nearest Neighbor (kNN) models showed competitive performance but generally lagged behind in key metrics like accuracy, recall, and AUC-ROC compared to Random Forest and Gradient Boosting.
- Support Vector Machine (SVM) exhibited strong specificity, but relatively lower recall compared to ensemble methods.

16. Conclusion

In conclusion, Random Forest and Gradient Boosting are recommended for this prediction task based on their comprehensive performance across multiple evaluation metrics. However, the choice of the best model may ultimately depend on specific objectives and considerations related to the application context.