

# Customer Purchase Prediction Report

## 1. Data Preprocessing Steps

### 1.1 Data Loading and Initial Exploration

- The dataset `online_shoppers_intention.csv` was loaded using pandas.
- The dataset contains **12,330 rows** and **18 columns**, including numerical, categorical, and boolean features.
- Basic exploration was performed using `.head()`, `.info()`, and `.describe()` to understand the structure and summary statistics of the data.

### 1.2 Handling Missing Values

- Missing values were checked using `.isnull().sum()`. No missing values were found in the dataset.

### 1.3 Encoding Categorical Variables

- One-hot encoding was applied to categorical variables (Month, OperatingSystems, Browser, Region, TrafficType, VisitorType, Weekend) to convert them into numerical format for modeling.

### 1.4 Feature Creating

- **New Features Created:**
  - `TimeSpentOnSite`: Sum of `ProductRelated_Duration`, `Administrative_Duration`, and `Informational_Duration` to capture total user engagement.
  - `VisitorType_New`: Derived from `VisitorType_Returning_Visitor` to simplify the analysis of returning visitors.

### 1.5 Normalization

- Normalization of the continuous variables `Administrative`, `ProductRelated`, and `BounceRates` using `MinMaxScaler`.

## 2. Exploratory Data Analysis (EDA) Findings

### 2.1 Distribution of Numerical Variables

- **Administrative Page Visits:**

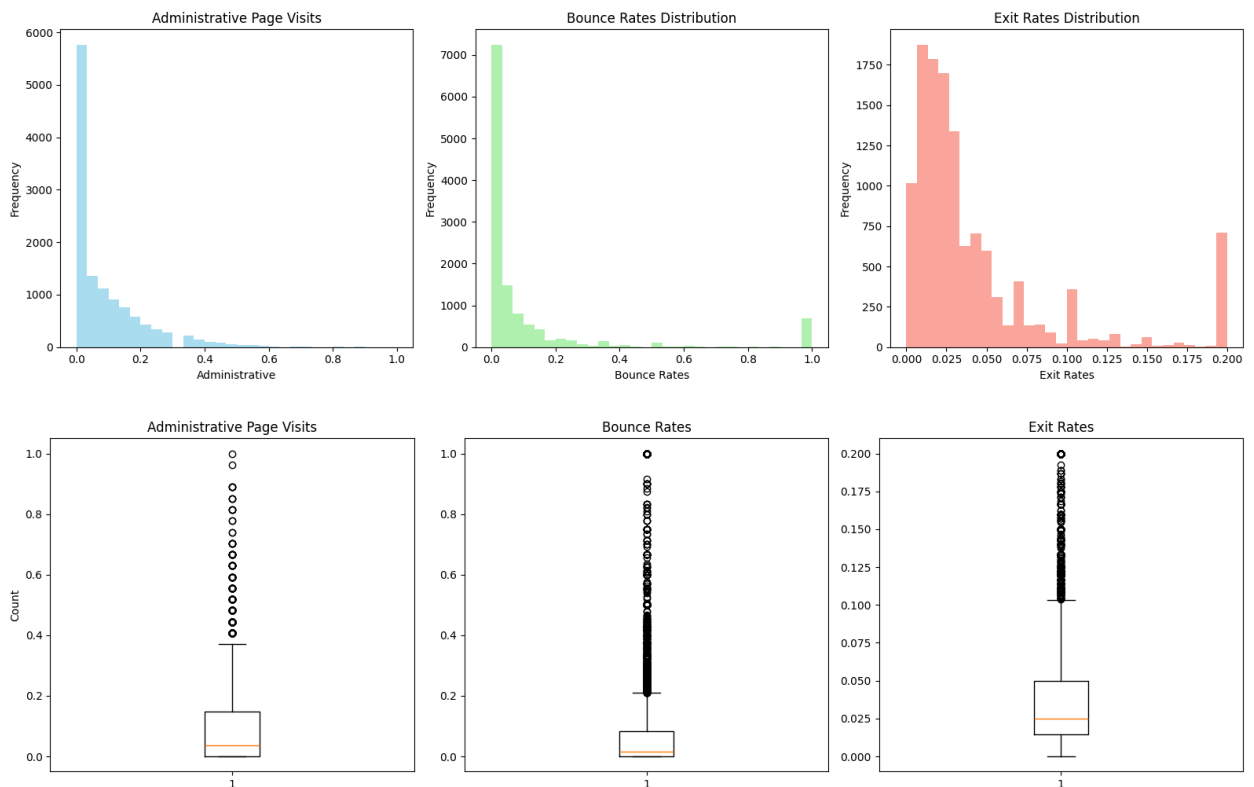
The histogram shows a right-skewed distribution, with most users visiting a small number of administrative pages. The box plot confirms the presence of outliers, indicating some users interact significantly more with administrative pages.

- **Bounce Rates:**

The histogram reveals a right-skewed distribution, with most users having low bounce rates. The box plot shows a few outliers, indicating some users have unusually high bounce rates.

- **Exit Rates:**

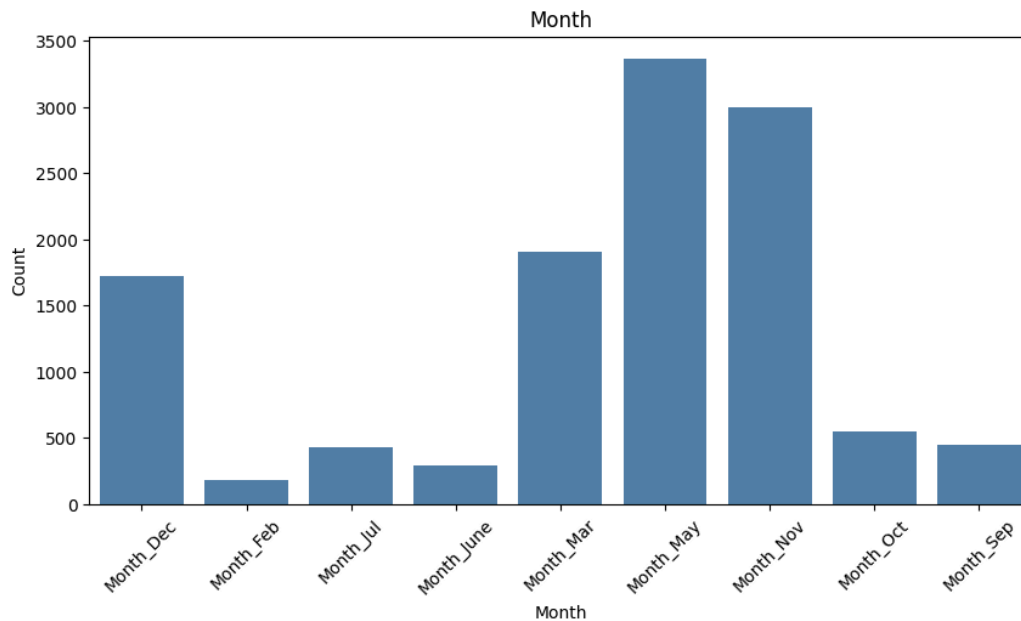
The histogram also shows a right-skewed distribution, with most users having low exit rates. The box plot highlights outliers, suggesting some users exit the site at a much higher rate.



## 2.2 Distribution of Categorical Variables

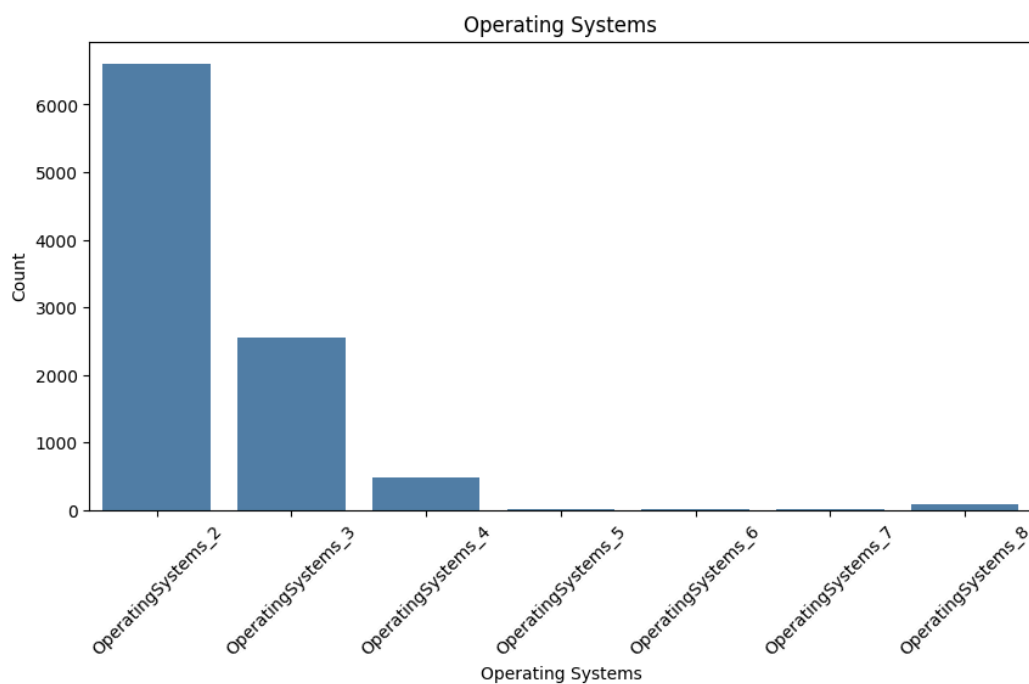
- **Month:**

The bar chart shows that visits are highest in **May** and **November**, indicating seasonal trends in user activity.



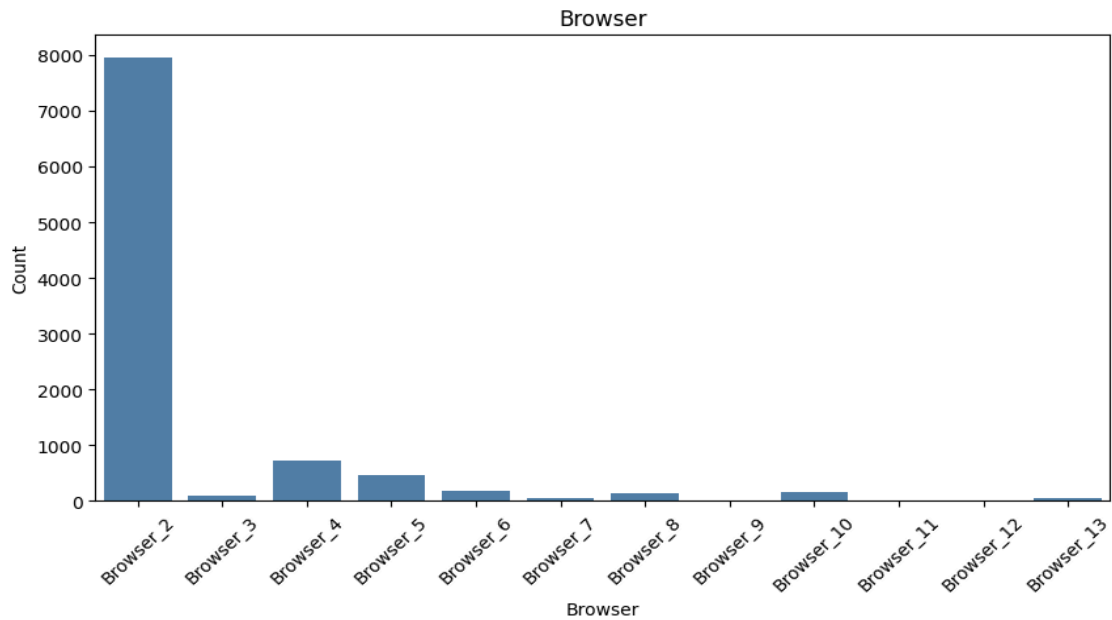
- **Operating Systems:**

The bar chart reveals that most users use **Operating System 2**, followed by **Operating System 3** and **4**.



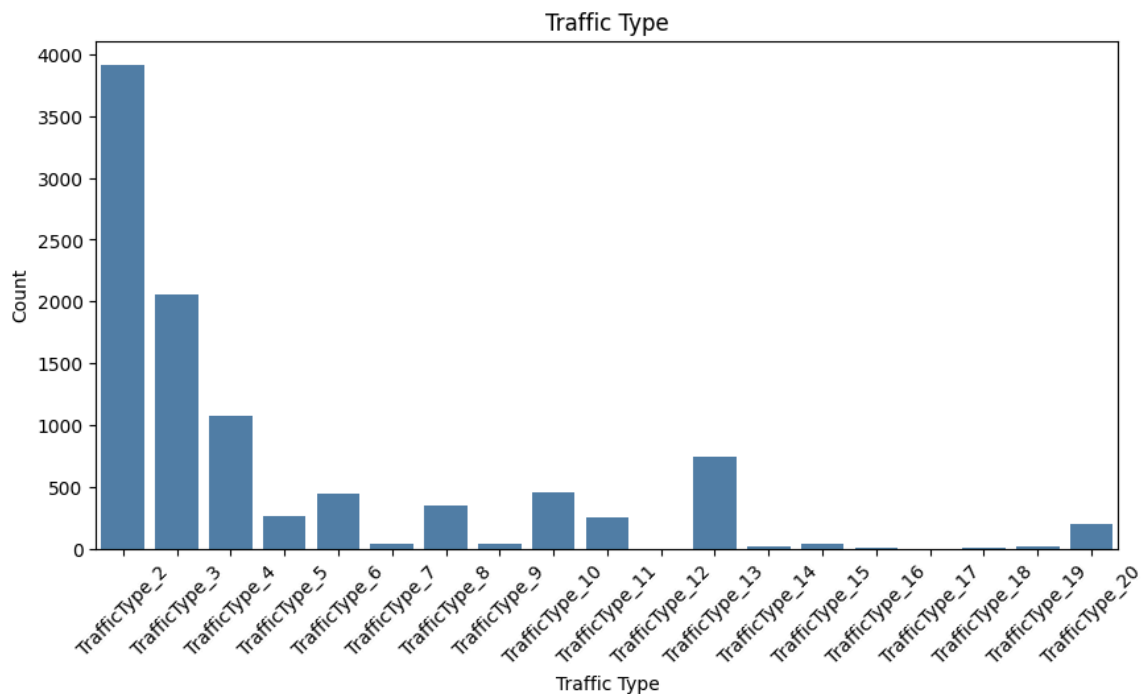
- **Browser:**

The bar chart highlights that **Browser 2** is the most popular, followed by **Browser 4** and **Browser 5**.



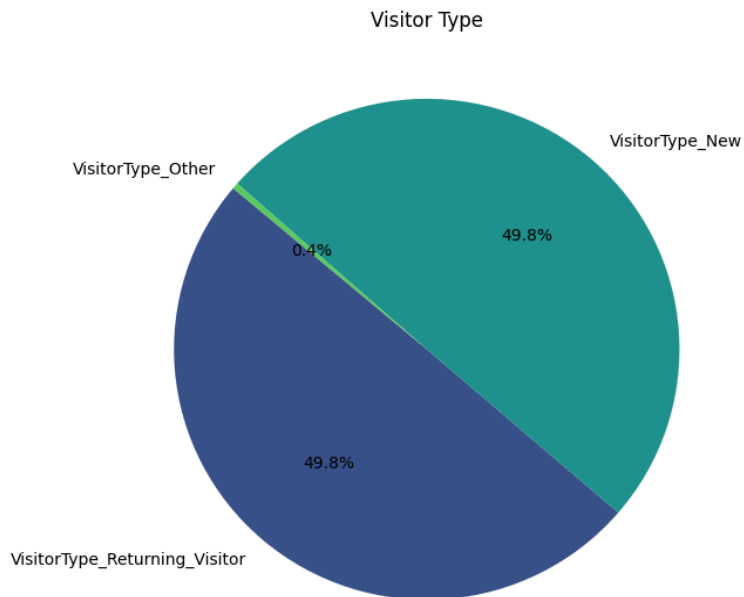
- **Traffic Type:**

The bar chart shows that **Traffic Type 2** is the most common, followed by **Traffic Type 3** and **Traffic Type 4**.



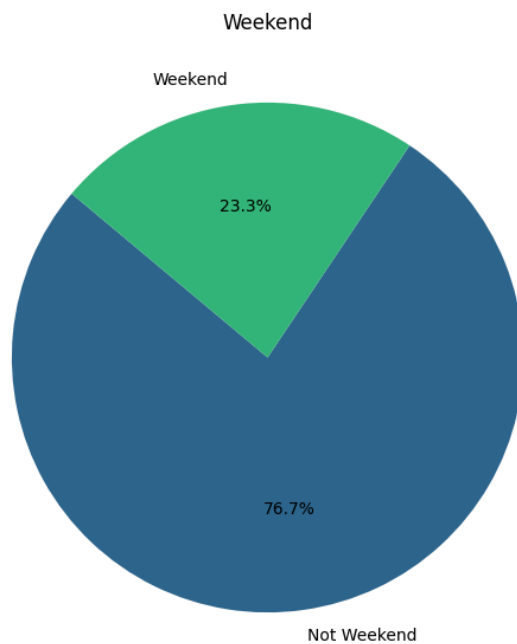
- **Visitor Type:**

The pie chart reveals that the dataset is evenly split between **Returning Visitors** and **New Visitors** (both 49.8%) and a small percentage of **Other** visitors.

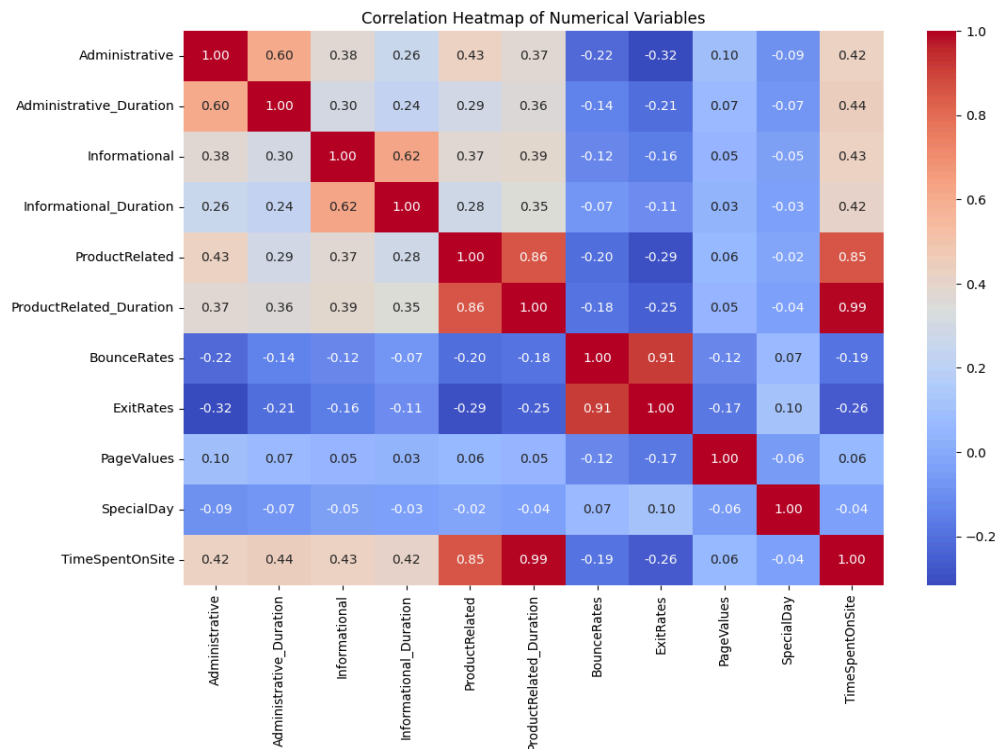


- **Weekend:**

The pie chart shows that **76.7%** of visits occur on weekdays, while **23.3%** occur on weekends.



## 2.3 Correlation Analysis



- The correlation heatmap reveals relationships between numerical variables:
  - **Strong Positive Correlations:**
    - ProductRelated and ProductRelated\_Duration (0.86): Users who view more product-related pages tend to spend more time on them.
    - Administrative and Administrative\_Duration (0.60): Users who visit more administrative pages also spend more time on them.
  - **Negative Correlations:**
    - Administrative and ExitRates (-0.32): Higher administrative rates are associated with lower exit rates.
    - ExitRates and ProductRelated (-0.26): Higher exit rates are linked to lower product related.

## 2.4 Purchase Rates: Weekend vs. Non-Weekend

- The bar plot compares purchase rates for weekend and non-weekend visits:
  - **Non-Weekend Visits:** Purchase rate is **15%**.
  - **Weekend Visits:** Purchase rate is **17.5%**.

- **Insight:** While weekdays drive the majority of purchases due to higher traffic, weekend visits have a slightly higher purchase rate, indicating that weekend users may be more intentional or engaged.

## 2.5 Relationship Between Numerical Variables and Revenue

- **ProductRelated\_Duration vs. Revenue:**  
The box plot shows that users who made a purchase (Revenue = Yes) spent significantly more time on product-related pages compared to non-purchasers (Revenue = No). This indicates that higher engagement with product content is strongly associated with conversions.
- **BounceRates vs. Revenue:**  
The box plot reveals that users who made a purchase (Revenue = Yes) had lower bounce rates compared to non-purchasers (Revenue = No). This suggests that users who engage beyond the landing page are more likely to convert.

## 3. Feature Engineering

### 3.1 New Features Created

- **TimeSpentOnSite:** Sum of ProductRelated\_Duration, Administrative\_Duration, and Informational\_Duration to capture total user engagement.
- **VisitorType\_New:** Derived from VisitorType\_Returning\_Visitor to simplify the analysis of returning visitors.
- **Normalization:** not required as numerical features were already scaled between 0 and 1.

### 3.2 Target Variable Creation

- **Creation of new variable Revenue as Binary Target** (1 for purchase, 0 for no purchase).

## 4. Model Development

### 4.1 Train-Test Split

- The dataset was split into training (80%) and testing (20%) sets, with stratification to ensure the proportion of buyers (Revenue = 1) and non-buyers (Revenue = 0) remained consistent in both sets.
  - **Training Set Class Distribution:**
    - Non-Buyers: **84.53%**
    - Buyers: **15.47%**
  - **Testing Set Class Distribution:**
    - Non-Buyers: **84.51%**
    - Buyers: **15.49%**

### 4.2 Models Evaluated

Three models were trained and evaluated:

- **Logistic Regression** (Scaled features using StandardScaler).
- **Random Forest**
- **XGBoost (Gradient Boosting)**

## 5. Model Comparison and Performance Metrics

### 5.1 Evaluation Results

The three models were evaluated using key metrics: **Accuracy**, **Precision**, **Recall**, **F1-Score**, and **ROC-AUC Score**. Below is a summary of their performance:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.8812	0.7380	0.3613	0.4851	0.8810
Random Forest	0.8978	0.7407	0.5236	0.6135	0.9168
<b>XGBoost (Gradient Boosting)</b>	<b>0.8950</b>	<b>0.6804</b>	<b>0.6073</b>	<b>0.6418</b>	<b>0.9213</b>

### 5.2 Best Model Selection

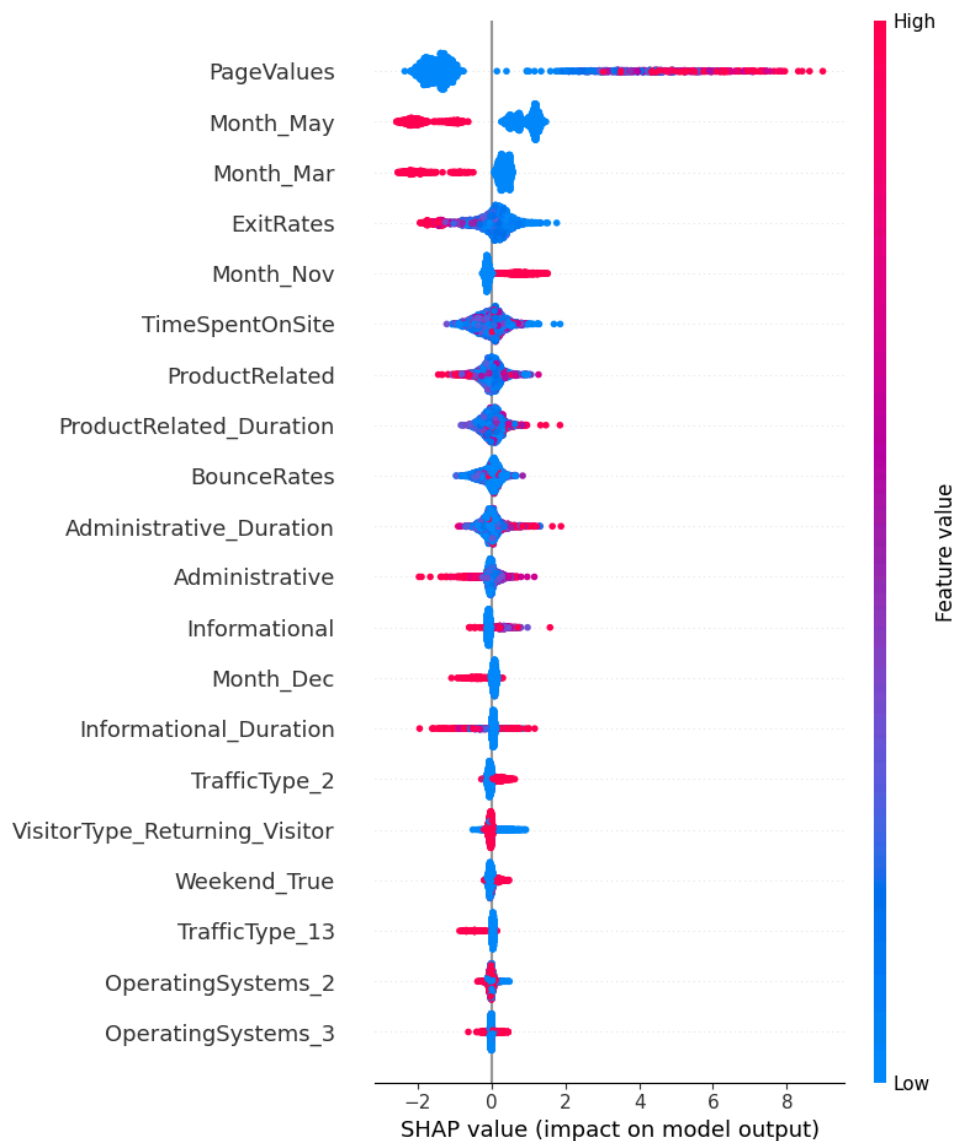
- **XGBoost** is the best-performing model overall due to its:
  - Highest **Recall (0.6073)**: Identifies more actual buyers.



- Highest **F1-Score (0.6418)**: Balances precision and recall effectively.
- Highest **ROC-AUC (0.9213)**: Best at distinguishing between buyers and non-buyers.

## 6. Key Business Insights

### 6.1 Variable Importance (SHAP Analysis)



The SHAP summary plot highlights the most influential features for predicting purchases:

1. **PageValues:**  
The average value of pages visited by a customer is the most critical factor. Higher page values correlate with higher purchase intent.
2. **Month\_May and Month\_Mar:**  
Visits in May and March are highly influential, indicating seasonal trends or promotions that drive purchases.
3. **ExitRates:**  
Lower exit rates are associated with higher purchase probabilities, emphasizing the importance of retaining users on the site.
4. **TimeSpentOnSite:**  
Total time spent on the site is a strong indicator of purchase intent. Engaged users are more likely to convert.
5. **ProductRelated and ProductRelated\_Duration:**  
Engagement with product-related pages significantly impacts purchase decisions.
6. **BounceRates:**  
Lower bounce rates are preferable, as users who leave after viewing only one page are less likely to purchase.
7. **VisitorType\_Returning\_Visitor:**  
Returning visitors are more likely to make purchases, highlighting the importance of customer retention strategies.
8. **Weekend\_True:**  
Weekend visits have a positive impact on purchase likelihood, likely due to users having more leisure time.

## 7. Recommendations to Improve Customer Retention and Purchases

### 7.1 Enhance Page Value

- Implement personalized recommendations and targeted promotions to increase the value of pages.

### 7.2 Seasonal Campaigns

- Run targeted marketing campaigns and promotions during high-impact months (May, March, December).

### **7.3 Reduce Exit and Bounce Rates**

- Improve website usability, content relevance, and page load speed to retain users and reduce early exits.

### **7.4 Engage Customers with Product Content**

- Ensure product-related pages are informative, visually appealing, and easy to navigate to encourage longer visits.

### **7.5 Retain Returning Visitors**

- Implement loyalty programs, personalized offers, and email marketing to encourage repeat visits and purchases.

### **7.6 Weekend Promotions**

- Introduce special weekend promotions, flash sales, or events to capitalize on higher purchase intent during weekends.

## **8. Actionable Recommendations for Marketing and Sales Teams**

### **8.1 Target High-Value Customers**

- Use the XGBoost model to identify high-value customers based on their engagement patterns (e.g., high PageValues, long TimeSpentOnSite).
- Focus on returning visitors and users who spend significant time on product-related pages.

### **8.2 Personalized Marketing**

- Leverage insights from SHAP analysis to create personalized marketing campaigns targeting users with high purchase intent.

### **8.3 Optimize Website Experience**

- Continuously monitor and optimize the website to reduce bounce and exit rates, ensuring a seamless user experience.

## Conclusion

- The **XGBoost model** is the best-performing model for predicting customer purchases, with strong performance in recall, F1-score, and ROC-AUC.
- Key drivers of purchase intent include **PageValues**, **TimeSpentOnSite**, and **ProductRelated\_Duration**.
- Actionable recommendations focus on improving user engagement, reducing bounce and exit rates, and leveraging seasonal trends to boost revenue.