### **1. Task and Data Description**

#### **Task Overview**

The goal of this task is to develop a machine learning model to predict whether a customer session on an online platform will result in a purchase (Revenue=True) based on customer behavior metrics. This prediction can help businesses identify potential buyers, optimize marketing strategies, and improve resource allocation.

#### **Dataset Overview**

The dataset contains 18 features derived from customer interaction data, categorized as follows:

* **Behavioral Metrics**:
  + Administrative, Administrative Duration: Count and total time spent on administrative pages.
  + Informational, Informational Duration: Count and total time spent on informational pages.
  + Product Related, Product Related Duration: Count and total time spent on product-related pages.
* **Engagement Metrics**:
  + Bounce Rate: Percentage of visitors who leave the site without further interaction.
  + Exit Rate: Percentage of visitors who leave after visiting a specific page.
  + Page Value: Average value of pages viewed prior to a purchase.
* **Temporal Features**:
  + Month: The month of the session.
  + Weekend: Indicates if the session occurred on a weekend.
  + Special Day: Indicates proximity to holidays like Mother’s Day.
* **User Characteristics**:
  + OperatingSystems: the type of OperatingSystems customers use
  + Browser: the type of Browser customers use
  + Visitor Type: Categorized as new, returning, or other.
  + Region: Geographic region of the visitor.
  + Traffic Type: The source of the user traffic.
* **Target Variable**:
  + Revenue: Boolean variable indicating whether a session resulted in a purchase.

### **2. Exploratory Data Analysis**

#### **2.1. Target Variable Distribution**

The Revenue variable is highly imbalanced:

* **Not Purchased (False)**: ~83.3% of the data.
* **Purchased (True)**: ~16.7% of the data.

This imbalance requires special handling, such as oversampling or adjusting classification thresholds.

#### **2.2. Visitor Type and Revenue**

* Returning visitors are significantly more likely to make a purchase compared to new visitors.
* A count plot of VisitorType with Revenue hue revealed this relationship.

#### **2.3. Bounce Rate and Page Value**

* Customers who made a purchase (Revenue=True) have lower Bounce Rate and higher Page Value, indicating that engaging content and valuable pages correlate positively with conversions.
* Density plots show clear separation between Revenue=True and Revenue=False.

#### **2.4. Temporal Patterns**

* Conversion rates vary by month, with notable spikes in holiday seasons.
* Sessions occurring on weekends and during special days exhibit higher purchase likelihood.

#### **2.5. PCA Visualization**

* Principal Component Analysis (PCA) was used to reduce high-dimensional data into 2D and 3D spaces. In this task, PCA was applied to visualize how the features differentiate customer sessions based on the target variable (Revenue).
* Scatter plots showed some clustering of Revenue=True sessions in specific regions, suggesting separability in feature space.

#### **2.6. Correlation Heatmap**

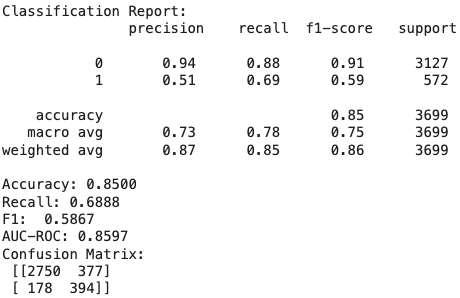
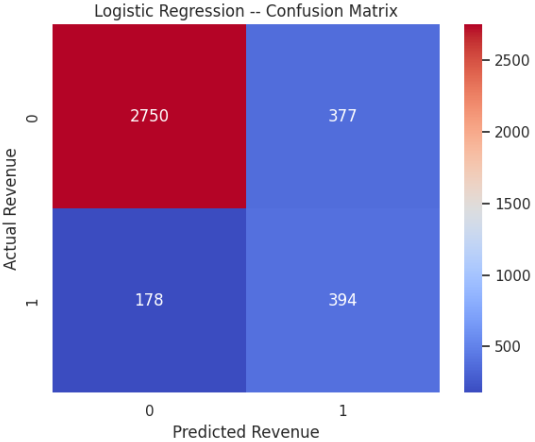
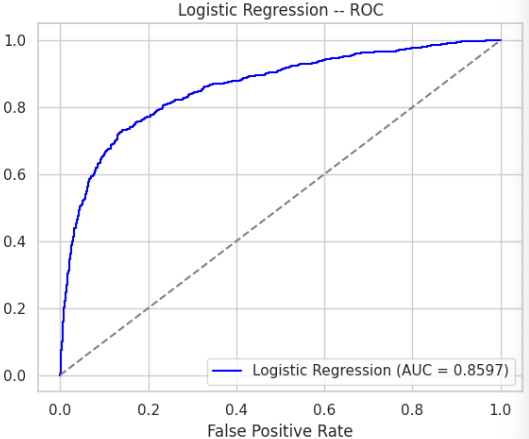
* Correlation Heatmap is a common visualization in data analysis, used to understand the relationships between numerical features and identify potential redundancies or key drivers for the target variable.
* Page Value may show a high positive correlation with Revenue, indicating that it is a key predictor.
* Features with high pairwise correlations (e.g.,Browser,TrafficType, Administrative, Administrative\_Duration ) might be redundant.

**Other Visualizations**:

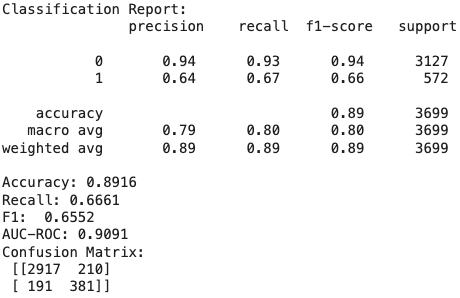
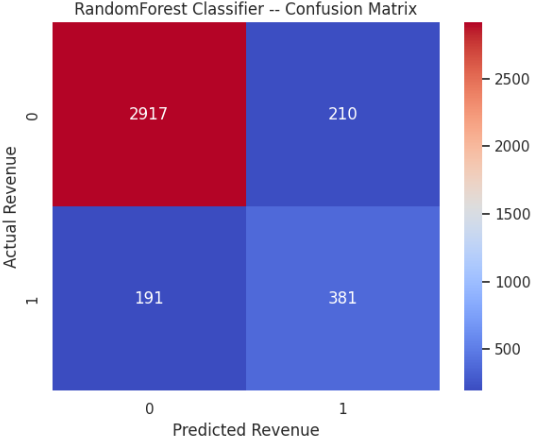
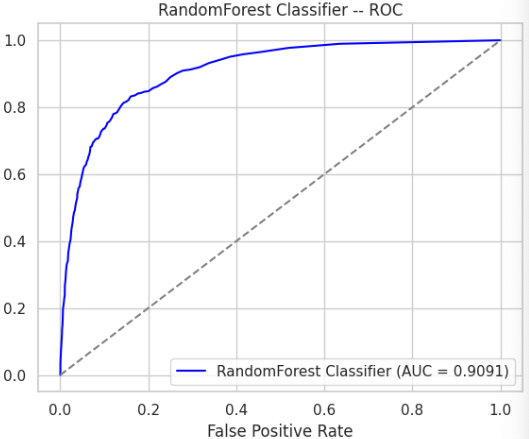
* Distribution plots for Bounce Rate and Page Value.

### **3. Algorithms Testing and Evaluation**

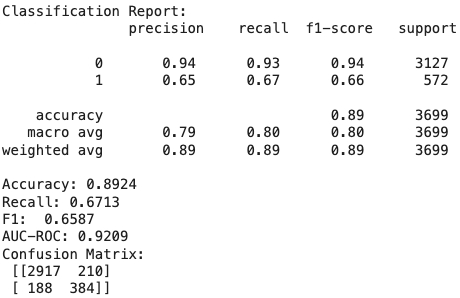
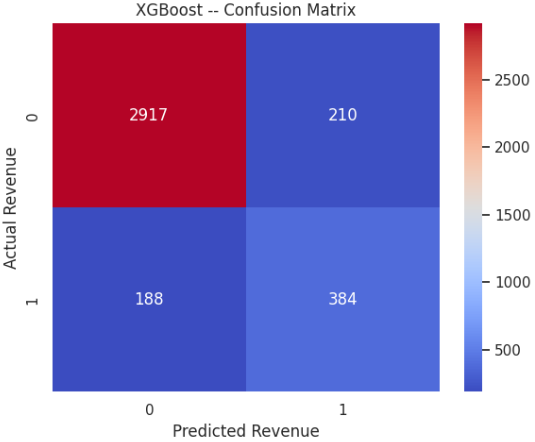
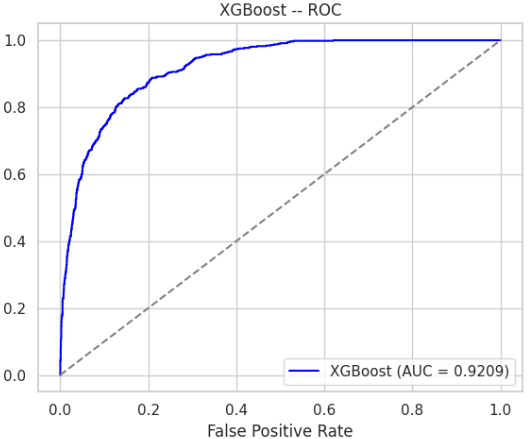
#### **3.1 Logistic Regression**

* **Model Description**: A baseline linear model with probabilistic outputs.
* **Performance Metrics**:
  + 
  + 
  + 
* **Strengths**:
  + Fast and interpretable.
  + Moderate performance on imbalanced data.

#### **3.2 Random Forest**

* **Model Description**: A robust ensemble model capable of capturing non-linear relationships.
* **Performance Metrics**:
  + ****
  + ****
  + ****
* **Strengths**:
  + Handles nonlinear interactions.
  + Even though the overall metrics are kind of similar to Logistic Regression, but it handles imbalanced data more effectively, the accuracy and recall of the unpurchased type is much higher than Logistic Regression.

#### **3.3 XGBoost**

* **Model Description**: A gradient boosting algorithm with hyperparameter tuning and SMOTE oversampling.
* **Hyperparameters**:
  + max\_depth=5, learning\_rate=0.1, n\_estimators=100.
* **Performance Metrics**:
  + ****
  + ****
  + ****
* **Strengths**:
  + A little better performance than random forest.
  + Handles imbalanced data effectively.

#### **3.4 Threshold Adjustment**

In this analysis, threshold adjustment was explored to address the class imbalance in the dataset. By default, models classify samples as positive (Revenue=True) if the predicted probability is greater than or equal to 0.5. To improve the recall for the minority class, several lower thresholds (e.g., 0.45, 0.4, 0.3) were tested to evaluate their impact on model performance.

Despite the adjustment, no threshold produced significantly better results compared to the default threshold of 0.5. While lowering the threshold marginally increased recall, it came at the cost of precision and overall F1 score, leading to no net improvement in the model's ability to balance precision and recall effectively.

This observation suggests that the model is already optimized for the given data distribution and the trade-offs inherent in the classification process. It may also indicate that further performance improvement would require:

* Enhanced feature engineering to provide the model with better discriminatory information.
* Alternative sampling methods (e.g., SMOTE combined with undersampling).
* Exploring more complex algorithms that inherently handle class imbalance better (e.g., cost-sensitive learning approaches).

### **4. Final Results and Model Selection**

#### **Final Model: XGBoost**

* **Why Selected**:
  + While the Random Forest and XGBoost may yield similar metrics (e.g., accuracy, F1-score, AUC-ROC) on the current dataset, XGBoost offers the following additional benefits:
    - Better control over class imbalance and regularization.
    - Detailed insights into feature importance for interpretation.
    - Greater computational efficiency and scalability for large datasets.
    - Advanced customization options, making it more flexible in future applications.
  + Final Decision: Choosing XGBoost as the best model provides not only comparable metrics but also strategic advantages in terms of efficiency, interpretability, and flexibility, making it the preferred choice for deployment.
* **Why are overall Metrics kind of Similar Across Models?**
  + Impact of Class Imbalance
    - In highly imbalanced datasets, most machine learning models tend to focus on correctly predicting the majority class because it dominates the data distribution. As a result, the overall metrics like accuracy or even AUC-ROC might not vary significantly across models, as these metrics are heavily influenced by the performance on the majority class.
    - In this case, 83% of the samples belong to the "Not Purchased" class, simply predicting "Not Purchased" most of the time can yield high accuracy without truly capturing the minority class behavior.
  + Dataset Complexity
    - If the dataset lacks strong feature separability, even advanced models like XGBoost may not outperform simpler models like Logistic Regression by a wide margin. This can happen if the key features are not sufficiently informative and the patterns in the data are inherently difficult to learn.
  + Similar Decision Boundaries
    - All three models could be producing similar decision boundaries because the underlying patterns in the data are relatively linear, which favors Logistic Regression. Or Non-linear patterns exist but are weak or sparsely distributed, which means Random Forest and XGBoost cannot leverage their strengths in capturing complex relationships effectively.
* **Future improvement direction**
  + Data category imbalance: apply more data augmentation or adjust classification thresholds.
  + Model optimization:
    - Use more powerful models (e.g., XGBoost, LightGBM).
    - Perform hyperparameter tuning.
  + Feature engineering: try feature interactions, nonlinear transformations, etc.
  + Evaluate performance using cross-validation and multiple metrics (F1, recall).
  + Check data quality and correlation between features.

#### **Business Impact**

* **Customer Targeting**: The model can identify high-potential customers, enabling personalized marketing efforts.
* **Improved Conversion Rates**: Insights into key features like Page Value can guide content optimization strategies.
* **Resource Allocation**: Focus efforts on high-value sessions during weekends and special days.