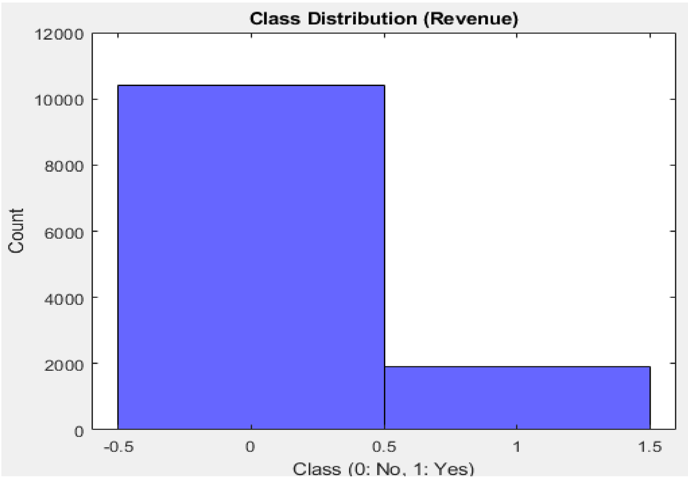
# Predicting Online Shopper Conversion: A Comparison of Logistic Regression and Gaussian Naive Bayes

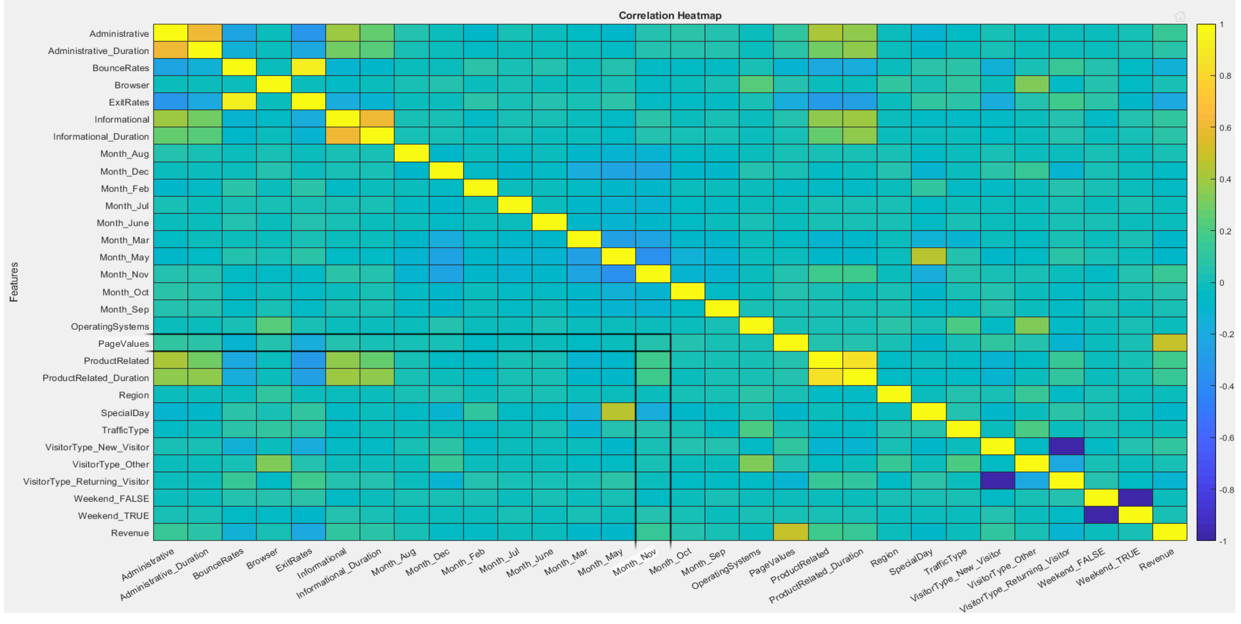
## Problem Description:

- Objective: This project aims to solve a binary classification problem predicting whether an online shopper will complete a purchase based on browsing behavior. The target variable is 'Revenue' (1 = Purchase, 0 = No Purchase).
- Relevance: Accurately predicting which shoppers will convert helps businesses to optimize marketing strategies, allocate resources more efficiently, and personalize shopper experiences to increase conversions.
- Real-World Impact: Accurate predictions enable targeted marketing, reducing costs, and increasing revenue by focusing efforts on high-conversion users.

## Dataset Overview:

- **Dataset**: Online Shoppers Purchasing Intention Dataset (UCI Machine Learning Repository)
  - Source: UCI Machine Learning Repository
  - Size: 12,330 rows, 17 features.
  - The target variable is 'Revenue' (1 = Purchase, 0 = No Purchase) [1]
- **Key Features**:
  - Numerical features:
    - Session duration (Administrative duration, Product-related views, Bounce rate, etc.)
    - Product-related duration, Exit rates, Special days, etc.
  - Categorical features:
    - Visitor type (New or Returning)
    - Weekend (whether it's a weekend visit)
    - Month (month of the year)



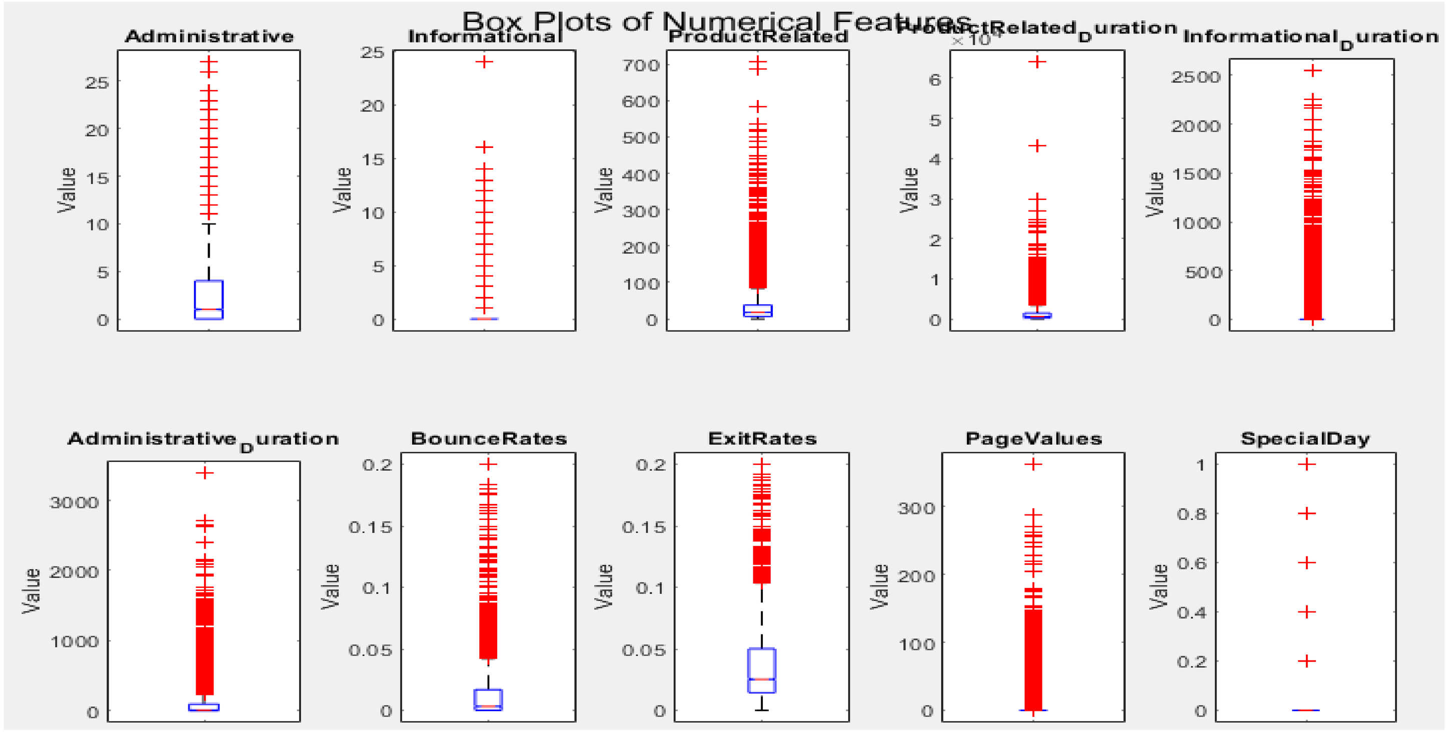



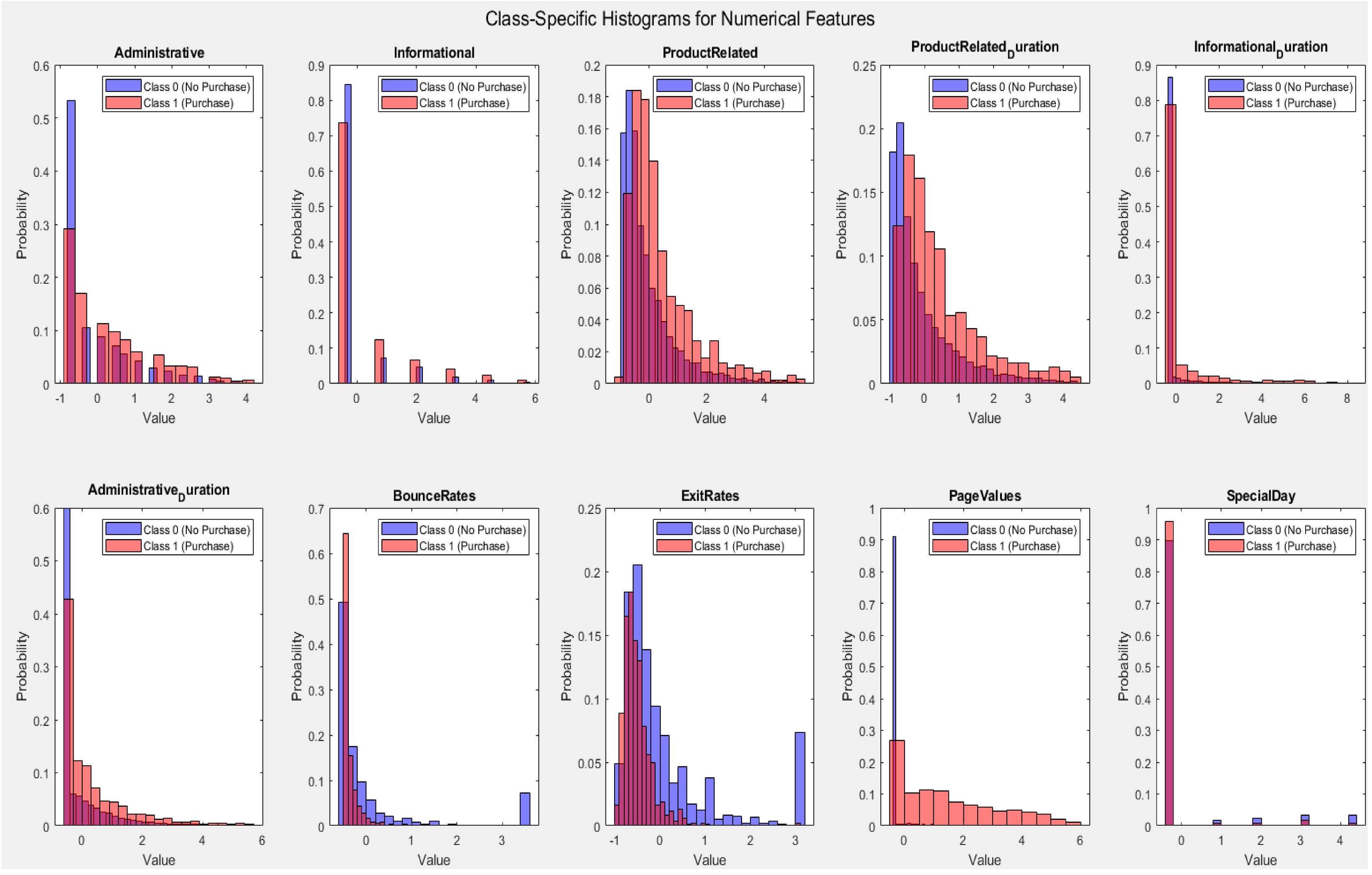

- **Preprocessing**:
  - Encoding Categorical Features: Categorical variables like 'Month' and 'VisitorType' were encoded using one-hot encoding.
  - Normalization: All numerical features (like Session duration, Bounce rates) were normalized to have zero mean and unit variance.
  - Handling Class Imbalance: The dataset suffers from class imbalance (most users don't complete a purchase) as shown above. We handled this by applying oversampling, undersampling, and class weights, UnderSampling delivered most higher performance .[2]
  - Outlier Removal: The top 2% of extreme values for numerical features were removed to improve model stability.
  - Feature Selection: We carefully selected relevant features for the model and removed less relevant or redundant ones to improve model performance. This step was crucial for reducing dimensionality, minimizing overfitting, and enhancing the interpretability of the model, you can see the correlation of all features with each other and the target column above.
    - Less Relevant Features Removed: Features such as 'Browser', 'OperatingSystems', and 'Region' were removed because they were either irrelevant to predicting purchase intent or did not provide useful variability.
    - Redundant Features Removed: Some features, like multiple 'Month' columns, were consolidated into one or removed because they provided redundant information.

## Methods Summary: Model Comparison

**Logistic Regression (LR) vs Gaussian Naive Bayes (GNB)**

| Aspect | Logistic Regression (LR) [3] | Gaussian Naive Bayes (GNB) [4] |
|---|---|---|
| Model Type | Linear, probabilistic classifier | Probabilistic classifier based on Bayes' theorem with Gaussian assumptions |
| Theory | Models the probability of a class as a linear combination of features | Assumes feature independence and models the probability using Gaussian distributions |
| Strengths | - Interpretable: Coefficients provide insight into feature importance.<br>- Performs well when the relationship between features and target is linear.<br>- Well-suited for problems with multicollinearity. | - Fast training and prediction times.<br>- Effective with imbalanced datasets.<br>- Simple and efficient with high-dimensional data.<br>- Works well for categorical and continuous features. |
| Weaknesses | - Struggles with non-linear relationships between features.<br>- Requires more computation for hyperparameter tuning.<br>- Sensitive to outliers in the dataset. | - Assumes feature independence, which is unrealistic for most real-world data.<br>- Assumes features follow a Gaussian distribution, which might not always be true.<br>- Limited in modeling complex relationships. |
| Interpretability | High: The model coefficients are easy to interpret and explain.<br>Each coefficient represents the log-odds of a feature affecting the target. | Moderate: Conditional probabilities give insight into feature importance but lack explicit coefficients. |
| Training Time | Moderate: Can be slower, especially with large datasets and cross-validation. | Fast: Extremely fast to train, particularly for large datasets. |
| Prediction Time | Moderate: Prediction time increases with the number of features. | Very fast: Efficient for large-scale predictions due to its simplicity. |
| Handling Class Imbalance | Requires Sampling techniques for imbalanced datasets. | Performs well with imbalanced datasets by naturally favoring the minority class during probability estimation. |
| Performance on Imbalanced Data | Less effective without handling imbalance, as it can be biased towards the majority class. | Good performance, particularly in capturing the minority class. |
| Feature Engineering | Requires careful feature engineering and selection.Correlated features can reduce model performance. | Assumes feature independence, which simplifies feature engineering. However, correlated features may negatively impact performance. |
| Scalability | Scales well with a moderate number of features but can be slow with very large datasets. | Excellent scalability for high-dimensional data, especially with sparse features. |




## Results:

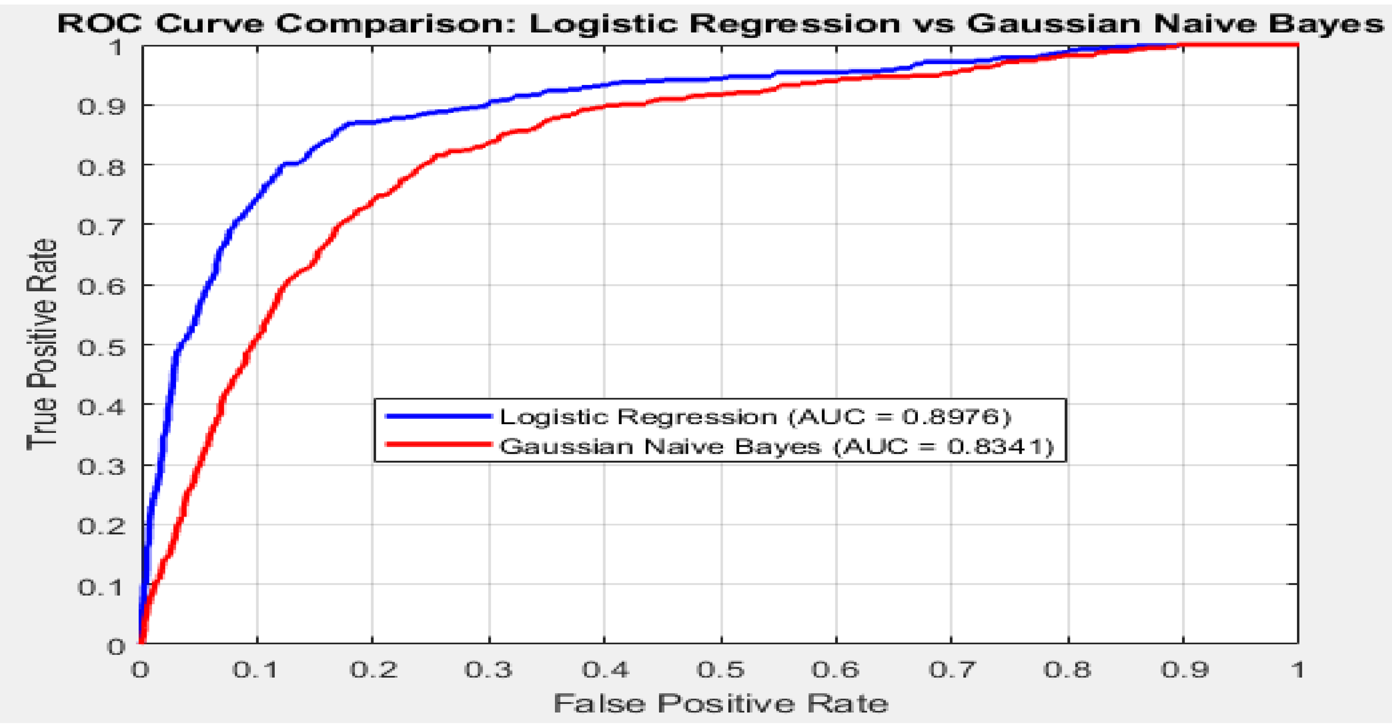
- Present the performance metrics for both models.
  - Metrics to show: Accuracy, Precision, Recall, F1-Score, AUC (Area Under Curve) .[5]

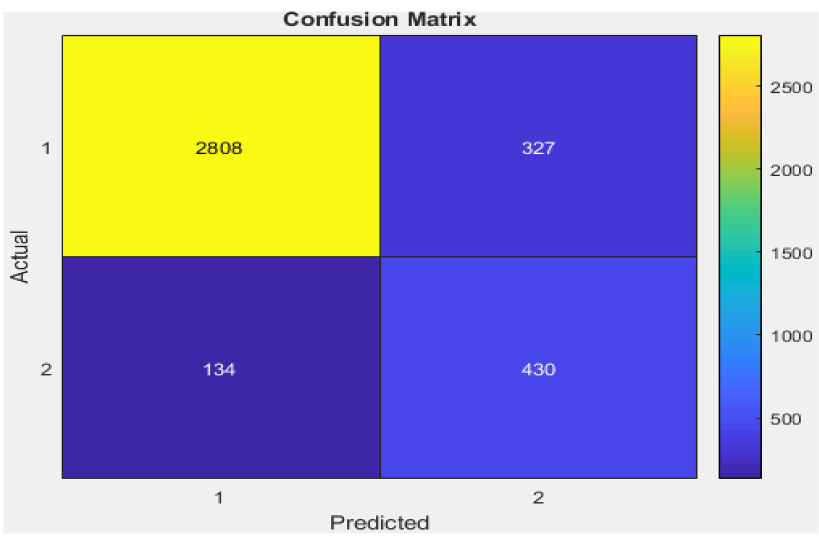
### Baseline Performance (Unoptimized models):

| Metric | Logistic Regression | Gaussian Naive Bayes |
|---|---|---|
| Accuracy | 0.8388 | 0.8388 |
| Precision | 0.7527 | 0.4767 |
| Recall | 0.3670 | 0.5816 |
| F1-Score | 0.4934 | 0.5240 |
| AUC | 0.8953 | 0.8351 |

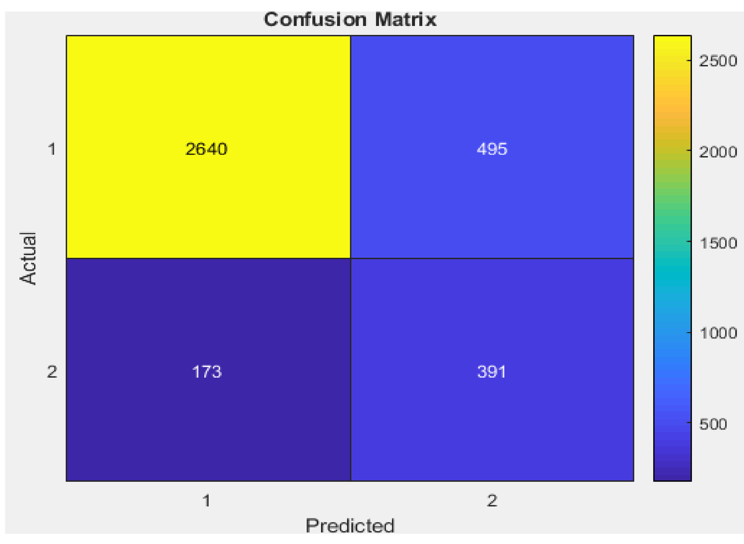
### Optimized Performance (After class balancing and hyperparameter tuning):

| Metric | Logistic Regression (Optimized) | Gaussian Naive Bayes (Optimized) |
|---|---|---|
| Accuracy | 0.87537 | 0.8194 |
| Precision | 0.56803 | 0.4413 |
| Recall | 0.76241 | 0.6933 |
| F1-Score | 0.65102 | 0.5393 |
| AUC | 0.89776 | 0.8415 |

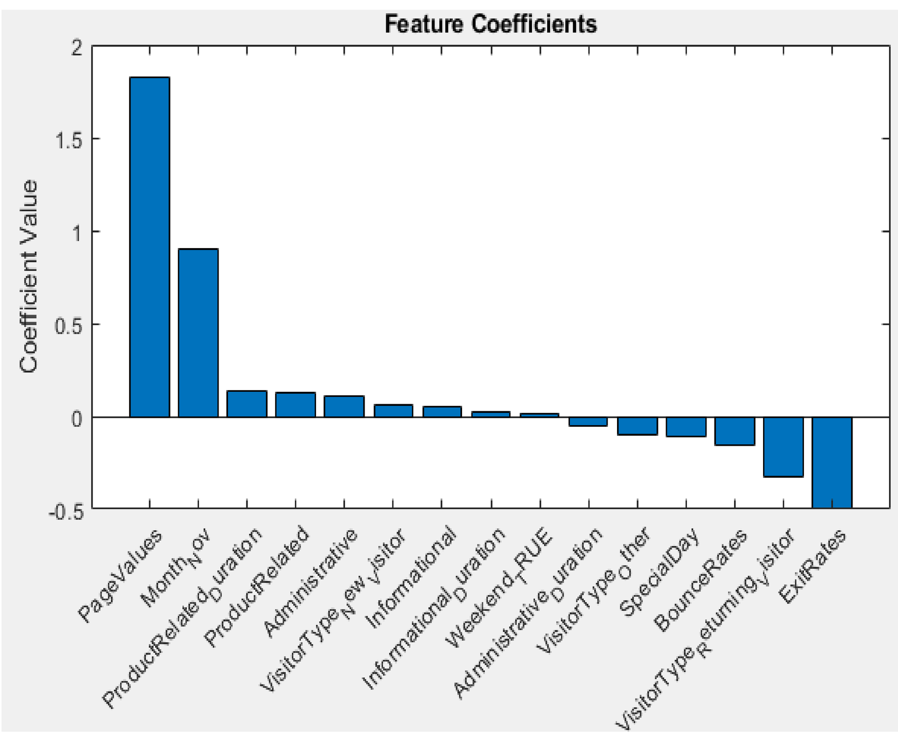


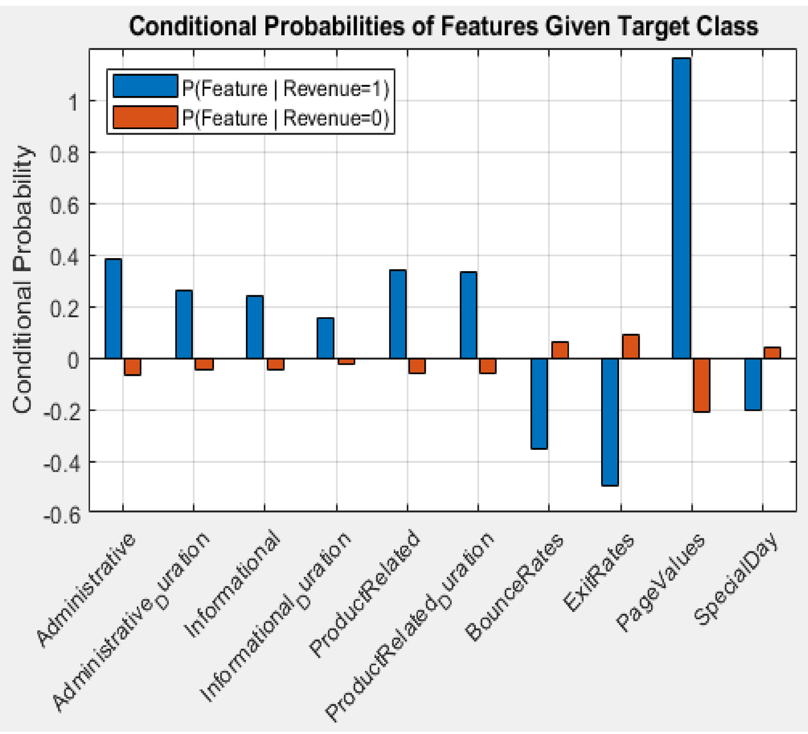




Confusion Matrix of Optimized LR model




Confusion Matrix of Optimized GNB model




## Discussion:

•**Model Comparison**:

- Logistic Regression outperforms Naive Bayes in terms of AUC and interpretability (due to its feature coefficients), but requires more computational resources for hyperparameter tuning.
- Naive Bayes performs faster and shows better recall for minority class detection, making it a viable choice for highly imbalanced datasets.
- When to Use:
  - Use Logistic Regression when feature interpretability and overall performance (AUC, Accuracy) are prioritized.
  - Use Gaussian Naive Bayes for quick modeling or when the focus is on recall and computational efficiency .[6]

•**Future Improvements**:

- Non-linear Models: Explore more complex models like Random Forest, XGBoost, or even Deep Learning models.
- Feature Selection: Implement more advanced techniques like Recursive Feature Elimination (RFE) or Principal Component Analysis (PCA) for dimensionality reduction.
- SMOTE: Use Synthetic Minority Over-sampling Technique (SMOTE) for generating synthetic samples for the minority class, improving model training.

## References:


1. Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository: Online Shoppers Purchasing Intention Dataset*. Link
2. Liu, X., et al. (2019). *Class Imbalance in Binary Classification: A Review*. IEEE Access, 7, 48179–48191. Link
3. Brownlee, J. (2020). *A Gentle Introduction to Logistic Regression*. Machine Learning Mastery.
4. Zhang, H. (2004). *The Optimality of Naive Bayes*. AAAI Conference on Machine Learning.
5. Raschka, S. (2018). *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning*. arXiv. Link
6. Ng, A. Y., & Jordan, M. I. (2001). *On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes*. Advances in Neural Information Processing Systems.