# E-COMMERCE CONVERSION OPTMIZATION ANALYSIS

## MSBA' 24 TEAM 13B


AMBUJ UPADHYAY <AMBUJU1@UCI.EDU>

AKHIL TANGUTUR <TANGUTUR@UCI.EDU>

PING-YEN CHUNG <PINGYEC@UCI.EDU>

YU-FANG CHANG <YUFANGC3@UCI.EDU>

ZIYUE WANG <ZIYUEW12@UCI.EDU>

# Table of contents

# 1. Introduction

## *1.1. Executive Summary*

In the digital age, the proliferation of online shopping has given rise to new dimensions in consumer behavior and economic activities. This evolution has brought with it an unprecedented level of competition among e-commerce businesses. To secure a competitive edge and thrive within this dynamic landscape, it is imperative for companies to decode and comprehend the factors driving consumer intentions.

We introduce an empirical model, expanding upon the traditional information system success framework. Through our research, we aim to equip e-commerce entities with actionable insights, enabling them to tailor their strategies to not only meet but anticipate consumer needs and behaviors. By understanding the critical pathways from intention to action, businesses can forge more meaningful connections with their customers, leading to sustained use and, ultimately, commercial success.

## *1.2. Business Ideas*

In this report, our primary objective is to analyze and predict online shopping behaviors by focusing on the consumer's intention to use, rather than just satisfaction metrics. This approach aims to provide e-commerce business with a more nuanced understanding of what drives consumer engagement and purchase.

Our project aimed at enhancing the online sales conversion rate for an e-commerce client, through detailed marketing analytics. The project involves analyzing customer behavior online shopping website to understand key performance indicators and metrics related to marketing. The focus is on devising marketing strategies to improve the conversion rate by guiding customers effectively through the marketing funnel. The analysis is based on data encompassing transactions, duration, and rates online activity over a year. This approach aims to leverage the growing popularity of online shopping to gain a competitive in the e-commerce industry by deeply understanding and satisfying to customer intentions and behaviors.

# 2. Data Description

## *2.1. Data Summary*

The Online Shoppers Purchasing Intention Dataset, sourced from the UCI Machine Learning Repository which was made publicly available on August 30, 2018, provides a broad analysis of 12,330 user sessions from an e-commerce platform over a one-year period. These sessions represent user interactions over a period of one year, ensuring diversity in the dataset and minimizing biases related to specific campaigns, special days, user profiles, or time periods. The dataset is categorized into two main classes based on the session outcome: sessions resulting in a purchase (positive class) and sessions without a purchase (negative class).

### *2.2. Class Distribution*

Negative Class (No Purchase): 84.5% (10,422 sessions)
Positive Class (Purchase Made): 15.5% (1,908 sessions)

### *2.3. Description of Features*

The dataset is composed of 18 attributes, comprising 10 quantitative and 8 qualitative variables. These attributes are categorized as follows:

***User Interaction Metrics:***

Administrative Visits: Counts visits to administrative sections.

Time on Administrative: Aggregate duration of visits to administrative sections.

Informational Visits: Counts visits to informational sections.

Time on Informational: Aggregate duration of visits to informational sections.

Product Related: Counts visits to product-related sections.

Time on Product Pages: Aggregate duration of visits to product-related sections.

These metrics gauge user engagement with different site components.

***Web Performance Indicators:***

Bounce Rate: The proportion of visits that are limited to one page, which can be a gauge of initial user interest.

Exit Rate: The rate at which a page is the last one visited in a session, shedding light on possible page-specific issues or disinterest. Page Value: An average estimation of the monetary value of a page, highlighting its role in generating revenue.

***Event-Related Attribute:***

Special Days: Indicates the session's nearness to significant events that can impact buying patterns, with variability based on the event's closeness.

***User and Session Information:***

Operating System: Identifies the operating system, which can influence compatibility and experience.

Browser: Specifies the browser type, which may affect the site's display and features.

Region: Identifies the user's location, offering insight into location-based market trends and preferences.

Traffic Source: Identifies how the user arrived at the site, revealing the marketing channel's effectiveness.

Visitor Classification: Distinguishes between new and returning users, important for assessing loyalty and engagement levels.

Weekend Browsing: Denotes whether the session occurred during the weekend, helping to distinguish between weekday and weekend user behavior.

Session Month: Identifies the month of the visit, aiding in the recognition of seasonal purchasing patterns.

### *2.4. Class Label*

The primary label of interest in this dataset is the `Revenue` attribute. This binary attribute denotes whether a session ended in a purchase (positive class) or not (negative class), making it the focal point for classification tasks.

# 3. Data Engineering

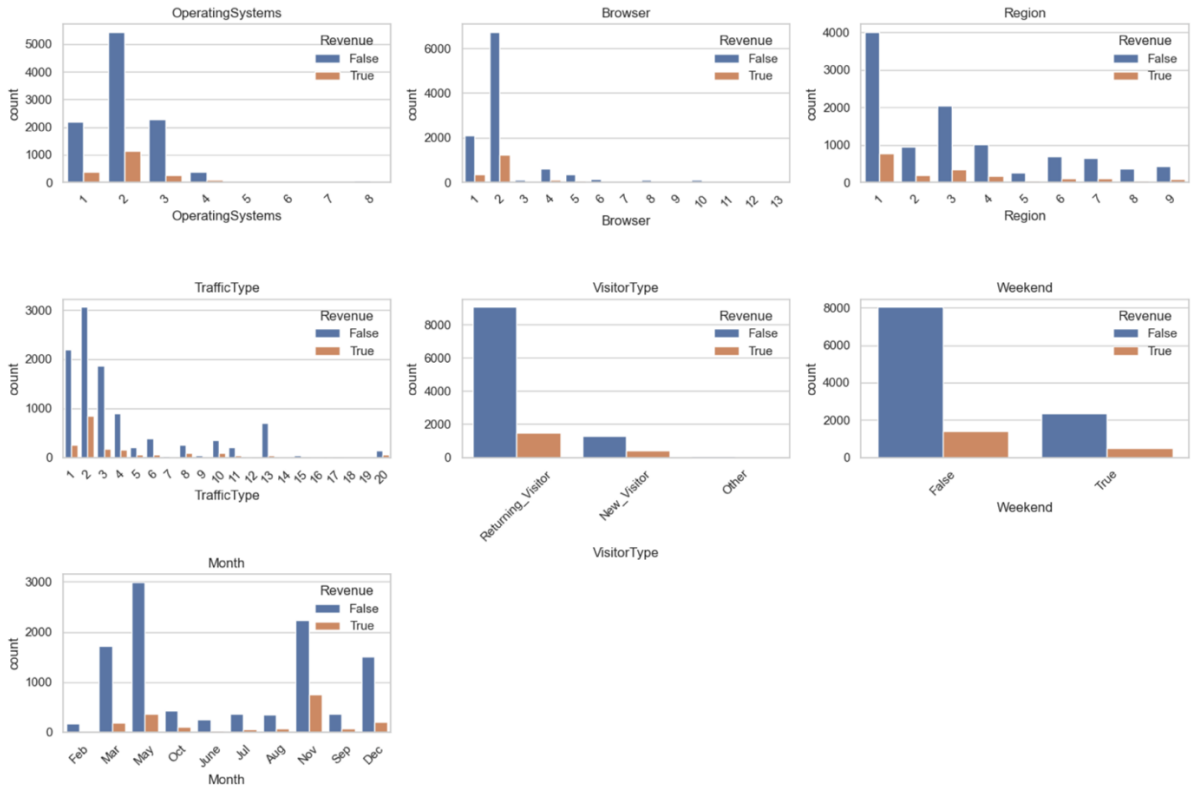
### *3.1.* Processing and Feature Engineering

The pre-processing phase involved several crucial steps:

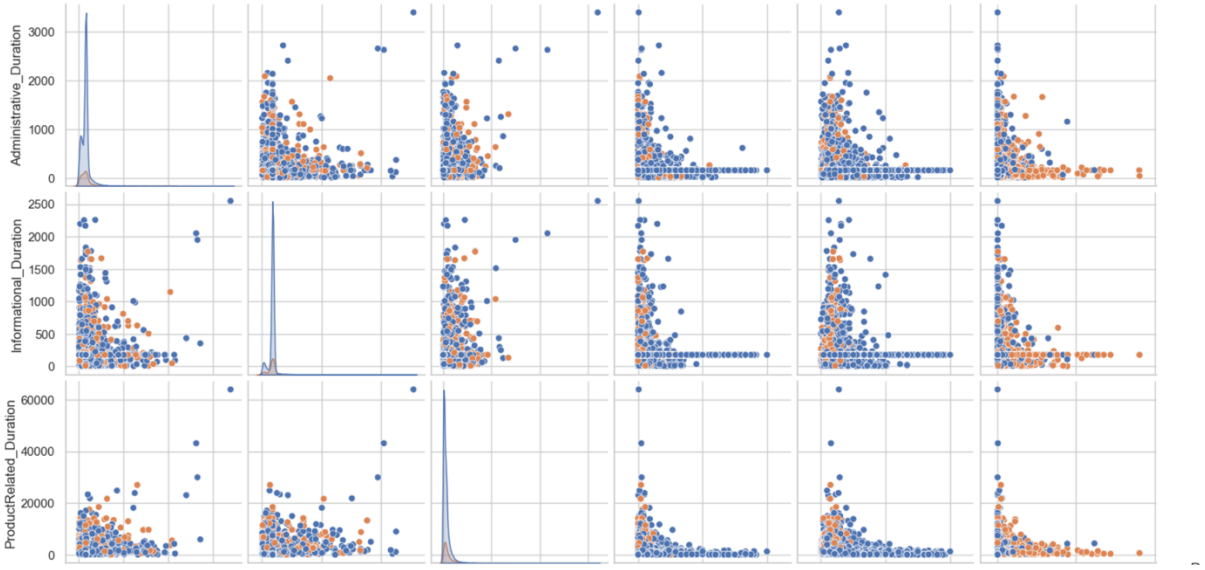
- Missing Value Analysis and Imputation**:** Recognizing that a count of zero in page-related features (Administrative, Informational, Product Related) essentially represents missing data, these were first converted to NaNs and then imputed with median values. This approach was chosen considering the categorical nature of these features.
- Outlier Detection and Removal**:** Outliers can skew model performance and were addressed using the IQR method for features such as 'Bounce Rates' and 'Exit Rates'. However, for 'Informational Duration' and 'Page Values', which exhibited extreme skews, outliers were retained to avoid loss of critical data.
- Feature Transformation: The 'Special Day' feature was transformed from a numerical to a binary variable, simplifying its interpretation and use in models. Additionally, categorical variables were converted to appropriate data types, enhancing model compatibility and efficiency.
- Feature Scaling: Given the substantial range differences among the numerical features, standardization was applied. This ensures all features contribute equally to the model, preventing any undue influence from features with larger scales.
- Label Encoding: The 'Month' and 'VisitorType' features were label-encoded, converting them into a format suitable for modeling, especially necessary for algorithms that require numerical input.

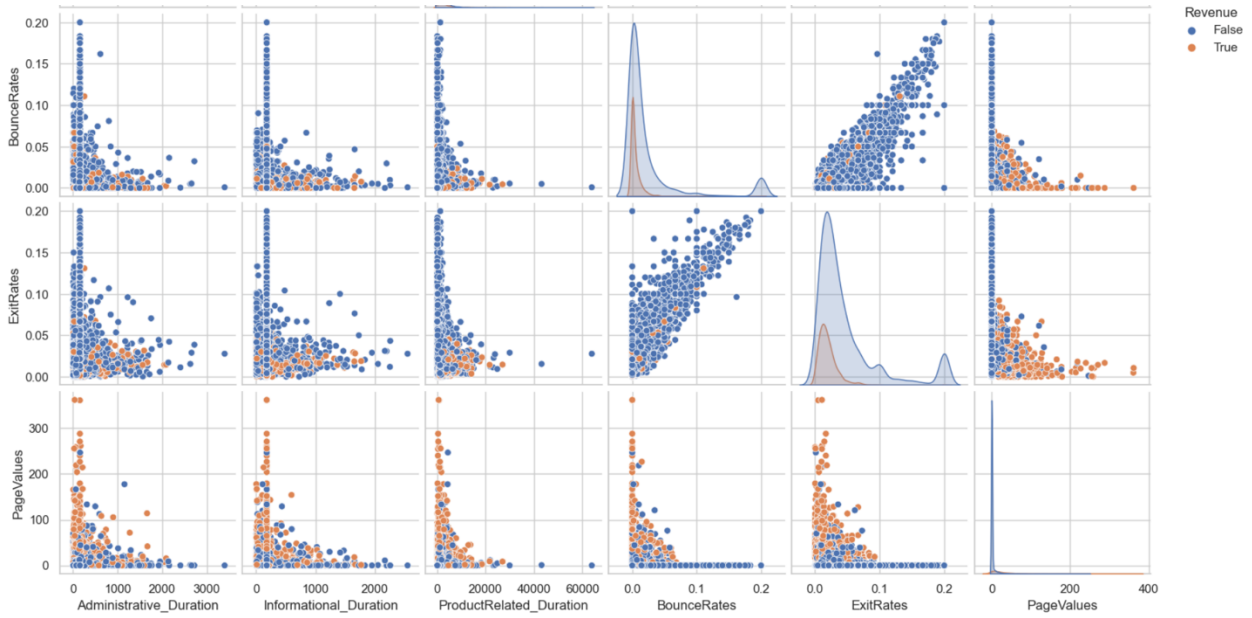
### *3.2.* Data Visualization and Key Takeaway

Visualization played a pivotal role in extracting insights from the dataset:

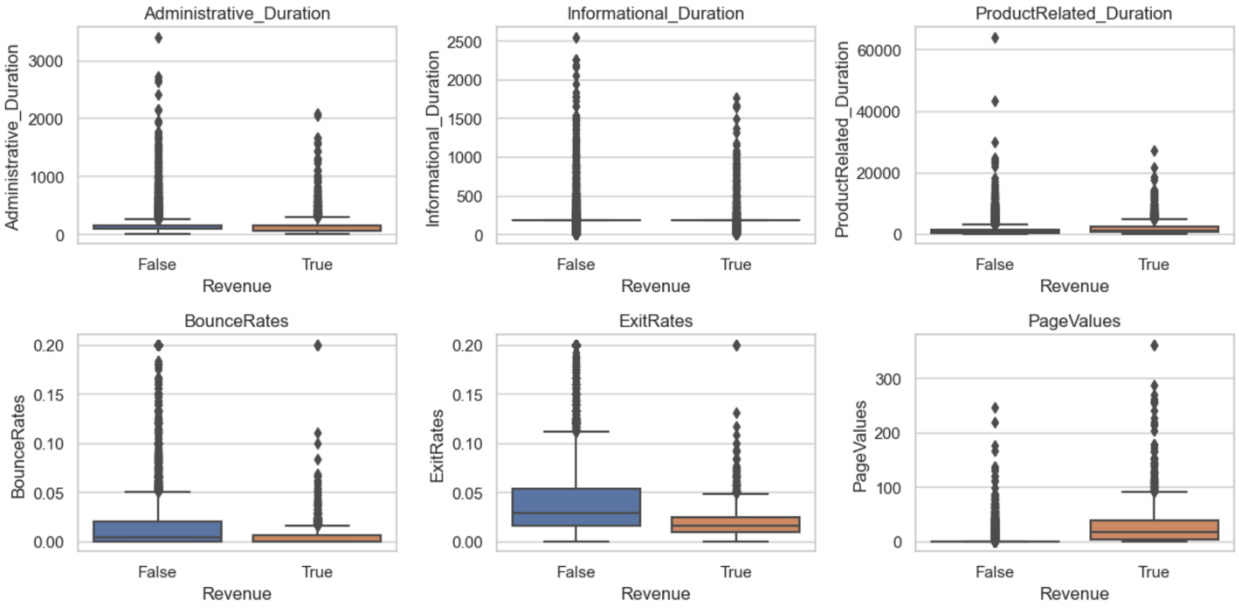



Count plots display the distribution of categorical variables, showing the count of observations in each category divided by the target variable 'Revenue'. They are helpful for understanding the impact of categorical features on purchasing decisions.



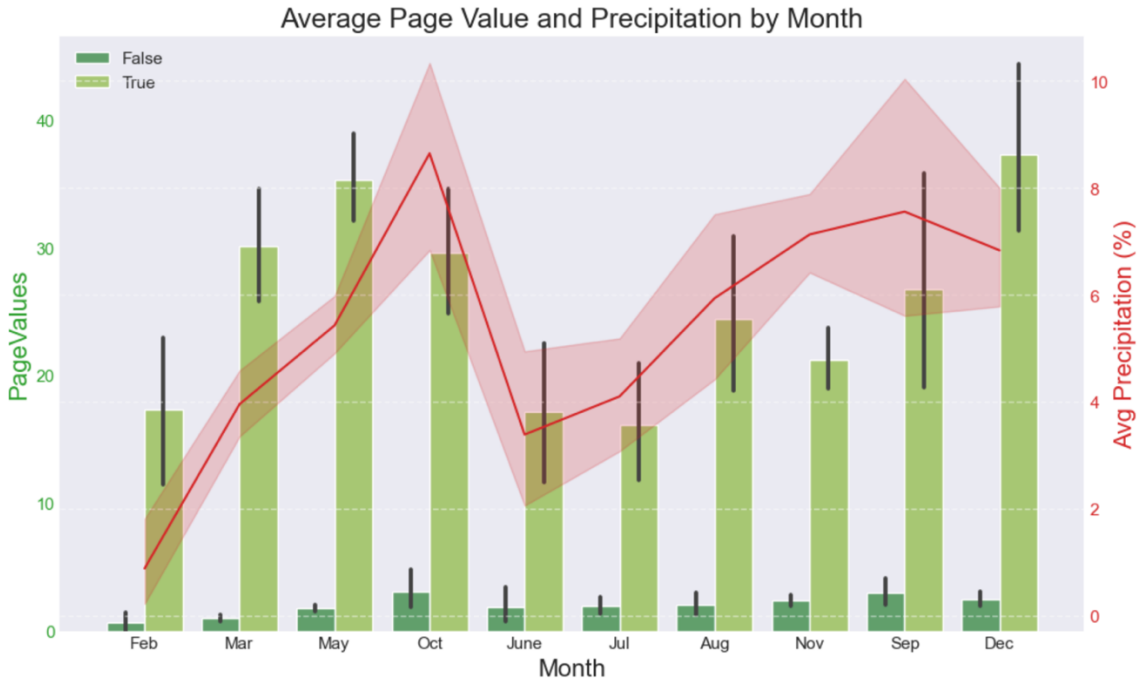



This plot shows pairwise relationships in the dataset, comparing different numerical features against each other and colored by the target variable 'Revenue'. It's useful for spotting correlations, trends, and outliers in the data.



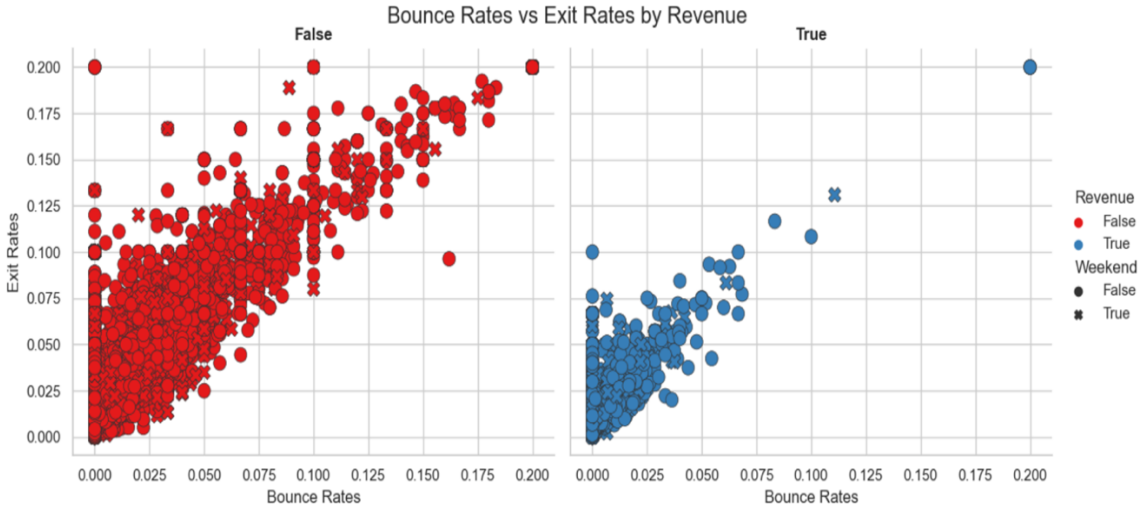

These box plots compare the distributions of various numerical features between sessions that resulted in revenue and those that did not. They help identify how different numerical variables behave in relation to the target variable.


By analysis the relationship of duration of page and revenue, it shows user behavior on an e-commerce website, indicating that while users are spending significant time on certain types of pages, this engagement is not necessarily leading to increased sales. Additionally, the closely linked nature of product engagement and time spent on product pages suggests a complex relationship.

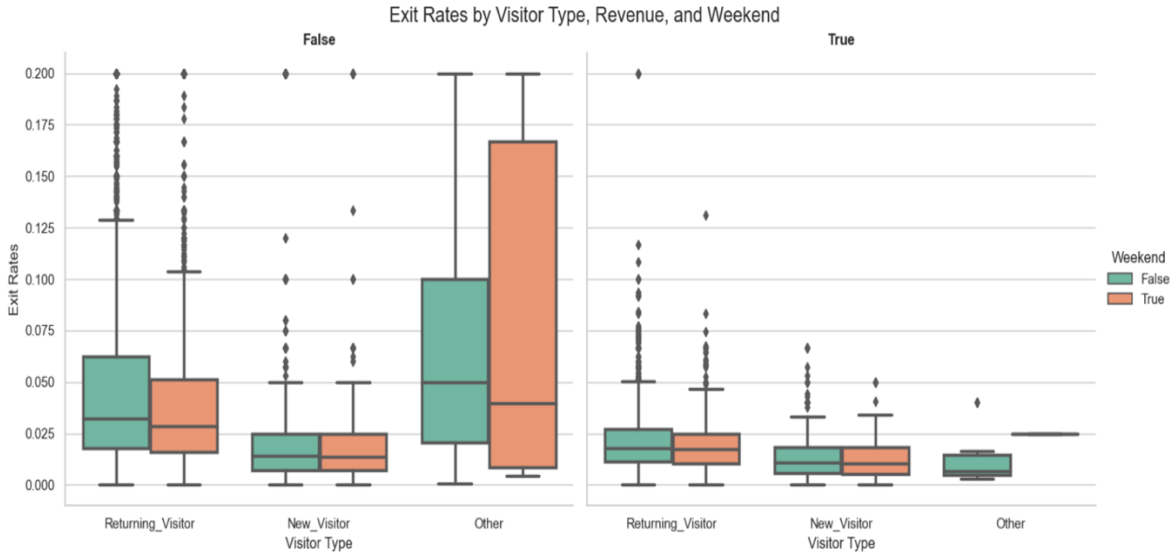



By analysis the relationship of page value vs. revenue by each month reveals a pattern of increasing page value over time, with a noticeable mid-year dip followed by a recovery. This trend suggests a seasonal impact on consumer activity. Significantly, there is a strong correlation between higher page values and instances of revenue generation, implying that pages with higher values are more

effective at converting visitors into paying customers. However, the notable variance in page values month-to-month indicates inconsistent engagement levels, which could be influenced by various marketing efforts or changes in consumer preferences.

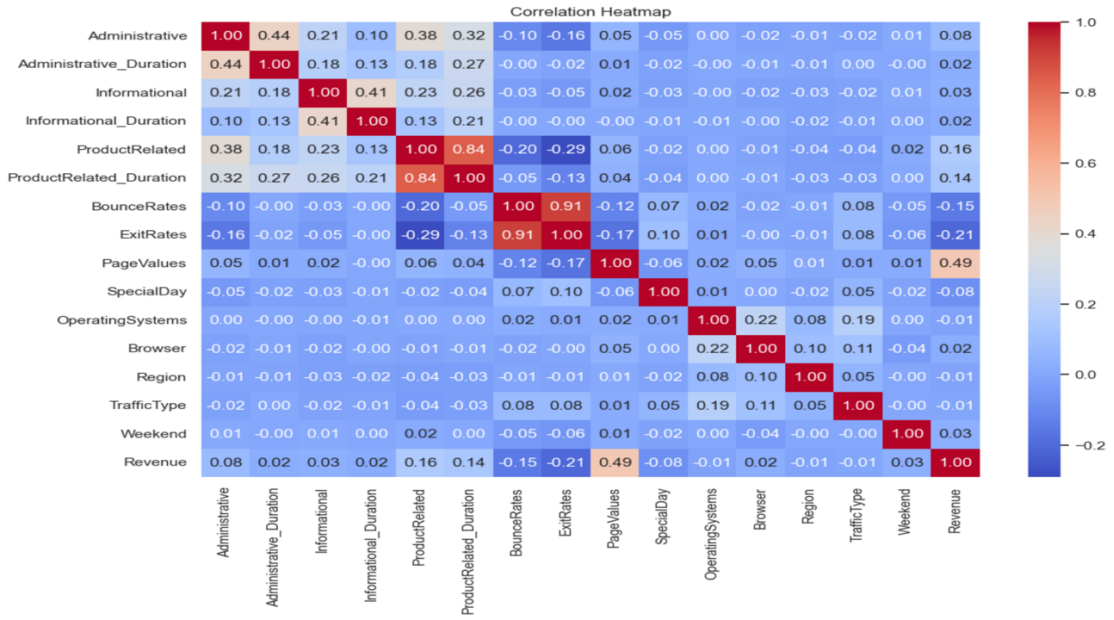



By analysis the relationship between bounce rate, exit rate and revenue, the scatter plot illustrates that sessions with higher bounce and exit rates generally do not lead to revenue generation, highlighted by a dense cluster of non-revenue sessions at elevated bounce and exit rates. There's a clear imbalance with a greater number of sessions failing to generate revenue compared to those that do. Additionally, the data indicates that higher bounce and exit rates are more prevalent during weekdays, as opposed to weekends, suggesting that user engagement is significantly different depending on the day of the week.

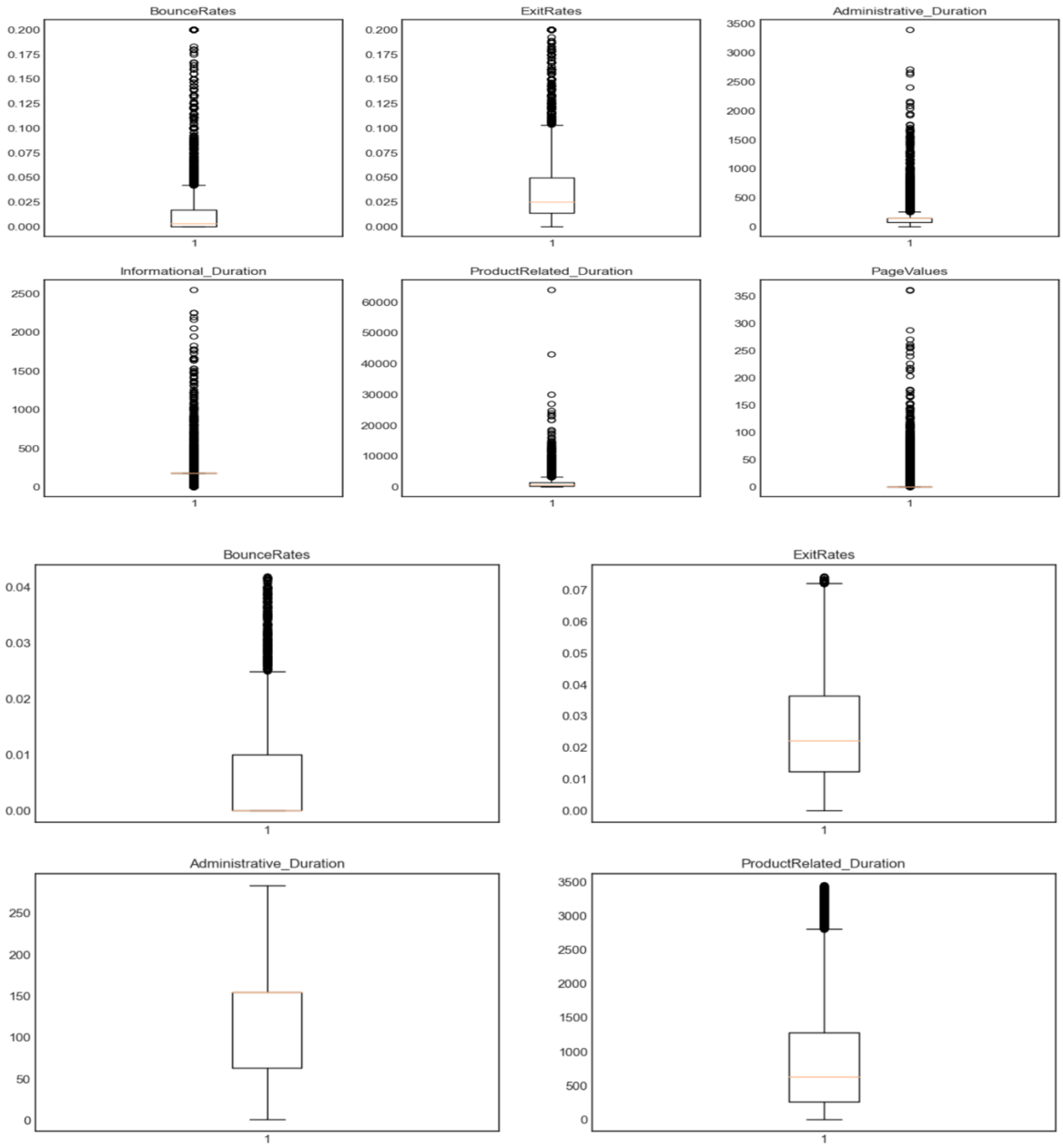



By analysis the relationship between visitor types, exit rate and revenue, it indicates that visitors who stay longer on the site are more likely to make a purchase. This trend is consistent regardless of the visitor type or whether the visit occurs on a weekend. Among the visitor categories, the

'Other' group displays a significant spread in exit rates during non-revenue sessions on weekends, suggesting a casual browsing behavior. New visitors exhibit low exit rates, which do not vary significantly between revenue and non-revenue sessions, highlighting successful engagement with first-time site users. Overall, the data underscores the importance of keeping exit rates low as a key factor in driving e-commerce revenue.




This heatmap visualizes the correlation matrix of the dataset, highlighting the strength and direction of relationships between different features. It's particularly useful for detecting multicollinearity and understanding feature interactions. From the map we can demonstrates, the page value is a strong predictor of revenue, with higher page values correlating positively with increased sales. Conversely, bounce rates negatively impact revenue, indicating that pages prompting users to leave after a single view can diminish earnings. Additionally, there is a pronounced positive correlation between bounce rates and exit rates, suggesting that pages that fail to retain visitors also see them exiting the site altogether. Crucially, product-related pages are pivotal in revenue generation, with a clear correlation between engagement on these pages and overall sales, highlighting them as key areas for optimization to boost the website’s profitability.




- **Page Interactions vs. Revenue:** Plots revealed a non-linear relationship between the duration spent on different types of pages and revenue generation. It highlighted the complexity of customer engagement and its impact on purchasing decisions.
- **Multicollinearity Insights:** The linear relationship detected between 'ProductRelated' and 'ProductRelated_Duration' indicated multicollinearity, suggesting a need for caution in model interpretation and feature selection.
- **Seasonal and Behavioral Trends:** The variation in 'PageValues' over months and its correlation with revenue generation highlighted the influence of time-based trends and specific page values on customer purchasing behavior.

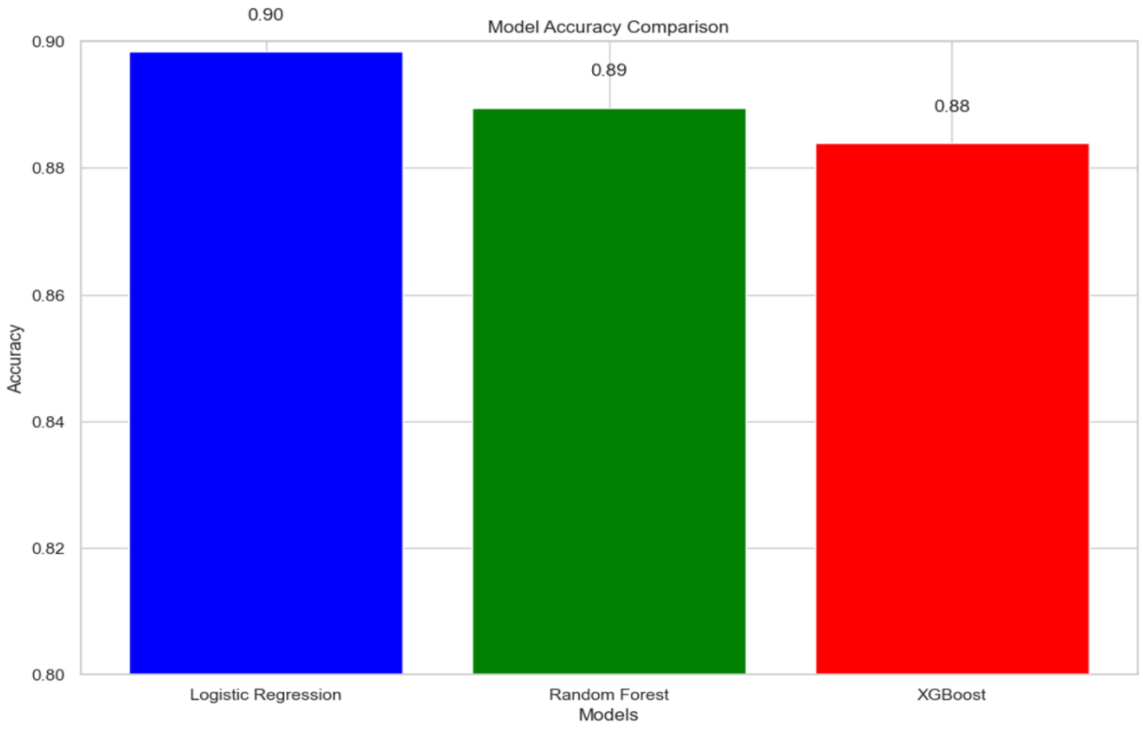
- **User Engagement Indicators:** 'BounceRates' and 'ExitRates', serving as indicators of user engagement, showed a strong inverse correlation with revenue, underscoring their potential as predictive features.

# 4. Modeling Results

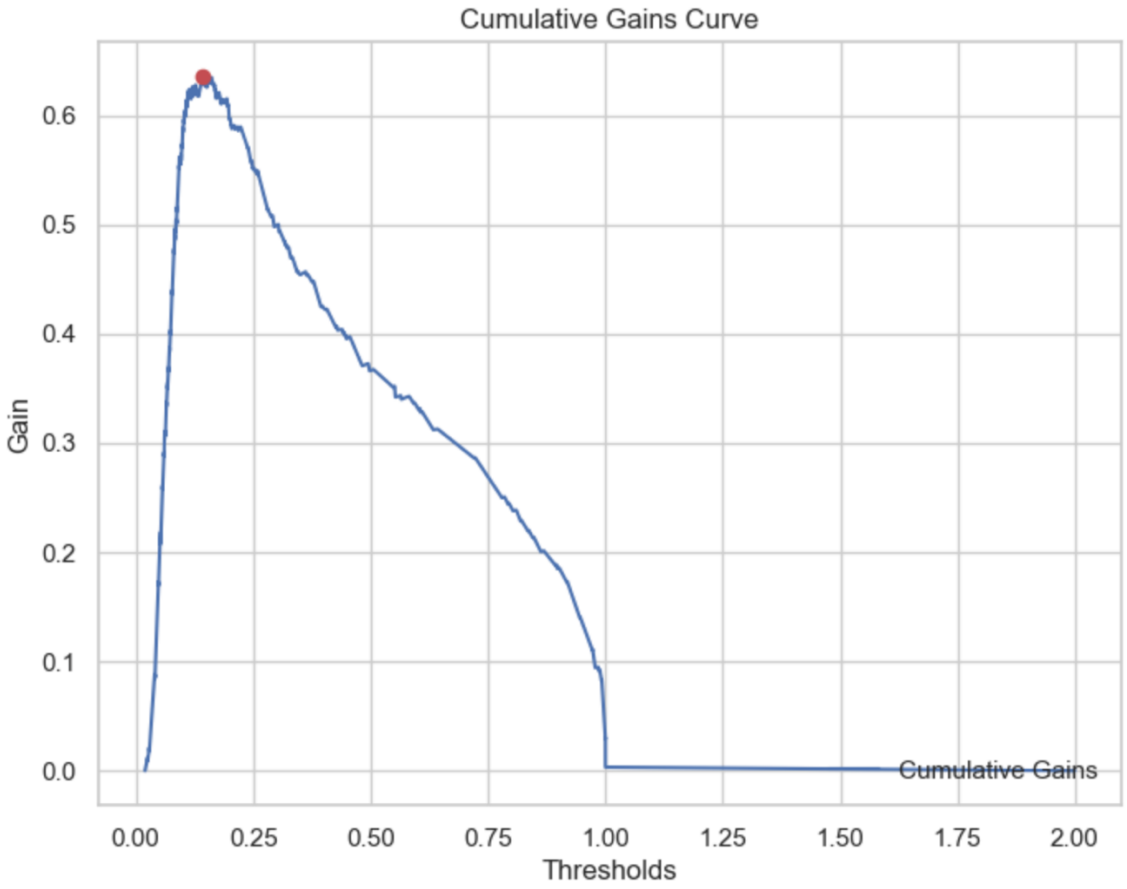
### *4.1. Benchmarking and Model Selection*

```
Logistic Regression Test Accuracy: 0.8983493697380547
Random Forest Test Accuracy: 0.8895639711969509
XGBoost Test Accuracy: 0.8839776531658883
```

The project initially evaluated Logistic Regression, Random Forest, and XGBoost. Logistic Regression outperformed others in cross-validation accuracy, suggesting its suitability for this dataset. This finding was critical, as it guided the subsequent focus on refining this model.



### *4.2. H*yperparameter Tuned

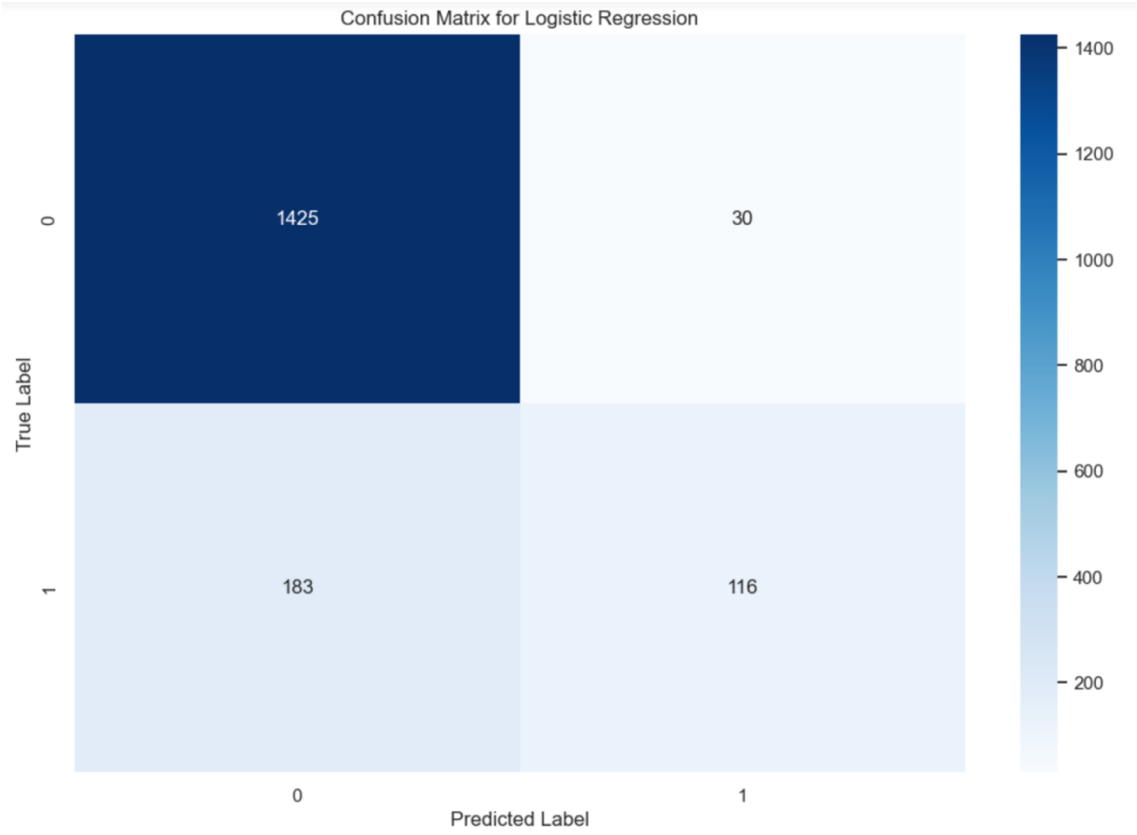


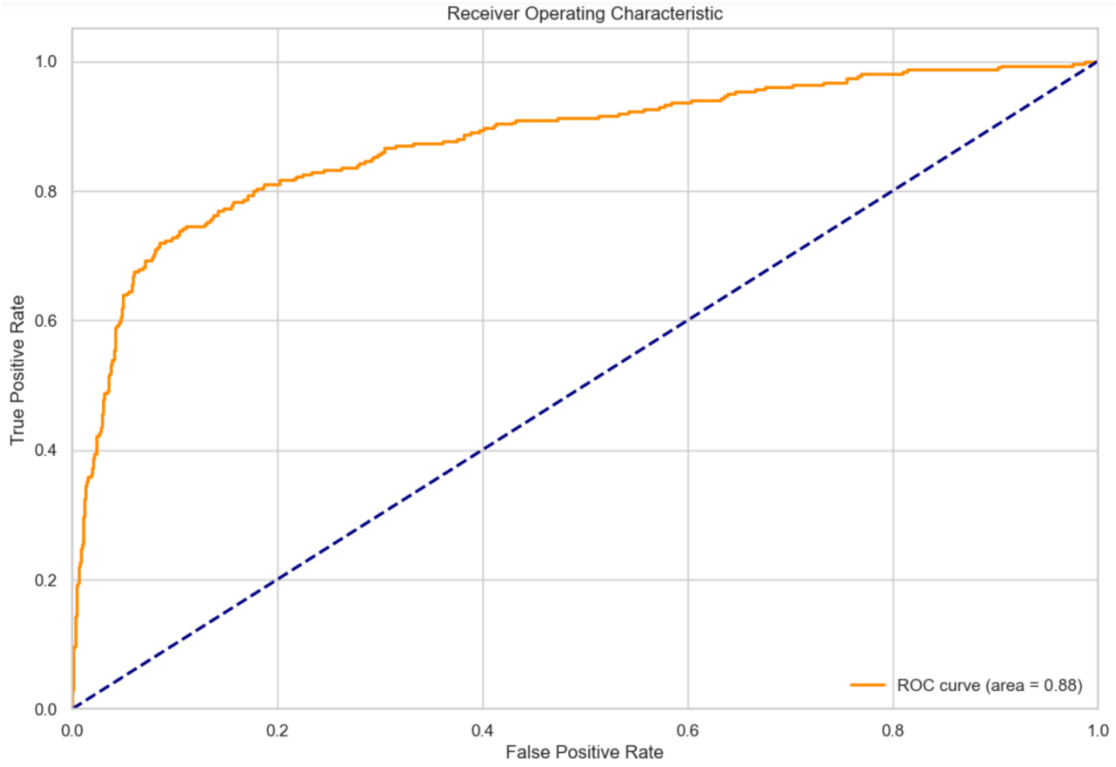

This curve illustrates the gains (true positive rate minus false positive rate) at different thresholds. It's valuable for identifying the threshold that maximizes the model's differentiation between the classes.

```
Fitting 10 folds for each of 30 candidates, totalling 300 fits
Best Parameters:  {'C': 1, 'solver': 'liblinear'}
```

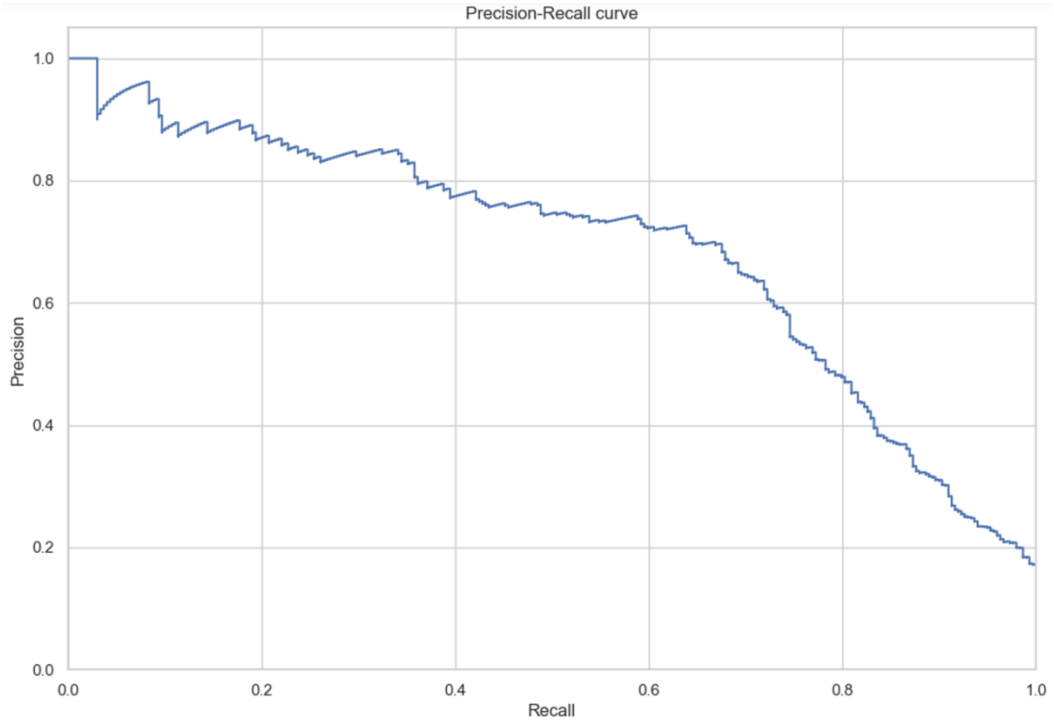
GridSearchCV was employed to fine-tune the Logistic Regression model. The process iterated over 30 combinations, with the best parameters being 'C': 1 and 'solver': 'liblinear'. The choice of 'liblinear' was particularly interesting, as it's more suited to smaller datasets and binary classification problems.

***4.3.*** Detailed Evaluation of the Optimized Model

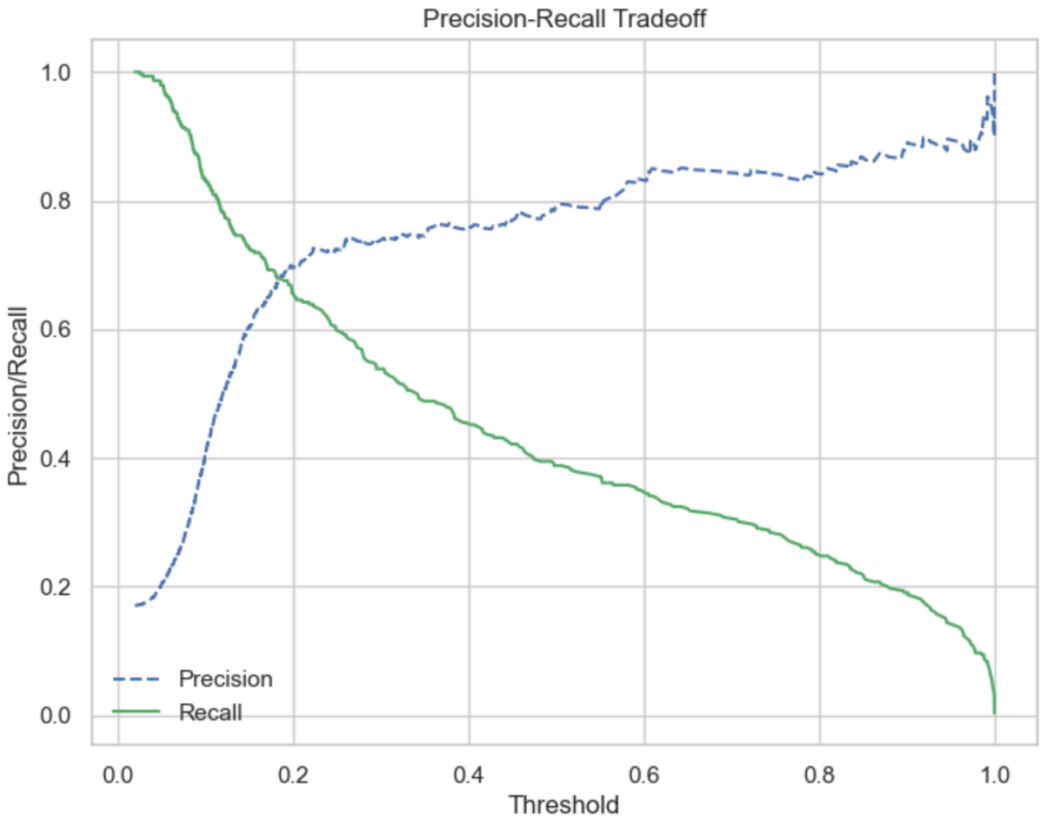





The Receiver Operating Characteristic (ROC) curve plots the true positive rate against the false positive rate at various threshold settings. The area under the curve (AUC) is a measure of the model's ability to distinguish between the classes.



This curve plots precision and recall for different threshold values. It's particularly useful for evaluating the performance of a classification model in cases where there is a significant class imbalance.

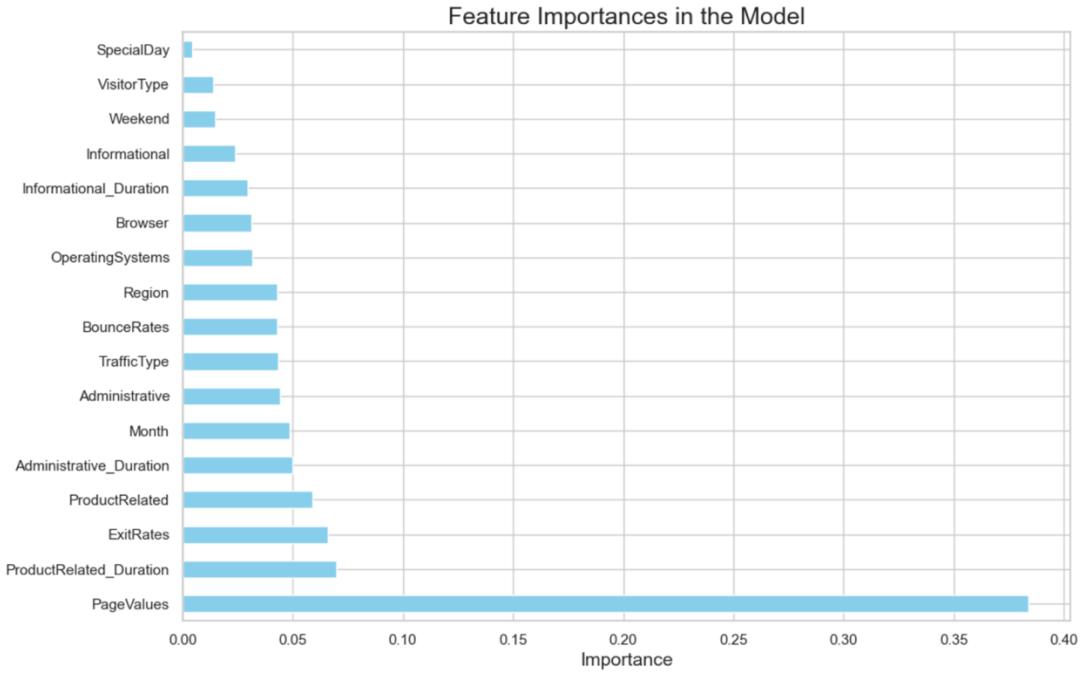



This plot shows the tradeoff between precision and recall for different thresholds in the Logistic Regression model. It's useful for understanding how changes in the threshold value affect the model's ability to correctly classify the positive class.

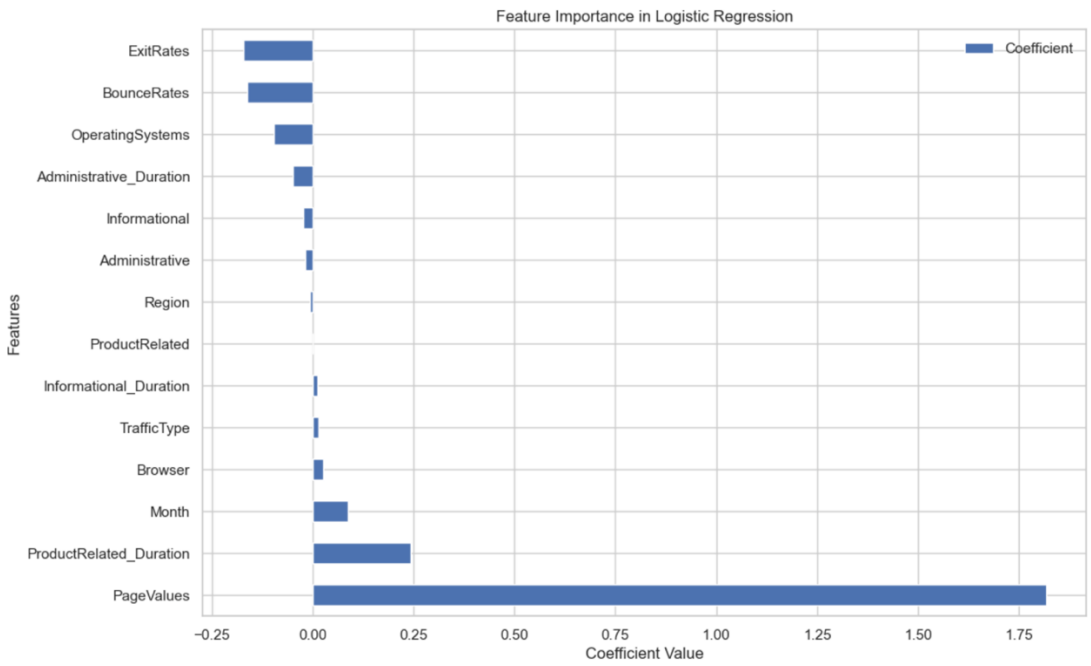
```
Accuracy on Test Set: 0.878563283922463
Confusion Matrix:
 [[1425   30]
 [ 183  116]]
Classification Report:
              precision    recall  f1-score   support

       False       0.89      0.98      0.93      1455
        True       0.79      0.39      0.52       299

    accuracy                           0.88      1754
   macro avg       0.84      0.68      0.73      1754
weighted avg       0.87      0.88      0.86      1754
```

- **Accuracy on Test Set:** Achieving 87.86% accuracy, the model showed robust performance in predicting non-purchases while being moderately successful in identifying actual purchases.
- **Precision-Recall Tradeoff:** The classification report revealed a high precision for non-purchases, suggesting the model's effectiveness in correctly identifying the majority class. However, the lower recall for purchases indicated a potential area for improvement in detecting the minority class.

### *4.4. Feature Importance and Interpretation*




| | Coefficient |
|---|---|
| PageValues | 1.817989 |
| ProductRelated_Duration | 0.242762 |
| Month | 0.088461 |
| Browser | 0.025569 |
| TrafficType | 0.013784 |
| Informational_Duration | 0.011014 |
| ProductRelated | 0.000176 |
| Region | -0.006302 |
| Administrative | -0.018825 |
| Informational | -0.024192 |
| Administrative_Duration | -0.048288 |
| OperatingSystems | -0.097329 |
| BounceRates | -0.161566 |
| ExitRates | -0.172372 |




The Logistic Regression model highlighted 'PageValues', 'ProductRelated_Duration', and 'Month' as significant predictors. The high coefficient for 'PageValues' emphasized its critical role in influencing purchase decisions, a valuable insight for strategizing content placement and design on the website.

### *4.5. Consistency Across Cross-Validation*

A cross-validated accuracy of approximately 88.47% underscored the model's consistency, indicating its general reliability and robustness against overfitting.

In conclusion of the logistic regression procesting, The preprocessing steps, particularly feature scaling and encoding, played a pivotal role in enhancing the model's predictive power. This iterative process of refining data and model parameters underscored the importance of thorough data preparation in achieving optimal model performance.

# 5. Inference and Suggestions

Based on the findings from the logistic regression analysis indicating the significance of Page Value, it's clear that enhancing our recommendation engine and creating more appealing bundle packages could lead to increased conversions. By leveraging the effect in the e-commerce strategy, we can further diversify our product offerings to tap into niche markets and drive additional revenue. To improve conversion rates, I suggest the following strategic actions:

- *Streamline User Interface*: Adopt a clean, minimalist user interface design that facilitates an effortless shopping experience, reducing cognitive load and decision fatigue for our customers.

- *Pricing Transparency*: Ensure that product pricing and information are presented clearly and early in the customer journey to foster trust and reduce cart abandonment rates.

- *Engagement Through Targeted Promotions*: Utilize data analytics to offer personalized discounts and promotions that encourage longer session durations and deeper engagement with our site.

- *Optimize Page Load Speeds:* By improving website performance and refresh rates, we can significantly decrease bounce rates. An attractive and compelling landing page that showcases products tailored to individual visitor preferences can further engage users.

- *Loyalty Incentives*: Implement a loyalty program that rewards returning visitors with exclusive offers, early access to new products, and personalized emails. This approach not only appreciates their loyalty but also encourages repeat business.

# 6. Conclusion

In conclusion, by our logistics regression analysis of e-commerce platform data has revealed a significant correlation between the Page Value feature and the likelihood of a customer completing a purchase. The Page Value, representing the average revenue generated by a page per user, indicates that customers are highly influenced by the variety and relevance of the products they are recommended. This finding suggests that customers are not just passively browsing but are actively seeking products that resonate with their needs and interests. To capitalize on this insight, it is imperative to invest in and refine our recommendation algorithms. By presenting customers with products that are closely aligned with their preferences and previous shopping behavior, we can increase the chance of conversion. Moreover, offering intelligently bundled packages that combine frequently bought together items or complementary products can simplify the purchasing decision process for the customer, leading to increased sales. This could attract a broader customer base and cater to more specialized needs, which are often underserved by mainstream e-commerce platforms, thus driving up the Page Value and, subsequently, the revenue.