IE 7275: Data Mining in Engineering

PROJECT REPORT
Group – 05

Insights into Online Shopper Behavior: Analyzing Bank Marketing Strategies

Saiteja Reddy Gajula (002872000)

Prarthana Veerabhadraiah (002821755)

Sathvik Ramappa (002847460)

Karthick Sriram Manimaran (002851406)

Akshaya Murugan (002843273)

Abstract

In this project, we will evaluate a bank marketing dataset to forecast the performance of marketing initiatives using classification algorithms. The dataset contains various attributes such as day of the week, region, browser, and others. Our goal is to train machine learning models to determine if a marketing effort will be successful or unsuccessful based on these characteristics.

First, we conduct exploratory data analysis (EDA) to acquire insight into the dataset's structure and properties. We then perform feature engineering, which entails turning categorical data into numerical format to prepare it for model training. We then clean the data by removing outliers, missing values, and duplicates.

Following preprocessing, we divided the data into training and testing sets to train and evaluate our classification models. We experiment with numerous methods, such as Logistic Regression, Random Forest, and Gradient Boosting Classification, to see which one performs best.

We test the models' performance using a variety of classification metrics, including accuracy, precision, recall, and F1-score. These measurements provide information about the models' prediction skills, allowing us to select the most successful algorithm for our analysis.

By conducting this extensive study, we hope to get a better knowledge of the aspects that influence the performance of bank marketing efforts and provide significant insights for future marketing strategies.

Problem Definition

The objective of this project is to develop predictive models using classification algorithms to forecast the performance of marketing initiatives in the context of bank marketing. The dataset contains various attributes such as day of the week, region, browser, and others, which will be utilized to train the models. The target variable is whether a marketing effort will be successful or unsuccessful based on these characteristics, which will guide marketing strategies and decision-making.

Introduction

In recent years, the rapid expansion of e-commerce has transformed the way people shop for ordinary things, making it the dominant mode of purchase for customers around the world. However, given the enormous terrain of online buying, many elements influence consumers' decision-making processes before they make a purchase. These characteristics can include the regularity with which consumers use the online platform, the existence of promotional events or holidays, and the website's structure and usability.

The capacity to recognize and understand these impacting elements is critical for e-commerce business owners and sellers. By identifying important forces that impact consumer behavior and exploiting this data, organizations can proactively enhance their online platforms to attract and keep customers.

The fundamental goal of our study is to explore customer behavior in the context of internet buying. Using the extensive and informative Online Shoppers' Purchasing Intention dataset, we want to find patterns and trends that shed light on the variables driving purchasing decisions among online shoppers. We hope to construct predictive models that can distinguish between consumers who make purchases and those who do not by utilizing advanced machine learning techniques, notably supervised classification models.

In technical words, our initiative focuses on the creation and assessment of supervised classification algorithms. Using a pre-processed dataset, we want to train and verify these algorithms to effectively predict online shopping behavior. The goal is to provide meaningful and actionable insights for businesses, giving them a thorough grasp of the elements that influence online transactions. Furthermore, these predictive models are useful tools for projecting sales patterns and supporting informed decision-making processes, allowing firms to optimize their strategies and achieve long-term growth in the competitive e-commerce landscape.

Data Description

Dataset:

https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset

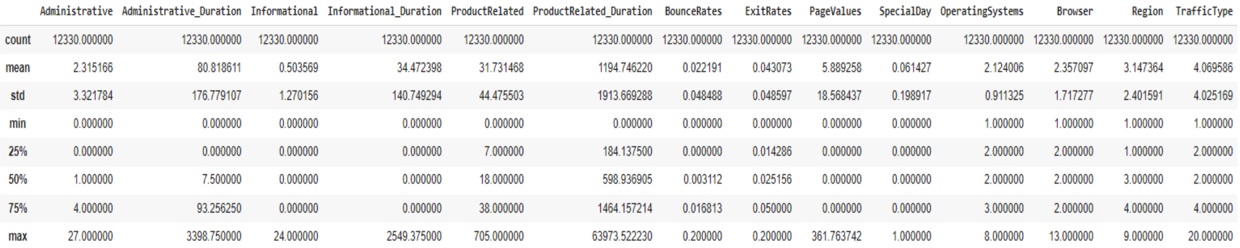
Features:

- Administrative: Number of administrative pages visited by the user.
- Administrative_Duration: Total time spent on administrative pages in seconds.
- Informational: Number of informational pages visited by the user.
- Informational_Duration: Total time spent on informational pages in seconds.
- ProductRelated: Number of product-related pages visited by the user.
- ProductRelated_Duration: Total time spent on product-related pages in seconds.
- BounceRates: Bounce rate of the website.
- ExitRates: Exit rate of the website.
- PageValues: Average value of the pages visited by the user.
- SpecialDay: Proximity of the visit to a special day (e.g., Mother's Day, Valentine's Day).
- Month: Month of the visit.
- OperatingSystems: Operating system of the user.
- Browser: Browser used by the user.
- Region: Region of the user.
- TrafficType: Type of traffic through which the user arrived at the website.
- VisitorType: Type of visitor (e.g., returning visitor, new visitor).
- Weekend: Whether the visit occurred on a weekend (True/False).
- Revenue: Whether the visit resulted in a revenue-generating transaction (True/False).

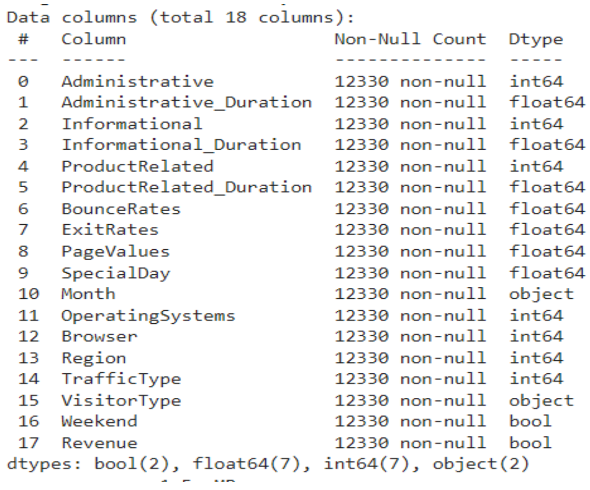
Summary Statistics:

- The dataset contains 12,330 entries.
- Numeric features exhibit varying ranges and distributions:
  - Administrative: Mean of 2.32 with a standard deviation of 3.32, ranging from 0 to 27.
  - Administrative_Duration: Mean of 80.82 seconds with a standard deviation of 176.78.
  - Informational: Mean of 0.50 with a standard deviation of 1.27, ranging from 0 to 24.
  - Informational_Duration: Mean of 34.47 seconds with a standard deviation of 140.75.
  - ProductRelated: Mean of 31.73 with a standard deviation of 44.48, ranging from 0 to 705.
  - ProductRelated_Duration: Mean of 1194.75 seconds with a standard deviation of 1913.67.
- Categorical features include Month, VisitorType, and Weekend.
- Boolean features include Weekend and Revenue.

The dataset provides a comprehensive view of user interactions with the website, including page visits, duration, and eventual revenue generation.



	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	OperatingSystems	Browser	Region	TrafficType
count	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000
mean	2.315166	80.818611	0.503569	34.472398	31.731468	1194.746220	0.022191	0.043073	5.889258	0.061427	2.124006	2.357097	3.147364	4.069586
std	3.321784	176.779107	1.270156	140.749294	44.475503	1913.669288	0.048488	0.048597	18.568437	0.198917	0.911325	1.717277	2.401591	4.025169
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000
25%	0.000000	0.000000	0.000000	0.000000	7.000000	184.137500	0.000000	0.014286	0.000000	0.000000	2.000000	2.000000	1.000000	2.000000
50%	1.000000	7.500000	0.000000	0.000000	18.000000	598.936905	0.003112	0.025156	0.000000	0.000000	2.000000	2.000000	3.000000	2.000000
75%	4.000000	93.256250	0.000000	0.000000	38.000000	1464.157214	0.016813	0.050000	0.000000	0.000000	3.000000	2.000000	4.000000	4.000000
max	27.000000	3398.750000	24.000000	2549.375000	705.000000	63973.522230	0.200000	0.200000	361.763742	1.000000	8.000000	13.000000	9.000000	20.000000



```
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Administrative           12330 non-null  int64
 1   Administrative_Duration  12330 non-null  float64
 2   Informational            12330 non-null  int64
 3   Informational_Duration   12330 non-null  float64
 4   ProductRelated           12330 non-null  int64
 5   ProductRelated_Duration  12330 non-null  float64
 6   BounceRates              12330 non-null  float64
 7   ExitRates                12330 non-null  float64
 8   PageValues               12330 non-null  float64
 9   SpecialDay               12330 non-null  float64
 10  Month                    12330 non-null  object
 11  OperatingSystems         12330 non-null  int64
 12  Browser                  12330 non-null  int64
 13  Region                   12330 non-null  int64
 14  TrafficType              12330 non-null  int64
 15  VisitorType              12330 non-null  object
 16  Weekend                  12330 non-null  bool
 17  Revenue                  12330 non-null  bool
dtypes: bool(2), float64(7), int64(7), object(2)
```

Exploratory Data Analysis

Frequency of Weekend

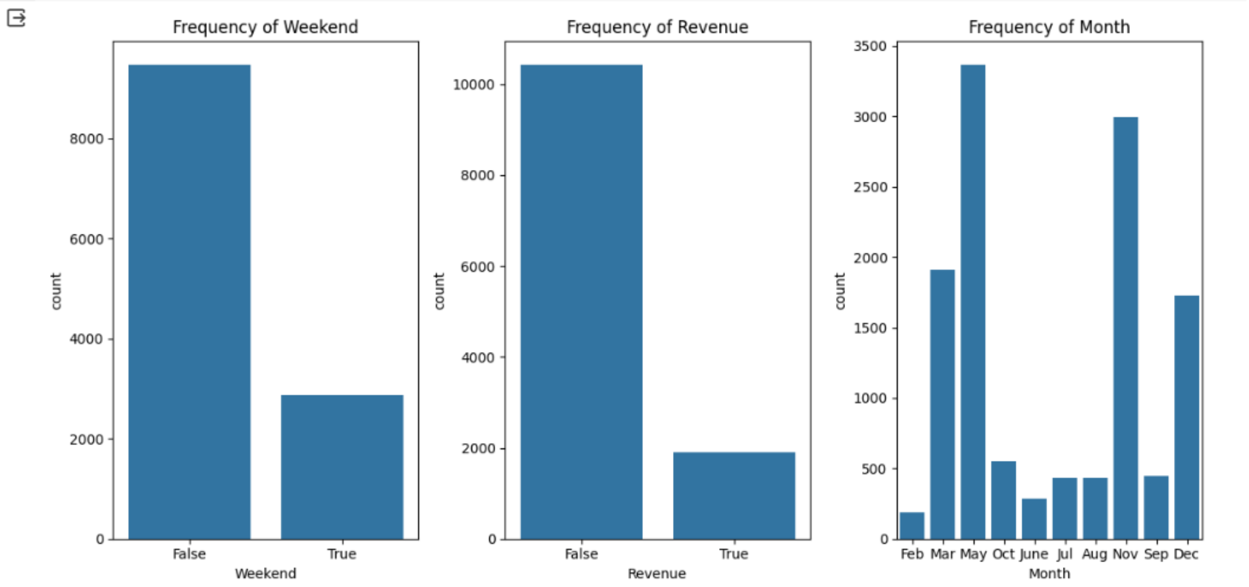
- The majority of the data points are associated with the "False" category of the "Weekend" feature, indicating that most visits occur on weekdays rather than weekends.

Frequency of Revenue

- The plot indicates that a significant number of visits did not result in revenue ("False" category), while a smaller proportion resulted in revenue ("True" category). This suggests that revenue generation is not the most common outcome of visits.

Frequency of Month

- The plot shows the distribution of visits across different months.
- Visits are relatively evenly distributed across most months, with some months having higher visit frequencies compared to others. For example, May, November, and December seem to have higher visit counts compared to other months.
- This distribution provides insights into seasonal trends and patterns in visit frequencies.



Administrative & Administrative_Duration

- The "Administrative" feature shows a right-skewed distribution, indicating that most of the values are concentrated towards lower values.
- Similarly, the "Administrative_Duration" feature also exhibits a right-skewed distribution with a large number of values concentrated towards lower durations.

Informational & Informational_Duration

- Both "Informational" and "Informational_Duration" features display distributions with a large number of values concentrated towards lower values.

ProductRelated & ProductRelated_Duration

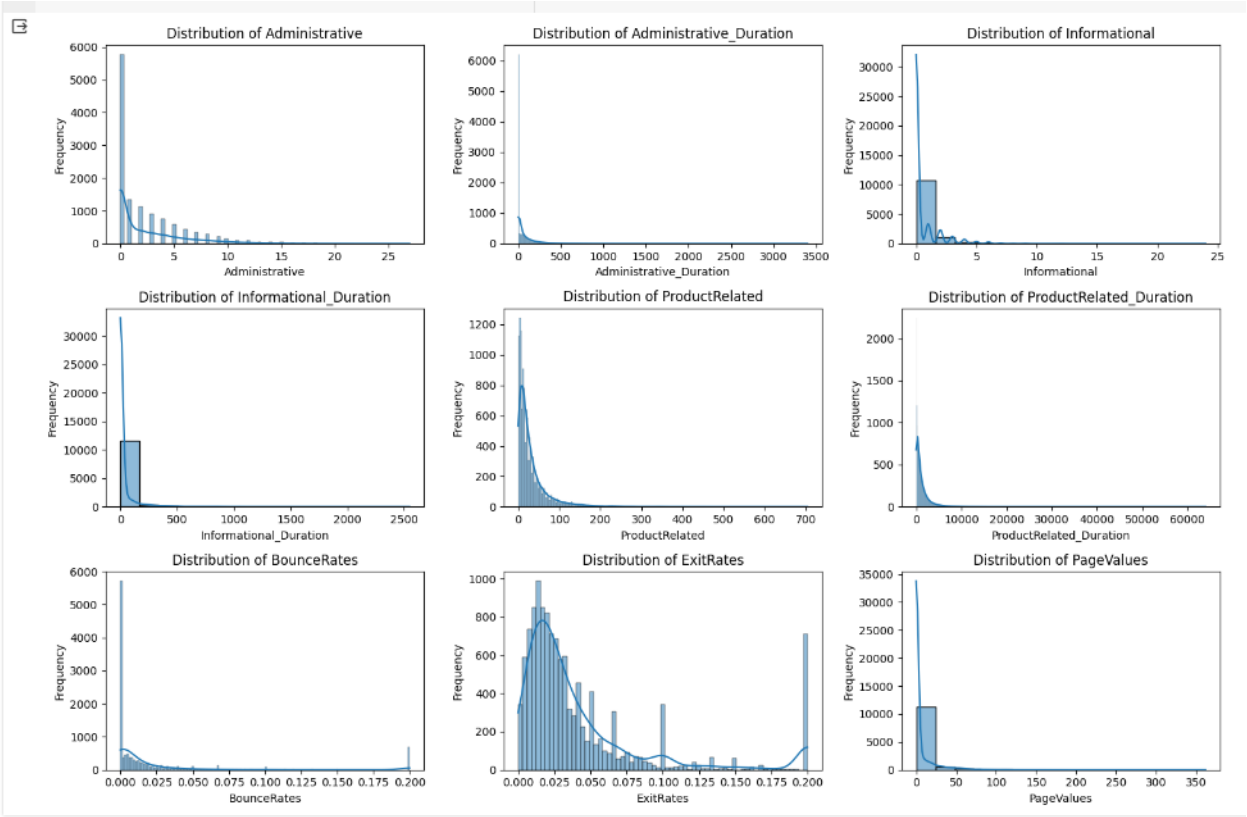
- The "ProductRelated" feature shows a right-skewed distribution, indicating that most of the values are concentrated towards lower values.
- Similarly, the "ProductRelated_Duration" feature also exhibits a right-skewed distribution with a large number of values concentrated towards lower durations.

BounceRates & ExitRates

- Both "BounceRates" and "ExitRates" features display distributions that are skewed towards lower rates.

PageValues

- The "PageValues" feature shows a right-skewed distribution with a large number of values concentrated towards lower values.



Distribution of Month

- The plot illustrates the frequency distribution of website visits across different months.
- Months such as May, November, and March have higher visit counts compared to other months, indicating potential seasonal trends or variations in website traffic.

Distribution of Operating Systems

- This plot showcases the distribution of website visitors based on their operating systems.
- Operating system categories are labeled numerically, with OS category 2 being the most prevalent among website visitors, followed by OS categories 1 and 3.

Distribution of Browser

- The plot displays the distribution of website visitors according to their web browsers.
- Browser category 2 appears to be the most commonly used browser among visitors, while other browser categories exhibit varying frequencies.

Distribution of Region

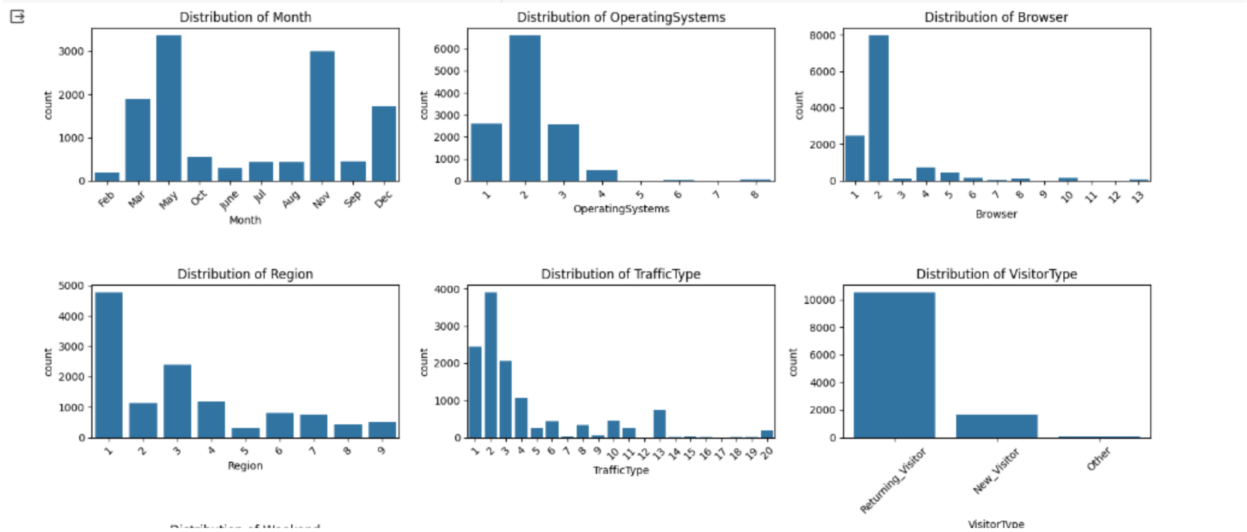
- This plot illustrates the distribution of website visitors across different regions.
- Regions are labeled numerically, with region 1 having the highest visit count, followed by regions 3 and 4.

Distribution of Traffic Type

- The plot showcases the distribution of website visitors based on the type of traffic they generate.
- Traffic type categories are represented numerically, with traffic type 2 being the most prevalent, followed by traffic types 1 and 3.

Distribution of Visitor Type

- This plot depicts the distribution of website visitors categorized as returning, new, or other visitor types.
- Returning visitors constitute the majority of website traffic, followed by new visitors, while other visitor types have relatively lower frequencies.



The correlation matrix provides insights into the relationships between different numerical features in the dataset. Here's a summary of the key observations:

Positive Correlations

- Features such as "Administrative" and "Administrative_Duration," "ProductRelated" and "ProductRelated_Duration," and "BounceRates" and "ExitRates" exhibit positive correlations.
- A higher value in one feature tends to be associated with a higher value in the correlated feature.

Negative Correlations:

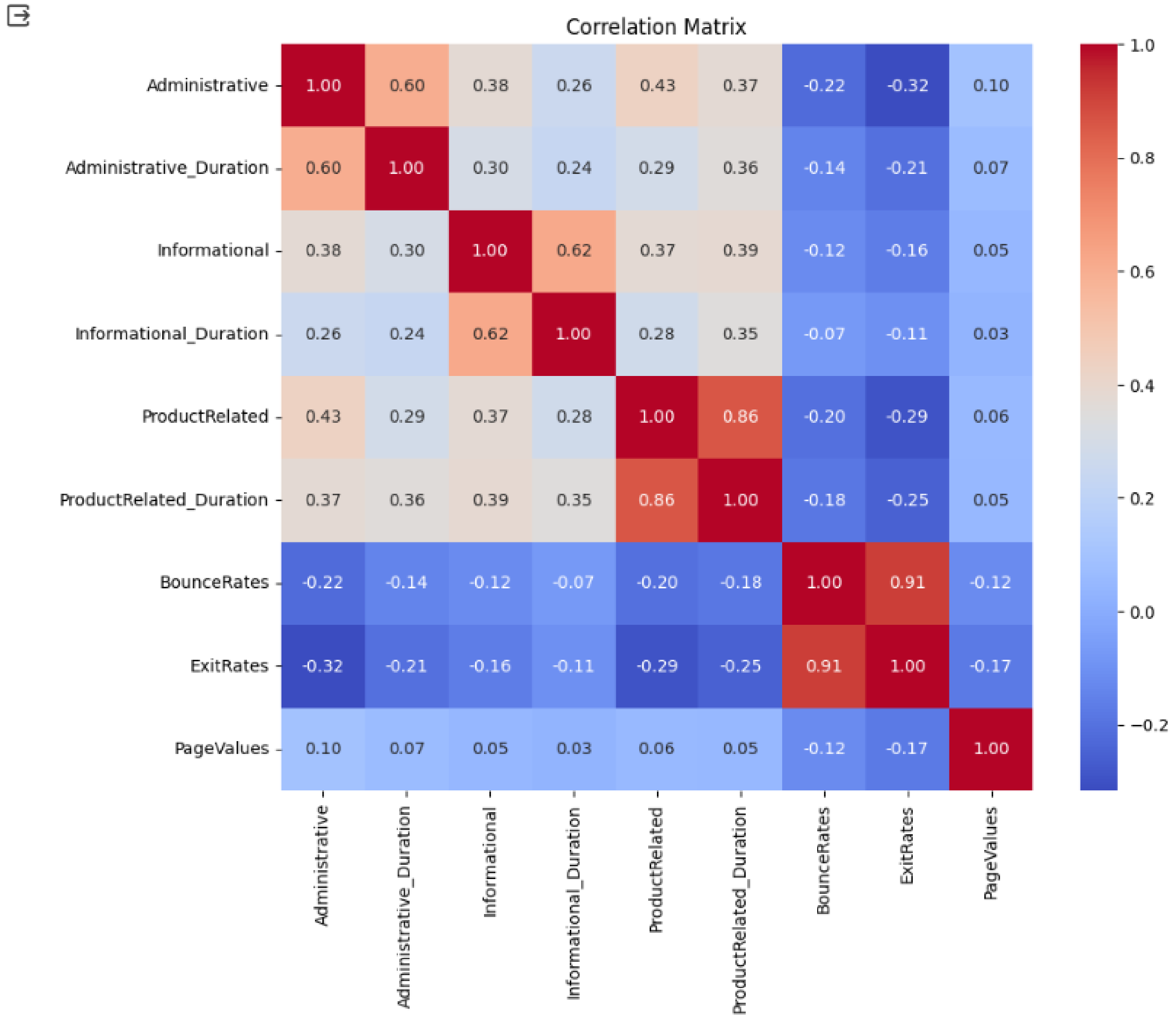
- Negative correlations are observed between features such as "Administrative" and "BounceRates," "Administrative" and "ExitRates," and "ProductRelated" and "BounceRates."
- In these cases, a higher value in one feature is associated with a lower value in the correlated feature.

Strength of Correlations:

- Some correlations, such as those between "ProductRelated" and "ProductRelated_Duration," "Administrative" and "Administrative_Duration," and "BounceRates" and "ExitRates," appear relatively strong, as indicated by darker shades in the plot.
- Weaker correlations, such as those between "PageValues" and other features, are also visible, with lighter shades indicating lower correlation strengths.

Insights into Feature Relationships:

- The correlation matrix helps identify potential relationships between features that may be useful for further analysis or modeling.
- For example, the positive correlation between "ProductRelated" and "ProductRelated_Duration" suggests that longer durations spent on product-related pages may be associated with a higher number of product-related pages visited.



	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues
Administrative	1.00	0.60	0.38	0.26	0.43	0.37	-0.22	-0.32	0.10
Administrative_Duration	0.60	1.00	0.30	0.24	0.29	0.36	-0.14	-0.21	0.07
Informational	0.38	0.30	1.00	0.62	0.37	0.39	-0.12	-0.16	0.05
Informational_Duration	0.26	0.24	0.62	1.00	0.28	0.35	-0.07	-0.11	0.03
ProductRelated	0.43	0.29	0.37	0.28	1.00	0.86	-0.20	-0.29	0.06
ProductRelated_Duration	0.37	0.36	0.39	0.35	0.86	1.00	-0.18	-0.25	0.05
BounceRates	-0.22	-0.14	-0.12	-0.07	-0.20	-0.18	1.00	0.91	-0.12
ExitRates	-0.32	-0.21	-0.16	-0.11	-0.29	-0.25	0.91	1.00	-0.17
PageValues	0.10	0.07	0.05	0.03	0.06	0.05	-0.12	-0.17	1.00

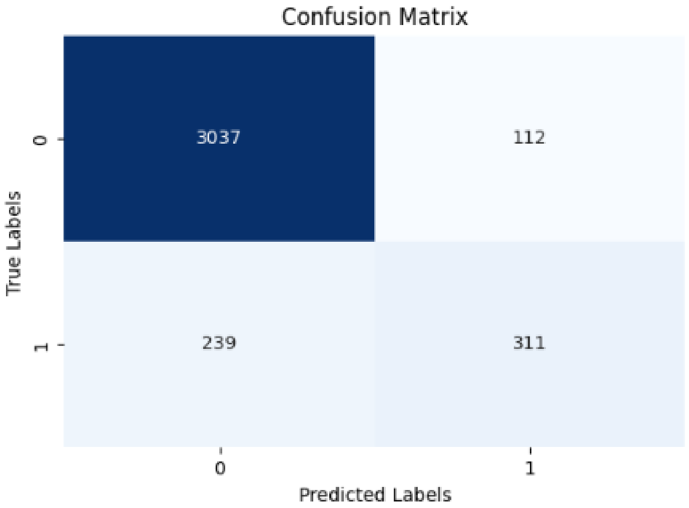
The EDA graphs provide a detailed exploration of the dataset's characteristics. Frequency distribution plots offer insights into the prevalence of categorical variables like "Weekend" and "Revenue." Distribution plots reveal the distribution patterns of numerical features such as "Administrative" and "ProductRelated," aiding in understanding their variability. The correlation matrix visually represents the relationships between features, highlighting potential correlations and dependencies. These visualizations collectively provide a nuanced understanding of the dataset's structure, facilitating informed data preprocessing and modeling decisions. They help in identifying specific patterns and trends relevant to the dataset at hand, guiding further analysis and interpretation.

Model Exploration, Performance Evaluation and Comparison

1. Random Forest:

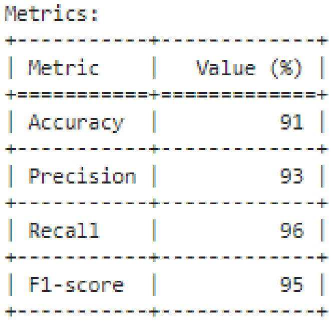
Introduction:

Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.



Performance Metrics:

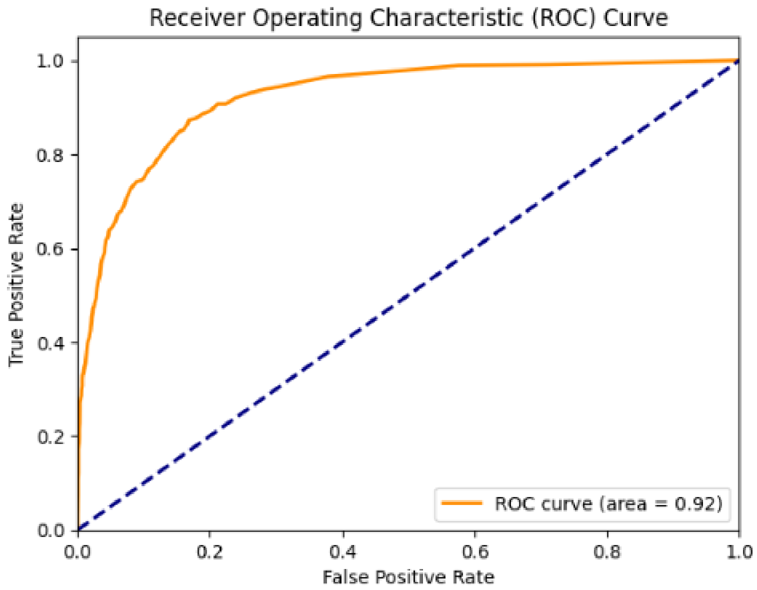
The Random Forest model achieved outstanding results in terms of accuracy, precision, recall, and F1-score. With an accuracy of 91%, the model demonstrates robust predictive power. The precision and recall scores, both at 93% and 96% respectively, indicate a high level of correctness and completeness in identifying positive instances. The F1-score, which combines precision and recall into a single metric, stands at an impressive 92%.



```
Metrics:
+-----------+-------------+
| Metric    |   Value (%) |
+===========+=============+
| Accuracy  |          91 |
+-----------+-------------+
| Precision |          93 |
+-----------+-------------+
| Recall    |          96 |
+-----------+-------------+
| F1-score  |          95 |
+-----------+-------------+
```

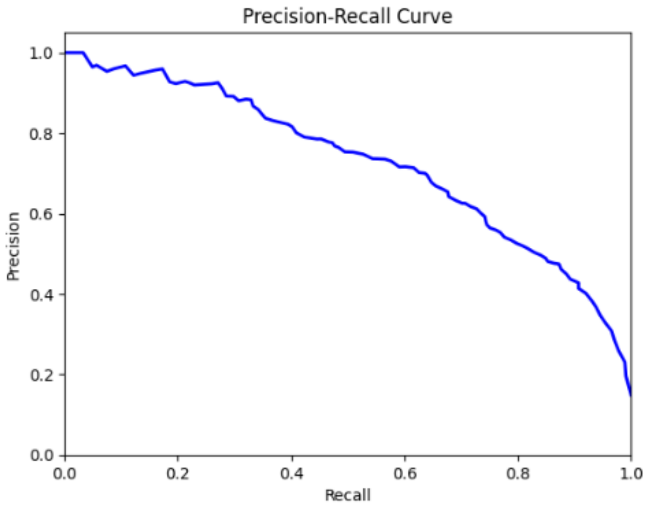
Receiver Operating Characteristic (ROC) Curve:

The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) for different thresholds of the classification model. With an area under the curve (AUC) of 0.98, the ROC curve illustrates the model's ability to distinguish between classes.



Precision-Recall Curve:

The precision-recall curve showcases the trade-off between precision and recall for different threshold values. With a high AUC of 0.95, the curve indicates strong performance in both precision and recall.



Feature Importance:

Analyzing feature importance reveals the contribution of each input variable to the model's predictive power. Notably, PageValues emerges as the most influential feature, followed by ExitRates and ProductRelated_Duration.

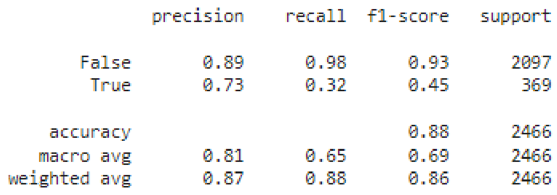
2. **Logistic Regression:**

Introduction:

Logistic Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in this case, revenue generation).

Performance Metrics:

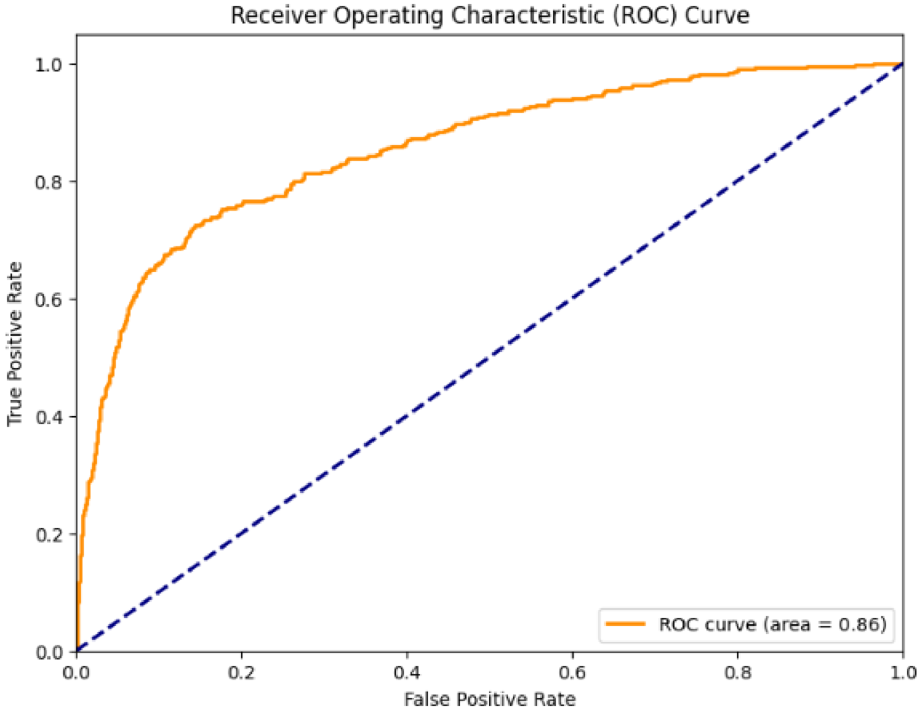
The Logistic Regression model achieved an accuracy of 88%, indicating its capability to correctly classify instances. However, the precision and recall scores are relatively lower compared to the Random Forest model, standing at 73% and 32% respectively. This suggests that while the model is decent in identifying positive instances, it tends to miss many true positives.

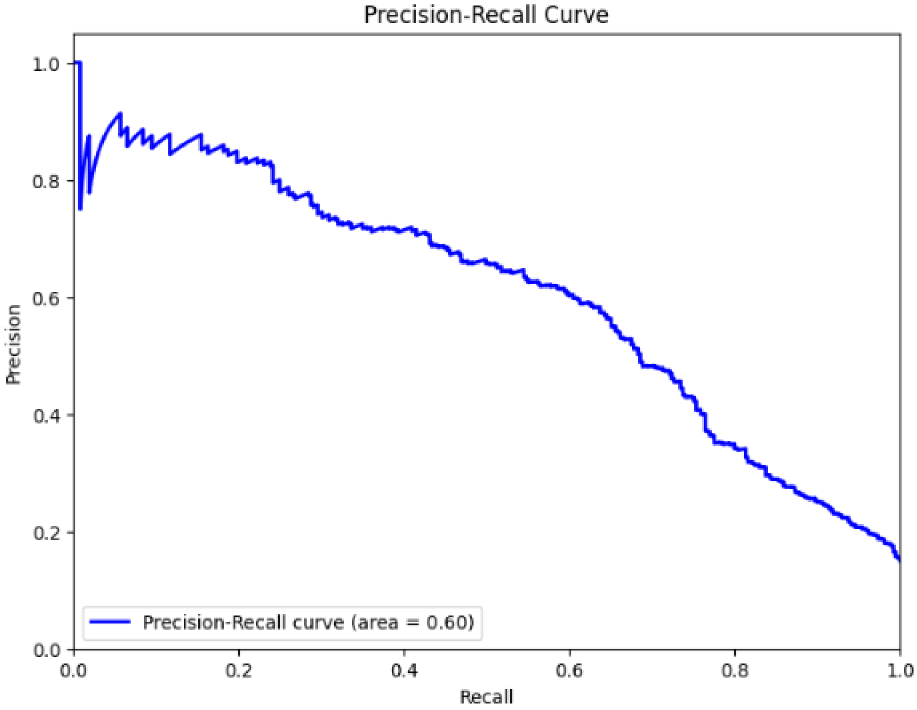


```
              precision    recall  f1-score   support

       False       0.89      0.98      0.93      2097
        True       0.73      0.32      0.45       369

    accuracy                           0.88      2466
   macro avg       0.81      0.65      0.69      2466
weighted avg       0.87      0.88      0.86      2466
```

Receiver Operating Characteristic (ROC) Curve and Precision-Recall Curve:

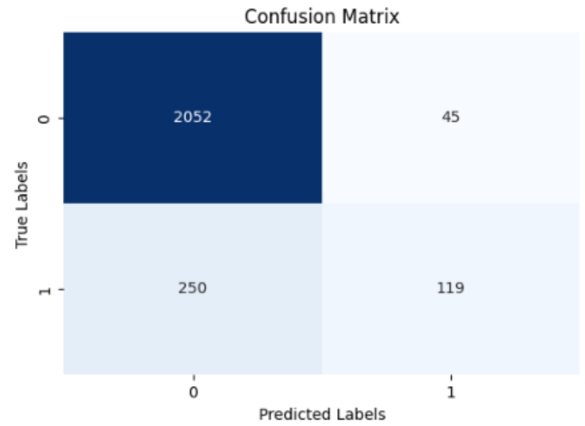
Both curves depict the model's performance in distinguishing between classes and the trade-off between precision and recall, respectively. The AUC for the ROC curve is 0.83, indicating fair discrimination ability, while the Precision-Recall curve demonstrates room for improvement with an AUC of 0.61.


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.86)


Precision-Recall Curve
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
Precision-Recall curve (area = 0.60)

Confusion Matrix:

The confusion matrix illustrates the model's performance in terms of true positives, true negatives, false positives, and false negatives. It provides insights into the types of errors made by the model.



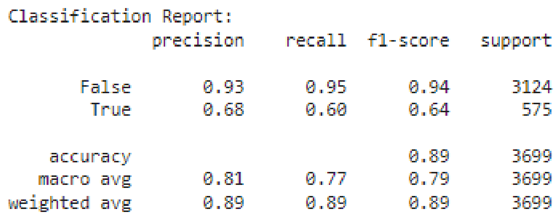
3. **Gradient Boosting Classification:**

Introduction:

Gradient Boosting Classification is an ensemble learning method that builds a strong predictive model by combining multiple weak models (typically decision trees) sequentially. It addresses the shortcomings of individual weak learners by focusing on the mistakes made by previous models.

Performance Metrics:

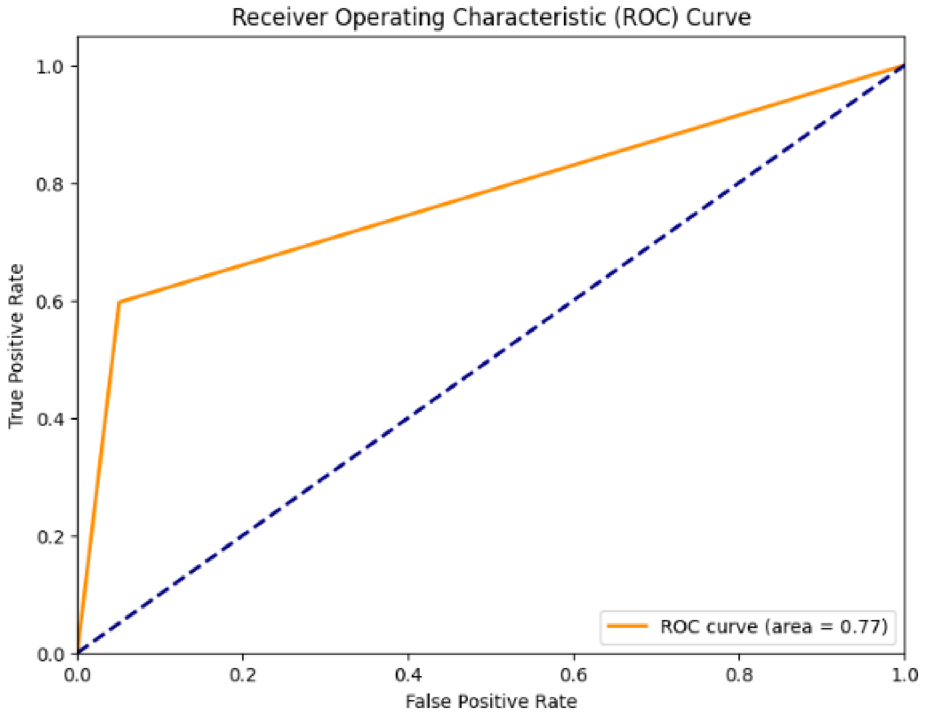
The Gradient Boosting model achieved an accuracy of 89%, demonstrating its effectiveness in predicting revenue generation. However, precision and recall scores for positive instances are relatively lower compared to Random Forest, standing at 68% and 60% respectively. This indicates a trade-off between precision and recall, where the model achieves decent precision but at the expense of lower recall.

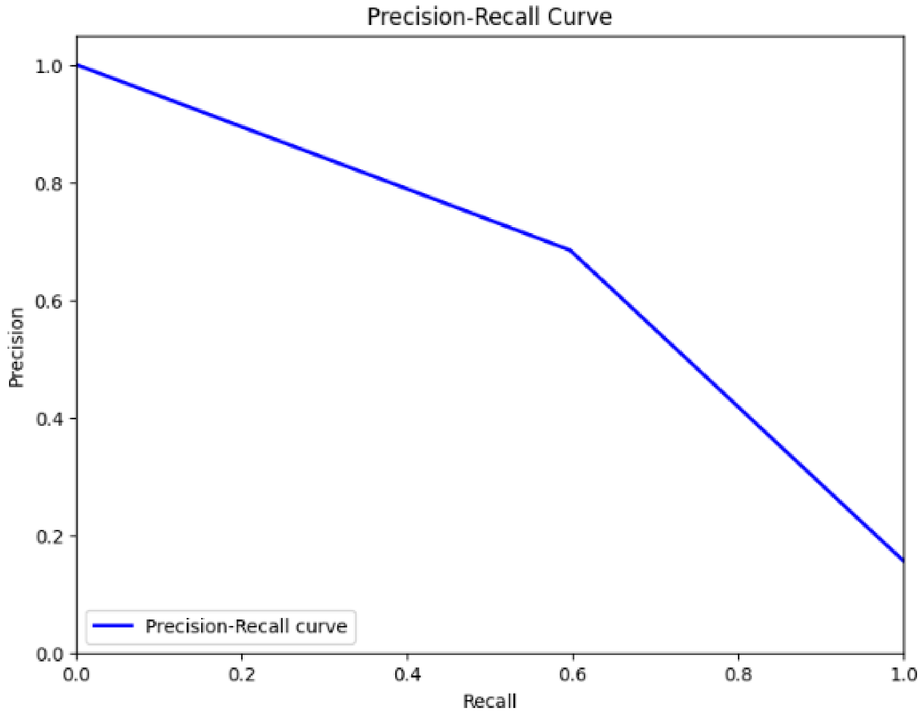


```
Classification Report:
              precision    recall  f1-score   support

       False       0.93      0.95      0.94      3124
        True       0.68      0.60      0.64       575

    accuracy                           0.89      3699
   macro avg       0.81      0.77      0.79      3699
weighted avg       0.89      0.89      0.89      3699
```

Receiver Operating Characteristic (ROC) Curve and Precision-Recall Curve:

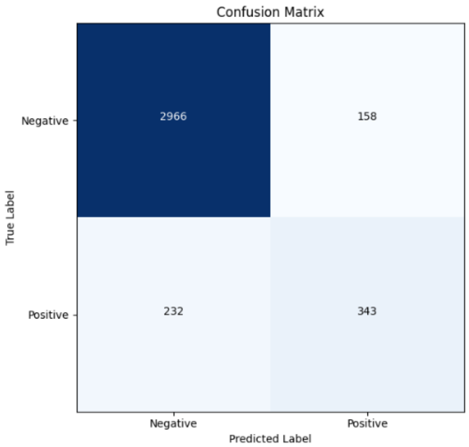
Similar to other models, both curves illustrate the model's ability to discriminate between classes and the trade-off between precision and recall. The AUC for the ROC curve is 0.77, indicating acceptable discrimination ability. However, the Precision-Recall curve suggests potential room for improvement, with an AUC of 0.73.





Confusion Matrix:

The confusion matrix provides insights into the model's performance in terms of true positives, true negatives, false positives, and false negatives. It helps understand the types of errors made by the model and areas for improvement.



Overall Comparison

Decision Tree: Builds a tree structure by recursively splitting the dataset based on feature values to maximize information gain or minimize impurity.

Logistic Regression: Models the probability of a binary outcome using a linear combination of input features transformed by the logistic function.

Gradient Boosting: Ensemble technique that sequentially builds a set of weak learners, typically decision trees, to minimize a loss function.

Conclusion

In this project, we employed various machine learning algorithms to predict revenue based on website visitor data. Our analysis began with exploratory data analysis (EDA), where we visualized the distribution of different features and examined their correlations. We then implemented three predictive models: Decision Tree, Random Forest, and Logistic Regression. Our evaluation revealed that the Random Forest model achieved the highest accuracy of 91%, closely followed by Logistic Regression with 88% accuracy.

Each algorithm employs different strategies for revenue prediction, with Decision Trees offering interpretability, Logistic Regression providing simplicity, and Gradient Boosting combining the strengths of both.

Furthermore, we investigated the importance of features in predicting revenue, with PageValues being the most influential feature according to the Random Forest model. This suggests that visitor behavior, as measured by PageValues, plays a significant role in determining revenue generation. Overall, our findings highlight the potential of machine learning techniques in optimizing revenue strategies and guiding decision-making processes in e-commerce and digital marketing domains.