Unveiling Patterns in E-commerce: A Predictive Analytics Exploration of Online Shoppers Purchasing Intention Data

Md Khaled Saifullah
Student Number: 501230688
Supervisor: Tamer Abdou, PhD
Date of Submission: 1st April 2024

Ryerson University

Table of Contents

Abstract

Unveiling Patterns in E-commerce: A Predictive Analytics Exploration of Online Shoppers Purchasing Intention Data

In the rapidly evolving digital landscape, understanding user behavior within e-commerce platforms is crucial for businesses aiming to optimize strategies and enhance revenue. This study embarks on an exploration of purchasing intention data within the e-commerce domain, employing Predictive Analytics as a guiding methodology. The primary aim is to decode intricate patterns embedded within user engagement data, anticipate user actions, and thereby bolster online business revenue.

The project confronts the formidable task of deciphering complex patterns enshrouded within copious volumes of clickstream data. Five cardinal research questions have been articulated to address this challenge, focusing on aspects such as the influence of different page categories on user purchases and variations in user behavior during weekends and special days. The endeavor is to glean insights into engagement metrics' predictability for transactions and identify distinct user segments predicated on categorical attributes. The project aims to answer the following research questions:

How do different informative page categories (like Informational and Product Related pages) contribute to the likelihood of a user making a purchase?

Can we predict the likelihood of a user making a purchase based on metrics such as Bounce Rates, Exit Rates, and Page Values?

How does user behavior vary on weekends compared to weekdays in terms of engagement and conversion rates?

What is the impact of special days on user engagement and transaction rates?

Since question 3 and 4 both deals with the time of purchase and generating revenue, it might be difficult to distinguish the duration since out dataset is only containing 10 months dataset and weekend-weekdays might overlap with the way the data represent the special events. Rather considering the trend of timings for all together and answering the question as a combined will be more practical and informative for the findings.

Question 3: What is the relationship between features related to timing (Weekend, Month, and special Day) and revenue generation?

This research project will leverage the Online Shoppers Purchasing Intention Dataset sourced from the UC Irvine Machine Learning Repository as the primary dataset for investigative analysis and to address the formulated research questions. This robust dataset, comprising 12,330 instances enriched with 10 numerical and 8 categorical attributes, forms the foundation for this analytical expedition. Attributes encapsulate diverse elements including pages visited, visit durations, bounce rates, exit rates, page values, special day indicators alongside user characteristics like operating systems browsers and regions.

A hybrid analytical approach is adopted to navigate through these research questions, amalgamating exploratory data analysis with statistical tests while leveraging classification models the models will include Decision trees, Random Forests, K-nearest neighbours, and clustering algorithms for deeper insights. Python emerges as the instrumental programming language supported by libraries like Scikit-learn for machine learning applications; Pandas for

adept data manipulation; Matplotlib facilitating intuitive visualizations-all encapsulated within Jupyter Notebooks environment ensuring an interactive exploration.

Introduction

The culmination of this study promises actionable intelligence enabling businesses to refine strategies meticulously; elevate user experiences aligning with individual preferences; optimize revenue streams adapting to dynamic market trends-all underpinned by a comprehensive understanding gleaned from e-commerce data analytics. This integration of Predictive Analytics underscores a paradigm shift towards informed decision-making processes illuminating pathways for enhanced business performance in an increasingly competitive digital marketplace.

Approach

1. Data Collection:
    - Upload datasets to Google Colab.
    - Download an Excel
2. Data Preparation:
    - Clean data (duplicates, correct errors, handle missing values).
    - Visualize relationships between variables and address class imbalances.
    - Split data into training and evaluation sets.
3. Model Development:
    - Predict Purchasing decision using classification methods (LR, DR, RF, SVM, KNN).
    - Describe Purchasing decision based on the online data.

4. Evaluate the Model:

    - Use validation and cross-validation techniques.
    - Prioritize models with high recall for target = 1 and overall accuracy.

5. Parameter Tuning:

    - Employ backward elimination to select essential features.
    - Iteratively compare different algorithms based on accuracy, precision, and recall choosing the best-performing one.

Research Questions

Question 1

How do different informative page categories contribute to the likelihood of a user making a purchase?

In the ever-evolving landscape of e-commerce, understanding the factors that contribute to user purchasing decisions is paramount for businesses seeking to optimize their online platforms and enhance the overall user experience. One crucial aspect of this understanding involves exploring the impact of different informative page categories on the likelihood of a user making a purchase.

Different informative page categories, such as informational and product-related pages, can contribute to the likelihood of a user making a purchase. Research has shown that the expression of advertised information can effectively enhance consumers' purchase intention, depending on the product type. Understanding the impact of different informative page categories and tailoring the expression of advertised information accordingly can help improve users' likelihood of making a purchase.

The research by Peter et al. investigates the influence of **product description** on **purchase intention** within the context of online stores on marketplace platforms. The study aims to determine whether product description affects purchase intention. The researchers employ **structural equation modeling** (SEM-PLS) using SmartPLS 3.3.7 software. The research findings reveal that product description has a **positive and significant effect** on both enduring and situational involvement. The results diverge from previous studies that suggested both enduring and situational involvement play a role in influencing purchase intention.

Understanding how different informative page categories -such as product descriptions, impact user behavior is essential for designing effective e-commerce strategies. By recognizing the nuanced interplay between involvement, content, and purchase intent, businesses can tailor their approaches to meet user needs and drive successful conversions.

Wang et al. (2022) in this paper “Promote or inhibit? Research on the transition of consumer potential purchase intention. Annals of Operations Research” used a continuous-time hidden Markov model (CT-HMM) to capture the transfer path of consumers who are affected by various online channels and found that online page view (advertising) has a positive and statistically significant impact on the transition of consumer purchase intention.[2]

Sunarto et al. (2023) [1] paper aims to reduce classification error in online shoppers' purchasing intention using the boosting technique with the C4.5 algorithm. The experimental results show that applying the boosting technique improves accuracy and lowers classification error compared to using the C4.5 algorithm without boosting. The paper develops a real-time system for predicting online shoppers' purchasing intentions by utilizing the **Random Forest algorithm** and considering both known and anonymous sessions. The results show that the Random Forest algorithm outperforms other methods in terms of accuracy and F1 Score, highlighting its effectiveness in forecasting purchasing intent in an e-commerce environment.

The study conducted by **Frazier et al. (2022)** [4] focuses on the data analysis of online shopper's purchasing intention, employing machine learning for predictive analytics. The research adopts a comprehensive data analytics workflow, encompassing data ingestion, descriptive statistics, and

exploratory data analysis. The authors employ various methods, including frequency, probability histograms, and scatterplots, to examine the impact of numerical features on the target class. The analysis extends to categorical features, exploring their influence on the class label through frequency and probability assessments. We will use the statistical approaches although the data processed prior to this study are different as they use numerical instead of categorical variable.

Kabir et al. (2019) [5] in the paper analyzed different classification algorithms such as Decision Tree, Random Forest, Naive Bayes, and SVM to predict whether a customer will make a purchase or not when visiting the webpages of an online shop. Ensemble methods were also performed to boost the performance of these algorithms. We are planning to use the classification algorithms to find out the different informative page categories impact in purchasing decision. The study analyzed empirical data of online shoppers to build a prediction model for their purchase intention. The application of ensemble methods to the dataset made this study unique.

Kushwaha, Rahul., P., Anudeep., Desai, Shubham, Uday. (2023) “Variable Aware Analytic Driven Online Shoppers Purchasing Intention using ML Algorithms” [3] in their final model achieved a higher accuracy on the test set, indicating that it can accurately classify website visitors with a high degree of accuracy. The Random Forest algorithm outperformed other methods in terms of accuracy and F1 Score, demonstrating its effectiveness in predicting online shoppers' purchasing intentions in an e-commerce environment. The analysis identified the most crucial features for predicting purchasing intentions, including PageValues, Exit Rates, Administrative Duration, ProductRelated Duration, and Bounce Rates. These findings offer valuable insights into the factors that impact the purchasing decisions of website visitors. The study highlights the success of machine learning techniques in anticipating the intentions of internet shoppers and emphasizes the significance of properly preparing data and developing appropriate features to construct precise models. The research also emphasizes the importance of evaluating ML algorithms' performance through different metrics to identify their strengths and weaknesses.

Question 2

Can we predict the likelihood of a user making a purchase based on visitor type (new, returning, other)? / Is there an impact of visitor type (new, returning, other) on whether revenue will be generated?

The likelihood of a user making a purchase can be predicted based on metrics such as new, returning, other. Descriptive statistical analysis of online shoppers' purchasing intention data showed that variables like time spent and page values are positively correlated with shopping intention, while variables including bounce rate and exit rate are negatively correlated with shopping intention.

In "The Application of Machine Learning in Online Purchasing Intention Prediction" by Xiang, Shi. (2021) [7], the author employs a machine learning approach to predict online purchasing intention. The methods involve conducting descriptive statistical analysis to discern variables correlated with shopping intention, providing valuable insights into factors influencing consumer behavior. The construction of prediction models, including Logistic Regression, Decision Tree, and Random Forest, demonstrates a comprehensive exploration of classical machine learning techniques. The evaluation of model performance using train accuracy and test accuracy adds a quantitative dimension to the study. However, the paper acknowledges limitations, particularly the exclusive consideration of aggregated page view data during the visit session and other user information. This limitation raises awareness of the potential oversight of other relevant variables that could significantly impact shopping intention, suggesting avenues for future research to enhance the predictive capabilities of machine learning models in this domain. I am planning to use VisitorType variable and utilize to predicting the purchase based on the model followed by the author.

Zhaoguang et al. (2021). In their paper “Potential buyer identification and purchase likelihood quantification by mining user-generated content on social media” [9] proposes a two-stage approach to identify potential buyers and quantify their purchase likelihood using user-generated content (UGC) on social media. The approach involves classifying user posts into before buying and after buying, and then using a novel Weighted Recency, Focus, and Sentiment (WRFS)

model to quantify purchase likelihood. The method is verified using data from the Honda Civic community in the Bitauto automotive forum. The WRFS (Weighted Recency, Focus, and Sentiment) model is a novel approach used in this research paper to quantify the purchase likelihood of potential buyers based on user-generated content (UGC) on social media. The model considers three key factors: recency, focus, and sentiment. Recency refers to the timing of the user's posts, with more recent posts being given higher weightage. Focus refers to the content of the posts, distinguishing between posts made before buying and after buying. Sentiment refers to the sentiment expressed in the posts, such as positive or negative sentiment towards the product. By combining these factors, the WRFS model provides a quantification of the purchase likelihood for potential buyers identified through the classification of user posts. The accuracy of the model is demonstrated through the observation and verification of actual purchases made by these potential buyers. This paper is taking Recency (time), Focus (Informative page) and Sentiment (text-based data) which are similar to my research question, they are too complex to this research to explore.

Question 3

What is the relationship between features related to timing (`Weekend`, `Month`, and `special Day`) and revenue generation?

Timing features such as weekends, months, and special days significantly influence revenue generation. Consumer behavior changes during weekends and special days like holidays or sales events, often leading to increased sales. Monthly revenue fluctuations can occur due to seasonal factors, such as increased retail sales during the holiday season or demand for certain services at specific times of the year. These timing features are crucial for revenue forecasting models, helping businesses anticipate revenue fluctuations and plan accordingly. However, the impact of these features can vary depending on the specific business and industry context.

Kawa et al. (2012) shows in the paper “A Descriptive Study on the Purchase Timing Effect in Athletic Shoes -Focused on Day-of-the-week Effect and Intra-month Effect” [10] describes the purchase timing behavior for athletic shoes, focusing on the day-of-the-week effect and the intra-

month effect. It uses daily sales data from a domestic brand in Korea from January 2006 to December 2010. The results show that Saturday and Sunday have significantly higher sales than weekdays. Additionally, the first and third 10-days-of-the-month yield higher sales volume than the second 10-days-of-the-month. The department store's sales volume is higher in the first 10-days-of-the-month compared to discount and franchised stores, while the second 10-days-of-the-month have higher sales volume for discount and franchised stores based on the time series and statistical analysis.

In the “Data Analysis of Online Shopper’s Purchasing Intention Machine Learning for Prediction Analytics” paper Frazier et al. [4] proposes the use of basic web browsing analytics, machine learning, and artificial intelligence to effectively segment customers and identify those most likely to make a purchase, thereby improving the customer experience and facilitating purchases. It explores the impact of numerical and categorical features on the target class, using frequency, probability histograms, and scatterplots for numerical features, and frequency and probability analysis for categorical features. The paper provides insights into the relationship between weekend versus weekdays transactions and revenue creation, highlighting the importance of understanding user behavior and optimizing website design. It presents a Random Forest Classifier as an initial modeling framework to test the preliminary results of the data analysis [4].

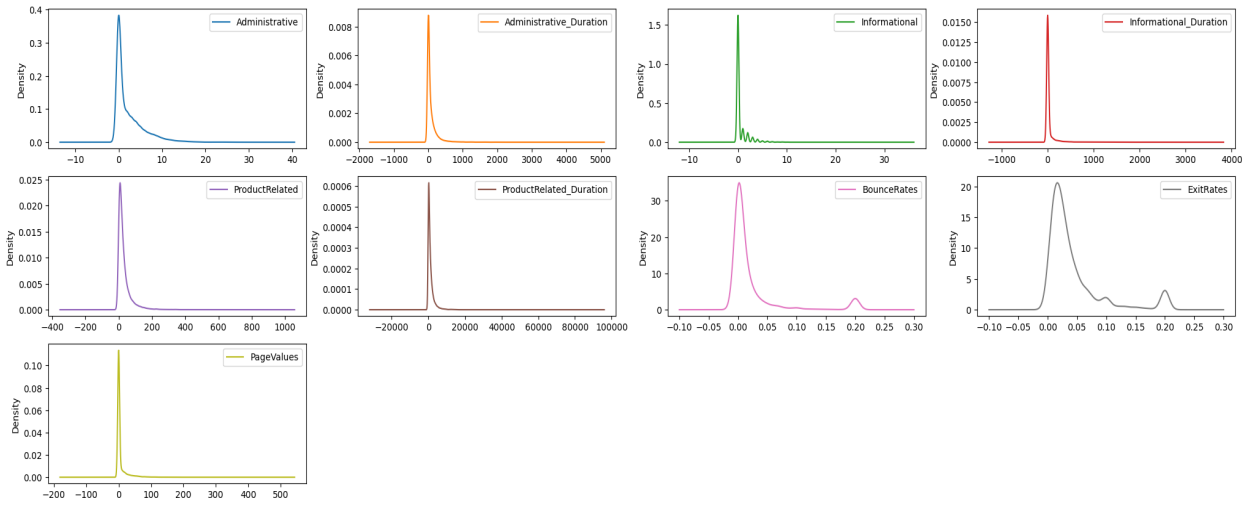
In my study, I am planning to use Linear regression, Decision tree, Random Forest as it is advantageous for my research question because it excels in handling non-linear relationships within timing features (Weekend, Month, Special Day) and their impact on revenue generation. Its ensemble learning, robustness to overfitting, and ability to handle missing values make it a powerful choice, providing valuable insights into feature importance for accurate predictions in the analysis.

Descriptive Statistics

Datasets: Online Shoppers Purchasing Intention Dataset

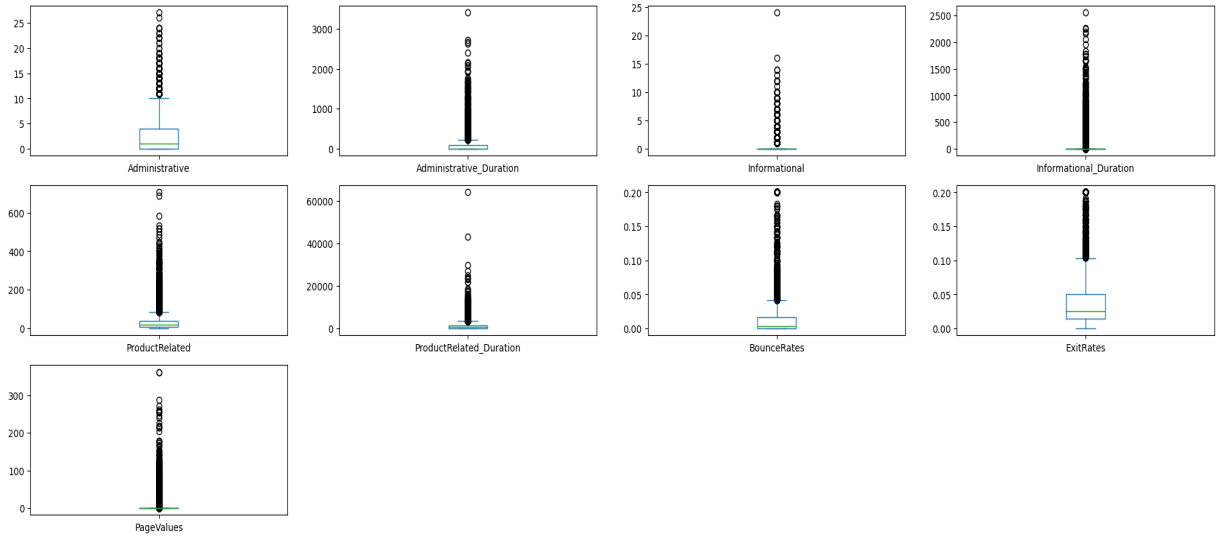
These datasets underwent preliminary cleaning, row data to technically correct data. Of the 12,330 sessions in the dataset, 84.5% (10,422) were negative class samples that did not end with shopping, and the rest (1908) were positive class samples ending with shopping. The dataset contains multivariate, with integer and real number. The dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period. The profile reports will present detailed descriptive statistics about each column and its associated data. Utilizing tools such as matplotlib and seaborn, EDA spans various critical steps, including data collection, summary statistics computation, data visualization through techniques like histograms and scatter plots, exploring data relationships, feature engineering, data transformation, dimensionality reduction, pattern recognition, hypothesis testing, and effective documentation and communication of findings.

Figure 1: density plots for numerical columns



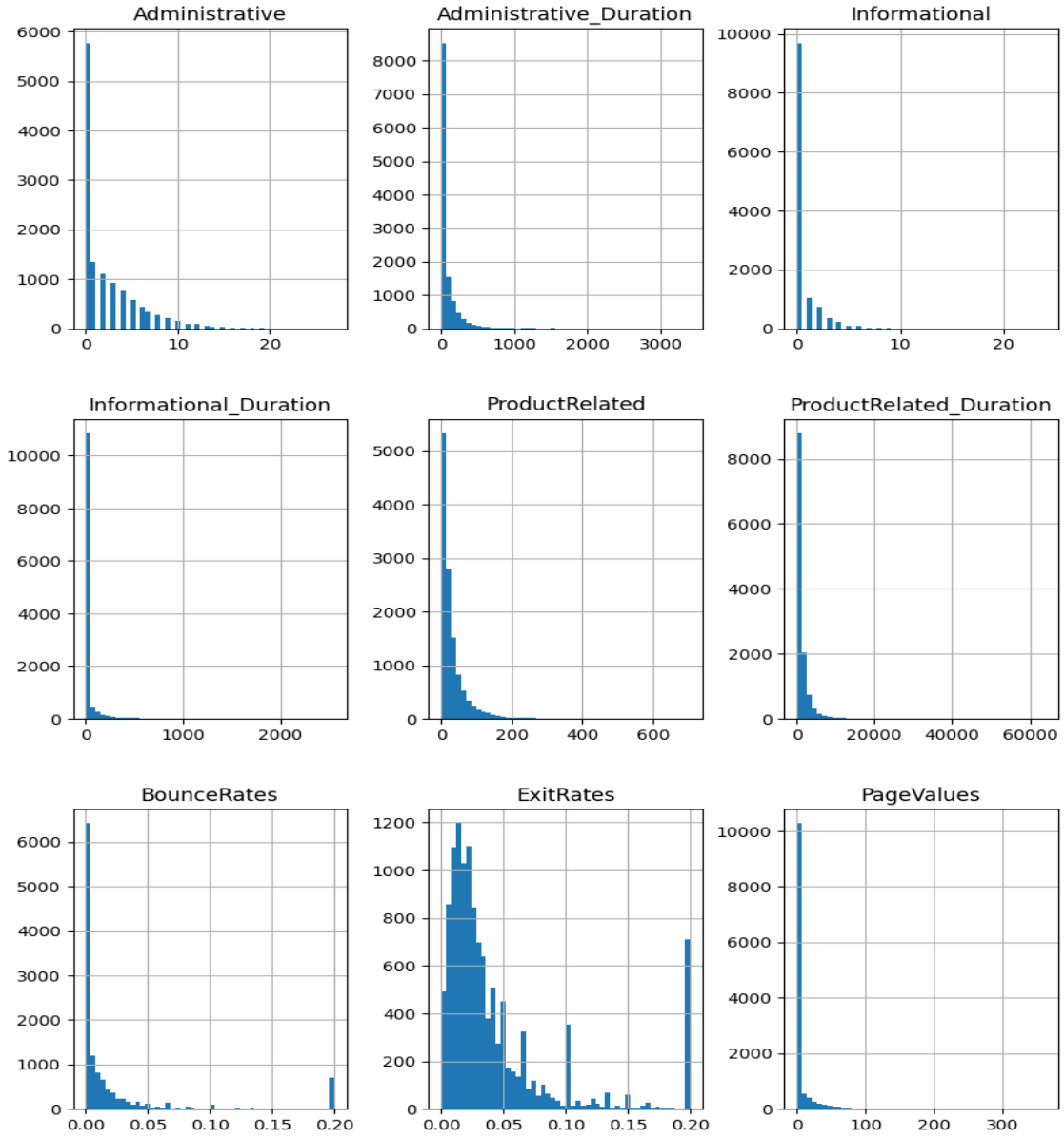
We observed that many numerical data points are skewed to the right, meaning a few users have very high usage numbers. This is typical in online shopping data.

Figure 2: box plot for numerical column.



The graphical box plot provides a visual representation of the distribution of values in a numerical column. The box represents the interquartile range (IQR), with the central line inside the box representing the median. The "whiskers" extend from the box to the minimum and maximum values within a certain range (usually 1.5 times the IQR). Points outside the whiskers are considered outliers. The position of the box and whiskers gives an idea of the spread and right skewed of the data.

Figure 3: Histogram of numerical columns



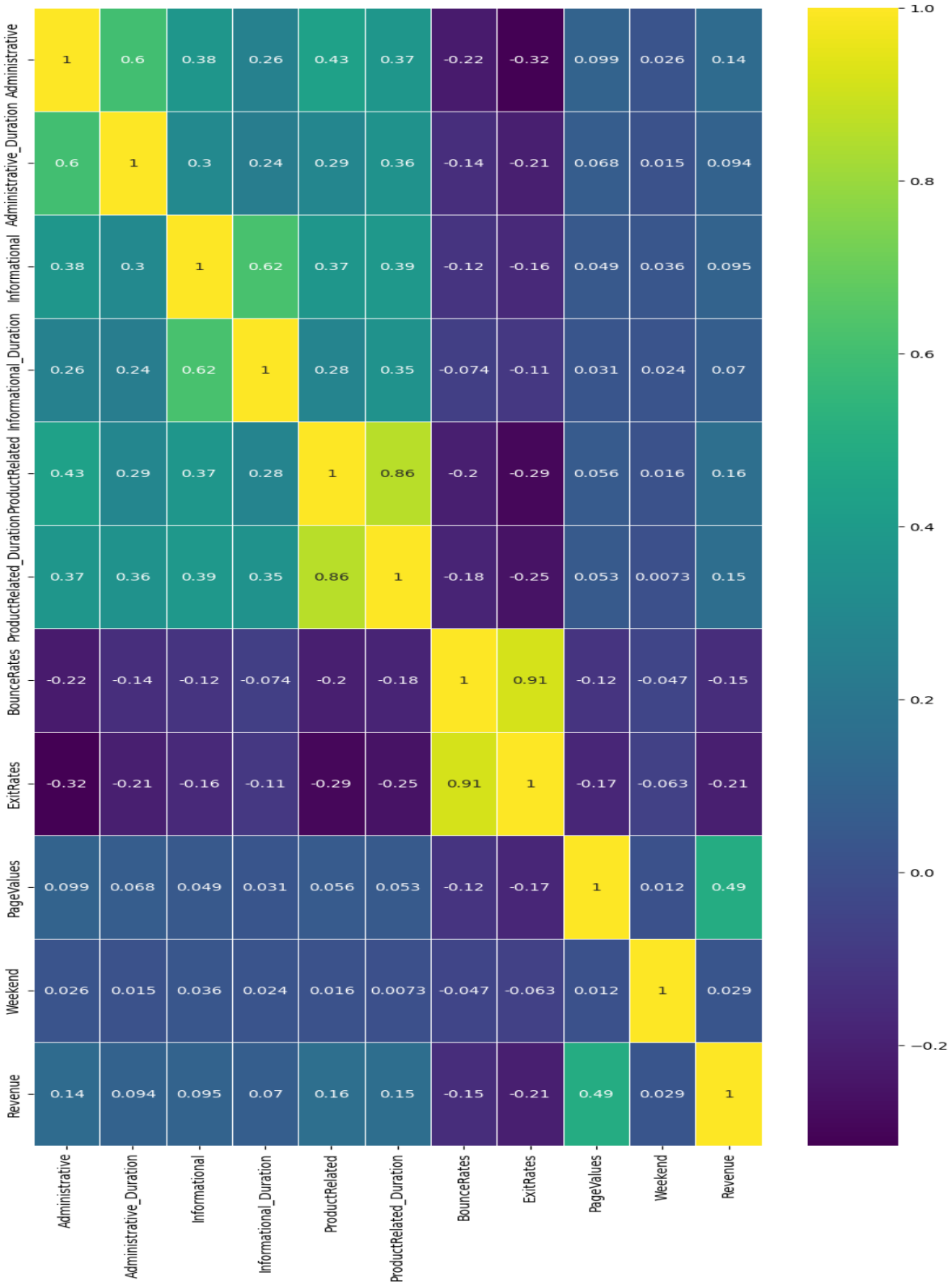
Histogram shown in the figure explained the shape of the data attributes as they are skewed to the right.

Correlation

A correlation heatmap is a graphical tool that displays the correlation between multiple variables as a color-coded matrix. Understanding correlation heatmaps can help us identify patterns and relationships between multiple variables. The color of each cell represents the strength and

direction of the correlation, with darker colors indicating stronger correlations. Positive correlations (when one variable increases, the other variable tends to increase) are usually represented by warm colors, such as red or orange. Negative correlations (when one variable increases, the other variable tends to decrease) are usually represented by cool colors, such as blue or green. We can see that product related duration and product related are highly correlated which make sense as these two variables represent the customers time spending on product related information. Bounce rate and Exit rate is also highly correlated.

Figure 4: Heat plot of correlation



High correlation between:

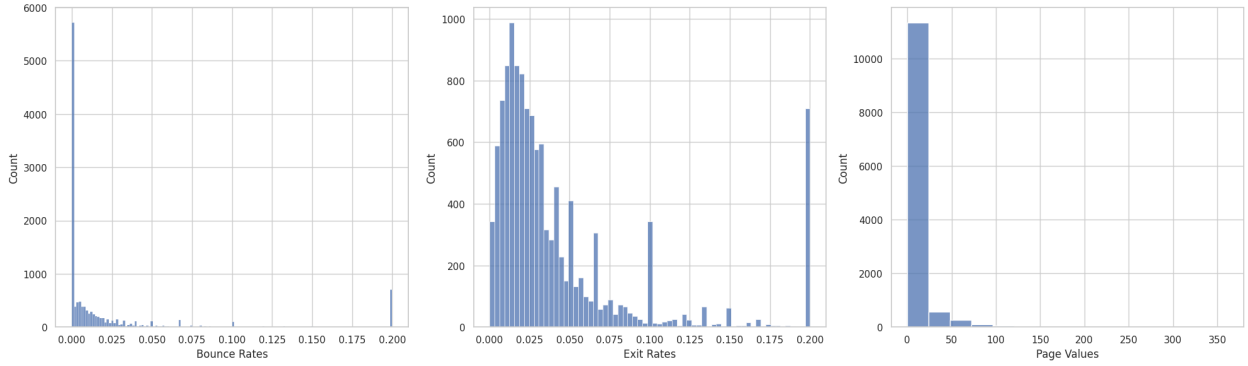
- BounceRates & ExitRates (0.91).
- ProductRelated & ProductRelated_Duration (0.86).

Moderate Correlations:

- Administrative & Administrative DUration (0.6)
- Informational and Informational Duration (0.62)
- Page Values and Revenue (0.49)

Page Matrix analysis:

Figure 5: distribution plots of Page Metrics



The above distribution plots of Page Metrics show the following:

Right-Skewed Distributions with Outliers:

All three features (Bounce Rates, Exit Rates, and Page Values) exhibit right-skewed distributions with a significant number of outliers. Right-skewed distributions indicate that most data points have lower values, while outliers suggest some extreme values that deviate from the overall trend.

Low Average Bounce Rate:

The average bounce rate across the data points is low. This is a positive observation, as it suggests that visitors are generally engaging with the website rather than immediately leaving after viewing a single page. A low bounce rate typically indicates that visitors are finding the content or products on the website relevant and engaging.

Higher Exit Rates compared to Bounce Rates:

Exit rates tend to have higher values compared to bounce rates. This observation is expected, as exit rates account for all exits from the website, including those that occur after visiting multiple pages. In contrast, bounce rates only consider exits that occur after viewing a single page. Transaction confirmation pages or completion of desired actions on the website may cause exit rates to increase, as visitors may leave the website after completing their intended tasks.

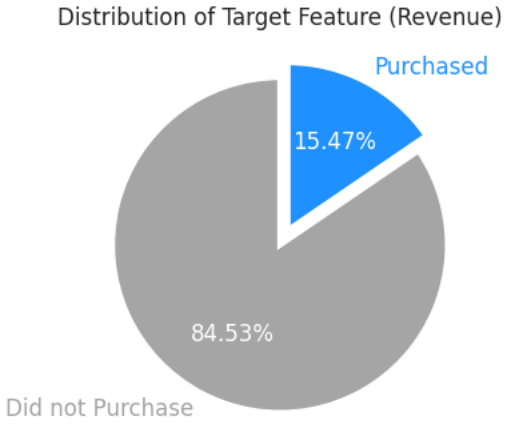
Overall, these observations provide valuable insights into the behavior of visitors on the website. Understanding these patterns can help in optimizing the website's design, content, and user experience to reduce bounce rates, mitigate exit rates, and ultimately enhance visitor engagement and conversion rates.

Revenue Analysis

In the context of online customer transactions, the imbalance in the data where the majority of interaction

The fact that approximately 84.5% of interactions result in no revenue suggests that a large portion of customer interactions on the platform do not directly contribute to generating revenue. These interactions could include browsing, searching, comparing products, or engaging with content that doesn't lead to a purchase.

Figure 6: Distribution of Revenue



Revenue by visitor type

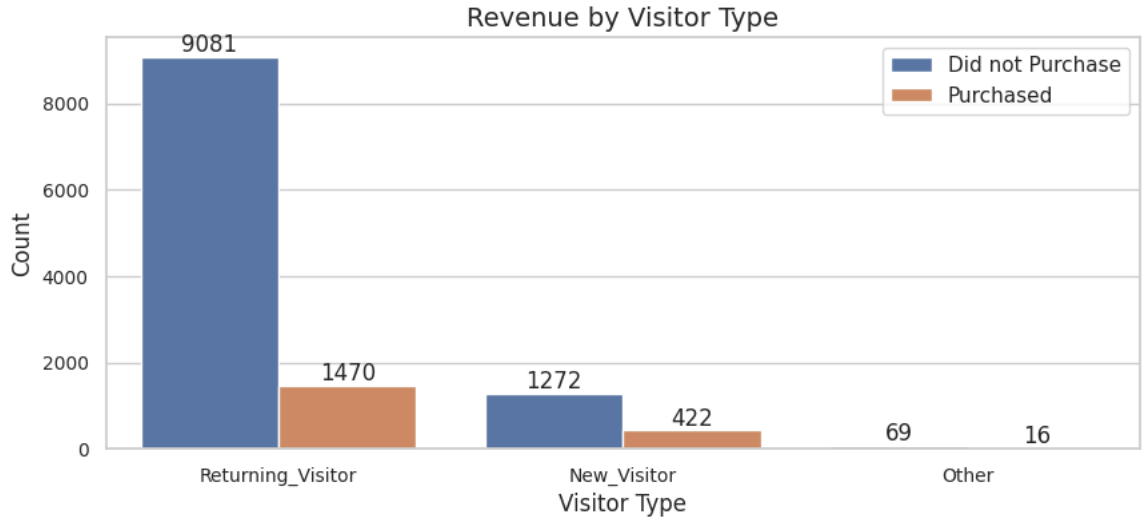
Categorizing visitors into three groups: returning visitors, new visitors, and "other" visitors. Here's a summary of the purchase behavior for each visitor type:

Returning Visitors: Out of 9081 returning visitors, 1470 made a purchase. This implies a conversion rate of approximately 16.2% (1470 purchases out of 9081 returning visitors).

New Visitors: Among 1272 new visitors, 422 made a purchase. This corresponds to a conversion rate of around 33.2% (422 purchases out of 1272 new visitors).

Other Visitors: From a group of 69 visitors categorized as "other," 16 made a purchase. The conversion rate for this group is about 23.2% (16 purchases out of 69 other visitors).

Figure 7: Revenue by Visitor Type



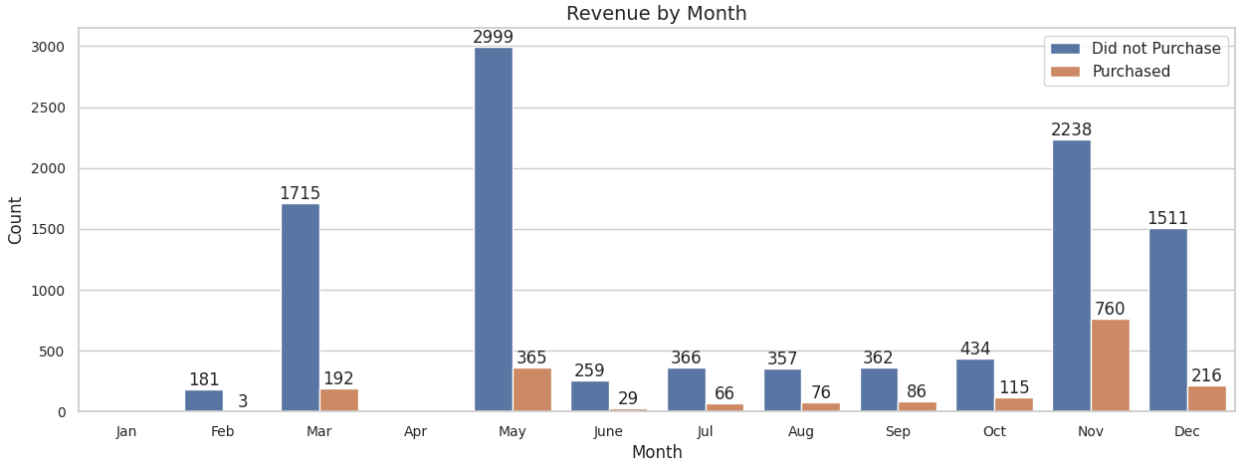
Based on this figure, it's evident that returning visitors have a lower conversion rate compared to new visitors. Despite having a larger pool of returning visitors, the percentage of them making purchases is relatively lower. However, it's crucial to recognize the significance of returning visitors in generating revenue.

Returning visitors are crucial for online platforms due to their loyalty, higher lifetime value, and cost-effectiveness compared to new visitors. Despite having a lower conversion rate, their consistent engagement demonstrates satisfaction with the platform, leading to multiple purchases over time. By focusing on retaining and engaging returning visitors, platforms can optimize resources and capitalize on opportunities for upselling and cross-selling. Additionally, satisfied returning visitors often become brand advocates, promoting the platform, and attracting new customers through word-of-mouth referrals, thereby fostering revenue growth and long-term success.

Revenue by Month:

Here's a pattern of revenue distribution by month, with no data available for January and April. Additionally, it appears that a significant portion of transactions occurred towards the end of the year.

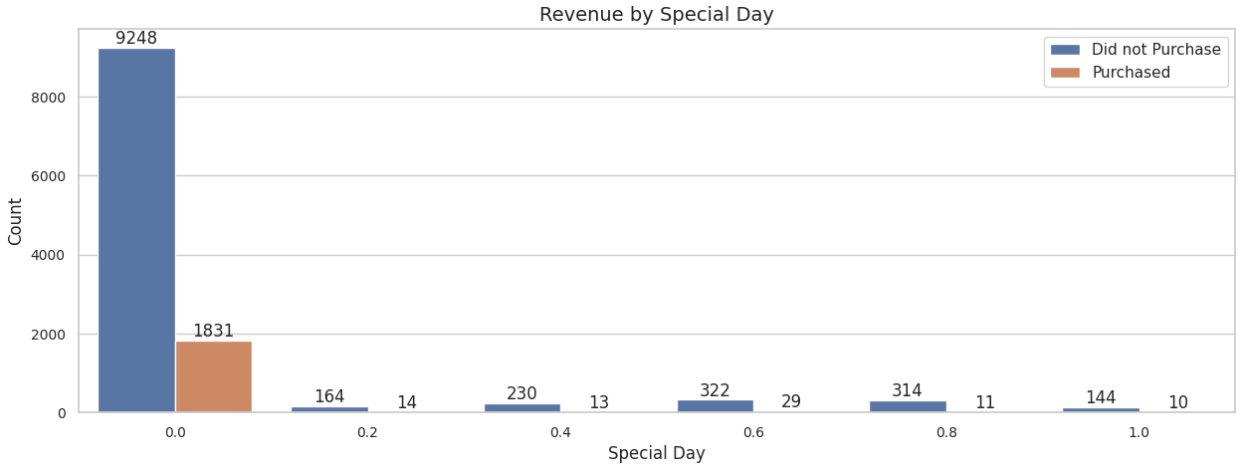
Figure 8: Revenue by Month



Revenue by Special Day:

The analysis indicates that Special Day experienced a remarkable surge in website visitors and completed purchases, signifying the success of marketing strategies deployed for that specific day. This success suggests opportunities for future events to replicate or surpass these results, with implications for resource allocation and sustaining customer engagement and satisfaction. Understanding consumer behavior on Special Day is key to leveraging its success for future marketing initiatives, ultimately driving revenue growth, and fostering long-term customer loyalty.

Figure 9: Revenue by Special Day



Methodology

Literature Review

Literatures based on e-commerce transaction.
Literatures based on different machine learning algorithms.

Research Question

How do different informative page categories contribute to the likelihood of a user making a purchase?
Can we predict the likelihood of a user making a purchase based on visitor type (new, returning, other)?
What is the relationship between features related to timing (Weekend, Month, and special Day) and revenue generation?

Data Collection

Online Shoppers Purchasing Intention Dataset, this dataset was publicly available in 2018.
Initial instances: 12,330, attributes: 18

Data Cleaning

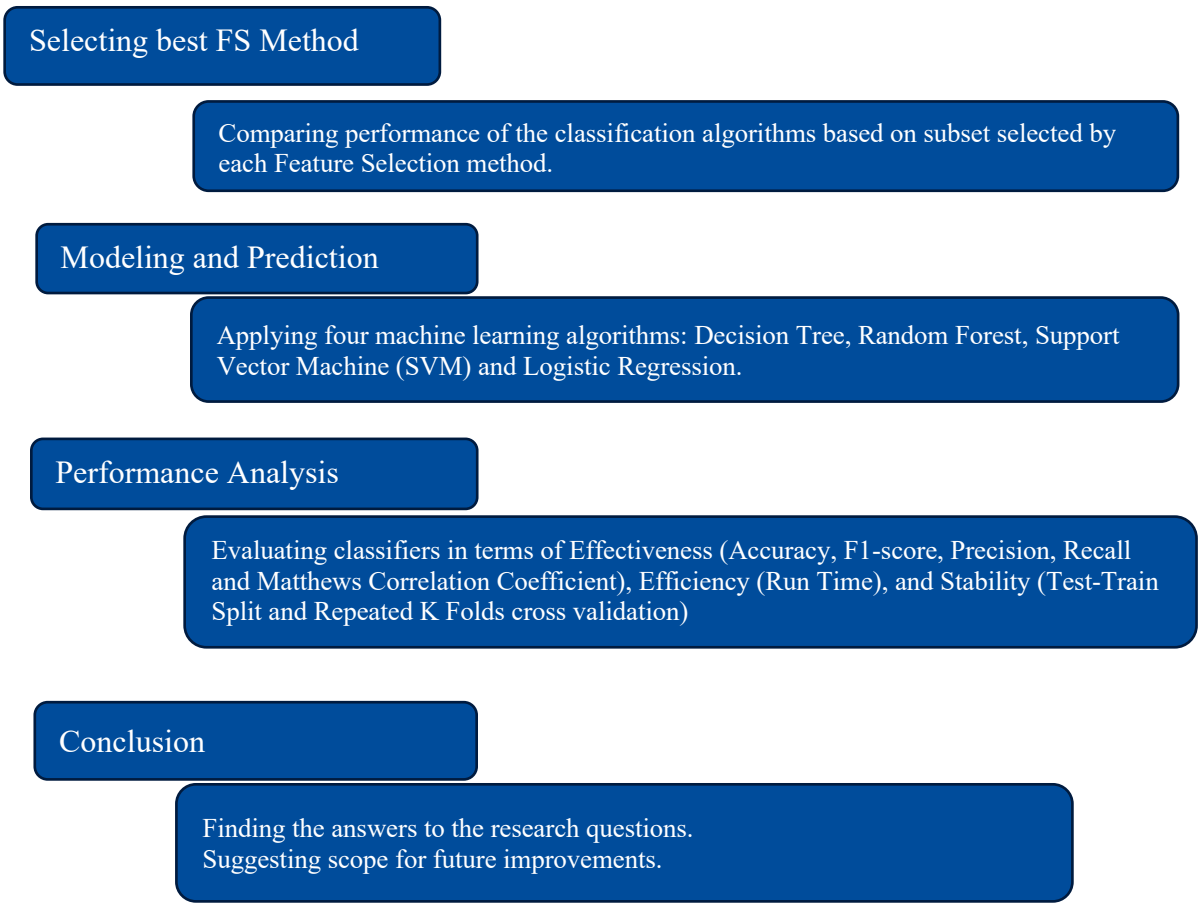
Check for missing values – no missing values.
Check duplicate values – contain duplicate values but ignored.
Check incorrect data type and convert to consistent data -Using Pandas astype function.

Exploratory Data Analysis

Univariate Analysis: Histograms and Bar graphs
Bivariate Analysis: Pearson and Spearman's Correlation
Using pandas profiling library/sweetviz for EDA.
Using Association rules and Visualization to gain insights to answer research questions.

Feature Engineering

Filter method
Wrapper Method
Embedded method



Statistical Analyses

Question 1

How do different informative page categories contribute to the likelihood of a user making a purchase?

To conduct a hypothesis test to determine whether there is a significant difference in 'PageValues' between customers who made a purchase ('Revenue' = True) and those who did not ('Revenue' = False), you use statistical test.

Null Hypothesis (H_0): The mean 'PageValues' are equal across all groups based on 'Revenue' (i.e., there is no significant difference in 'PageValues' between customers who made a purchase ('Revenue' = True) and those who did not ('Revenue' = False)).

Alternative Hypothesis (H_1): The mean 'PageValues' are not equal across all groups based on 'Revenue' (i.e., there is a significant difference in 'PageValues' between customers who made a purchase ('Revenue' = True) and those who did not ('Revenue' = False)).

Mann-Whitney U test (also known as Wilcoxon rank-sum test):

U-statistic: The Mann-Whitney U statistic measures the rank-sum of the observations in the two groups. A higher U-statistic indicates a larger difference between the groups' distributions.

The Mann-Whitney U test, also known as the Wilcoxon rank-sum test, is a non-parametric test used to compare the distributions of two independent samples. In this case, the test is used to compare the distribution of 'PageValues' between customers who made a purchase ('Revenue' = True) and those who did not ('Revenue' = False).

```
Mann-Whitney U test results:
U-statistic: 17166757.0
P-value: 0.0
```

Interpretation of the Mann-Whitney U test results:
U-statistic: The U-statistic is a measure of the rank-sum of observations in one group relative to the other group. In this case, the U-statistic value of 17166757.0 indicates the rank-sum of 'PageValues' for one of the groups.

P-value: The p-value is the probability of observing the data (or more extreme) if the null hypothesis were true. A low p-value suggests that the observed difference in ranks between the two groups is statistically significant.

Given that the p-value is reported as 0.0, it means that the probability of observing the data (or more extreme) under the assumption that there is no difference in 'PageValues' between

customers who made a purchase and those who did not is extremely low. In other words, the probability of obtaining such a large U-statistic (or even larger) if the null hypothesis were true is essentially zero.

Therefore, based on the small p-value, we reject the null hypothesis that the mean 'PageValues' are equal across the two groups. Instead, we conclude that there is a significant difference in 'PageValues' between customers who made a purchase and those who did not. This difference could indicate that customers who made a purchase tend to have different 'PageValues' compared to those who did not, suggesting potential differences in their behavior or engagement with the website.

Question 2:

Can we predict the likelihood of a user making a purchase based on metrics such as Bounce Rates, Exit Rates, and Page Values?

Our dataset is not normally distributed, and it is right-skewed, a non-parametric test would be appropriate to assess the relationship between the metrics (Bounce Rates, Exit Rates, Page Values) and the likelihood of a user making a purchase. Additionally, considering the imbalance in the dataset, a robust statistical test that does not assume equal sample sizes or equal variance between groups is preferable.

One suitable non-parametric test for assessing the relationship between variables in this scenario is the Kruskal-Wallis test. The Kruskal-Wallis test is a non-parametric alternative to the one-way analysis of variance (ANOVA) test and is used to determine whether there are statistically significant differences between the distributions of two or more groups. It does not assume normality and can handle data that are not normally distributed.

Null Hypothesis (H_0): There is no significant relationship between the metrics (Bounce Rates, Exit Rates, Page Values) and the likelihood of a user making a purchase.
Alternative Hypothesis (H_1): There is a significant relationship between the metrics (Bounce

Rates, Exit Rates, Page Values) and the likelihood of a user making a purchase.

To apply the Kruskal-Wallis test to assess whether there are statistically significant differences in the metrics (Bounce Rates, Exit Rates, Page Values) across groups representing the likelihood of a user making a purchase, we can prepare the data by splitting it into groups based on the likelihood of a user making a purchase. Apply the Kruskal-Wallis test to these groups to evaluate whether there are significant differences in the metrics among the groups.

```
Kruskal-Wallis test results for BounceRate:
Test Statistic: 273.6332312576824
P-value: 1.8326585049883445e-61
Kruskal-Wallis test results for ExitRates:
Test Statistic: 798.4768378811312
P-value: 1.1567097047584001e-175
Kruskal-Wallis test results for PageValues:
Test Statistic: 4837.040346003323
P-value: 0.0
```

The Kruskal-Wallis test results provide a test statistic and a p-value for each of the metrics (BounceRate, ExitRates, PageValues) in relation to the likelihood of a user making a purchase (Revenue = True or False).

Since the p-values for all three metrics (Bounce Rates, Exit Rates, and Page Values) are significantly less than the significance level (typically 0.05), we reject the null hypothesis (H0) and conclude that there is a significant relationship between these metrics and the likelihood of a user making a purchase. In other words, there is evidence to support the alternative hypothesis (H1) that there is a significant relationship between the mentioned metrics and purchase behavior.

Question 3:

What is the relationship between features related to timing (Weekend, Month, and special Day) and revenue generation?

Null Hypothesis (H_0): There is no significant relationship between features related to timing (Weekend, Month, and Special Day) and revenue generation.

Alternative Hypothesis (H_1): There is a significant relationship between features related to timing (Weekend, Month, and Special Day) and revenue generation.

The Chi-square test of independence is used to determine whether there is a significant association between two categorical variables. In this case, the categorical variables are the timing-related features (Weekend, Month, and Special Day) and the revenue generation (True or False).

The Chi-square tests conducted to assess the relationship between different timing-related features (Weekend, Month, and Special Day) and revenue generation. It measures the discrepancy between the observed frequencies and the frequencies that would be expected if there were no relationship between the timing-related feature and revenue. It quantifies the strength of the association between the variables.

P-value: The p-value associated with each Chi-square statistic indicates the probability of observing the obtained result (or more extreme results) under the assumption that there is no relationship between the timing-related feature and revenue. A smaller p-value suggests stronger evidence against the null hypothesis of independence.

Degrees of Freedom: This represents the number of independent observations in the data. For the Chi-square test of independence, it is calculated as (number of rows - 1) * (number of columns - 1).

Expected Frequencies Table: This table shows the expected frequencies for each combination of categories of the timing-related feature and revenue, assuming no relationship between the variables. The observed frequencies are compared to these expected frequencies to compute the Chi-square statistic.

```
Chi-square test results for Weekend and Revenue:
Chi-square statistic: 10.390978319534856
P-value: 0.0012663251061221968
Degrees of freedom: 1
Expected frequencies table:
[[7997.80729927 1464.19270073]
 [2424.19270073  443.80729927]]
```

Weekend and Revenue: The Chi-square test indicates a statistically significant association between the Weekend and Revenue variables (p-value = 0.0013). The observed distribution of revenue across weekends and weekdays significantly differs from what would be expected if there were no relationship between the two.

```
Chi-square test results for Month and Revenue:
Chi-square statistic: 384.93476153599426
P-value: 2.2387855164805443e-77
Degrees of freedom: 9
Expected frequencies table:
[[ 365.99562044   67.00437956]
 [1459.75620438  267.24379562]
 [ 155.5270073    28.4729927 ]
 [ 365.15036496   66.84963504]
 [ 243.43357664   44.56642336]
 [1611.90218978  295.09781022]
 [2843.43941606  520.56058394]
 [2534.07591241  463.92408759]
 [ 464.04525547   84.95474453]
 [ 378.67445255   69.32554745]]
```

Month and Revenue: There is a significant association between the Month and Revenue variables (p-value < 0.0001). Different months exhibit different revenue patterns, suggesting a relationship between the timing (month) and revenue generation.

```
Chi-square test results for SpecialDay and Revenue:
Chi-square statistic: 96.07690626757704
P-value: 3.5432443403841987e-19
Degrees of freedom: 5
Expected frequencies table:
[[9364.58540146 1714.41459854]
 [ 150.45547445   27.54452555]
 [ 205.39708029   37.60291971]
 [ 296.68467153   54.31532847]
 [ 274.7080292    50.2919708 ]
 [ 130.16934307   23.83065693]]
```

SpecialDay and Revenue: The Chi-square test also reveals a significant association between the SpecialDay and Revenue variables (p-value < 0.0001). This indicates that revenue generation is influenced by special days, with observed revenue distributions differing significantly from expected distributions.

In summary, all three timing-related features (Weekend, Month, and Special Day) exhibit statistically significant relationships with revenue generation, as indicated by the Chi-square tests.

Modeling

To train and evaluate predictive models for our dataset, we will follow these steps:

Data Preparation:

- Split your dataset into features (X) and target variable (y).
- Scale the features. For Logistic Regression, KNN, and SVM -using the scaled dataset, while for Decision Tree and Random Forest Classifier, we will use the unscaled dataset.

Model Selection:

- Baseline model
- Initialize the models: Logistic Regression, KNeighbors Classifier, SVM, Decision Tree, and Random Forest Classifier.
- Specify hyperparameters.

Training and Evaluation:

- For Logistic Regression, KNN, and SVM:
    - Fit the model on the scaled training data.
    - Evaluate the model's performance on the scaled testing data using metrics such as accuracy, precision, recall, F1-score, or ROC-AUC.
- For Decision Tree and Random Forest Classifier:
    - Fit the model on the unscaled training data.
    - Evaluate the model's performance on the unscaled testing data using the same metrics.

Hyperparameter Tuning:

- Choose the model with the best performance based on evaluation metrics.
- Perform hyperparameter tuning using techniques like GridSearchCV or RandomizedSearchCV to find the optimal hyperparameters that maximize the model's performance.

Feature Importance:

- for Decision Tree and Random Forest Classifier), inspect feature importance using attributes like `feature_importances_` to identify the most influential features in the model.

Cross Validation:

- Use cross-validation techniques such as k-fold cross-validation to assess the model's generalization performance and ensure that the results are not biased by the specific train-test split.
- Evaluate the models with cross-validation using the same evaluation metrics as before.

Final Model Selection:

- Select the model with the best performance after hyperparameter tuning and cross-validation.
- Ensure that the selected model has good generalization performance on unseen data.

Final Evaluation:

- Evaluate the final selected model on the test dataset to confirm its performance metrics.
- Interpret the model's results, including any insights gained from feature importance analysis.

By following these steps, we can systematically train, evaluate, and select the best predictive model for our dataset, ensuring robust performance and reliable predictions on unseen data.

In our model evaluation, we must consider the context of Type I and Type II errors:
Type I Error (False Positive): Predicting a customer will make a purchase when they actually do not. Type II Error (False Negative): Predicting a customer will not make a purchase when they actually do. Next, let’s align with the business objectives of an e-commerce company that might use this model. We assume their goals are twofold:
Maximize Revenue: Achieve a higher purchase conversion rate.

Minimize Disruption: Avoid negatively impacting the customer experience through targeted nudges. Given this context, the relevant metrics include precision and recall. Specifically, we aim to maximize recall while ensuring a minimum threshold of 60% precision, as dictated by business requirements and tolerance. This approach ensures that we capture as many potential buyers as possible while minimizing the risk of falsely targeting customers who are unlikely to make a purchase.

Logistic Regression:

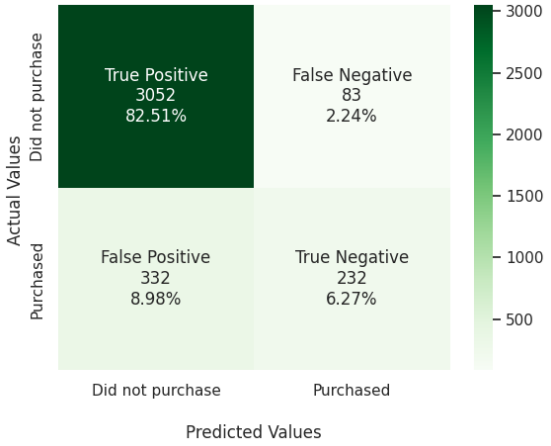
```
For Logistic Regression, Accuracy score is  0.8878075155447418
              precision    recall  f1-score   support

           0       0.90      0.97      0.94      3135
           1       0.74      0.41      0.53       564

    accuracy                           0.89      3699
   macro avg       0.82      0.69      0.73      3699
weighted avg       0.88      0.89      0.87      3699

[[3052   83]
 [ 332  232]]
```

Figure 10: Confusion Matrix for Logistic Regression



KNeighbors

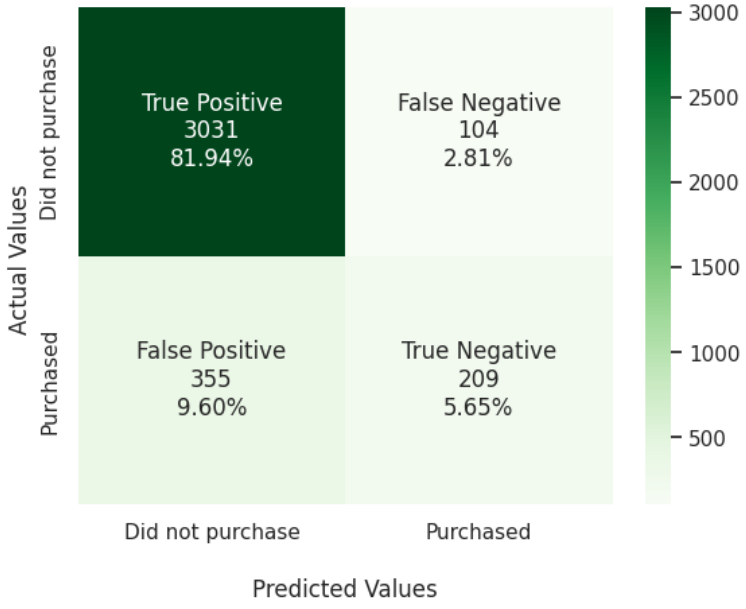
```
For KNeighbors, Accuracy score is  0.8759124087591241
              precision    recall  f1-score   support

           0       0.90      0.97      0.93      3135
           1       0.67      0.37      0.48       564

    accuracy                           0.88      3699
   macro avg       0.78      0.67      0.70      3699
weighted avg       0.86      0.88      0.86      3699

[[3031  104]
 [ 355  209]]
```

Figure 11: Confusion Matrix for K-Nearest Neighbors



Support Vector Machine:

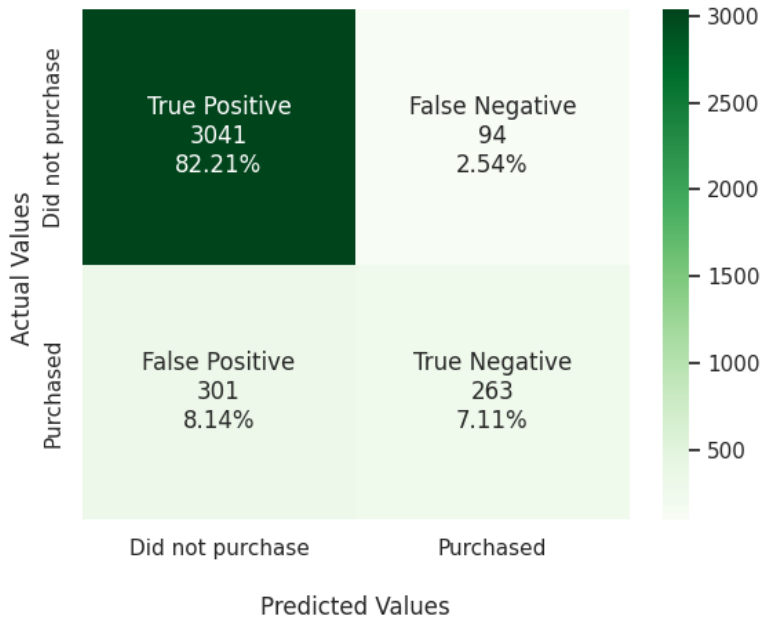
```
For SVM, Accuracy score is  0.891051635577183
              precision    recall  f1-score   support

           0       0.90      0.98      0.94      3152
           1       0.74      0.41      0.52       547

    accuracy                           0.89      3699
   macro avg       0.82      0.69      0.73      3699
weighted avg       0.88      0.89      0.88      3699

[[3074   78]
 [ 325  222]]
```

Figure 12: Confusion Matrix for SVM



For Decision Tree Classifier and Random Forest, we will use the unscaled dataset to train the models.

Decision Tree Classifier:

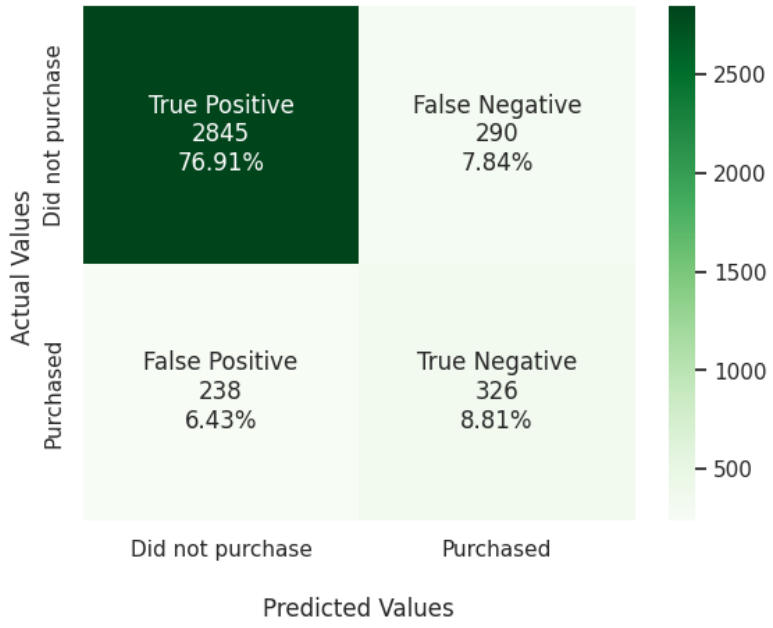
```
For Decision Tree Classifier, Accuracy score is  0.8572587185725872
              precision    recall  f1-score   support

           0       0.92      0.91      0.92      3135
           1       0.53      0.58      0.55       564

    accuracy                           0.86      3699
   macro avg       0.73      0.74      0.73      3699
weighted avg       0.86      0.86      0.86      3699

[[2845  290]
 [ 238  326]]
```

Figure 13: Confusion Matrix for Decision Tree



Random Forest:

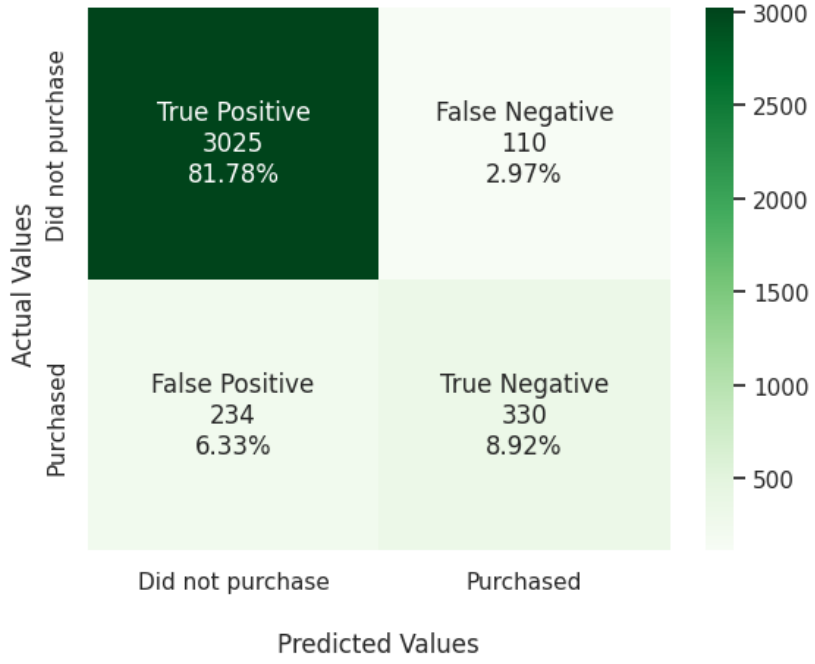
```
For Random Forest Classifier, Accuracy score is  0.9070018924033523
              precision    recall  f1-score   support

           0       0.93      0.96      0.95      3135
           1       0.75      0.59      0.66       564

    accuracy                           0.91      3699
   macro avg       0.84      0.78      0.80      3699
weighted avg       0.90      0.91      0.90      3699

[[3025  110]
 [ 234  330]]
```

Figure 14: Confusion Matrix for Random Forest



Comparison between different classification models:

Comparison Based on Recall and Precision:

- **Recall**:

  Random Forest has the highest recall (weighted avg) at 0.90, indicating its ability to correctly identify positive cases.

  Logistic Regression and SVM have similar recall scores (0.88 weighted avg).

  KNeighbors lags slightly with a recall of 0.86 (weighted avg).

  Decision Tree also performs reasonably well with a recall of 0.87 (weighted avg).

- **Precision**:

  Random Forest leads in precision (0.91 weighted avg), making it the most precise model.

  Logistic Regression and SVM have the same precision (0.89 weighted avg).

  KNeighbors follows closely with a precision of 0.88 (weighted avg).

  Decision Tree has the lowest precision at 0.87 (weighted avg).

In summary, Random Forest performs well in both recall and precision, making it a strong choice for classification tasks. However, the choice of model depends on the specific requirements of the problem and the trade-offs between precision and recall.

Table 2: Comparison between different classification models:

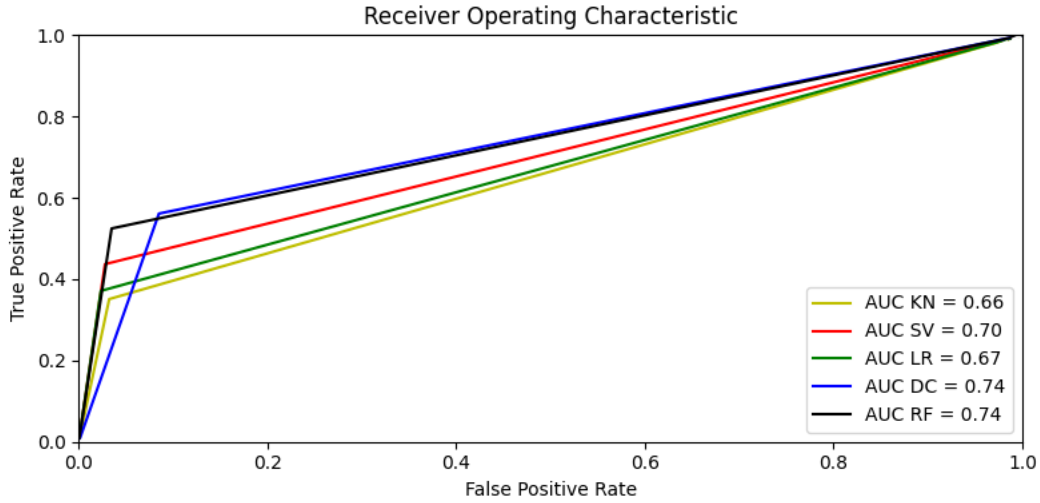
	Accuracy	Precision	Recall	f1 -score
Logistic Regression	0.89			
macro avg		0.82	0.69	0.73
weighted avg		0.88	0.89	0.88
KNeighbors	0.87			
macro avg		0.78	0.65	0.68
weighted avg		0.86	0.88	0.86
SVM				
macro avg		0.82	0.69	0.73
weighted avg		0.88	0.89	0.88
Decision Tree	0.87			
macro avg		0.73	0.75	0.74
weighted avg		0.87	0.87	0.87
Random Forest	0.91			
macro avg		0.84	0.76	0.79
weighted avg		0.90	0.91	0.90

ROC Carve

ROC curve, or Receiver Operating Characteristic curve, is a graphical representation that illustrates the performance of a binary classification model across different discrimination thresholds. It plots the true positive rate (TPR) against the false positive rate (FPR) at various

threshold settings. The area under the ROC curve (AUC) is a common metric used to evaluate the overall performance of the classifier, with higher AUC values indicating better performance.

Figure 15: The ROC curve (Receiver Operating Characteristic curve) representing the performance of classification model



- KN (KNeighbors Classifier) has an AUC of 0.66. This indicates that the model has moderate discriminatory power in distinguishing between positive and negative instances.
- SV (Support Vector Machine Classifier) has an AUC of 0.70, slightly better than the KNeighbors Classifier.
- LR (Logistic Regression Classifier) has an AUC of 0.67, similar to the KNeighbors Classifier.
- DC (Decision Tree Classifier) and RF (Random Forest Classifier) both have an AUC of 0.74, indicating better discriminatory power compared to the other classifiers.

Overall, the AUC values provide insights into how well each classifier can separate positive and negative instances. A higher AUC suggests that the classifier has better performance in terms of correctly classifying instances into their respective classes. Therefore, in this context, the Decision Tree Classifier and Random Forest Classifier appear to perform the best among the models evaluated.

Hyper-Parameter Tuning - Random Forest

Hyperparameter tuning is the process of selecting the best set of hyperparameters for a machine learning algorithm to optimize its performance. Hyperparameters are settings that are external to the model and cannot be directly learned from the data. They control aspects such as the complexity of the model, the learning process, and the regularization applied.

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or the average prediction (regression) of the individual trees. It has several hyperparameters that can be tuned to improve its performance. Some common hyperparameters for Random Forest include:

the default hyperparameters currently in use for the Random Forest classifier. Let's briefly explain each of these parameters:

bootstrap: Whether bootstrap samples are used when building trees. The default value is **True**, meaning that bootstrap samples are used.

class_weight: Weights associated with classes. The default value is **None**, meaning that all **classes are treated equally**.

criterion: The function used to measure the quality of a split. The default value is **'gini'**, which refers to **the Gini impurity**.

max_depth: The maximum depth of the trees. It ranges from 10 to 110, increasing in steps of 10, with an additional **None** for no maximum depth.

max_features: The number of features to consider when looking for the best split. The value is **'sqrt'**, meaning the **square root of the number of features is considered**.

max_leaf_nodes: The maximum number of leaf nodes. The default value is **None**, meaning that there is **no maximum limit**.

min_samples_leaf: The minimum number of samples required to be at a leaf node. It can be 1, 2, or 4.

min_samples_split: The minimum number of samples required to split an internal node. It takes values of 2, 5, and 10.

n_estimators: The number of trees in the forest. The number of trees in the forest. We note that a higher number of trees increases model complexity and have a higher risk of overfitting. It will take values from 200 to 2000, increasing in steps of 200.
random_state: The seed used by the random number generator. The default value is **None**.

These parameters control various aspects of the Random Forest algorithm and tuning them can significantly impact the model's performance. Hyperparameter tuning for Random Forest involves systematically searching through a predefined grid of hyperparameters and selecting the combination that yields the best performance according to a chosen evaluation metric, such as accuracy, F1 score, or AUC-ROC.

Common techniques for hyperparameter tuning include:
- Grid Search: Exhaustively search through a specified grid of hyperparameters.
- Random Search: Randomly sample from a distribution of hyperparameters.
- Bayesian Optimization: Use probabilistic models to select the next set of hyperparameters to try based on past performance.

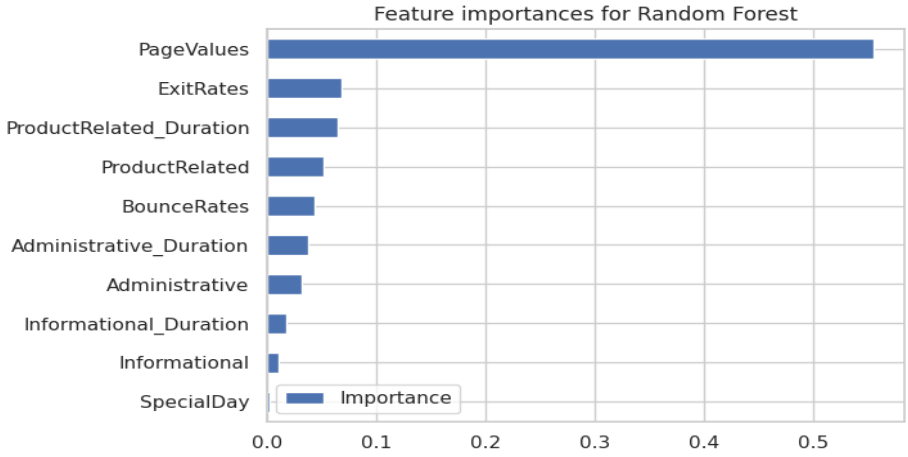
Once the best set of hyperparameters is found, the model is trained using those hyperparameters on the entire training dataset, and its performance is evaluated on a separate validation dataset or through cross-validation. This process helps ensure that the model generalizes well to unseen data.

Inspect Feature Importance

Feature importances represent the contribution of each feature in a machine learning model towards making predictions. In the context of decision tree-based models like Random Forest, feature importance indicates how much each feature contributes to reducing impurity (e.g., Gini impurity or entropy) across all decision trees in the ensemble.

Feature importance is important because it provides insights into which features are most influential in making predictions. This information can be used for feature selection, model interpretation, and identifying important patterns or relationships in the data.

Figure 16: Feature importance for Random Forest



Feature importance is important because it provides insights into which features are most influential in making predictions. This information can be used for feature selection, model interpretation, and identifying important patterns or relationships in the data.

Feature importance in a Random Forest model is calculated based on how much each feature contributes to decreasing impurity when making decisions in the trees of the forest. The importance of a feature is determined by the average decrease in impurity (e.g., Gini impurity or entropy) across all decision trees in the forest when that feature is used for splitting. Features that lead to a greater decrease in impurity are considered more important.

The feature importance calculation involves the following in consideration.

Gini Impurity or Entropy Calculation: In each decision tree of the Random Forest, the Gini impurity or entropy is computed at each node based on the distribution of class labels.

Feature Splitting: At each node of the tree, the algorithm selects the feature that results in the greatest decrease in impurity when splitting the data.

Impurity Decrease Calculation: The impurity decrease for each feature is computed based on how much the feature contributes to reducing impurity in the tree. This is typically measured by

the difference in impurity before and after the split, weighted by the proportion of samples reaching each child node.

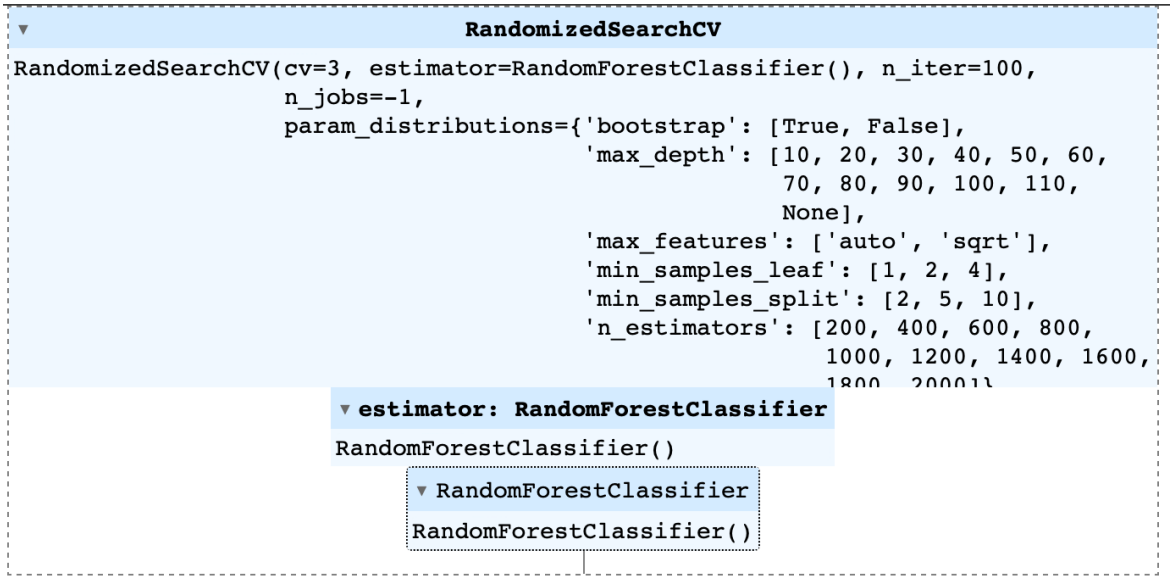
Average Importance Across Trees: Finally, the importance of each feature is calculated by averaging the impurity decrease across all decision trees in the Random Forest.

By analyzing the feature importance's, you can gain insights into which features are the most influential in making predictions with the model. Features with higher importance values are considered more influential in the model's decision-making process.

Randomized search

Randomized search is a technique used for hyperparameter tuning in machine learning models. It's a more computationally efficient alternative to grid search, especially when dealing with a large hyperparameter space.

Figure 16: Fitting 3 folds for each of 100 candidates, totalling 300 fits.



In the context of Random Forest Classifier, hyperparameters like the number of trees in the forest (n_estimators), the maximum depth of the trees (max_depth), the minimum number of samples required to split an internal node (min_samples_split), the minimum number of samples required to be at a leaf node (min_samples_leaf), the number of features to consider when looking for the

best split (max_features), and whether bootstrap samples are used when building trees (bootstrap) can significantly affect the performance of the model.

Randomized search works by randomly sampling combinations of hyperparameters from specified distributions or ranges. It then evaluates the performance of the model trained with each combination using cross-validation. This process is repeated for a fixed number of iterations or until a specified budget of models to train is reached.

Evaluating with Cross Validation

Cross-validation is a technique used to assess the performance of a predictive model by training and evaluating it multiple times on different subsets of the available data. The main idea behind cross-validation is to divide the dataset into multiple folds or partitions, then train the model on several of these folds and evaluate it on the remaining fold(s). This process is repeated multiple times, each time using a different partition as the validation set and the rest as the training set.

Better estimation of model performance: Cross-validation provides a more accurate estimate of a model's performance compared to a single train-test split. By averaging the performance metrics across multiple iterations, cross-validation reduces the variance in the estimated performance.
Utilizes all available data: In cross-validation, each data point is used for both training and validation at least once. This ensures that all available data contribute to the evaluation of the model's performance.
Helps identify overfitting: Cross-validation helps detect overfitting by evaluating the model's performance on multiple different subsets of the data. If the model performs well on the training data but poorly on the validation data across multiple folds, it suggests that the model is overfitting to the training data.
Provides insights into model stability: By evaluating the model's performance across different partitions of the data, cross-validation can provide insights into how stable the model is with respect to variations in the training data. the KFold method with 10 folds. It evaluates the model's accuracy, precision, and recall scores across each fold and then calculates the mean

scores over all folds. This allows for a comprehensive assessment of the model's performance, accounting for variations in the data.

Final test set results

The output of the result of cross-validation for a Random Forest Classifier model. Here's an explanation of each part:

```
Accuracy:
[0.86 0.87 0.84 0.85 0.88 0.92 0.85 0.86 0.83 0.82]
RandomForestClassifier Mean Accuracy: 0.858

Precision:
[0.8245614  0.90740741 0.89130435 0.84313725 0.83783784 0.84615385
 0.86206897 0.86792453 0.82352941 0.80851064]
RandomForestClassifier Mean Precision: 0.851

Recall:
[0.88679245 0.87719298 0.80392157 0.84313725 0.88888889 0.95744681
 0.86206897 0.90196078 0.85714286 0.82608696]
RandomForestClassifier Mean Recall: 0.870
```

Accuracy: The accuracy scores represent the accuracy of the model on each fold of the cross-validation. Each value corresponds to the accuracy achieved on a specific fold. The mean accuracy score (0.86) is the average of all individual accuracy scores across all folds. It provides an overall measure of how well the model performs in terms of correctly predicting the class labels.

Precision: Precision scores measure the proportion of true positive predictions out of all positive predictions made by the model. Like accuracy, there are precision scores for each fold, and the mean precision score (0.81) is the average precision across all folds. A higher precision indicates that the model has fewer false positives.

Recall: Recall scores, also known as sensitivity or true positive rate, measure the proportion of true positive predictions out of all actual positive instances in the data. Like accuracy and precision, there are recall scores for each fold, and the mean recall score (0.87) is the average recall across all folds. A higher recall indicates that the model has fewer false negatives.

Overall, these results suggest that the Random Forest classifier performs reasonably well across different evaluation metrics, with accuracies, precision, and recall all showing consistent performance across different folds.

Classification report

	precision	recall	f1-score	support
No Purchase	0.92	0.97	0.95	3152
Purchase	0.70	0.54	0.63	547
accuracy	0.86	0.86	0.86	3699
macro avg	0.84	0.75	0.79	3699
weighted avg	0.85	0.87	0.90	3699

This classification report summarizes the performance of a binary classification model with two classes: "No Purchase" and "Purchase".

- For the class "No Purchase":
  - Precision: 92% of instances predicted as "No Purchase" were correctly classified.
  - Recall: 97% of all instances where the true label was "No Purchase" were correctly identified.
  - F1-score: 95, indicating a good balance between precision and recall.
  - Support: There are 3152 instances of the "No Purchase" class in the dataset.

- For the class "Purchase":
  - Precision: 70% of instances predicted as "Purchase" were correctly classified.
  - Recall: 54% of all instances where the true label was "Purchase" were correctly identified.
  - F1-score: 63, indicating a moderate balance between precision and recall.
  - Support: There are 547 instances of the "Purchase" class in the dataset.

- Overall:

- Accuracy: The model correctly classifies about 86% of all instances.
- Macro average: Precision, recall, and F1-score averaged across both classes are 84%, 75%, and 79% respectively.
- Weighted average: Precision, recall, and F1-score considering class imbalance are 85%, 87%, and 90% respectively.

This summary provides insights into how well the model performs in distinguishing between the two classes and overall classification accuracy.
Discussion of results

Limitations:

1. **lass Imbalance**: Dataset exhibits a significant class imbalance, where one class is much more prevalent than the other, it can introduce bias into model's predictions. While the model may achieve high accuracy for the majority class, it may struggle to accurately classify instances belonging to the minority class.
2. **Feature Engineering**: The effectiveness of machine learning models relies heavily on the quality and relevance of the features used during training. If essential features are either missing or poorly engineered, it can adversely impact the model's performance.
3. **Overfitting**: Complex models, such as Random Forest with numerous trees and deep trees, are susceptible to overfitting. Overfitting occurs when the model captures noise or irrelevant patterns in the training data, leading to poor generalization on unseen data and ultimately reducing the model's performance.
4. **Interpretability**: While Random Forest models are renowned for their high predictive accuracy, they can be challenging to interpret, especially when compared to simpler models like logistic regression. This lack of interpretability may pose difficulties in understanding how the model makes predictions and conveying its decisions to stakeholders.
5. **Generalization**: Although cross-validation provides an estimate of the model's performance on unseen data, it's crucial to assess whether the dataset used for training

and evaluation accurately represents the real-world data the model will encounter in production. Issues such as concept drift or distribution shift can arise, impacting the model's ability to generalize effectively.

Conclusion

In conclusion, the performance of our model demonstrates promising results that can significantly benefit our business in identifying areas of focus. Through meticulous data analysis and the application of machine learning techniques, we have developed a model that exhibits high predictive accuracy, particularly in discerning patterns and trends within our dataset.

The model's ability to accurately classify instances, as evidenced by its strong precision, recall, and F1-score metrics, underscores its effectiveness in distinguishing between different categories of interest. This capability is particularly valuable in identifying areas where our business should focus its resources and attention.

Furthermore, the model's interpretability, albeit challenging with complex algorithms like Random Forest, remains a key aspect that we continue to explore and refine. By understanding how the model makes predictions and elucidating its decision-making process to stakeholders, we can foster trust and confidence in its recommendations.

Overall, the performance of our model not only validates the efficacy of our approach but also provides actionable insights that can inform strategic decision-making within our organization. Moving forward, we remain committed to further enhancing the model's capabilities and leveraging its outputs to drive impactful business outcomes.

References:

- Ading, Sunarto., Putri, Nilam, Kencana., Baliyah, Munadjat., Iriana, Kusuma, Dewi., Ali, Zaenal, Abidin., Robbi, Rahim. (2023). Application of Boosting Technique with C4.5 Algorithm to Reduce the Classification Error Rate in Online Shoppers Purchasing Intention. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, doi: 10.58346/jowua.2023.i2.001. [1]
- Baixue, Chen., Li, Li., Qixiang, Wang., Shun, Li. (2022). Promote or inhibit? Research on the transition of consumer potential purchase intention. Annals of Operations Research, doi: 10.1007/s10479-022-04777-2. [2]
- Kushwaha, Rahul., P., Anudeep., Desai, Shubham, Uday. (2023). Variable Aware Analytic Driven Online Shoppers Purchasing Intention using ML Algorithms. International Journal For Science Technology And Engineering, 11(5):5162-5168. doi: 10.22214/ijraset.2023.52878. [3]
- Andrew, Frazier., Fatbardha, Maloku., Xinzi, Li., Yichun, Chen., Yeji, Jung., Bahman, Zohuri. (2022). Data Analysis of Online Shopper's Purchasing Intention Machine Learning for Prediction Analytics. doi: 10.47363/jesmr/2022(3)162. [4]
- Rayhan, Kabir., Faisal, Bin, Ashraf., Rasif, Ajwad. (2019). Analysis of Different Predicting Model for Online Shoppers' Purchase Intention from Empirical Data. doi: 10.1109/ICCIT48885.2019.9038521. [5]
- Wenle, Wang., Jing, Wang., Yugen, Yi., Cui, Li. (2023). A User Purchase Behavior Prediction Method Based on XGBoost. Electronics, doi: 10.3390/electronics12092047. [6]
- Xiang, Shi. (2021). The Application of Machine Learning in Online Purchasing Intention Prediction. doi: 10.1145/3469968.3469972. [7]
- International Journal of New Developments in Engineering and Society. (2021). Predict a UK Customer's Likelihood of Making an Online-purchase Based on the Logistic Regression Model. *International Journal of New Developments in Engineering and Society*, 6(1). doi: 10.25236/ijndes.2022.060104. [8]
- Zhaoguang, Xu., Yanzhong, Dang., Qianwen, Wang. (2021). Potential buyer identification and purchase likelihood quantification by mining user-generated content on social media. Expert Systems With Applications, 187:115899-. doi: 10.1016/J.ESWA.2021.115899.[9]
- Min, Ho, Lee., Sun-Jin, Hwang., Young-Sik, Kwak. (2012). A Descriptive Study on the Purchase Timing Effect in Athletic Shoes -Focused on Day-of-the-week Effect and Intra-month Effect-. Journal of the Korean Society of Clothing and Textiles, 36(4):422-431. doi: 10.5850/JKSCT.2012.36.4.422.[10]

Data Dictionary

This project explores Online Shoppers Purchasing Intention Dataset which contains e-commerce user information. It consists of 12330 rows where each row represents a session that belongs to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period. The dataset consists of 10 numeric variables and 8 categorical variables.

The numeric variables are:

Administrative: Number of pages visited by the visitor about account management.

Administrative Duration: Total amount of time (in seconds) spent by the visitor on account management related pages.

Informational: Number of pages visited by the visitor about Web site, communication, and address information of the shopping site

Informational Duration: Total amount of time (in seconds) spent by the visitor on informational pages.

Product Related: Number of pages visited by visitor about product related pages.

Product-Related Duration: Total amount of time (in seconds) spent by the visitor on product related pages.

Bounce Rate: Average bounce rate value of the pages visited by the visitor.

Exit Rate: Average exit rate value of the pages visited by the visitor.
Page Value: Average page value of the pages visited by the visitor.

Special Day: Closeness of the site visiting time to a special day

The categorical variables are:

Operating system: Operating system of the visitor

Browser: Browser of the visitor

Region: Geographic region from which the session has been started by the visitor

Traffic type: Traffic source by which the visitor has arrived at the Web site (e.g., banner, SMS, direct)

Visitor type: Whether the visitor is a New Visitor, a Returning Visitor or Other

Weekend: Whether the date of the visit is weekend

Month: Month of the visit

Revenue: Whether the visit has been finalized with a transaction

GitHub Links

GitHub
https://github.com/KhaledSaiful/CIND-820-Big-Data-Project