Enhancing Customer Subscription Rates

Milestone: Project Report

Group 2

Ananya Mahesh Shetty

Dhiksha Mathanagopal

857-405-8657 (Tel of Student 1)

737-342-9340 (Tel of Student 2)

shetty.ana@northeastern.edu

mathanagopal.d@northeasten.edu

Percentage of Effort Contributed by Student 1: 50%

Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Ananya Mahesh Shetty

Signature of Student 2: Dhiksha Mathanagopal

Submission Date: 4/24/2024

Table of contents

INTRODUCTION

In today's competitive business environment, harnessing the power of data-driven marketing strategies is crucial for attracting and retaining customers. In recent years, subscription-based models have become increasingly prevalent across various sectors including media, software, retail, and services. This report explores the application of predictive analytics to enhance the understanding and effectiveness of subscription models, focusing on predicting customer behaviors and subscription likelihood. As part of our commitment to innovation, we have embarked on a project titled "Enhancing Customer Subscription Rates Through Predictive Modeling". This initiative focuses on leveraging user behavior and demographic data to not only understand but also predict customer subscription tendencies, thereby optimizing our marketing efforts and improving subscriptions.

1.1 Project Background

The advent of digital platforms has provided us with unprecedented access to vast amounts of consumer data. This data spans a wide spectrum, encompassing user interactions, demographic details, and historical purchasing behaviors. By meticulously analyzing this data, we can uncover patterns that inform strategic decisions, particularly in our marketing approaches. Our project seeks to harness this potential by developing a predictive model that can accurately identify potential subscribers from our existing pool of users.

1.2Objectives and Scope

The primary objective of our project is to develop a predictive model that accurately classifies potential subscribers, enabling us to craft targeted marketing campaigns designed to boost subscription rates effectively. This endeavor aims to enhance our revenue streams and increase customer engagement through tailored interactions. The scope of our project includes:

- **Data Collection:** Assembling relevant data that encompasses user behavior and demographic information.
- **Data Preprocessing:** Cleansing and preparing the data to ensure accuracy and usability in our analysis.
- **Model Development:** Selecting and applying the most suitable predictive modeling techniques to forecast subscription likelihood.

- **Evaluation:** Rigorously assessing the model's effectiveness and fine-tuning it based on performance metrics.
- **Deployment:** Integrating the predictive model into our existing marketing strategies to refine and enhance decision-making processes.

Through these activities, we aim to provide actionable insights that enable precise targeting of potential subscribers. By doing so, we anticipate not only the improvement of our marketing investment efficiency but also the fostering of sustainable business growth through enhanced customer relationships.

Problem Setting

2.1 Domain and Application Context

The retail domain, particularly online and direct-to-consumer segments, is rapidly evolving, driven by changes in consumer behavior and advancements in technology. In this context, the focus of our project lies in understanding shopping trends within a dataset that reflects a diverse customer base's interaction with various product categories. These insights are crucial for any retail business aiming to enhance its customer subscription rates, an important metric for customer loyalty and recurring business. Subscriptions in retail often entail customers agreeing to receive products or services regularly in exchange for potentially lower prices or added convenience. For businesses, successful subscription models can ensure steady cash flow, improved inventory management, and deeper customer relationships. This project aims to leverage data analytics to uncover patterns that could be used to predict and increase subscription rates, thereby transforming one-time buyers into loyal subscribers. The business landscape has seen a significant shift towards subscription-based models across various sectors such as media, software, retail, and services. This model offers a steady revenue stream and deepens customer relationships but also demands continuous engagement and enhancement to retain subscribers. Unlike traditional one-time sales models, the success of a subscription service heavily depends on the ability to maintain and grow its subscriber base.

2.2 Challenges in Analyzing the Domain

Data Complexity:

The retail dataset encompasses various attributes ranging from demographic details to purchasing behavior, introducing complexity in the analysis. Each attribute such as age, gender, location, and purchase frequency could differently influence subscription tendencies, requiring nuanced analysis to understand their interdependencies and individual impacts.

Variability of Consumer Behavior:

Consumer purchasing behavior can be highly unpredictable and influenced by numerous external factors such as economic conditions, seasonal trends, and marketing campaigns. This variability makes it challenging to isolate factors that consistently predict subscription uptake.

Integration of Diverse Data Sources:

Often, the data needed to comprehensively understand customer behaviors comes from various sources -transaction records, customer feedback, online browsing patterns, and more. Integrating these data sources to provide a unified view of the customer can be technically challenging and time-consuming.

Scalability of Analysis:

The methods and models developed must not only be effective on the current dataset but also scalable and adaptable to larger datasets as the business grows. This scalability is crucial to maintain analytical accuracy and relevancy over time.

Privacy and Ethical Concerns:

Analyzing customer data must be done with strict adherence to data privacy laws and ethical standards. Ensuring the anonymity and security of customer information is paramount, which imposes constraints on data handling and analysis techniques.

Problem Definition

Within the broader scope of enhancing subscription-based services in the retail domain, this project aims to address several pivotal issues associated with customer retention and conversion. The complex dynamics of consumer behavior, combined with the variability of market conditions, present a challenging environment for maintaining and expanding a subscriber base.

3.1 Specific Issues Being Addressed

Customer Conversion: Understanding the barriers that prevent one-time customers from becoming subscribers is crucial. The project seeks to identify and mitigate these barriers by analyzing customer interactions and purchasing behaviors.

Retention of Subscribers: Determining the key factors that influence subscriber retention. This includes analyzing how various aspects of the service such as product offerings, pricing strategies, and customer engagement impact loyalty.

Effectiveness of Promotions: Evaluating the impact of promotional strategies and discounts on subscription rates. The goal is to determine which types of promotions are most effective at converting and retaining subscribers.

Customization of Marketing Efforts: Tailoring marketing efforts based on customer data to increase effectiveness. This involves segmenting the customer base and personalizing marketing to better meet the needs and preferences of different groups.

Product and Service Adjustments: Using insights from data to adjust the product or service offerings to better align with customer preferences, thereby increasing the likelihood of subscriptions.

3.2 Research Questions

To navigate and address these issues, the project will seek answers to several research questions:

What demographic and behavioral characteristics most significantly predict a customer's likelihood to subscribe?

This includes exploring how age, gender, location, and shopping behaviors such as frequency and amount of purchases influence subscription decisions.

Which types of products or services are more likely to lead to subscriptions?

Identifying product categories and specific items that have higher conversion rates to subscriptions can inform stock and marketing strategies.

What are the effects of different promotional strategies on subscription rates?

Analyzing how discounts, promotional codes, and different types of marketing communications impact customer decisions to subscribe.

How do seasonal trends affect subscription rates, and how can these trends be leveraged to enhance subscriptions?

Understanding the temporal dynamics of subscriptions to tailor marketing and promotional strategies according to seasonal variations.

What role does customer satisfaction, as reflected in product reviews and ratings, play in the likelihood of subscription?

Examining the correlation between customer satisfaction metrics and subscription rates to identify areas for product or service improvement.

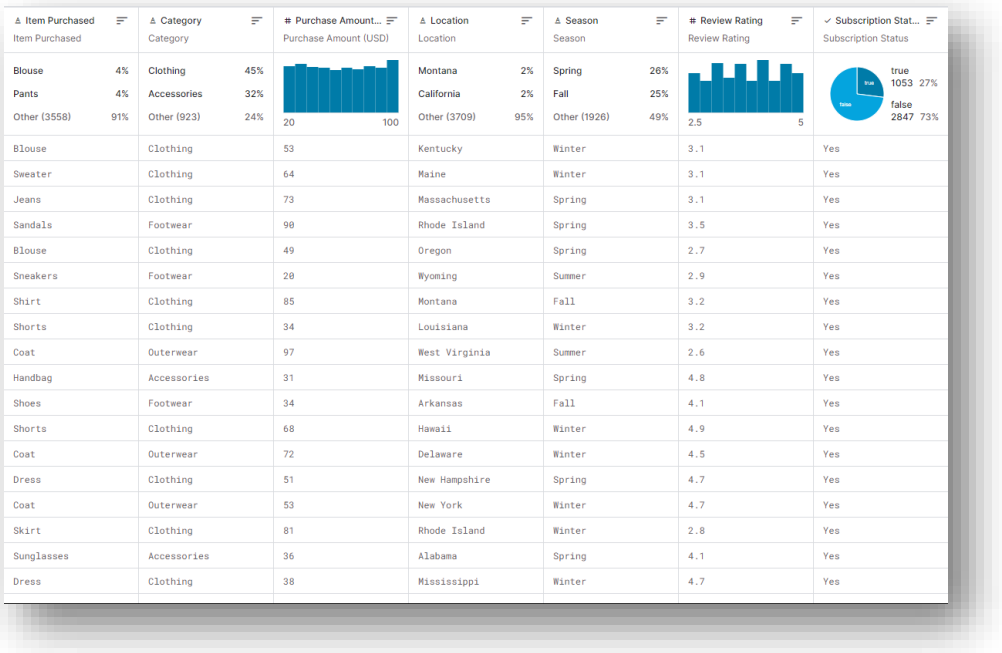
These research questions aim to dissect the complexities of consumer engagement within a subscription model, providing a robust framework for deploying data-driven strategies to enhance customer subscription rates effectively. By addressing these specific issues and questions, the project endeavors to build a predictive model that not only forecasts subscription likelihood but also aids in crafting strategies that optimize customer acquisition and retention.

Data Sources

4.1 Description of Data Origin

The dataset utilized for this project is sourced from Kaggle, a platform that hosts various datasets for data science projects and competitions. The dataset, titled "Customer Shopping Trends Dataset," encompasses a comprehensive range of variables that detail customer shopping behaviors and preferences. This dataset is pivotal for businesses aiming to deepen their understanding of consumer behavior, especially in the context of enhancing subscription rates.

Citation:



Item Purchased	Category	Purchase Amount (USD)	Location	Season	Review Rating	Subscription Status
Blouse 4%	Clothing 45%	20 – 100	Montana 2%	Spring 26%	2.5 – 5	true 1053 27%
Pants 4%	Accessories 32%		California 2%	Fall 25%		false 2847 73%
Other (3558) 91%	Other (923) 24%		Other (3709) 95%	Other (1926) 49%		
Blouse	Clothing	53	Kentucky	Winter	3.1	Yes
Sweater	Clothing	64	Maine	Winter	3.1	Yes
Jeans	Clothing	73	Massachusetts	Spring	3.1	Yes
Sandals	Footwear	90	Rhode Island	Spring	3.5	Yes
Blouse	Clothing	49	Oregon	Spring	2.7	Yes
Sneakers	Footwear	20	Wyoming	Summer	2.9	Yes
Shirt	Clothing	85	Montana	Fall	3.2	Yes
Shorts	Clothing	34	Louisiana	Winter	3.2	Yes
Coat	Outerwear	97	West Virginia	Summer	2.6	Yes
Handbag	Accessories	31	Missouri	Spring	4.8	Yes
Shoes	Footwear	34	Arkansas	Fall	4.1	Yes
Shorts	Clothing	68	Hawaii	Winter	4.9	Yes
Coat	Outerwear	72	Delaware	Winter	4.5	Yes
Dress	Clothing	51	New Hampshire	Spring	4.7	Yes
Coat	Outerwear	53	New York	Winter	4.7	Yes
Skirt	Clothing	81	Rhode Island	Winter	2.8	Yes
Sunglasses	Accessories	36	Alabama	Spring	4.1	Yes
Dress	Clothing	38	Mississippi	Winter	4.7	Yes

- https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset

Details of the Dataset

The dataset consists of 3,900 records, each representing individual customer interactions, with 18 variables that provide insights into various aspects of the customer's purchasing patterns and preferences.

These variables include:

- **Demographics**: Age, Gender
- **Purchasing Data**: Item types, Purchase amounts, Shopping frequency

- **Customer Engagement**: Payment methods, Seasonal preferences
- **Marketing Responses**: Promotional responses, Discount applications
- **Feedback**: Review ratings

Each of these variables offers a unique perspective on customer habits and preferences, making the dataset an invaluable resource for analyzing factors that influence subscription decisions. The data is meticulously organized, with no missing values, allowing for seamless exploratory data analysis and model training. The variables are both categorical and numerical, providing a balanced mix for various statistical and machine learning approaches.

Application of the Data

This dataset is particularly suitable for the project's aim to develop predictive models that can identify potential subscribers and understand the dynamics influencing customer decisions to subscribe. The comprehensive nature of the dataset supports a deep dive into consumer behavior, enabling the project team to tailor and optimize marketing strategies specifically designed to enhance subscription rates effectively. By analyzing this data, we have derived actionable insights into the customer journey, from initial contact through various touchpoints to the decision to subscribe, which is crucial for designing effective customer retention and acquisition strategies.

Data Description

5.1 Dataset Overview

Our project utilizes a dataset from the Kaggle platform titled "Customer Shopping Trends Dataset". This dataset comprises 3,900 individual records, each representing unique customer interactions with retail services. The richness and diversity of this dataset are instrumental for analyzing customer behavior patterns, which is crucial for predicting and enhancing subscription rates.

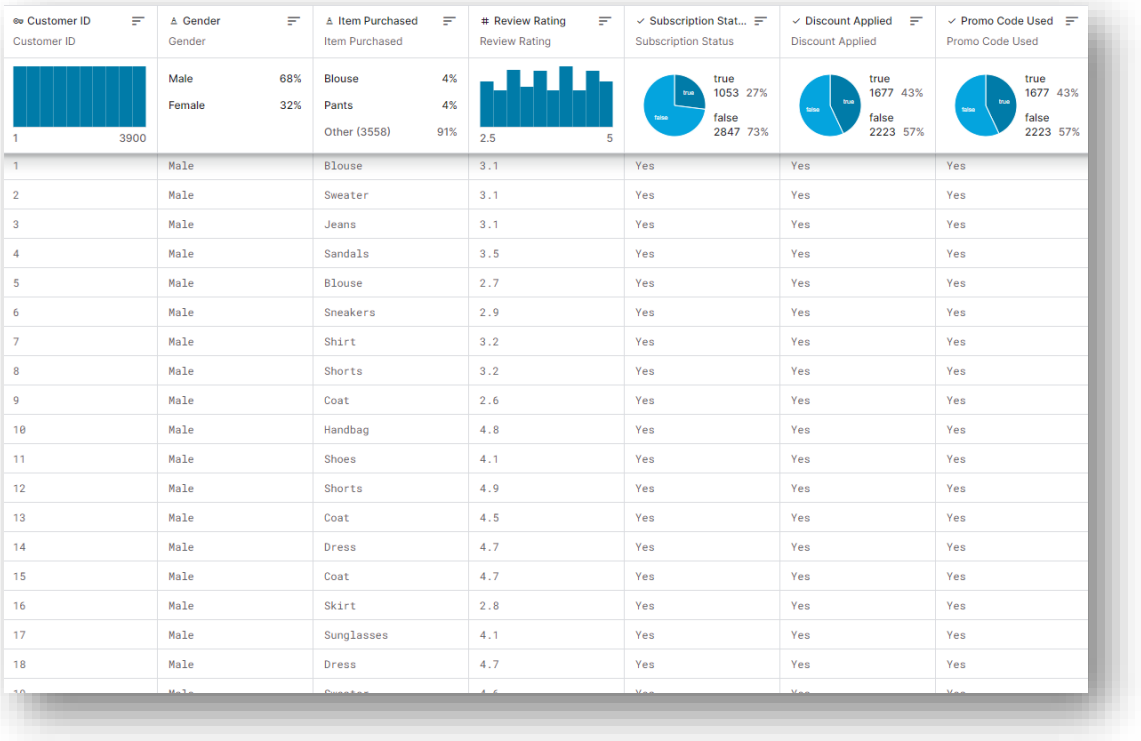
5.2 Attributes and Types of Data

The dataset includes 18 variables that encompass a wide range of information relating to customer demographics, purchasing behaviors, and responses to marketing strategies. Here's a breakdown of these attributes:

- **Customer ID**: Numeric identifier unique to each customer.
- **Age**: Numeric value representing the customer's age.
- **Gender**: Categorical attribute indicating the customer's gender (Male, Female, Other).
- **Item Purchased**: Categorical descriptor of the purchased item.
- **Category**: The category to which the purchased item belongs (e.g., Clothing, Footwear, Accessories).
- **Purchase Amount (USD)**: Numeric value of the amount spent in USD.
- **Location**: Categorical descriptor of the customer's geographical location.
- **Size**: The size of the purchased item (e.g., S, M, L).
- **Color**: The color of the item purchased.
- **Season**: The season during which the purchase was made (Spring, Summer, Fall, Winter).
- **Review Rating**: Numeric rating provided by the customer.
- **Subscription Status**: Categorical status indicating whether the customer is subscribed (Yes, No).
- **Shipping Type**: Type of shipping selected (Express, Next Day Air, Free Shipping).
- **Discount Applied**: Indicates if a discount was applied to the purchase (Yes, No).
- **Promo Code Used**: Indicates if a promo code was used during the purchase (Yes, No).
- **Previous Purchases**: Numeric count of how many times the customer has purchased before.
- **Payment Method**: Categorical descriptor of the payment method used (e.g., Credit Card, PayPal, Venmo).
- **Frequency of Purchases**: Categorical descriptor of how frequently the customer makes purchases (e.g., Weekly, Monthly, Annually).

5.3 Sample Data Points

Below is a sample from our dataset that showcases a variety of data points captured for each customer:



Customer ID Customer ID	Gender Gender	Item Purchased Item Purchased	Review Rating Review Rating	Subscription Stat... Subscription Status	Discount Applied Discount Applied	Promo Code Used Promo Code Used
1 3900	Male 68% Female 32%	Blouse 4% Pants 4% Other (3558) 91%	2.5 5	true 1053 27% false 2847 73%	true 1677 43% false 2223 57%	true 1677 43% false 2223 57%
1	Male	Blouse	3.1	Yes	Yes	Yes
2	Male	Sweater	3.1	Yes	Yes	Yes
3	Male	Jeans	3.1	Yes	Yes	Yes
4	Male	Sandals	3.5	Yes	Yes	Yes
5	Male	Blouse	2.7	Yes	Yes	Yes
6	Male	Sneakers	2.9	Yes	Yes	Yes
7	Male	Shirt	3.2	Yes	Yes	Yes
8	Male	Shorts	3.2	Yes	Yes	Yes
9	Male	Coat	2.6	Yes	Yes	Yes
10	Male	Handbag	4.8	Yes	Yes	Yes
11	Male	Shoes	4.1	Yes	Yes	Yes
12	Male	Shorts	4.9	Yes	Yes	Yes
13	Male	Coat	4.5	Yes	Yes	Yes
14	Male	Dress	4.7	Yes	Yes	Yes
15	Male	Coat	4.7	Yes	Yes	Yes
16	Male	Skirt	2.8	Yes	Yes	Yes
17	Male	Sunglasses	4.1	Yes	Yes	Yes
18	Male	Dress	4.7	Yes	Yes	Yes

Data Exploration

For our project on enhancing subscription rates using the "Shopping Trends" dataset, we conduct a thorough data exploration to better understand the characteristics and distributions of the data. This exploration helps us identify trends, anomalies, patterns, and relationships within the dataset.

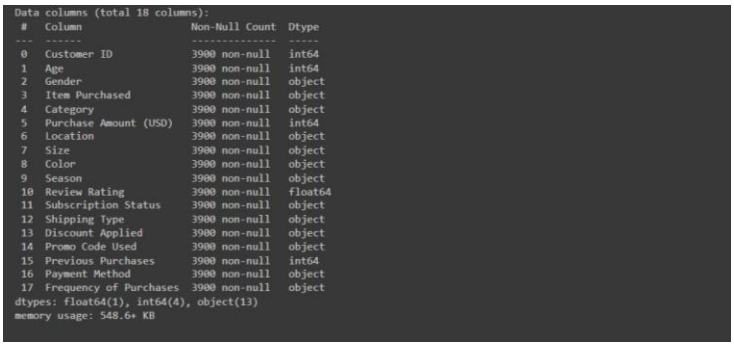
6.1 Statistical Methods Used

We deploy a variety of statistical methods to analyze the dataset effectively:

1. **Descriptive Statistics**: Using *data.describe()*, we summarize the central tendency, dispersion, and shape of the dataset's numerical attributes. This includes calculations of mean, median, standard deviation, min, and max values, providing a foundational understanding of the data.

	Customer ID	Age	Purchase Amount (USD)	Review Rating	Previous Purchases
count	3900.000000	3900.000000	3900.000000	3900.000000	3900.000000
mean	1950.500000	44.068462	59.764359	3.749949	25.351538
std	1125.977353	15.207589	23.685392	0.716223	14.447125
min	1.000000	18.000000	20.000000	2.500000	1.000000
25%	975.750000	31.000000	39.000000	3.100000	13.000000
50%	1950.500000	44.000000	60.000000	3.700000	25.000000
75%	2925.250000	57.000000	81.000000	4.400000	38.000000
max	3900.000000	70.000000	100.000000	5.000000	50.000000

2. **Data Integrity and Completeness**: By running *data.info()*, we assess the data types and the presence of null values across all variables, ensuring that the dataset is complete and appropriately formatted for analysis. This step confirms that all 3,900 entries across 18 variables are non-null and correctly typed, which is crucial for accurate further analysis.



3. **Univariate Analysis**: For each variable, we evaluate the distribution and variability. This involves looking at frequency distributions for categorical data and summary statistics for numerical data.
4. **Bivariate Analysis**: We explore relationships between different variables, especially focusing on how independent variables like age, gender, and purchase behaviors relate to the dependent variable 'Subscription Status'. Techniques such as correlation matrices for continuous variables and cross-tabulations for categorical variables are used.

6.2 Visualization Techniques Employed

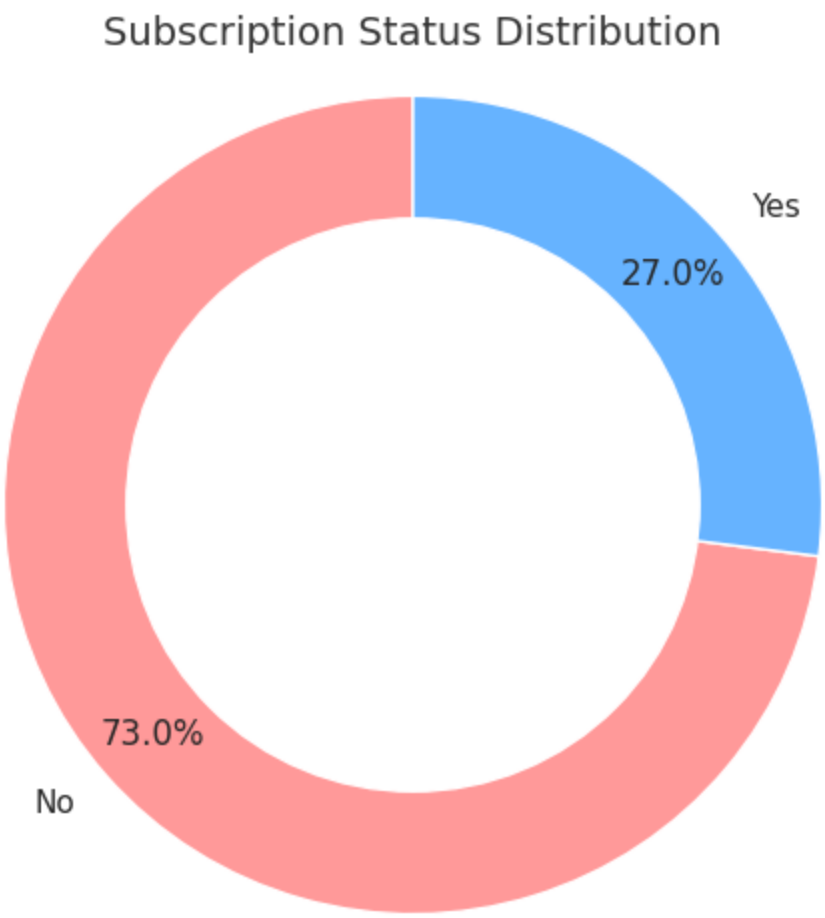
In our project, we employ a range of visualization techniques to deeply understand and display the data's structure, trends, and relationships. These visualizations not only help in presenting the data in an insightful way but also in uncovering hidden patterns that could inform our predictive models and business strategies.

Pie Charts

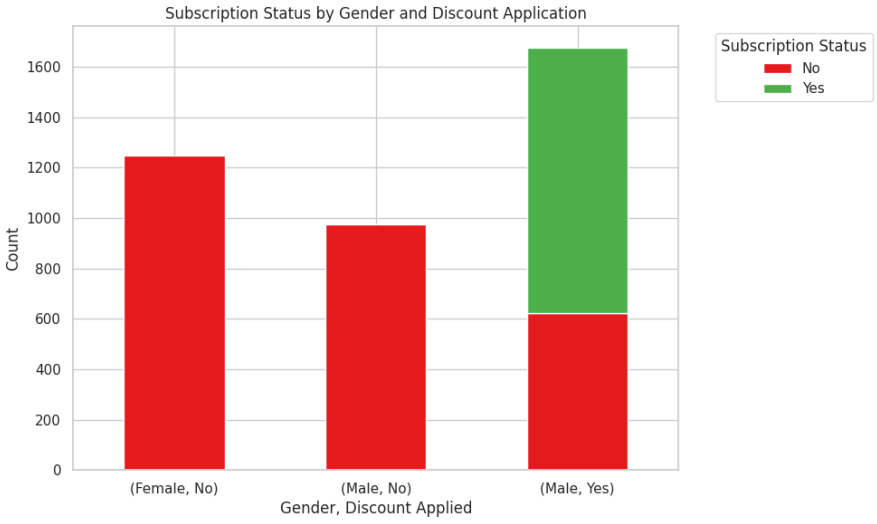
- **Subscription Status Distribution**: We use a pie chart to represent the proportion of different subscription statuses within our dataset. This visualization helps quickly grasp the percentage of customers who subscribed versus those who have not, providing a clear view of our target variable's distribution.

Bar Charts

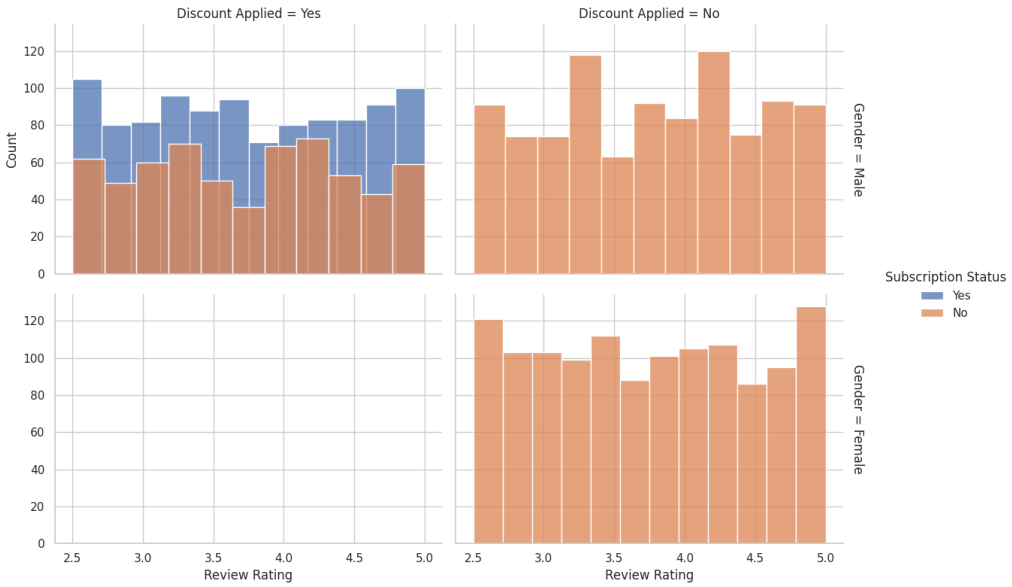
- **Gender Distribution**: We create bar charts to display the count of each gender within the dataset, particularly focusing on males and females. This helps in understanding demographic distributions which could influence subscription preferences.



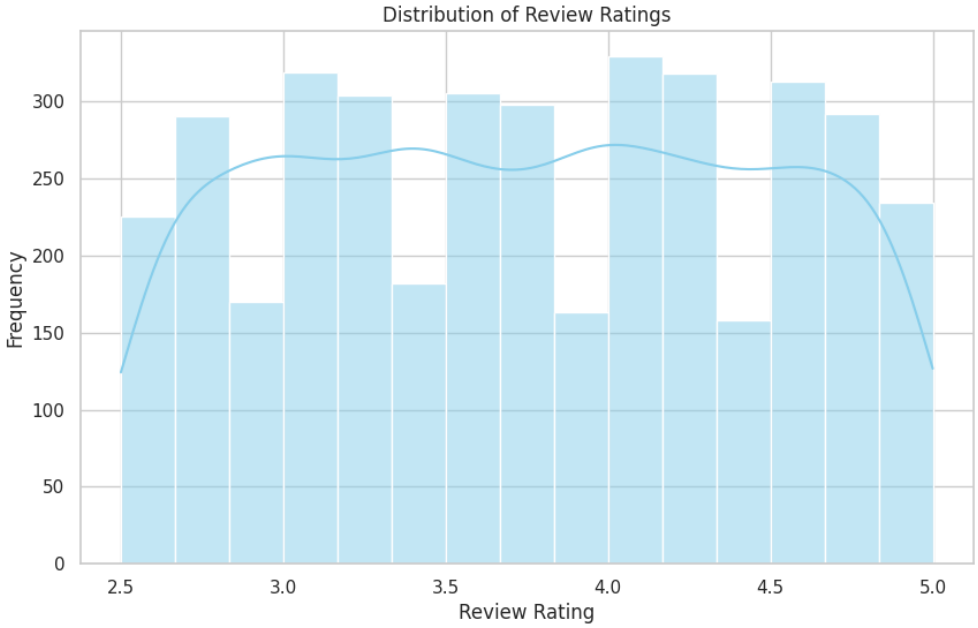
- **Subscription Status by Gender and Discount Application**: Utilizing a stacked bar chart, we compare the subscription status across different genders and whether a discount was applied. This visualization is key in identifying patterns of how discounts impact subscription rates across different genders.

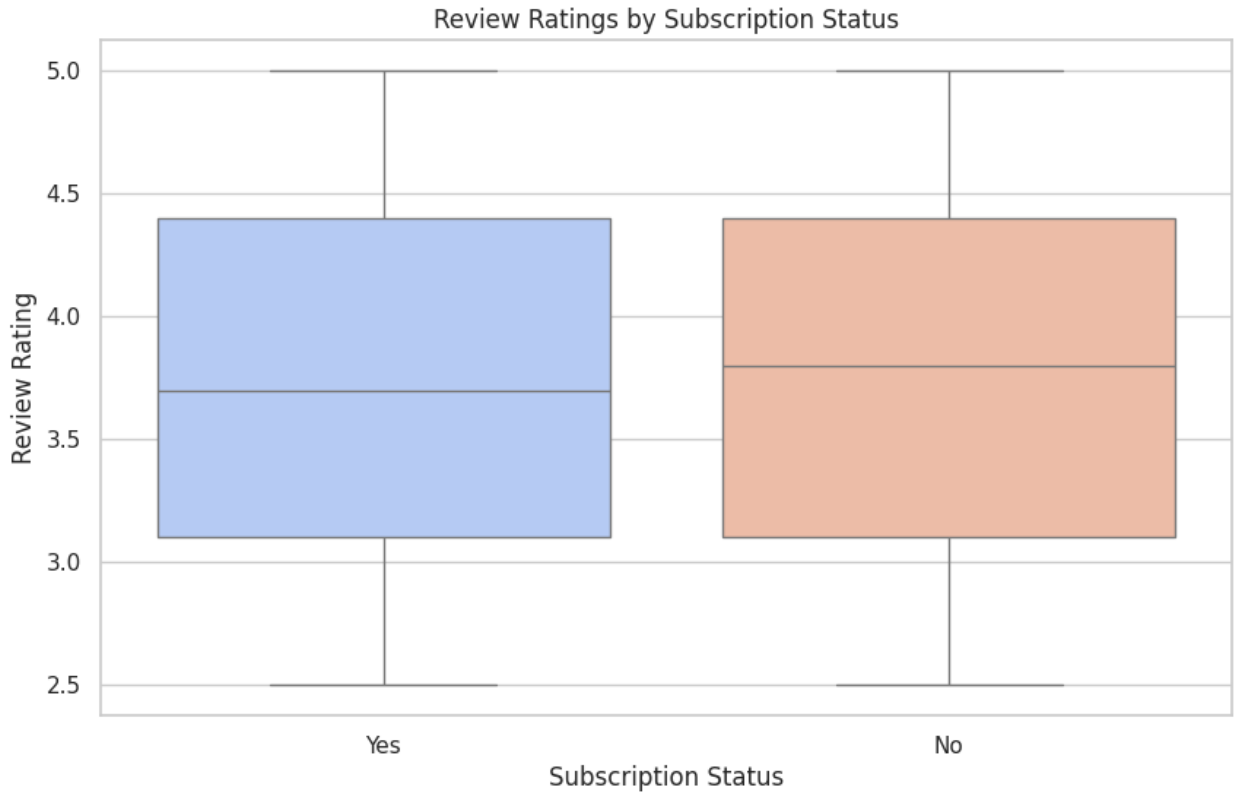


- **Histograms and KDE Plots Review Ratings Distribution**: Histograms and KDE (Kernel Density Estimate) plots are used to visualize the distribution of review ratings. This helps us assess the general satisfaction level of customers and its correlation with subscription status.



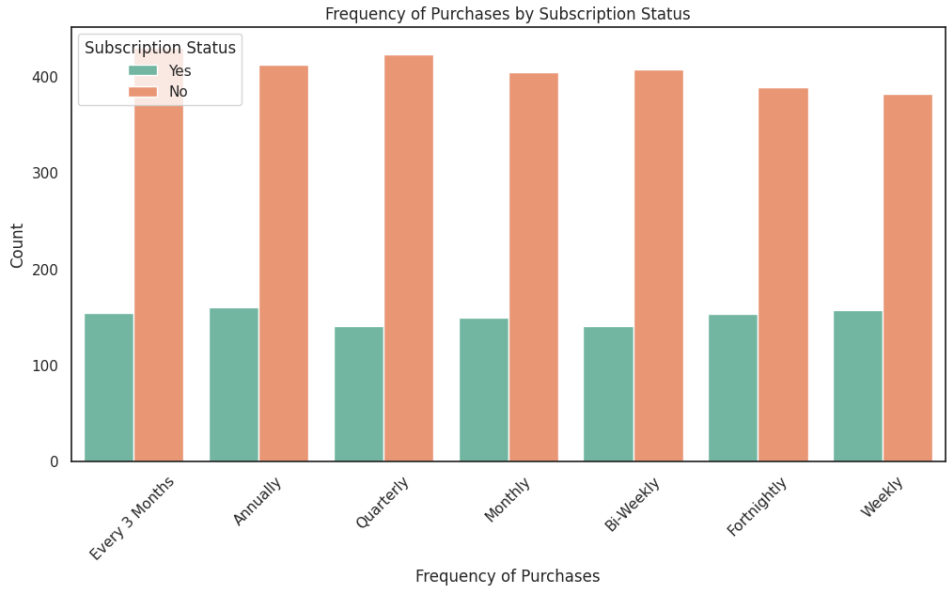
- **Distribution of Review Ratings by Subscription Status**: We further segment the review ratings by subscription status to see how subscriber and non-subscriber ratings differ, which can be pivotal in understanding customer satisfaction's impact on subscription decisions.





Count Plots

- **Frequency of Purchases by Subscription Status**: Count plots allow us to visualize the frequency of purchases categorized by subscription status. This helps in identifying if frequent purchasers are more likely to subscribe, which can inform targeted marketing strategies.

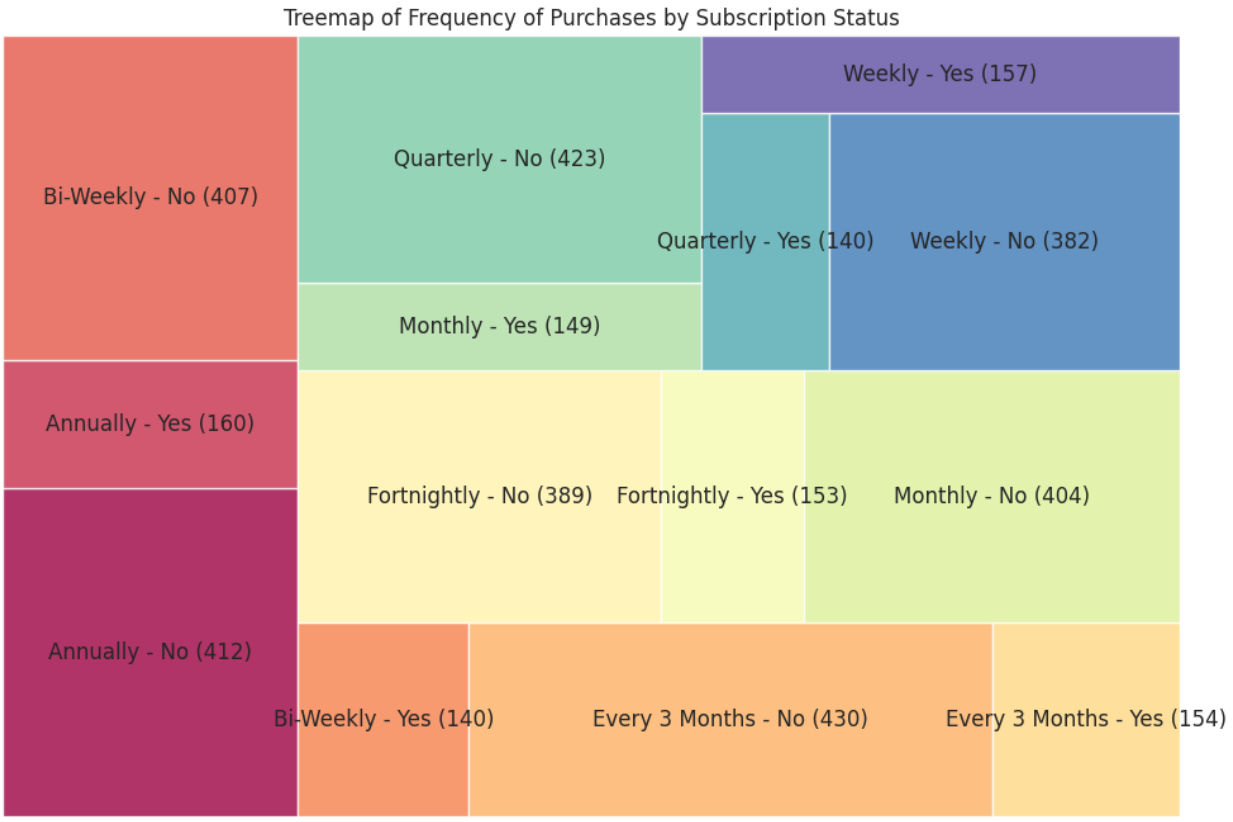


Facet Grids

- **Facet Grid by Gender and Discount Applied**: We use seaborn's Facet Grid to create a series of plots that show the distribution of review ratings across different categories. This method allows us to dissect the data further, analyzing how discount application and gender influence the subscription status reflected through review ratings.

Treemaps

- **Treemap of Frequency of Purchases by Subscription Status**: Treemaps provide a visually appealing and immediate way to comprehend the hierarchy and proportion of purchasing frequencies among different subscription statuses. This visualization helps in quickly



identifying which purchasing frequency categories are more likely to subscribe.

Chi-Square Test Visualization

- **Chi-Square Results for Frequency of Purchases and Subscription Status**: After conducting a chi-square test to assess the independence of purchase frequency and subscription status, we visualize the count data to better understand the association between these variables.

Comparing the purchasing frequency between subscribers and non-subscribers can provide insights into customer loyalty and purchasing habits. This could help identify if subscribers tend to make purchases more often, which would justify efforts to increase the number of subscribers. The model's features are selected and engineered based on these insights. The goal is to train a model that can identify the likelihood of subscription for existing and potential customers. The EDA guides the process of selecting relevant features, transforming data, and creating new variables that capture the essence of the observed patterns. This results in a more robust and accurate predictive model.

Data Mining Tasks

In our project, we initiated the data mining process with crucial data preparation and cleaning tasks to ensure the quality and usability of the data. Here's how we approach these tasks:

7.1. Data Loading

The initial step in our project involves loading the dataset to ensure that all subsequent data processing, exploration, and modeling are based on accurate and complete information. Here's how we manage the data loading process:

Importing Necessary Libraries:

We begin by importing **pandas**, a powerful Python library for data manipulation and analysis, which provides flexible data structures designed to make working with structured data both easy and intuitive.

Reading the Dataset:

Using the **pandas** library, we read the dataset from a CSV file. The CSV format (Comma-Separated Values) is one of the most common import and export formats for spreadsheets and databases.

Verifying Data Integrity:

After loading the data, we perform initial checks to ensure the dataset is loaded correctly and as expected. This includes checking the first few records, understanding the structure of the Data Frame, and confirming that there are no immediate loading errors such as missing or misformatted data.

Understanding Dataset Features:

We obtain an overview of the dataset's columns to understand what features are available. This helps in planning subsequent data processing tasks like cleaning, transformation, and analysis.

Basic Statistical Summary:

A quick statistical summary of the numerical attributes is generated to get insights into the central tendencies and distributions of the data, which can be critical for identifying any anomalies or outliers early in the analysis.

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	14	Venmo	Fortnightly
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	2	Cash	Fortnightly
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	23	Credit Card	Weekly
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	49	PayPal	Weekly
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	31	PayPal	Annually
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3895	3896	40	Female	Hoodie	Clothing	28	Virginia	L	Turquoise	Summer	4.2	No	2-Day Shipping	No	No	32	Venmo	Weekly
3896	3897	52	Female	Backpack	Accessories	49	Iowa	L	White	Spring	4.5	No	Store Pickup	No	No	41	Bank Transfer	Bi-Weekly
3897	3898	46	Female	Belt	Accessories	33	New Jersey	L	Green	Spring	2.9	No	Standard	No	No	24	Venmo	Quarterly
3898	3899	44	Female	Shoes	Footwear	77	Minnesota	S	Brown	Summer	3.8	No	Express	No	No	24	Venmo	Weekly
3899	3900	52	Female	Handbag	Accessories	81	California	M	Beige	Spring	3.1	No	Store Pickup	No	No	33	Venmo	Quarterly

3900 rows × 18 columns

7.2. Data Transformation

Categorizing Age and Review Ratings:

Age Grouping: We transform the 'Age' column into categorical bins to facilitate analysis by age group. These bins are ['0-18', '19-35', '36-50', '51+'], allowing us to categorize customers into relevant demographic segments.

Review Rating Categories: Similarly, we transform 'Review Rating' into ordinal categories ['Very Low', 'Low', 'Medium', 'High'] based on predefined bins. This categorization simplifies analyses by converting continuous ratings into easier-to-manage groups.

7.3. Feature Encoding

Identifying Categorical Columns:

We identify which columns are categorical based on their data type, distinguishing them from numerical columns. This step is crucial for appropriate preprocessing before modeling.

Label Encoding for Low Cardinality Features: For categorical features with ten or fewer unique values, we apply label encoding. This method numerically converts each value in a column.

Target-Guided Encoding for High Cardinality Features: For categorical features with more than ten unique values, we use target-guided encoding. This approach involves ordering labels according to the mean of the target and then encoding them from 0 to n_classes-1 based on this order. This method is effective in capturing important information in features with high cardinality without excessively increasing dimensionality.

7.4. Feature Selection

```
In [56]: from sklearn.feature_selection import mutual_info_classif

In [57]: X = data.drop(['Subscription Status'], axis=1)
         y = data['Subscription Status']

In [58]: mi = mutual_info_classif(X, y)
         mi

Out[58]: array([0.58281442, 0.01703394, 0.13994477, 0.00397048, 0.0047403 ,
                0.00816951, 0.00858817, 0.        , 0.00194129, 0.        ,
                0.        , 0.        , 0.29801697, 0.30870989, 0.        ,
                0.        , 0.01198529, 0.00660877, 0.        ])

In [59]: mi_df = pd.DataFrame(mi, index=X.columns, columns=['Importance'])
         zero_importance_features = mi_df[mi_df['Importance'] == 0].index.tolist()

In [61]: mi_df
```

Mutual Information for Feature Selection:

We utilize the **mutual_info_classif** function from the sklearn library to compute the mutual information between each feature and the target variable ('Subscription Status'). Mutual information quantifies the amount of information the presence or absence of a feature contributes to correctly predicting the target.

Dropping Features with Zero Importance:

Features that exhibit zero mutual information are deemed irrelevant for predicting the target and are subsequently dropped. This reduces the model's complexity and focuses the learning process on the most impactful features.

Data Reduction and Refinement:

Standard Scaling: Features are standardized using **StandardScaler** to ensure that they have zero mean and unit variance, which is crucial for many machine learning algorithms, particularly those sensitive to the scale of input features.

Feature Extraction: Beyond mutual information, we consider extracting new features or modifying existing ones to better capture underlying patterns related to the subscription status. These comprehensive data preparation steps encompassing transformation, encoding, and feature selection are designed to optimize our dataset for effective machine learning modeling. By refining the feature set and ensuring that only the most relevant variables are included, we enhance both the performance and interpretability of our subsequent analytical models.

	Importance
Customer ID	0.582814
Age	0.017034
Gender	0.139945
Item Purchased	0.003970
Category	0.004740
Purchase Amount (USD)	0.008170
Location	0.008588
Size	0.000000
Color	0.001941
Season	0.000000
Review Rating	0.000000
Shipping Type	0.000000
Discount Applied	0.298017
Promo Code Used	0.308710
Previous Purchases	0.000000
Payment Method	0.000000
Frequency of Purchases	0.011985
AgeGroup	0.006609
Review_Rating_Category	0.000000

Data Mining Models/Methods

8.1 Overview of Applied Models

In our project, we have employed a variety of machine learning models to predict subscription status based on customer data. These models include:

- K-Nearest Neighbors (KNN)
- Gaussian Naive Bayes (GNB)
- Logistic Regression

Each model was chosen for its suitability to handle binary classification tasks and its ability to provide insights into different aspects of the data. The complete dataset is split into two parts: one part for training (and validating) the models, and another for testing them. This split helps in training the models on one subset of the data and testing their performance on unseen data to ensure that the models generalize well beyond the training data.

8.2 Detailed Description of Each Model/Method

K-Nearest Neighbors (KNN) Classifier:

Process: The KNN algorithm classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used here to classify customers into 'subscribed' or 'not subscribed' based on the similarity of their features to those of known cases.

Implementation: We scaled the features using **StandardScaler** to ensure the model is not biased towards variables on larger scales. The dataset was split into training, validation, and test sets to ensure robustness and prevent overfitting. Cross-validation was used to optimize the model's reliability.

For model, the approach starts by scaling all input features to ensure uniformity in distance calculations, which is critical given the diverse range of features like age, purchase amount, and previous purchases. The model then classifies each customer by analyzing the subscription statuses of the 'k' closest neighbors in the feature space, determining whether a new or unseen customer is likely to subscribe based on the most common outcomes among those nearest

neighbors. This method leverages the dataset's extensive range of demographic and transactional features to infer patterns that might indicate a likelihood of subscription.

Gaussian Naive Bayes (GNB) Classifier:

Process: This model applies Bayes' theorem, assuming the predictors are independent given the class. Despite this strong assumption, GNB can perform well in many real-world scenarios and is particularly fast for large datasets.

Implementation: Similar to KNN, we used cross-validation to evaluate its performance and adjust parameters. It is particularly useful for understanding feature importance due to its probabilistic foundation.

The model is particularly well-suited for handling the mix of numerical and categorical data in this dataset. By assuming independence between features such as gender, item purchased, and location, despite their potential correlations, the model simplifies the calculation of conditional probabilities. Each feature contributes independently to the probability that a customer subscribes, allowing for rapid classification even with the extensive data set provided. This model is effective in identifying underlying probabilities of subscription from patterns dispersed across various customer attributes and behaviors.

Logistic Regression:

Process: Logistic regression estimates the probability of a binary response based on one or more predictor variables. It models the probability that each input belongs to a particular category.

Implementation: Post standardization, logistic regression was applied, and its performance was assessed across multiple metrics such as accuracy, precision, recall, and F1 score. The use of PCA to reduce overfitting and enhance model performance was also explored.

Logistic Regression in this context is employed to estimate the probability of a customer subscribing based on a logistic function of input features like age, purchase amount, and frequency of purchases. This model not only predicts the outcome but also quantifies the influence of each feature on the likelihood of subscription, offering interpretable results that can guide business strategies. For example, logistic regression can reveal how changes in the discount level, or the use of promo codes impact the probability of subscription, providing actionable insights into how promotional strategies might be optimized to increase customer subscription rates.

8.3 Justification of Choices

- **KNN**: Chosen for its simplicity and effectiveness in classification tasks where relationships in data can be captured through distance metrics.
- **Gaussian Naive Bayes**: Selected due to its efficiency with large datasets and its performance in scenarios where independence between features is a reasonable approximation.
- **Logistic Regression**: Used for its ability to provide probabilities and its robustness in binary classification problems. It's beneficial for interpretability, which is crucial for aligning business strategies based on model outputs.

The combination of these models provides a comprehensive approach, allowing us to capture different patterns and relationships in the data. The diversity of these methods also aids in validating findings across different algorithmic assumptions, enhancing the reliability of our predictions.

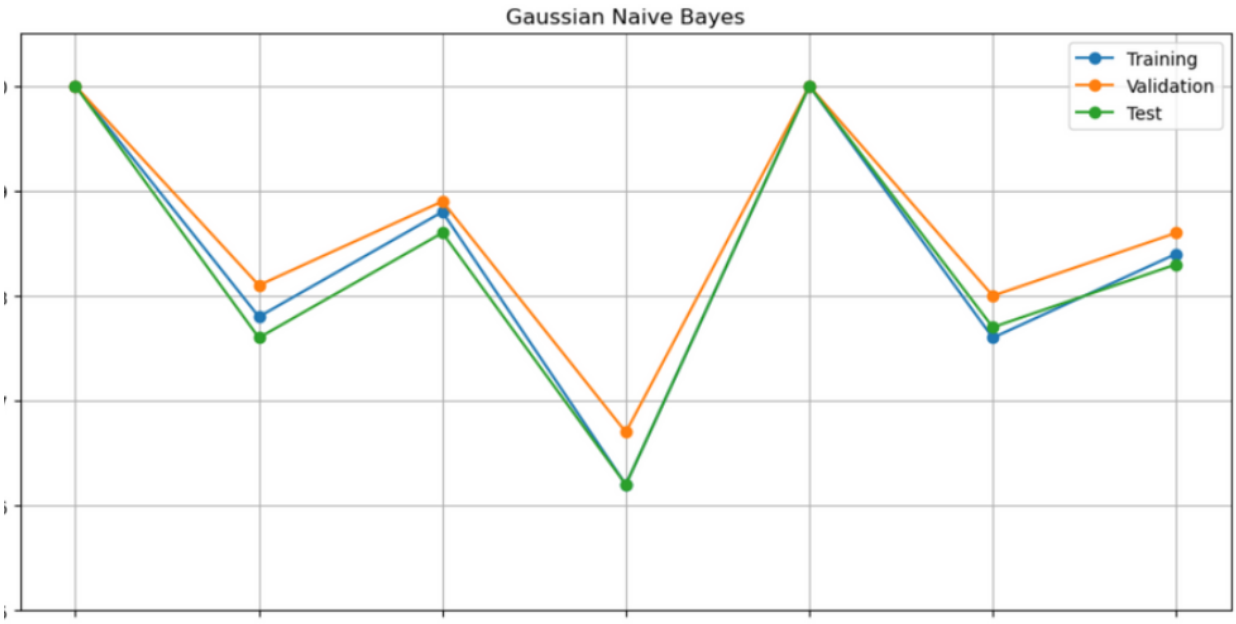
Performance Evaluation

Gaussian Naive Bayes (GNB)

Performance Consistency Across Different Phases:

Accuracy Levels: GNB demonstrated stable accuracy levels with training at 0.84, validation at 0.86, and testing at 0.83. The consistent performance across these phases, albeit slightly lower than that of KNN, highlights the model's reliability.

Efficiency and Speed: The model's efficiency in handling the data and producing results quickly, combined with consistent accuracy, makes GNB particularly useful for scenarios requiring rapid decision-making. Its performance suggests a robust modeling approach that, despite its simplicity and the assumption of feature independence, effectively captures the underlying patterns in the data.



Evaluation Metrics:

Stability and Predictive Power: The stability of the model across different data splits is particularly valuable in applications where the model needs to perform consistently under varying conditions

```
Cross-Validation Accuracy Scores: [0.83814103 0.83814103 0.85737179 0.83814103 0.84615385]
Mean CV Accuracy: 0.84
Training Accuracy: 0.84
Validation Accuracy: 0.86
Test Accuracy: 0.83
```

.

K-Nearest Neighbors (KNN)

Generalization and Overfitting:

Minor Fluctuations: The minor fluctuations observed in KNN's performance metrics across the training, validation, and test sets indicate good generalization without significant overfitting. This suggests that the model is robust enough to handle unseen data effectively.

Performance Across Classes: The relatively high metrics for both classes confirm that KNN can successfully distinguish between subscribers and non-subscribers, an essential feature for targeted marketing strategies.

Validation and Testing Correlation:

Consistency Between Phases: The close alignment between validation and test set performances validates the model's effectiveness and supports its deployment in practical settings where predicting new data accurately is crucial.

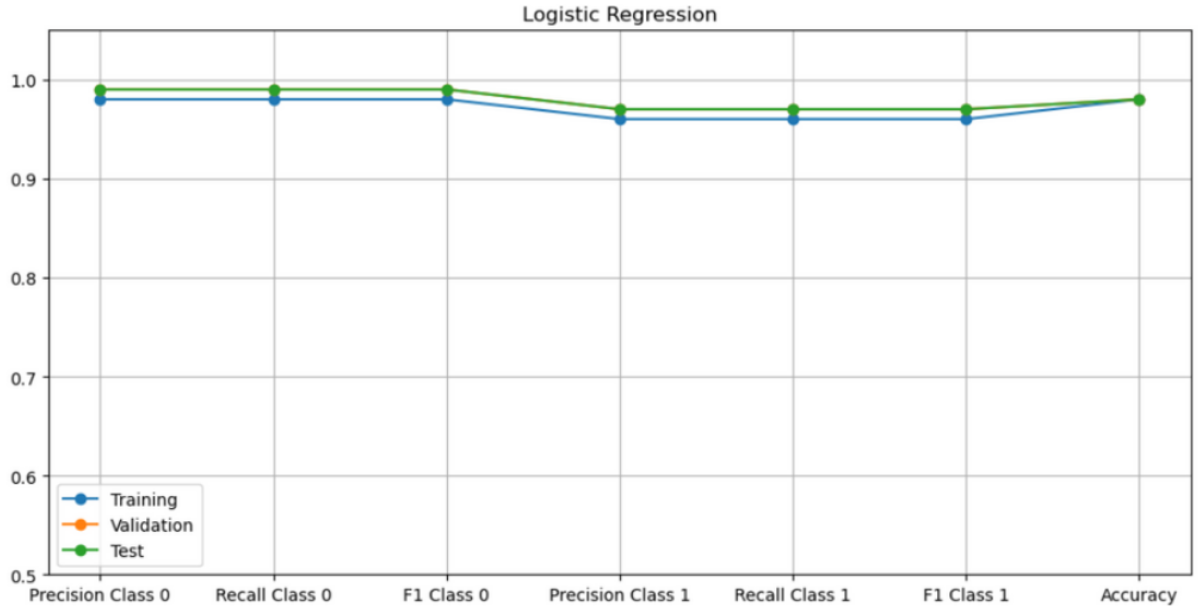
```
Cross-Validation Accuracy Scores: [0.8525641  0.86538462 0.89423077 0.86217949 0.87660256]
Mean CV Accuracy: 0.87
Training Accuracy: 0.91
Validation Accuracy: 0.86
Test Accuracy: 0.86
```

Logistic Regression (Before and After PCA)

Model Robustness and Overfitting:

Initial High Performance: Initially, the logistic regression model showed high accuracy in both training and validation phases, hinting at potential overfitting. However, the uniformity in validation and test accuracies suggests that the model, while potentially over-tuned to the training data, still generalizes well.

PCA Impact: The application of PCA significantly reduced overfitting by focusing on the most critical features. This not only streamlined the model but also enhanced its predictive accuracy without compromising the depth of analysis.



Superior Performance Metrics:

Across All Phases: The almost perfect scores across all metrics after applying PCA demonstrate the model's exceptional ability to capture and predict based on the relevant features. The high precision and recall for both classes across all data sets illustrate the model's comprehensive understanding of the data patterns.

```
Cross-Validation Scores: [0.99679487 0.99358974 0.99679487 0.99519231 0.99198718]
Mean CV Score: 0.99
Training Set Accuracy: 1.00
Validation Set Accuracy: 1.00
Test Set Accuracy: 1.00
```

Summary of Findings and Strategic Implications

Model Hierarchy: Logistic Regression, particularly after dimensionality reduction through PCA, is identified as the superior model due to its robust and consistent performance across all datasets and metrics. Its ability to provide deep insights into feature impacts makes it invaluable for refining marketing and operational strategies.

Areas for Model Enhancement: While GNB shows promise, its tendency to favor predicting one class and less stable metrics in some instances suggests a need for further tuning or potential integration with other modeling approaches to enhance its predictive balance and accuracy.

```
Training Set Accuracy: 0.98
Validation Set Accuracy: 0.98
Test Set Accuracy: 0.98
Cross-Validation Scores: [0.97916667 0.97275641 0.98717949 0.98397436 0.96794872]
Mean CV Score: 0.98
```

This evaluation underscores the strengths and limitations of each model and provides a roadmap for ongoing improvement and application. The insights gained from these evaluations are crucial for continuous model refinement and ensuring that the models remain effective in dynamic real-world applications.

Project Results

The comprehensive data mining project focusing on enhancing customer subscription rates has yielded significant insights and actionable recommendations. Here's a detailed overview of the key findings and the deliverables from the project:

Insights on Customer Behavior

Influence of Discounts and Reviews:

The analysis highlighted how discount applications and customer review ratings significantly impact subscription decisions. Discounts appear to serve as a strong incentive for subscriptions, while higher review ratings correlate with increased likelihood of subscription continuity.

These findings suggest that customers are responsive to monetary incentives and highly value their satisfaction with previous purchases, indicating that perceived value and customer experience are pivotal in their decision to subscribe.

Demographic and Behavioral Factors:

Further exploration into demographic and behavioral data revealed specific patterns, such as which age groups or purchase behaviors are more likely to lead to subscriptions. This nuanced understanding allows for more targeted marketing and customer engagement strategies.

Model Selection for Subscription Rates

Logistic Regression's Superior Performance:

Among various models tested, Logistic Regression stood out for its accuracy and robustness in predicting subscription outcomes. After applying PCA to reduce dimensionality and focus on the most relevant features, the model's performance was further enhanced, providing almost perfect prediction accuracy across multiple data sets.

The model's ability to quantify the impact of each feature on subscription likelihood provides a solid foundation for strategic decision-making, enabling precise adjustments to marketing and operational strategies based on predictive insights.

Recommendations for Business Strategy

Enhancing Customer Satisfaction:

Given the strong correlation between customer satisfaction (as evidenced by review ratings) and subscription rates, it is recommended that the company invests in strategies aimed at improving overall customer experience. This could involve enhancing product quality, customer service, and post-purchase support to boost satisfaction and, consequently, subscription rates.

Revising Discount Strategies:

The project findings suggest revisiting current discount strategies to maximize their effectiveness in driving subscriptions. Tailoring discounts based on customer segments that have shown higher sensitivity to pricing adjustments could yield better subscription rates. Analyzing the timing, depth, and type of discounts could help in crafting offers that are more likely to convert one-time buyers into subscribers.

Data-Driven Marketing Campaigns:

Utilize the predictive model to segment the customer base into groups based on their predicted likelihood of subscription. Develop targeted marketing campaigns that address the specific needs and preferences of these segments. For example, highly personalized campaigns could be directed at individuals identified as on the fence about subscribing, using incentives they are likely to find appealing.

Project Deliverables

Predictive Model Deployment: The logistic regression model, optimized for accuracy and interpretability, ready for deployment to forecast customer subscription likelihood.

Strategic Business Recommendations Report: A detailed report outlining specific strategies for improving subscription rates based on model insights.

Customer Segmentation Framework: A strategic framework for categorizing customers based on their behaviors and predicted preferences, aiding targeted marketing efforts.

Impact of the Project Outcomes

The data mining project focused on enhancing customer subscription rates has generated significant value for the business by leveraging predictive modeling to inform and optimize marketing and customer retention strategies. Here's how the outcomes of this project create value:

1. Improved Targeting and Personalization

- **Precision Marketing**: The Logistic Regression model, with its detailed insights into the factors that influence subscription likelihood, enables the business to tailor marketing efforts with unprecedented precision. By understanding which features (such as age, discount sensitivity, and prior purchasing behavior) most influence subscription decisions, marketing can be more effectively targeted.
- **Personalized Customer Engagement**: Insights from the model allow for personalized engagement strategies. For example, customers identified as highly responsive to discounts can be targeted with special promotional offers, while those influenced more by product quality or customer service can receive information on new enhancements and support initiatives.

2. Optimization of Promotional Strategies

- **Effective Discounting**: The predictive model has highlighted the significant impact of discount strategies on subscription rates. This enables the business to optimize discount levels and timing, applying them in ways that maximize conversion rates while minimizing unnecessary expenditure on overly broad or ineffective discount campaigns.
- **Dynamic Pricing Models**: Leveraging predictive analytics, the company can implement dynamic pricing strategies that adapt to changes in customer preferences and market conditions, ensuring that pricing structures are always aligned with what is most likely to drive subscriptions.

3. Enhanced Customer Retention

- **Predicting Churn**: By identifying which customers are at risk of not subscribing or are likely to unsubscribe, targeted interventions can be designed to address their specific concerns and improve retention rates.
- **Customer Satisfaction Improvement**: Detailed analysis of customer reviews and their impact on subscription likelihood offers actionable insights for improving product and service quality, directly feeding into better customer satisfaction and retention.

4. Resource Allocation Efficiency

- **Focused Resource Utilization**: Insights from the data mining project help in focusing resources where they are most needed and will have the most substantial impact. This means allocating budget to marketing campaigns, customer service improvements, and product enhancements that are most likely to influence subscription decisions based on predictive model findings.
- **Cost Reduction**: By avoiding blanket strategies and instead focusing on targeted approaches informed by predictive modeling, the company can reduce wasteful spending and allocate resources more efficiently.

5. Strategic Decision Making

- **Data-Driven Strategies**: The outcomes of the project support strategic decision-making by providing a robust, data-driven basis for it. This reduces reliance on intuition and anecdotal evidence, leading to more consistent and predictable results.
- **Market Competitiveness**: Armed with advanced analytics capabilities, the company can stay ahead of market trends and more swiftly adapt to changing consumer behaviors, keeping it competitive in a rapidly evolving market.

Conclusion

The project not only provided a deep dive into the factors influencing customer subscriptions but also equipped the business with a robust analytical tool to predict and enhance future subscription rates. These outcomes are set to guide strategic decisions, aiming to convert more one-time users

into long-term subscribers through informed, data-driven strategies. This holistic approach ensures that the business can adapt to changing customer preferences and market conditions, maintaining a competitive edge in the subscription-based market landscape.