Predicting Customer Purchase Behavior: A Binary Classification Model for E-Commerce Websites

Draft of Project Report

Group 34

Student 1 : Saisreeja Yalavarthi
Student 2 : Sanidhya Mathur

(857)-330-0756 (Tel of Student 1)
(857)-268-8284 (Tel of Student 2)

yalavarthi.sa@northeastern.edu
mathur.san@northeastern.edu

Percentage of Effort Contributed by Student 1: 50%
Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Saisreeja Yalavarthi
Signature of Student 2: Sanidhya Mathur

Submission Date: 04/07/2023

1. The Problem

1.a Problem Setting:

In today's world, online shopping has reached each and every doorstep adopted by all the age groups alike. However, we all have found ourselves casually surfing around various categories but eventually end up not buying any product. Such a situation becomes key for the company to analyze and determine what are the various tasks or actions which will push the customers to buy their product.

1.b Problem Definition:

We are proposing to build a model that will predict if a customer makes a purchase during their website session by using features such as session duration, region and browser type. The target variable for the model will be whether or not a purchase was made and the input features are other columns in the dataset. The model's performance will be evaluated based on its accuracy in predicting the "Revenue" column. Thus, we shall develop a binary classification model to predict customer purchase behavior on a given website. This model can be helpful for the sales & marketing department to analyze what attracts a user and also incentivize with curated coupons or offers.

2. The Data

2.a Data Sources:

The dataset for the project is extracted from the UCI repository: http://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset#

2.b Data Description:

The dataset includes features from over 12,330 sessions with 18 attributes (10 numerical and 8 categorical) that provide insight into the behavior of visitors on an e-commerce website, including the number of pages visited and time spent on each type of page. Moreover, metrics such as "Bounce Rate", "Exit Rate", and "Page Value" measured by "Google Analytics" are also taken into consideration. Additional features include "Special Day", which indicates the closeness of the site visiting time to a specific special day, and "Page Value" which represents the average value for a web page that a user visited before completing an e-commerce transaction. The dataset also includes information about the visitor's operating system, browser, region, traffic type, visitor type, whether the date of the visit is a weekend, and the month of the year.

Sr. No	Data Name	Data Type	Data Information
1	Administrative	Numerical	Number of administrative pages visited in a session
2	Administrative Duration	Numerical	Duration of administrative page visited
3	Informational	Numerical	Number of informational pages visited in a session
4	Informational Duration	Numerical	Duration of informational page visited
5	Product Related	Numerical	Number of product-related pages visited in a session
6	Product Related Duration	Numerical	Duration of product-related page visited
7	Bounce Rates	Numerical	Visitor's percentage during a session who enter the site but do not trigger any request to analytics server before exit
8	Exit Rates	Numerical	Visitor's percentage during a session who leave the site from this particular page
9	Page Values	Numerical	Google Analytics' value assigned to a page that a user visited after which the target (buying) was achieved
10	Special Day	Numerical	0-1 value defining closeness of the visit to a certain special day (as transactions are usually completed more often)
11	Month	Categorical	Month of the transaction
12	Operating System	Categorical	Operating System used by the user for shopping
13	Browser	Categorical	Browser used by the user
14	Region	Categorical	Region of the user from where they are browsing
15	Traffic Type	Categorical	20 classifications for user traffic on the page
16	Visitor Type	Categorical	Classified as Returning or New Visitor
17	Weekend	Categorical	Boolean Value specifying weekend as True or False
18	Revenue	Categorical	Boolean Value specifying whether revenue was generated on a particular visit or not. This serves as our Class Label

3. **Data Exploration and Data Visualization**:

The Customer Purchase Behavior dataset has over 12,000 session records, each contributing to the ultimate question, "Was Revenue made from this session?" based on the 18 numerical and categorical attributes. It is thus important to analyze how these records are distributed across the variables.

We note that there are no missing values in the dataset as each session has been characterized by the type of session, their purpose, region, type of day, operating system etc. Now, we start our exploration for numerical variables first.

3.a Numerical Variables

The columns store "counts of visits" when the purpose was "administrative" or "informational". However, we observe that the graph is dominated by "0" values, as naturally, out of the 12,330 sessions, there are several which are not administrative or informational.

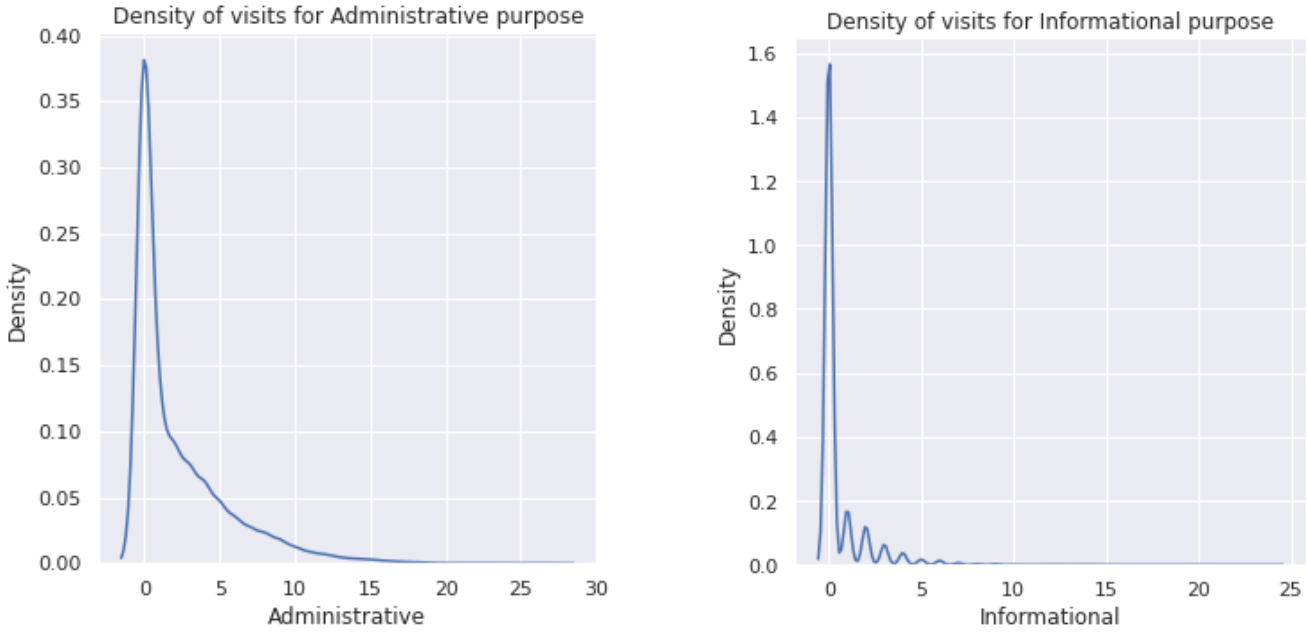


Fig i. Density of visits for Administrative and Informational Purpose

As the trend is extremely skewed right, we are unable to make sense of the pattern for a large set of values. It gives a false image that most of the values are clustered towards the right, that is near “zero-value” visits.

However, we infer from Chapter 4 of the book “ Data Mining for Business Analytics “, that Zooming and Panning the information to our interest helps reveal patterns and outliers in a better manner.

Thus, let us zoom and observe the pattern for non-zero values.

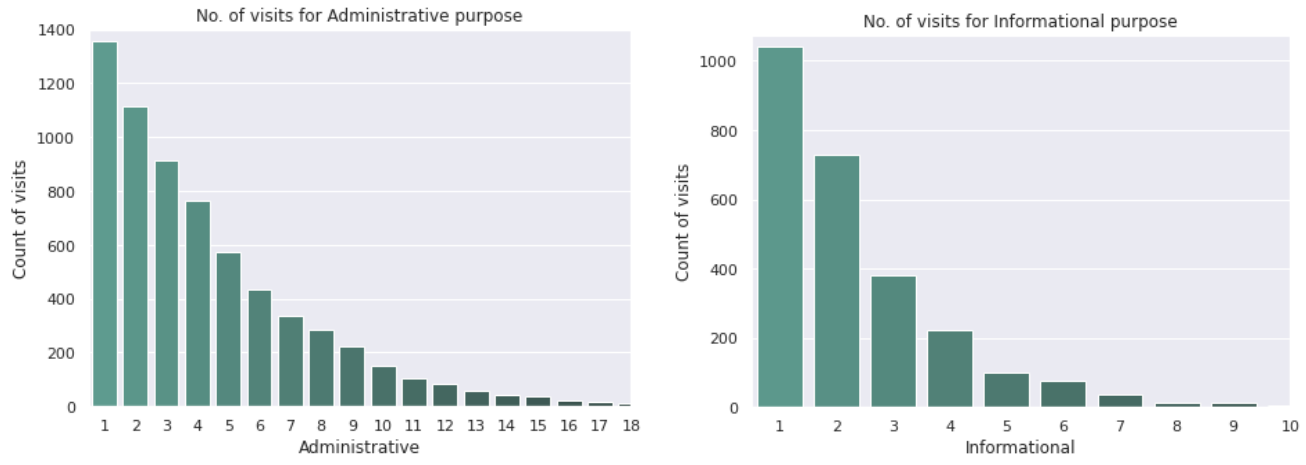


Fig ii. No. of visits for Administrative and Informational Purpose

Thus, we see a clear pattern of maximum number of visits for administrative purposes being around 1 to 5, while for informational purposes being 1 to 3.

For a customer’s catalog and shopping website, the purpose of visits being administrative and informational too are centered around the products on display. Hence, it is of utmost importance to observe the trend of the number of visits in a session, which are related to products available on the website.

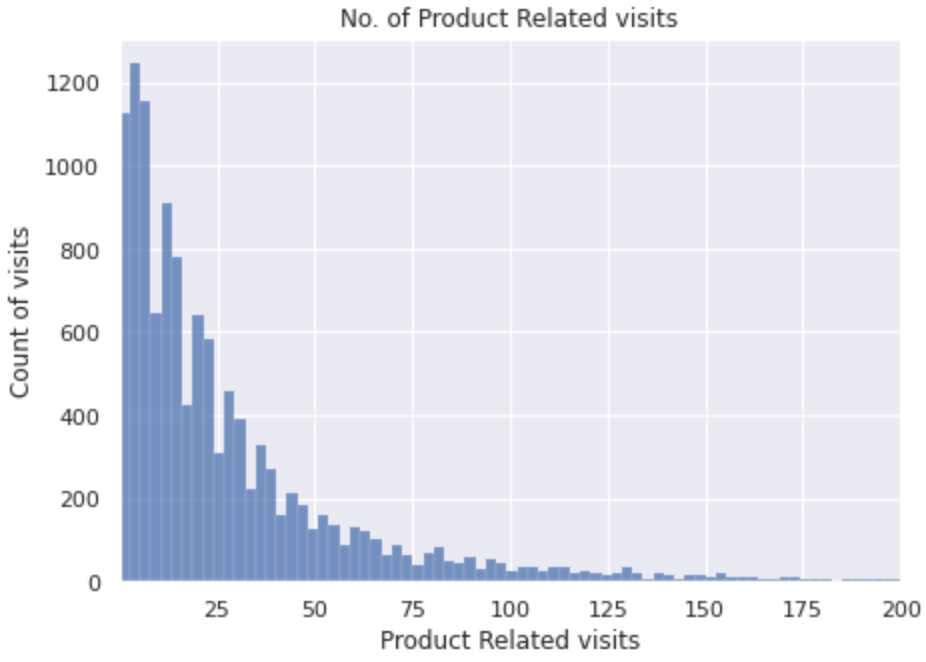
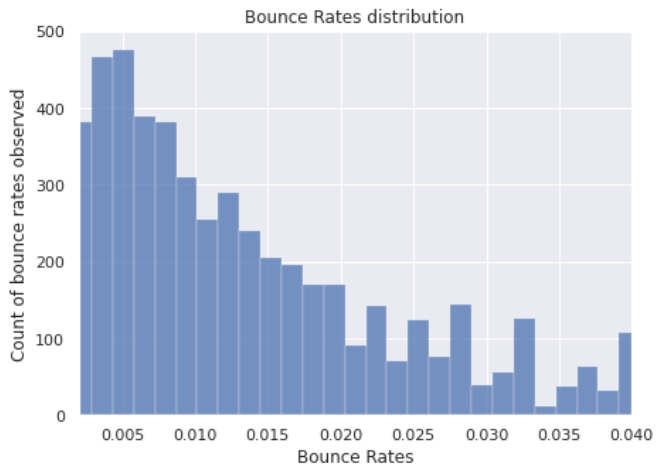


Fig iii. No. of visits related to Products

We can clearly see that most visits related to products are around 0 to 25 in a particular session.

Bounce Rates and Exit Rates form two key parameters which determine whether Revenue will be made in a particular session or not. Bounce Rates are visitor's percentage during a session who enter the site but do not trigger any request to the analytics server before exit. Exit Rates are Visitor's percentage during a session who leave the site from this particular page. Both these values help understand the customer sentiment about a particular page. Thus we can customize the pages in order to have a higher chance of the session eventually yielding revenue.



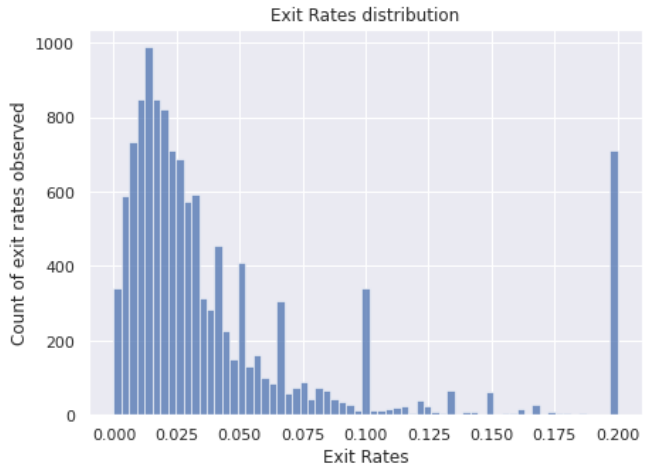


Fig iv. Distribution of Bounce Rates and Exit Rates

Both the rates are skewed-right, with Exit Rates being heavily skewed but yet having a bulk of values at Exit Rate = 0.200. Although we note the above distributions, we are yet not sure whether these sessions will yield revenue. Hence we plot the distribution of Bounce Rates and Exit Rates versus Revenue.

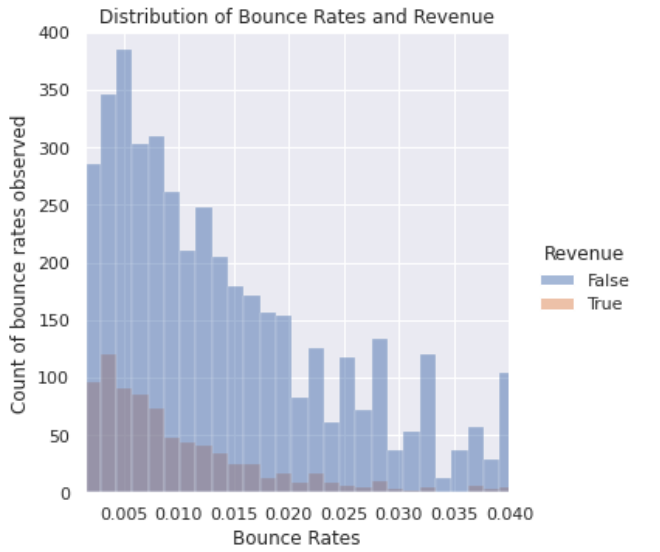


Fig v. Distribution of Bounce Rates, Exit Rates and Revenue

From the above plot of Bounce Rates vs Revenue we can interpret that Revenue is generated merely from 0.005 till 0.015 indicating that Revenue is hardly generated over the gradually increasing Bounce Rates. Similarly for the Exit Rates vs Revenue plot we can observe that users mostly try to find out and analyze the information of the products they are interested in rather than making a purchase.

Joint Plot helps us better understand the data, as we can observe the frequency distribution of both Exit Rates and Bounce Rates, both showing outliers at the end. Along with this, we also note the revenue generated over the scatterplot of Bounce and Exit rates.

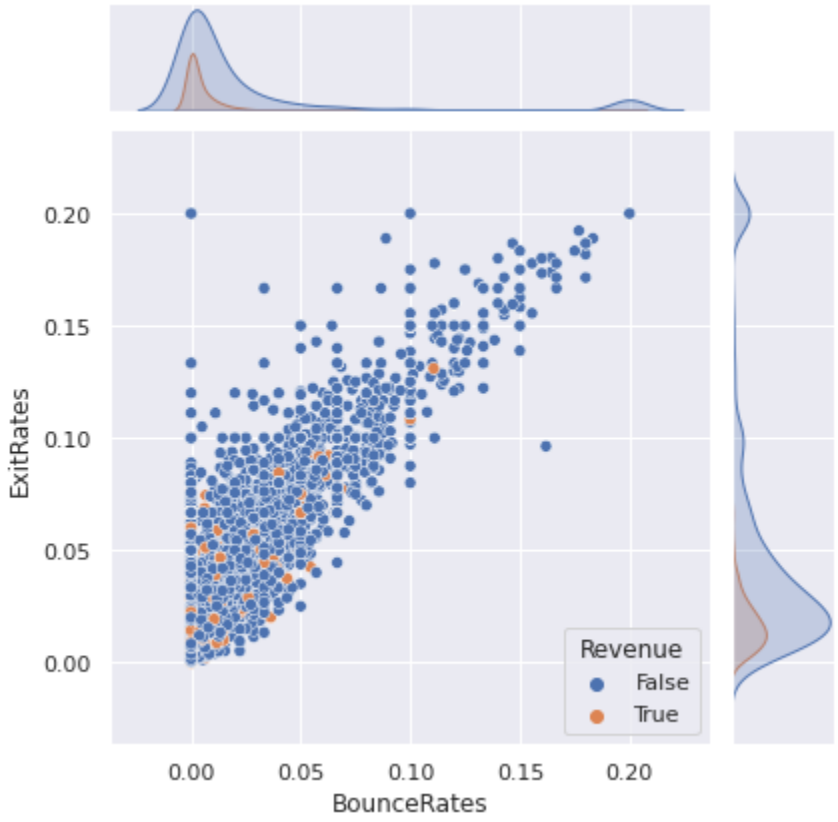


Fig vi. Joint Plot for Bounce Rates versus Exit Rates

3.b Categorical Variables

Special Day refers to occasions or days like Mother's Day, Father's Day, Valentine's Day etc., during which customers tend to make purchases and result in increasing revenue over those days. Offers during Christmas and Thanksgiving contribute to high revenue generation during the Holiday Season.

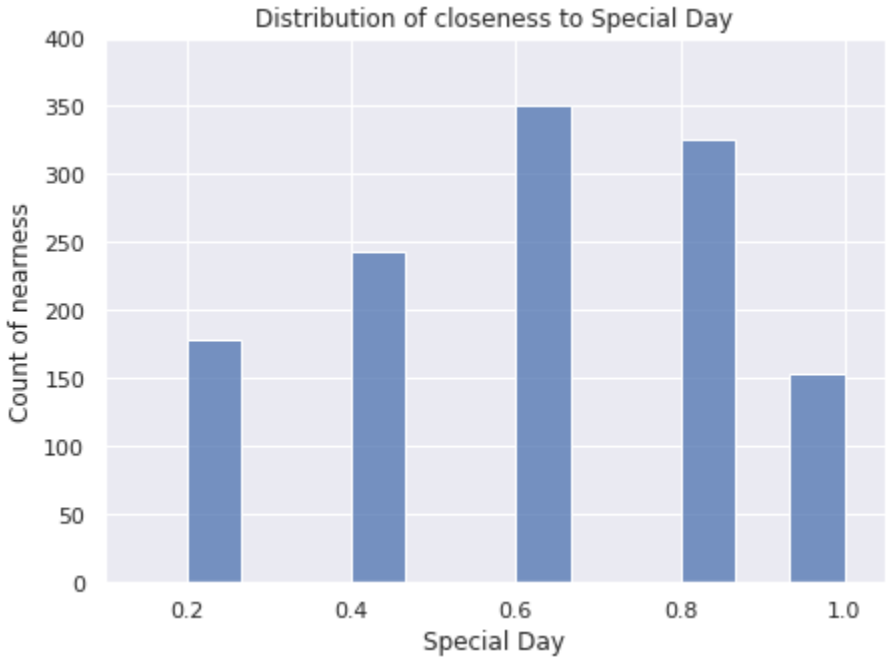


Fig vii. Nearness of number of visits to Special Days

The bar graph clearly indicates that on a scale from 0-1, customers are most likely to make purchases when special occasions are nearby as maximum purchases have a rating of 0.6 and 0.8 with a count of 350 and 300. Even during the "Special Day", there are a considerable number of visits by customers.

Revenue generation varies from region to region. The dataset is divided into multiple regions with each region having varied number of visits and hence varied response to whether revenue was generated or not.

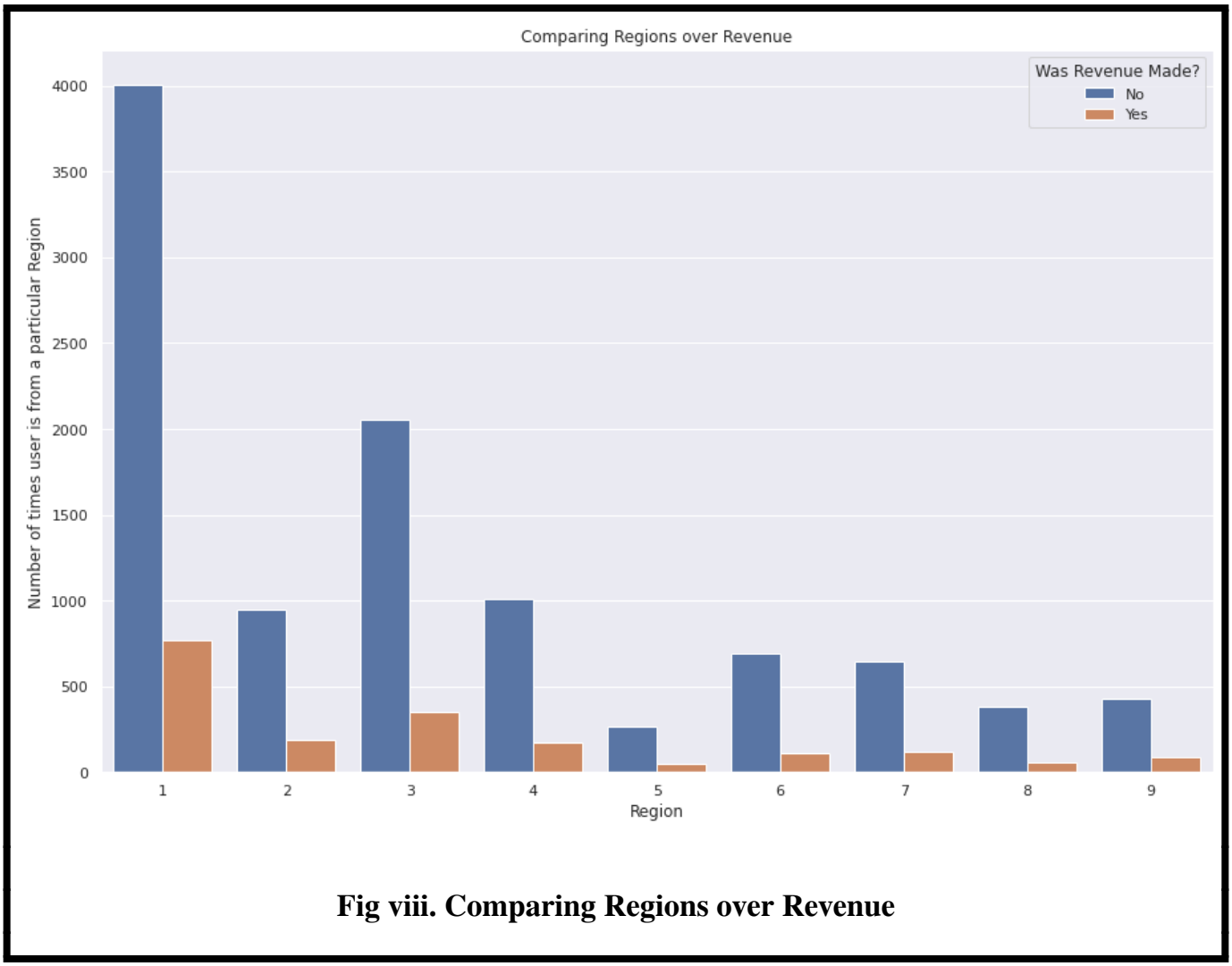


Fig viii. Comparing Regions over Revenue

It can be observed that users from Region 1 generate more revenue as compared to other regions and users from Region 5 and 8 are the lowest revenue genrators. It can also be inferred that Region 1 has the highest number of users logging in followed by Region 3 whether Revenue is generated or not.

Regions having a high number of visits can be optimized by customized suggestions on products they should buy. On the other hand, regions having lower number of visits, should be invited to the website by targeted and lucrative advertisements.

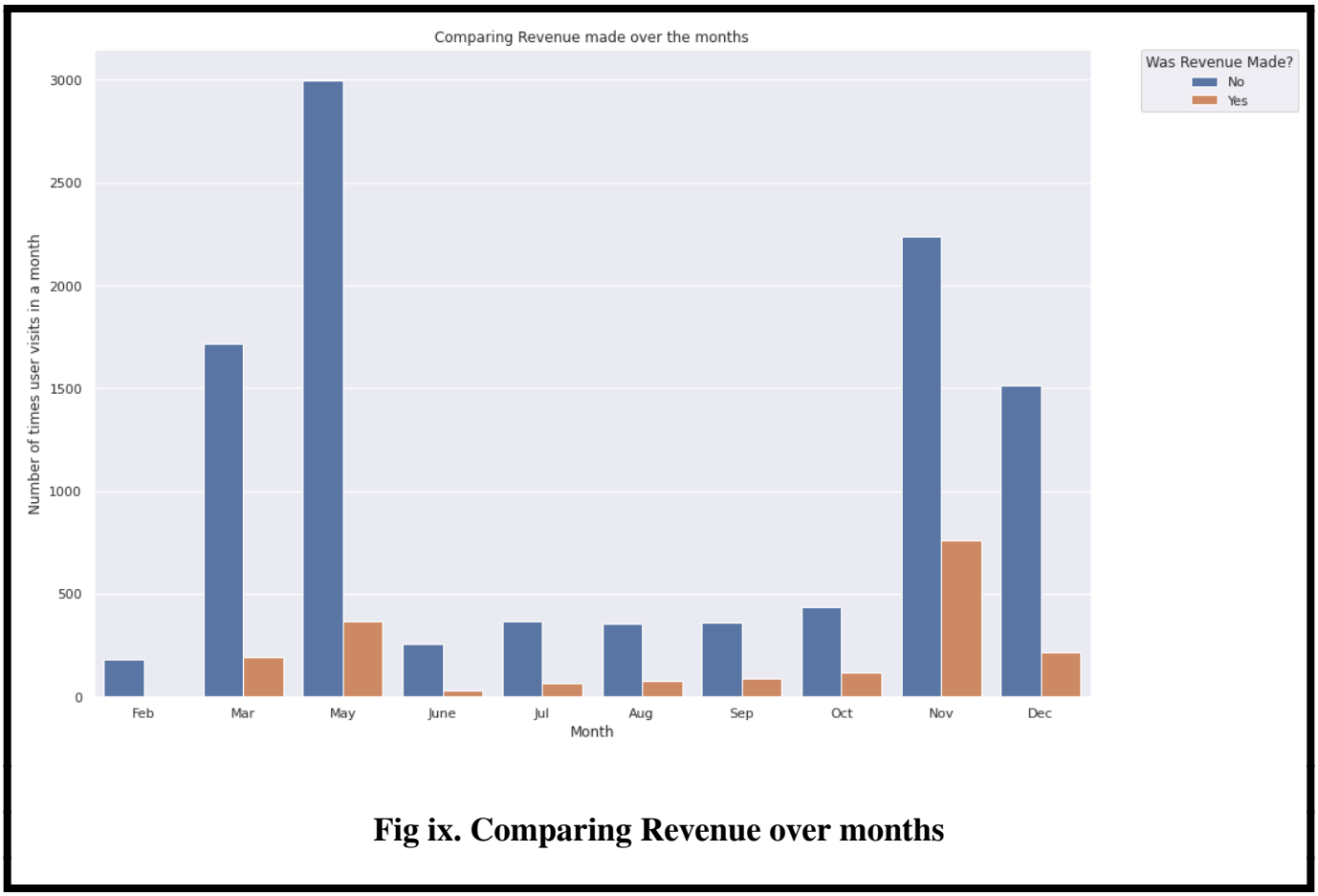


Fig ix. Comparing Revenue over months

May is the month, which tends to be the holiday season, tops with the highest number of user visits followed by November, March and May. Consequently, these months also serve as the most revenue generating months. November being the Thanksgiving and start of Christmas season with high offers and deals results in higher revenue generation followed by May and December respectively.

We observe a correlation matrix between numerical variables to make the dataset efficient by dropping attributes that have high correlation with other attributes.

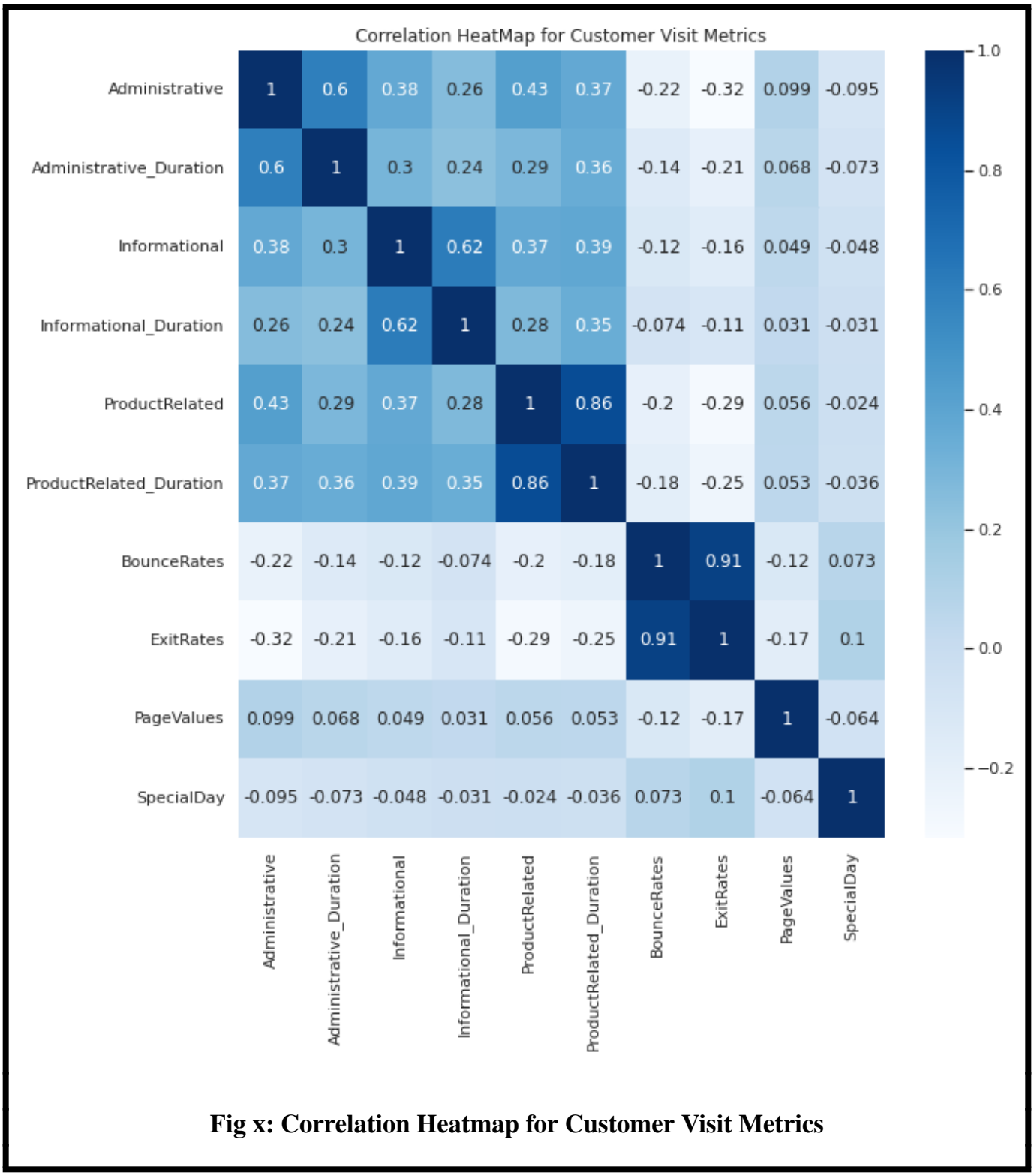


Fig x: Correlation Heatmap for Customer Visit Metrics

Correlation plot clearly indicates a strong affinity between Bounce Rates and Exit Rates. The columns Product Related and Product Related Duration are also highly correlated.
It is to be noted, that we are not willing to drop Bounce Rates and Exit Rates as “Customization of Pages - in order to have a higher chance of the session eventually yielding revenue” is highly dependent on these two factors. However, we drop the column “Product Related Duration” as the information provided is highly correlated with the column “Product Related”.

Similar trend can be observed in the below Pair Plot.

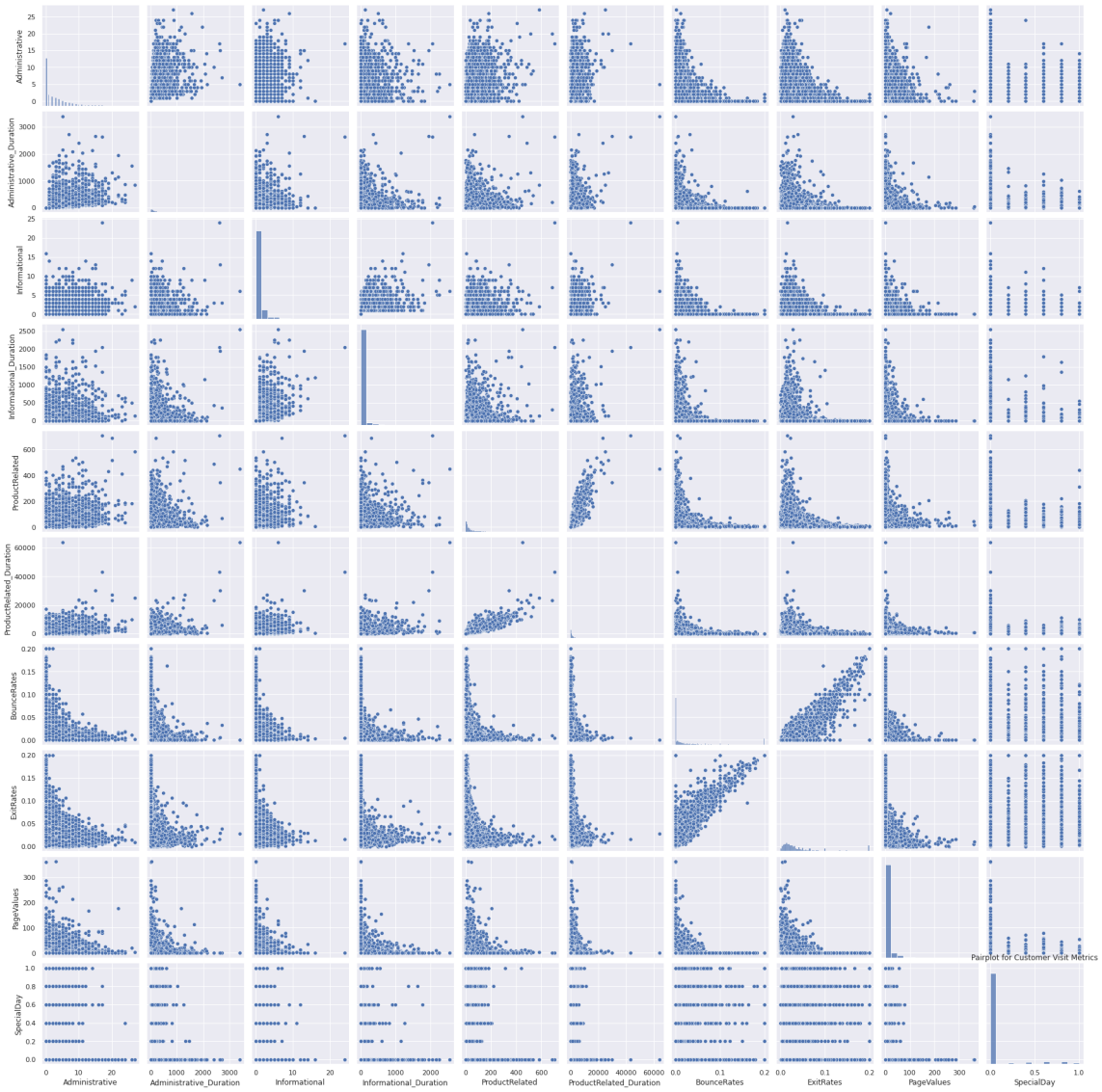


Fig xi: Pairplot for Customer Visit Metrics

Focussing on the below two charts from the above Pair Plot.

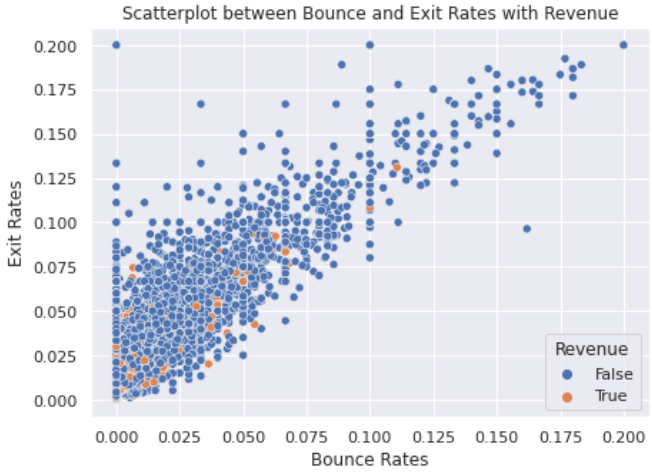


Fig xii (a)

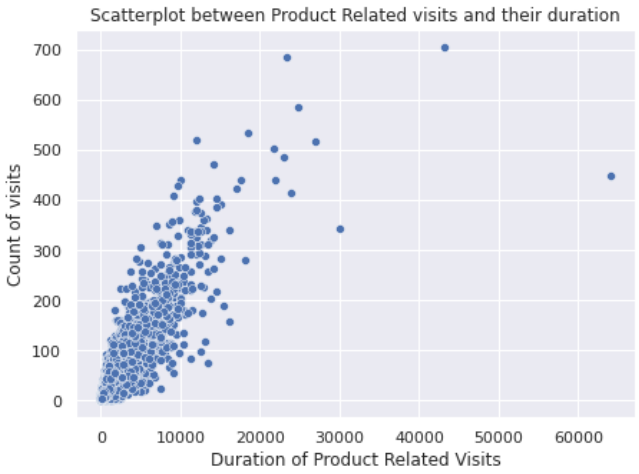


Fig xii (b)

Fig xii (a): Scatterplot of Bounce Rates and Exit Rates with Revenue
Fig xii (b): Scatterplot of Duration of Product related visits and their distribution

Revenue is generated only in specks of visits, with most revenue being generated when the bounce and exit rates are at the lower end. As the rates keep increasing, customers lose interest and eventually drop off the website without any purchase.

Also, as seen from the correlated matrix, it is validated from the scatterplots that the two columns “ProductRelated_Duration” and “ProductRelated” are highly correlated conveying similar information. Thus we shall drop the column “ProductRelated_Duration” while feeding the values into our model.

4. Exploration of Candidate Data Mining Models, and Select the Final Model or Models

4.a Data Mining Models

Data mining can be explained as a series of processes to derive trends and correlations through a large database using mathematical and statistical techniques. Such analyses guide in obtaining logical and insightful inferences and thus drive further actions based on them. Data Mining Models can be primarily categorized into Descriptive (Explanatory) models or Predictive models [*Data Mining for Business Analytics - Galit, Peter, Mia, Nitin*].

Our Project - "Predicting Customer Purchase Behavior" is a Binary Classification Model to analyze various parameters of a customer's online visit and classify whether it will result in revenue for the e-commerce website or not. As we have studied, for **Classification Problem**, we can adopt models like **Logistic Regression**, **Decision Trees**, **Naïve Bayes**, **K-Nearest Neighbour**. As our's is a classification model, we ***cannot use models like Multiple Linear Regression or Time Series Analysis***.

4.b Data Splitting

Data Splitting is the practice of dividing the dataset into training and validation (testing) sets so as to estimate the performance of the data mining model on new data. Thus, we expect the model to be fitted with known predictors shared with it during the training and then classify the target variable to the best of its ability on the new dataset which it has not encountered before.

After analyzing and visualizing the customer's online behavior, we are now ready to apply the various data mining models and classify the **target variable "Revenue"**. We first partition our dataset into a standard **70-30%**, that is, 70% of the data (**8631 records**) will be put into training the model, while the remaining 30% of the data (**3699 records**) will be used to test the model.

4.c Standardizing the data

We used StandardScaler() to homogenize the data over all records and scale all the features across the dataset as a preprocessing step before applying any model.

4.d Exploration of Models:

I. Logistic Regression

Logistic regression is widely used for binary classification problems where it models the probability of the target variable (Revenue in our case) based on the several input features. As required, our Revenue variable is binary in nature - revenue is made or not made.

It will thus model the probability estimation of revenue variable being equal to 1 i.e., "Revenue is generated" - based on the various customer behaviors captured.

Advantages:

i. **Interpretability**: Logistic Regression is straightforward with the results as its coefficients are nothing but independent variables' effect on the probability of the outcome variable.

ii. **Suitable for binary classification**: Our target variable is "Revenue", which is a binary outcome thus logistic regression is best suited for our problem statement.

iii. **Accommodates outliers:** It is suitable for a dataset with extreme values or outliers.

iv. **Smooth Decision Making**: It calculates the probability of the outcome variable "Revenue", will be useful for the e-commerce website to make a decision.

Disadvantages:

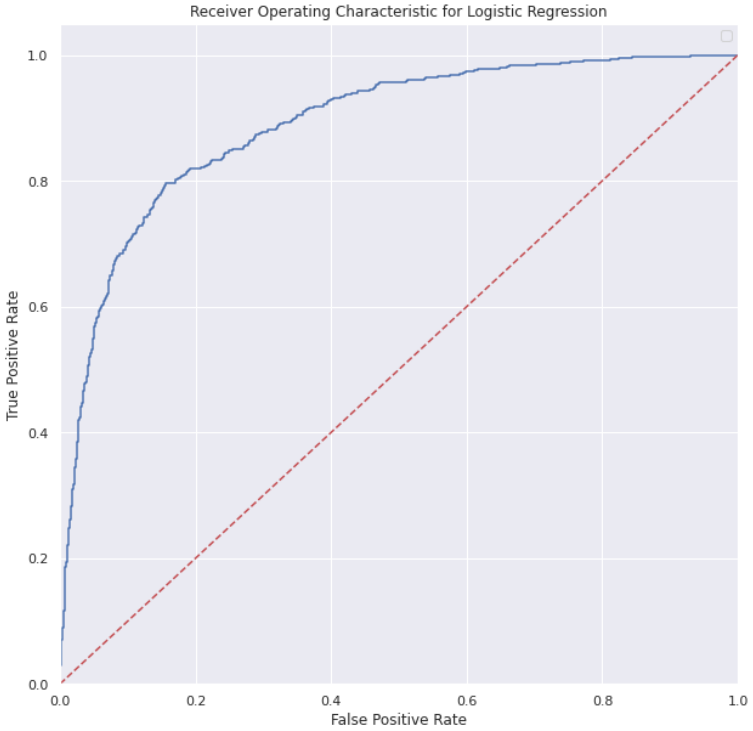
i. **Not suitable for non-linear dataset**: However, the dataset in our project has a linear relationship throughout, thus overriding the disadvantage.

ii. **Needs large dataset**: Our dataset comprises over 12,500 records, thus the model is reliable.

Confusion Matrix

3030	77
372	220

ROC Curve



Area under the Curve	0.6734
Accuracy score	0.8786
Error	0.1214
Sensitivity	0.9752
PPV	0.8906
F1 Score	0.4949

II. K-Nearest Neighbor

It is a classification algorithm where a newly introduced data point will be classified based on its similarity with the already existing clusters in its neighborhood. Here, the model will compare the distances between the new data point and every other K number of datapoints in its neighborhood and thus assign the category having highest frequency. The distance between the points will be calculated by the Euclidean distance formula.

Advantages:

i. **Flexible Model**: As KNN is a non-parametric algorithm it does not assume the distribution of the data. Thus, datasets having complexities across its attributes can also be modeled.

ii. **Functional for Nonlinear relationships**: KNN is suitable for both nonlinear and linear relationships across dependent and independent variables.

Disadvantages:

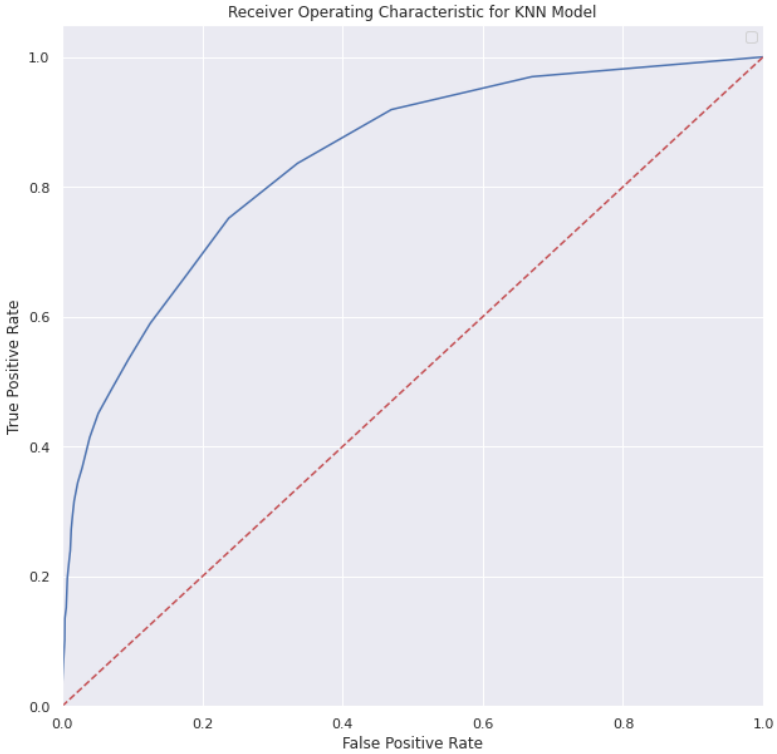
i. **High Computation**: As KNN uses the Euclidean distance between the newly introduced data point and already existing clusters, it is computationally expensive.

ii. **Sensitive to outliers**: Outliers will affect the classification of newly introduced points.

Confusion Matrix

3054	53
405	187

ROC Curve



Area under the Curve	0.6494
Accuracy score	0.8762
Error	0.1238
Sensitivity	0.9829
PPV	0.8829
F1 Score	0.4495

III. Naïve Bayes

Based on Bayes' Theorem, it is usually preferred to solve multi-class prediction problems. The "naive assumption" here is that each feature is independent of each other and calculates the conditional probability based on prior knowledge.

However, in our project several features share decent correlation amongst themselves which can be clearly seen in the less accuracy of the model as mentioned below.

Advantages:

i. **Time Efficient**: Adopts a fast algorithm which requires less time to train and quicker computational time.

ii. **Efficient with smaller datasets**: It considers independence of the features thus is faster with the process.

iii. **Accurate Fitting**: Can ignore features which are not relevant to the target variable and thus fits the model accurately.

Disadvantages:

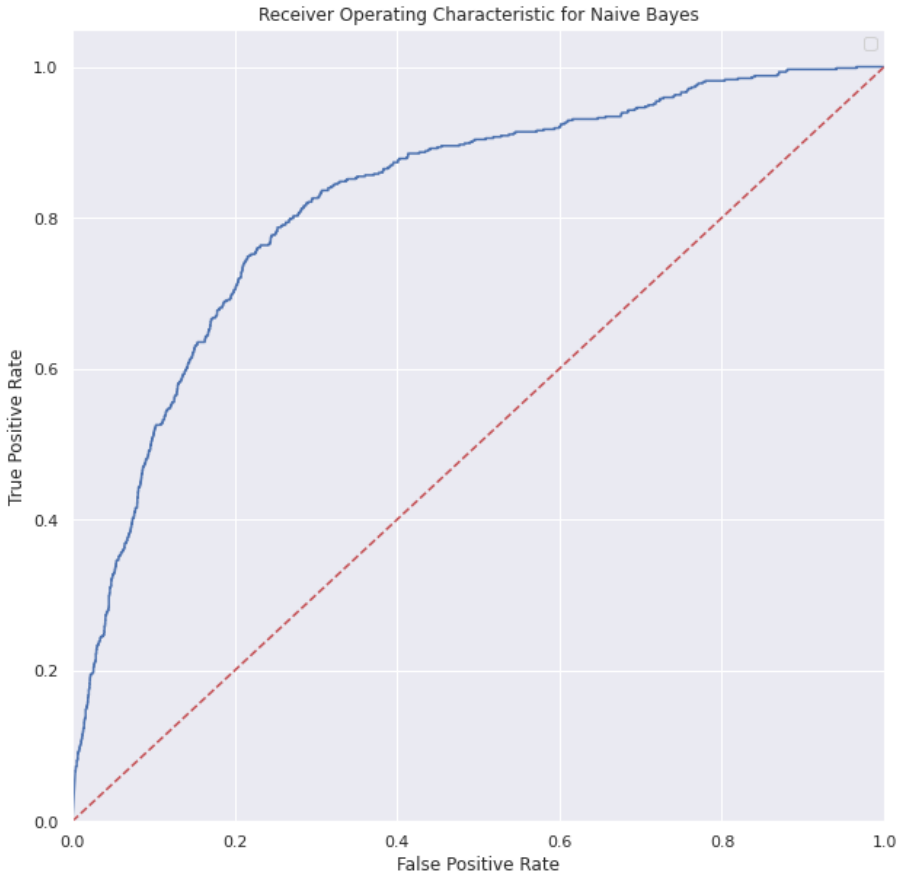
i. **Inaccurate predictions**: As it strongly assumes independence of the features, it may not hold true in all the real world applications.

ii. **Not fit for datasets with missing values**: Thus the dataset needs a series of data imputation procedures before proceeding.

Confusion Matrix

373	2734
8	584

ROC Curve



Area under the Curve	0.7466
Accuracy score	0.8018
Error	0.1981
Sensitivity	0.8278
PPV	0.9285
F1 Score	0.5180

IV. Decision Trees

A classification decision tree partitions the input data into several subsets of distinct target class labels. Decision tree selects a feature that best splits the data into subsets, where the information gain is maximum and impurity is least. This process is repeated until all subsets belong to the same variable groups. The purity of such a distinct subset is calculated by an impurity measure, called Gini index, which measures the degree of class heterogeneity.

Advantages:

i. **Flexibility**: Ability to handle both numerical and categorical data. To automatically classify various interactions between features.

ii. **Non-parametric**: Minimal to no assumptions about the distribution of data.

iii. **Minimal Data Cleaning**: Can effectively handle missing values & outliers in the input data.

iv. **Low computational cost**: Thus suitable for large datasets, with huge records and features.

Disadvantages:

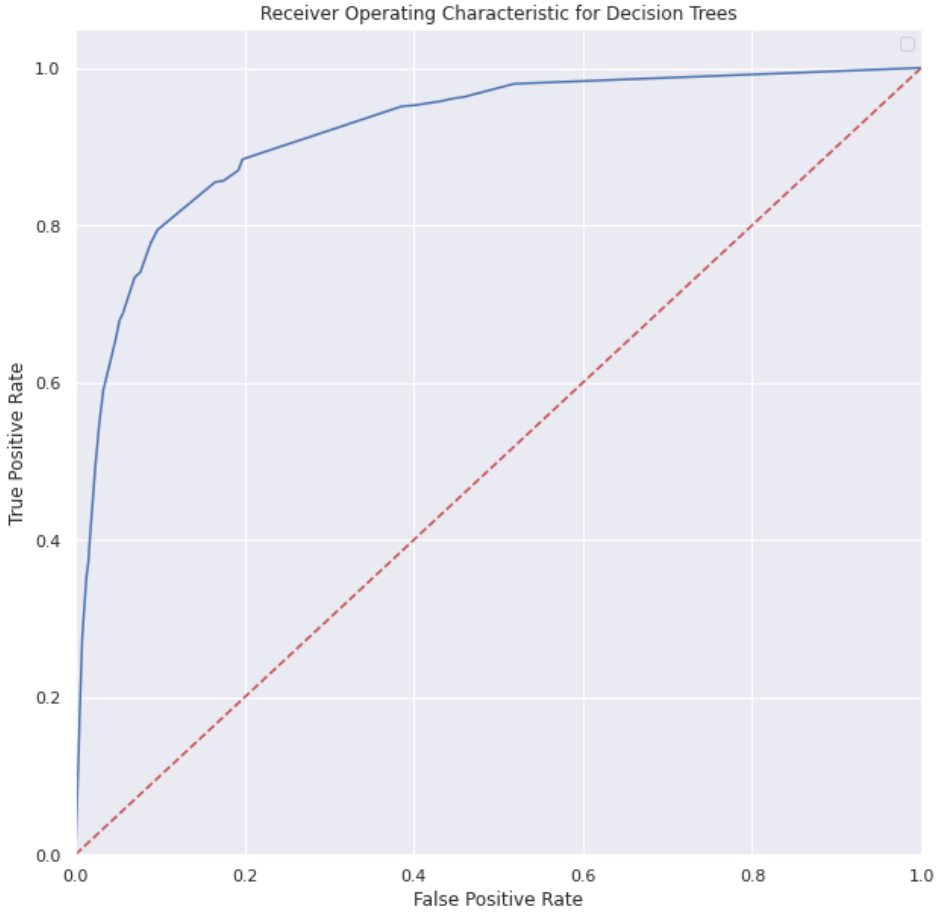
i. **Overfitting**: Being a complex algorithm due to the no. of iterations for each subset, an imbalance dataset will overfit the model.

ii. **Biased Model**: Categories with higher cardinality or multiple subcategories drive the model. Additionally, Decision trees can be satisfied with a local optimal solution rather than exploring a global optimal solution.

Confusion Matrix

3005	102
243	349

ROC Curve



Area under the Curve	0.7783
Accuracy score	0.9067
Error	0.0933
Sensitivity	0.9671
PPV	0.9248
F1 Score	0.6692

5. **Hyperparameter Tuning on Logistic Regression Parameters:**

Hyperparameter tuning is the process of finalizing an optimal set of parameters for ML algorithms which gives the best performance on a given task. For **logistic regression**, hyperparameters are parameters that are not learned from the data but rather set before training the model.

'newton-cg' is a solver algorithm to determine which algorithm to use in the optimization process to find the coefficients that **minimize the loss function**.

L2 penalty, or **Ridge regularization**, is used to **prevent overfitting of a model** by adding a penalty term to the loss function. In logistic regression, the L2 penalty is applied to the coefficients of the model.

The output of the Hyperparameter tuning was
Best: 0.997927 using {'C': 100, 'penalty': 'l2', 'solver': 'newton-cg'}

Modified & Improved Metrics After Hyperparameter Tuning:

Area under the Curve	0.6821
Accuracy score	0.8792
Error	0.1208
Sensitivity	0.9762
PPV	0.8910
F1 Score	0.4961

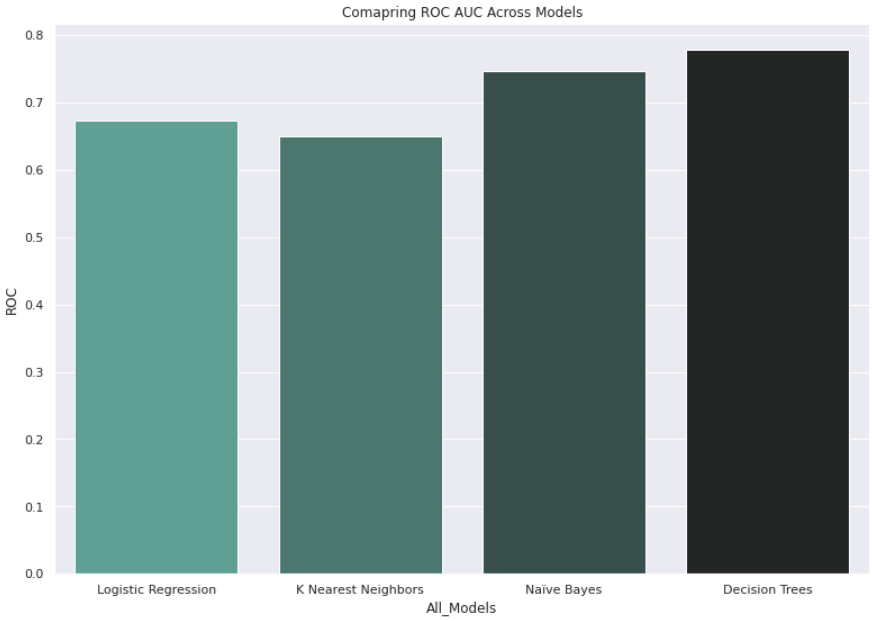
Summary

Sr. No	Model	AUC	Accuracy	Error	Sensitivity	PPV	F1 Score
1	**Logistic Regression**	0.6734	0.8786	0.1214	0.9752	0.8906	0.4949
2	**KNN**	0.6494	0.8762	0.1238	0.9829	0.8829	0.4495
3	**Naïve Bayes**	0.7466	0.8018	0.1981	0.8278	0.9285	0.5180
4	**Decision Trees**	0.7783	0.9067	0.093	0.9671	0.9248	0.6692
5	**Logistic Regression Tuned**	0.62810	0.8792	0.12080	0.9760	0.8910	0.4961

<u>Area under the Curve for every model's ROC Curve</u>

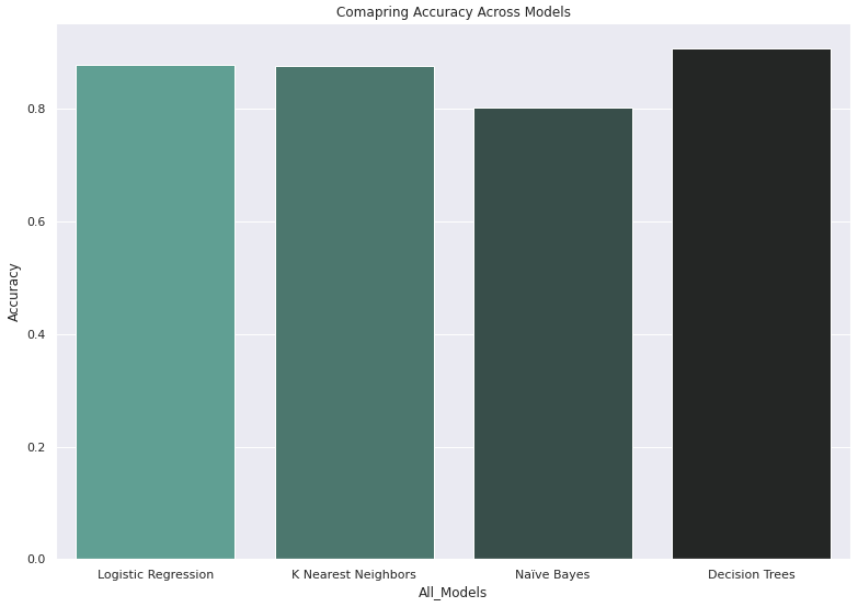
An **ROC curve** displays the relationship between the true positive rate (TPR) and the false positive rate (FPR) at various classification thresholds. The **area under the curve** calculates the model performance across all possible classification thresholds.

An AUC of 0.5 stipulates that the model's performance is as good as random guess, whereas an AUC of 1 indicates perfect classification performance.



Accuracy of all the Models

Accuracy of a model shows the proportion of observations that are correctly classified. In our problem, it is the percentage of cases where the model was successful to predict **whether a web session resulted in revenue or no**.

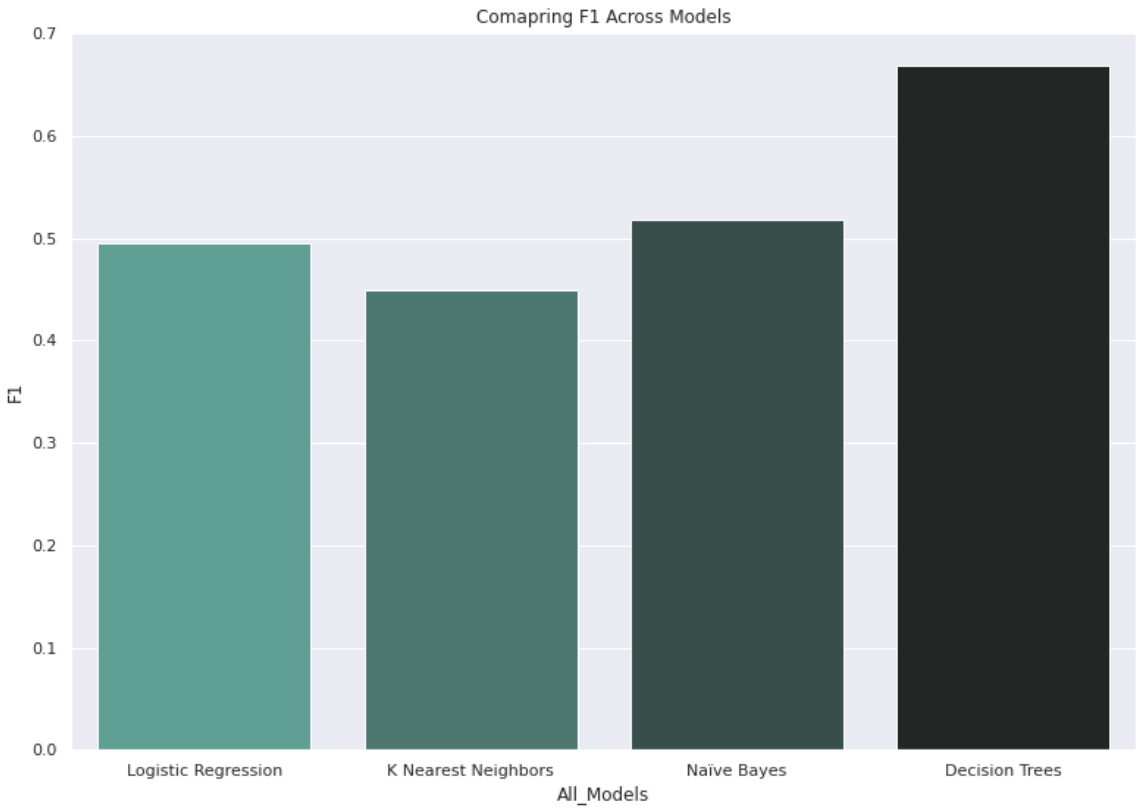


While the overall objective of our project is to identify **whether or not revenue was made in a web session**, based on the several input features, we are **more concerned to identify those web sessions that yield in revenue**. Revenue attribute as the target variable, has **only 18.31%** of records yielding in revenue, thus it is a **highly unbalanced dataset**.

Why are Sensitivity and PPV important?

Sensitivity calculates the ability of the model to detect the members of the class of interest from all the class of interest members. Similarly PPV calculates the proportion of correctly identified **class of interest members**. Therefore, it is crucial to give **importance to F1 score** which factors both **Sensitivity** and **PPV**.

$$\text{F1 Score} = 2\,x\,\frac{Sensitivity\ x\ PPV}{Sensitivity + PPV}$$



6. **<u>The Winning Model</u>**

After evaluating different models such as Logistic Regression, K-Nearest Neighbors, Naïve Bayes, and Decision Trees, we conclude that the **Decision Tree Model has the highest F1 Score**, making it our winning model. In real-world applications, our project can be useful for monitoring various numerical parameters, such as the time spent on each product-related page, bounce rates, and exit rates during a session. These parameters can help determine the likelihood of a customer making a purchase during their session. Additionally, by using the Decision Tree Model, we can monitor browser and regional performances while factoring in additional revenue generated during weekends or special-day offers.

Finally, the **Decision Tree Model** provides a credible and accurate way to analyze complex datasets such as Customer Purchase Behavior and thus make predictions based on different input parameters. Though we tested the models in the field of e-commerce, we can also extend the study to various fields such as healthcare, finance, and marketing.