UNIVERSITY OF TEXAS AT ARLINGTON

MS in APPLIED STATISTICS AND DATA SCIENCE

ASDS 6302 – MACHINE LEARNING AND APPLICATIONS

TEAM – 7

ONLINE SHOPPERS PURCHASE AND INTENTION

PRESENTED BY

HARIHARAN SELVAM (1002174644)

Table of contents

ONLINE SHOPPERS PURCHASE AND INTENTION

- 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.

PROBLEM STATEMENT:

- Identify the factors that will help to improve the customers count on an online shopping site.

DATASET OVERVIEW:

- Source: https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset
- The dataset has 12330 observations with 18 features.
- Out of these 18 features, 14 are numerical and 4 are categorical.

INFORMATION ABOUT ATTRIBUTES:

- Administrative, Administrative_Duration, Informational, Informational_Duration, ProductRelated, ProductRelated_Duration- number of different types of pages visited by visitor in that session and total time spent in each of these page categories
- BounceRates - Entering site and exits without any impact
- ExitRates - Chances of site being the last page they visited
- PageValues- Average value of web page user visited before completing an e-commerce transaction
- SpecialDay, - Special Day or event
- OperatingSystems , Browser , Region, TrafficType – Servers
- Month - Months of purchase
- Visitor Type - Type of visitors
- Weekend - Purchase of weekends
- Revenue- Revenue of the site.

DATA PREPROCESSING:

- Data preprocessing is the process of converting a raw data into understandable format

Dropping duplicates:

- After dropping duplicates, no. of observations got reduced to 12205.

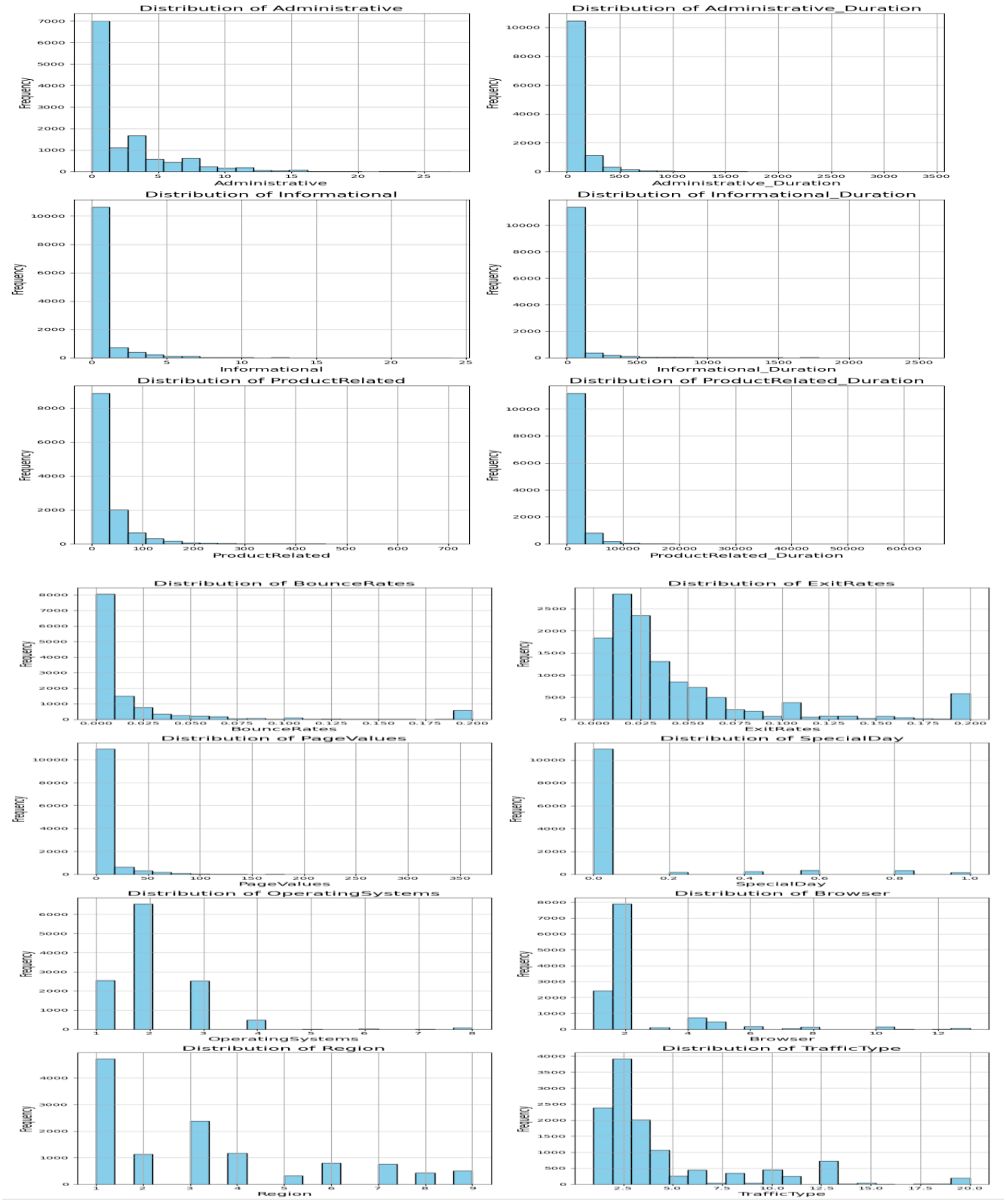
Checking for Missing values:

- No missing values in this dataset.

List of Numerical Variables:

- Administrative', 'Administrative_Duration', 'Informational', 'Informational_Duration', 'ProductRelated', 'ProductRelated_Duration', 'BounceRates', 'ExitRates', 'PageValues', 'SpecialDay', 'OperatingSystems', 'Browser', 'Region', 'TrafficType

Distribution of Numerical Variables:

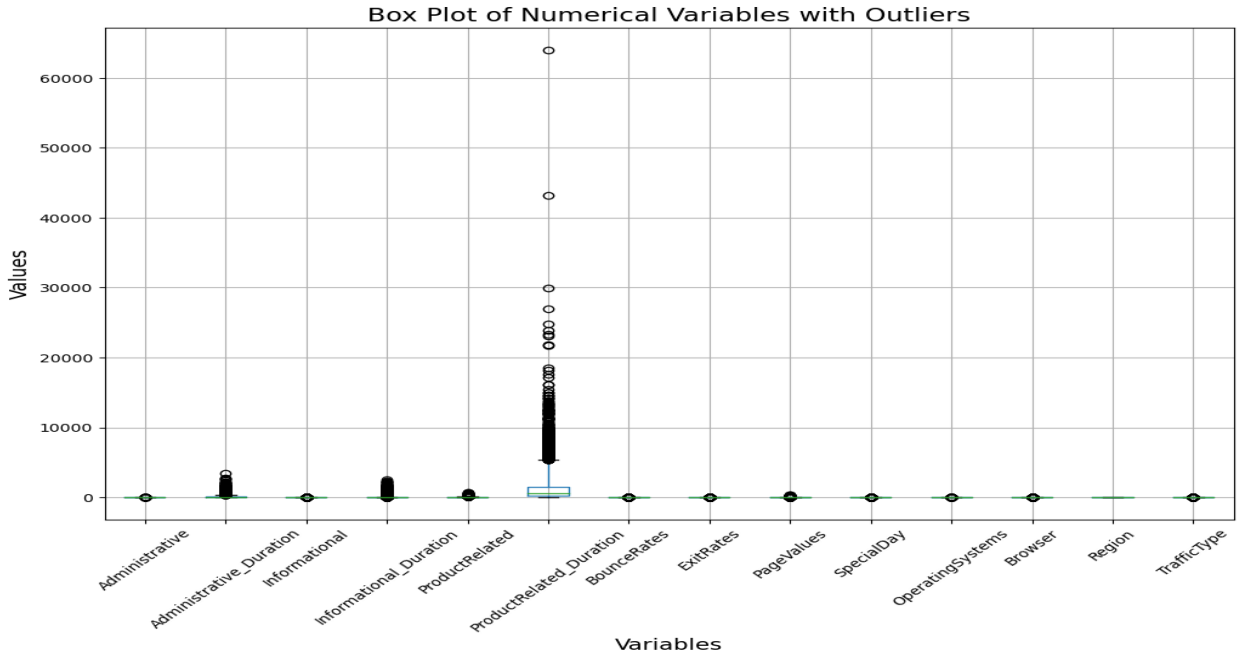


List of Categorical variables:

- 'Month', 'VisitorType', 'Weekend', 'Revenue'.
- Cardinality of Categorical variables:
- Month have 10 unique values such as February, March, April, May, June, July, August, October, November, December.
- VisitorType have 3 unique values such as Returning visitor, New visitor, Other.
- Weekend have 2 unique values such as Yes, No.
- Revenue have 2 unique values such as Yes, No.

Detection of Outliers:

- I have found some outliers in my dataset.
- Here is the boxplot to show my outliers



- After removing outliers, no. of observations got reduced to 7667.
- Here is the update of my cleaned data.

  Rows: 7667

  Columns: 18

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7667 entries, 185 to 12329
Data columns (total 18 columns):
 #    Column                   Non-Null Count  Dtype
---   ------                   --------------  -----
 0    Administrative           7667 non-null   float64
 1    Administrative_Duration  7667 non-null   float64
 2    Informational            7667 non-null   float64
 3    Informational_Duration   7667 non-null   float64
 4    ProductRelated           7667 non-null   float64
 5    ProductRelated_Duration  7667 non-null   float64
 6    BounceRates              7667 non-null   float64
 7    ExitRates                7667 non-null   float64
 8    PageValues               7667 non-null   float64
 9    SpecialDay               7667 non-null   float64
 10   Month                    7667 non-null   object
 11   OperatingSystems         7667 non-null   float64
 12   Browser                  7667 non-null   float64
 13   Region                   7667 non-null   float64
 14   TrafficType              7667 non-null   float64
 15   VisitorType              7667 non-null   object
 16   Weekend                  7667 non-null   bool
 17   Revenue                  7667 non-null   bool
dtypes: bool(2), float64(14), object(2)
memory usage: 1.0+ MB
```

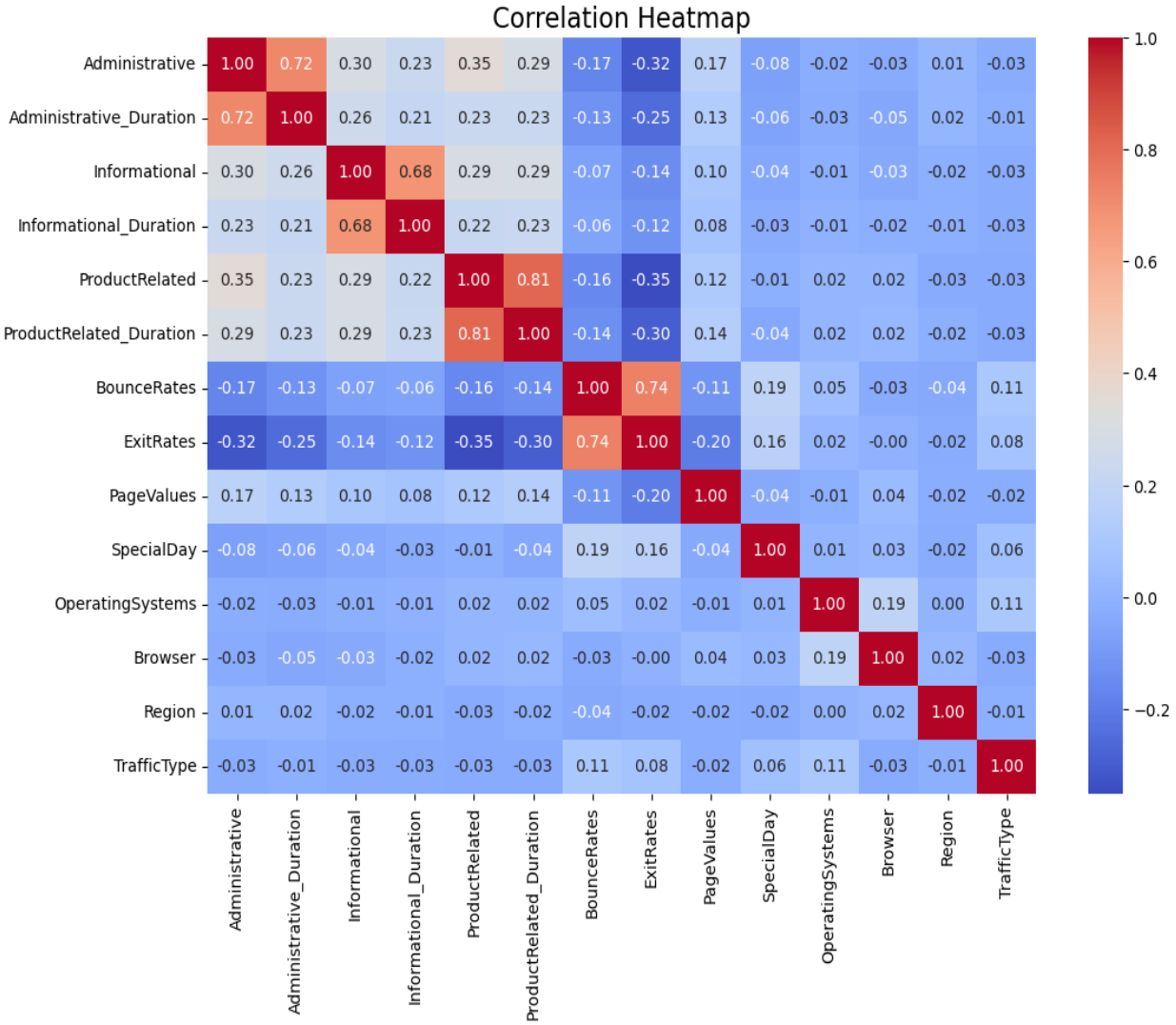
- My cleaned dataset does'nt have any null values, missing values, outliers

Exploratory Data Analysis:

- Exploratory Data Analysis (EDA) is an analysis approach that identifies general patterns and insights in data.

Correlation Heatmap:

- Correlation heatmap is created for numerical variables to gain insights.
- Here is my correlation heatmap



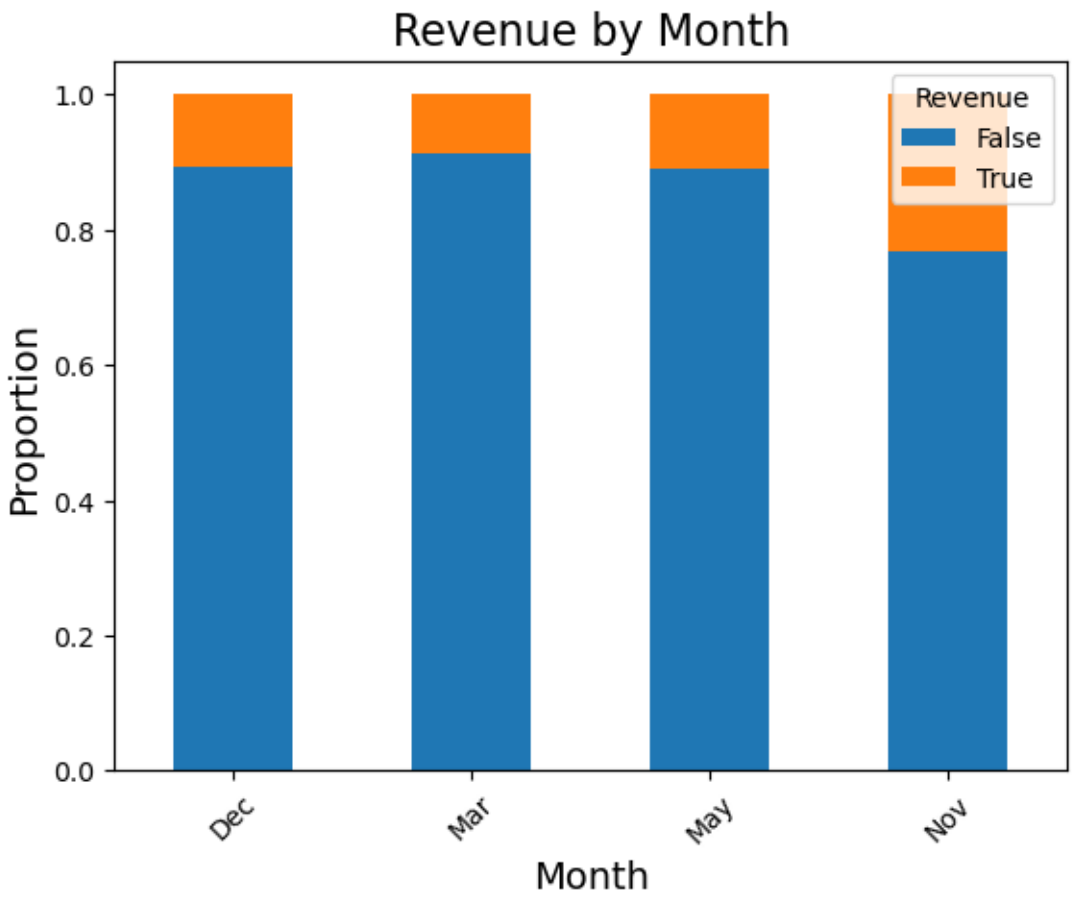
- With the help of correlation matrix, I got to know that
- There is a strong positive correlation between
  - Administrative and Administrative_Duration
  - ProductRelated and ProductRelated_Duration.
  - Bounce Rates and Exit Rates
- There is a moderate positive correlation between
  - Informational and ProductRelated
  - Administrative and Informational
  - Administrative and ProductRelated
  - Administrative and PageValues
- There is a weak negative correlation between
  - ProductRelated and Bounce Rates
  - ProductRelated and Exit Rates
  - Informational and Bounce Rates

- Informational and Exit Rates

Stacked Bar plot:

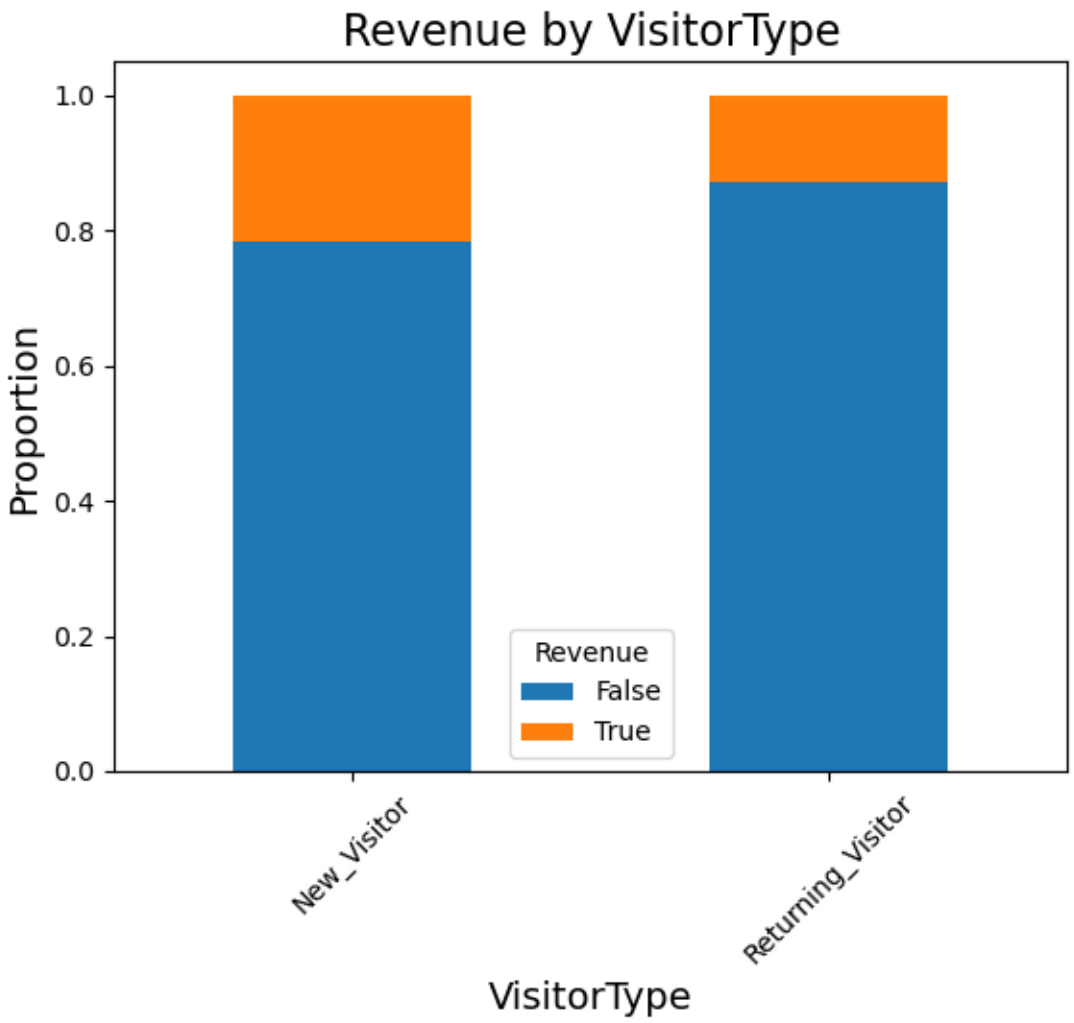
- A stacked bar plot is a type of chart that represents data in rectangular bars, with each bar subdivided into segments that represent different categories or groups.

- Relationship between revenue and month



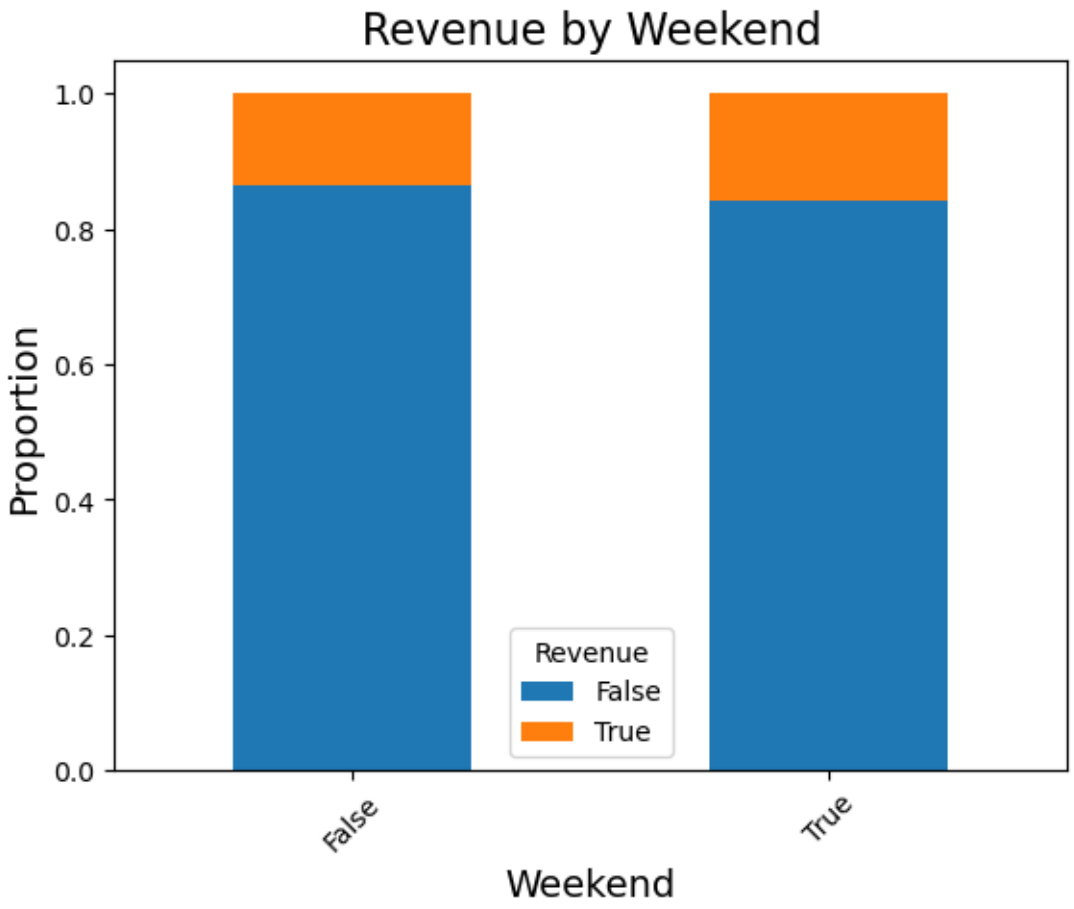
- Highest revenue is from March. Lowest revenue is from November.

Relationship between visitorType and revenue



- High revenue got produced with the help of returning visitors than new visitor

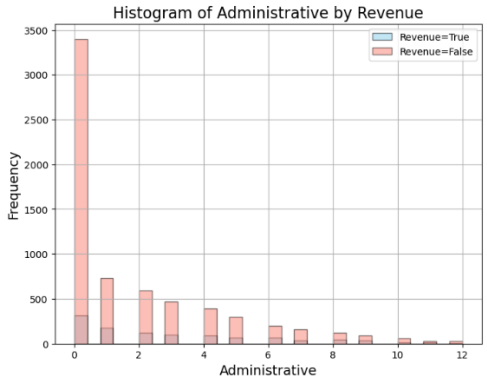
Relationship between weekend and revenue

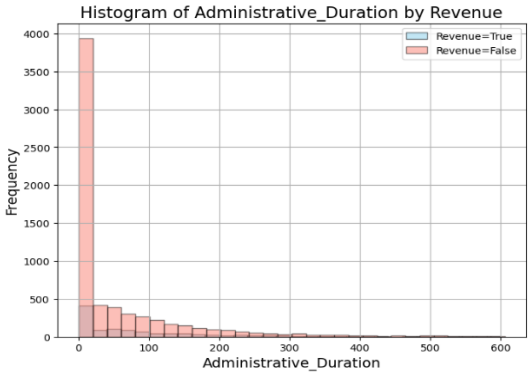


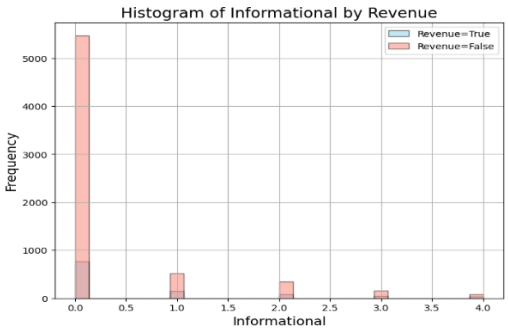
- Both false and true condition almost have similar results.

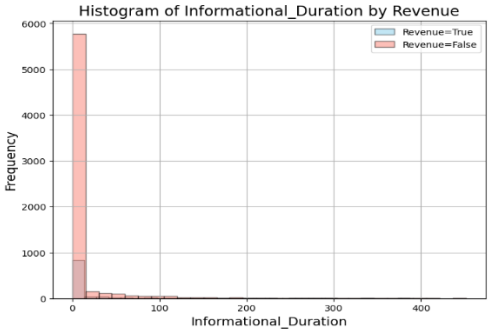
Histogram:

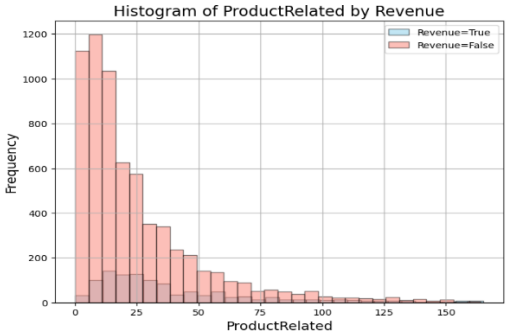
- A histogram is a graphical representation of the distribution of numerical data.
- Histogram to show relationship between numerical variables and revenue

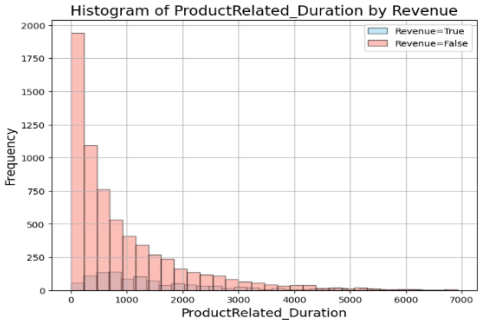


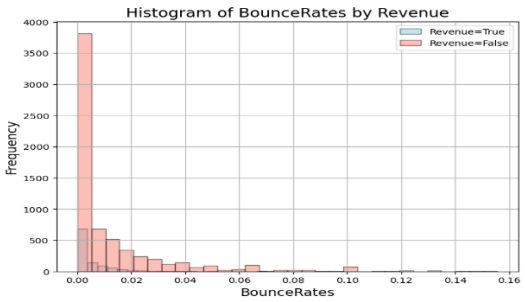


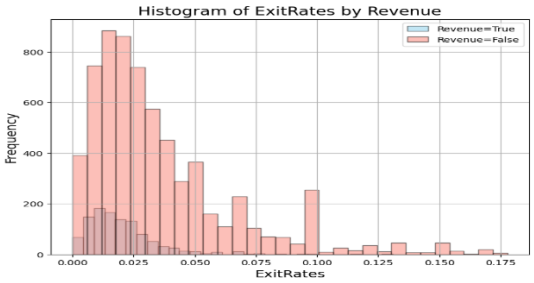


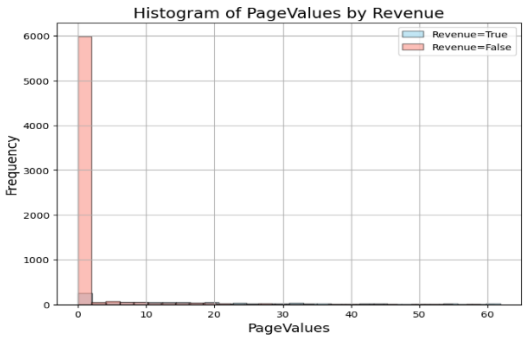


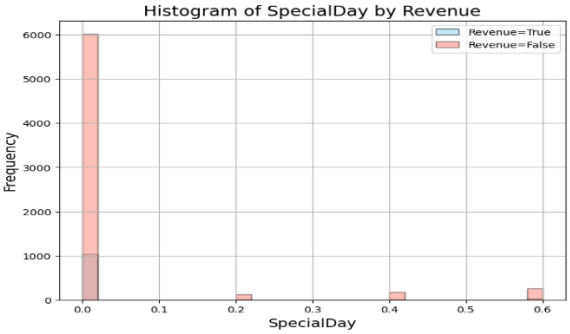












Valuable Insights:

- From the Exploratory Data Analysis, I got some insights.
- websites with high informational content tend to also have users who spend more time on informational pages
- websites with more product-related content might have slightly lower bounce rates and exit rates.
- websites with high bounce rates also tend to have high exit rates

- website features might have a small positive impact on user behavior.type of traffic a website receives might influence the administrative tasks associated with managing that traffic
- Returning visitors help for revenue more than new visitors, which is one of the positives for the websites.

MODEL BUILDING:

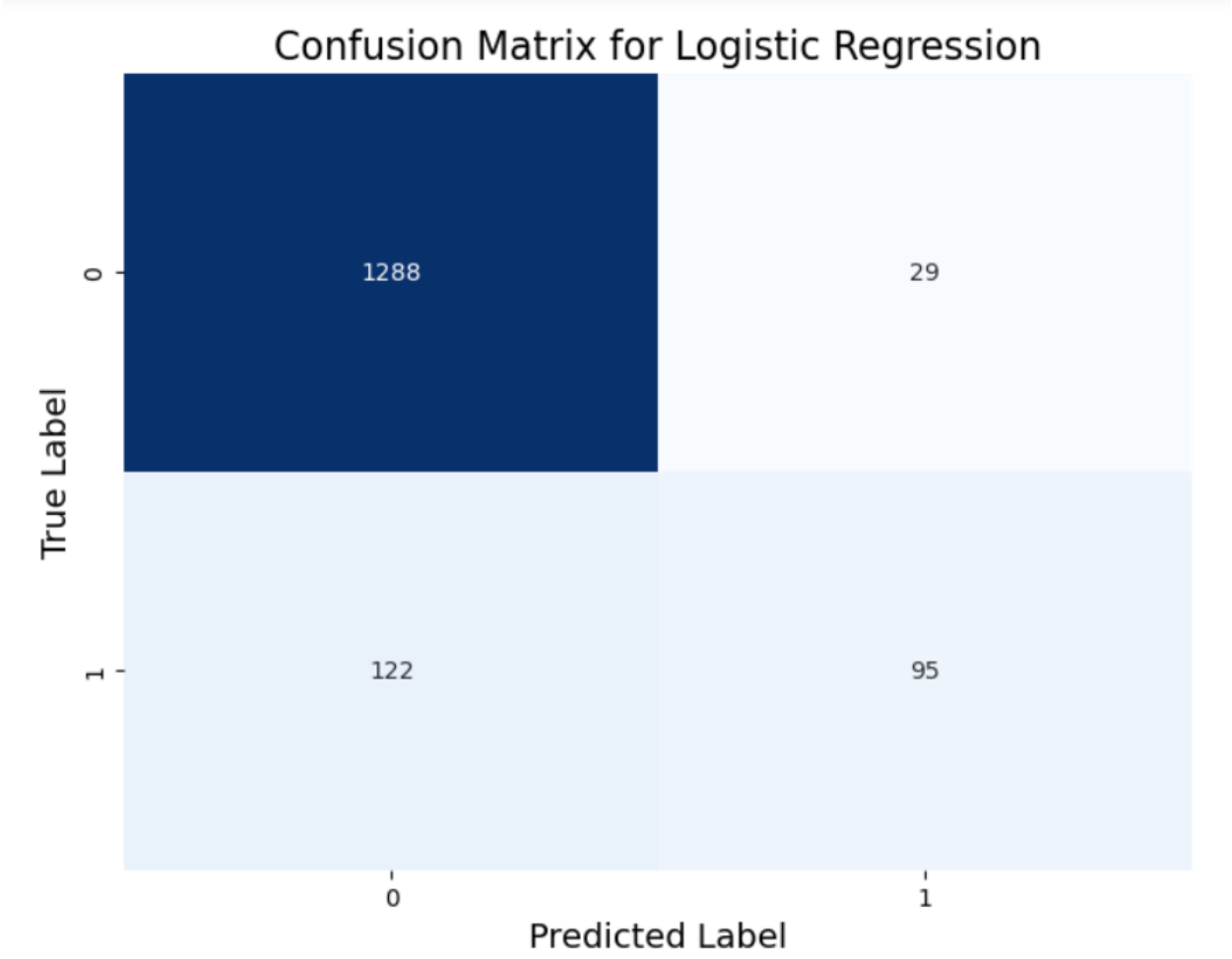
- Here, I'm going to build Logistic Regression and Support Vector Machine Classification models.
- Target :Revenue
- Features: All other attributes.
- Train – test ratio: 70:30

Logistic Regression:

- Logistic regression is a statistical model used for binary classification tasks, where the target variable has two possible outcomes or classes.
- Here is the evaluation of my Logistic Regression model

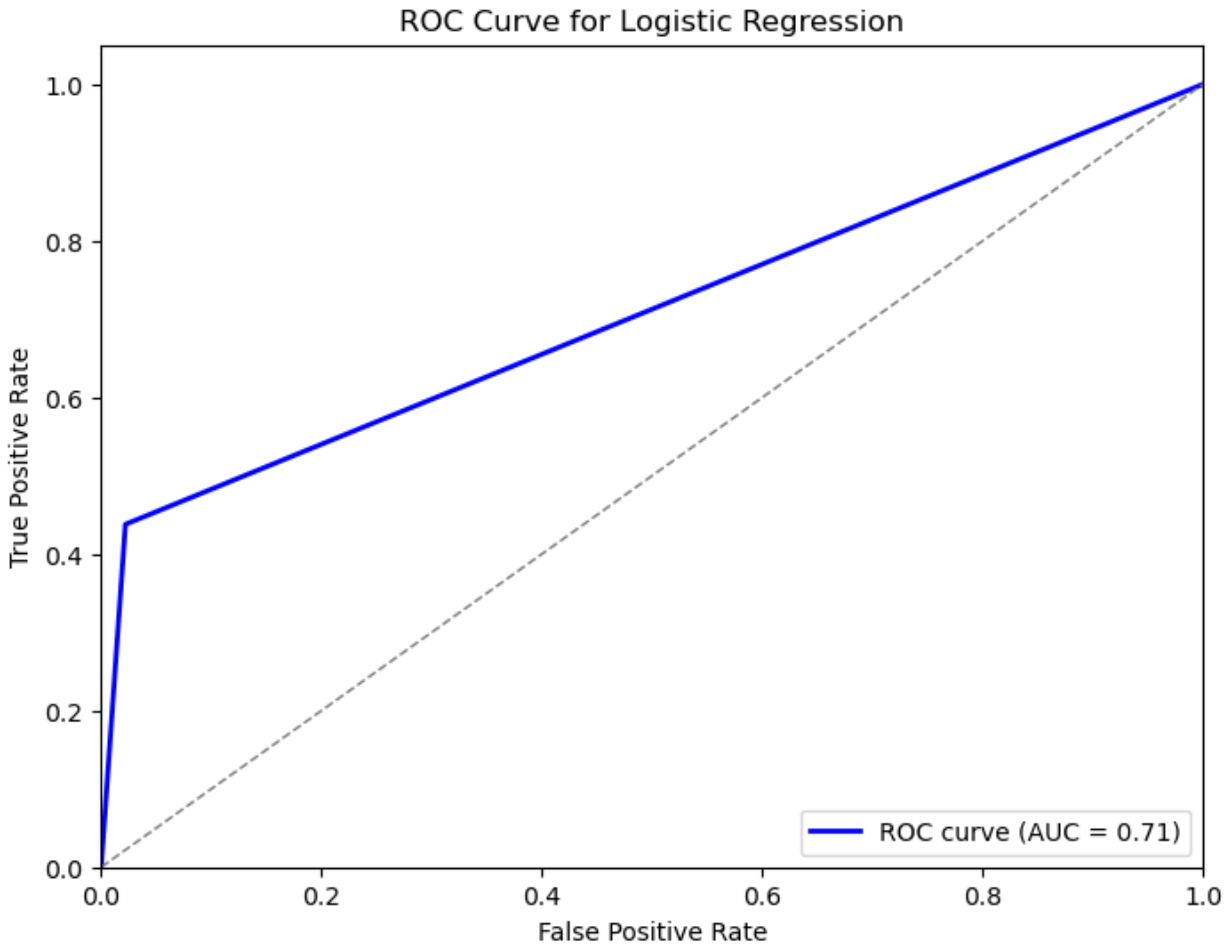
```
Logistic Regression Accuracy: 0.90
Classification Report for Logistic Regression:
              precision    recall  f1-score   support

       False       0.91      0.98      0.94      1317
        True       0.77      0.44      0.56       217

    accuracy                           0.90      1534
   macro avg       0.84      0.71      0.75      1534
weighted avg       0.89      0.90      0.89      1534

Confusion Matrix for Logistic Regression:
[[1288   29]
 [ 122   95]]
```

- Here is the heatmap for my confusion matrix



- 1288 – Correctly classified True Positive
- 29 – Incorrectly Classified False Positive
- 122- Incorrectly Classified False Negative
- 95 – Correctly classified True Negative

Here is the ROC curve with AUC value for Logistic Regression



- Here, I got the ROC curve with AUC value of 0.71
- Which proves that the model has a pretty good performance on the dataset.

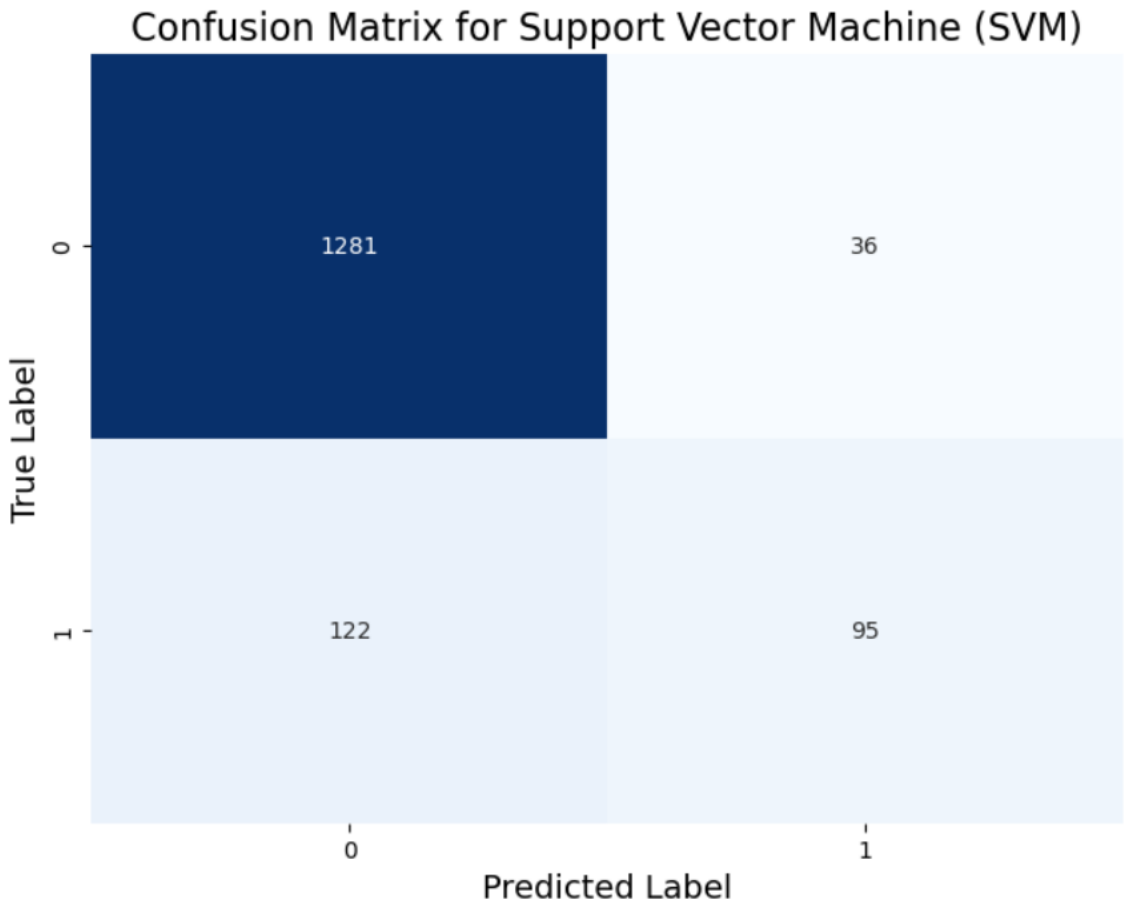
Support Vector Machines:

Support Vector Machines (SVMs) are supervised learning models used for classification and regression tasks.

Here is the evaluation of my SVM model

```
Support Vector Machine (SVM) Accuracy: 0.90
Classification Report for Support Vector Machine (SVM):
              precision    recall  f1-score   support

       False       0.91      0.97      0.94      1317
        True       0.73      0.44      0.55       217

    accuracy                           0.90      1534
   macro avg       0.82      0.71      0.74      1534
weighted avg       0.89      0.90      0.89      1534

Confusion Matrix for Support Vector Machine (SVM):
[[1281   36]
 [ 122   95]]
```

Here is the heatmap for confusion matrix



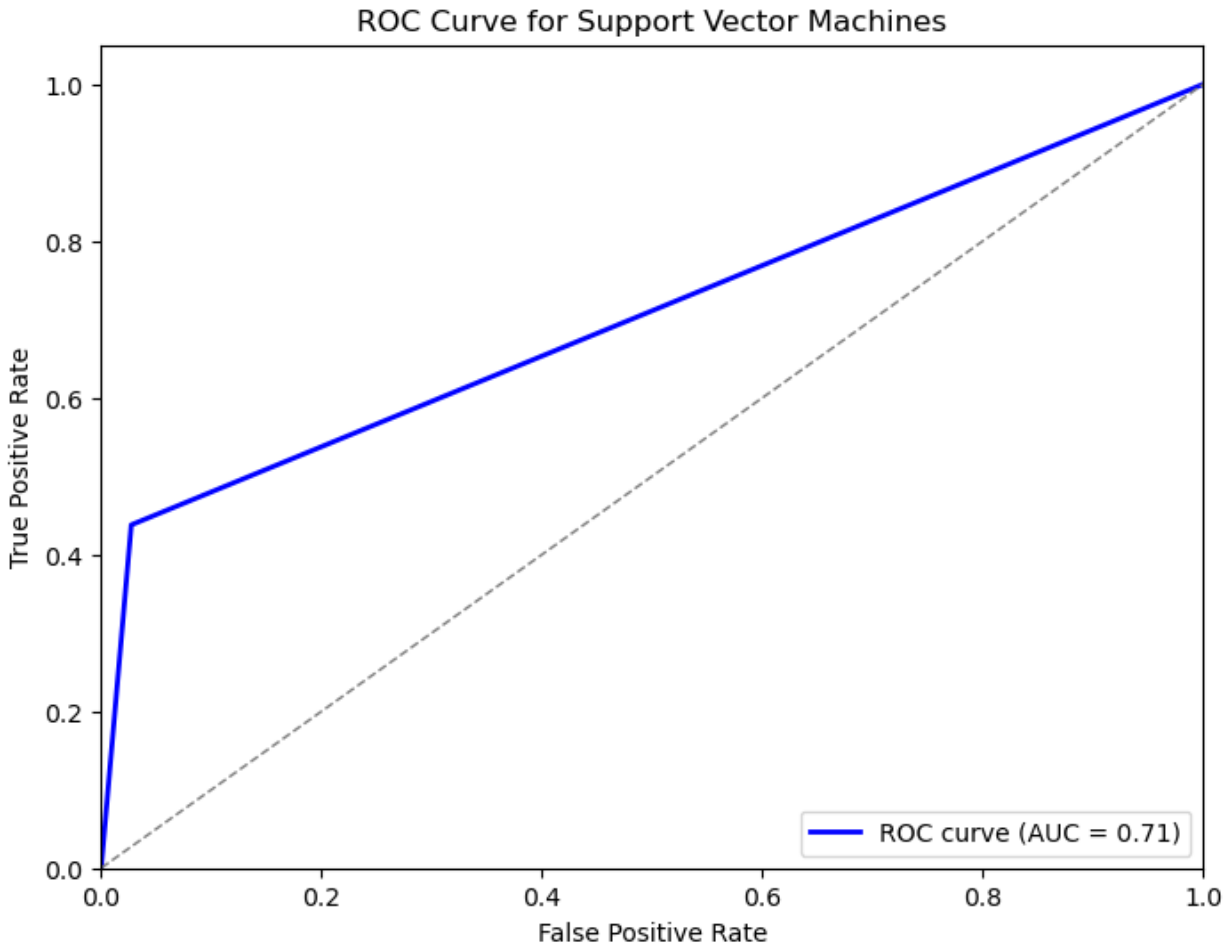
1281 – Correctly classified True Positive

36 – Incorrectly classified False positive

122-Incorrectly classified False negative

95 – Correctly classified True negative

Here is the ROC curve with AUC value for SVM model



Here, I got an ROC curve with AUC value of 0.71

Conclusion:

- Both Logistic Regression and Support Vector Machines perform in a similar manner.
- But I would like to go with Logistic Regression as my best fitted model
- Pagevalues, visitortype, special day, month, weekend will help to improve revenue for online shopping site.

- These features capture various aspects of user behaviour, engagement and timing which will be crucial for marketing strategies.

Future Recommendations:

- Leverage Recommendation engines.
- Seasonal promotions and discounts
- Time based promotions
- SEO and Social Media Advertising
- Landing page optimization
- Technical operation and user interface
- Newcomer and referral discounts
- Personalized e-mail retargeting

Reference:

Informatory paper:

https://www.semanticscholar.org/paper/Real-time-prediction-of-online-shoppers%E2%80%99-purchasing-Sakar-Polat/747e098f85ca2d20afd6313b11242c0c427e6fb3