Analysing and Predicting Online Purchase Behaviour with Machine Learning.

Solomon OKORO

github.com/soulsuv/analysing-online-purchase-with-ML

OBJECTIVES

- Analyse the E-commerce site data to determine important factors influencing customer purchase.
- Perform detailed Exploratory Data Analysis for missing values & outlier removal.
- Perform Feature Selection for the variables to be used in the Machine Learning Models.
- Predict the purchase tendency of the users with different Machine Learning Models and compare the Models' performance.

DATA SUMMARY

Dataset has **14,731** rows and **22** columns.

Page Visits and Duration: Number of times the Home page, Landing page and Product Description pages were visited with their respective durations.

Google Metrics: Bounce rate (% of users who visit one page and exit without exploring other pages), Exit rate (Percentage of users who exit the website after exploring multiple pages, excluding the first page) and Page value (Average value for a page that a user visited before landing on the goal page or completing an e-commerce transaction).

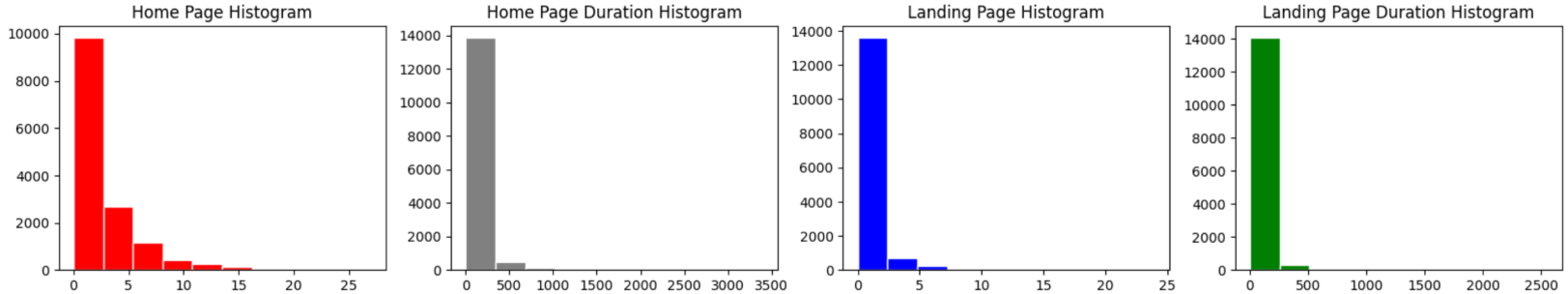
Seasonal Purchase and Month of Purchase: Indicator to track seasonal purchases and respective month of purchase. Also, weekend purchase tracker.

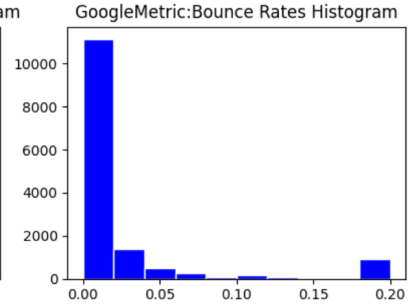
User demographics: Zone (region), Gender, Education, Visitor Type and Marital Status.

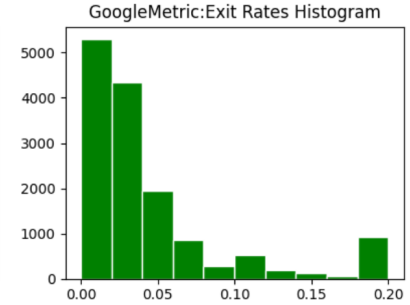
Device information: Cookies, Search Engine and Operating System.

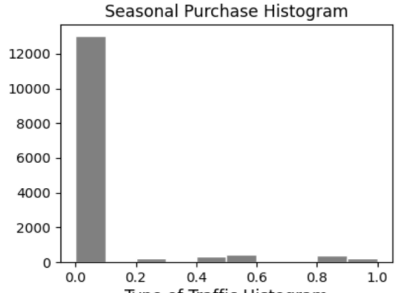
Made Purchase: True or False for Purchase made (The Target Variable).

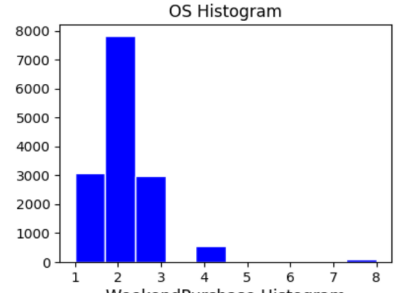
Overview of Numeric Data

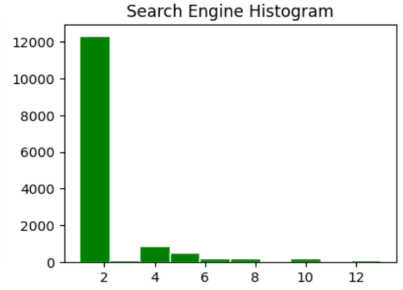


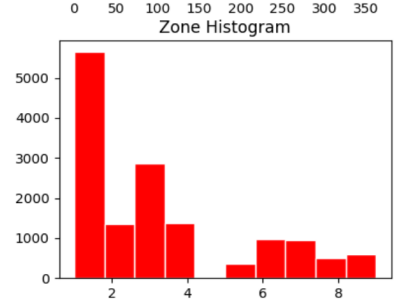


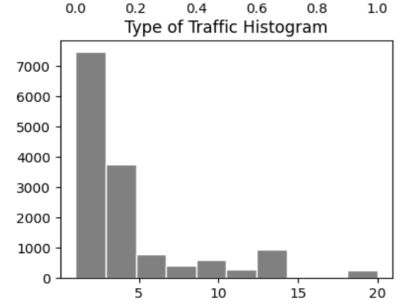


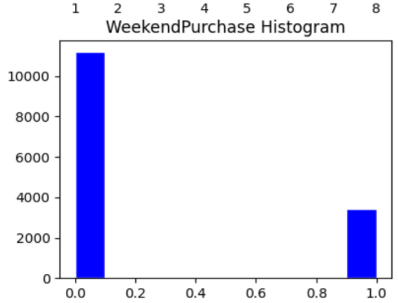








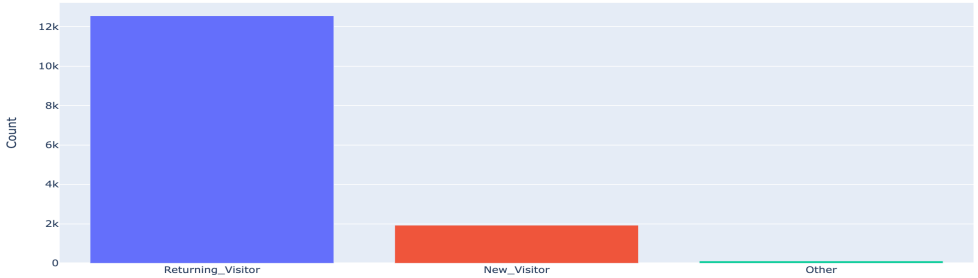




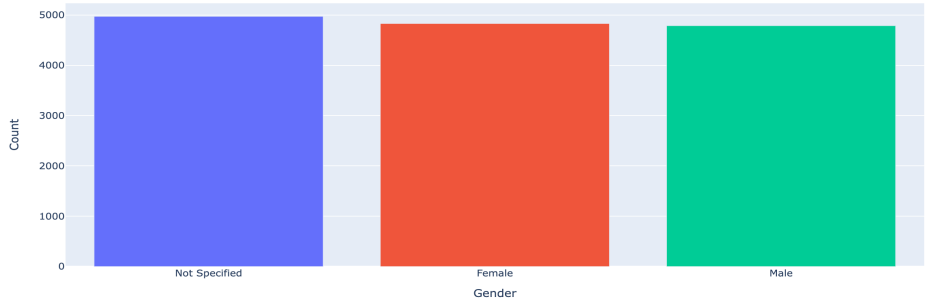
- Continuous data can be described by its median as the dataset is skewed.
- Encoded categorical data can be described by its Mode.
- Therefore, Median and Modes would be used for Imputation.

Overview of String Data

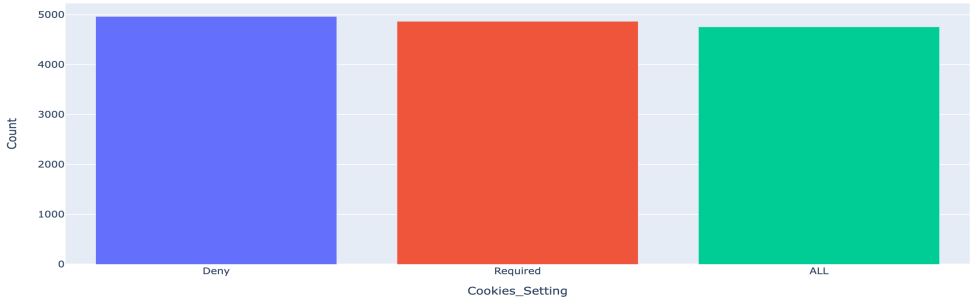
Customer Type Counts



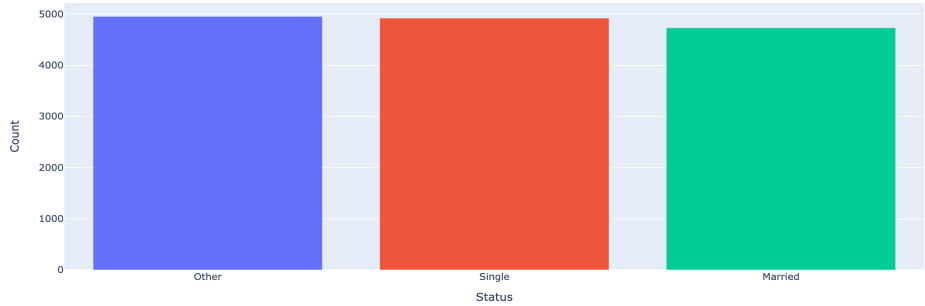
Gender Counts



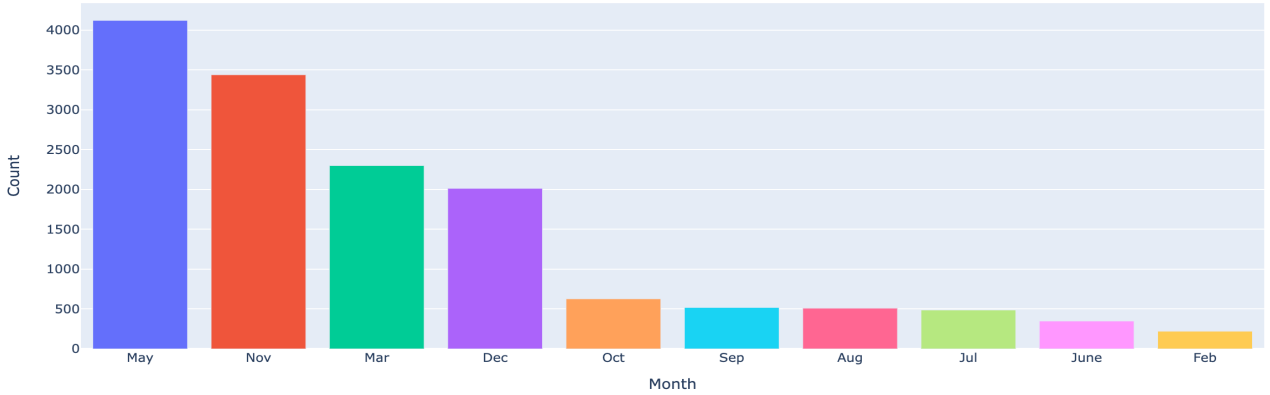
Cookies Setting Type Counts



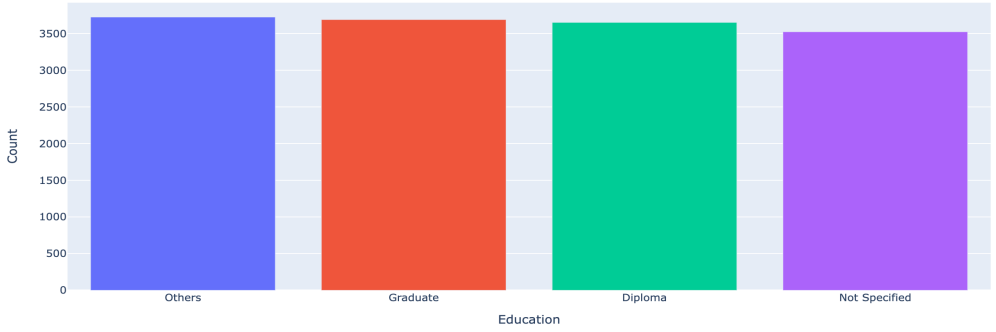
Marital Status Counts



Month Seasonal Purchase Counts



Education Status Counts



Most of the String Data are evenly distributed except for Month and Customer Type.

EDA and Pre-processing

- Null values in each column ranged from 0% to 1.13% with the Nulls Missing Completely at Random.
- Deleting all rows with Null values reduced the rows from 14,731 to 11,808 (~20% reduction). This is above the 10% threshold for dropping Null Values.

- Following the analysis of the numeric data, deployed imputation to fill Null values with the median and modes as described.
- After filling numeric values, dropped remaining nulls coming from string data. Rows reduced to 13,888 (~6% reduction).

- Label Encoding for converting String Data to Numerical data

```
columns_to_encode = [
'Month_SeasonalPurchase', 'Customer Type', 'Gender',
'Cookies Setting', 'Education', 'Marital Status', Made
Purchase'
]

labelencoder = LabelEncoder()

for column in columns_to_encode:
data[column] = labelencoder.fit_transform(data[column])
```

Outlier Detection & Scaling

- Inter Quantile Range (IQR) of the Numeric Data

```
HomePage                          3.000000
HomePage_Duration                90.025000
LandingPage                       0.000000
LandingPage_Duration              0.000000
ProductDescriptionPage           30.000000
ProductDescriptionPage_Duration  1248.398333
GoogleMetric:Bounce Rates         0.017647
GoogleMetric:Exit Rates           0.035417
GoogleMetric:Page Values          0.000000
SeasonalPurchase                  0.000000
Month_SeasonalPurchase            2.000000
Zone                              3.000000
CustomerType                      0.000000
Gender                            2.000000
Cookies Setting                   2.000000
Education                         3.000000
Marital Status                    2.000000
Made_Purchase                     1.000000
dtype: float64
```

- As can be observed, some of the variables had an IQR of 0 which implies the boxplot method of removing outliers would not suffice for those. Used the IQR method twice for `['HomePage_Duration', 'ProductDescriptionPage_Duration']`

- For `LandingPage_Duration & GoogleMetric:Page Values:`
- Deleted the outliers beyond the 99th percentile.

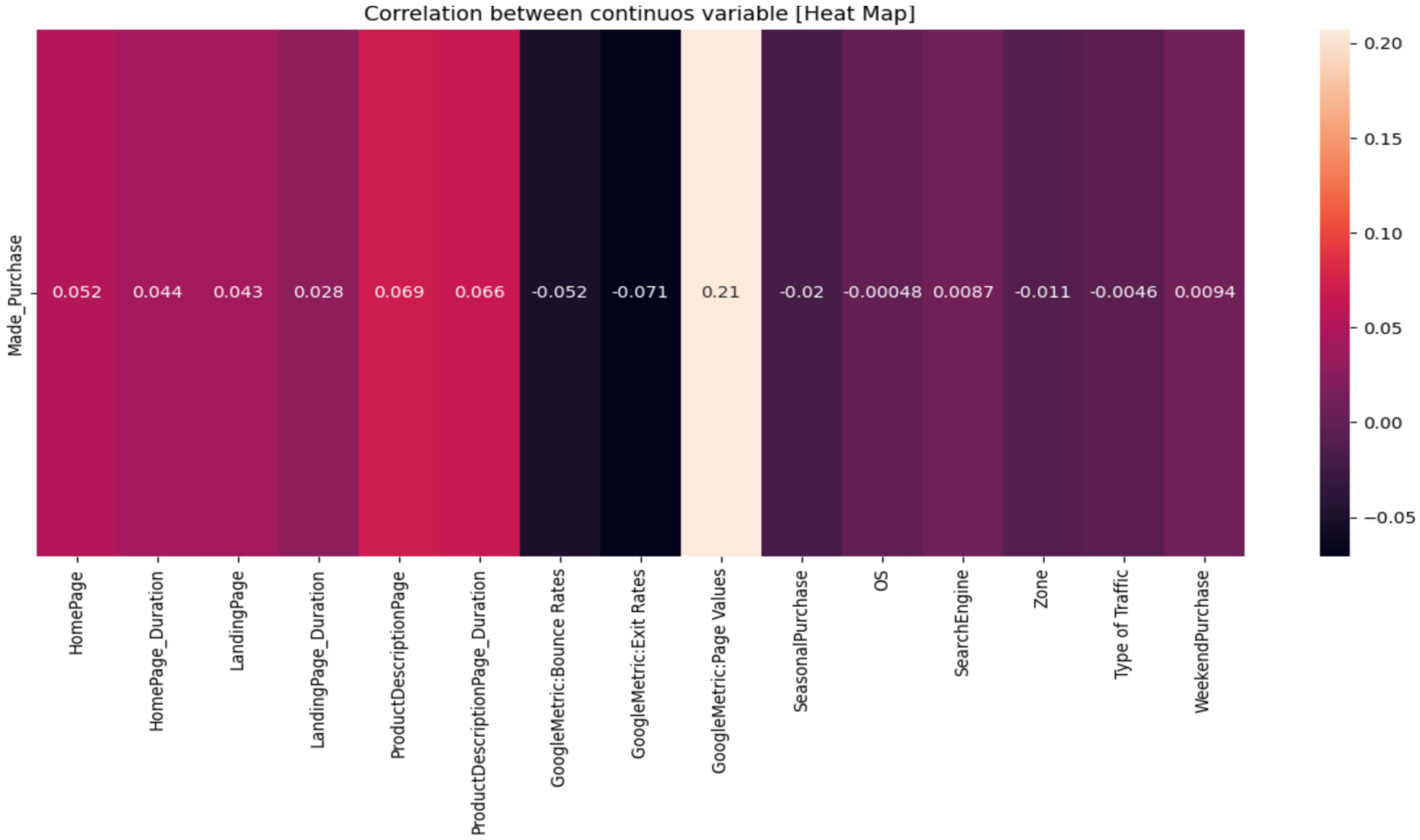
- Final dataset: 9954 rows

- Scaling the dataset

`scaler = RobustScaler()` : for duration numeric data.

`scaler = MinMaxScaler()` : for other numeric data.

Feature Selection (Continuous to Target)



- The Google Metric for Page Values has the highest correlation with the Target (Made Purchase). The low correlations to be dropped are OS type, Search Engine, Type of Traffic and Weekend Purchase. Also dropped the Landing Page Duration column.

Machine Learning (Final Data Set & ML models)

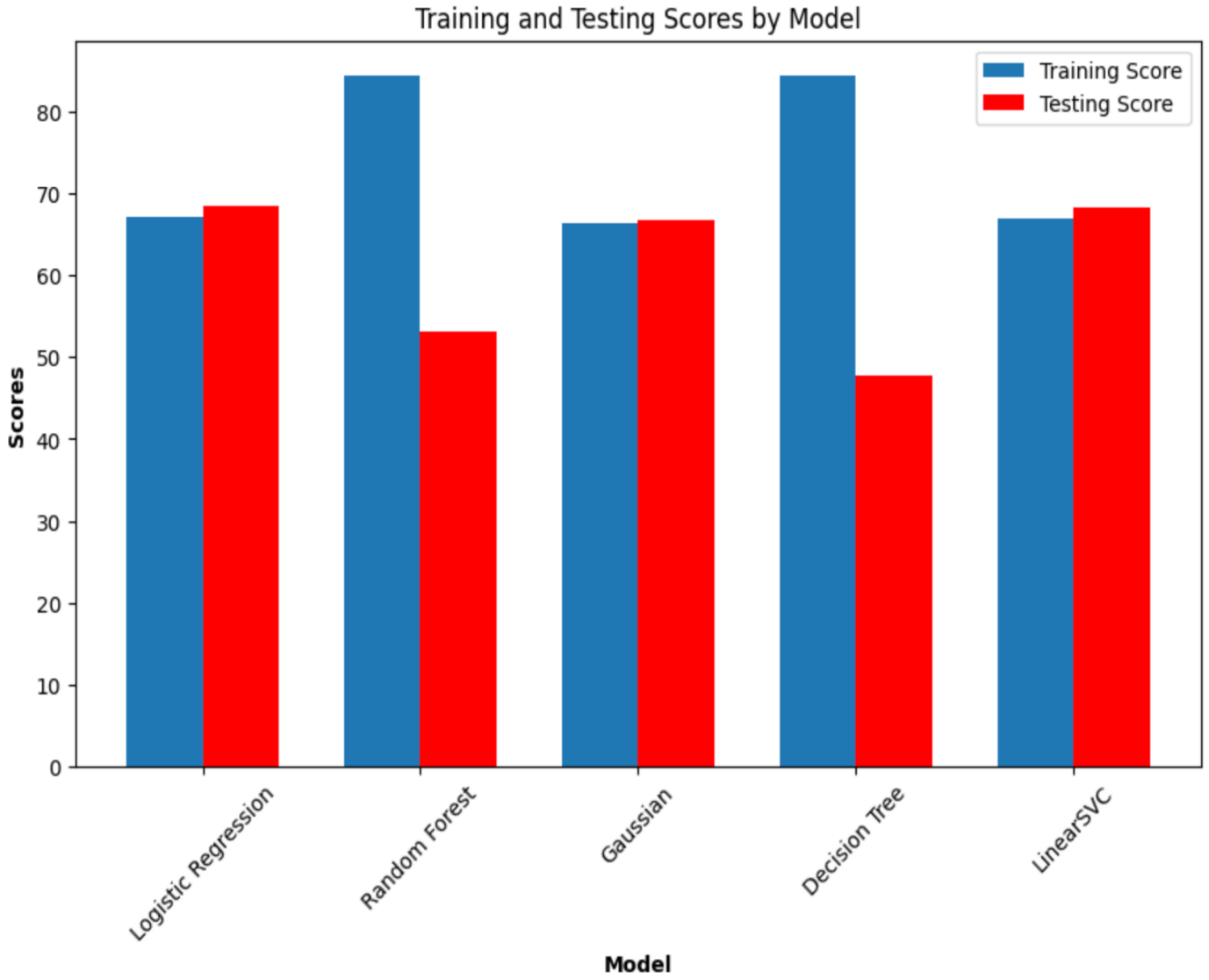
	count	mean	std	min	25%	50%	75%	max
HomePage	9954.0	0.078561	0.131327	0.000000	0.000000	0.000000	0.142857	1.000000
HomePage_Duration	9954.0	0.651224	1.098966	0.000000	0.000000	0.000000	1.000000	4.253125
LandingPage	9954.0	0.021820	0.071975	0.000000	0.000000	0.000000	0.000000	1.000000
ProductDescriptionPage	9954.0	0.110443	0.110757	0.000000	0.031447	0.075472	0.150943	1.000000
ProductDescriptionPage_Duration	9954.0	0.239984	0.747750	-0.512833	-0.368152	0.000000	0.631848	2.519298
GoogleMetric:Bounce Rates	9954.0	0.146426	0.284279	0.000000	0.000000	0.017391	0.125000	1.000000
GoogleMetric:Exit Rates	9954.0	0.266438	0.276019	0.000000	0.084750	0.166667	0.333333	1.000000
GoogleMetric:Page Values	9954.0	0.037330	0.132085	0.000000	0.000000	0.000000	0.000000	1.000000
SeasonalPurchase	9954.0	0.072132	0.213380	0.000000	0.000000	0.000000	0.000000	1.000000
Month_SeasonalPurchase	9954.0	0.564229	0.255888	0.000000	0.444444	0.666667	0.777778	1.000000
Zone	9954.0	0.266124	0.300583	0.000000	0.000000	0.250000	0.375000	1.000000
CustomerType	9954.0	1.720916	0.686941	0.000000	2.000000	2.000000	2.000000	2.000000
Gender	9954.0	1.026622	0.817518	0.000000	0.000000	1.000000	2.000000	2.000000
Cookies Setting	9954.0	1.005927	0.814113	0.000000	0.000000	1.000000	2.000000	2.000000
Education	9954.0	1.501406	1.126415	0.000000	0.000000	1.000000	3.000000	3.000000
Marital Status	9954.0	1.020293	0.813449	0.000000	0.000000	1.000000	2.000000	2.000000
Made_Purchase	9954.0	0.361965	0.480593	0.000000	0.000000	0.000000	1.000000	1.000000

Shape: (9954, 17)

Machine Learning Models to be Deployed:

- Logistic Regression
- Decision Tree
- Random Forest
- GaussianNB
- LinearSVC

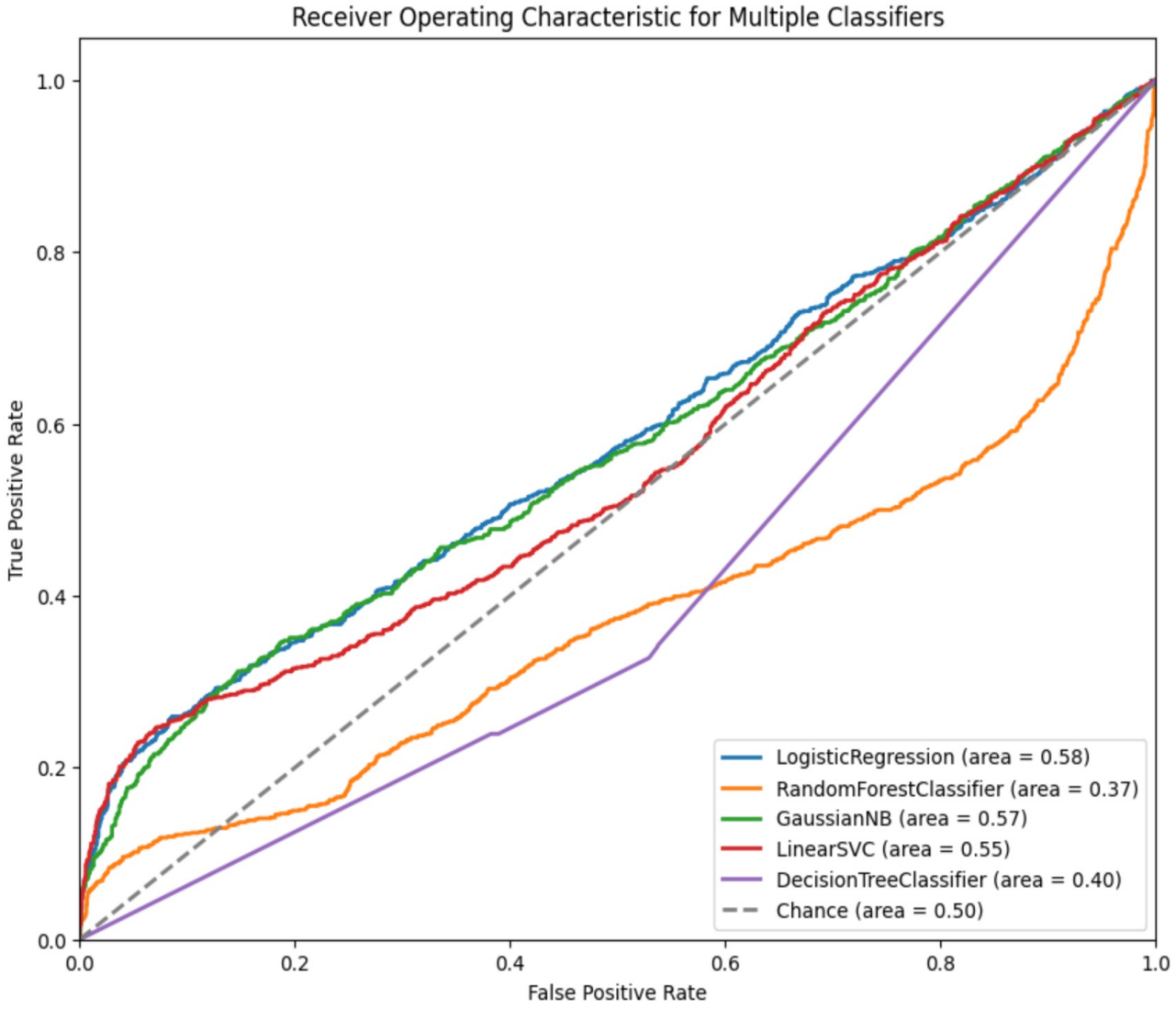
ML Models Performance



- The Random Forest and Decision Tree models have the highest training accuracy but perform poorly with the test data showing overfitting of the model.

- The Logistic Regression, LinearSVC and GaussianNB (in that order) perform lower in training accuracy than the earlier models but have better testing score accuracy making them better for predicting the likelihood of the customer purchasing an item from the website.

ML Models Performance – ROC, AUC.



- In line with the analysis of the training and testing accuracy scores, the Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC) values also show that the Logistic Regression model performs best, followed by the GaussianNB and LinearSVC models.

- The Decision Tree and Random Forest models perform worse than speculative guesses.

ML Models Performance – Cross Validation

- Cross validation performed with cv = 3.
- Cross Validation also confirms the performance of the models. The Accuracy of the Random Forest and Decision Tree models drops significantly after Cross Validation.

Models	cv1 score	cv2 score	cv3 score	Previous Training Accuracy
Decision Tree	0.50	0.50	0.51	0.85
Random Forest	0.55	0.57	0.58	0.85
Logistic Regression	0.67	0.67	0.68	0.67
Gaussian NB	0.66	0.66	0.66	0.66
Linear SVC	0.67	0.66	0.68	0.67

CONCLUSION

1.Influential Factors: The Google Metric for Page Values stands out as the most influential factor in predicting online purchase behaviour. This underscores the importance of engaging content and effective page design in driving e-commerce transactions.

2.Data Pre-processing: The application of median and mode for imputation in handling missing values proved effective, especially given the skewness of the dataset. This approach, along with careful outlier management, optimized the dataset for machine learning analysis without excessive loss of data.

3.Feature Selection: The elimination of low-correlation features such as OS type, Search Engine, Type of Traffic, Weekend Purchase, and Landing Page Duration was crucial in refining the model inputs, thereby enhancing the predictability and performance of the machine learning models.

4.Model Performance: The Logistic Regression, LinearSVC, and GaussianNB models demonstrated a commendable balance between training and testing accuracy, indicating their superior generalization capability and suitability for predicting customer purchase decisions on the website.

5.Overfitting Challenge: The high training accuracy of the Random Forest and Decision Tree models did not translate to the testing phase, highlighting a significant overfitting issue with these models. This serves as a reminder of the importance of model validation and the potential pitfalls of over-reliance on training accuracy.

6.Predictive Reliability: The ROC curve and AUC values reinforced the findings on model performance, with Logistic Regression emerging as the most reliable model for predicting online purchase behaviour, followed closely by GaussianNB and LinearSVC. The consistent performance across different evaluation metrics underscores the robustness of these models in capturing the nuances of online purchasing patterns.