Lead Scoring Case Study using logistic regression

Problem Statement and **Goals**

Business Context**:** X Education sells online courses to industry professionals.
Lead Generation**:** Professionals interested in courses visit the website and provide their contact information, becoming leads.
Lead Acquisition**:** Leads are generated through various sources like website visits, referrals, and marketing.
Conversion Rate**:** Currently, the lead conversion rate is around 30% (i.e., out of 100 leads, only 30 convert into paying customers).

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

Solution Methodology

- **Data cleaning and data manipulation.**
- **Check and handle duplicate data: Ensure the dataset does not contain duplicate entries that could analysis**
- **Check and handle NA values and missing values: Address missing values to avoid errors in analysis and modeling ,**
- **Drop columns, if it contains large amount of missing values and not useful for the analysis: Remove rows or columns with missing values if they are insignificant. Imputation of the values if necessary Fill missing values with appropriate methods (mean, median, mode).**
- **Check and handle outliers in data and**
- **Identify the outliers**

EDA (Exploratory Data Analysis)

Univariate data analysis: statistical method that involves analyzing a single variable at a time. It helps us understand the distribution, central tendency, and variability of a dataset. Using univariate analysis we can find value count, distribution of variable etc.

Bivariate data analysis: statistical method that involves analyzing two variables at a time to understand their relationship. It helps us identify patterns, correlations, and associations between the two variables.Used to find Correlation coefficients and pattern between the variables etc.

Feature Scaling : Normalize features so that they contribute equally to the model and improve convergence speed
- Standardization**:** Common for algorithms that assume data is normally distributed (e.g., Linear Regression).
- Normalization: Useful for algorithms that rely on distances (e.g., K-Nearest Neighbors).
- Robust Scaling: Preferred when dealing with outliers.

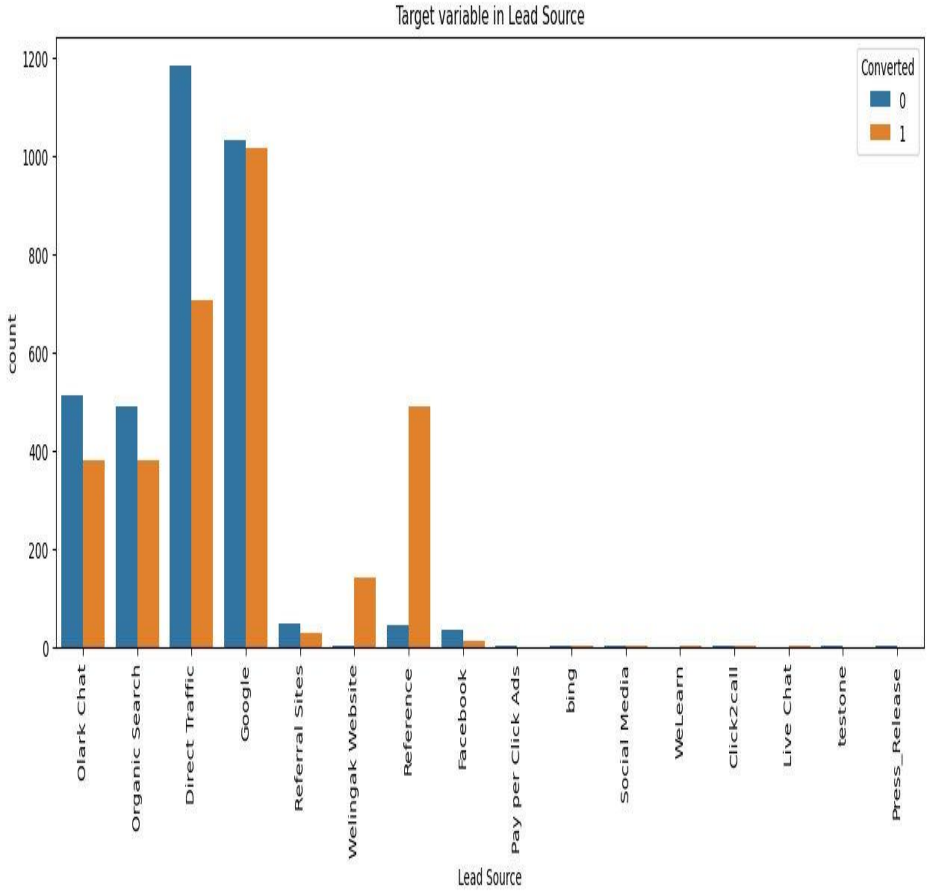
Dummy Variables and encoding of the data: Convert categorical variables into numerical format to use in machine learning models.

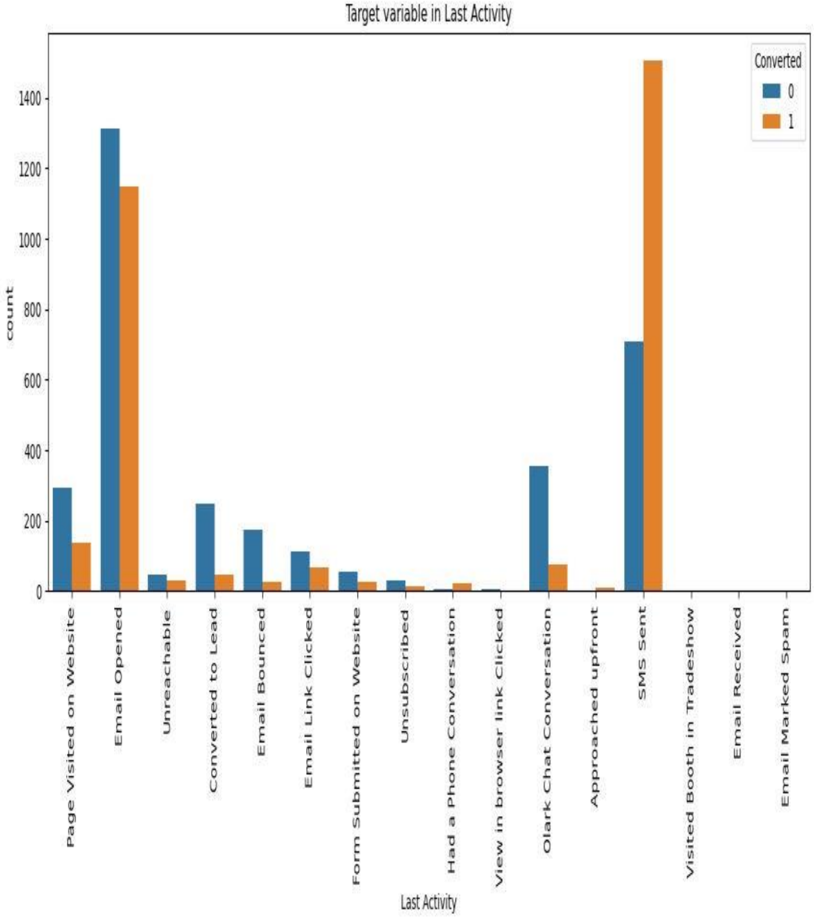
- One Hot Encoding: Converts categorical variables into a set of binary columns.
- Binary Encoding: Combines the benefits of One-Hot and Label Encoding, suitable for high cardinality features. • Label Encoding :Assigns a unique integer to each category.

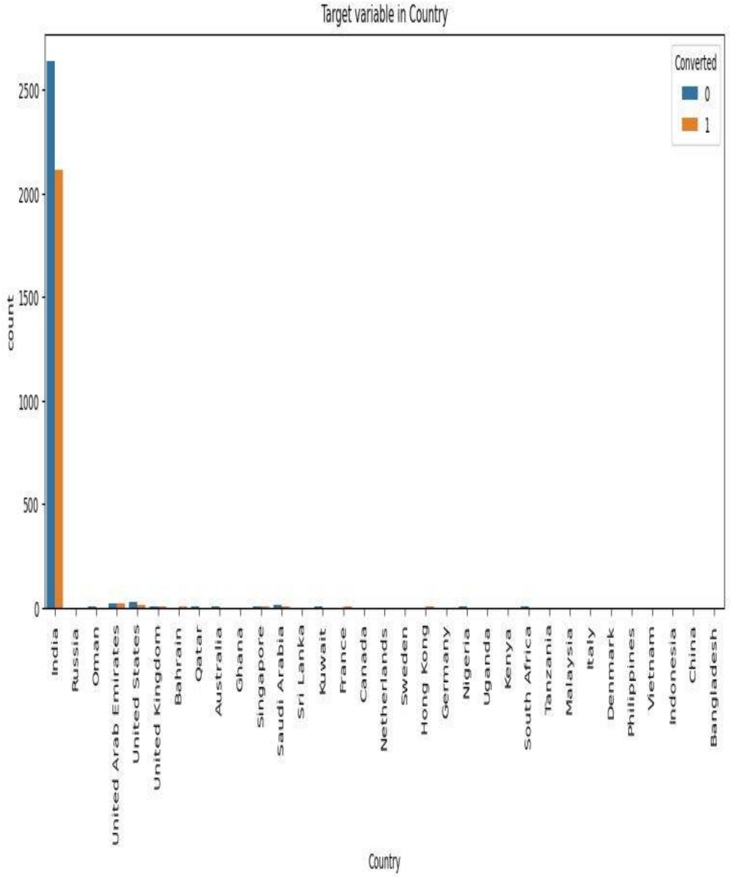
Data Manipulation

- **Total Number of Rows =37, Total Number of Columns =9240.**
- **Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”**
- **Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.**
- **Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.**
- **After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.**
- **Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.**

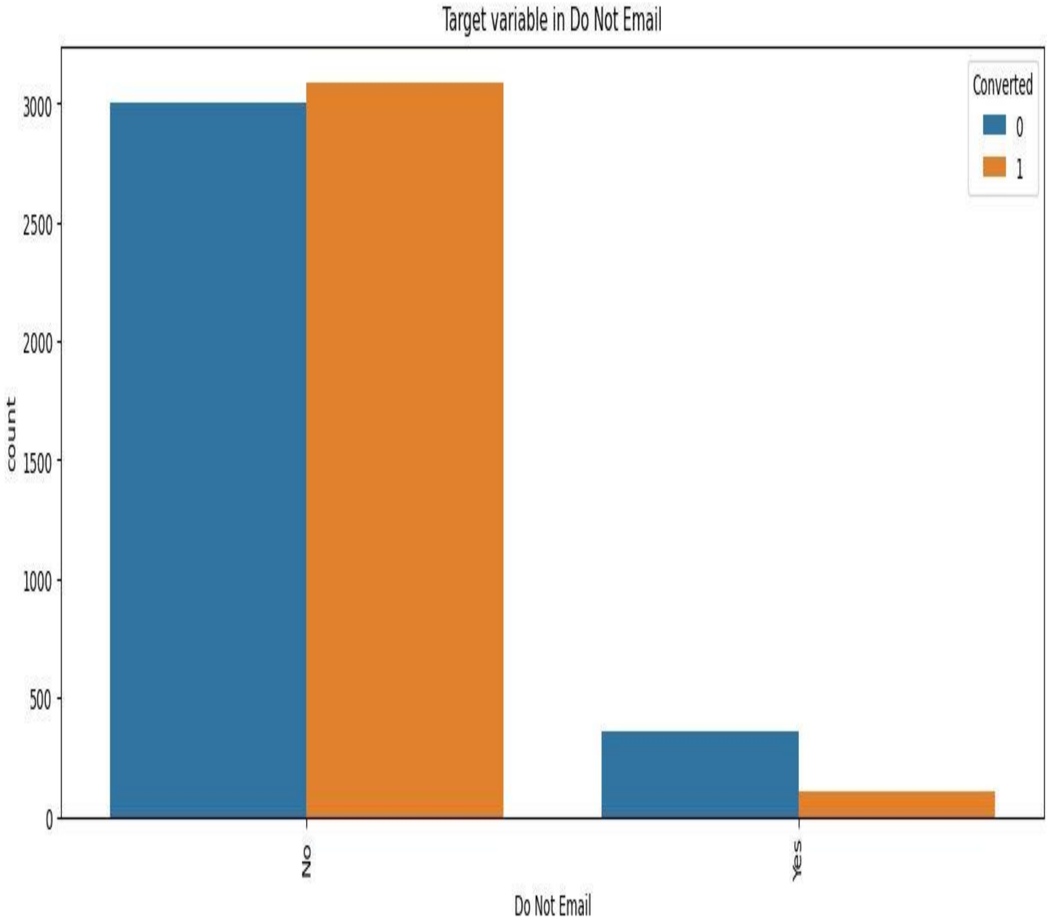
Visualization

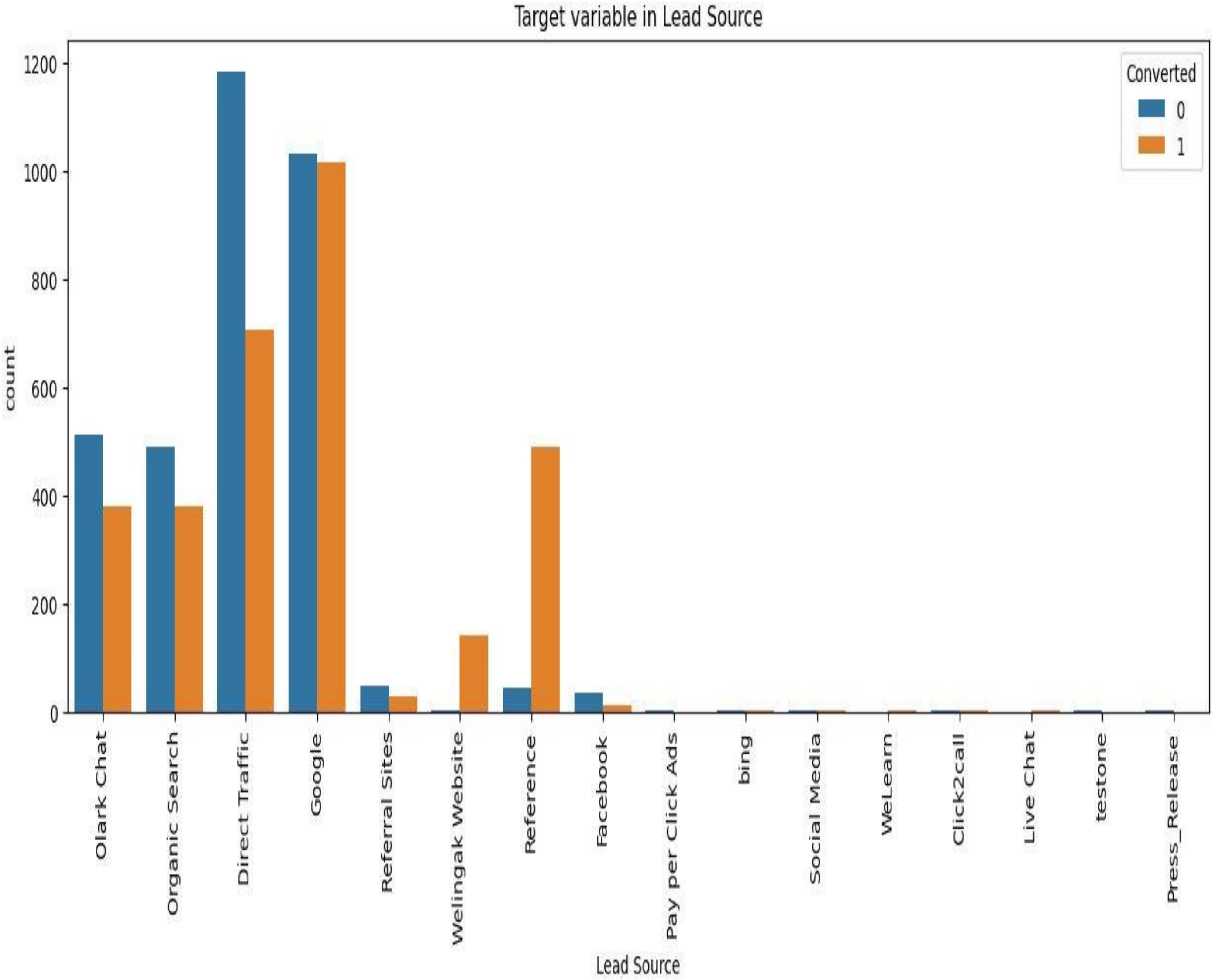




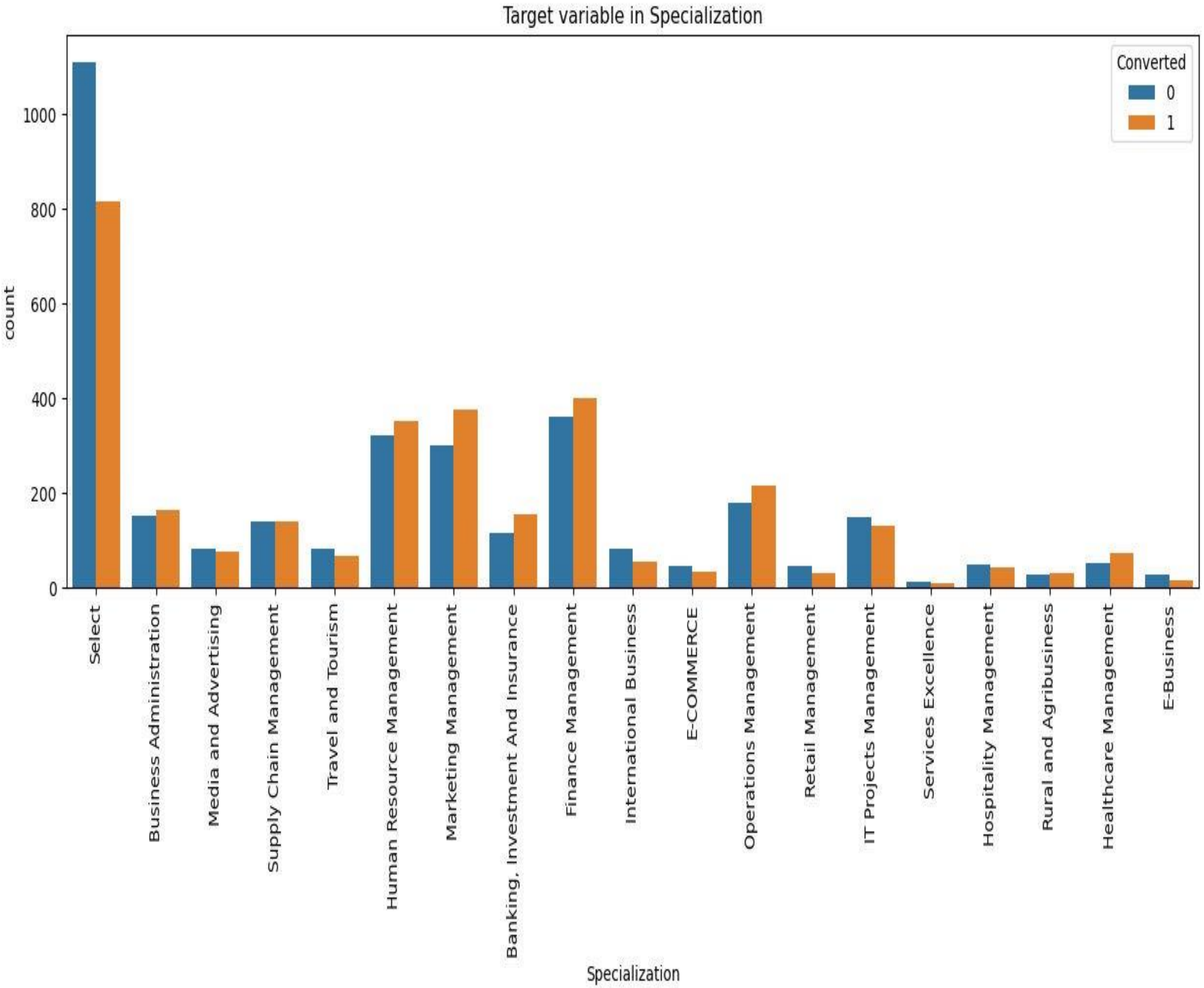


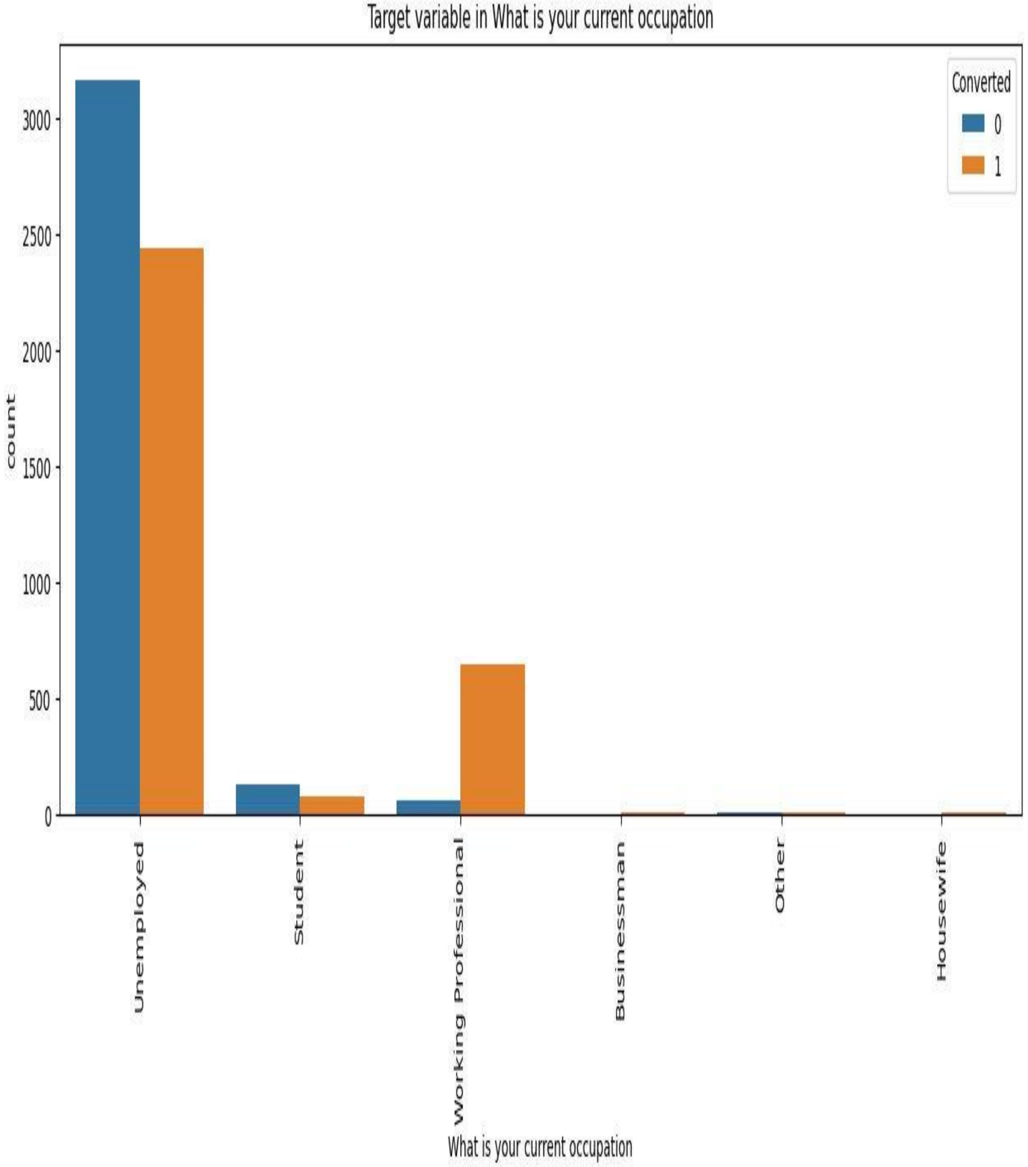
Visualization





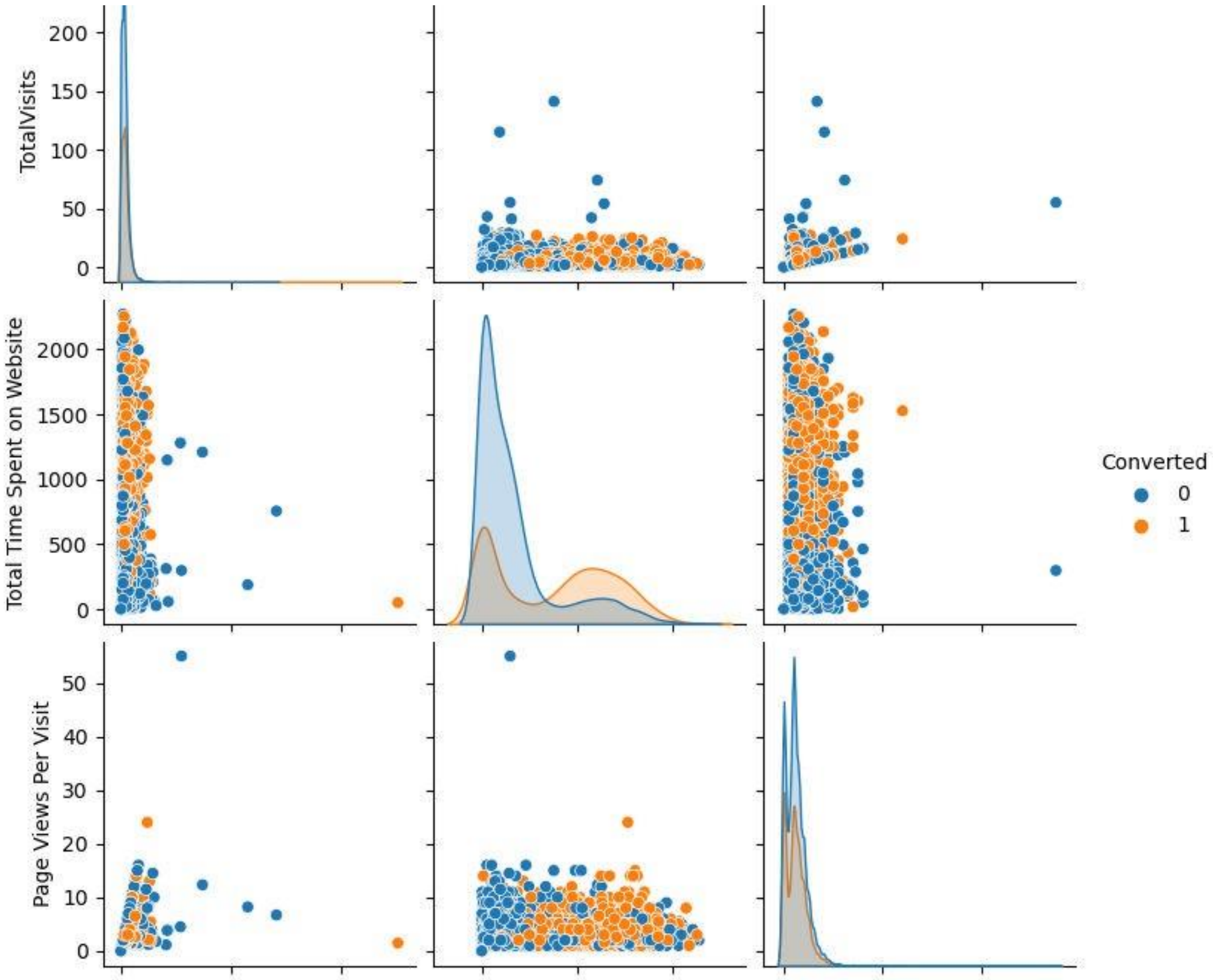
Visualization with Target Variable





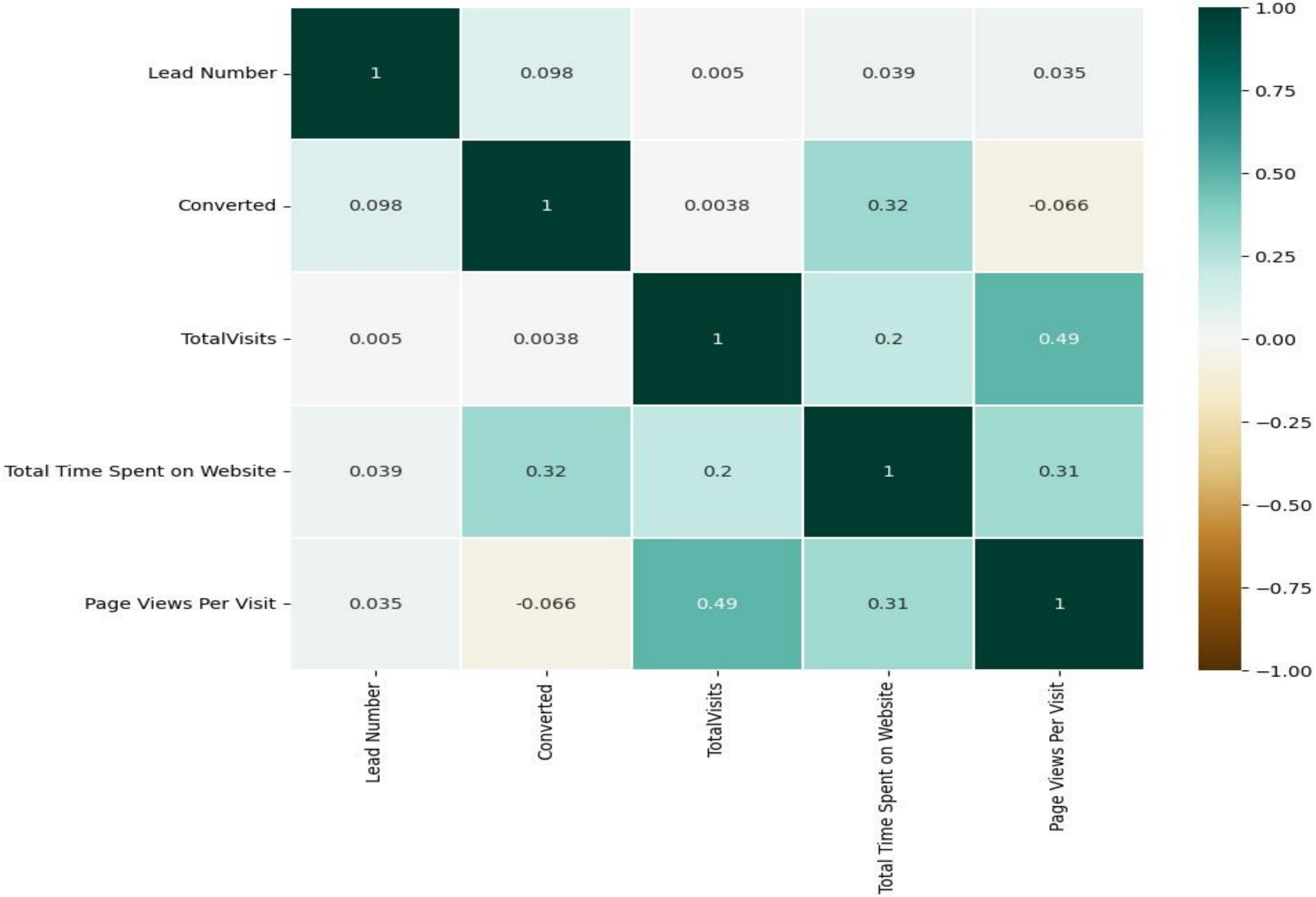
Data Visualization and Analysis

- Image is a pair plot, which is a multivariate data visualization technique.
- It displays the relationships between multiple variables simultaneously, allowing for a comprehensive understanding of their distributions and interdependencies. • Here "Total Visits" variable, providing insights into its frequency distribution, central tendency, and variability.



Correlation Matrix

- Correlation matrix, which visualizes the relationships between different variables.



	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Lead Number	1	0.098	0.005	0.039	0.035
Converted	0.098	1	0.0038	0.32	-0.066
TotalVisits	0.005	0.0038	1	0.2	0.49
Total Time Spent on Website	0.039	0.32	0.2	1	0.31
Page Views Per Visit	0.035	-0.066	0.49	0.31	1

- The variables represent various metrics related to lead generation and website behavior.

- The correlation matrix provides insights into the relationships between different variables related to lead generation and website behavior.
- These Visualizations can be used to identify potential areas for improvement and optimize marketing and sales strategies.

Data Conversion

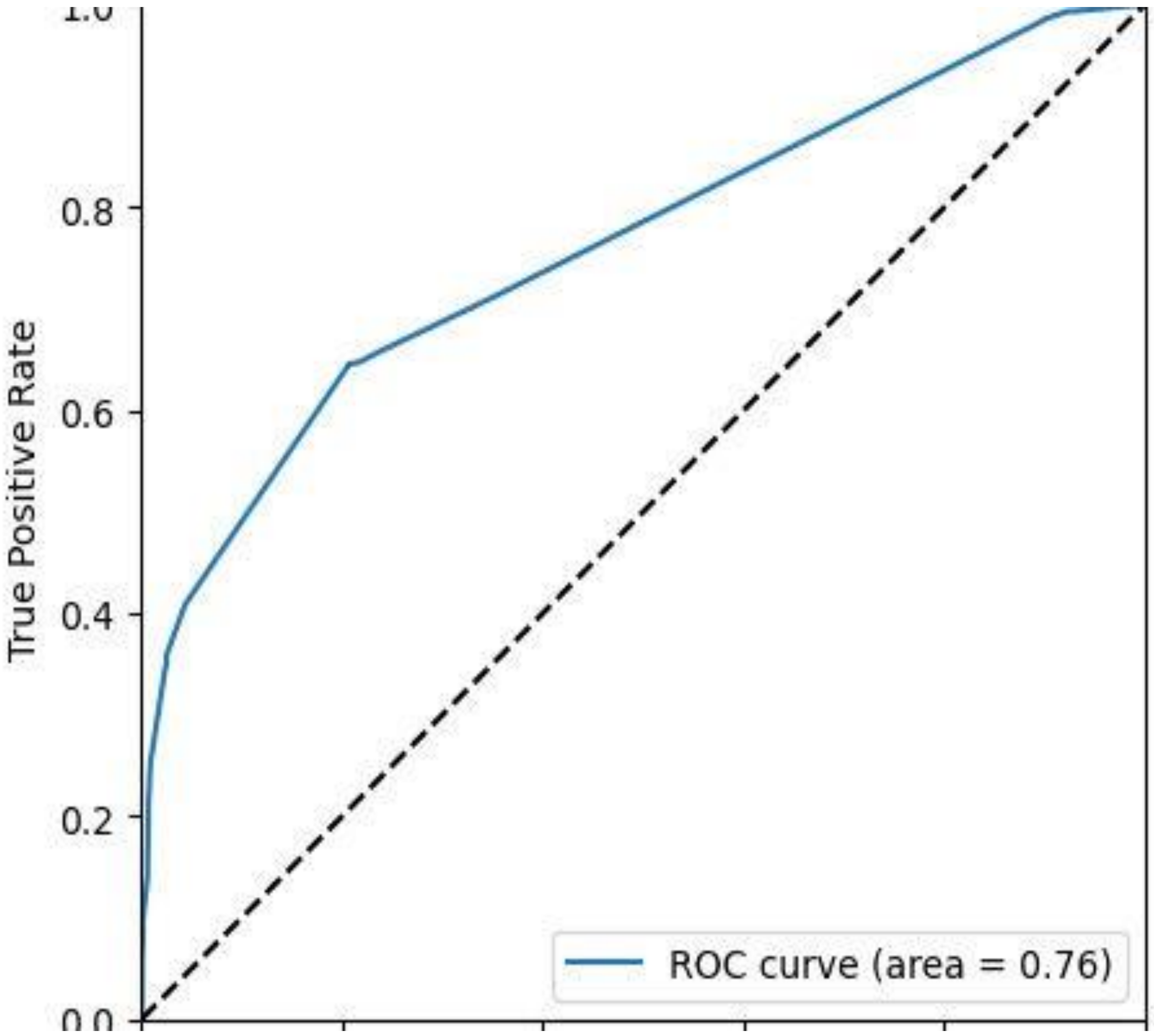
- **Numerical Variables are Normalised**
- **Dummy Variables are created for object type variables**
- **Total Rows for Analysis: 8792**
- **Total Columns for Analysis: 43**

Model Building

- **Splitting the Data into Training and Testing Sets**
-
- **The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.**
- **Use RFE for Feature Selection**
- **Running RFE with 15 variables as output**
- **Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5**
- **Predictions on test data set**
- **Overall accuracy is Good Receiver Operating Characteristic (ROC) curve.**

ROC Curve

ROC curves are used to assess the performance of binary classification models. • Diagonal line: This represents a random classifier. A model that performs no better than random guessing would follow this line.



- Curve shape: The shape of the ROC curve indicates the model's performance. • A curve that is closer to the top-left corner indicates better performance, as it has a high TPR and a low FPR.

- Area under the curve (AUC): This metric quantifies the overall performance of the model. An AUC of 1 indicates perfect classification, while an AUC of 0.5 indicates random guessing. • The ROC curve provides a valuable tool for evaluating the performance of binary classification models.

- By understanding the components of the curve and the AUC, you can assess the model's ability to distinguish between positive and negative cases.

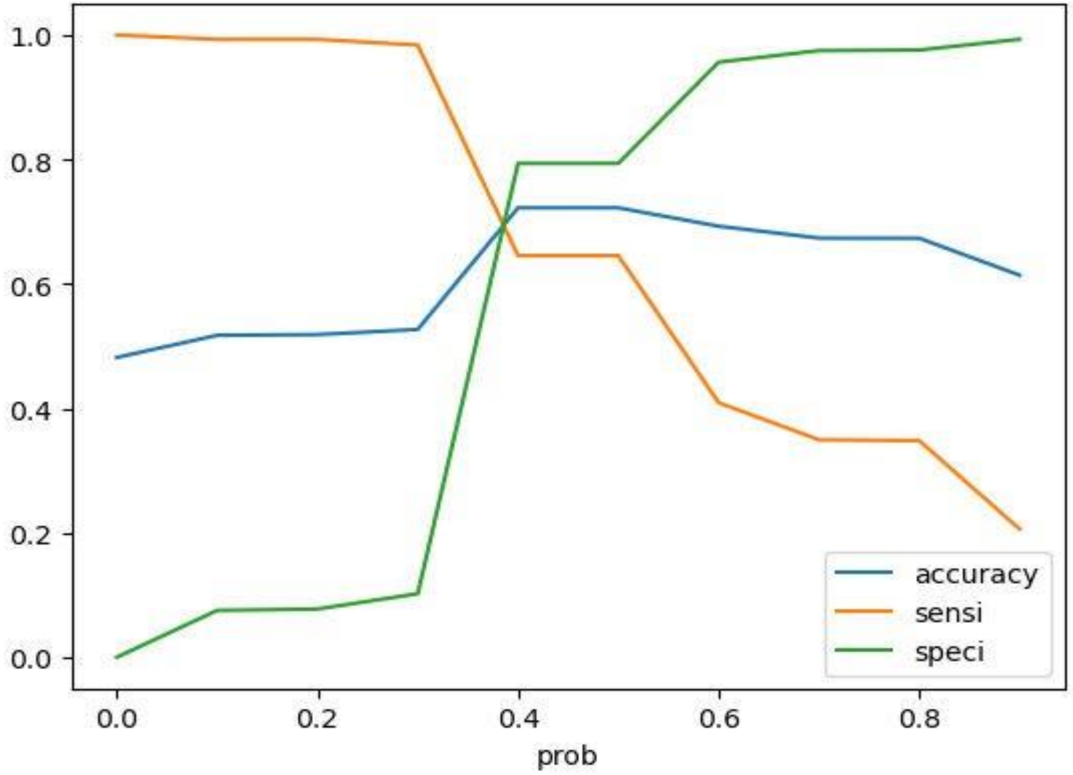
Classification Plot

The plot provides a valuable tool for understanding how the performance of a binary classification model changes with different probability thresholds.

By analyzing the trends in accuracy, sensitivity, and specificity, you can select the optimal threshold that balances the trade-off between these metrics and meets the specific needs of application. Accuracy=0.72

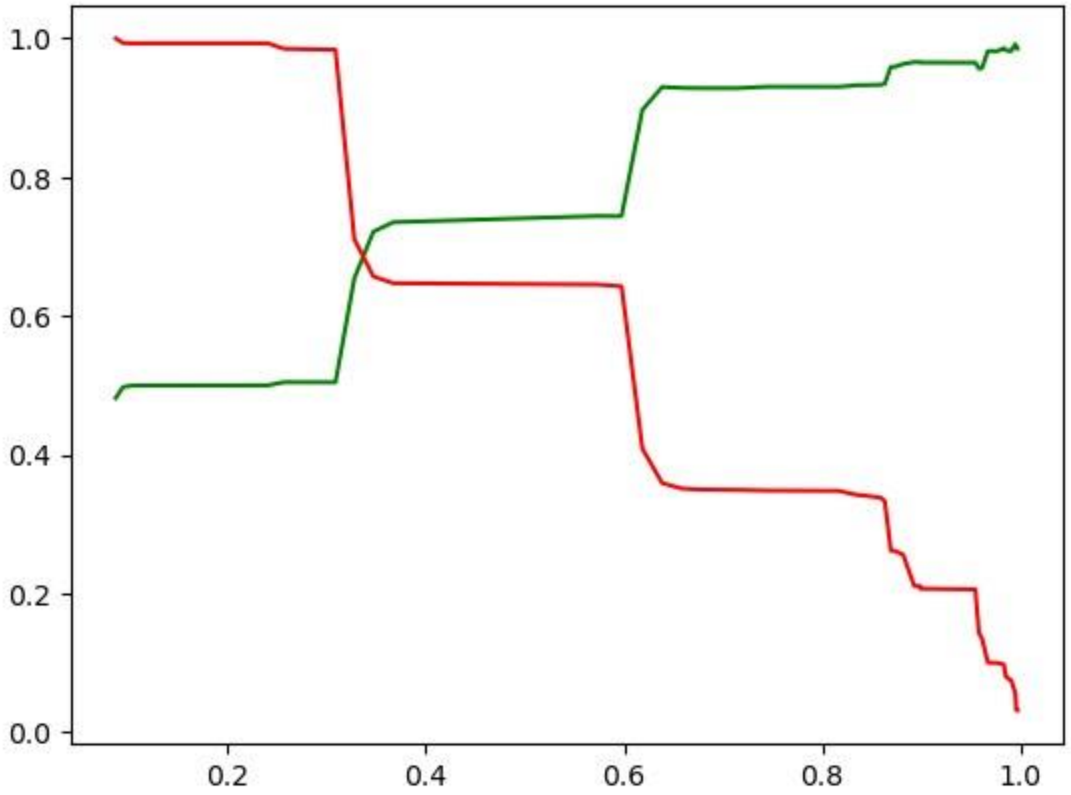
Sensitivity=0.64

Specificity=0.79



Classification Chart

Chart provides a valuable tool for understanding the performance of classification models in terms of their ability to target positive cases.
By analyzing the shape of the curve and the lift, assess the model's effectiveness in capturing the most valuable segments of the target population.
Accuracy:0.72
Precision :0.74 Recall :0.64



Conclusion /Summary

It was found that the variables that mattered the most in the potential buyers are (In descending order)

- The total time spend on the Website.
- Total number of visits.

Lead source

a. Google
b. Direct traffic
c. Organic search
d. Welingak website

SMS

- Olark chat conversation
- When the lead origin is Lead add format.
- When their current occupation is as a working professional.
- Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses..