Lead Score Case Study-Upgrad

Group Members:

Avinandita Mahanti

Ashwini Mukati

Approach towards the Solution

1. **Check and handle duplicate data:**
    1. Identify and remove any duplicate records from the dataset to ensure data integrity and avoid bias in analysis.
2. **Check and handle NA values and missing values:**
    1. Identify columns with missing values (NA) and decide on appropriate strategies to handle them (e.g., imputation, removal, or treating them as a separate category).
3. **Drop columns with a large amount of missing values and not useful for analysis:**
    1. Identify columns with a high percentage of missing values and consider dropping them if they don't contribute meaningful information.
4. **Imputation of missing values if necessary:**
    1. For columns with missing values, consider imputing the missing values using techniques like mean, median, mode, or regression imputation, based on the data's characteristics.
5. **Check and handle outliers in data:**
    1. Identify outliers in the data and decide on appropriate strategies for handling them (e.g., removing, transforming, or treating them as special cases).
6. **Exploratory Data Analysis (EDA)**
    1. Perform univariate data analysis: Examine the distribution, value counts, and summary statistics of individual variables to gain insights into their characteristics.
    2. Perform bivariate data analysis: Analyze the relationships between pairs of variables, such as correlation coefficients, scatter plots, or other visualizations, to understand their patterns and potential dependencies.

7. Feature Scaling & Dummy Variables and encoding of the data:

1. If needed, apply feature scaling (e.g., normalization or standardization) to ensure all variables are on a similar scale.
2. Create dummy variables or perform encoding to convert categorical variables into numeric format suitable for the model.

8.Classification technique: Logistic Regression

1. Train a logistic regression model on the preprocessed data to predict the binary outcome variable.

9.Validation of the model:

1. Split the dataset into training and testing sets or use cross-validation techniques to evaluate the model's performance on unseen data.
2. Calculate relevant metrics like accuracy, precision, recall, F1-score, and ROC-AUC to assess the model's performance.

10.Model presentation:

- Present the logistic regression model and its results, including coefficients, significance, and model performance metrics.

11.Conclusions and recommendations:

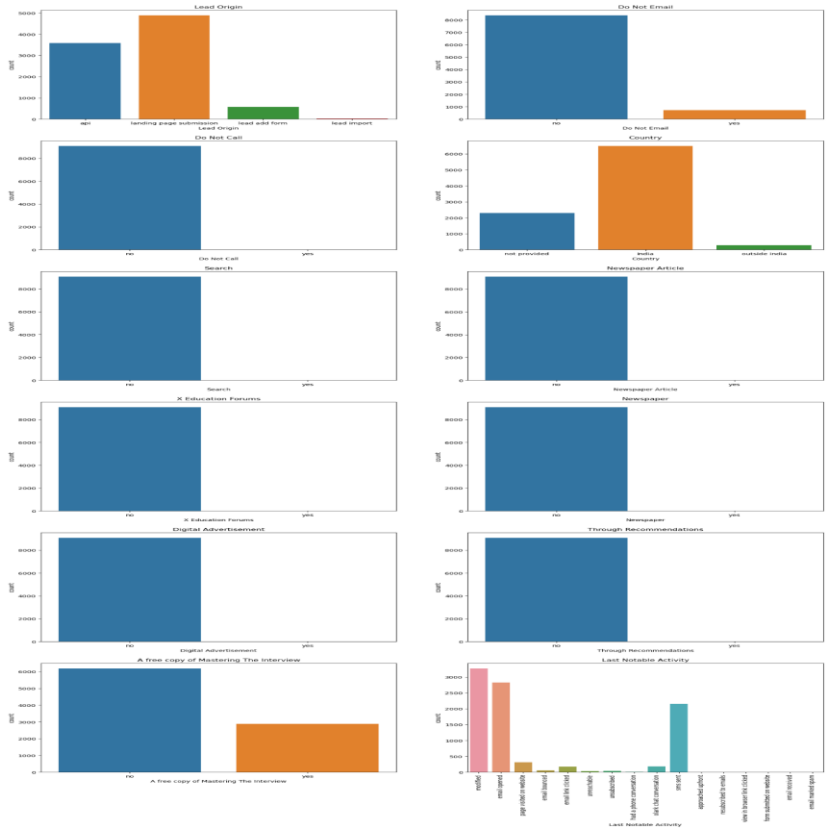
- Summarize the findings from the analysis, highlighting important variables and their impact on the target variable.
- Provide actionable recommendations based on the insights gained from the model to improve decision-making and achieve the desired outcomes.
- It's important to note that each step may require iterations and adjustments based on the nature of the data and the objectives of the analysis. Additionally, proper validation and interpretation of the model results are essential for making informed decisions.

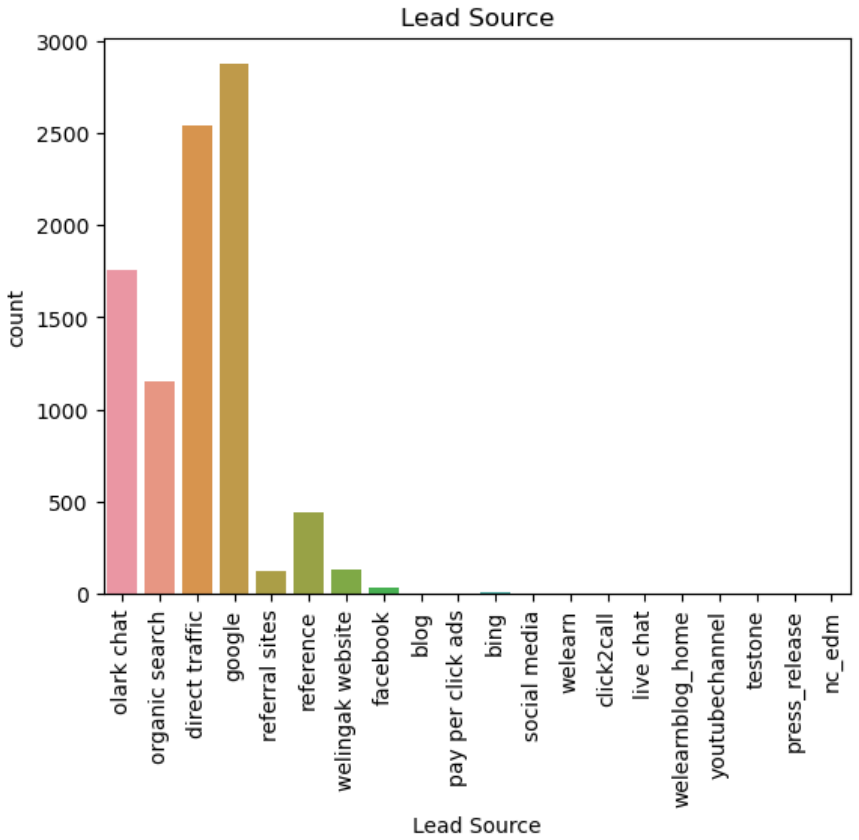
Data Pre-Processing

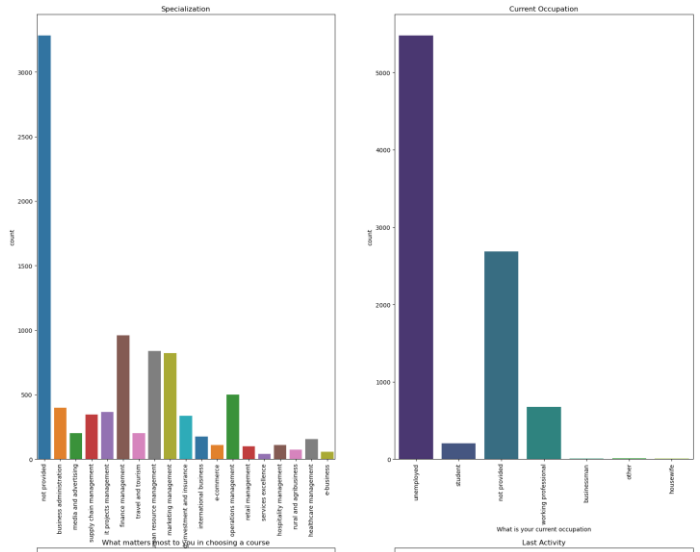
1. **Data Size:**
    1. Total Number of Rows: 37
    2. Total Number of Columns: 9240
2. **Dropping Single Value Features:**
    1. Drop features with a single unique value, such as "Magazine," "Receive More Updates About Our Courses," "Update me on Supply Chain Content," "Get updates on DM Content," "I agree to pay the amount through cheque," etc.
3. **Removing Unnecessary Columns:**
    1. Remove "Prospect ID" and "Lead Number" columns, which are not necessary for the analysis.
4. **Dropping Features with Low Variance:**
    1. Drop features with low variance, such as "Do Not Call," "What matters most to you in choosing course," "Search," "Newspaper Article," "X Education Forums," "Newspaper," "Digital Advertisement," etc. These features may not provide valuable information for analysis due to lack of variation.
5. **Dropping Columns with High Missing Values:**
    1. Drop columns with more than 35% missing values, such as 'How did you hear about X Education' and 'Lead Profile.' These columns have a significant amount of missing data, making them less informative for the analysis.

After performing these data cleaning and manipulation steps, you will be left with a more streamlined and relevant dataset that can be used for further analysis, feature engineering, and building the logistic regression model for classification. Remember to keep track of the changes made to the dataset and document the preprocessing steps for transparency and reproducibility.

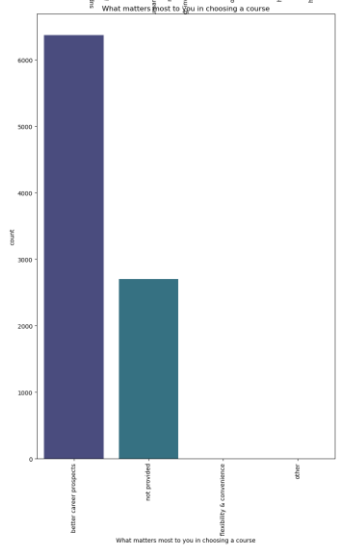
Graphs from Exploratory Data Analysis

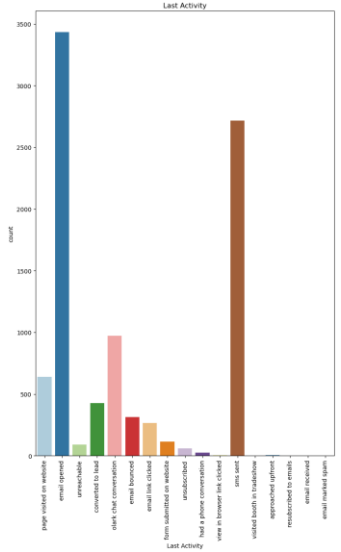


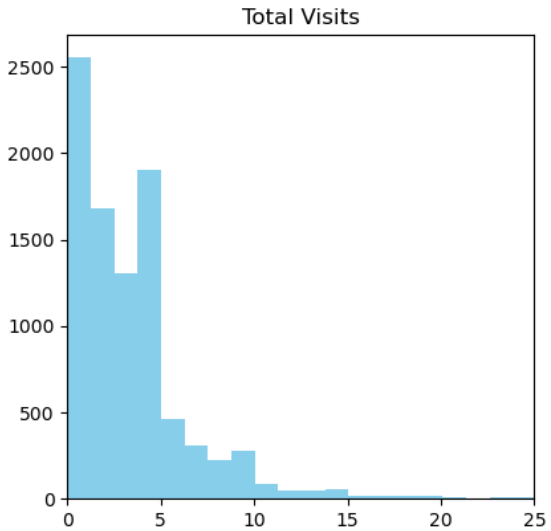


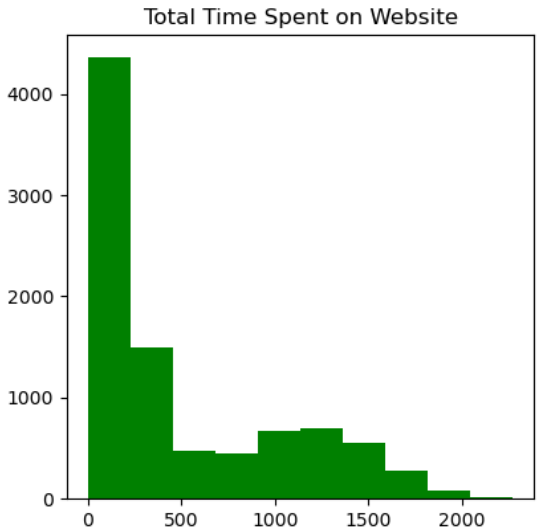

Specialization
count
not provided
business administration
media and advertising
supply chain management
IT projects management
finance management
travel and tourism
human resource management
marketing management
retail management
international business
e-commerce
operations management
retail management
services excellence
hospitality management
rural and agribusiness
healthcare management
e-business

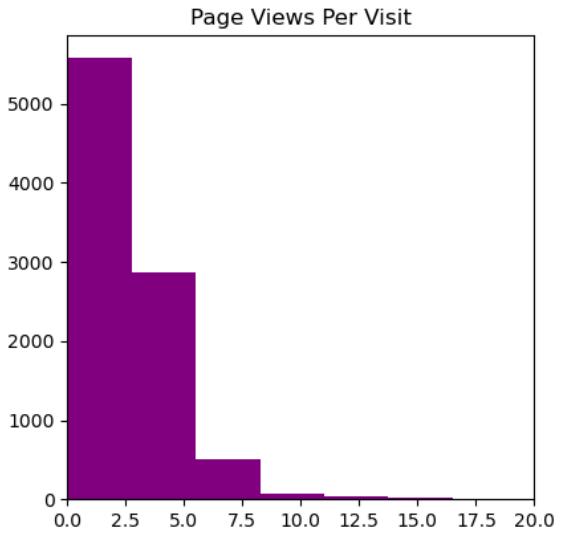
Current Occupation
count
unemployed
student
not provided
working professional
businessman
other
housewife
What is your current occupation

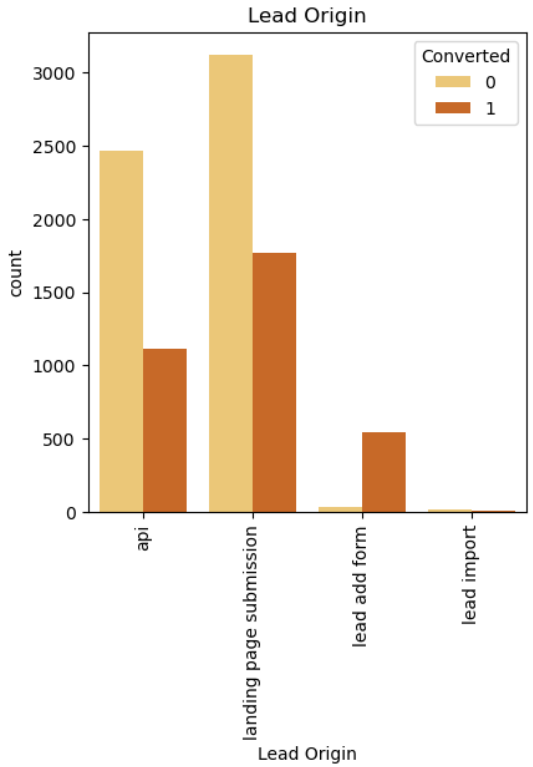

What matters most to you in choosing a course
count
better career prospects
not provided
flexibility & convenience
other
What matters most to you in choosing a course

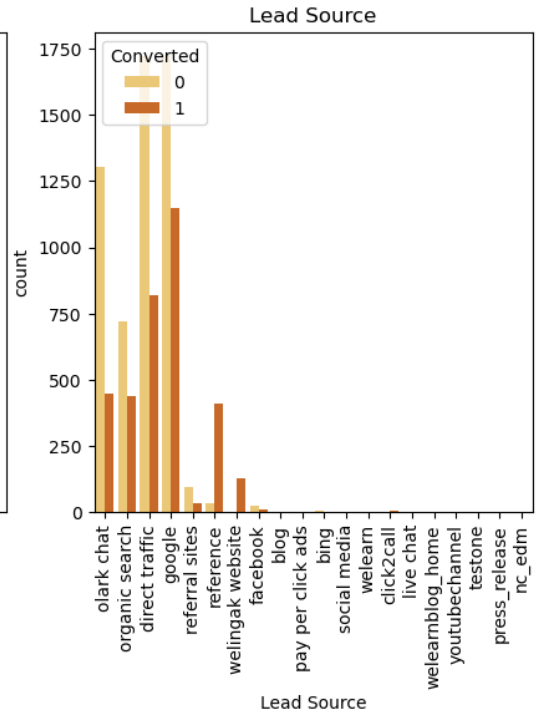

Last Activity
count
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity

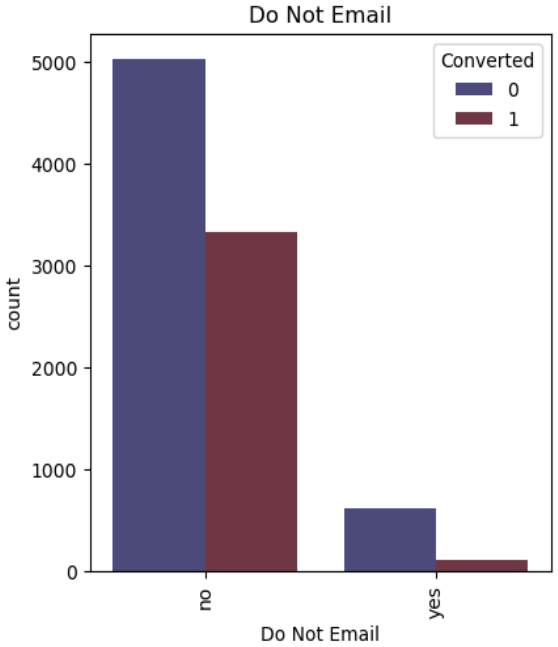

Total Visits
2500
2000
1500
1000
500
0
0
5
10
15
20
25

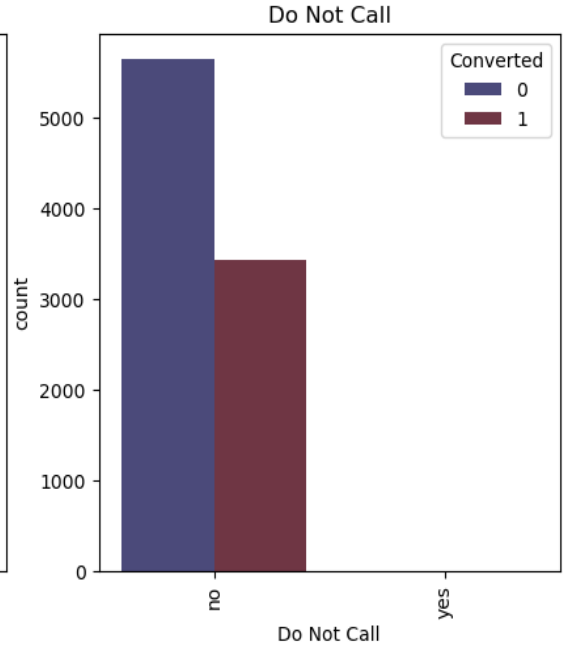

Total Time Spent on Website
4000
3000
2000
1000
0
0
500
1000
1500
2000

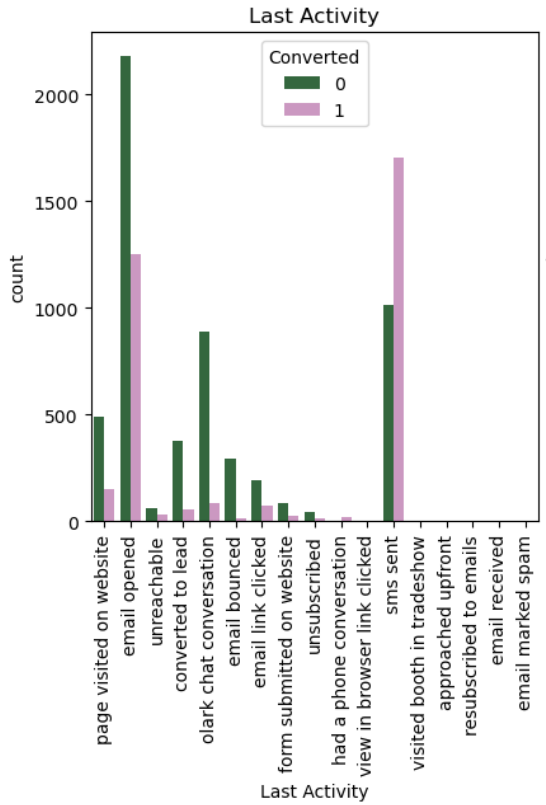

Page Views Per Visit
5000
4000
3000
2000
1000
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

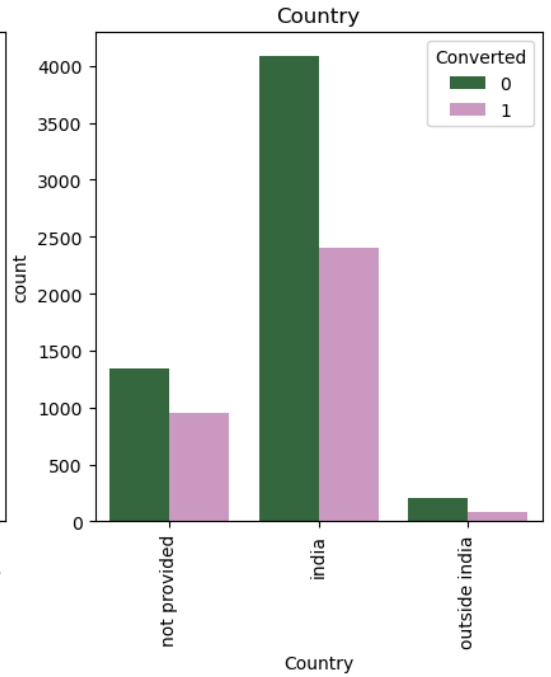

Lead Origin
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
api
landing page submission
lead add form
lead import
Lead Origin

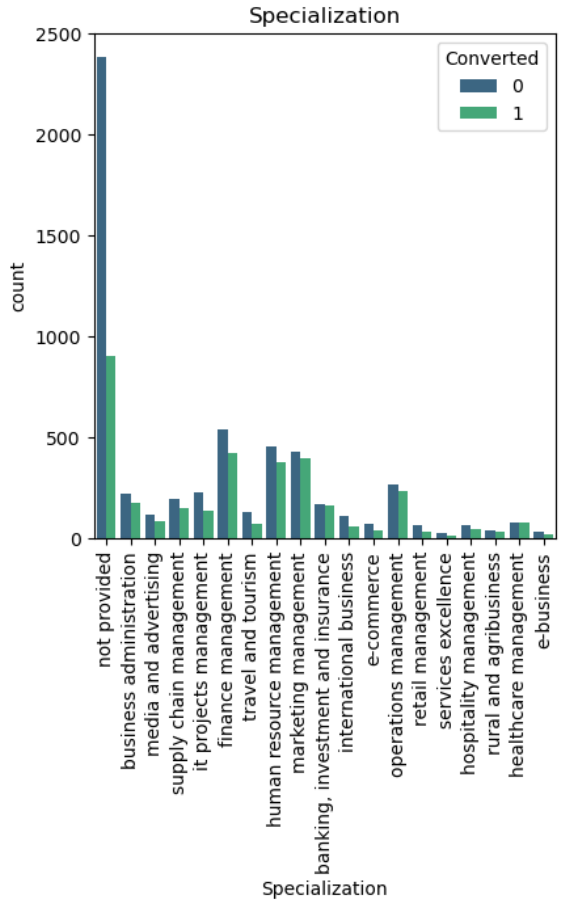

Lead Source
Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
olark chat
organic search
direct traffic
google
referral sites
reference
welingak website
facebook
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source

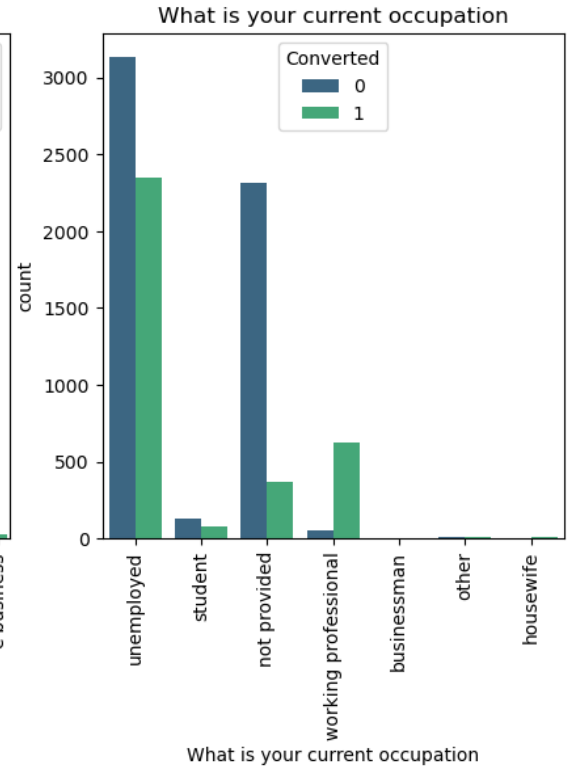

Do Not Email
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Do Not Email

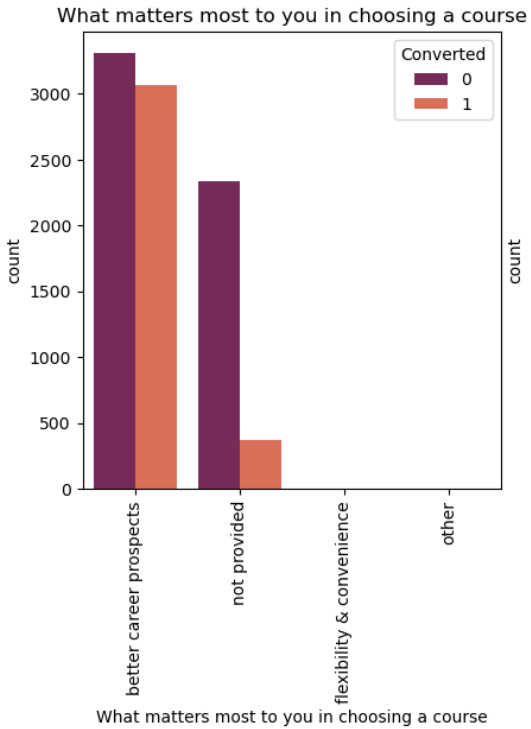

Do Not Call
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Do Not Call

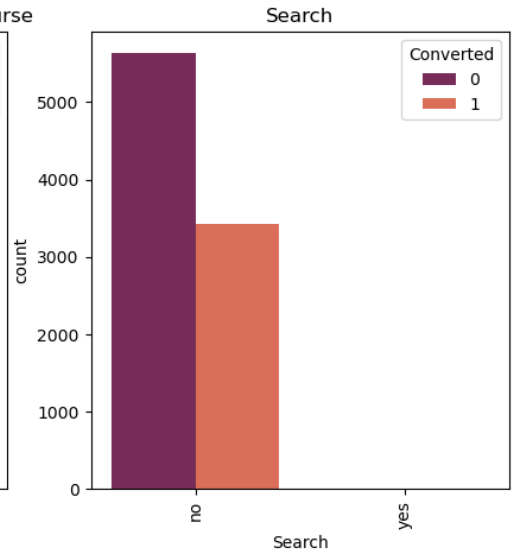

Last Activity
Converted
0
1
count
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity

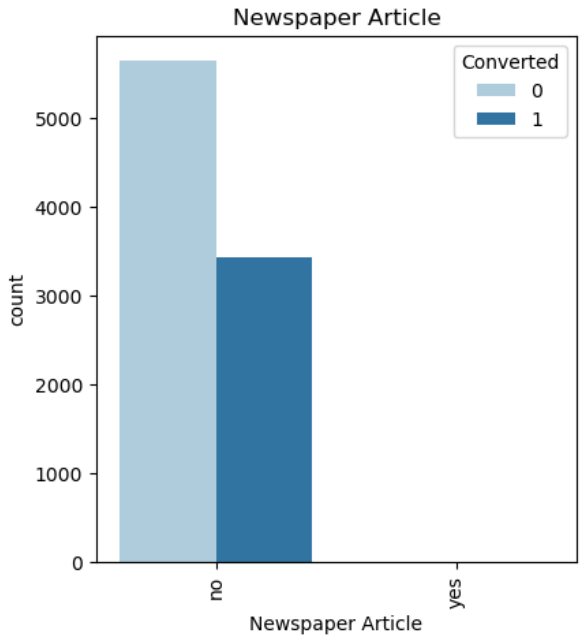

Country
Converted
0
1
count
not provided
india
outside india
Country

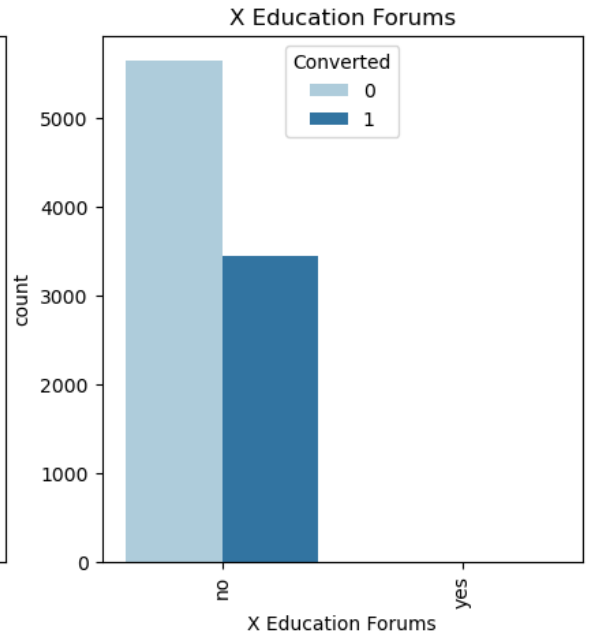

Specialization
Converted
0
1
count
not provided
business administration
media and advertising
supply chain management
it projects management
finance management
travel and tourism
human resource management
marketing management
banking, investment and insurance
international business
e-commerce
operations management
retail management
services excellence
hospitality management
rural and agribusiness
healthcare management
e-business
Specialization

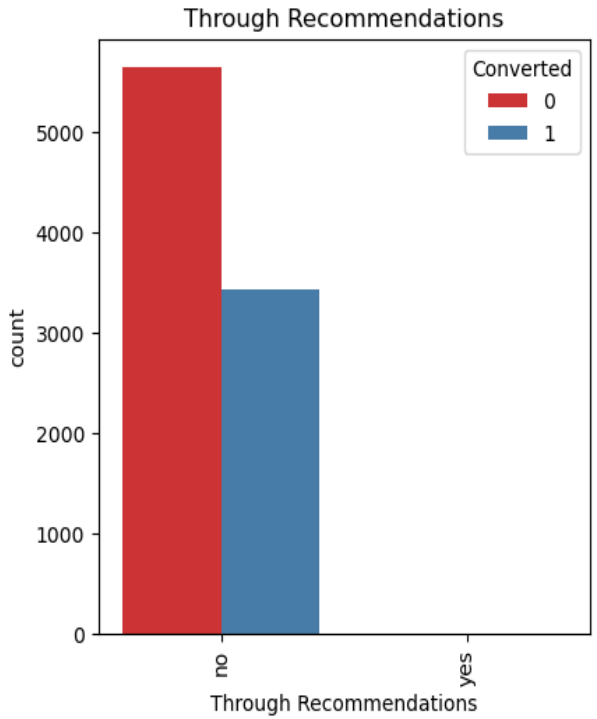

What is your current occupation
Converted
0
1
count
unemployed
student
not provided
working professional
businessman
other
housewife
What is your current occupation

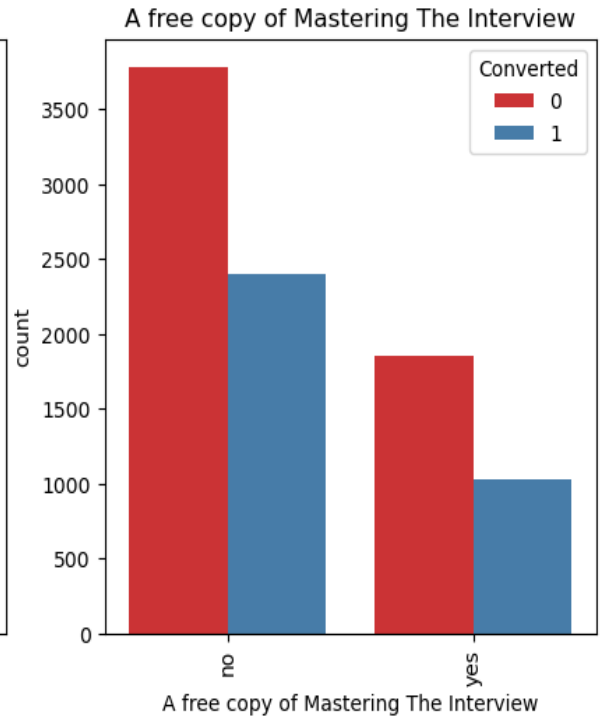

What matters most to you in choosing a course
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
better career prospects
not provided
flexibility & convenience
other
What matters most to you in choosing a course

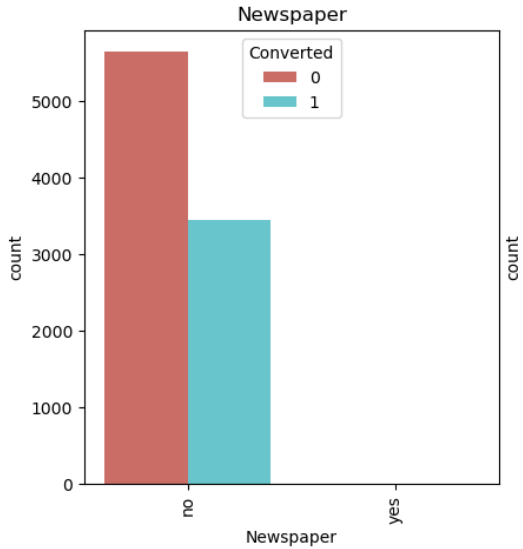

Search
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Search

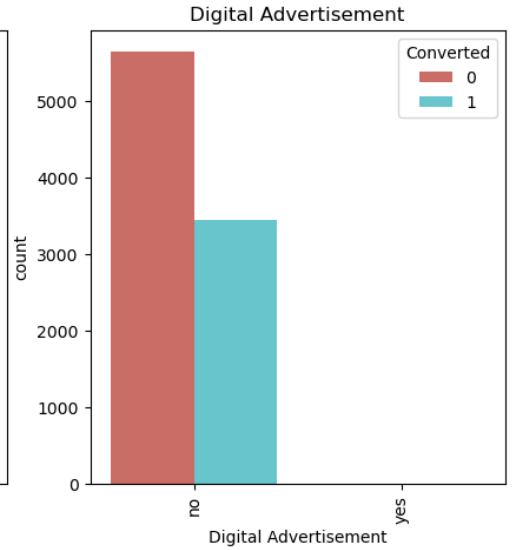

Newspaper Article
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Newspaper Article

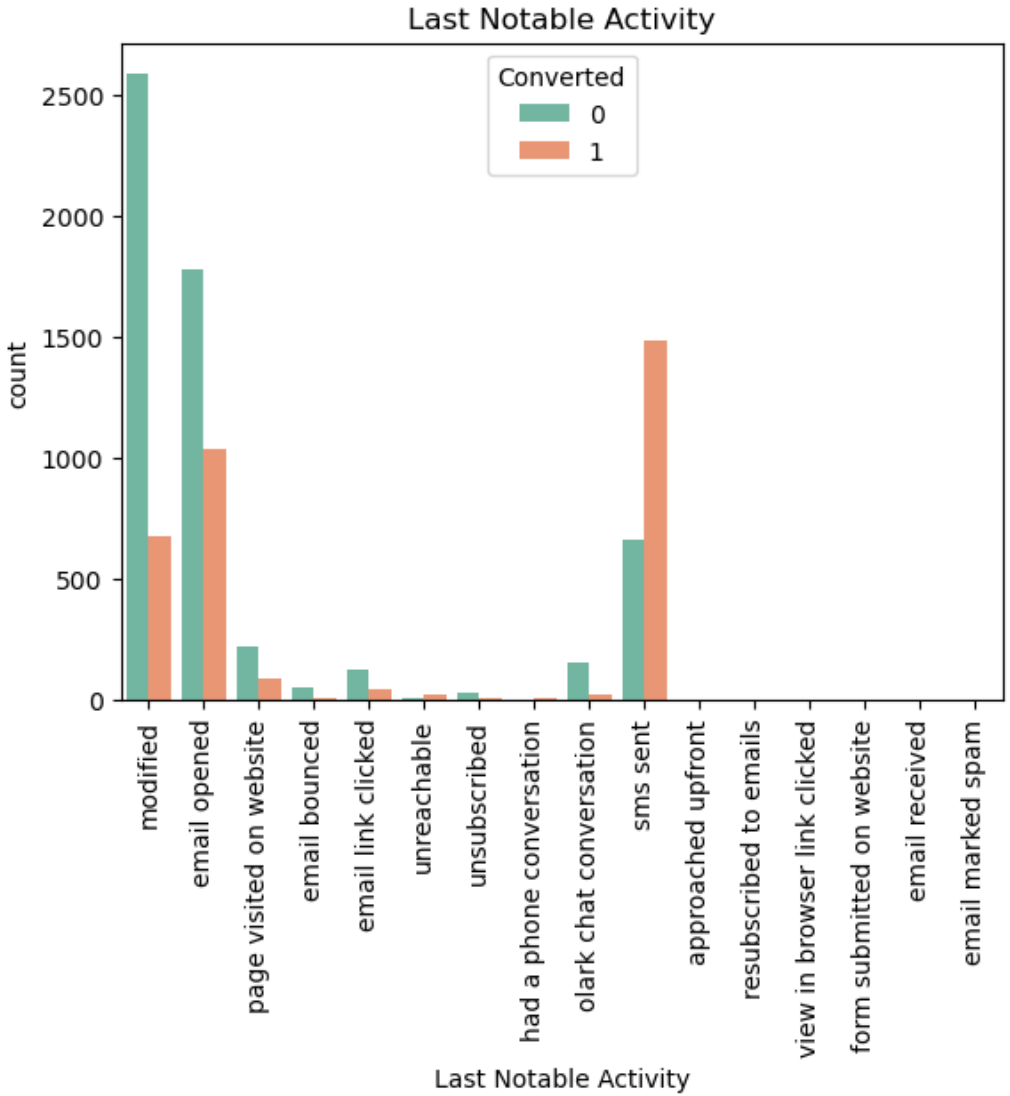

X Education Forums
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
X Education Forums

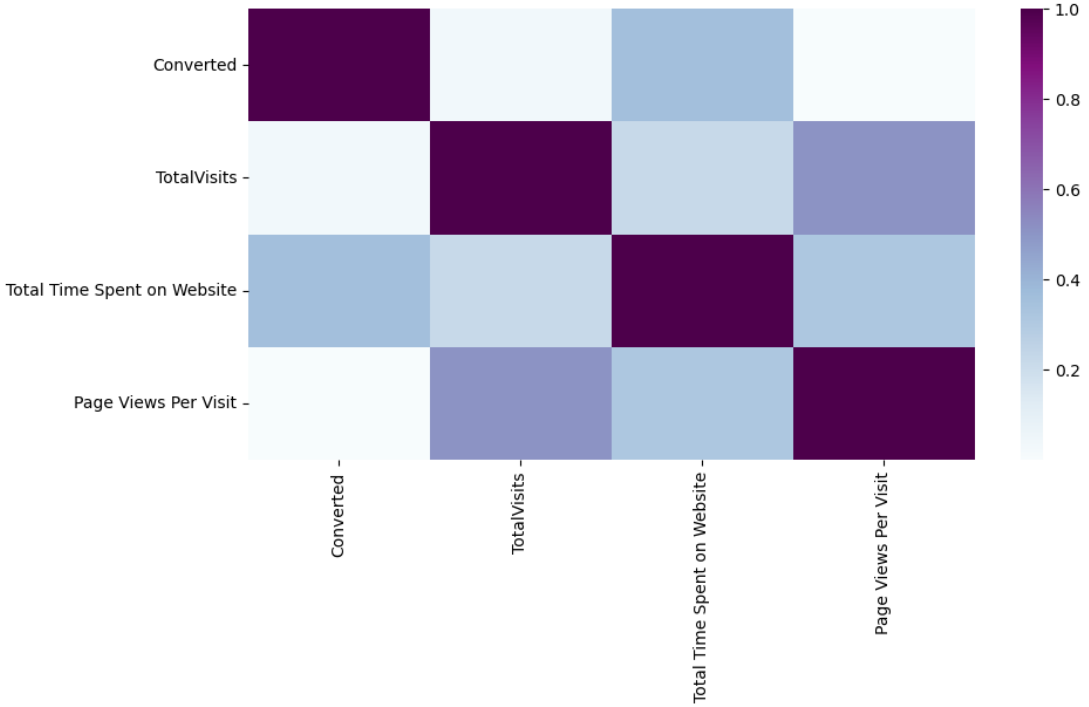

Through Recommendations
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Through Recommendations


A free copy of Mastering The Interview
Converted
0
1
count
3500
3000
2500
2000
1500
1000
500
0
no
yes
A free copy of Mastering The Interview


Newspaper
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Newspaper


Digital Advertisement
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Digital Advertisement


Last Notable Activity
Converted
0
1
count
2500
2000
1500
1000
500
0
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted on website
email received
email marked spam
Last Notable Activity


Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1.0
0.8
0.6
0.4
0.2
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit

TotalVisits
Total Time Spent on Website
Page Views Per Visit
Lead Origin_landing page submission
Lead Origin_lead add form
Lead Origin_lead import
Specialization_business administration
Specialization_e-business
Specialization_e-commerce
Specialization_finance management
Specialization_healthcare management
Specialization_hospitality management
Specialization_human resource management
Specialization_international business
Specialization_it projects management
specialization_marketing management
specialization_media and advertising
specialization_operations management
Specialization_retail management
Specialization_rural and agribusiness
Specialization_services excellence
Specialization_supply chain management
Specialization_travel and tourism
Lead Source_blog
Lead Source_click2call
Lead Source_direct traffic
Lead Source_facebook
Lead Source_google
Lead Source_live chat
Lead Source_nc_cdm
Lead Source_olark chat
Lead Source_organic search
Lead Source_pay per click ads
Lead Source_press_release
Lead Source_reference
Lead Source_referral sites
Lead Source_social media
Lead Source_testzone
Lead Source_welearn
Lead Source_welearnblog_home
Lead Source_welingak website
Lead Source_youtubechannel
Do Not Email_yes
Last Activity_converted to lead
Last Activity_email bounced
Last Activity_email link clicked
Last Activity_email marked spam
Last Activity_email opened
Last Activity_email received
Last Activity_form submitted on website
Last Activity_had a phone conversation
Last Activity_olark chat conversation
Last Activity_page visited on website
Last Activity_resubscribed to emails
Last Activity_sms sent
Last Activity_unreachable
Last Activity_unsubscribed
Last Activity_view in browser link clicked
Last Activity_visited booth in tradeshow
What is your current occupation_housewife
What is your current occupation_other
What is your current occupation_student
What is your current occupation_unemployed
What is your current occupation_working professional
A free copy of Mastering The Interview_yes
Last Notable Activity_email bounced
Last Notable Activity_email link clicked
Last Notable Activity_email marked spam
Last Notable Activity_email opened
Last Notable Activity_email received
Last Notable Activity_form submitted on website
Last Notable Activity_had a phone conversation
Last Notable Activity_modified
Last Notable Activity_olark chat conversation
Last Notable Activity_page visited on website
Last Notable Activity_resubscribed to emails
Last Notable Activity_sms sent
Last Notable Activity_unreachable
Last Notable Activity_unsubscribed
Last Notable Activity_view in browser link clicked
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Conversion of Data

1. **Numerical Variables Normalization:**
    1. Numerical variables have been normalized, which means that their values have been transformed to a common scale to ensure that they have similar ranges. Common normalization techniques include Min-Max scaling or Z-score normalization.
2. **Creation of Dummy Variables:**
    1. For object type variables (categorical variables), dummy variables have been created. Dummy variables are binary representations of each category within a categorical variable. They are used to convert categorical data into a numerical format suitable for analysis and modeling.
3. **Total Rows for Analysis:**
    1. After the data conversion, the dataset contains 8792 rows, which are now suitable for analysis.
4. **Total Columns for Analysis:**
    1. The dataset now consists of 43 columns that have been prepared for analysis after the data conversion process.

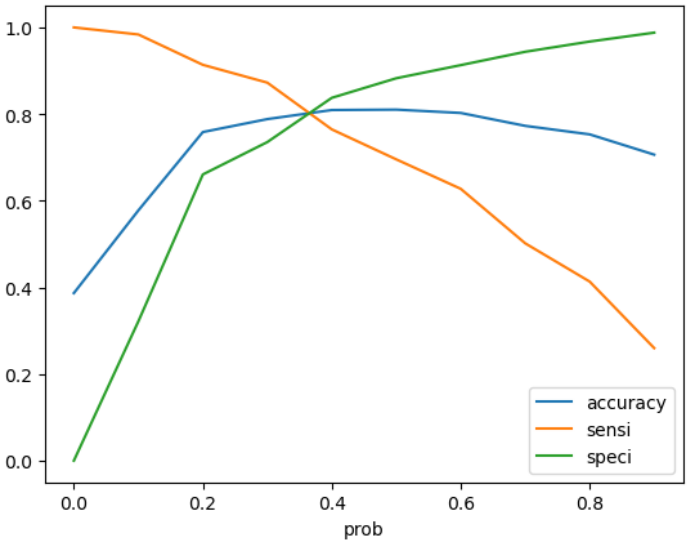
With numerical variables normalized and object type variables converted into dummy variables, the dataset is ready for further analysis and modeling. You can now proceed with building and validating the logistic regression model or any other classification technique based on your objectives and the characteristics of the data.

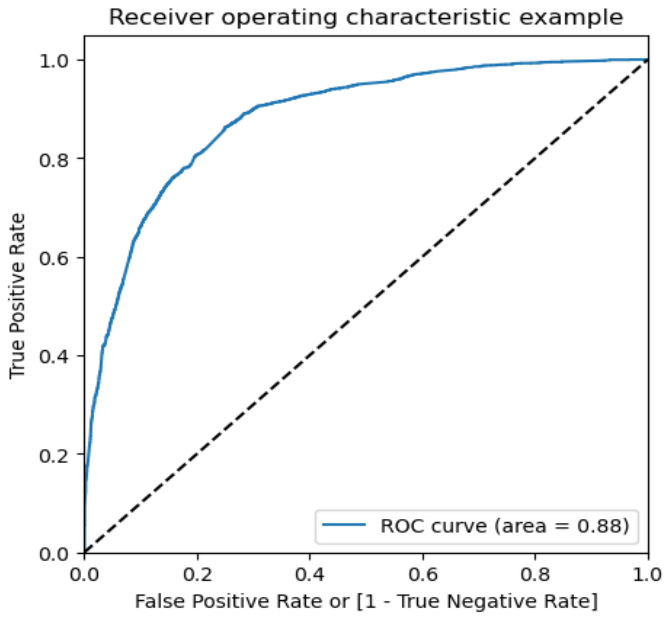
Keep in mind to handle class imbalances (if any) and perform thorough feature engineering to maximize the model's performance. Additionally, consider using appropriate evaluation metrics to assess the model's accuracy and performance on unseen data during the validation process.

Model Building

1. **Splitting the Data into Training and Testing Sets:**
    1. The dataset is divided into two subsets: the training set, which is used to train the model, and the testing set, which is used to evaluate the model's performance on unseen data. The chosen split ratio is 70:30, with 70% of the data used for training and 30% for testing.
2. **Using RFE for Feature Selection:**
    1. RFE (Recursive Feature Elimination) is a feature selection technique that recursively removes the least significant features from the dataset until the desired number of features is reached. It helps identify the most relevant features for building the model.
3. **Running RFE with 15 Variables as Output:**
    1. RFE is applied with the aim of selecting 15 most important variables as the model's input features. These 15 variables are considered to have the highest impact on the target variable.
4. **Building Model by Removing Insignificant Variables:**
    1. After RFE, the model is built using the selected 15 variables. However, some of these variables may have p-values greater than 0.05, indicating they are not statistically significant in predicting the target variable. Additionally, some variables may have high Variance Inflation Factor (VIF) values (greater than 5), suggesting potential multicollinearity issues. To address these concerns, the variables with high p-values and VIF values are removed from the model.
5. **Predictions on the Test Dataset:**
    1. After building the model with the selected significant variables, it is used to make predictions on the test dataset (30% of the original data). This allows for an evaluation of the model's performance on unseen data.
6. **Overall Accuracy of 81%:**
    1. The accuracy of the model, when applied to the test dataset, is found to be 81%. This indicates that approximately 81% of the model's predictions on the test data are correct.

- It's important to remember that model building is an iterative process, and the choice of variables to include or exclude should be based on various statistical tests, domain knowledge, and the specific objectives of the analysis. Additionally, further fine-tuning and validation are crucial to ensure the model's reliability and generalizability to new data.





Finding Optimal Cutoff

- Optimal cut-off probability balances sensitivity and specificity in a classification model.
- The chosen optimal cut-off is 0.35, providing a favorable trade-off between true positive and true negative rates.
- Consider the problem's context and misclassification costs when selecting the cut-off.
- Validate the model's performance on a separate dataset to ensure generalizability.

Case Study Conclusion

Based on the analysis, X Education can significantly improve its chances of converting potential buyers into actual customers by prioritizing certain key variables. These variables, ranked in descending order of importance, include:

1. The total time spent on the Website.
2. Total number of visits.
3. Lead source (Google, Direct traffic, Organic search, and Welingak website).
4. Last activity (SMS and Olark chat conversation).
5. Lead origin as Lead add format.
6. Current occupation as a working professional.

- By focusing on these crucial factors, X Education can fine-tune its marketing strategies, cater to potential buyers' preferences, and enhance its chances of success in enticing them to enroll in their courses. This strategic approach will likely lead to increased customer conversion rates and overall growth for the organization.