Lead Score Case Study

Group Members
1. Siddharth Chauhan
2. Nitin Kumar
3. Nibedita Sahoo

Problem Statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use**.**

Solution Methodology

- Data cleaning and data manipulation.
  1. Check and handle duplicate data.
  2. Check and handle NA values and missing values.
  3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
  4. Imputation of the values, if necessary.
  5. Check and handle outliers in data.
- EDA
  1. Univariate data analysis: value count, distribution of variable etc.
  2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

Data Manipulation

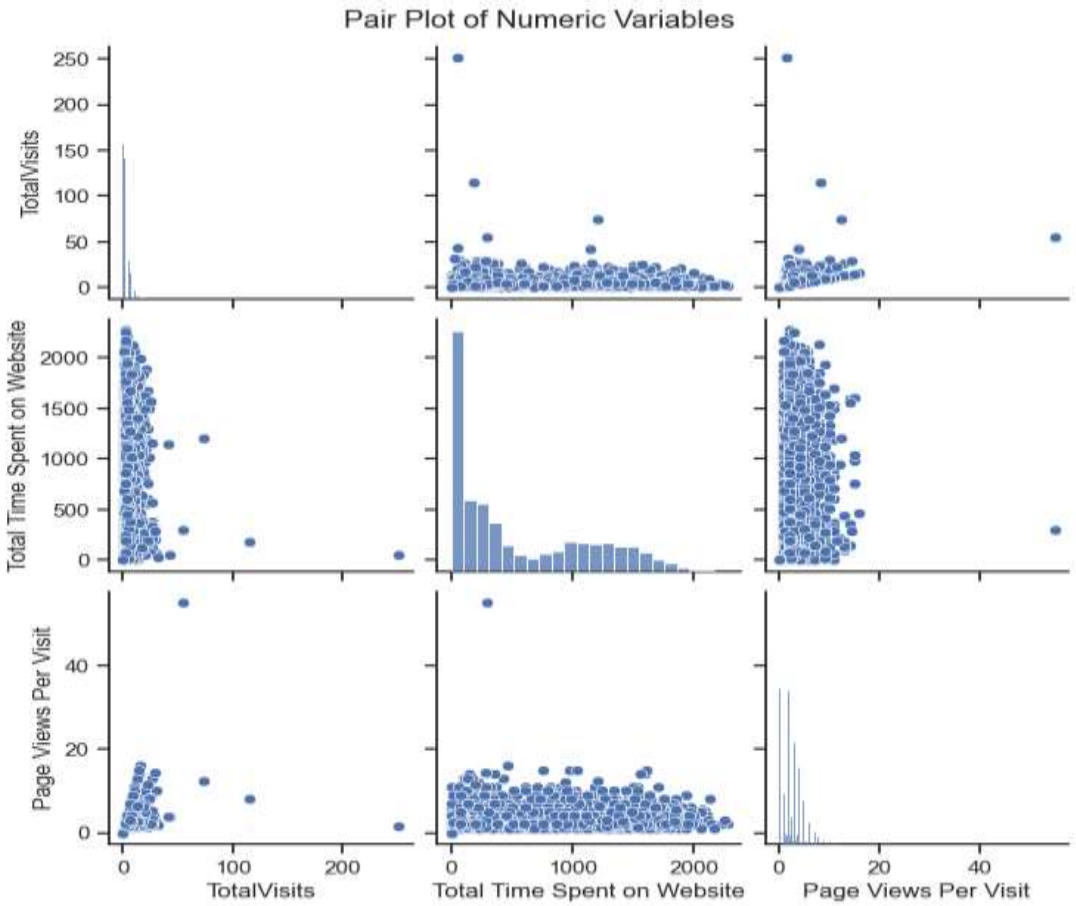
- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like : 'Tags', 'Lead Quality', 'Asymmetrique Activity Index', 'Asymmetrique Profile Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score' which have null values greater than 30% have been dropped.
- Removing 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: "Magazine" , "Search," "Newspaper Article," "X Education Forums," "Newspaper," "Digital Advertisement," "Through Recommendations," "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," and "What matters most to you in choosing a course" 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'.

Data Manipulation

- Based on the provided information, it appears that there are no columns with inappropriate data types, and there is no "Date_of_birth" or "Birthday" column to check for DATE TIME type. This suggests that the existing data types in the dataset are suitable and do not require any adjustments or conversions.
- After manipulation of the data and removing the columns that doesn’t give the insight for the analysis we have 6391 columns and 15 rows.

Further we need to visualize the data and make the updates on the dataset to get towards the analysis part / EDA.

EDA

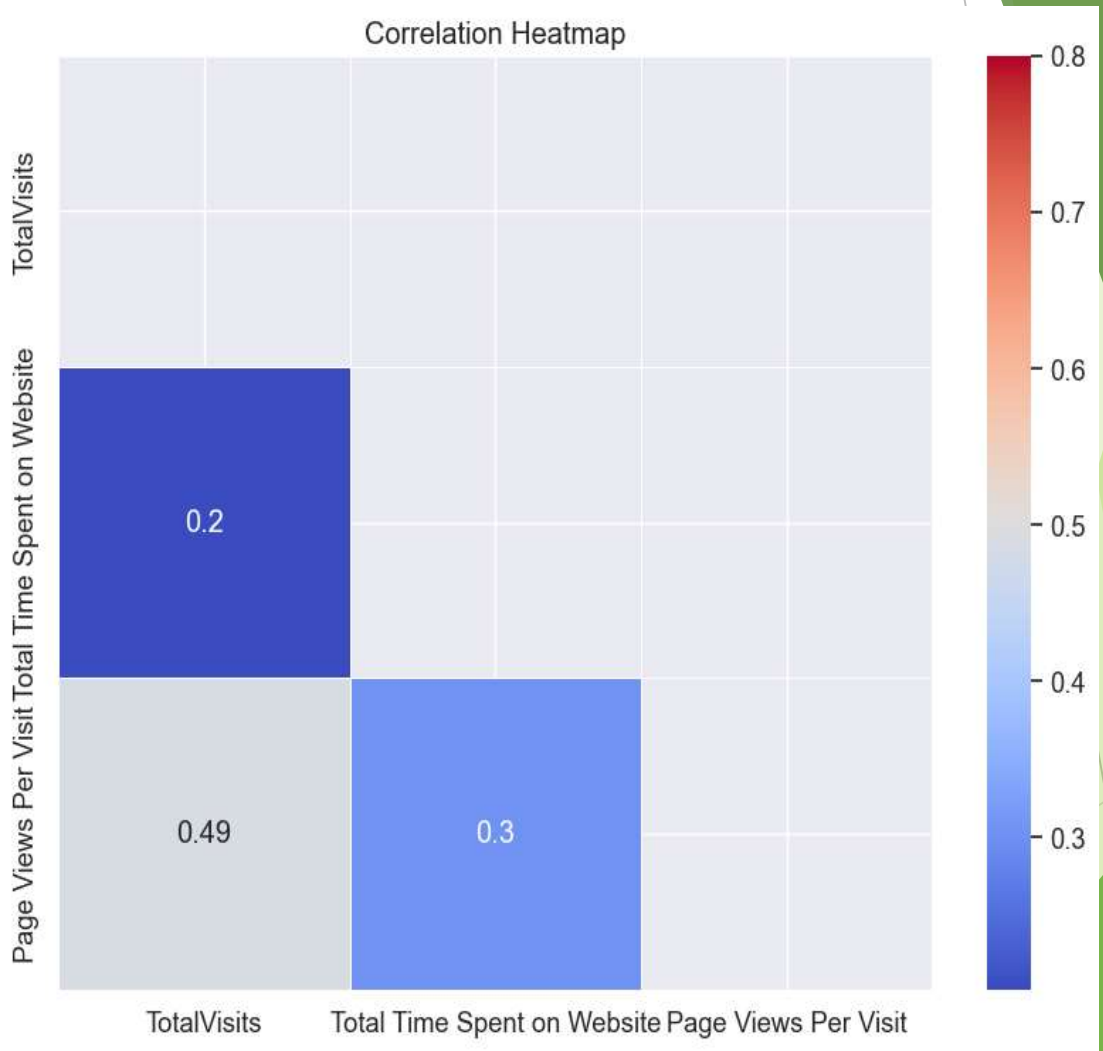


Drawing a pair plot to understand the relationship between the Numeric variables.

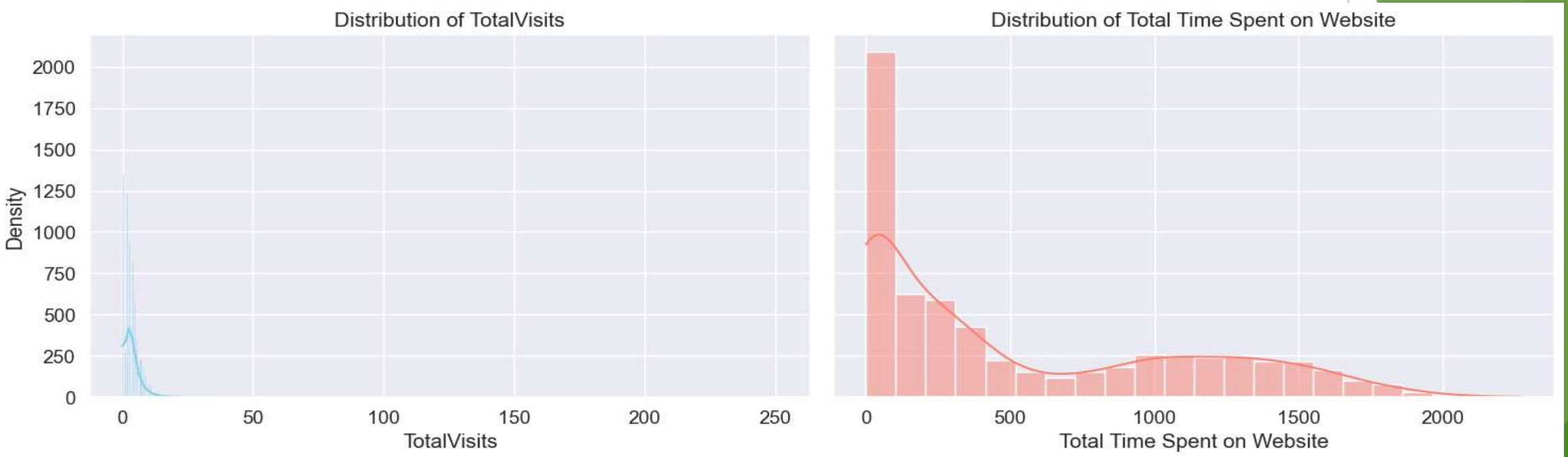
EDA

Using a Correlation Heatmap to check the correlation between the numeric variables : ‘TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit.

We observe that there is - No multicollinear relationship.

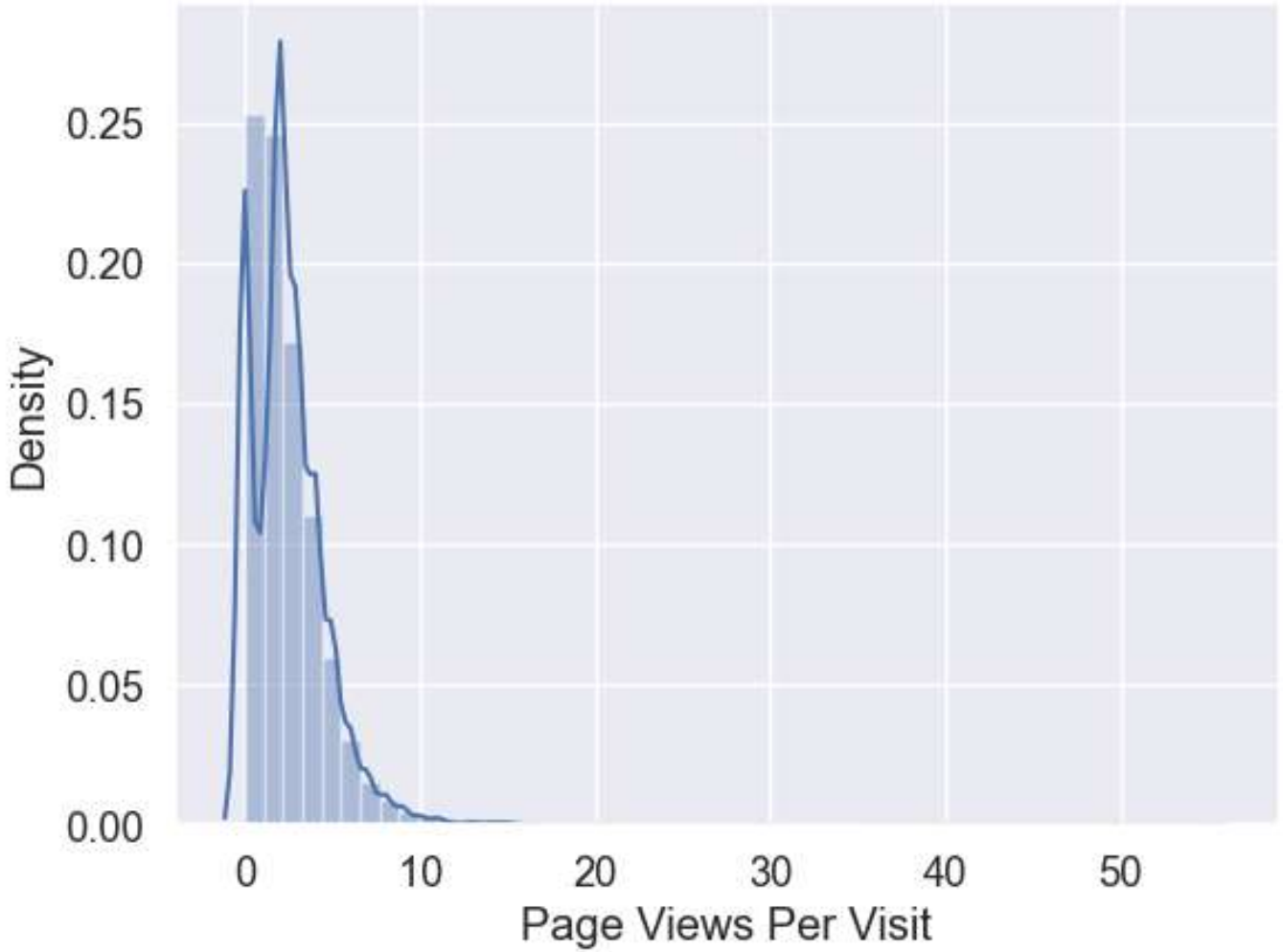


EDA



Using Subplots to display the Distribution of the Total Visits and Distribution of the Total Time Spent on the Website with respect to the density.

EDA



Creating a Distplot to display the Page Views Per visit vs Density insights.

EDA

Insights from the data analysis:

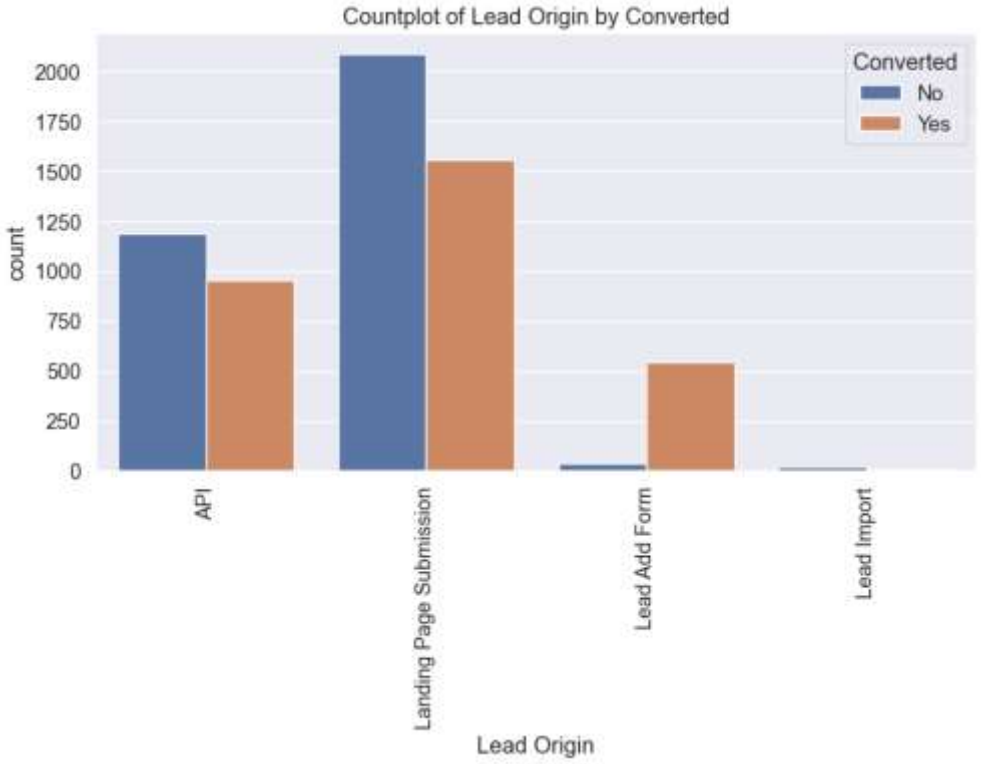
1. Visitor Frequency:
   - The majority of customers visited the website fewer than 20 times. This indicates that a significant portion of the customer base consists of infrequent visitors.

2. Page Views per Visit:
   - A large proportion of customers viewed less than 5 pages per visit. This suggests that many visitors may have specific content or tasks in mind, leading to shorter browsing sessions.

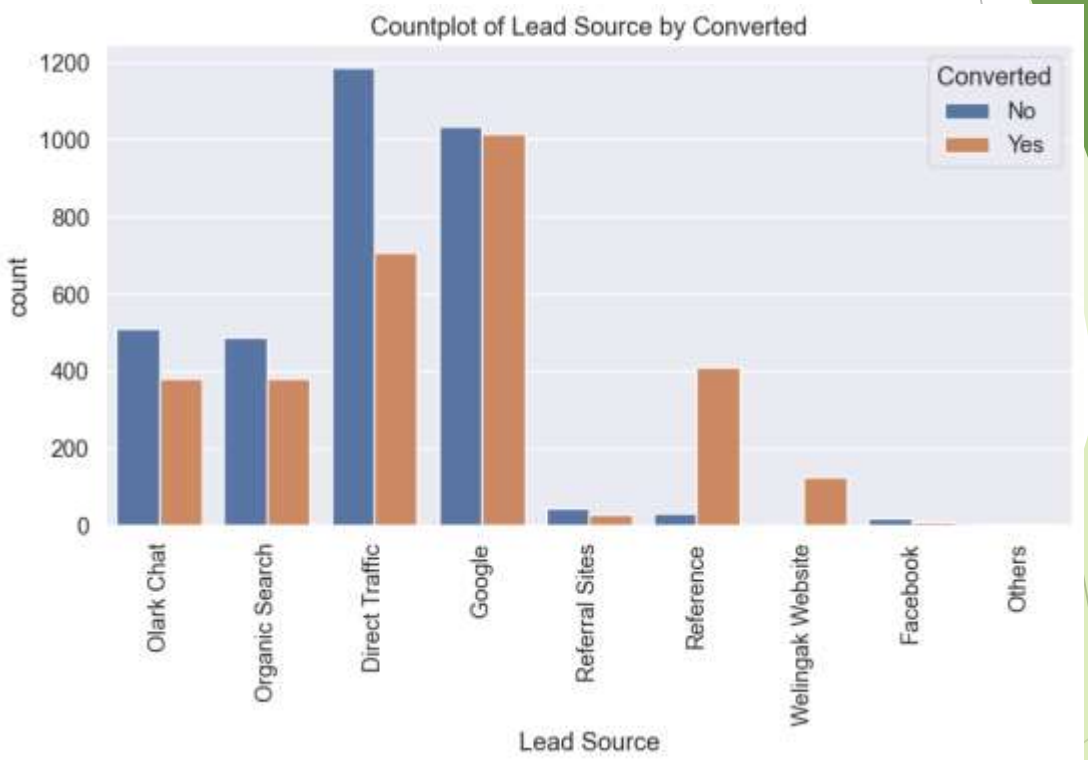
3. Total Time Spent on the Website:
   - Most customers spent less than 2,000 units of time (minutes or seconds) in total on the website. This indicates that the majority of visitors had relatively short engagement durations on the website.

These insights provide an overview of customer behavior on the website, highlighting trends in visitor frequency, page views, and time spent, which can be valuable for tailoring marketing strategies and website content to better meet customer needs.

Categorical Variable Relation

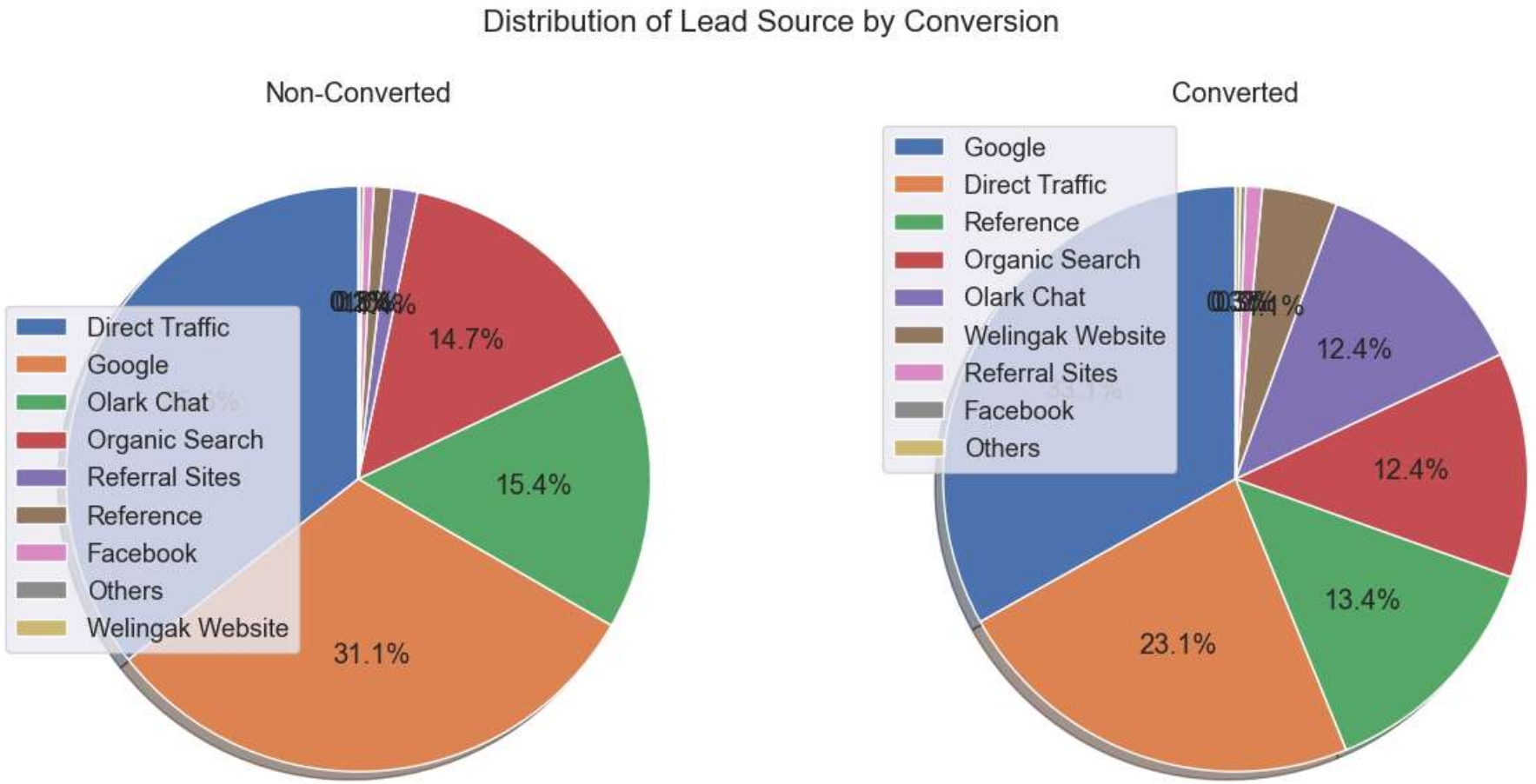


LEAD ORIGIN

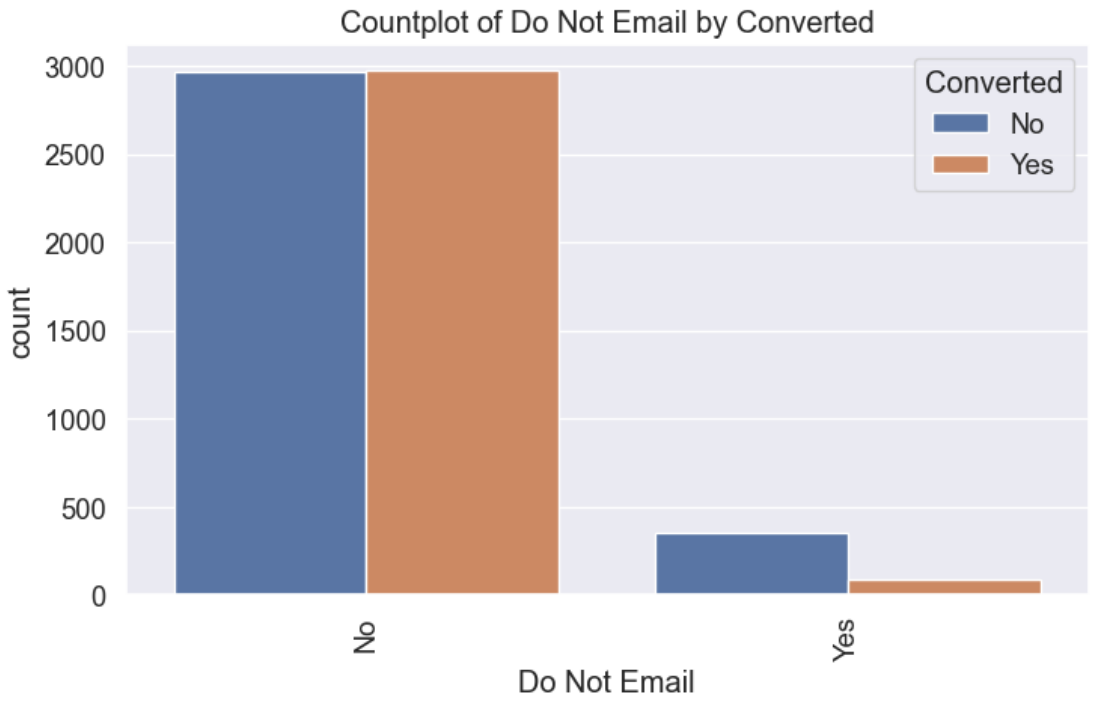


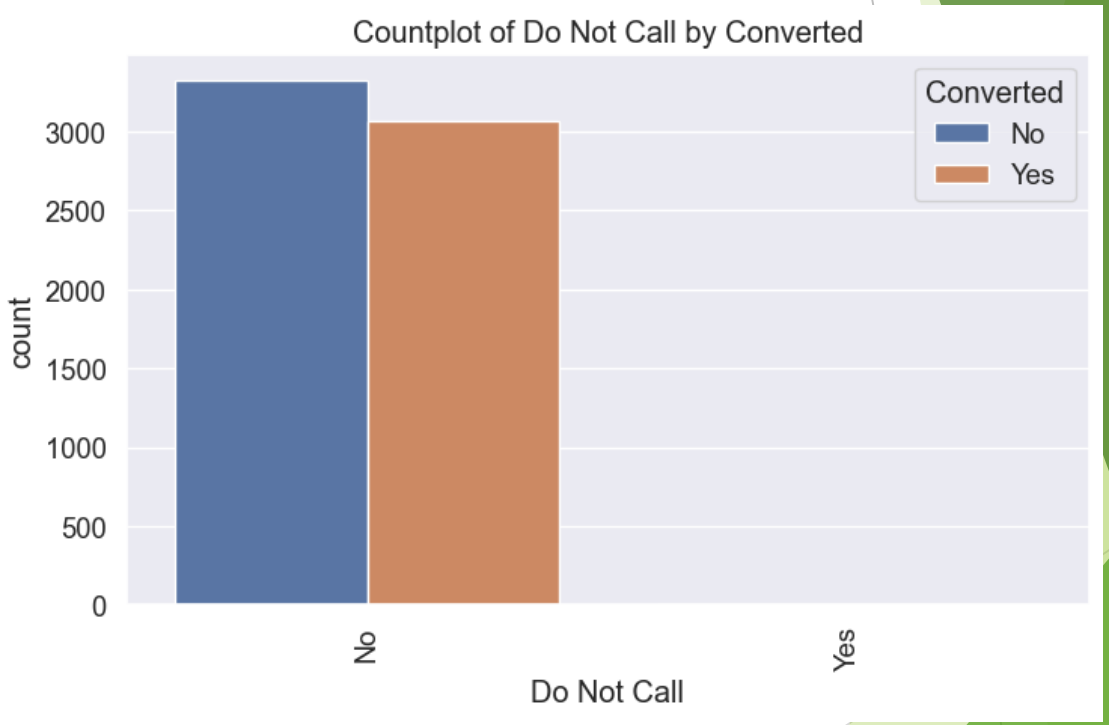
LEAD SOURCE

Categorical Variable Relation

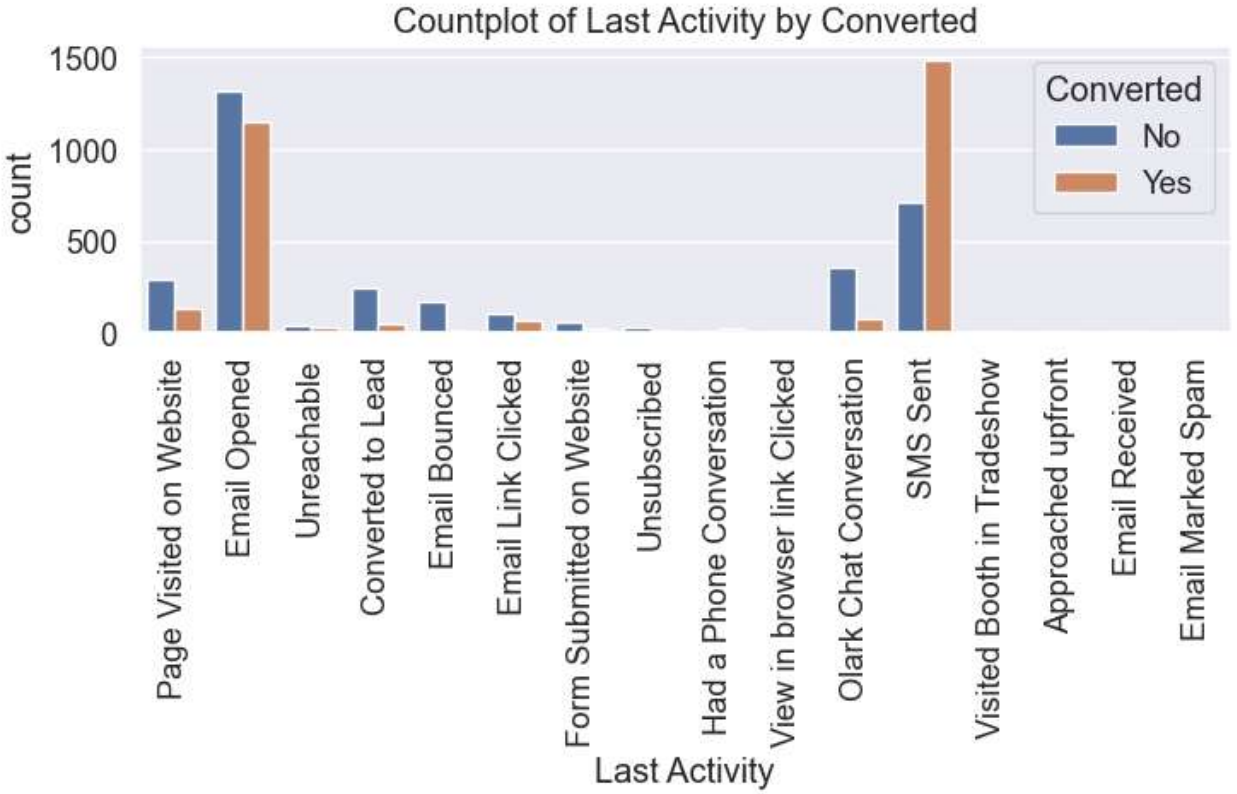


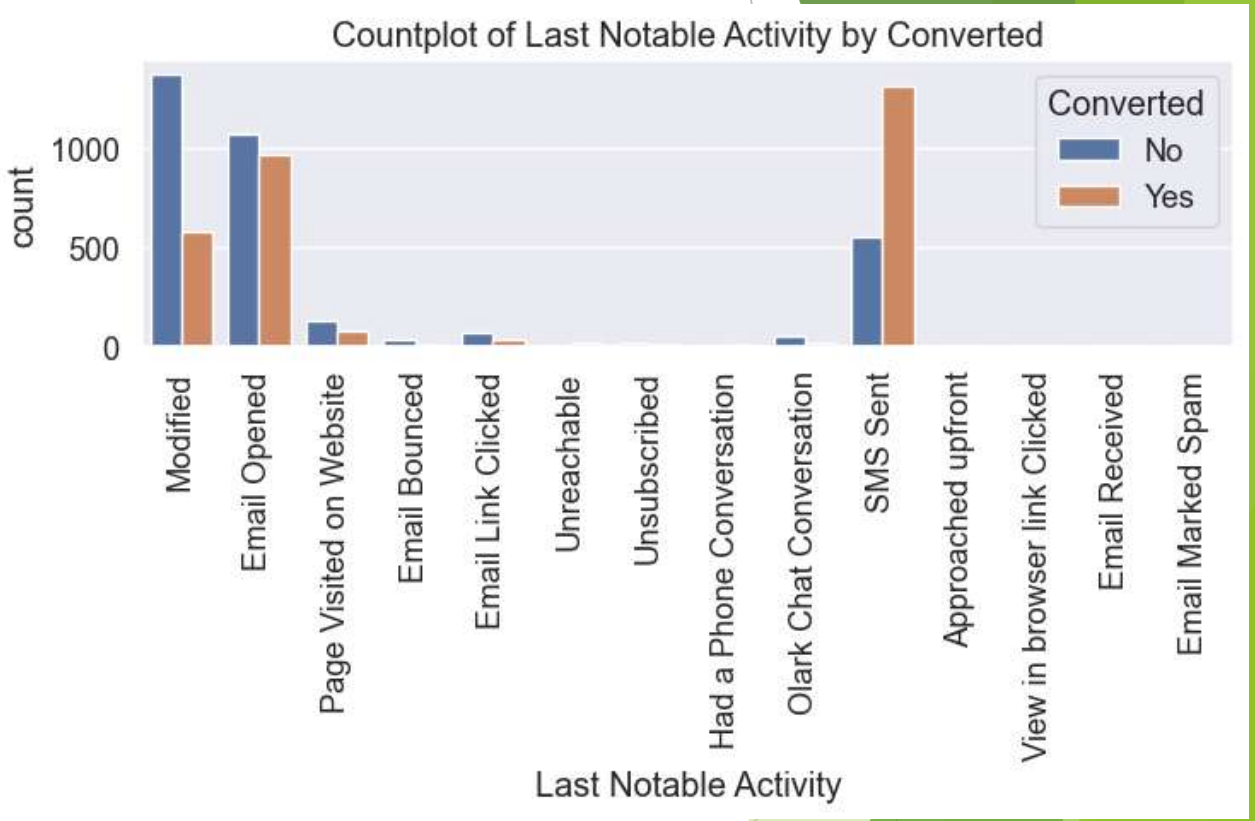
Do not email & Do not call:



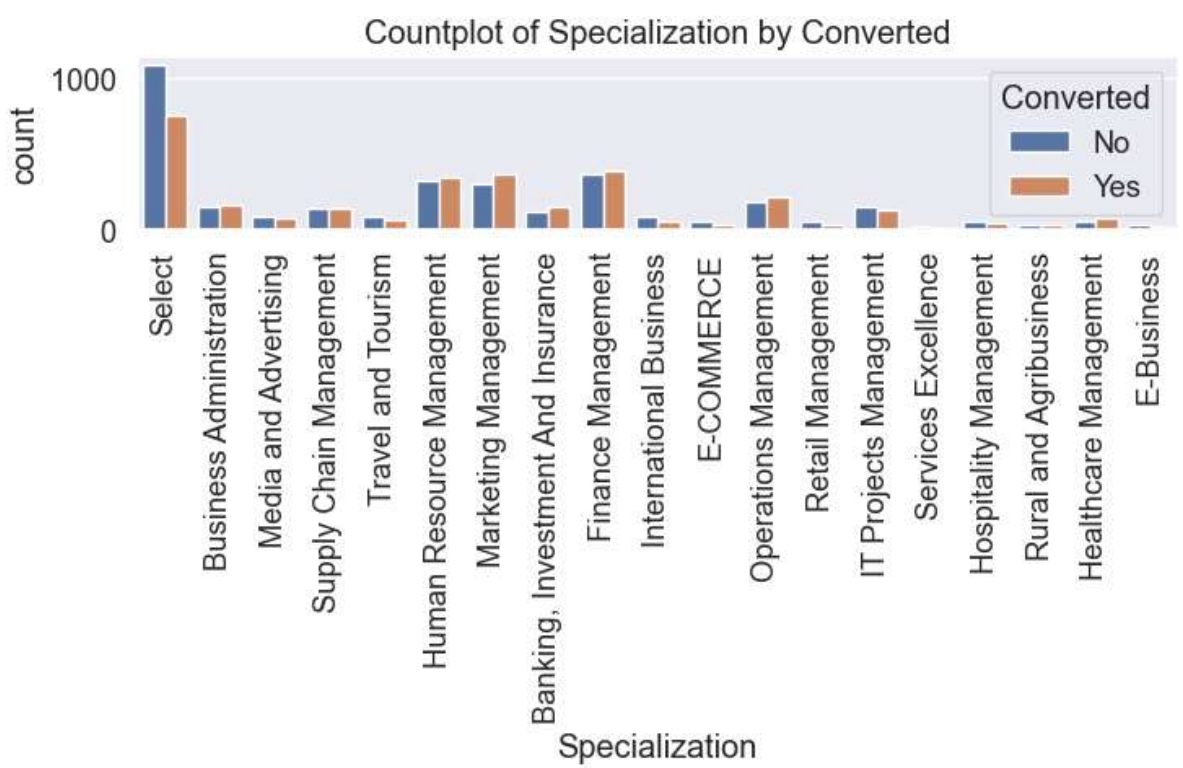


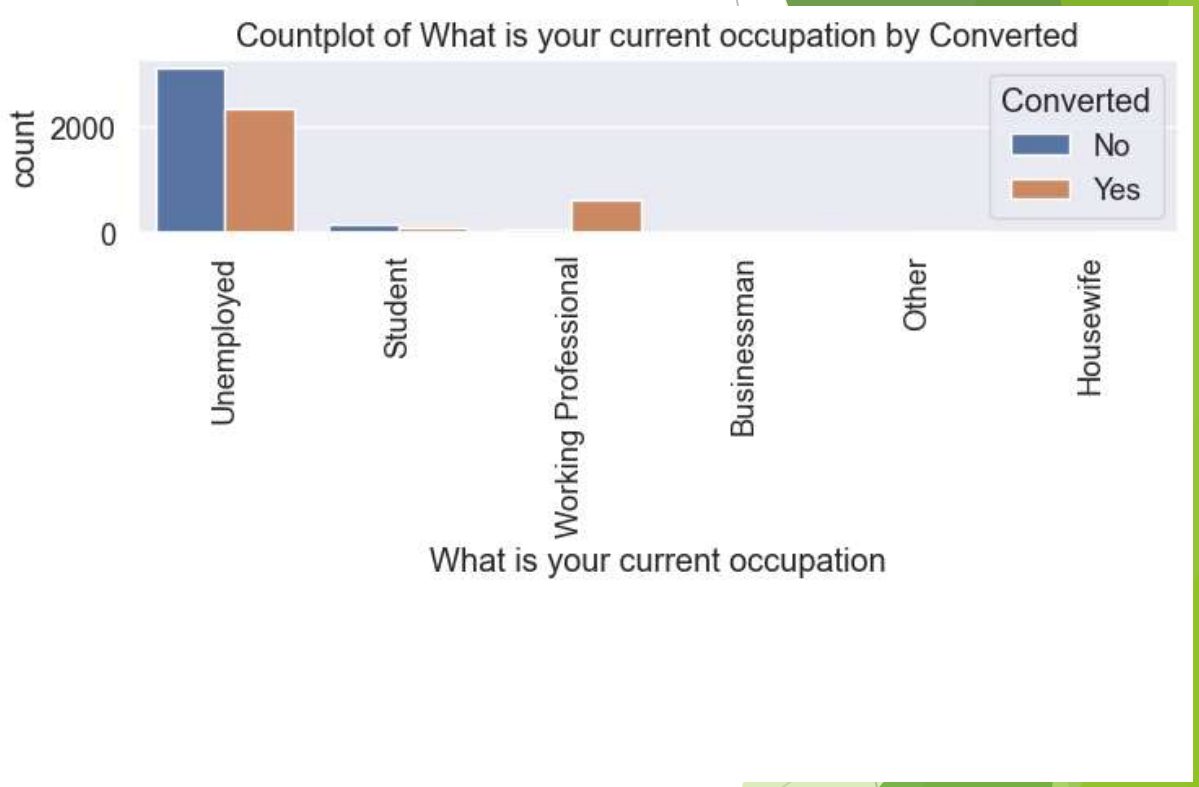
Last activity & Last notable activity:



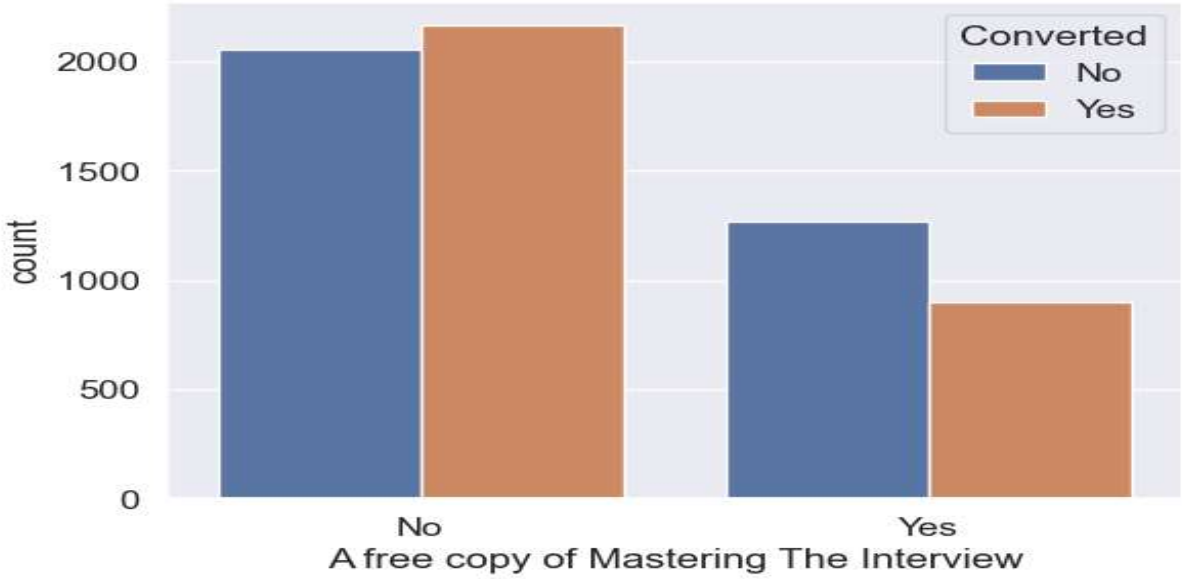


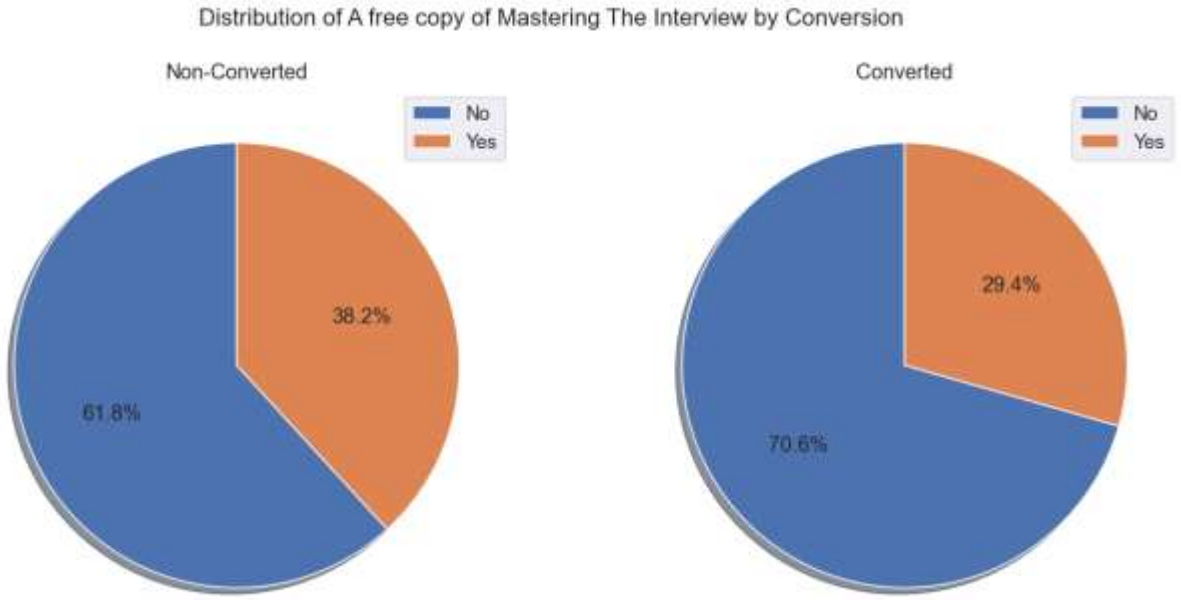
Occupation:


Countplot of Specialization by Converted
count
1000
0
Converted
No
Yes
Select
Business Administration
Media and Advertising
Supply Chain Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
Finance Management
International Business
E-COMMERCE
Operations Management
Retail Management
IT Projects Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization


Countplot of What is your current occupation by Converted
count
2000
0
Converted
No
Yes
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

A free copy of mastering the interview:





Insights from the data analysis:

1. Lead Origin:
- The majority of customers were identified as leads through three primary sources: API, Landing Page Submission, and Lead Add Form. Among these, Lead Add Form appears to be the most reliable identifier for determining converted customers.

2. Traffic Sources:
- The predominant sources of website traffic are Google and Direct Traffic. Interestingly, "Reference" and "Welingak Website" sources have a higher proportion of converted customers compared to non-converted customers, suggesting their effectiveness in driving conversions.

3. Contact Preferences:
- A significant number of customers, whether converted or not, have selected "Do not call" and "Do not email" as their contact preferences, indicating a preference for limited communication.

4. Last Activity:
- For converted customers, the last activity and last notable activity tend to be "Sent SMS," while for non-converted customers, the last activity is often "Opened Email." This suggests different engagement patterns for these two groups.

5. Employment Status:
- The majority of customers are unemployed, indicating a potential market segment that may require tailored marketing approaches.

6. Specialization:
- Most customers did not choose a specialization, while those who did primarily work in management fields. This information can be used to refine marketing strategies and course offerings.

7. Interest in Free Copy:
- Most customers, both converted and non-converted, do not express interest in receiving a free copy of "Mastering The Interview." This preference is consistent across both groups.

These insights provide valuable information for marketing and sales teams to refine their strategies and target the most promising leads effectively.

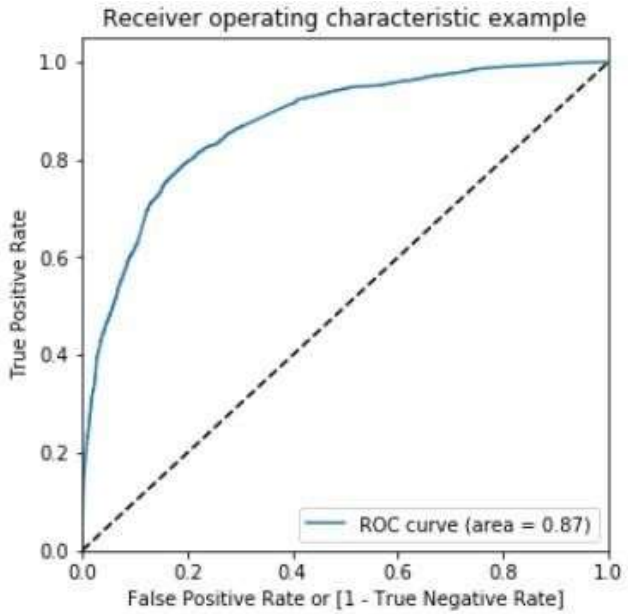
Data Conversion

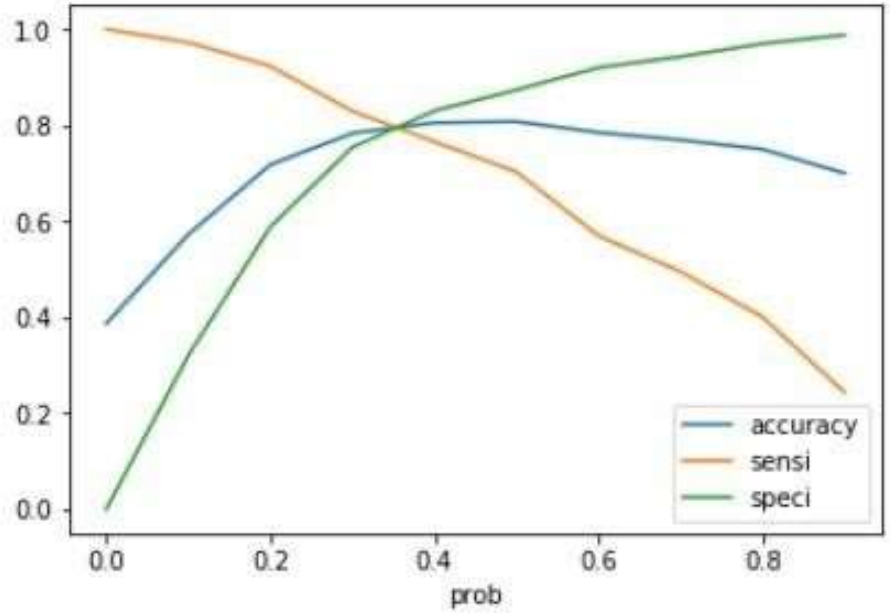
- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8792
- Total Columns for Analysis: 43

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 81%

ROC Curve





- **Finding Optimal Cut off Point**
- Optimal cut off probability is that
- probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.35.

Conclusion

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.
- When the lead source was:
  a. Google
  b. Direct traffic
  c. Organic search
  d. Welingak website
- When the last activity was:
  a. SMS
  b. 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.