LEAD SCORE CASE STUDY

GROUP MEMBER

SHREYA ARON
SHUBHRANGI SHARMA
SIDDHANT SINGH

TABLE OF CONTENT

PROBLEM STATEMENT

X Education, an online course provider for professionals, faces a low lead conversion rate of around 30%. Despite acquiring numerous leads through marketing campaigns and referrals, only a small percentage of these leads turn into paying customers. The company aims to improve its conversion rate to 80% by identifying "Hot Leads"—prospects with the highest potential to convert. This requires building a predictive model to assign a lead score to each prospect, enabling the sales team to prioritize and focus their efforts on nurturing the most promising leads through targeted communication and education.

SOLUTION METHODOLOGY

1. Importing Libraries and Data
   a) Import Libraries, Suppress Warnings and Set Display
   b) Reading the Data
2. Data Understanding and Inspection
3. Data Cleaning & Preparation
   a) Treatment for 'Select' values
   b) Handling Missing Values ,Outliers and unwanted data
4. Data Analysis (EDA)
   a) Univariate Analysis
   b) Bivariate Analysis
5. Data Preparation
   a) Dummy Variables

SOLUTION METHODOLOGY

6 Test-Train Split

7 Feature Scaling
a) Looking at Correlations

8 Model Building
a) Feature Selection Using RFE

9 Model Evaluation

10 Making Predictions on test set

DATA MANIPULATION

Total Number of Rows =37, Total Number of Columns =9240

The value 'Select' in the data is treated as equivalent to a null value. Dropping the columns having more than 40% as missing value and total 7 columns were dropped

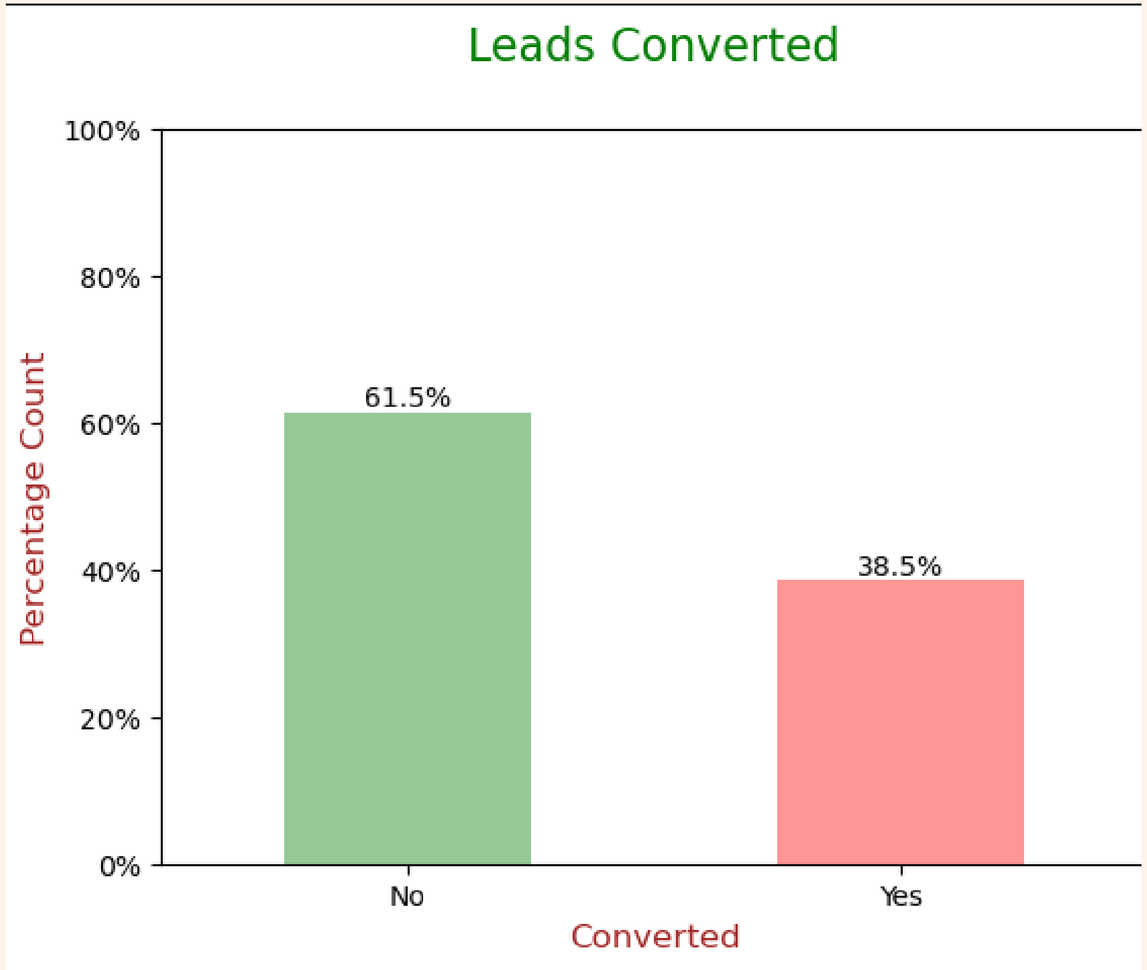
checked the count of values in each categorical column and then decide whether impute or drop the missing values for that particular column

Removing Unwanted Columns such as columns with only one unique value , Dropping columns of no use for modeling and Dropping Category Columns that are Skewed

Also handle the outliers in columns with numerical values and Fixing Invalid values and Standardising Data in columns

EDA

UNIVARIATE ANALYSIS



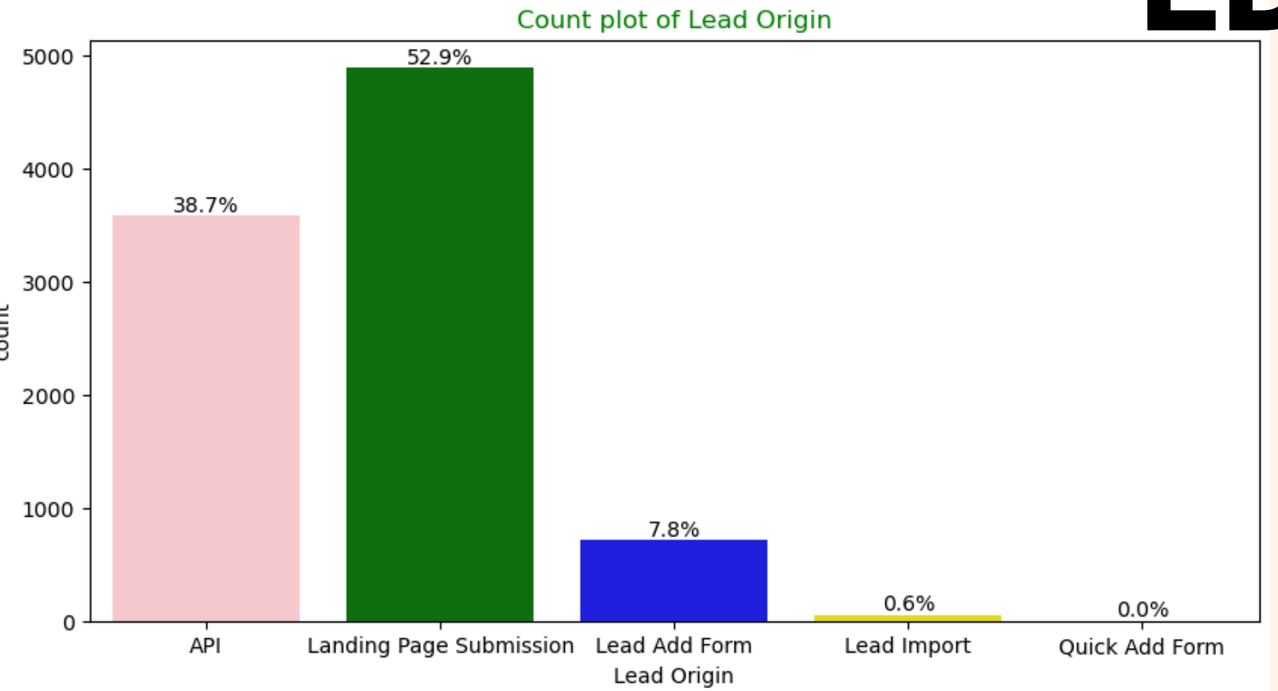
1. Lead Conversion Rate: About 38.5% of the leads are converted into paying customers, which aligns with the company's current average conversion rate.

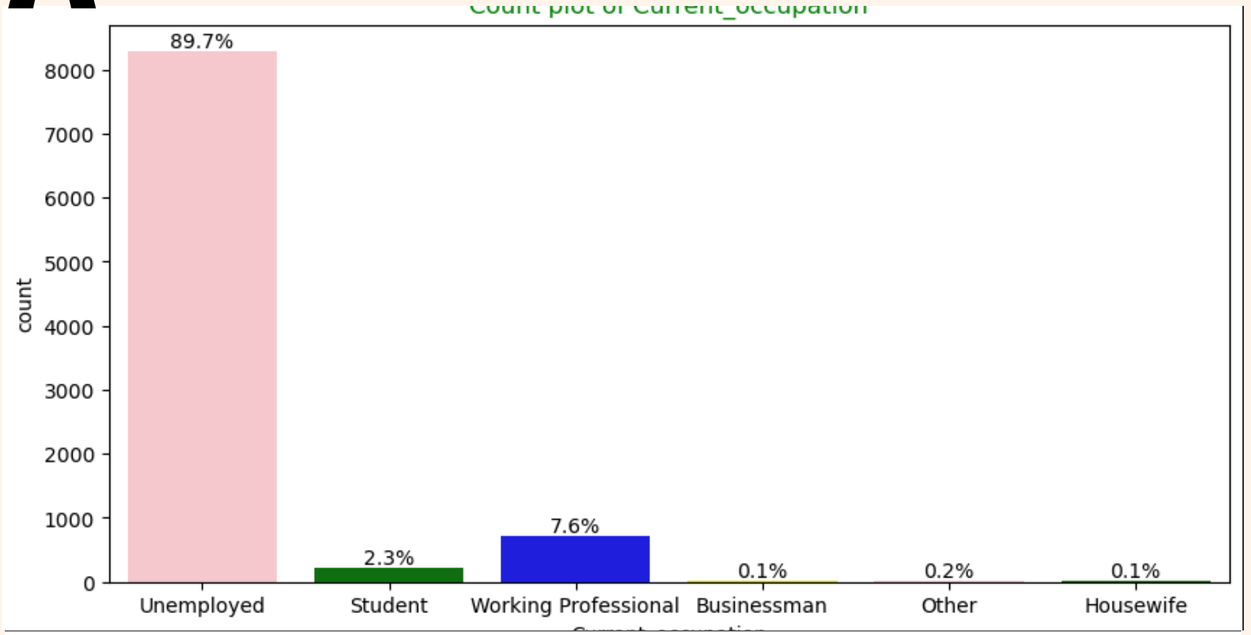
2. Non-converted Leads: A significant portion (61.5%) of leads do not convert, indicating inefficiencies in the current process.

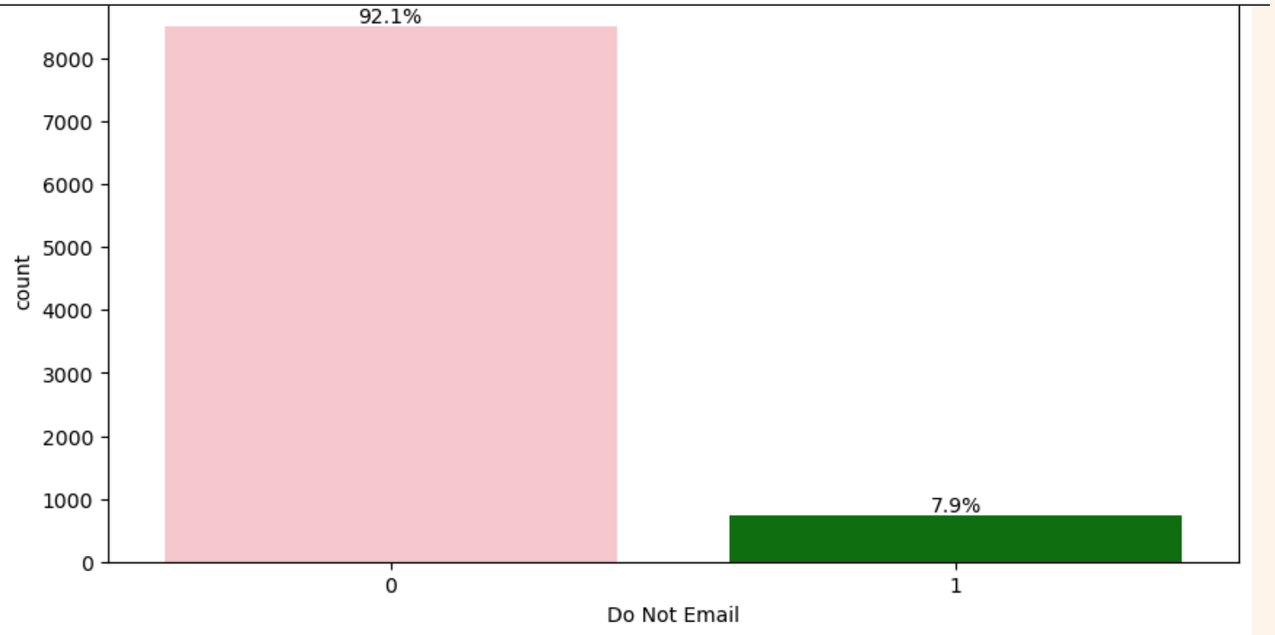
3. Visualization Enhancements: The use of color coding (green for "No" and red for "Yes") and percentage annotations provides a clear, intuitive understanding of conversion rates.

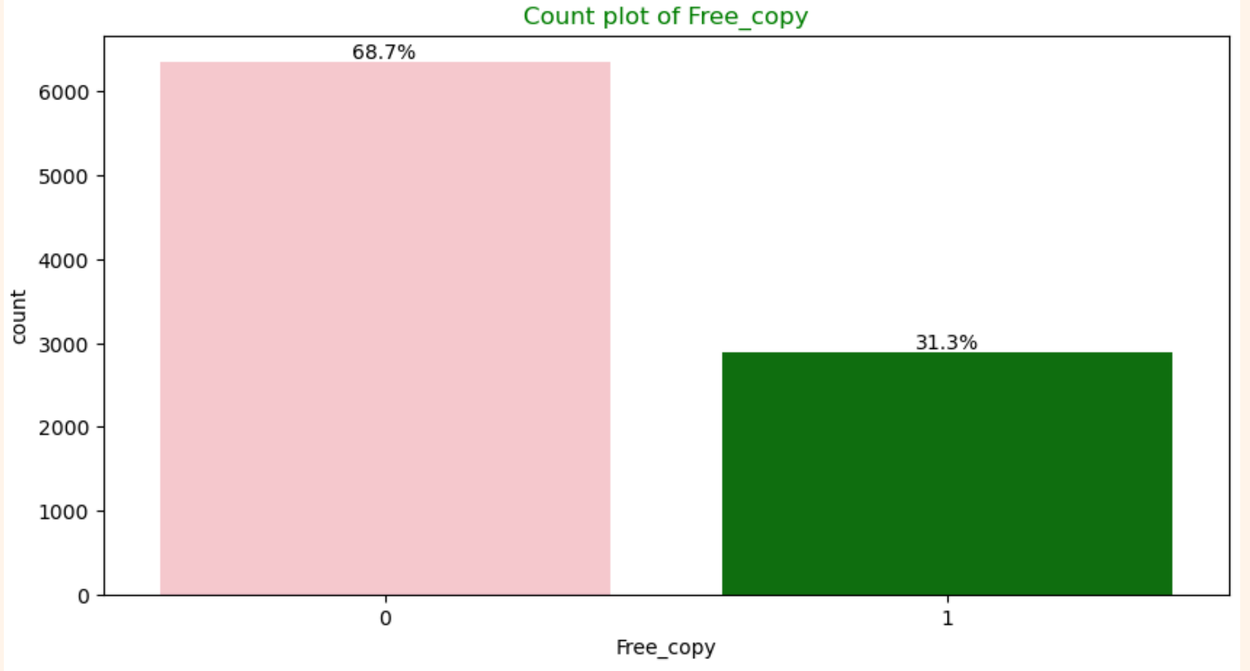
4.Opportunity for Improvement: The substantial gap between converted and non-converted leads highlights a clear area where focused efforts on potential leads could yield significant improvements in conversion.

EDA

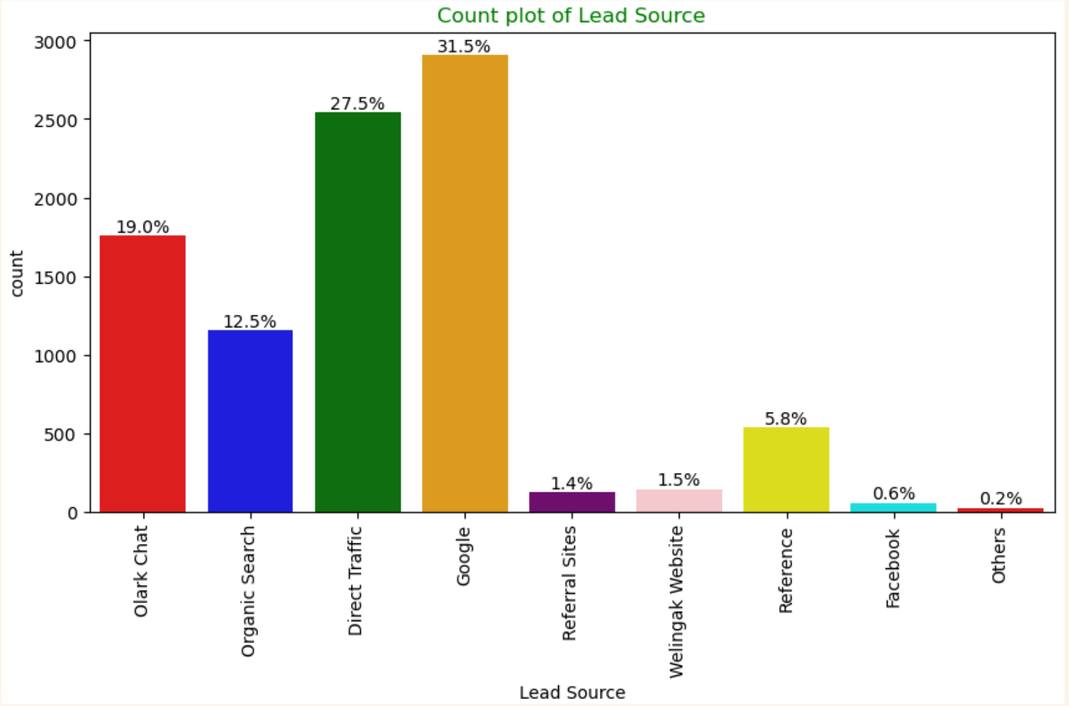

Count plot of Lead Origin
count
Lead Origin
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
38.7%
52.9%
7.8%
0.6%
0.0%
0
1000
2000
3000
4000
5000

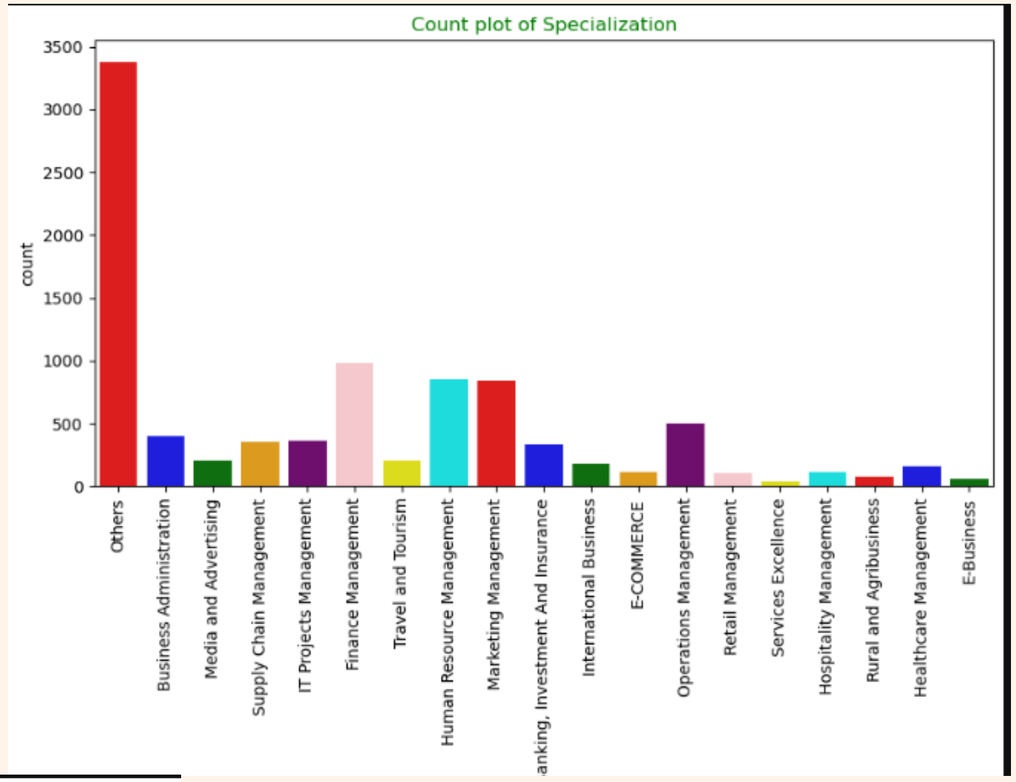

count
Unemployed
Student
Working Professional
Businessman
Other
Housewife
89.7%
2.3%
7.6%
0.1%
0.2%
0.1%
0
1000
2000
3000
4000
5000
6000
7000
8000


count
Do Not Email
0
1
92.1%
7.9%
0
1000
2000
3000
4000
5000
6000
7000
8000


Count plot of Free_copy
count
Free_copy
0
1
68.7%
31.3%
0
1000
2000
3000
4000
5000
6000

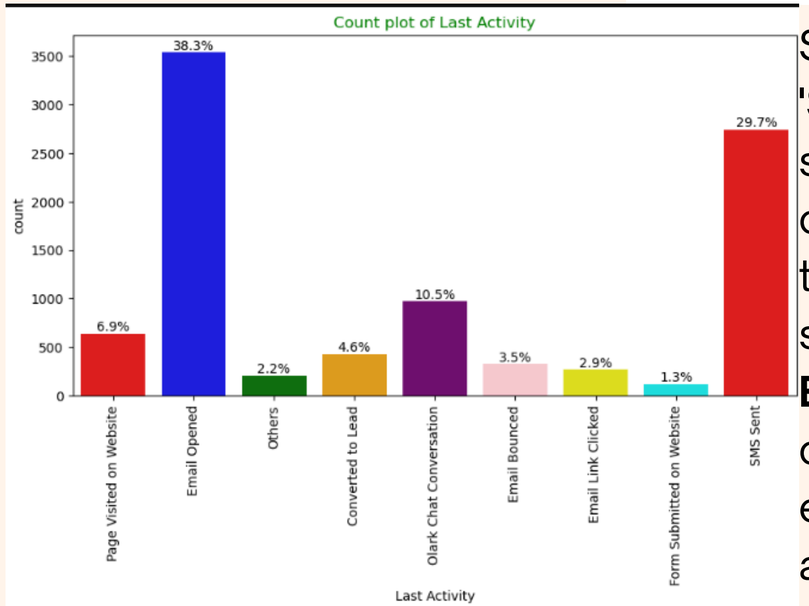
EDA





Distribution Diversity: The categorical columns showcase varied distributions, with certain categories dominating in some columns, indicating key trends in the dataset.

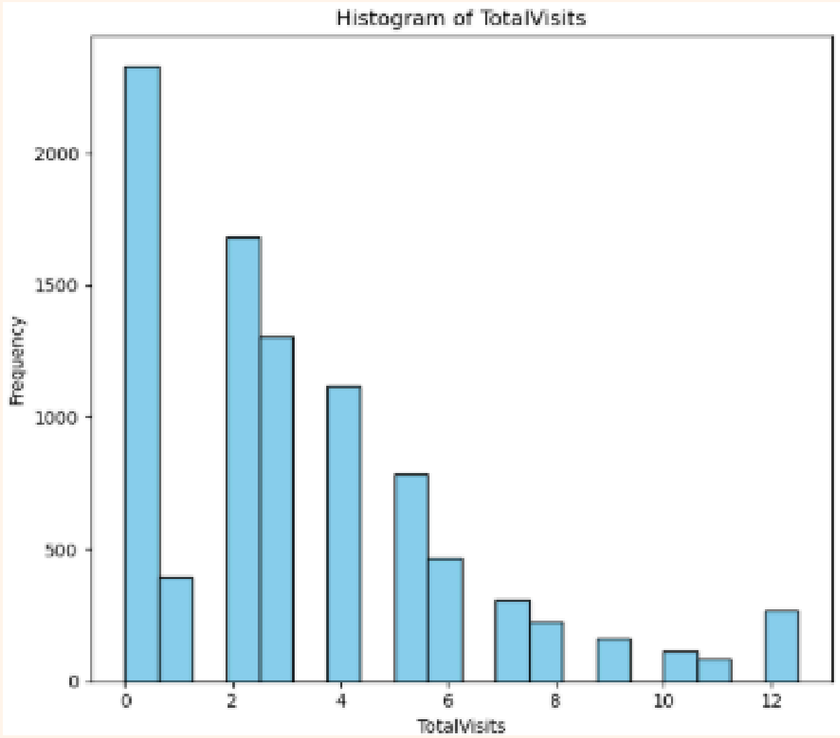
Lead Source Insights: Columns like 'Lead Source' highlight the most common channels through which leads are acquired, helping prioritize effective marketing strategies.



Specialization Concentration: The 'Specialization' column shows a significant skew towards certain specializations, with others having relatively low representation, though annotations for this column were skipped.

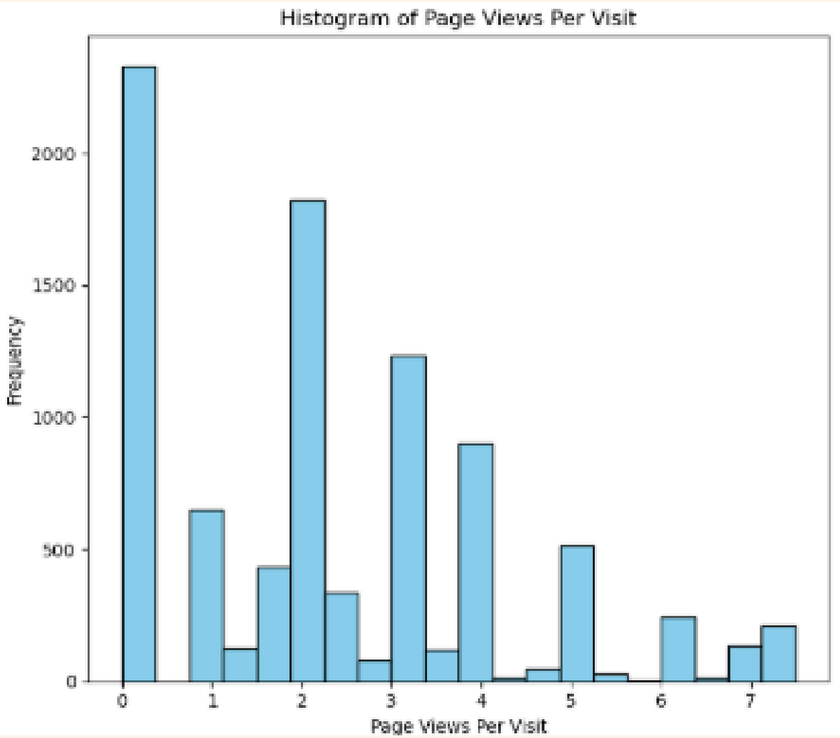
Enhanced Visualization: The use of a custom color palette and percentage annotations ensures clear, visually engaging insights, aiding in easier interpretation of trends.

EDA



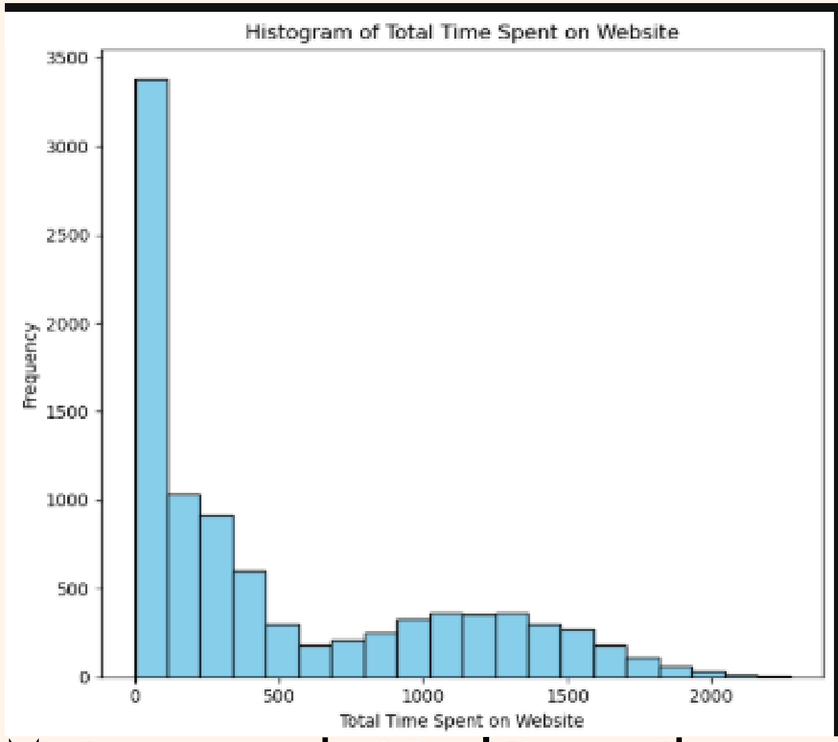
Most users have very few visits, with a significant spike at 0 or 1 visit.

The frequency decreases rapidly as the number of visits increases, indicating only a few users revisit the website frequently.



A majority of users view only 1 to 2 pages per visit, suggesting limited engagement with the website content.

There's a small subset of users who view multiple pages, indicating potential interest or intent.

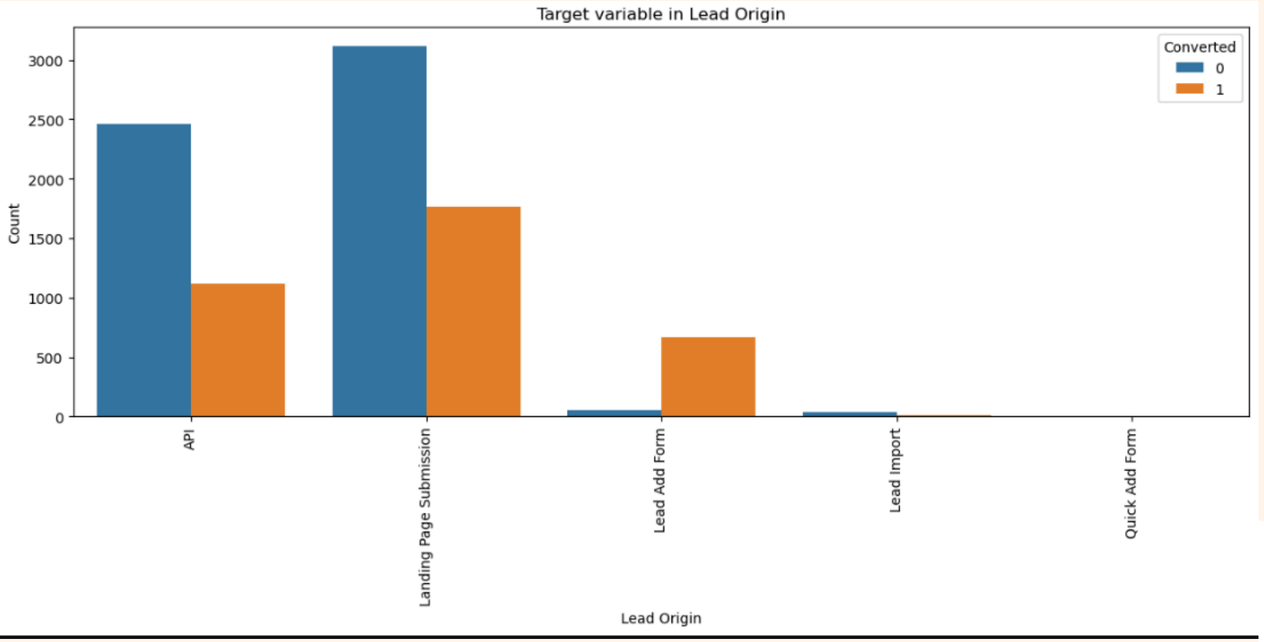


Most users spend minimal time on the website (0-100 seconds), reflecting quick exits or disinterest.

A long tail distribution suggests some users spend significantly more time, likely indicating higher engagement or exploration.

EDA

BIVARIATE ANALYSIS

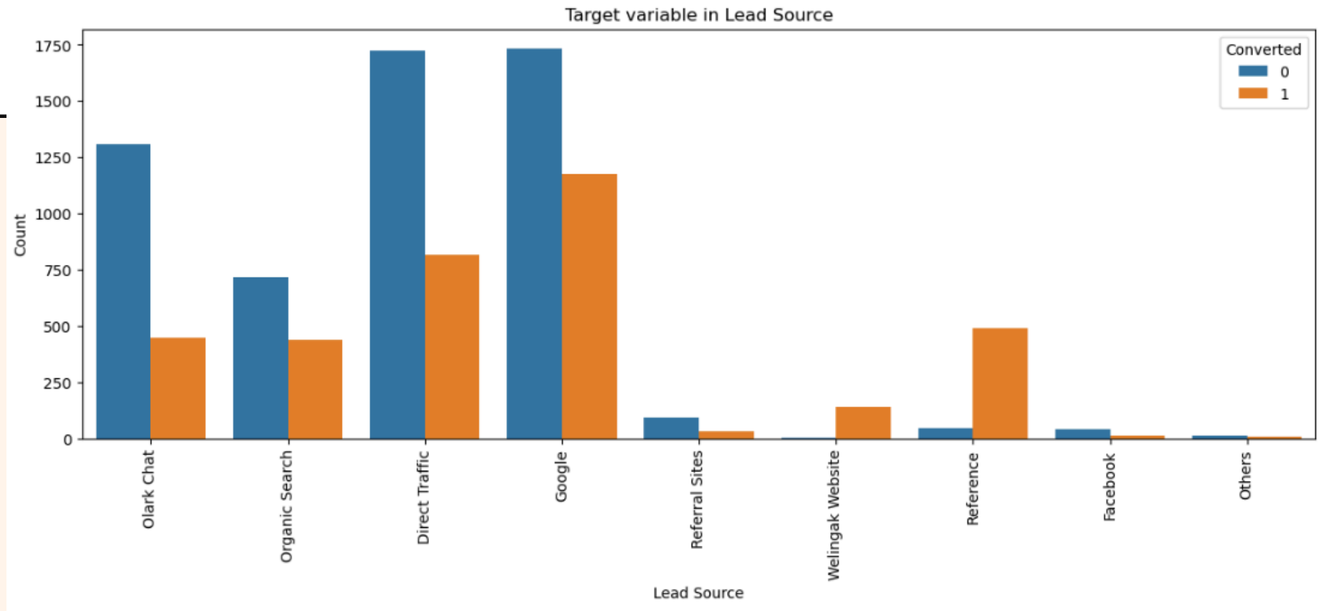


EFFECTIVENESS OF LANDING PAGE SUBMISSIONS: "LANDING PAGE SUBMISSION" IS THE MOST SIGNIFICANT LEAD ORIGIN, WITH A LARGE NUMBER OF LEADS AND A NOTICEABLE PORTION BEING CONVERTED. THIS INDICATES ITS STRONG PERFORMANCE AS A LEAD GENERATION CHANNEL.

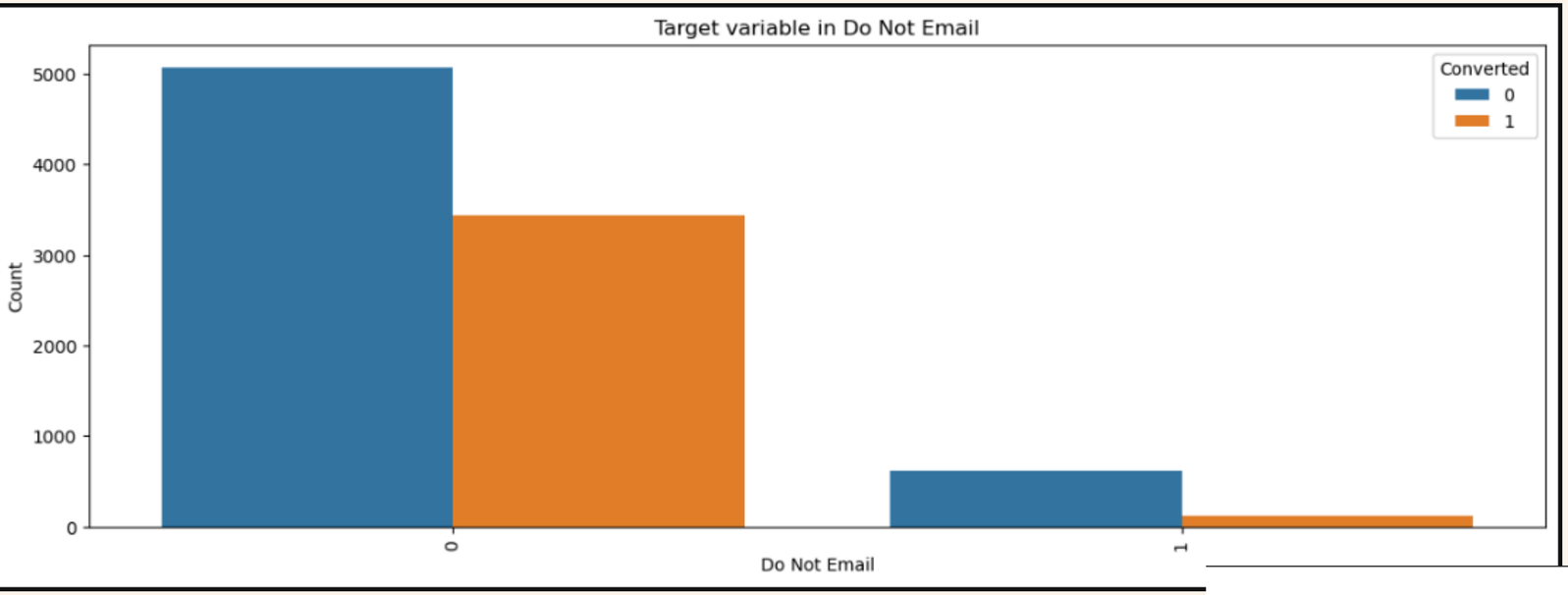
HIGH CONVERSION RATE FOR LEAD ADD FORM: ALTHOUGH THE TOTAL COUNT OF LEADS FROM THE "LEAD ADD FORM" IS LOWER, IT DEMONSTRATES A HIGHER PROPORTION OF CONVERSIONS. THIS SUGGESTS THAT LEADS FROM THIS SOURCE ARE MORE LIKELY TO CONVERT, MAKING IT AN EFFICIENT CHANNEL DESPITE ITS SMALLER VOLUME.

STRONG PERFORMANCE OF GOOGLE AS A LEAD SOURCE: GOOGLE CONTRIBUTES SIGNIFICANTLY TO BOTH CONVERTED (ORANGE BAR) AND NON-CONVERTED (BLUE BAR) LEADS, WITH A NOTICEABLE NUMBER OF CONVERSIONS. THIS HIGHLIGHTS GOOGLE AS A DOMINANT AND IMPACTFUL LEAD SOURCE IN THE OVERALL FUNNEL.

HIGH CONVERSION RATE OF REFERENCE LEADS: DESPITE HAVING A LOWER TOTAL COUNT COMPARED TO SOURCES LIKE GOOGLE OR DIRECT TRAFFIC, THE REFERENCE CATEGORY SHOWS A RELATIVELY HIGHER PROPORTION OF CONVERTED LEADS. THIS INDICATES THAT REFERENCES ARE A HIGHLY EFFICIENT SOURCE FOR DRIVING CONVERSIONS.



EDA

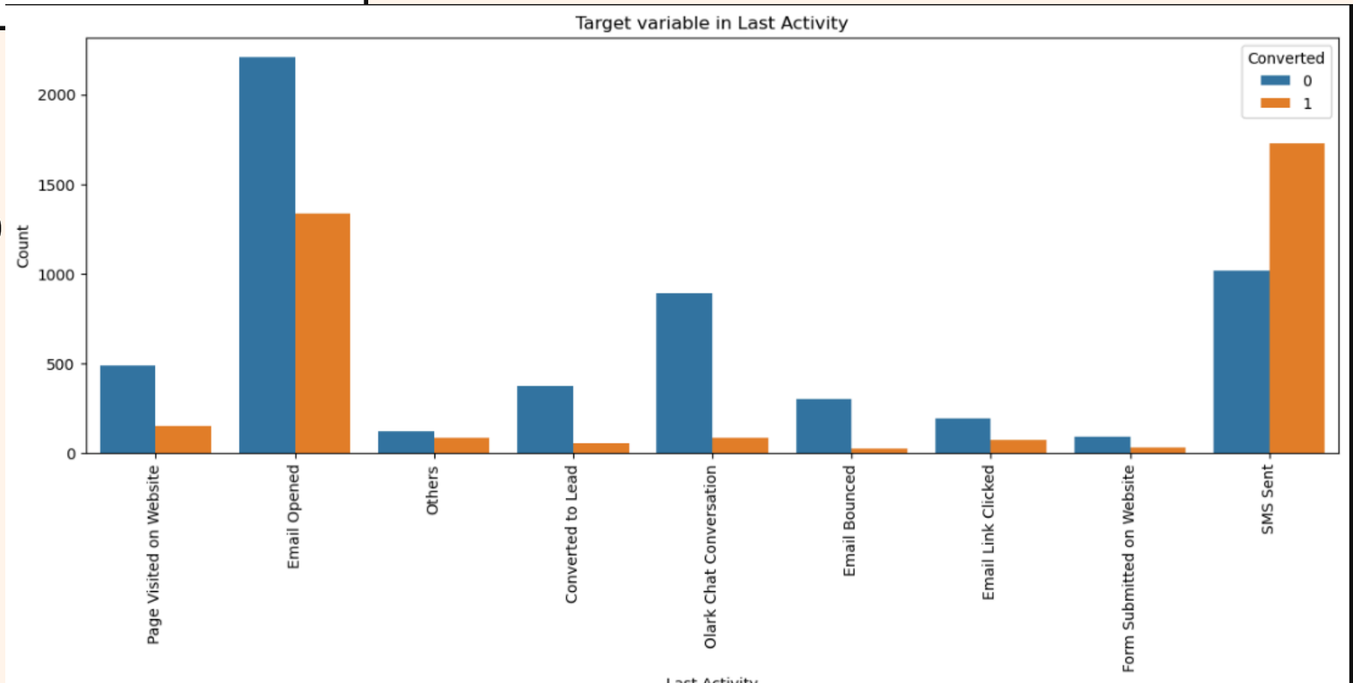


PREFERENCE FOR COMMUNICATION VIA EMAIL: IN THE "DO NOT EMAIL = 0" CATEGORY, THE COUNT OF INDIVIDUALS WHO DID NOT CONVERT (BLUE) IS HIGHER THAN THOSE WHO CONVERTED (ORANGE). THIS SUGGESTS THAT MOST INDIVIDUALS WHO ALLOW EMAIL COMMUNICATION ARE LESS LIKELY TO CONVERT.

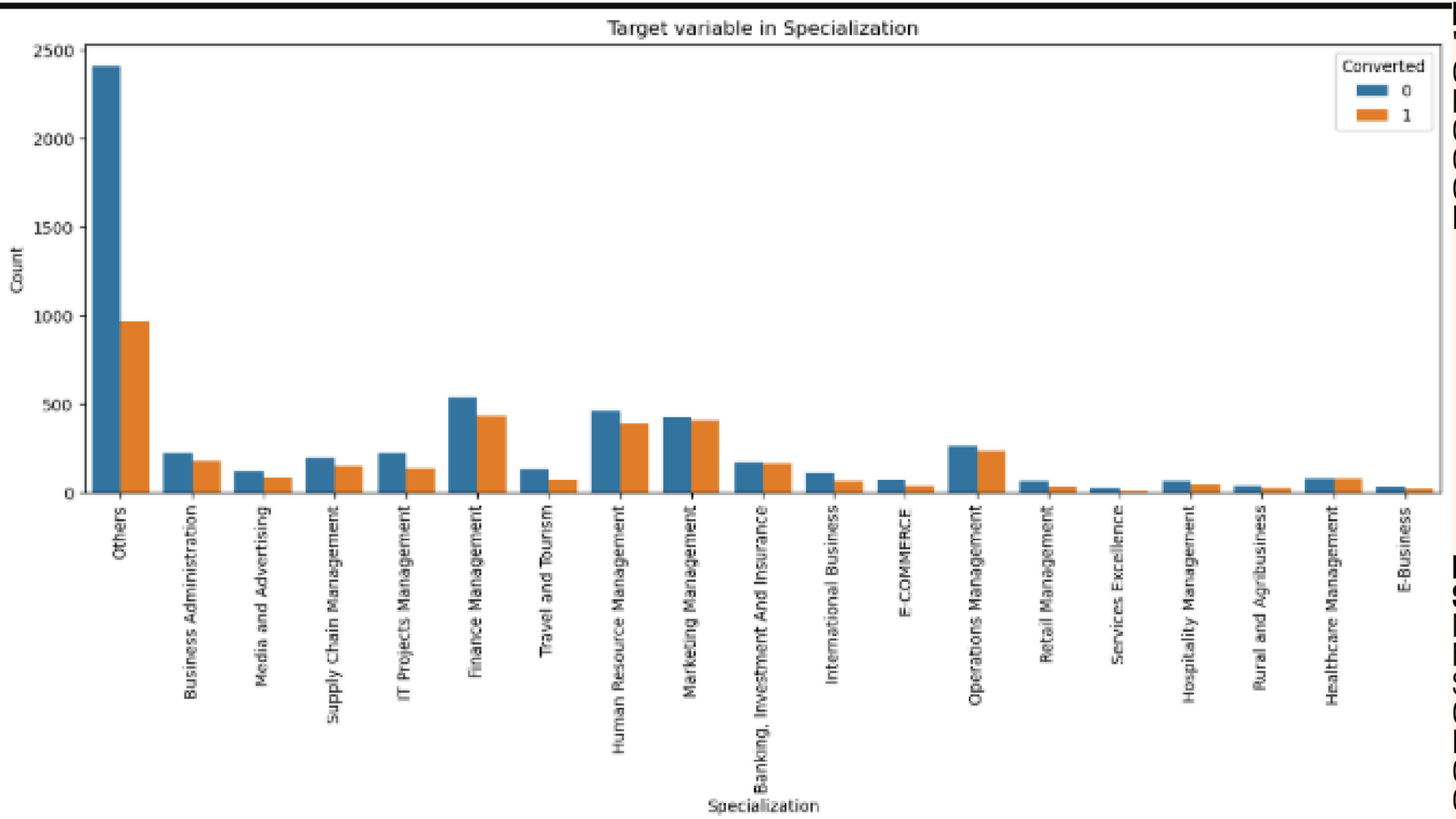
LIMITED DATA IN "DO NOT EMAIL = 1": THE "DO NOT EMAIL = 1" CATEGORY SHOWS SIGNIFICANTLY FEWER INDIVIDUALS OVERALL, WITH A SLIGHTLY HIGHER COUNT OF NON-CONVERTED LEADS. THIS INDICATES THAT FEWER USERS OPT OUT OF EMAIL COMMUNICATION, MAKING IT LESS IMPACTFUL IN THE CONTEXT OF CONVERSIONS.

"EMAIL OPENED" DOMINANCE: THE "EMAIL OPENED" CATEGORY HAS THE HIGHEST OVERALL COUNT OF ACTIVITIES, REGARDLESS OF CONVERSION. HOWEVER, THE MAJORITY OF LEADS IN THIS CATEGORY ARE NOT CONVERTED (BLUE BAR), INDICATING THAT SIMPLY OPENING AN EMAIL DOES NOT STRONGLY CORRELATE WITH CONVERSIONS.

EFFECTIVENESS OF SMS SENT: THE "SMS SENT" CATEGORY SHOWS A HIGHER PROPORTION OF CONVERTED LEADS (ORANGE BAR) COMPARED TO NON-CONVERTED LEADS. THIS SUGGESTS THAT SENDING AN SMS AS THE LAST ACTIVITY IS MORE EFFECTIVE IN DRIVING CONVERSIONS COMPARED TO OTHER ACTIVITIES.



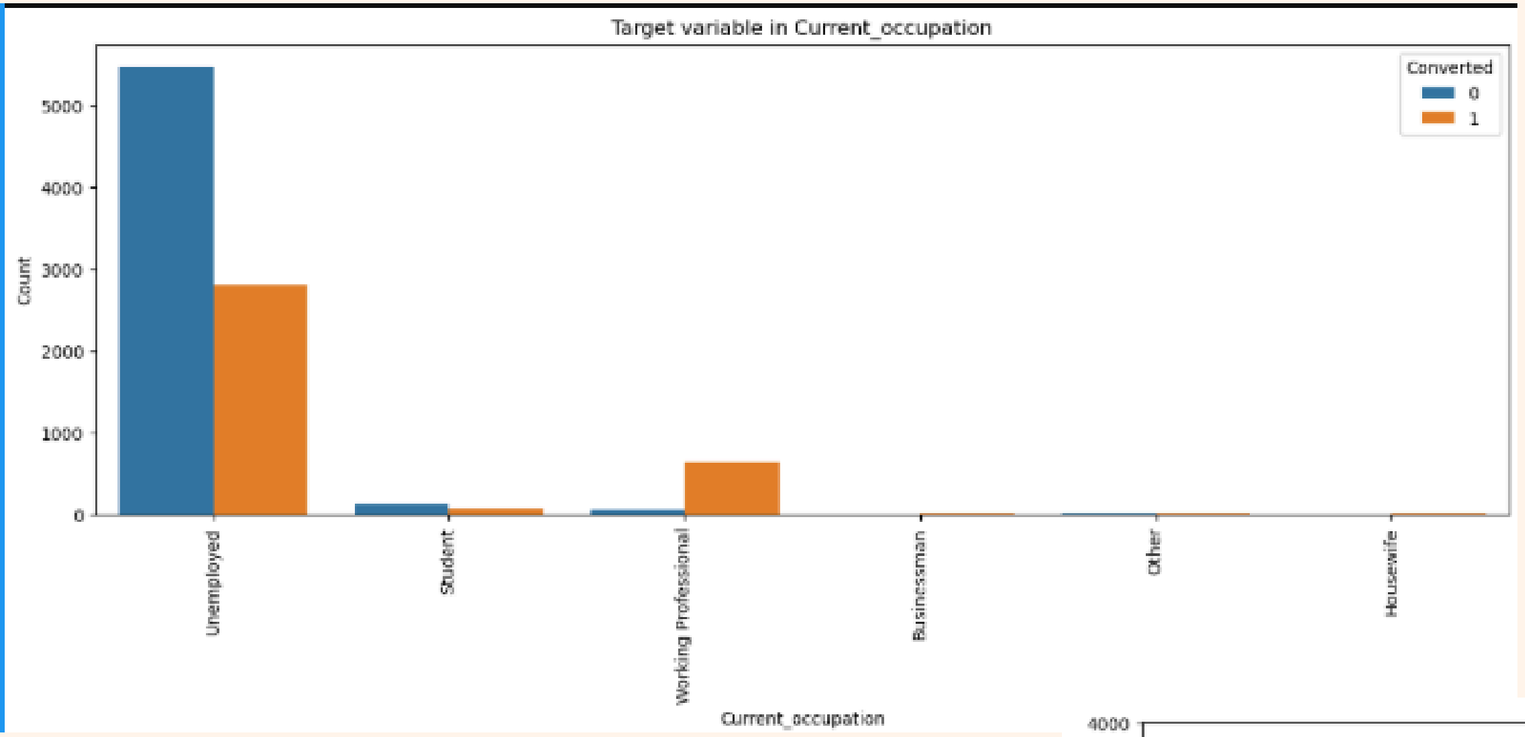
EDA



DOMINANCE OF "OTHERS" SPECIALIZATION: THE "OTHERS" SPECIALIZATION HAS THE HIGHEST COUNT OF INDIVIDUALS, WITH A SUBSTANTIAL PORTION OF BOTH CONVERTED AND NON-CONVERTED USERS. THIS SUGGESTS THAT THIS CATEGORY HAS BROAD REPRESENTATION AND COULD BE KEY TO IDENTIFYING TRENDS OR IMPROVEMENT OPPORTUNITIES.

BALANCED CONVERSIONS IN CERTAIN SPECIALIZATIONS: FIELDS LIKE "FINANCE MANAGEMENT," "HUMAN RESOURCE MANAGEMENT," AND "MARKETING MANAGEMENT" SHOW A RELATIVELY BALANCED DISTRIBUTION OF CONVERTED AND NON-CONVERTED USERS. THIS INDICATES POTENTIAL STABILITY AND CONSISTENCY IN CONVERSION RATES FOR THESE CATEGORIES.

EDA

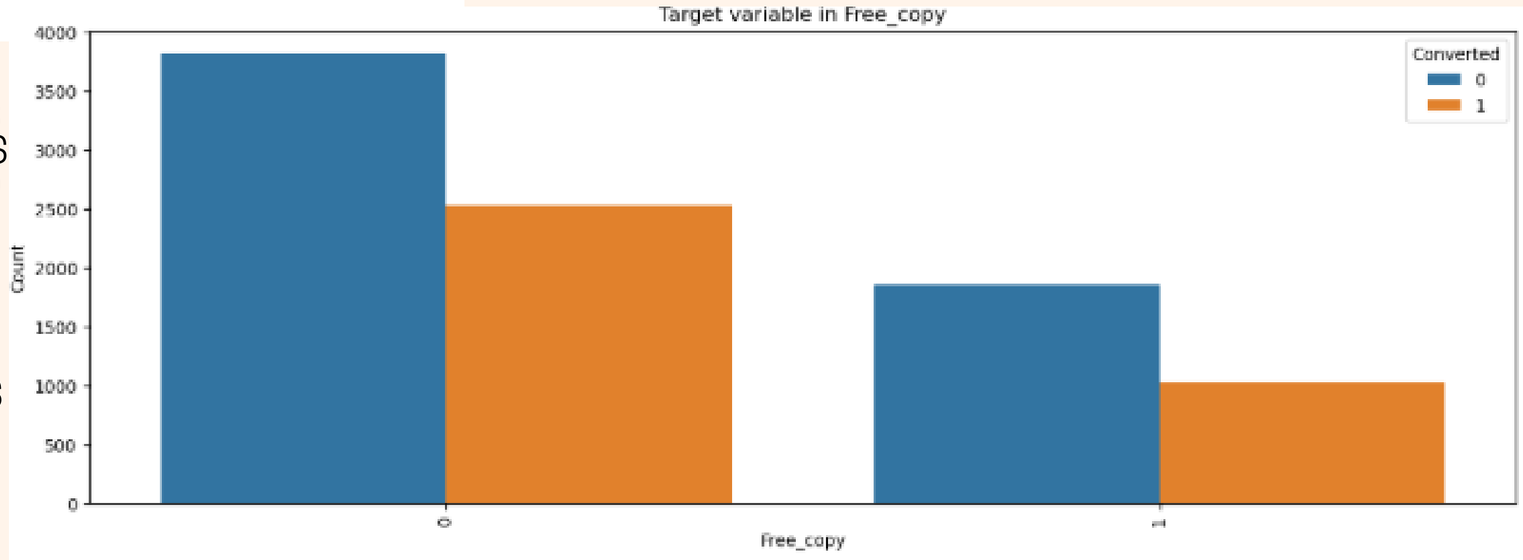


DOMINANCE OF UNEMPLOYED INDIVIDUALS: THE "UNEMPLOYED" CATEGORY HAS THE HIGHEST COUNT, WITH MOST INDIVIDUALS NOT CONVERTING (BLUE BAR). THIS HIGHLIGHTS THAT UNEMPLOYED INDIVIDUALS FORM THE MAJORITY OF THIS DATASET BUT ARE LESS LIKELY TO CONVERT.

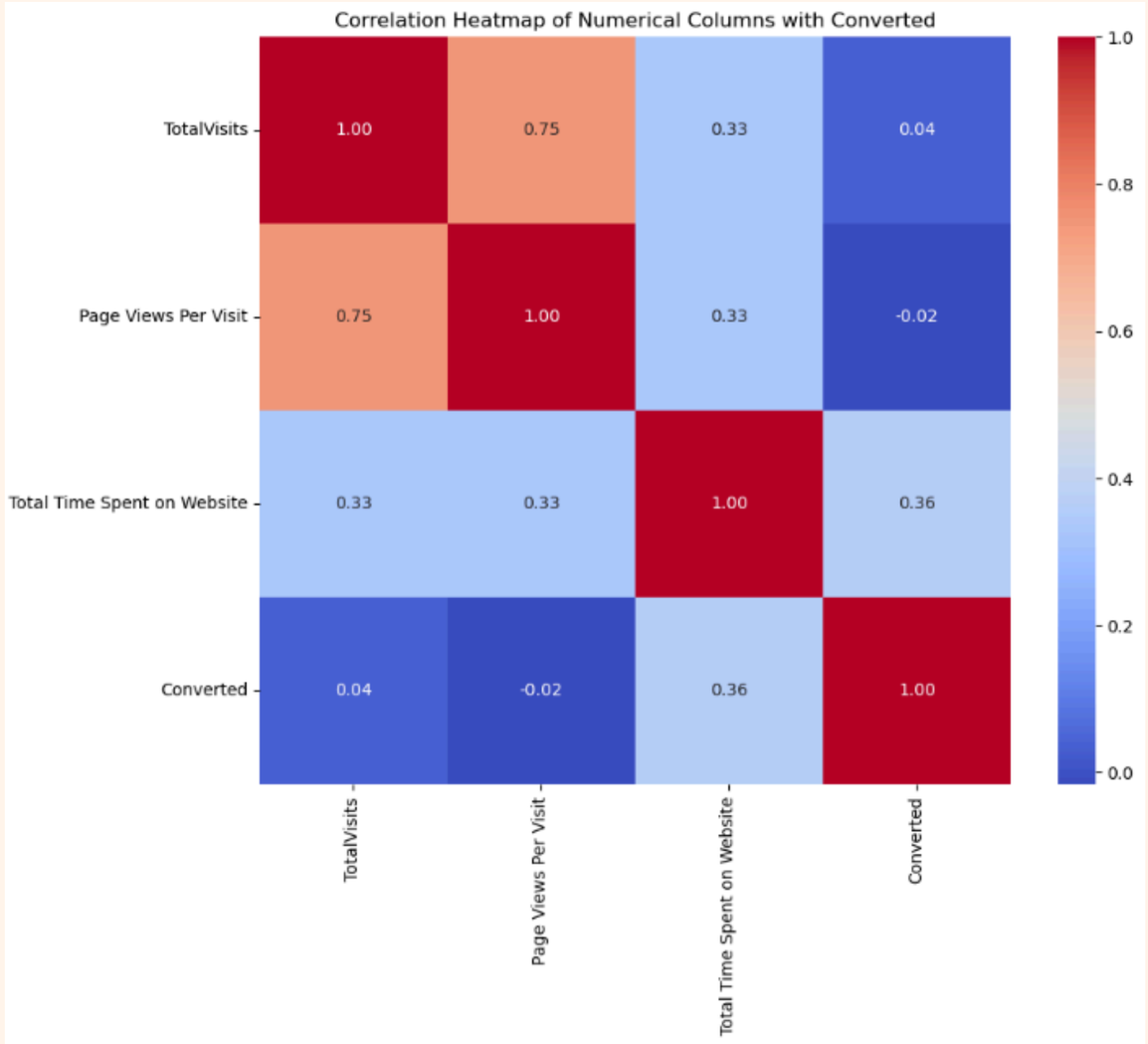
HIGHER CONVERSION AMONG WORKING PROFESSIONALS: AMONG THE "WORKING PROFESSIONAL" GROUP, THERE IS A NOTICEABLY HIGHER NUMBER OF CONVERSIONS (ORANGE BAR), SUGGESTING THAT THIS CATEGORY HAS BETTER ENGAGEMENT OR INTEREST COMPARED TO OTHERS LIKE "STUDENT" OR "HOUSEWIFE," WHERE CONVERSIONS ARE MINIMAL.

IMPACT OF FREE COPY ON CONVERSION: WHEN NO FREE COPY IS OFFERED (FREE_COPY = 0), THE COUNT OF NON-CONVERTED USERS (BLUE) IS SIGNIFICANTLY HIGHER THAN THE COUNT OF CONVERTED USERS (ORANGE). THIS INDICATES THAT THE ABSENCE OF A FREE COPY CORRELATES WITH LOWER CONVERSIONS.

INCREASED CONVERSION LIKELIHOOD WITH FREE COPY: WHEN A FREE COPY IS PROVIDED (FREE_COPY = 1), THERE IS A NOTICEABLE INCREASE IN THE PROPORTION OF CONVERTED USERS (ORANGE), EVEN THOUGH NON-CONVERSIONS STILL DOMINATE. THIS SUGGESTS THAT OFFERING A FREE COPY POSITIVELY INFLUENCES THE LIKELIHOOD OF CONVERSION.



EDA

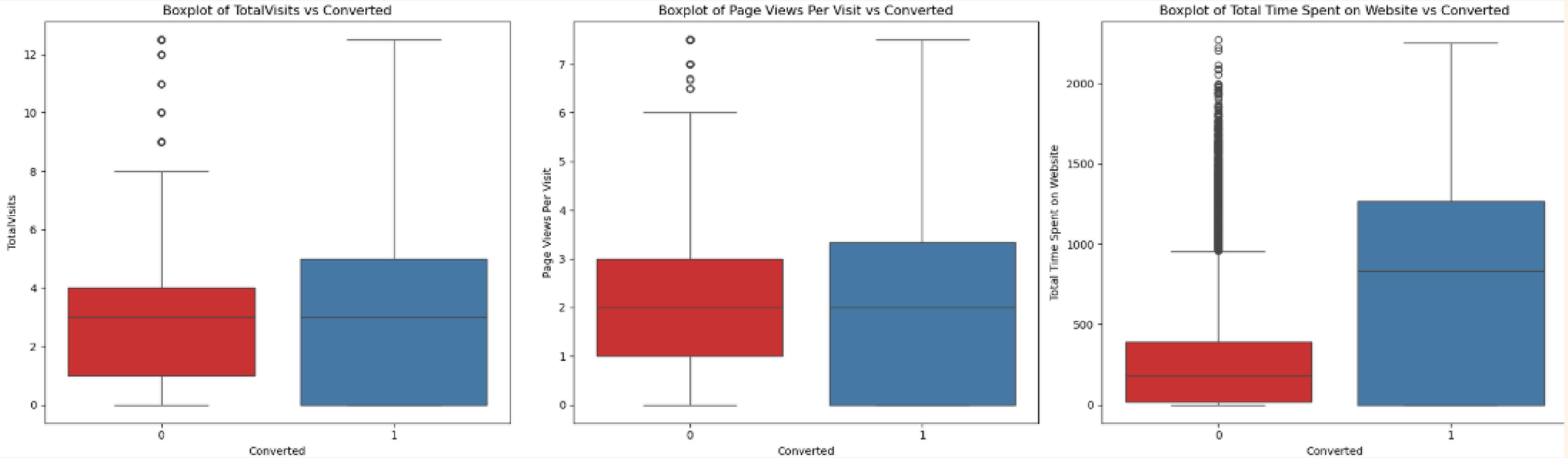


TOTAL TIME SPENT ON WEBSITE AND CONVERSION:
THERE IS A MODERATE POSITIVE CORRELATION (APPROXIMATELY 0.36) BETWEEN "TOTAL TIME SPENT ON WEBSITE" AND "CONVERTED." THIS SUGGESTS THAT USERS WHO SPEND MORE TIME ON THE WEBSITE HAVE A HIGHER LIKELIHOOD OF CONVERTING.

TOTALVISITS AND PAGE VIEWS PER VISIT:
A STRONG POSITIVE CORRELATION (AROUND 0.75) EXISTS BETWEEN "TOTALVISITS" AND "PAGE VIEWS PER VISIT." THIS INDICATES THAT USERS WHO MAKE MORE VISITS ALSO TEND TO VIEW MORE PAGES PER VISIT, REFLECTING A PATTERN OF INCREASED ENGAGEMENT.

PAGE VIEWS PER VISIT AND CONVERSION:
THE CORRELATION BETWEEN "PAGE VIEWS PER VISIT" AND "CONVERTED" IS VERY WEAK AND SLIGHTLY NEGATIVE (APPROXIMATELY -0.02). THIS IMPLIES THAT THE NUMBER OF PAGES VIEWED PER VISIT HAS LITTLE TO NO SIGNIFICANT EFFECT ON WHETHER A USER CONVERTS.

EDA



CONVERTED USERS TEND TO HAVE A HIGHER MEDIAN NUMBER OF TOTAL VISITS COMPARED TO NON-CONVERTED USERS.

NON-CONVERTED USERS EXHIBIT A SIGNIFICANT NUMBER OF OUTLIERS, INDICATING SOME USERS VISIT MANY TIMES BUT FAIL TO CONVERT.

CONVERTED USERS HAVE A SLIGHTLY HIGHER MEDIAN OF PAGE VIEWS PER VISIT THAN NON-CONVERTED USERS.

NON-CONVERTED USERS SHOW SEVERAL OUTLIERS, SUGGESTING SOME USERS VIEW A LARGE NUMBER OF PAGES BUT DO NOT CONVERT.

CONVERTED USERS CLEARLY SPEND MORE MEDIAN TIME ON THE WEBSITE COMPARED TO NON-CONVERTED USERS.

NON-CONVERTED USERS DISPLAY MANY OUTLIERS, SHOWING THAT SOME SPEND CONSIDERABLE TIME ON THE SITE WITHOUT CONVERTING.

DATA CONVERSION

Numerical Variables are Normalised

Dummy Variables are created for object type variables

Total Rows for Analysis: 9240

Total Columns for Analysis: 49

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

Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5

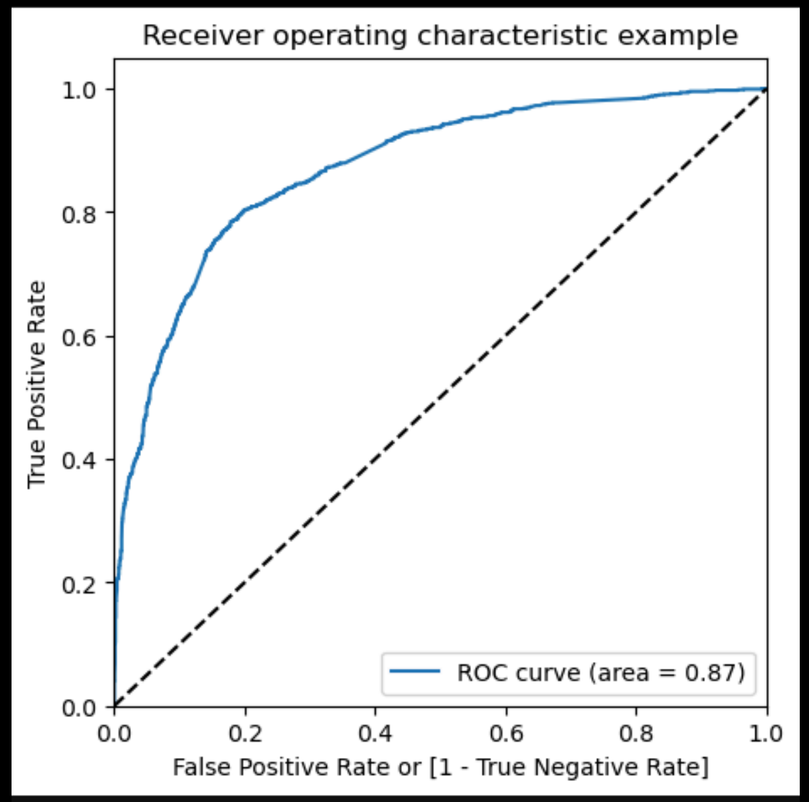
Predictions on test data set
&
Overall accuracy 80.47%

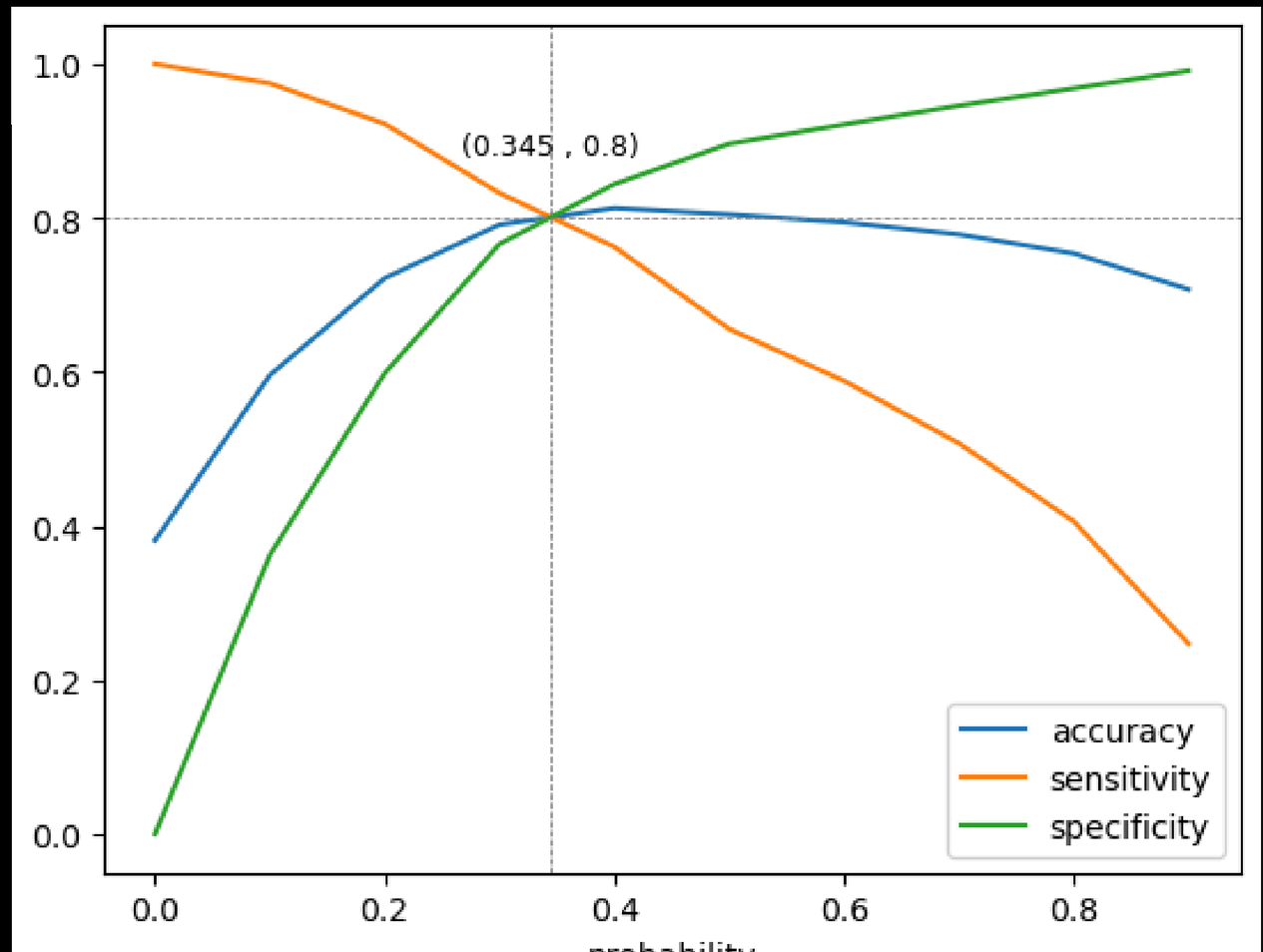
ROC CURVE

Strong Model Performance (ROC Curve): The ROC curve demonstrates an AUC (Area Under the Curve) of 0.87, indicating that the model has excellent discriminative power, effectively distinguishing between positive and negative classes.

Optimal Threshold Identified: In the second graph, the optimal threshold probability is around 0.345, where accuracy, sensitivity, and specificity intersect at approximately 0.8. This balance is crucial for achieving an effective trade-off between identifying true positives and minimizing false positives.

Sensitivity-Specificity Trade-off: The model achieves a high sensitivity and specificity of approximately 0.8 at the optimal threshold, showcasing its capability to minimize both false negatives and false positives effectively. This balance is vital for reliable decision-making





CONCLUSION

Train Data Set

- Accuracy: 80.47%
- Sensitivity: 80.13%
- Specificity: 80.68%

Test Data Set

- Accuracy: 80.16%
- Sensitivity: 79.82% ≈ 80%
- Specificity: 80.38%

Notes

- The model achieved a sensitivity of 80.13% in the train set and 79.82% in the test set, using a cut-off value of 0.345.
- Sensitivity in this case indicates how many leads the model identify correctly out of all potential leads which are converting
- The CEO of X Education had set a target sensitivity of around 80%.
- The model also achieved an accuracy of 80.47%, which is in line with the study's objectives.

RECOMMENDATIONS

To increase our Lead Conversion Rates:

- Focus on features with positive coefficients for targeted marketing strategies.
- Develop strategies to attract high-quality leads from top-performing lead sources.
- Engage working professionals with tailored messaging.
- Optimize communication channels based on lead engagement impact.
- More budget/spend can be done on Welingak Website in terms of advertising, etc.
- Incentives/discounts for providing reference that convert to lead, encourage providing more references.
- Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.

To identify areas of improvement:

- Analyze negative coefficients in specialization offerings.
- Review landing page submission process for areas of improvement.


THANK
YOU