LEAD SCORING CASE STUDY

BY

Shivakumar Dwarapogu

&

Vaishali vishwakarma

BUSINESS PROBLEM STATEMENT:

- X Education, an online course provider, seeks to boost its lead conversion rate, which currently stands at 30%.
- The company attracts professionals through various channels, including website visits, form submissions, and referrals.
- To optimize this process, they intend to pinpoint high-potential leads, known as 'Hot Leads,' to enhance their lead conversion rate.
- Aiming for an 80% conversion rate, the company aims to build a lead scoring model that assigns scores reflecting the likelihood of conversion, allowing the sales team to focus on prospects with the highest conversion potential.

GOAL OF THE CASE STUDY

- 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.

SUMMARY SOLUTION

- Reading and understanding the data
- Cleaning the data and handling the data.
- Performing exploratory data analysis and visualization
- Data preparation for model building and model evaluation
- Final predictions
- Conclusion and improvement areas

IMPORTING LIBRARIES, READING AND UNDERSTANDING THE DATA

- ✓The dataset's structure reveals a total of 9240 rows and 37 columns.
- ✓The dataset prominently features several categorical variables, necessitating the creation of dummy variables. Furthermore, a noticeable presence of null values calls for appropriate treatment strategies.
- ✓Several columns within the data frame exhibit a notable count of missing or null values. Decisions regarding their handling will be made during the data cleaning and imputation phase.

DATA CLEANING AND HANDLING

- As indicated in the Problem Statement, several categorical variables include a category labeled 'Select' which essentially denotes a null value. This scenario often arises when customers haven't made a selection from the available options. Consequently, these columns retain the default 'Select' value to represent this absence of choice
- After evaluating the null values, it was determined that columns with over 40% null values should be dropped.
- Following the column drop, measures were taken to manage null values. Non-contributory columns were eliminated, while imputation was carried out where necessary to enhance the module's efficacy.
- The columns that follow display pronounced skewness: 'Do Not Call', 'Search', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', and 'Through Recommendations'. Considering their minimal impact on model improvement and the potential to distort logistic regression models, these columns will be removed. Skewed variables can introduce bias or inaccuracy into parameter estimates, affecting the model's reliability.
- Moving on to outlier treatment, both "TotalVisits" and "Page Views Per Visit" display outliers. These outliers have been addressed by applying capping and flooring techniques.
- Essential actions involve addressing invalid values and standardizing data across columns. Invalid values were corrected, and less common value levels were aggregated under the category "Others".

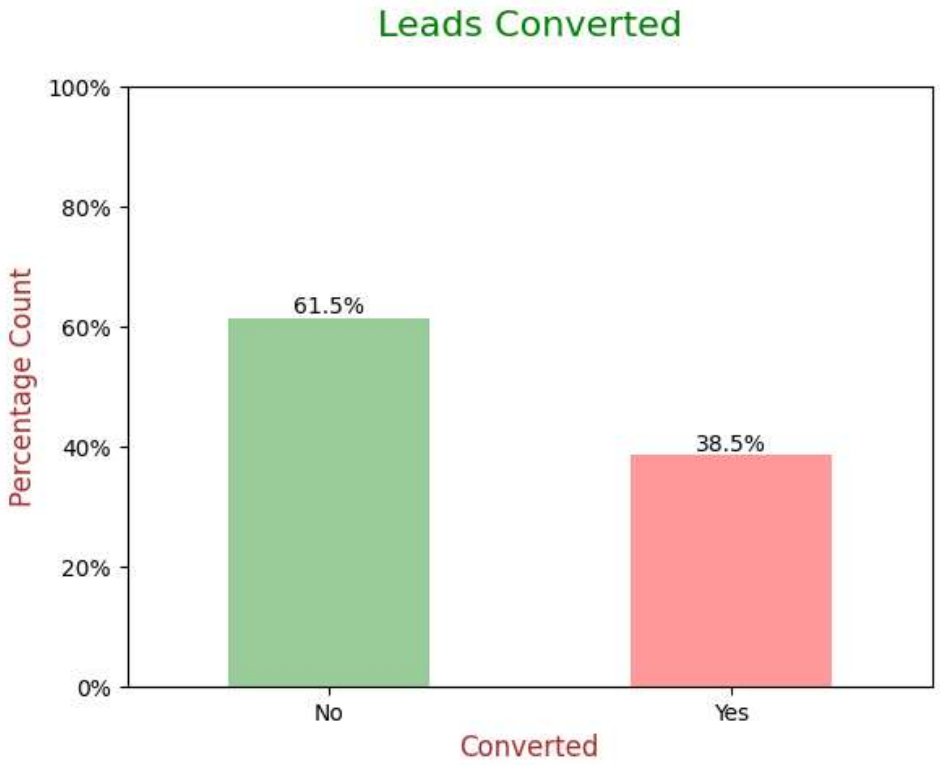
EXPLORATORY DATA ANALYSIS AND VISUALIZATION

Data Imbalance

Data imbalance occurs when one value, often the majority, is substantially more prevalent than another value, representing the minority. This leads to an uneven distribution of observations within the dataset.

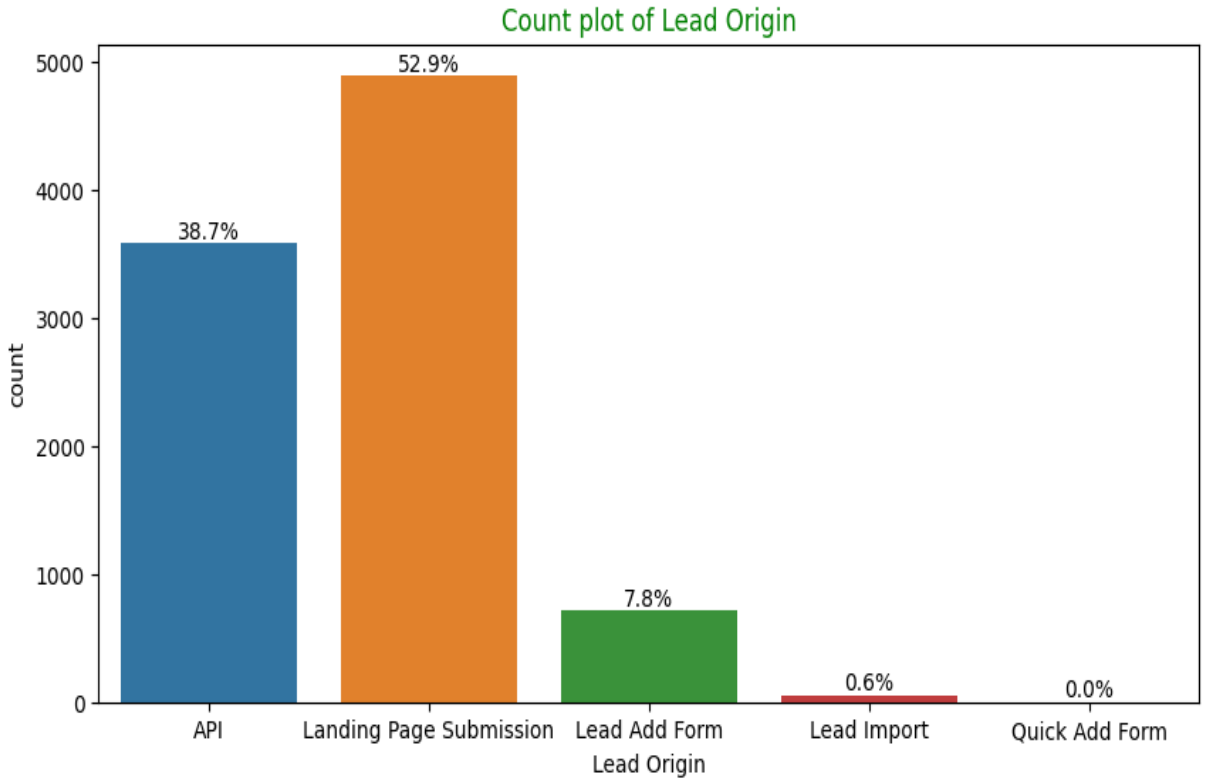
Observations

The conversion rate is 38.5%, indicating that a minority, specifically 38.5% of individuals, have transformed into leads. In contrast, the majority, approximately 61.5% of individuals, did not convert into leads.

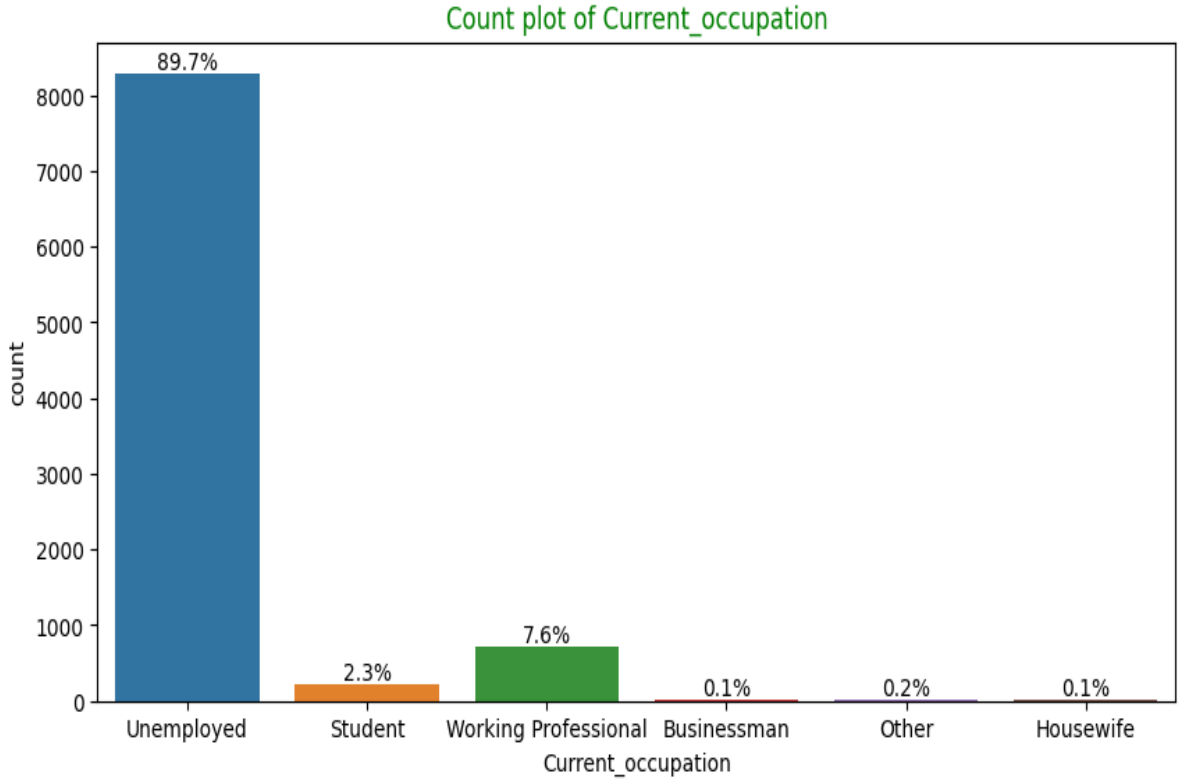


UNIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES

Lead Origin: "Landing Page Submission" accounted for 53% of the customers, while "API" represented 39%.

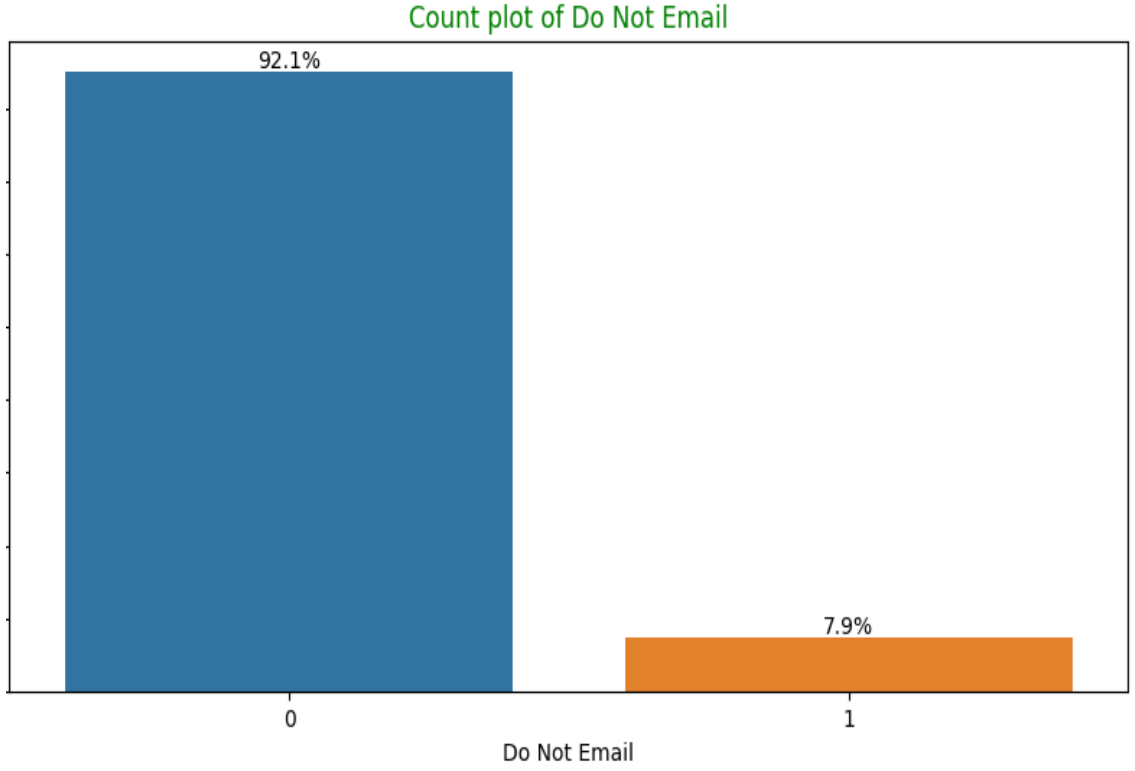


Current Occupation: The majority, or 90%, of customers were categorized as "Unemployed."

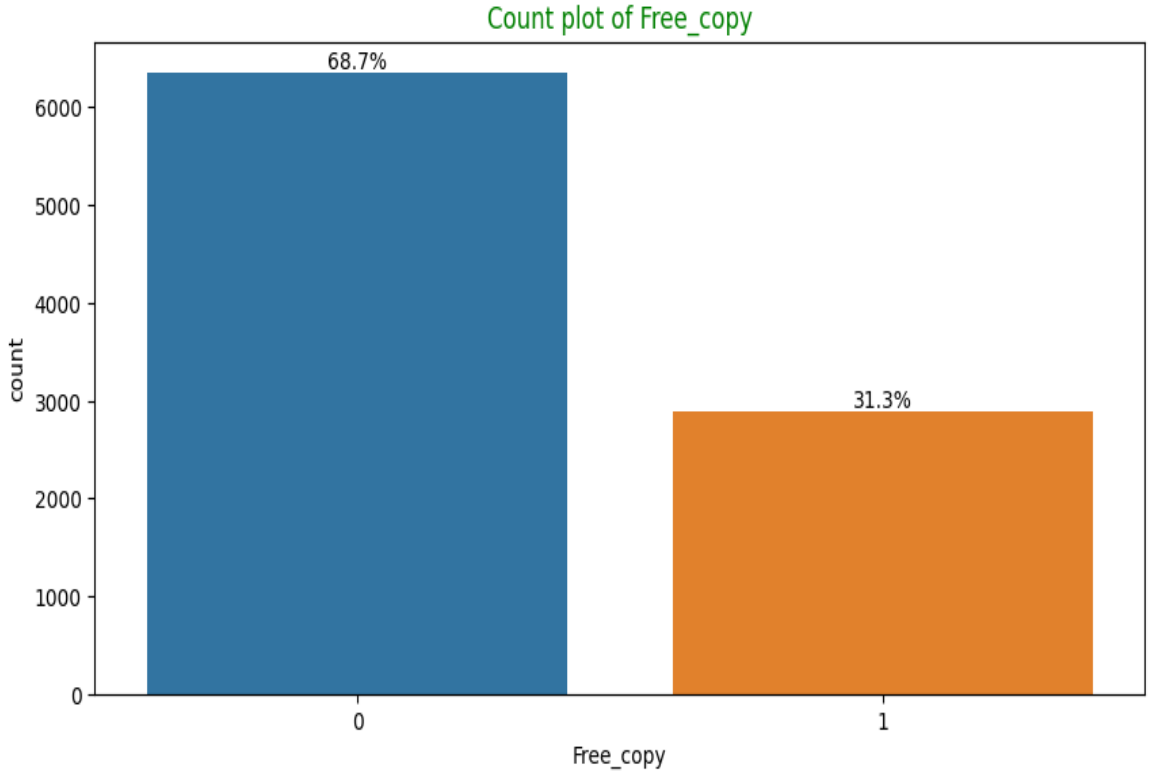


UNIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES

Do Not Email: Approximately 92% of individuals indicated their preference to not receive emails about the course.



Free copy: Only 31.3% of individuals have chosen to receive a free copy of the "Mastering the Interview" resource.

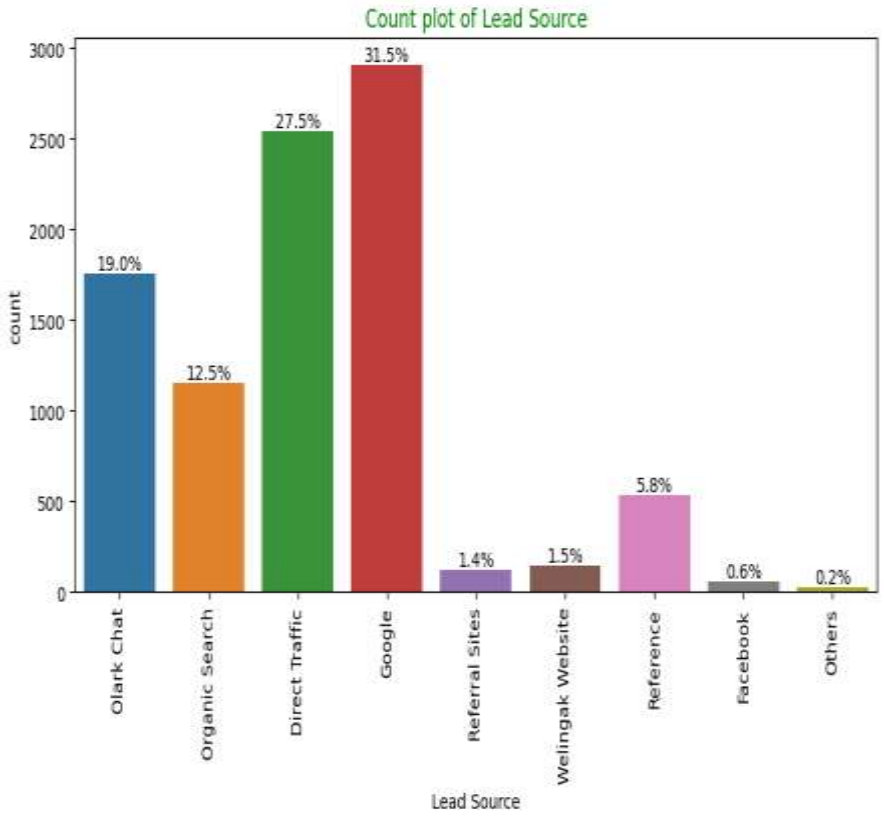


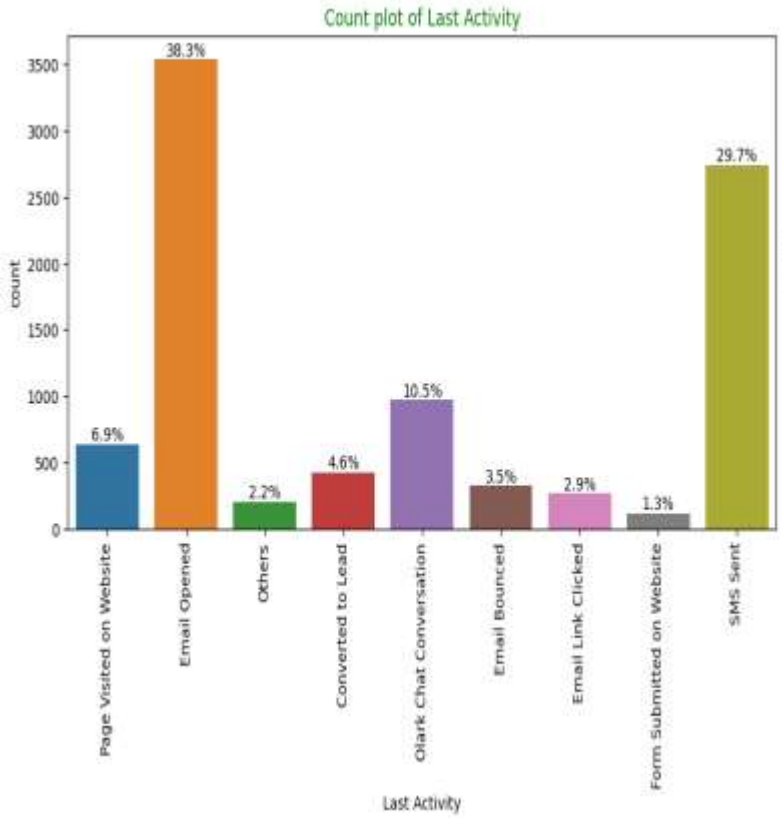
UNIVARIATE ANALYSIS USING COUNT PLOT

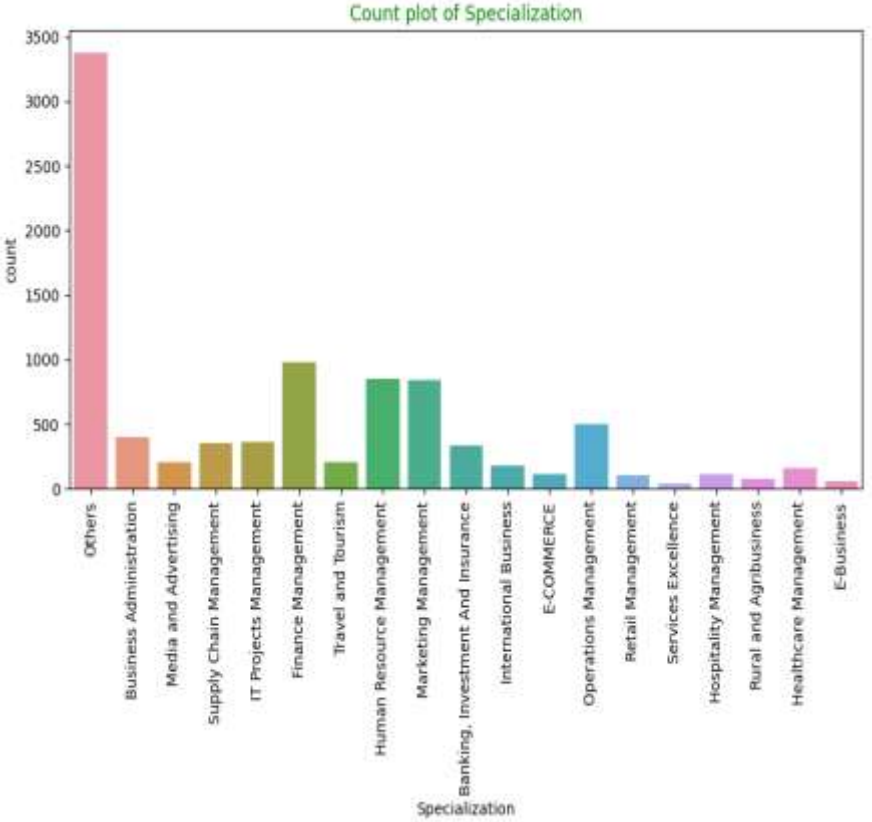
Lead Source: The combined percentage for "Google" and "Direct Traffic" as the lead sources was 58%.

Last Activity: 68% of customers were involved in activities related to "SMS Sent" and "Email Opened."

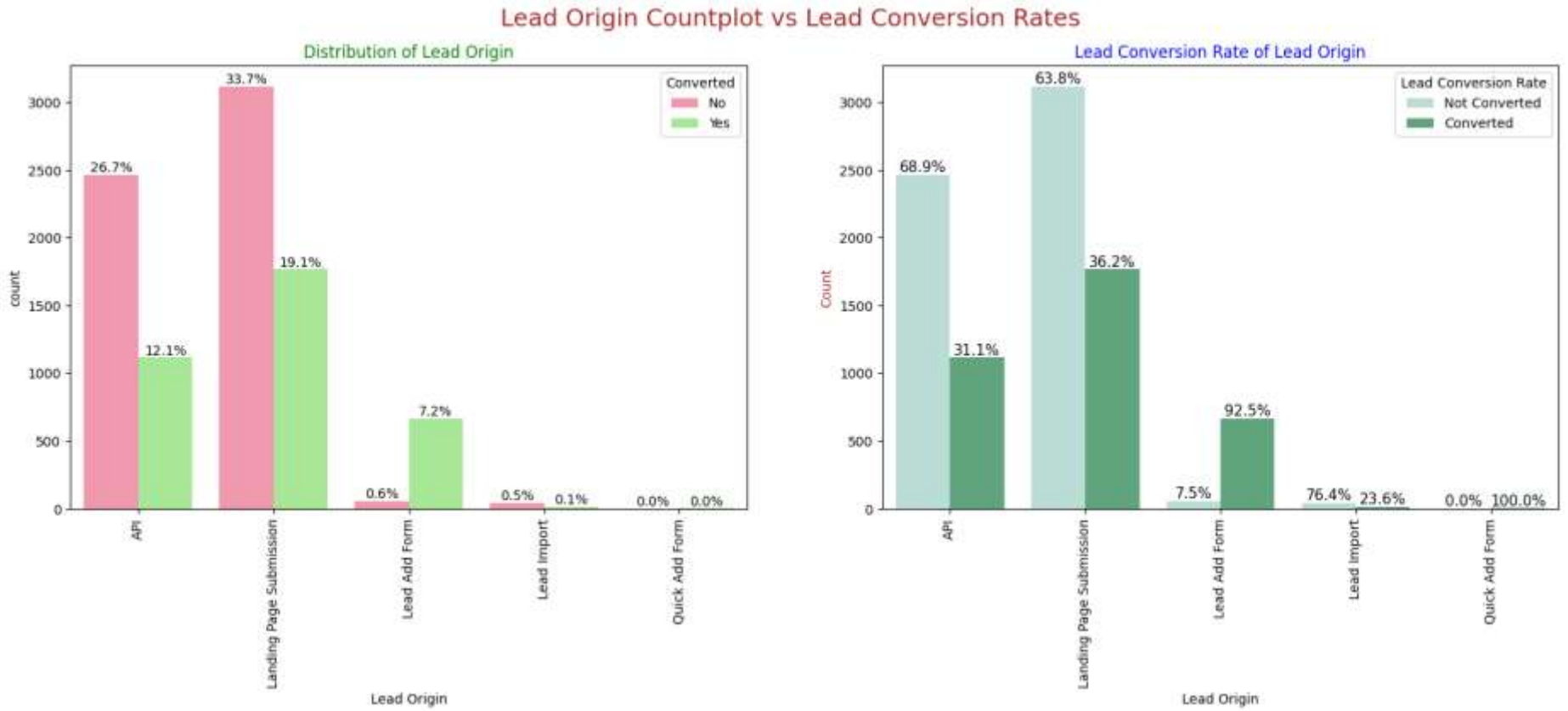
Specialization: A substantial number of individuals have not opted for any specialization.







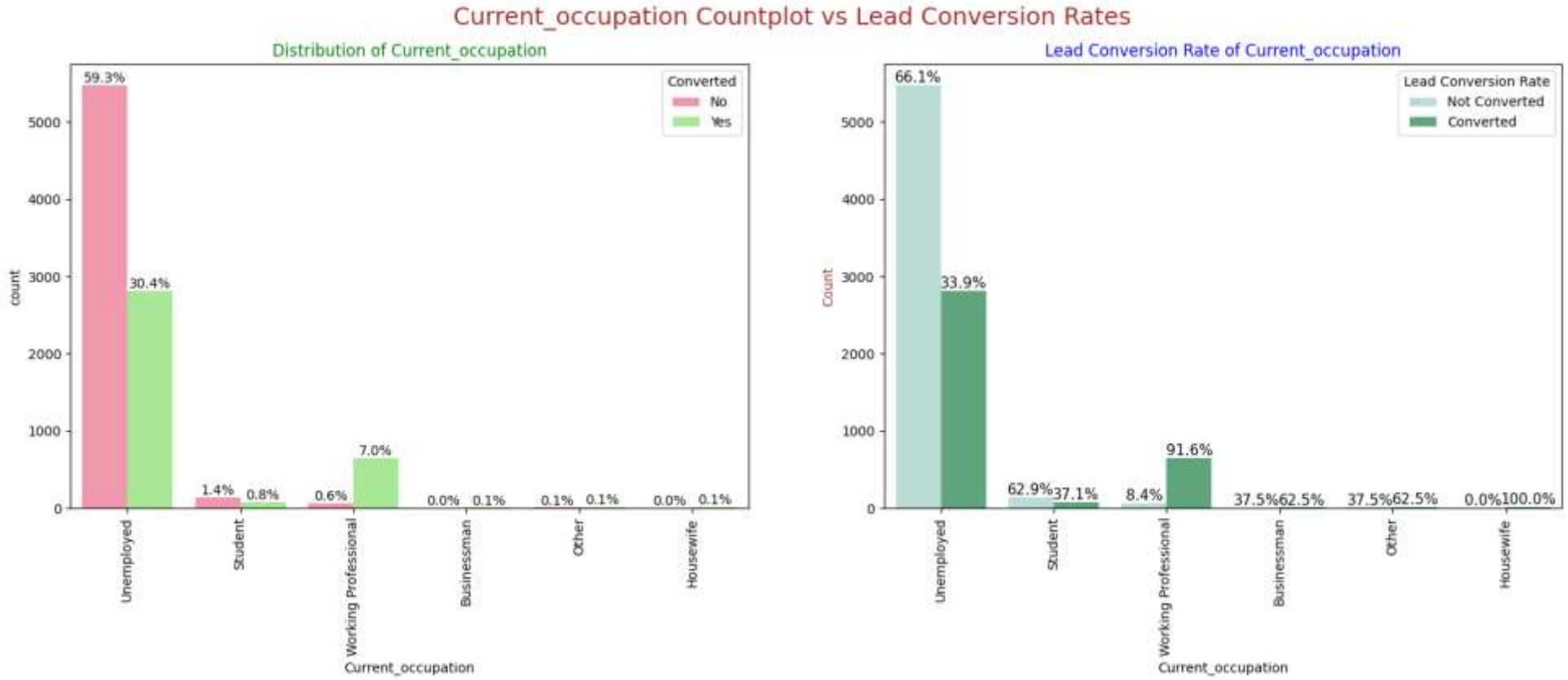
BIVARIATE ANALYSIS ON LEAD SCORE AND CONVERSION RATES



Roughly 52% of all leads originated from "Landing Page Submission," exhibiting a lead conversion rate (LCR) of 36%.

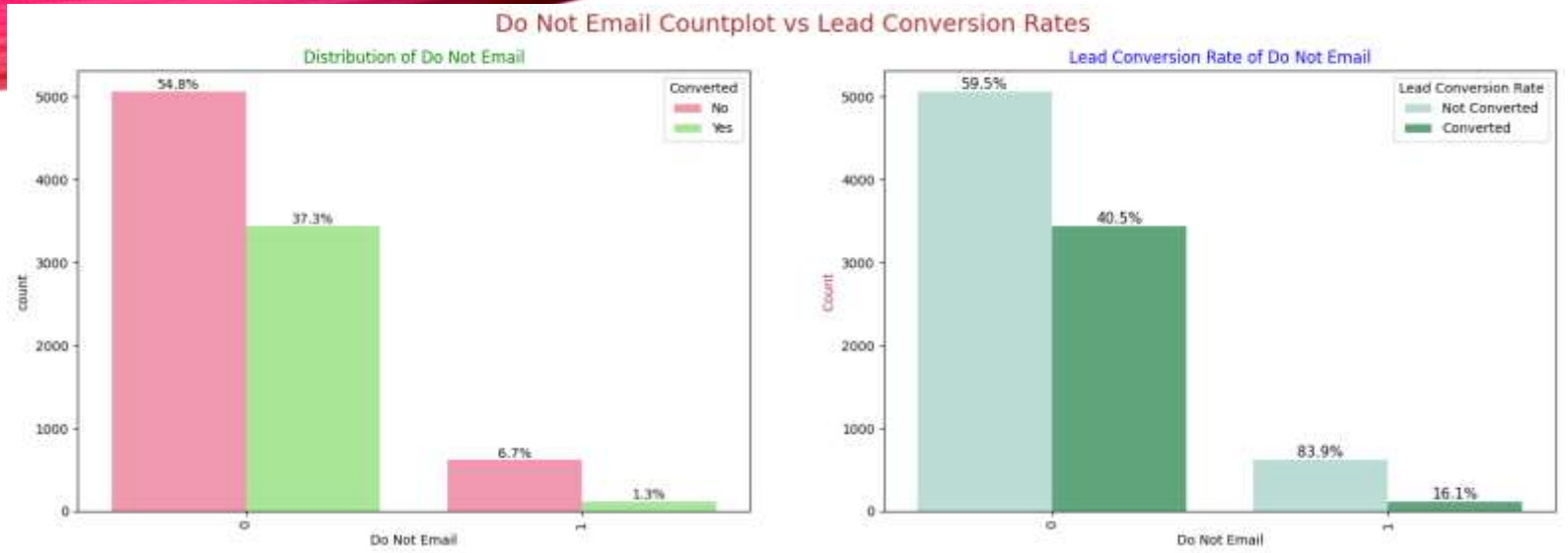
Meanwhile, the "API" accounted for approximately 39% of customers, demonstrating a lead conversion rate (LCR) of 31%.

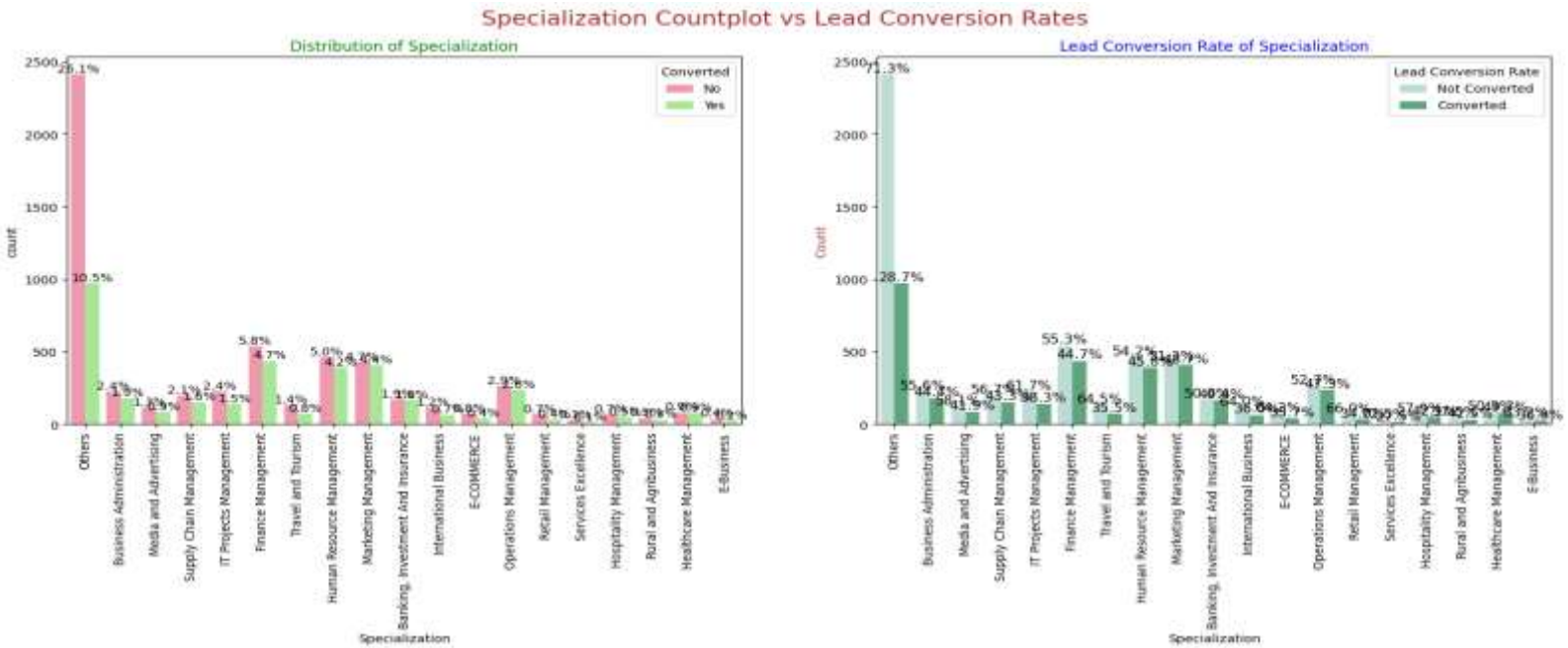
BIVARIATE ANALYSIS ON CURRENT OCCUPATION AND LEAD CONVERSION RATES



About 90% of customers fall into the "Unemployed" category, boasting a lead conversion rate (LCR) of 34%. In contrast, "Working Professionals" make up a mere 7.6% of the total customer base, yet exhibit an impressive lead conversion rate (LCR) of nearly 92%.

BIVARIATE ANALYSIS





Do Not Email VS Lead conversion rates

- An overwhelming 92% of individuals have chosen to opt out of receiving course-related emails.

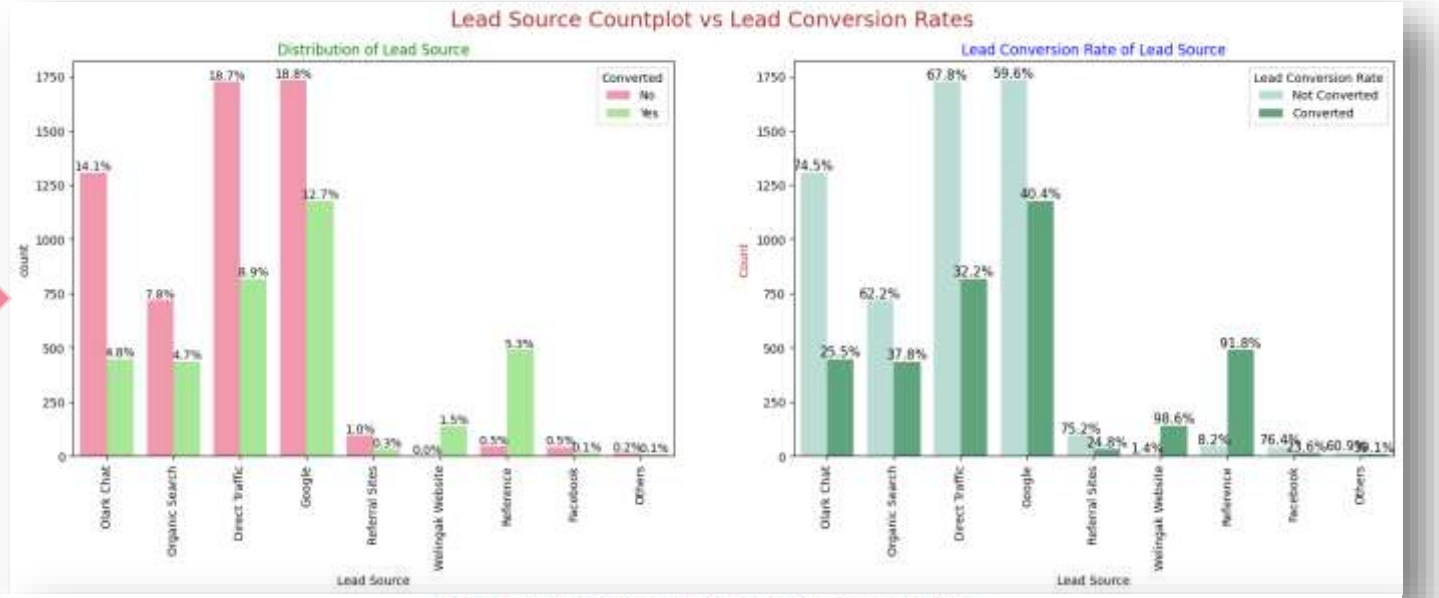
Specialization VS Lead conversion rates

- Disciplines such as Marketing Management, HR Management, and Finance Management exhibit substantial contributions.

BIVARIATE ANALYSIS

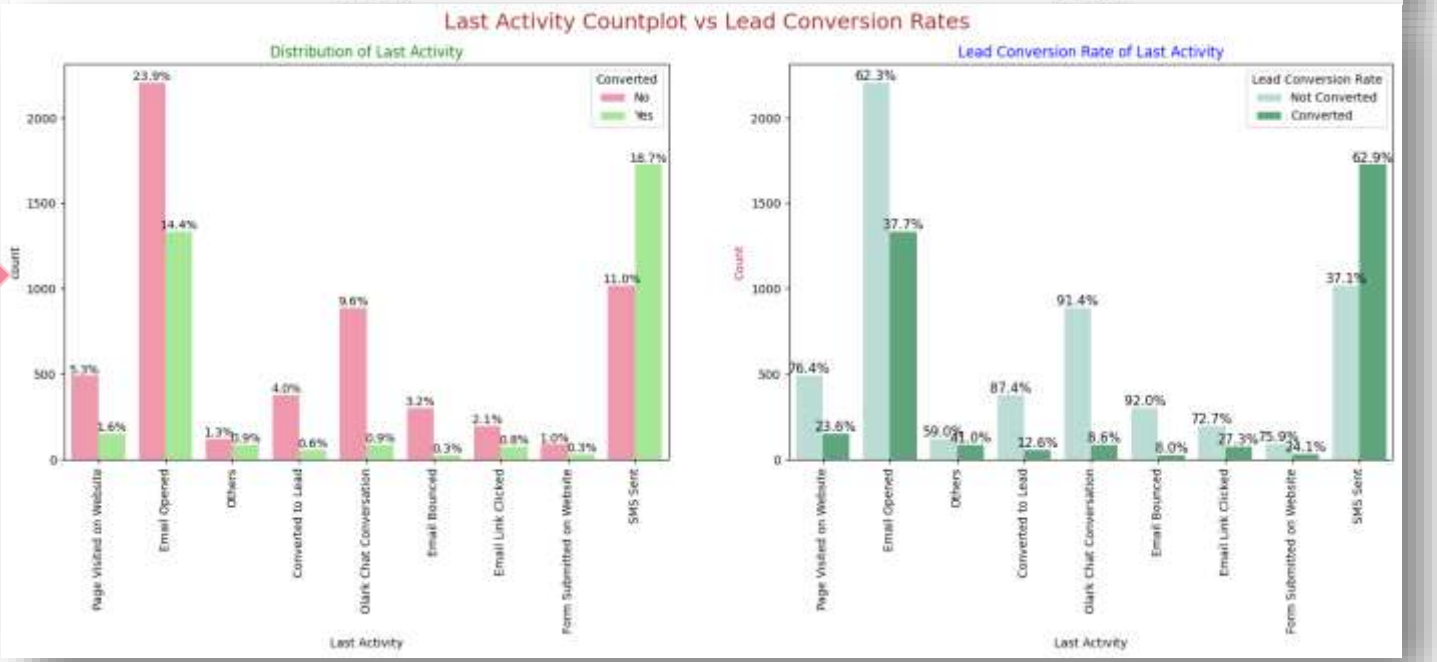
Lead source Vs Conversion rates

- Among the customers, Google boasts an impressive lead conversion rate (LCR) of 40%, originating from 31% of the customer pool. Direct Traffic contributes a LCR of 32%, with 27% of customers associated with this source, albeit lower than Google.
- Organic Search yields a noteworthy LCR of 37.8%, but its contribution comes from just 12.5% of customers. Reference showcases a remarkable LCR of 91%, yet only around 6% of customers are attributed to this lead source.

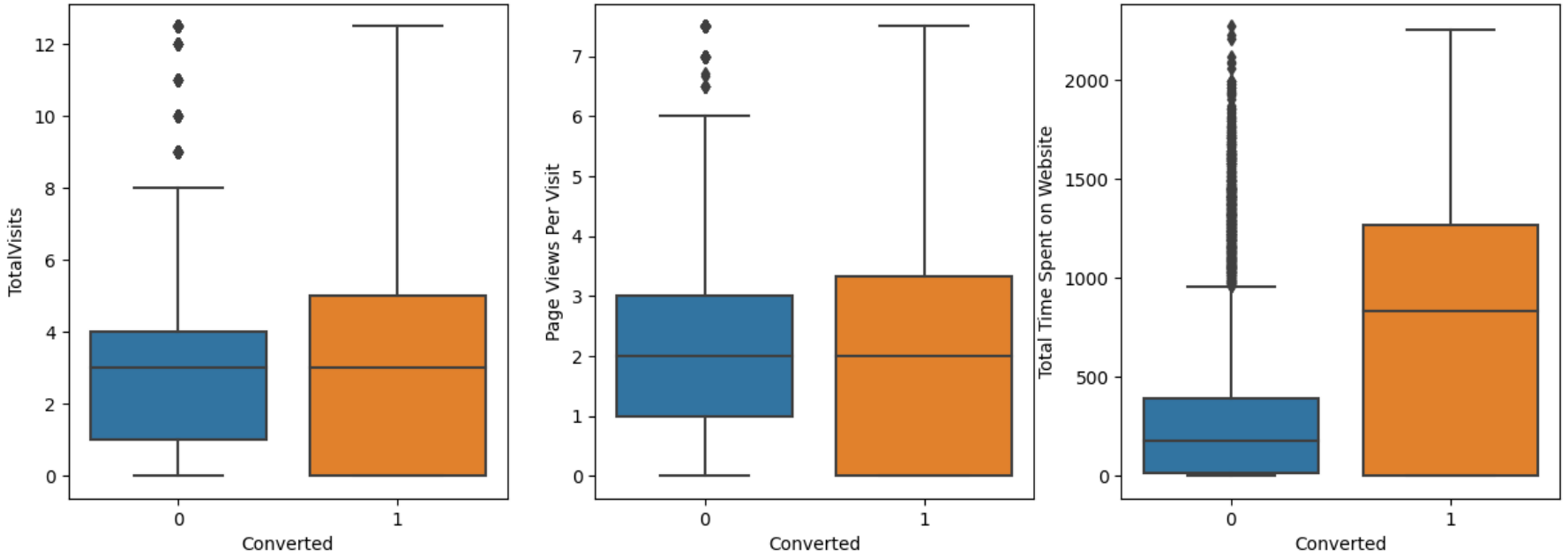


Last Activity Vs Conversion rates

- The activity 'SMS Sent' demonstrates a notably high lead conversion rate of 63%, stemming from a 30% contribution in terms of last activities.
- 'Email Opened' activities account for 38% of the last activities conducted by customers, yielding a 37% lead conversion rate.



BIVARIATE ANALYSIS FOR NUMERICAL VARIABLES



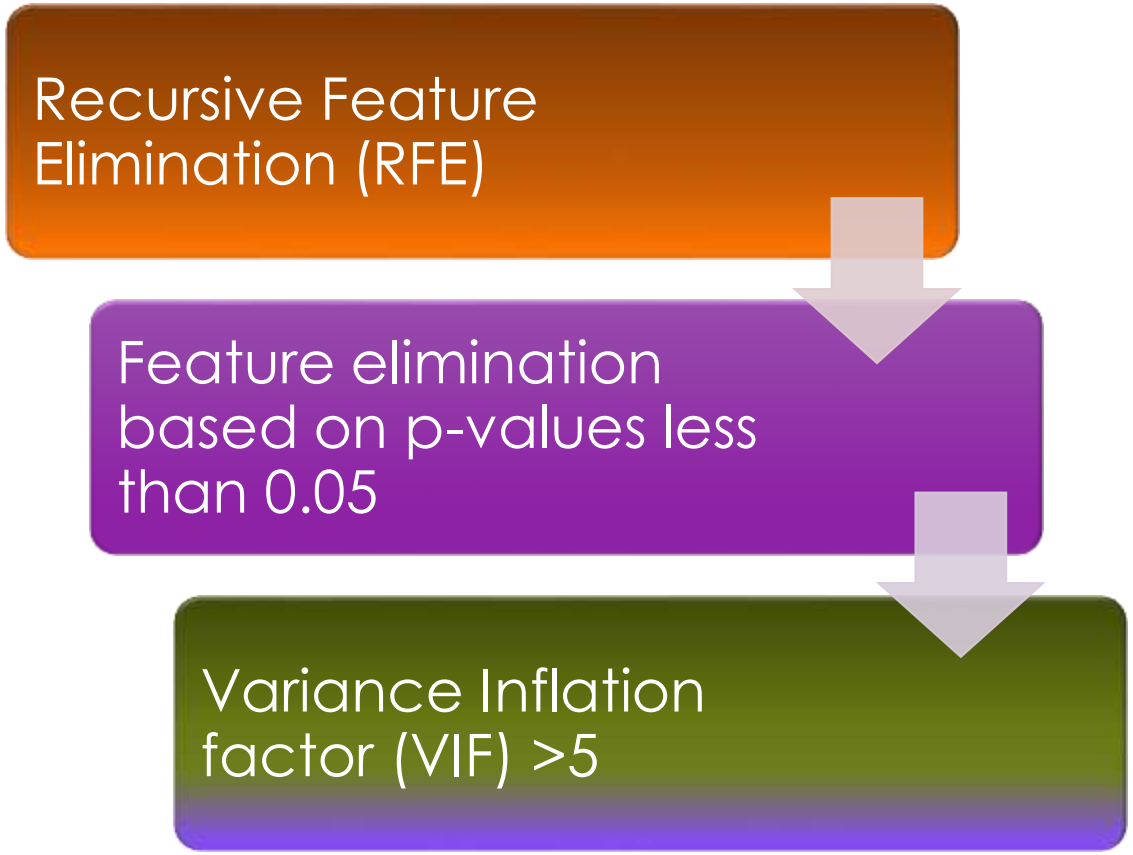
The boxplot demonstrates that past leads who spend more time on the website are more successfully converted compared to those who spend less time.

DATA PREPARATION FOR MODEL BUILDING

- For categorical variables with multiple levels, dummy features were generated through one-hot encoding. As binary-level categorical columns were already mapped to 1s and 0s in prior steps, the process involved creating dummy variables and subsequently dropping the original variables.
- Subsequent to the creation of dummy variables, the data was split into test and train datasets, adhering to a 70:30 ratio.
- After the split, x and y variables were generated to facilitate the construction of the model.
- Feature data underwent standard scaling using the Standard Scaler.

MODEL BUILDING

- The model construction initiated by utilizing the Recursive Feature Elimination (RFE) method for feature selection. In this process, 15 features were selected using the parameter n_features_to_select.
- An iterative examination of the P-values was conducted to progressively eliminate features, culminating in the development of the final model.
- In Model 4, notable p-values were observed within the designated threshold (p_value > 0.05), and the Variance Inflation Factor (VIF) values remained favorable, all falling below 5.



FINAL MODEL

P values of Final Model4

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6455
Model Family:	Binomial	Df Model:	12
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2743.1
Date:	Mon, 14 Aug 2023	Deviance:	5486.1
Time:	03:34:26	Pearson chi2:	8.11e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3819
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0236	0.143	-7.145	0.000	-1.304	-0.743
Total Time Spent on Website	1.0498	0.039	27.234	0.000	0.974	1.125
Lead Origin_Landing Page Submission	-1.2590	0.125	-10.037	0.000	-1.505	-1.013
Lead Source_Olark Chat	0.9072	0.118	7.701	0.000	0.676	1.138
Lead Source_Reference	2.9253	0.215	13.615	0.000	2.504	3.346
Lead Source_Welingak Website	5.3887	0.728	7.399	0.000	3.961	6.816
Last Activity_Email Opened	0.9421	0.104	9.022	0.000	0.737	1.147
Last Activity_Olark Chat Conversation	-0.5556	0.187	-2.974	0.003	-0.922	-0.189
Last Activity_Others	1.2531	0.238	5.259	0.000	0.786	1.720
Last Activity_SMS Sent	2.0519	0.107	19.106	0.000	1.841	2.262
Specialization_Hospitality Management	-1.0944	0.323	-3.391	0.001	-1.727	-0.462
Specialization_Others	-1.2033	0.121	-9.950	0.000	-1.440	-0.966
Current_occupation_Working Professional	2.6697	0.190	14.034	0.000	2.297	3.042

VIFs of all variables in the Final Model 4

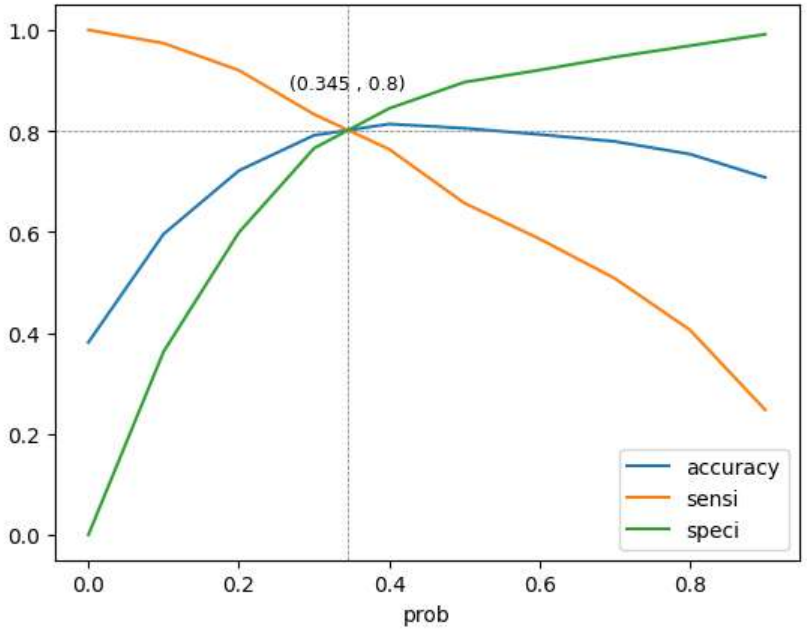
	Features	VIF
0	Specialization_Others	2.47
1	Lead Origin_Landing Page Submission	2.45
2	Last Activity_Email Opened	2.36
3	Last Activity_SMS Sent	2.20
4	Lead Source_Olark Chat	2.14
5	Last Activity_Olark Chat Conversation	1.72
6	Lead Source_Reference	1.31
7	Total Time Spent on Website	1.24
8	Current_occupation_Working Professional	1.21
9	Lead Source_Welingak Website	1.08
10	Last Activity_Others	1.08
11	Specialization_Hospitality Management	1.02

MODEL EVALUATION

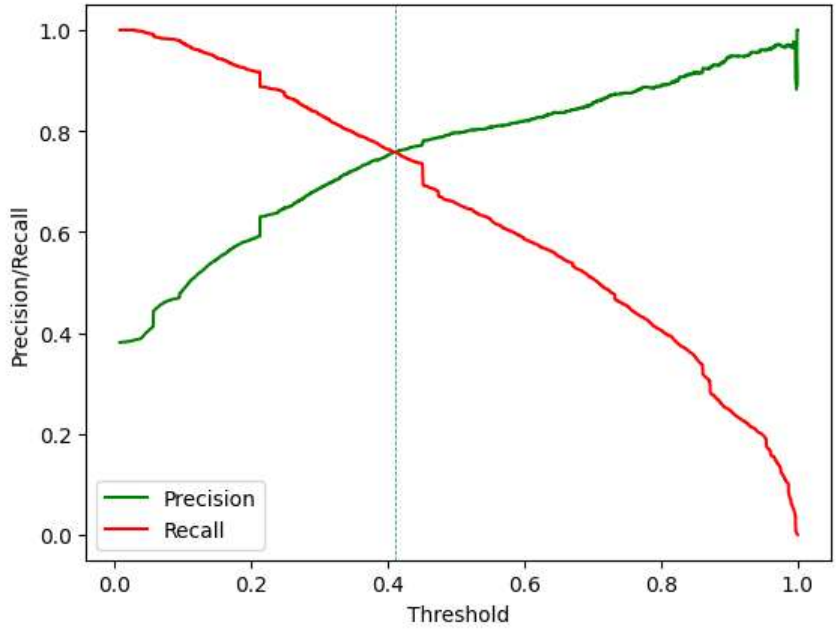
- The predicted values of the final model were implemented on the training set.
- With the sensitivity-specificity threshold set at 0.345, the model accomplished the aimed 80% target for the True Positive Rate, aptly aligning with the predefined business objectives.
- After careful consideration, the ultimate decision was to adopt the sensitivity-specificity approach as the optimal cutoff point, ensuring precise predictions.

MODEL EVALUATION ON TRAIN SET

Accuracy, Sensitivity, Specificity intersect at 0.345



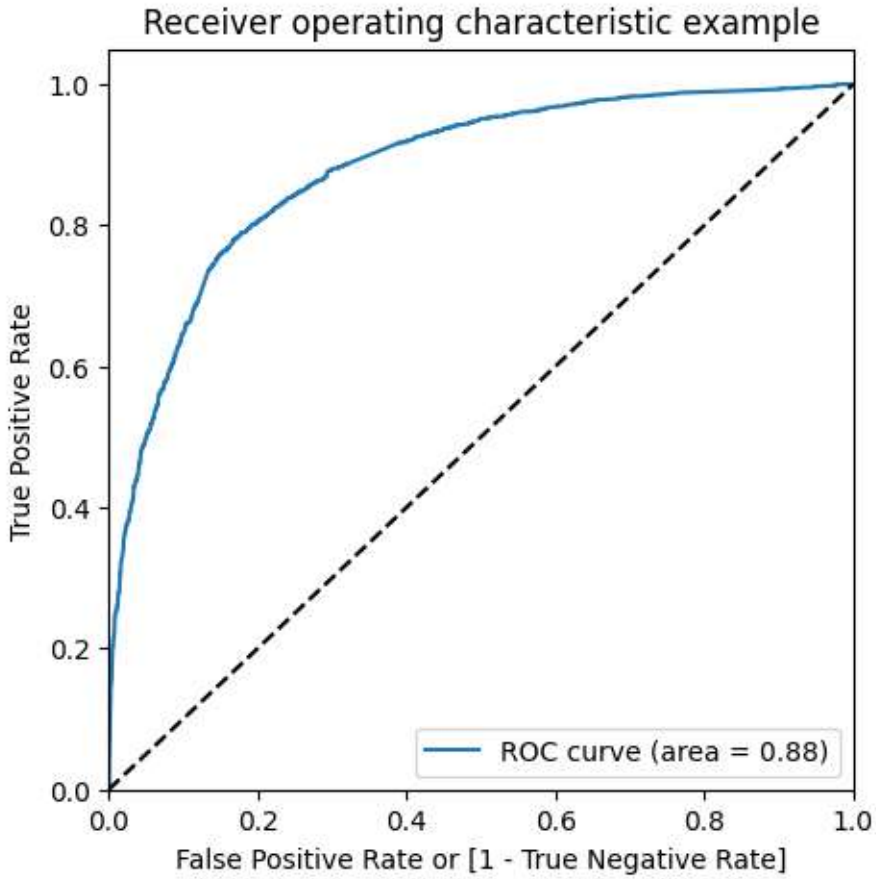
Precision and recall probability threshold is approx. 0.41



PREDICTIONS

- ✓ The model demonstrated robust predictive capability, evident from its ROC curve area of 0.87 out of 1.
- ✓ Utilizing the converted probabilities, the final predicted values were derived, poised for utilization in lead conversion strategies.
- ✓ The inclusion of converted probabilities yielded the final predicted values, now prepared for utilization in lead conversion endeavors.

ROC Curve Plot



CONCLUSION

- Consistent Performance: The evaluation metrics show remarkable consistency between the test and train datasets, indicating the model's steady performance across diverse evaluation criteria.
- Target Sensitivity Attained: The model successfully reached a sensitivity of 80.05% in the training dataset and 79.82% in the test dataset, employing a threshold of 0.345. This accomplishment aligns harmoniously with the CEO's established target sensitivity of approximately 80%.
- Accurate Identification: Sensitivity's function as a metric for correctly identifying potential converting leads underscores the model's precision in recognizing valuable prospects.
- Objective Accomplished: With an accuracy of 80.46%, the model aligns precisely with the predefined study objectives, confirming its effectiveness in lead prediction.
- The model highlights four pivotal features that significantly contribute to predicting hot leads:

❖ Lead source welingak Website

❖ Lead Source_Reference

❖ Current_occupation_Working Professional

❖ Last Activity_SMS Sent

RECOMMENDATIONS FOR ENHANCING LEAD CONVERSION RATES

- ✓ Focus on Positive Coefficients: Prioritize features with positive coefficients for targeted and strategic marketing campaigns.
- ✓ Target High-Quality Leads: Devise strategies to attract top-performing leads from successful lead sources.
- ✓ Engage Professionals: Tailor messaging to engage working professionals, capitalizing on their higher conversion potential.
- ✓ Optimize Communication: Optimize communication channels based on their effectiveness in engaging leads.
- ✓ Allocate Budget: Allocate more budget to Welingak Website advertising to maximize its impact.
- ✓ Encourage References: Introduce incentives or discounts for successful reference conversions to stimulate more referrals.
- ✓ Target Working Professionals: Aggressively target working professionals due to their higher conversion rates and stronger financial position.
- ✓ Leverage Last Activity: Utilize the feature Last Activity_SMS Sent for better monitoring and improvement of conversion rates.

Areas for Improvement:

- **Specialization Enhancement:** Examine negative coefficients related to specialization offerings for potential improvements.
- **Landing Page Evaluation:** Review the landing page submission process to identify and address areas for enhancement.

These recommendations offer actionable insights to elevate lead conversion rates and fine-tune the overall lead management strategy.