

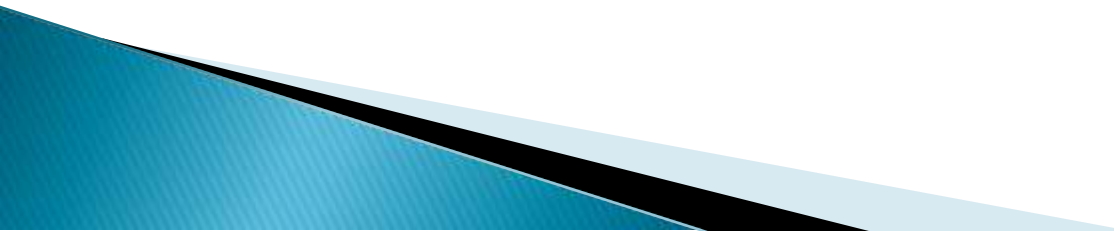
Lead Score Case Study

Amey Basangoudar

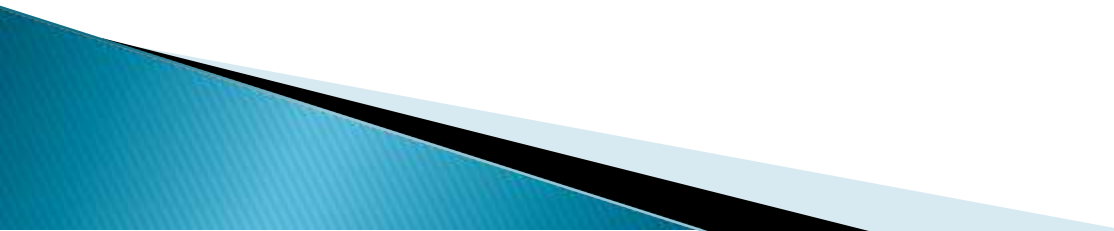
Bhuvanesh Prasad

Shreya Singh

Problem Statement

- X Education offers online courses for industry professionals.
 - Despite generating a significant number of leads, the company struggles with a low lead conversion rate. For instance, out of 100 acquired leads per day, only around 30 are converted.
 - The company aims to enhance efficiency by pinpointing the most promising leads, referred to as 'Hot Leads.'
 - Successful identification of these high-potential leads is expected to boost the lead conversion rate. This improvement is attributed to the sales team's increased focus on communicating with potential leads, rather than making calls to everyone indiscriminately.
- 

Business Goals

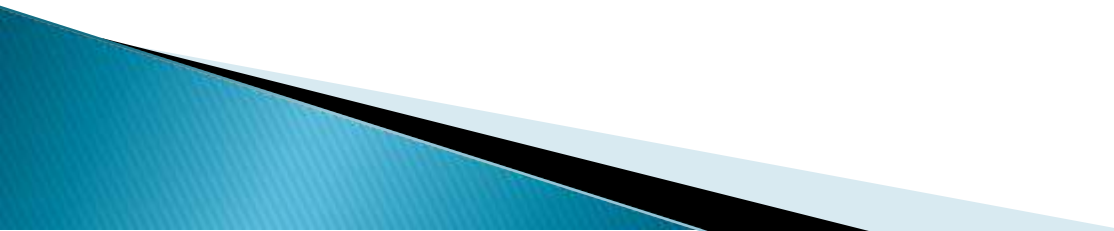
- The company needs a model to be constructed for the purpose of selecting the most promising leads.
 - Implement a lead scoring system to assign a score to each lead, reflecting its potential for conversion.
 - The lead score serves as an indicator, with higher scores suggesting greater potential for conversion and lower scores indicating reduced chances of conversion.
 - The goal is to build a model that achieves a lead conversion rate of approximately 80%.
- 

Methodology Overview

➤ **Data Cleaning and Manipulation:**

- ❖ Address duplicate data.
- ❖ Handle NA values and missing data.
- ❖ Drop columns with a significant amount of missing values and no relevance for analysis.
- ❖ Impute values as needed.
- ❖ Check and manage outliers in the dataset.

➤ **Exploratory Data Analysis (EDA):**

- ❖ Conduct univariate data analysis, including value counts and variable distributions.
 - ❖ Perform bivariate data analysis, examining correlation coefficients and patterns between variables.
- 

➤ **Feature Scaling, Dummy Variables, and Data Encoding:**

- ❖ Implement feature scaling.
- ❖ Create dummy variables and encode the data to prepare for analysis.

➤ **Model Presentation:**

- ❖ Present the developed model, highlighting key features and outcomes.

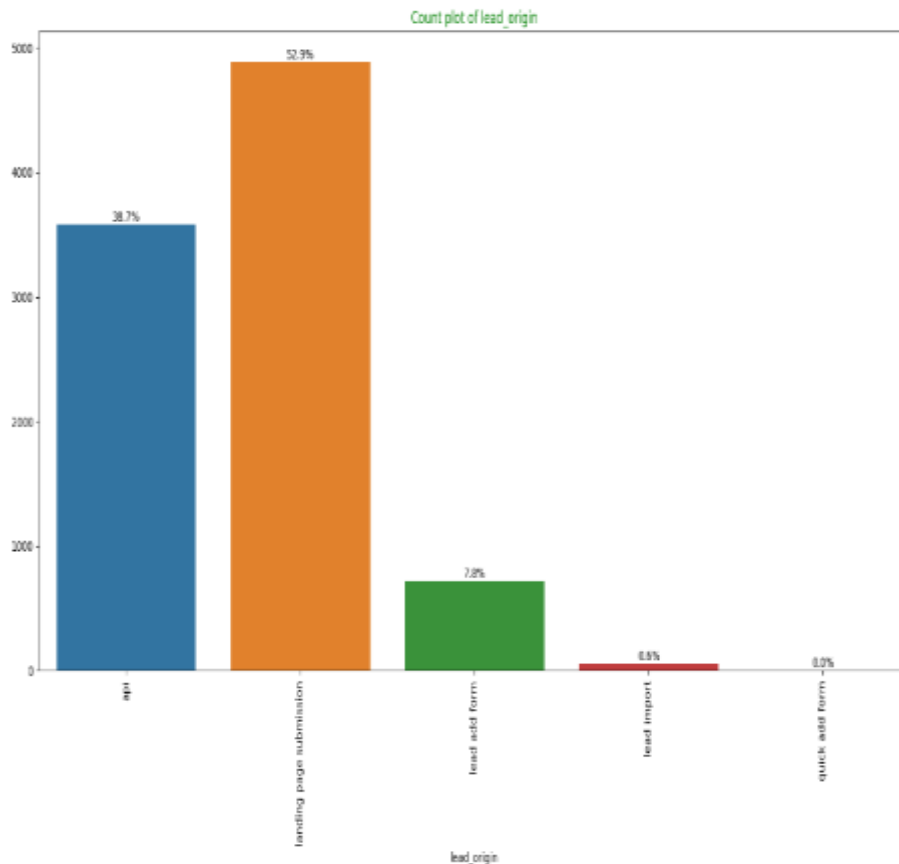
➤ **Model Evaluation:**

- ❖ Assess the model using various measures and metrics.

➤ **Conclusions :**

- ❖ Draw conclusions based on the analysis.
- 

EDA - Exploratory Data Analysis

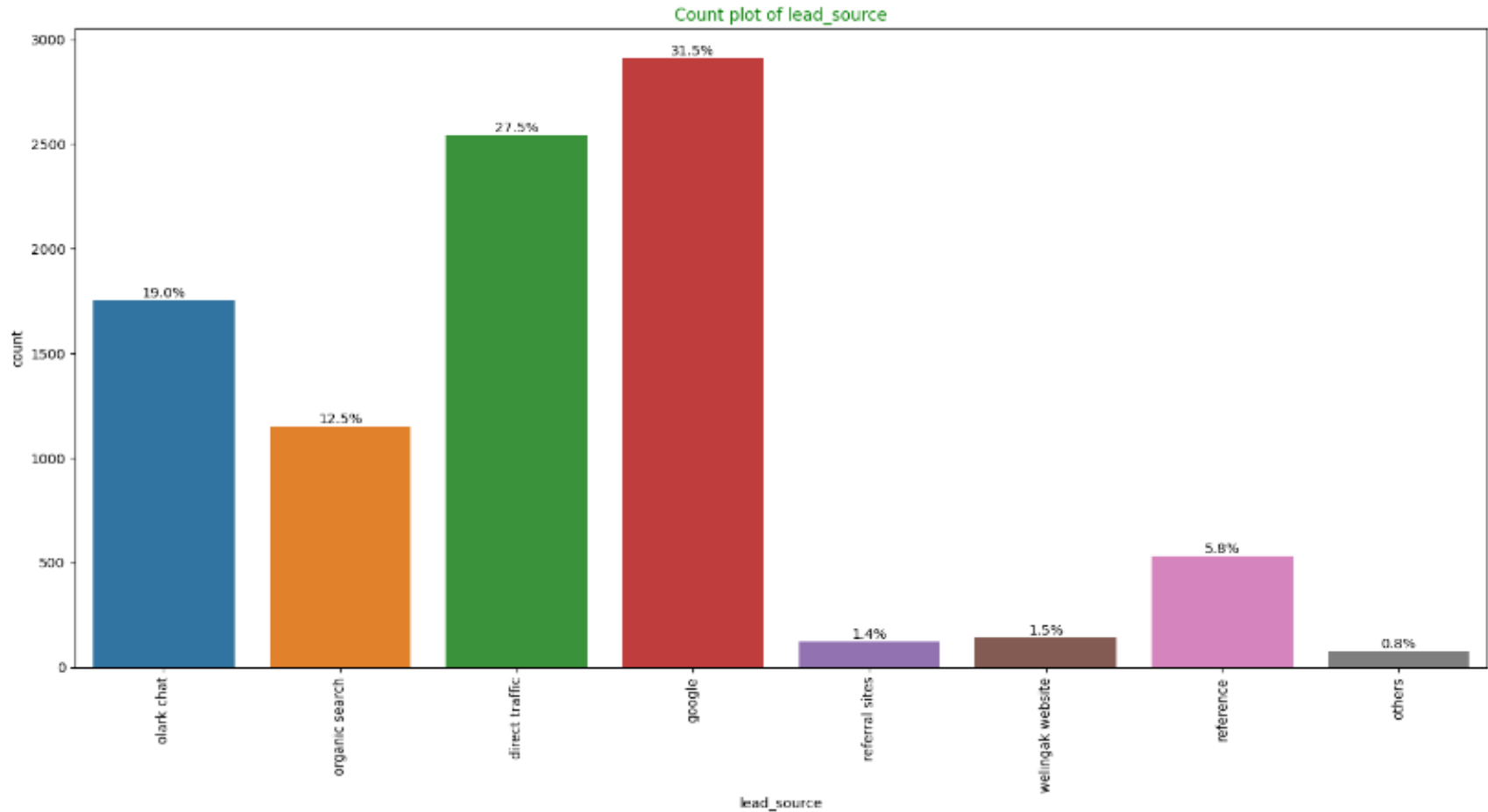


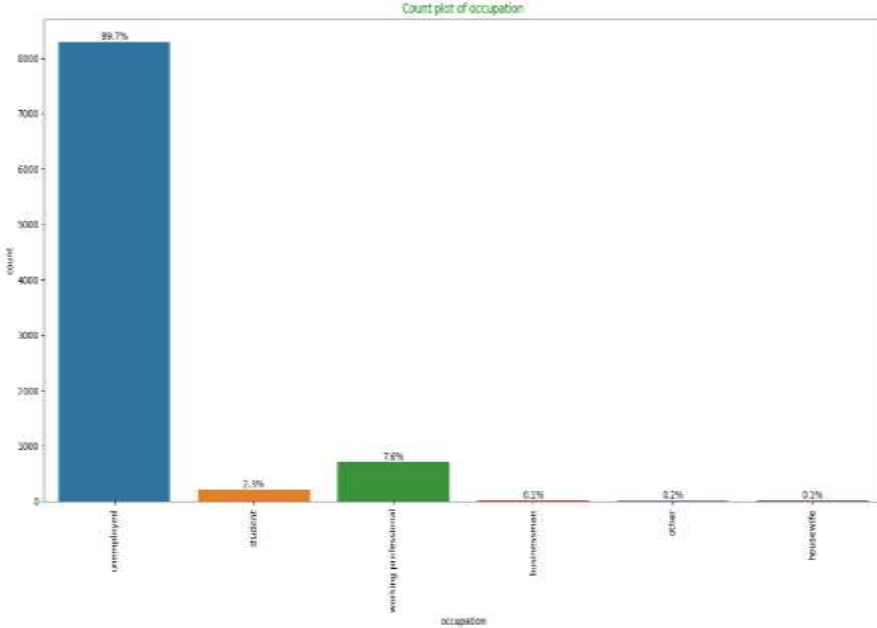
Lead Origin:

➤ Most leads, 53%, originate from 'landing_page_submission,' while 'api' closely follows as the second most prevalent, accounting for 39% of customers.

Lead Source:

- A significant portion of leads (58%) is attributed to a blend of "google" and "direct_traffic."



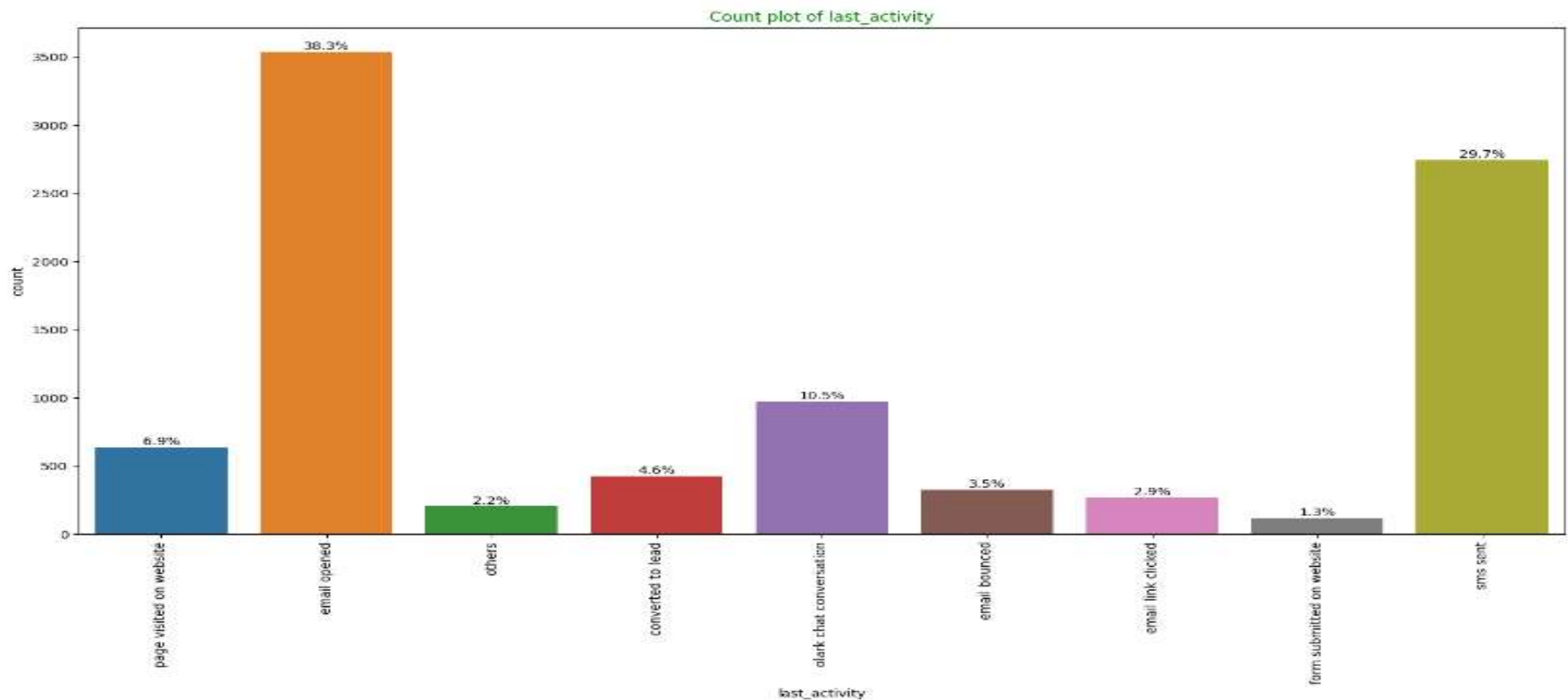


Last Activity:

➤ Approximately 68% of customer interactions are associated with activities like "sms_sent" and "email_opened".

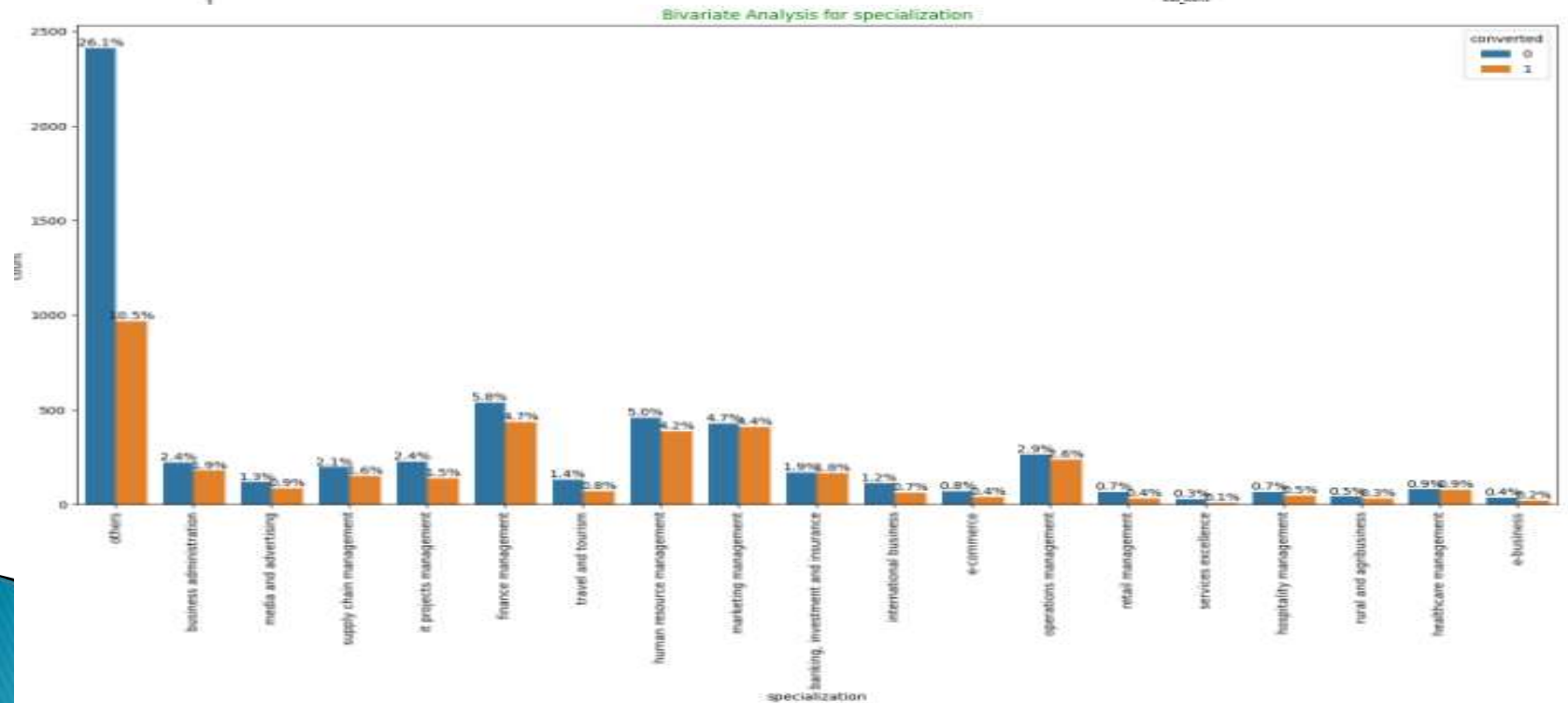
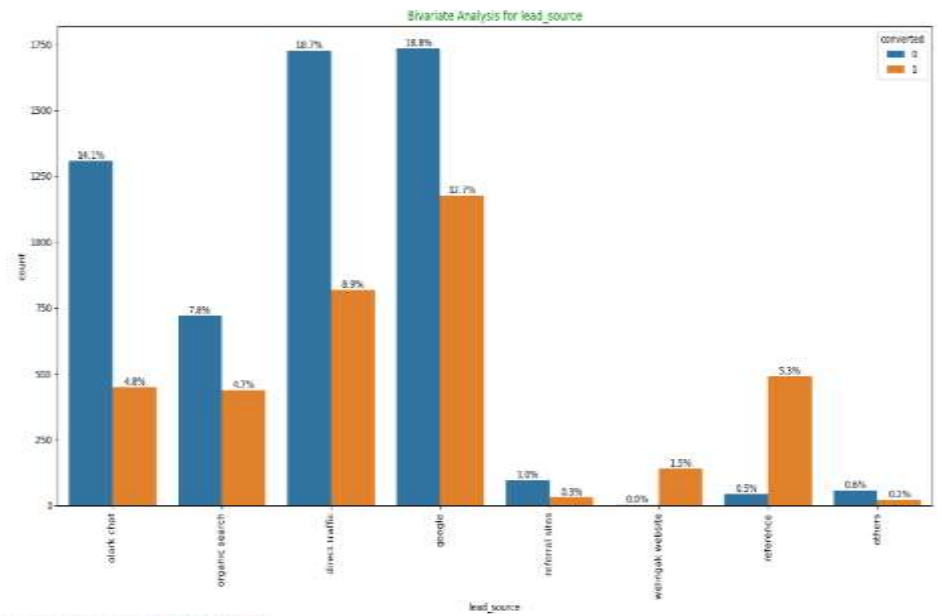
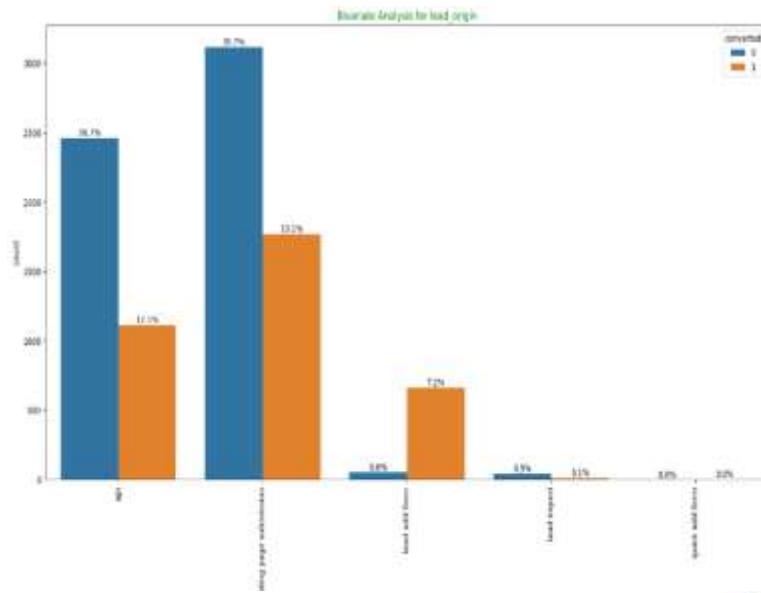
Current Occupation:

➤ The majority, around 90% of customers, are categorized as "unemployed" in the current occupation field.

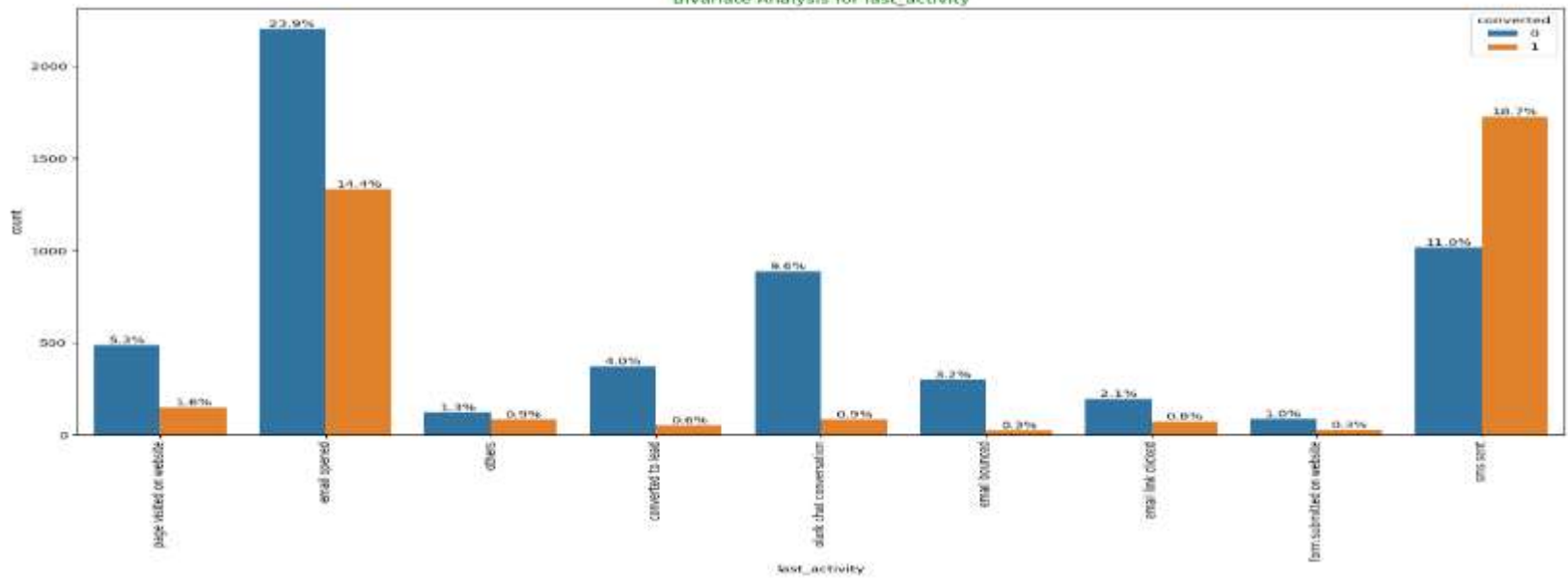


BiVariate Analysis

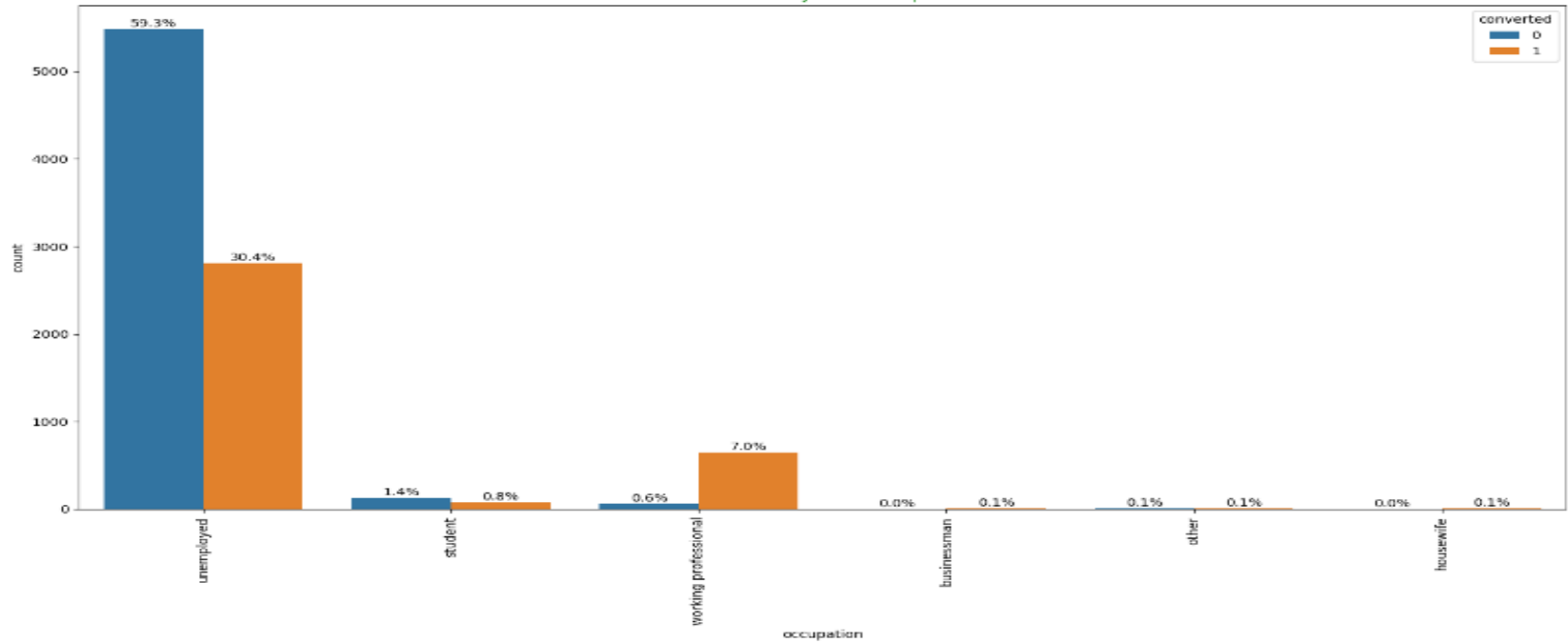
- Lead Origin: "Landing Page Submission" - 52%, LCR 36%; "API" - 39%, LCR 31%.
- Occupation distribution: 90% "Unemployed" with a 34% LCR; 7.6% "Working Professionals" with a high 92% LCR.
- Do Not Email:- 92% opt-out of emails.
- Lead Source distribution: "Google" (31%, LCR 40%), "Direct Traffic" (27%, LCR 32%), "Organic Search" (12.5%, LCR 37.8%), "Reference" (6%, high LCR 91%).
- Last Activities: 'SMS Sent' - 30%, LCR 63%; 'Email Opened' - 38%, LCR 37%.
- Specialization: Positive contributions from Marketing, HR, and Finance Management.
- Lead conversion rate is calculated as the percentage of conversions divided by total leads, represented by the formula $\frac{\%converted}{total_leads}$; for example, the API's conversion rate is approximately 31% ($\frac{12.1\%}{38.8\%}$).



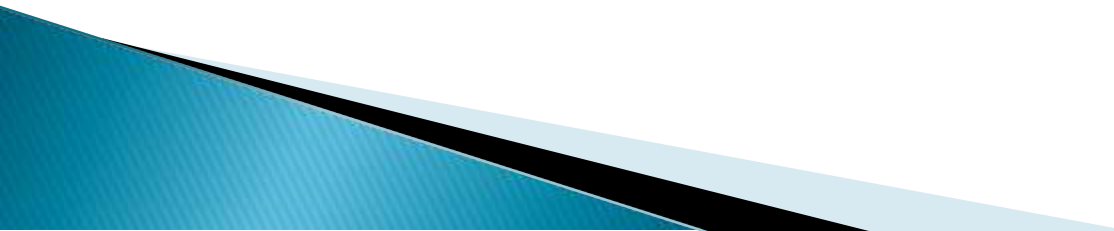
Bivariate Analysis for last_activity



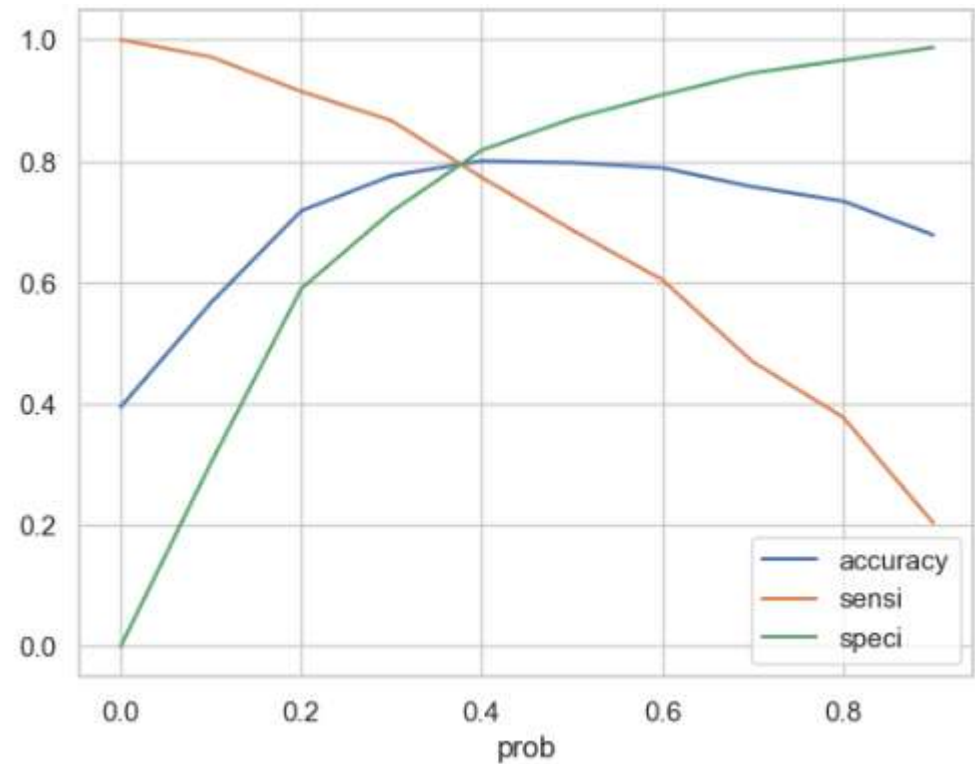
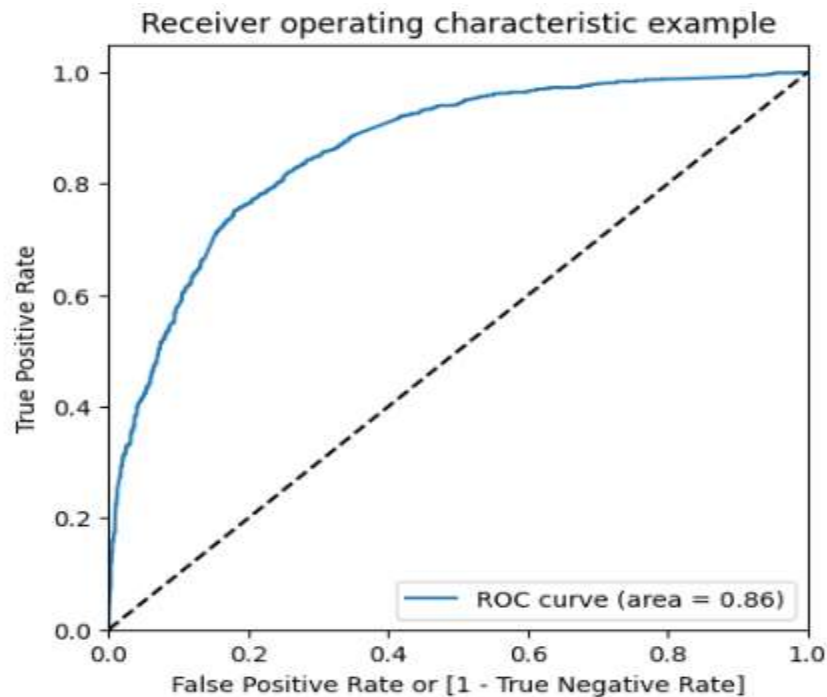
Bivariate Analysis for occupation



Model Building

- Perform a train-test split with a 70:30 ratio as the initial step for regression.
 - Utilize Recursive Feature Elimination (RFE) for feature selection.
 - Run RFE with an output of 15 variables.
 - Build the model by eliminating variables with a p-value greater than 0.05 and a VIF value greater than 5.
 - Make predictions on the test dataset.
 - Achieve an overall accuracy of 80%.
- 

Model Evaluation:(ROC Curve)



➤ An area under the ROC curve of 0.88 signifies the model's effectiveness.

➤ Based on the provided curve, the optimal cutoff probability is identified at 0.35.

Conclusion

➤ Training Dataset Metrics:

- ❖ Accuracy: 80.88%
- ❖ Sensitivity: 80.61%
- ❖ Specificity: 81.04%

➤ Testing Dataset Metrics:

- ❖ Accuracy: 77.49%
- ❖ Sensitivity: 79.84%
- ❖ Specificity: 75.95%

- The proximity of accuracy, sensitivity, and specificity between the training and testing datasets indicates a robust model. Additionally, the achieved sensitivity of around 80% aligns with the CEO's target, with an overall accuracy of 80.88%, meeting the study's objectives.

Thank You
»»