Lead Score Case Study

Shivam Yadav
Chandra Shekhar
Shraddha Singh

Problem Statement

- - X Education offers specialized online courses designed for professionals in various industries.
- - The company is committed to optimizing its operational efficiency by identifying the most promising leads, commonly referred to as 'Hot Leads.'
- - Despite generating a substantial volume of leads, the company faces the challenge of a low lead conversion rate. To put it into perspective, out of every 100 acquired leads per day, only approximately 30 successfully convert.
- - The successful identification of these high-potential leads is expected to significantly elevate the lead conversion rate. This anticipated improvement is attributed to the sales team's enhanced focus on targeted communication with potential leads, prioritizing meaningful engagement over indiscriminate outreach to everyone.

Business Goals

- - The company is seeking the creation of a model designed to identify the most promising leads.
- - By implementing a lead scoring system, each lead will be assigned a score, providing insights into its potential for conversion.
- - The lead score functions as an indicator, where higher scores signify a greater likelihood of conversion, while lower scores suggest diminished chances.
- - The primary goal is to construct a model that achieves an ambitious lead conversion rate of approximately 80%.

Methodology Overview

Data Cleaning and Manipulation:

- - Identify and address duplicate data entries.
- - Handle NA values and missing data through appropriate methods.
- - Drop columns with a substantial number of missing values that hold no significance for analysis.
- - Impute values where necessary to enhance data completeness.
- - Check for outliers in the dataset and implement appropriate management strategies.

Exploratory Data Analysis (EDA):

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

Feature Scaling, Dummy Variables, and Data Encoding:

- - Normalize feature values through feature scaling.
- - Create dummy variables and encode the data for streamlined analysis.

Model Presentation:

- - Highlight the developed model, emphasizing crucial features and outcomes.

Model Evaluation:

- - Assess the model's performance using various measures and metrics.

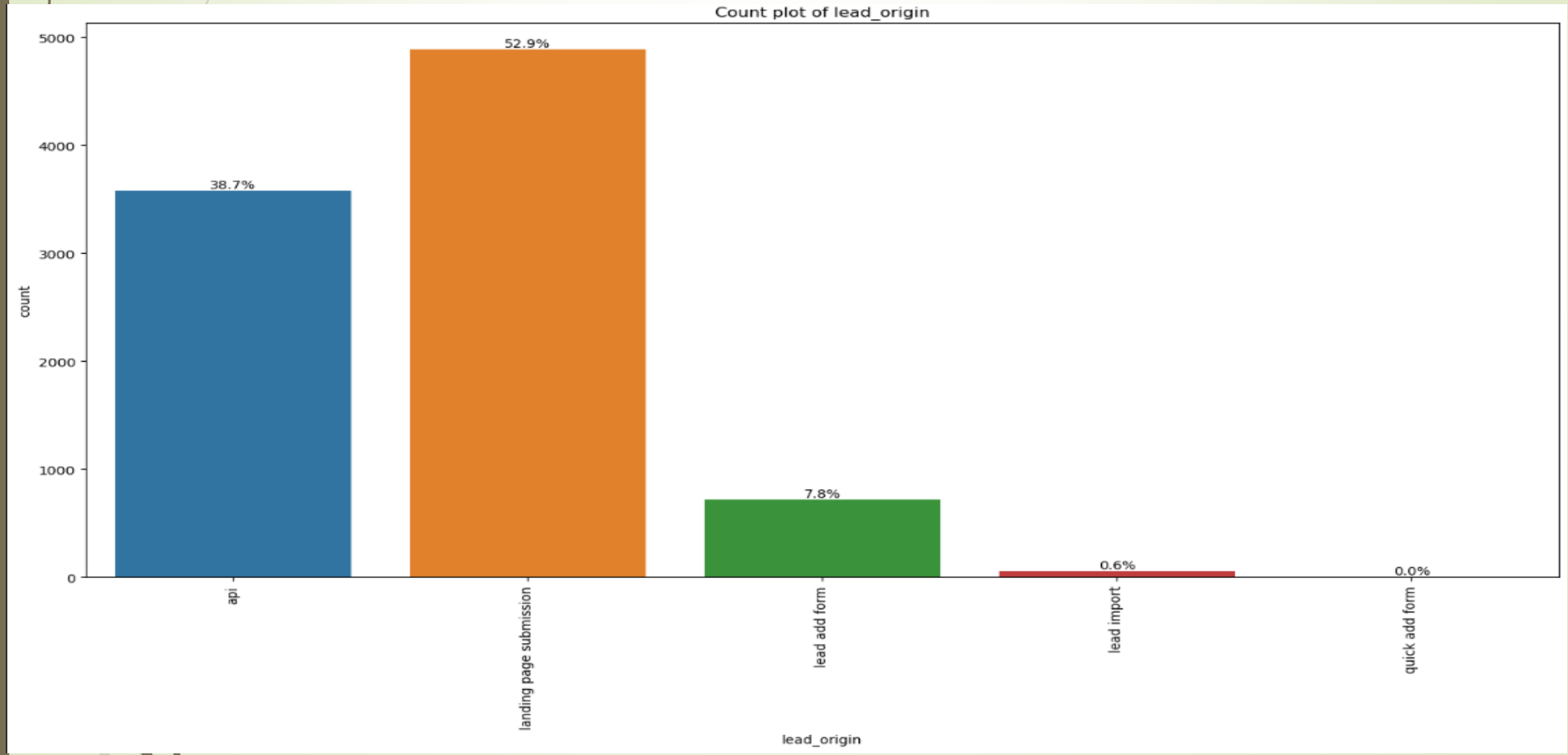
Conclusions:

- - Formulate insightful conclusions based on the analysis findings.

EDA - Exploratory Data Analysis

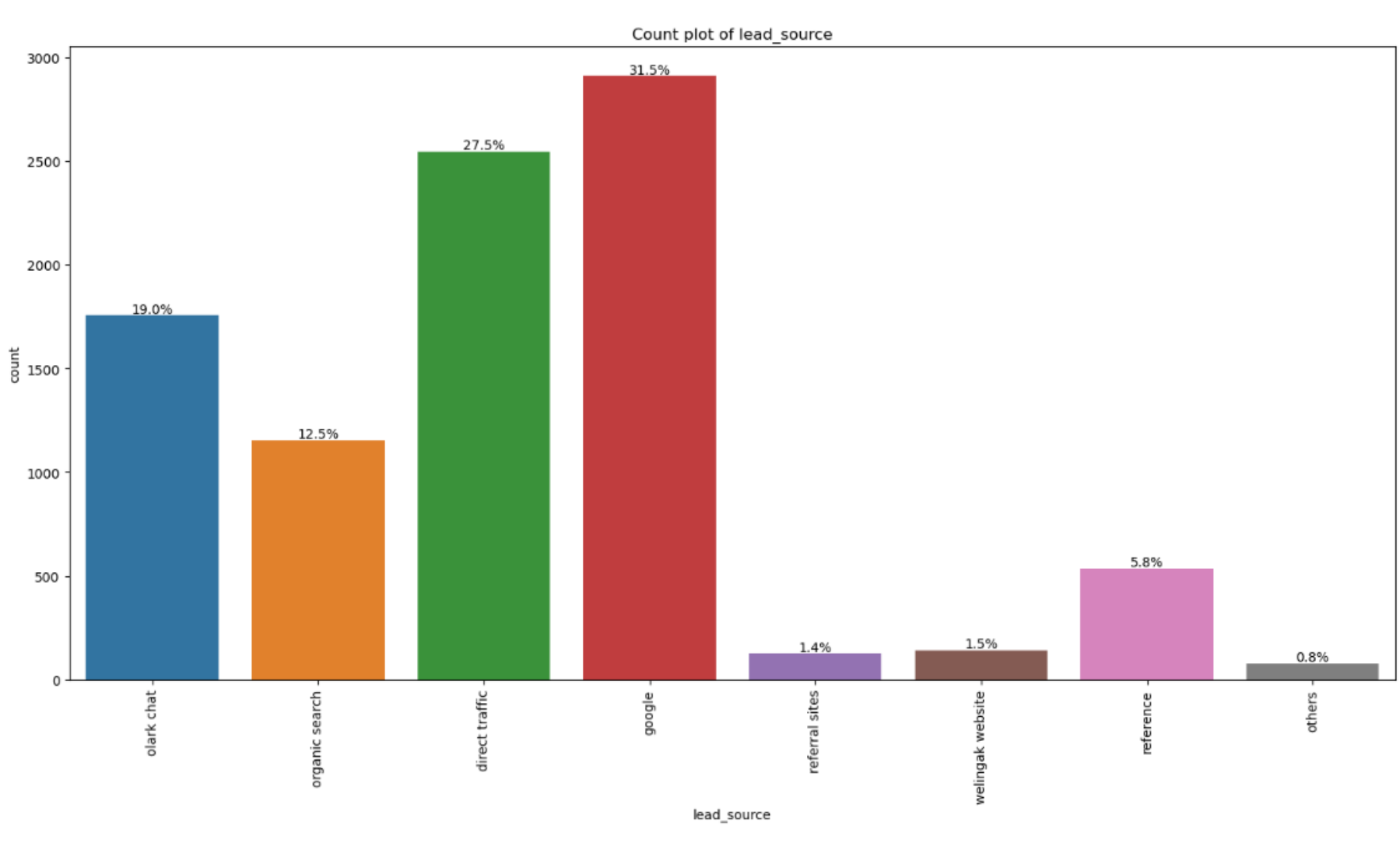
Lead Origin:

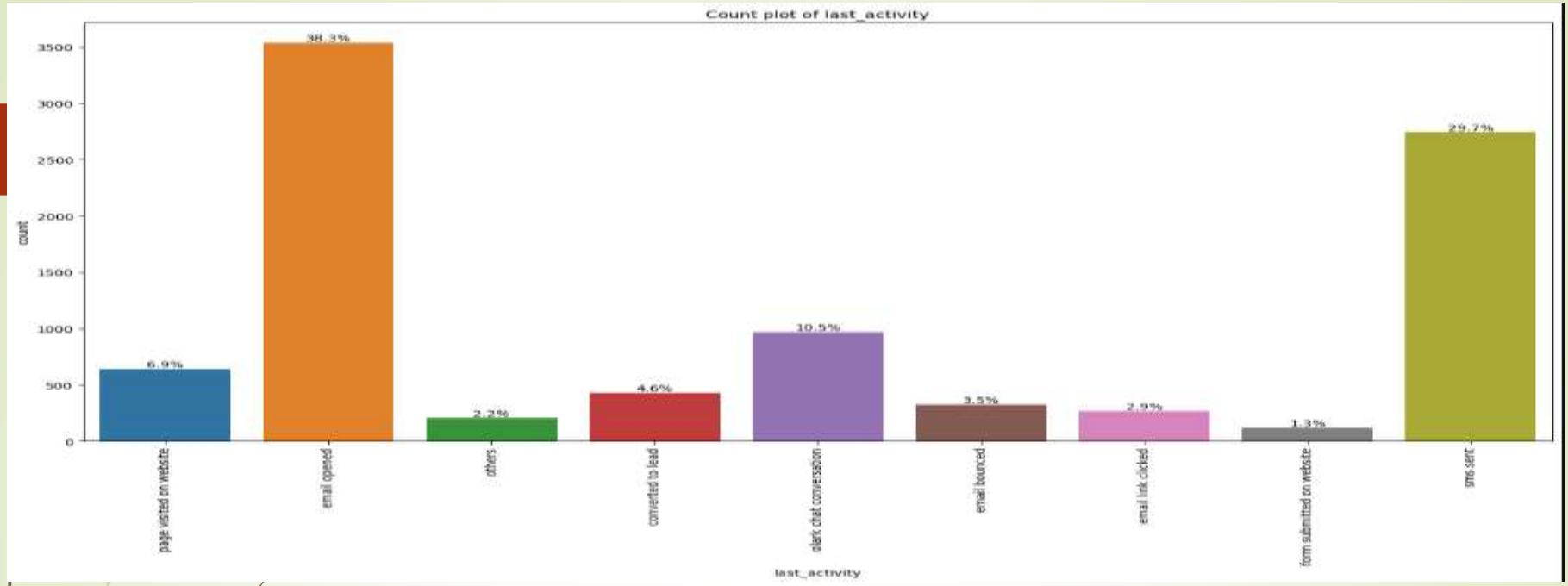
- The majority of leads, constituting 53%, initiate from 'landing_page_submission,' with the second most prevalent source being 'api,' representing 39% of customers.

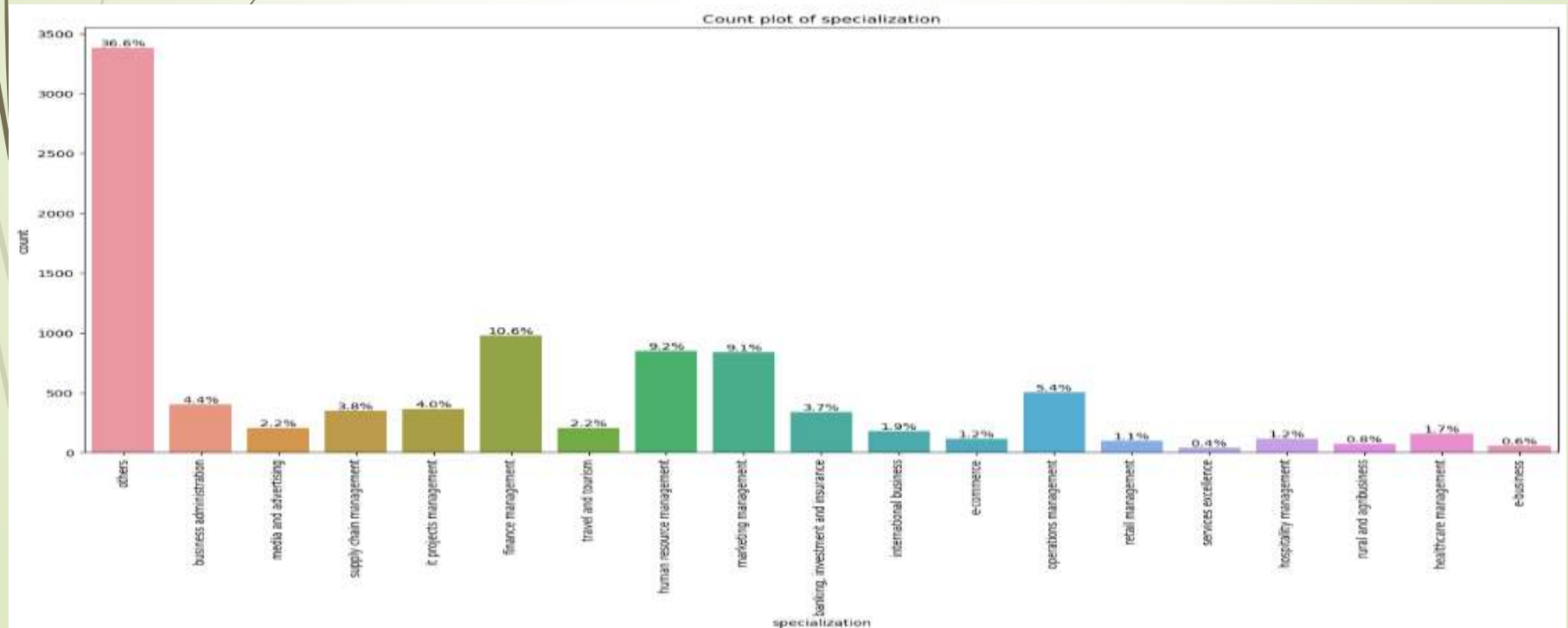


Lead Source:

- A considerable proportion of leads, amounting to 58%, is a combination of "google" and "direct_traffic."




Count plot of last_activity
count
last_activity
38.3%
29.7%
10.5%
6.9%
4.6%
3.5%
2.9%
2.2%
1.3%
page visited on website
email opened
others
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
sms sent


Count plot of specialization
count
specialization
36.6%
10.6%
9.2%
9.1%
5.4%
4.4%
4.0%
3.8%
3.7%
2.2%
2.2%
1.9%
1.2%
1.1%
1.2%
1.7%
0.4%
0.8%
0.6%
others
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

Observations:

- Roughly 68% of customer interactions are linked to activities like "sms_sent" and "email_opened.“

Current Occupation:

- The prevailing designation in the current occupation field is "unemployed," representing approximately 90% of customers.

BiVariate Analysis

- **Lead Origin:**
  - "Landing Page Submission" accounts for 52%, with a Lead Conversion Rate (LCR) of 36%.
  - "API" constitutes 39%, demonstrating a 31% LCR.

- **Occupation Distribution:**
  - Approximately 90% fall under "Unemployed" with a 34% LCR.
  - "Working Professionals" represent 7.6% but boast a high 92% LCR.

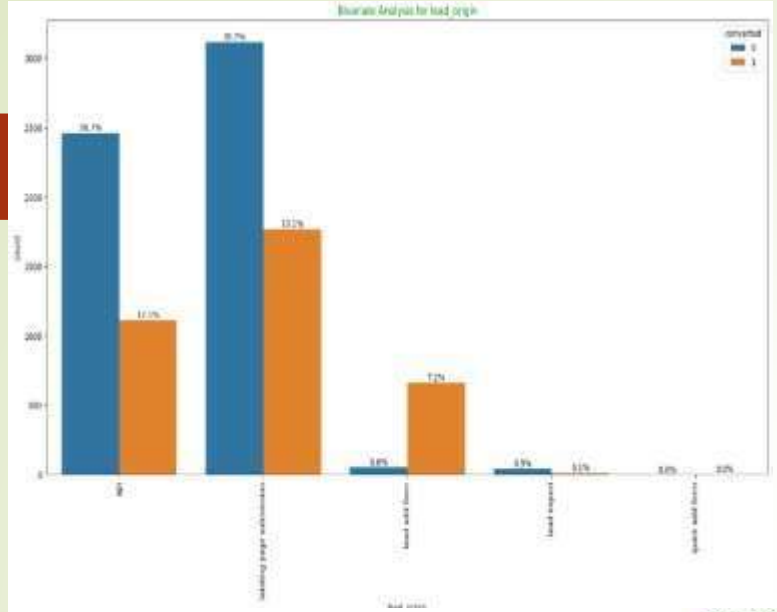
- **Do Not Email:**
  - A significant 92% choose to opt-out of emails.

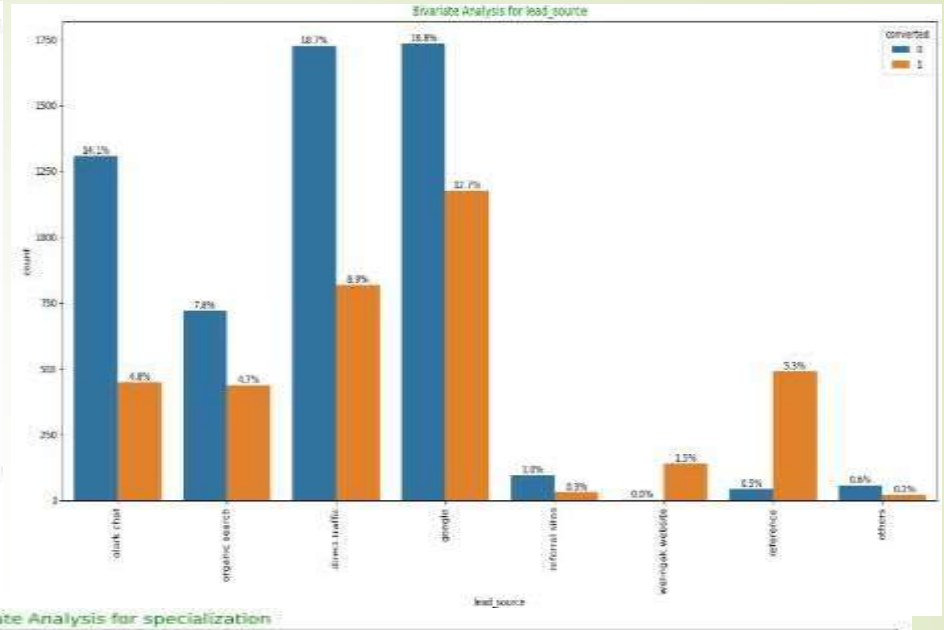
- **Lead Source Distribution:**
  - "Google" holds 31%, exhibiting a 40% LCR.
  - "Direct Traffic" contributes 27%, with a 32% LCR.
  - "Organic Search" represents 12.5%, achieving a 37.8% LCR.
  - "Reference" makes up 6% but has a notable LCR of 91%.

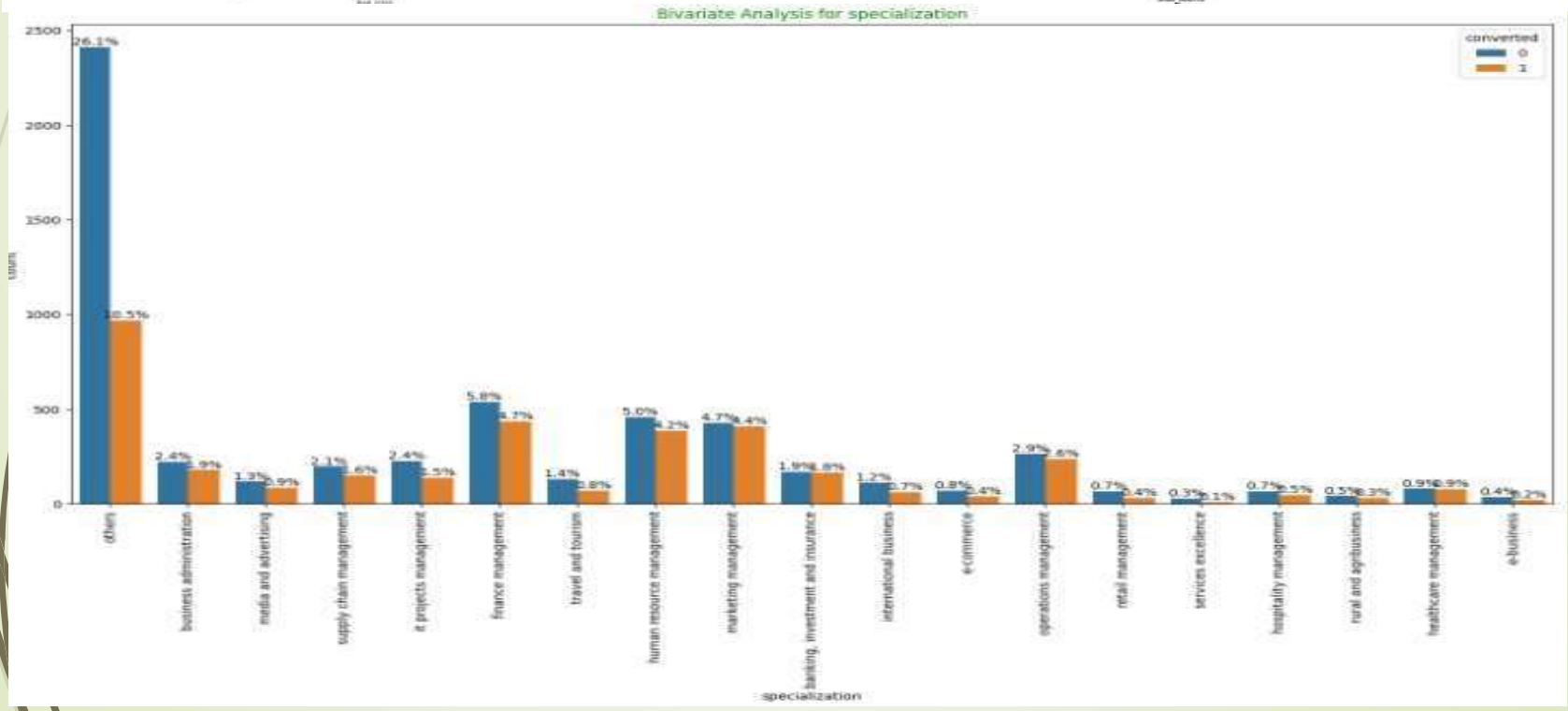
- **Last Activities:**
  - 'SMS Sent' comprises 30%, demonstrating a 63% LCR.
  - 'Email Opened' accounts for 38%, with a 37% LCR.

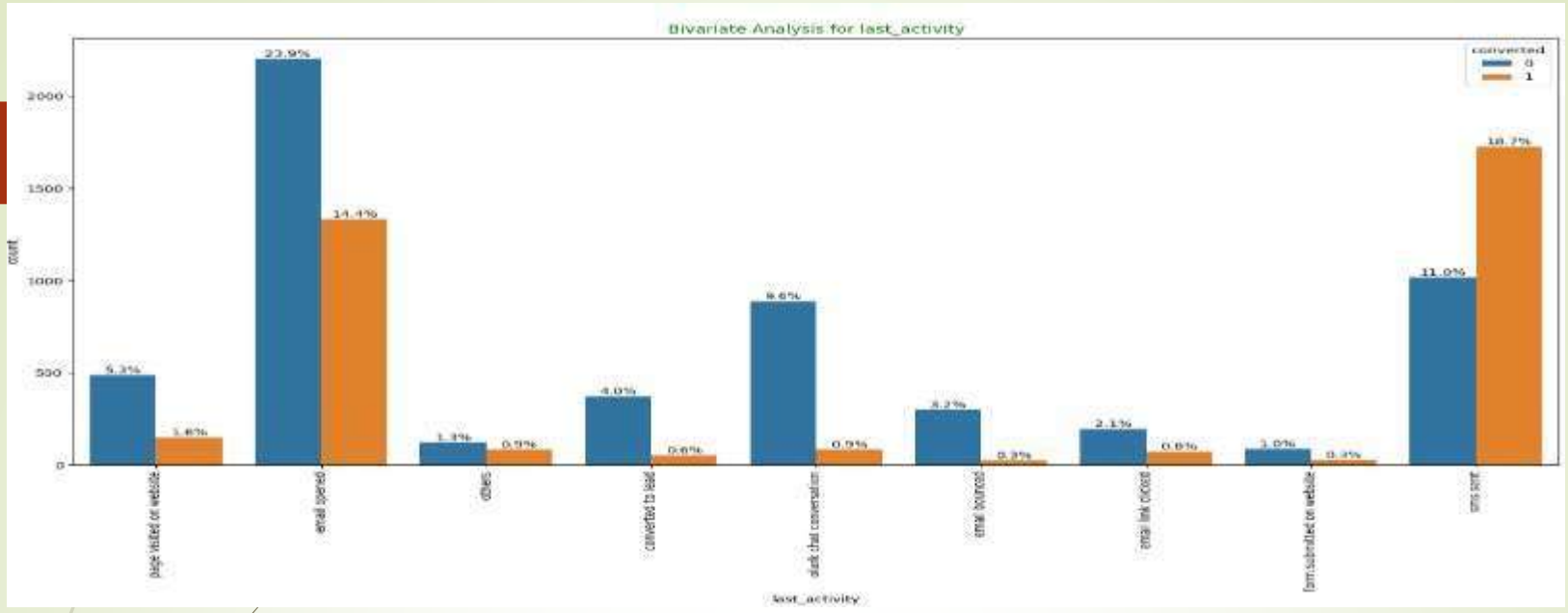
- **Specialization:**
  - Positive contributions are observed from Marketing, HR, and Finance Management.

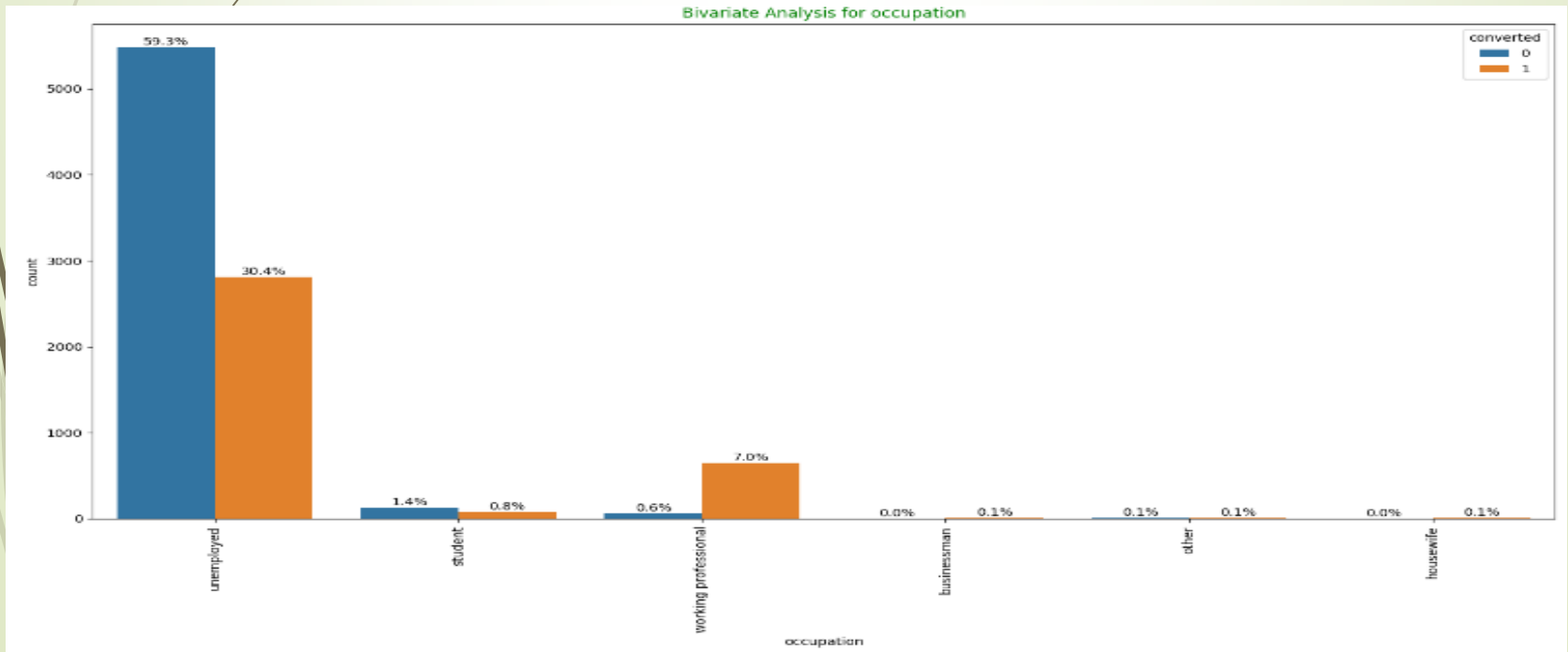
- **Lead Conversion Rate:**
  - Calculated as the percentage of conversions divided by total leads, expressed as %converted / total_leads.
  - Example: The API's conversion rate is approximately 31% (12.1% / 38.8%).




Bivariate Analysis for lead_source
converted
0
1
count
lead_source
olark chat
organic search
direct traffic
google
referral sites
welingak website
reference
others


Bivariate Analysis for specialization
converted
0
1
count
specialization
others
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


Bivariate Analysis for last_activity
converted
0
1
count
last_activity
page visited on website
email opened
others
converted to lead
chat conversation
email bounced
email link clicked
form submitted on website
sms sent
22.9%
14.4%
11.0%
18.7%

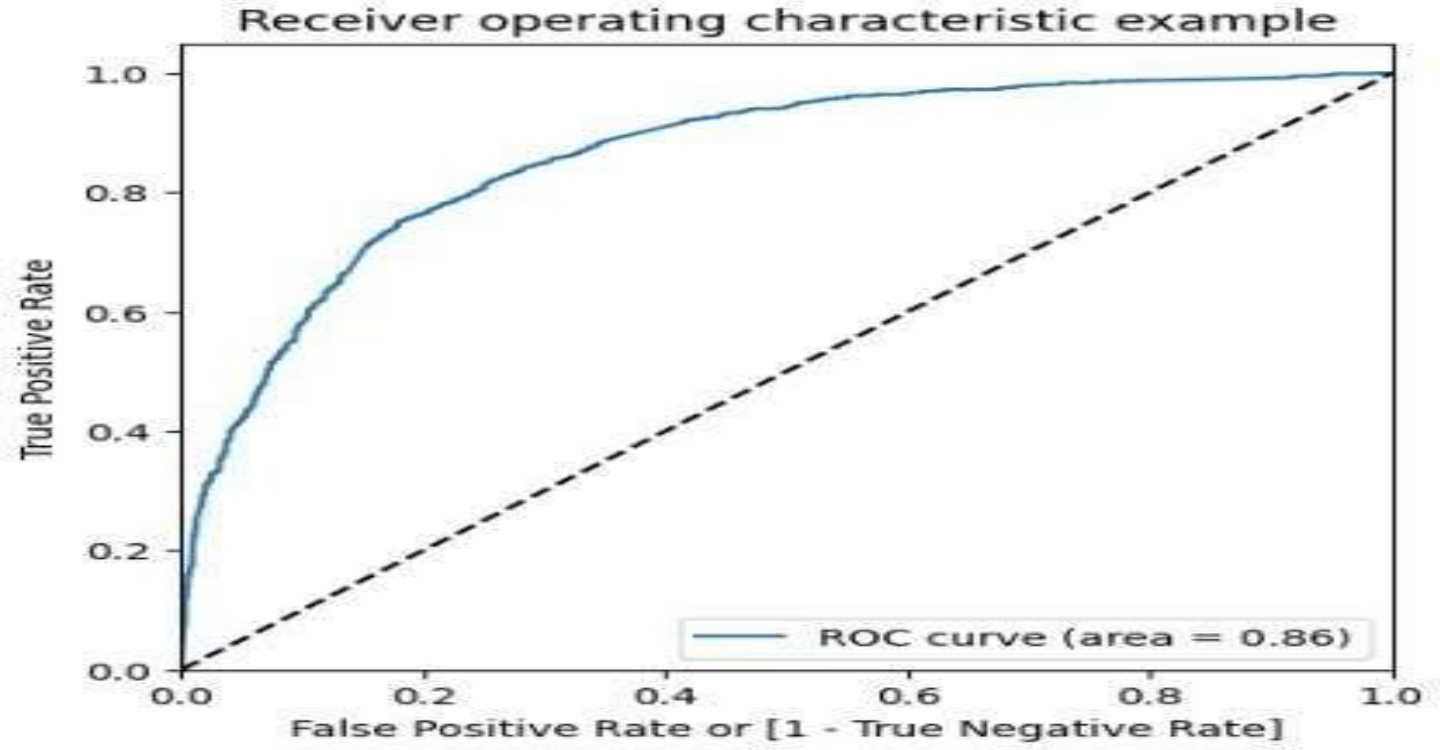

Bivariate Analysis for occupation
converted
0
1
count
occupation
unemployed
student
working professional
businessman
other
housewife
59.3%
30.4%
1.4%
0.8%
0.6%
7.0%
0.0%
0.1%
0.1%
0.1%
0.0%
0.1%

Model Building

- - Begin the regression analysis by conducting a train-test split with a 70:30 ratio.
- - Employ Recursive Feature Elimination (RFE) for feature selection, aiming for an output of 15 variables.
- - Construct the model by excluding variables with a p-value exceeding 0.05 and a Variance Inflation Factor (VIF) surpassing 5.
- - Utilize the model to make predictions on the test dataset.
- - Strive to achieve an overall accuracy of 80%.

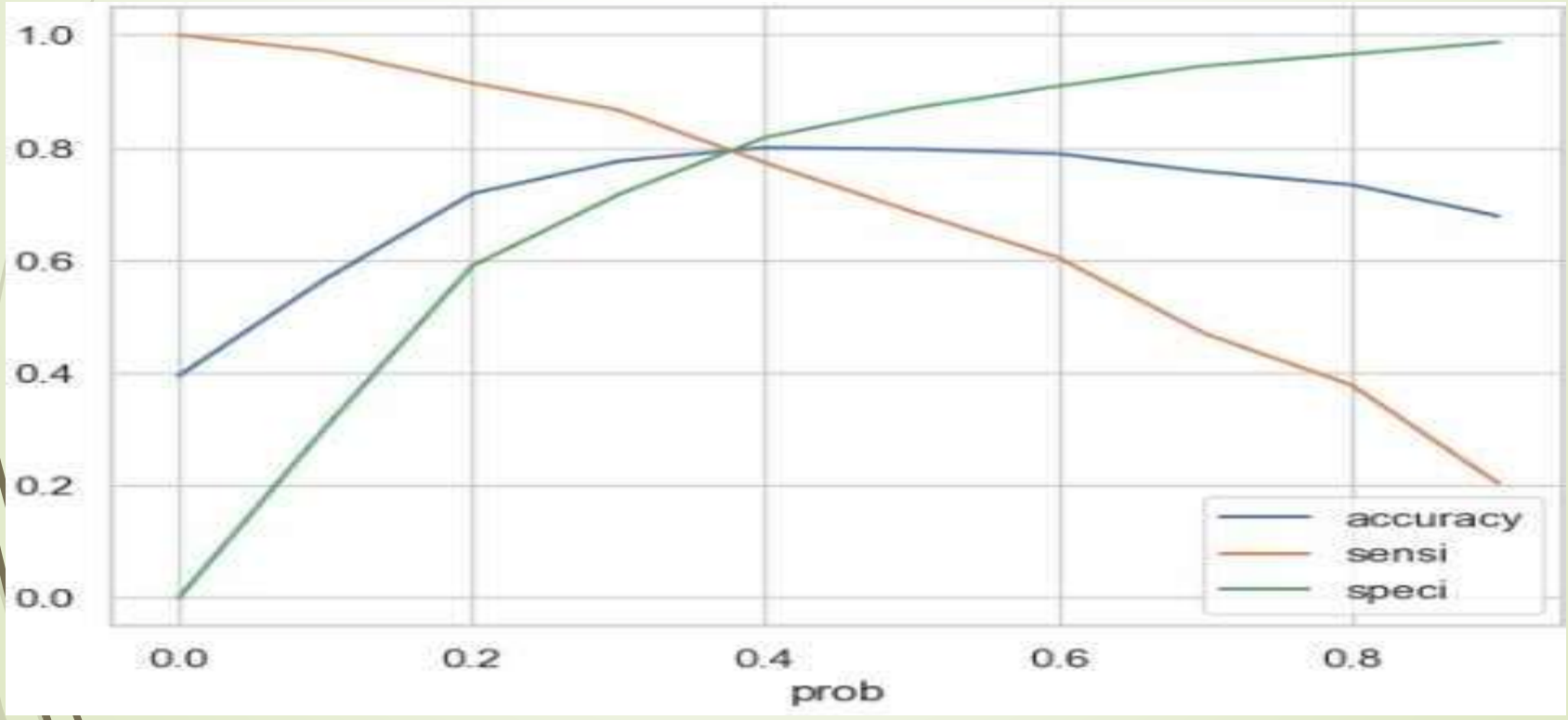
Model Evaluation:(ROC Curve)

- A ROC curve area of 0.88 indicates the model's efficacy.



Model Evaluation:(ROC Curve)

- From the given curve, the optimal cutoff probability is determined to be 0.35.



INSIGHTS AND 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 similarity in accuracy, sensitivity, and specificity values between the training and testing datasets indicates the model's robustness.

- The CEO of X Education aimed for a sensitivity of approximately 80%, and the model met this target while achieving an overall accuracy of 80.88%, aligning well with the study's objectives.

Thank You