CASE STUDY – LEAD SCORING

BY:-
Anshuma Mishra
Manikandan Krishnan
Aparna Ankalkote

PROBLEM STATEMENT:-

- X Education provides online courses tailored for professionals in various industries.
- Despite generating a substantial number of leads, the company faces challenges with a low conversion rate, converting only about 30 out of 100 acquired leads per day.
- To improve efficiency, the company plans to identify the most promising leads, referred to as 'Hot Leads,' with the goal of enhancing the lead conversion rate.
- This improvement is anticipated through the sales team's increased emphasis on communicating with potential leads, as opposed to making indiscriminate calls to everyone.

BUSINESS GOALS:-

- The organization requires the development of a model designed to identify the most favorable leads.
- Create and apply a lead scoring mechanism to assess each lead, indicating its likelihood of conversion.
- Higher lead scores signify increased potential for conversion, while lower scores suggest a diminished likelihood of conversion.
- The objective is to establish a model that attains an approximately 80% lead conversion rate.

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.

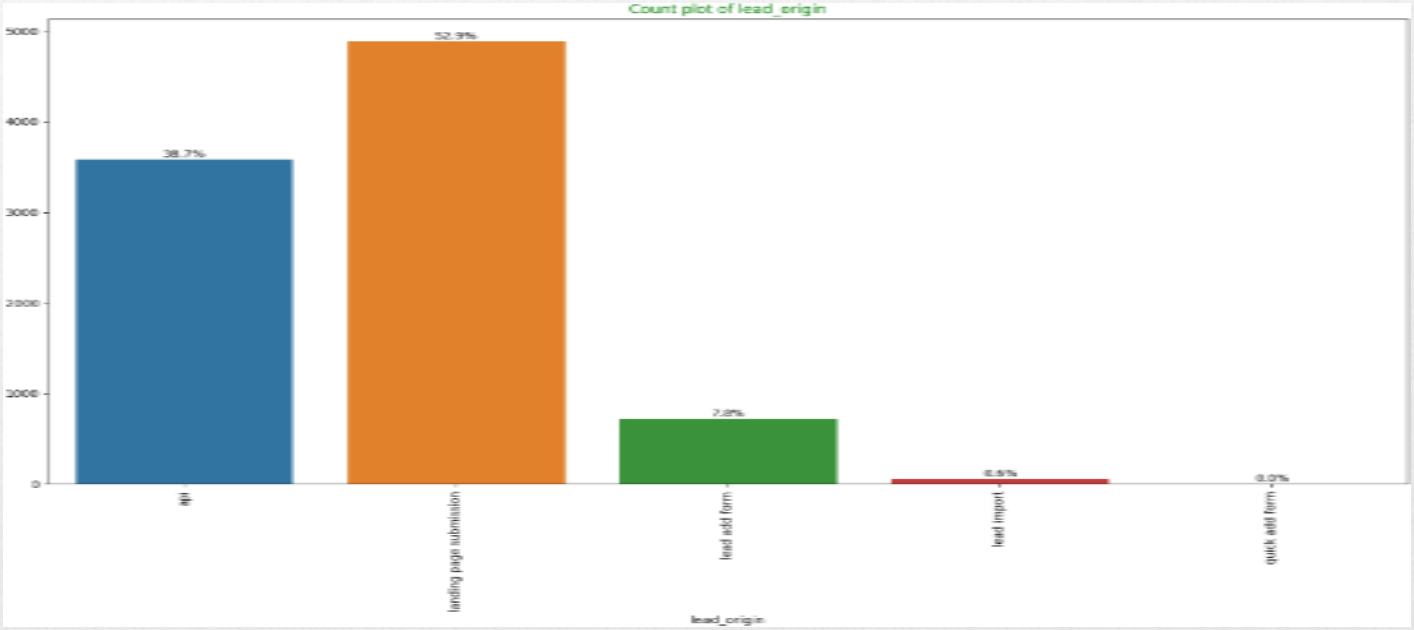
- **Feature Scaling, Dummy Variables, and Data Encoding:**
  - Implement feature scaling.
  - Create dummy variables and encode the data to prepare for analysis.
- **Model Presentation:**
  - Present the developed model, highlighting key features and outcomes.
- **Model Evaluation:**
  - Assess the model using various measures and metrics.
- **Conclusions :**
  - Draw conclusions based on the analysis.

Exploratory Data Analysis – EDA:-

- **Univariate Analysis**

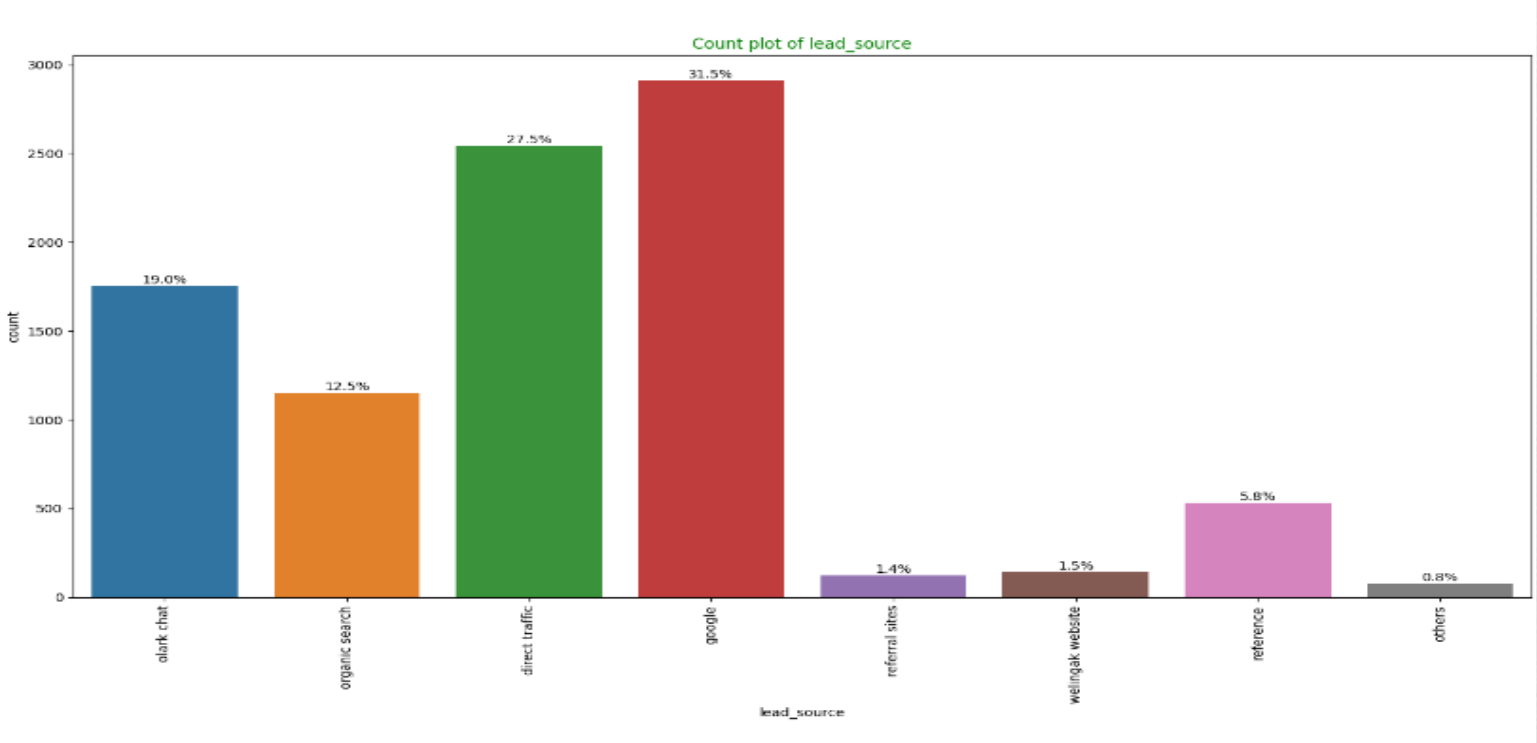
1. *Lead Origin:-*

For lead_origin, "landing_page_submission" emerges as the predominant choice, constituting 53% of the customer base. Following closely is "api," making up 39% of customers.



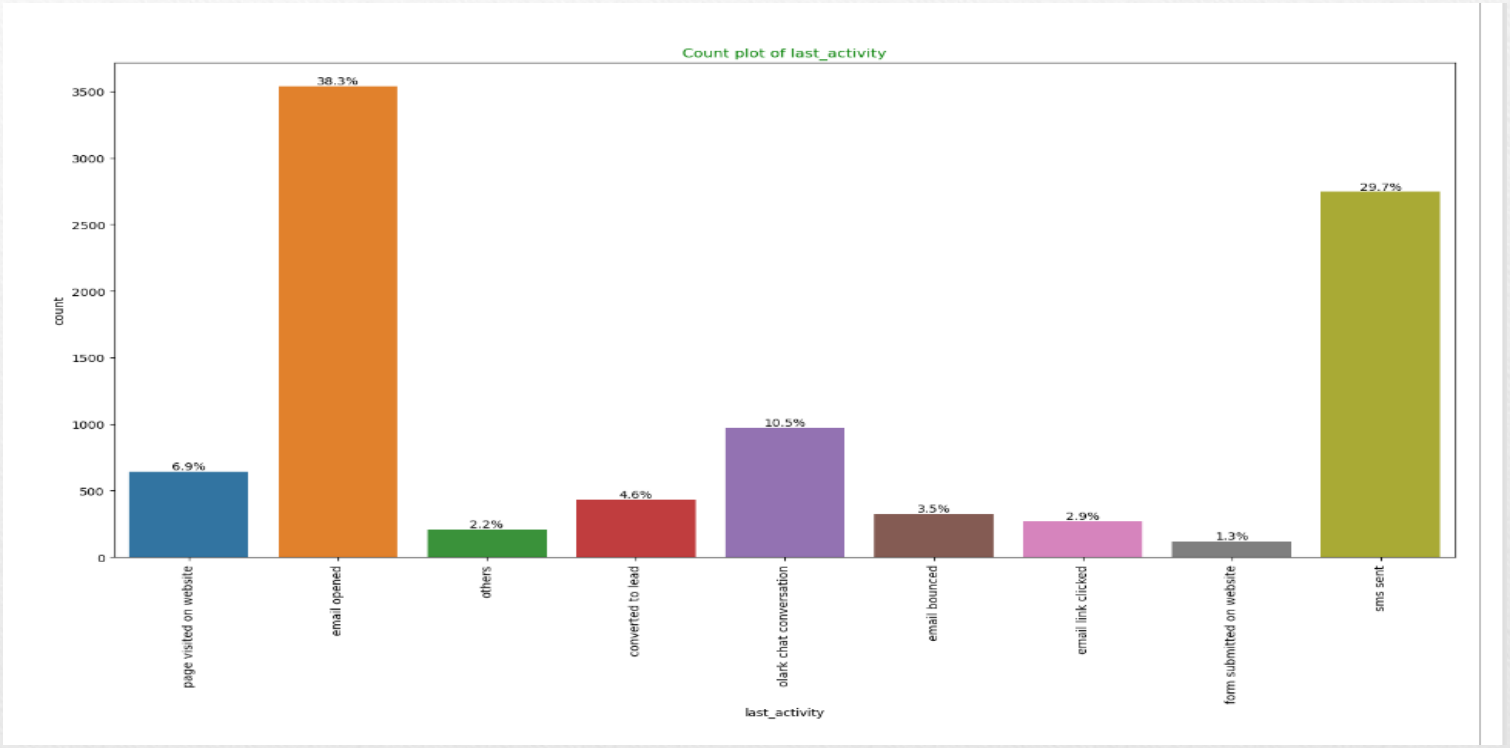
2. *Lead Source:-*

In the lead_source category, the majority of leads (58%) originate from a combination of "google" and "direct_traffic."



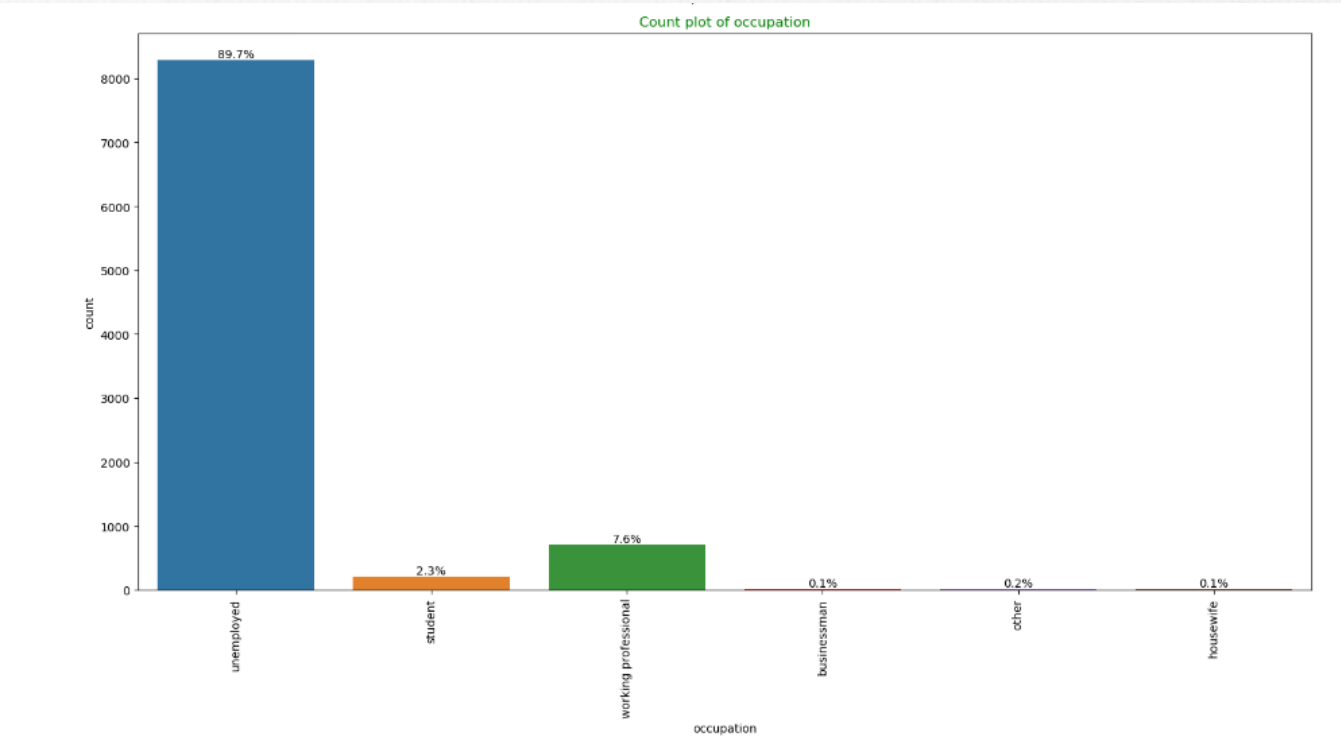
3. *Last Activity:-*

In the last_activity dimension, approximately 68% of customer engagements involve activities like sma_sent and email_opened.



4. *Current Occupation:-*

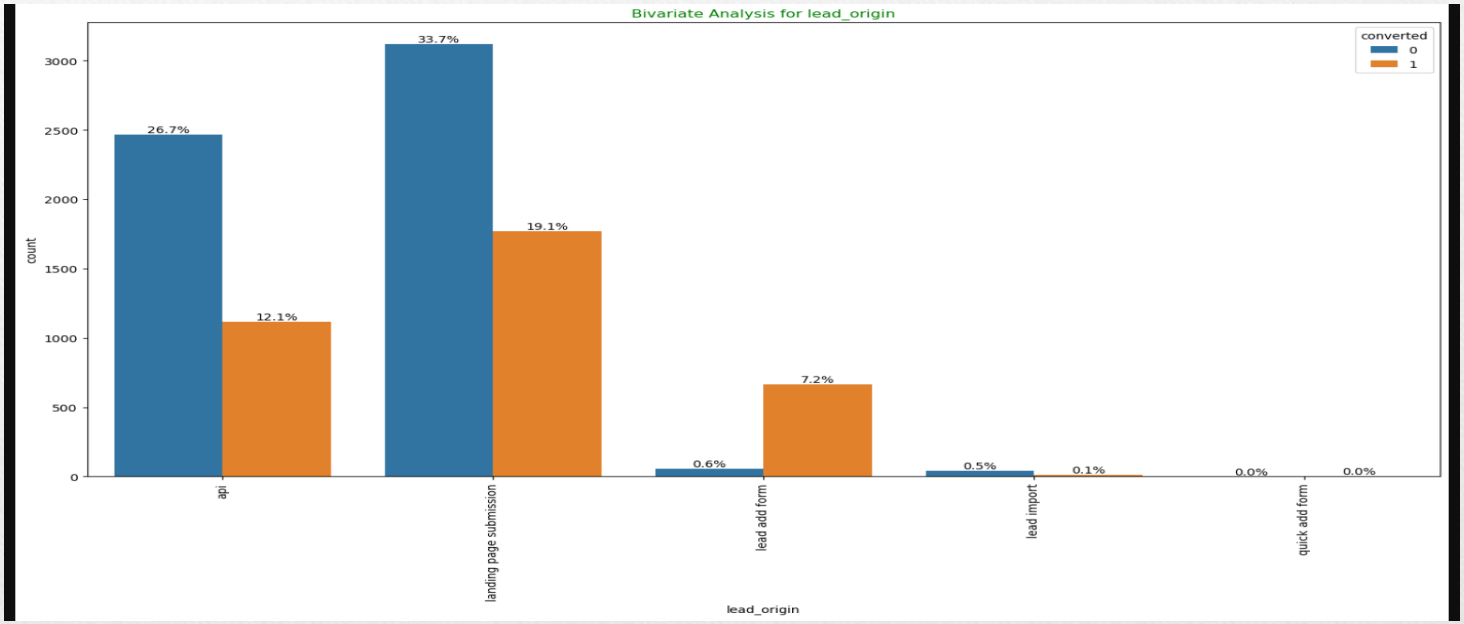
Current Occupation analysis reveals that around 90% of customers fall under the "unemployed" classification in the occupation category.



- **Bivariate Analysis**

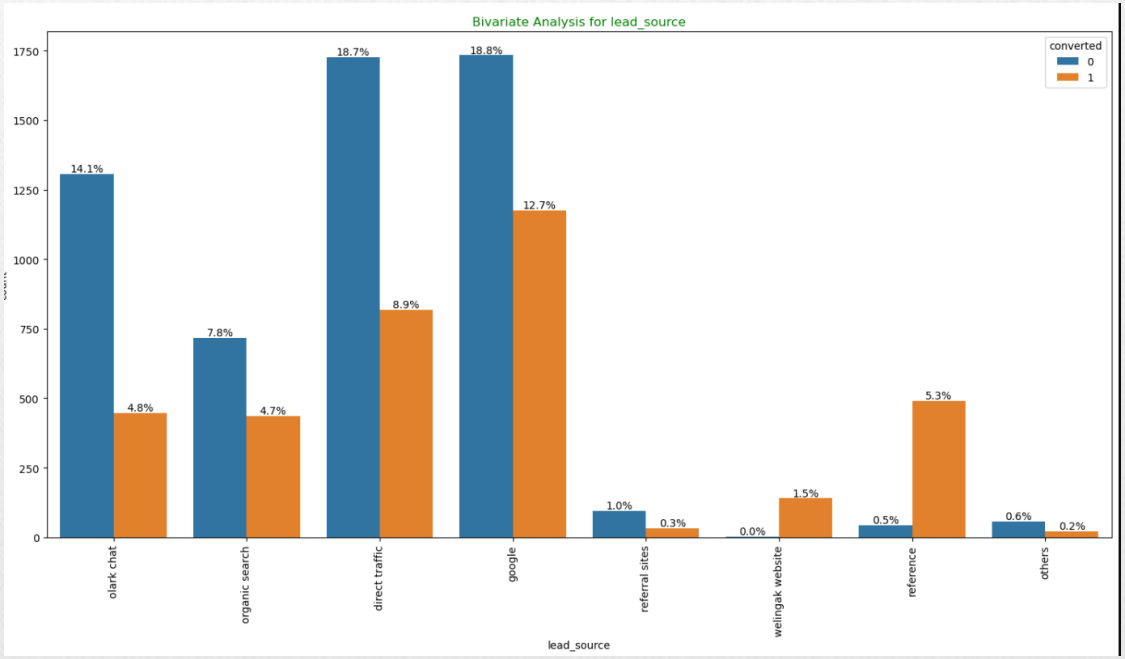
1.<u>Lead Origin</u>:-

Approximately 52% of all leads originated from "Landing Page Submission," boasting a lead conversion rate (LCR) of 36%. The "API" identified around 39% of customers with a corresponding LCR of 31%.



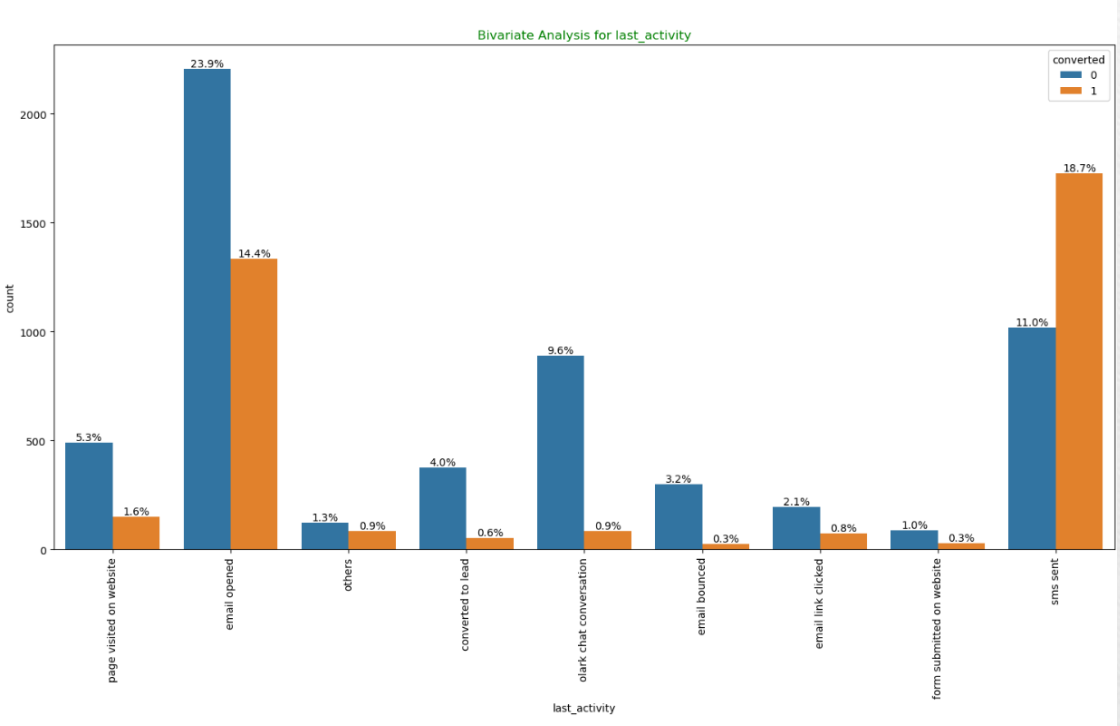
2.*Lead Source*:-

Lead Source: Google stands out with a high LCR of 40% among its 31% customer base. Direct Traffic follows with a 32% LCR and 27% of customers, while Organic Search contributes a 37.8% LCR with a smaller 12.5% of customers. Despite Reference having an LCR of 91%, it comprises only approximately 6% of customers through this Lead Source.



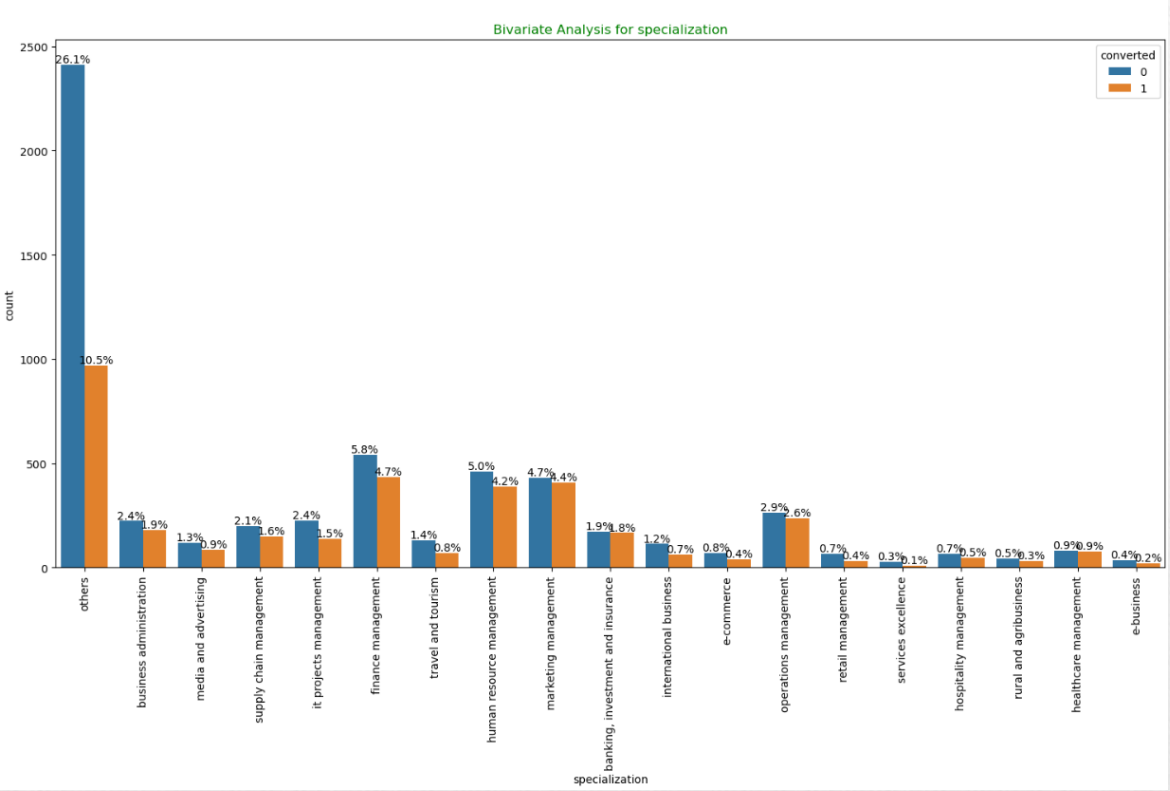
3.*Last Activity*:-

'SMS Sent' holds a notable lead conversion rate of 63%, contributing to 30% of the last activities. 'Email Opened' represents 38% of the last activities performed by customers, accompanied by a 37% lead conversion rate.



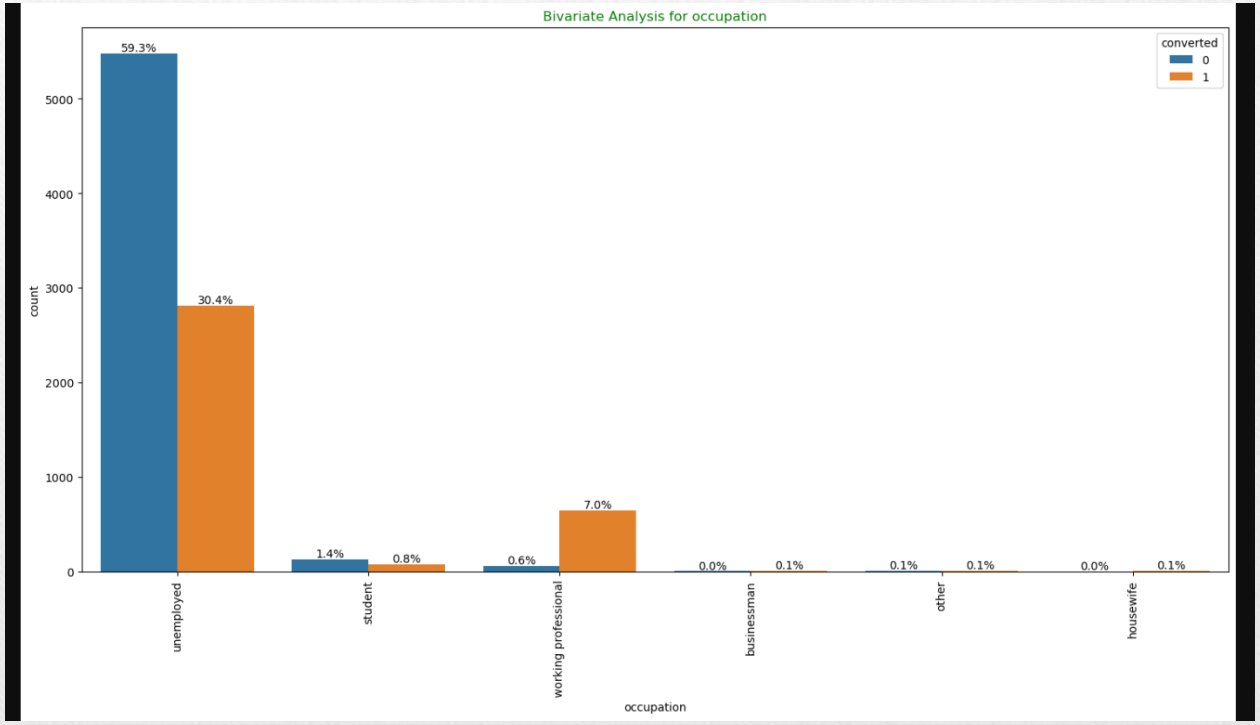
4.*Specialization*:-

Marketing Management, HR Management, and Finance Management demonstrate substantial contributions to lead conversion rates.



5. *Occupation distribution:-*

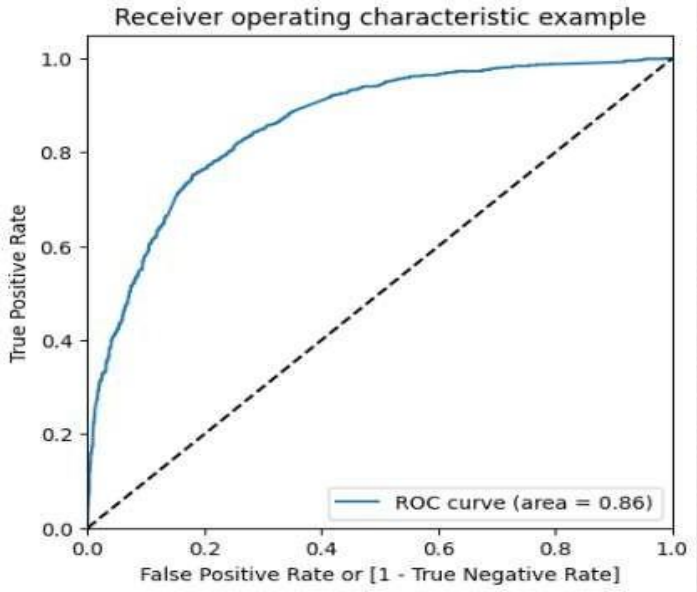
A significant 90% of customers are unemployed, yet they exhibit a commendable lead conversion rate (LCR) of 34%. On the contrary, Working Professionals constitute only 7.6% of total customers but display a remarkably high LCR of almost 92%.

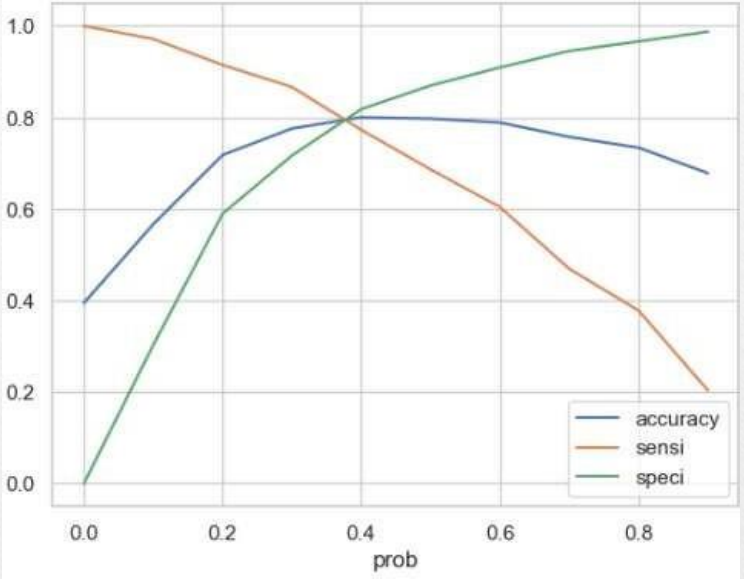


MODEL BUILDING:-

- Begin by dividing the data into training and testing sets, adopting a 70:30 ratio for regression analysis.
- Employ Recursive Feature Elimination (RFE) to select features, specifying an output of 15 variables.
- Construct the model by excluding variables with a p-value exceeding 0.05 and a VIF value surpassing 5.
- Generate predictions using the test dataset.
- Attain a comprehensive accuracy rate 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 close alignment of these metrics (accuracy, sensitivity and specificity) between the train and test sets indicates the effectiveness of the model. Furthermore, the achieved sensitivity of approximately 80% aligns with the CEO's target, and the overall accuracy of 80.88% aligns with the study's objectives as well.

THANK YOU