Lead Scoring Case Study

Submitted by

Chetan kabburi

Rahul Shetty

Rahul Singh

Business Objective

- Assist X Education in identifying high-potential leads that have the greatest likelihood of converting into paying customers.
- Develop a logistic regression model that assigns a lead score ranging from 0 to 100, helping the company efficiently target prospective leads.

Problem Statement

Company Overview:

- **X Education** is an e-learning company offering online courses for industry professionals.

Context:

- The company promotes its courses via online platforms such as **search engines and websites**.
- Visitors explore different courses, watch videos, or complete inquiry forms on the website.
- Those who submit their details via forms are classified as **leads**.

Challenges:

- The company receives a **large volume of leads**, but only **38% convert** into customers.
- To improve efficiency, **X Education** seeks to develop a system that assigns a **lead score** to determine the likelihood of conversion.
- A **higher lead score** indicates a greater probability of conversion, while a **lower lead score** suggests a weaker potential.

Data Cleaning

Handling Missing Data:

- Removed columns with more than **40% missing values**, including:
  - **How did you hear about X Education?** (78.46% missing)
  - **Lead Profile** (74.18% missing)
  - **Lead Quality** (51.60% missing)
  - **Asymmetrique Profile Score, Activity Score, Activity Index, Profile Index** (45% missing)

Eliminating Redundant Columns:

- Dropped unnecessary variables:
  - **Updates**
  - **Last Activity**
  - **Prospect ID**

Filling Missing Values:

Numerical data filled using **median values**.

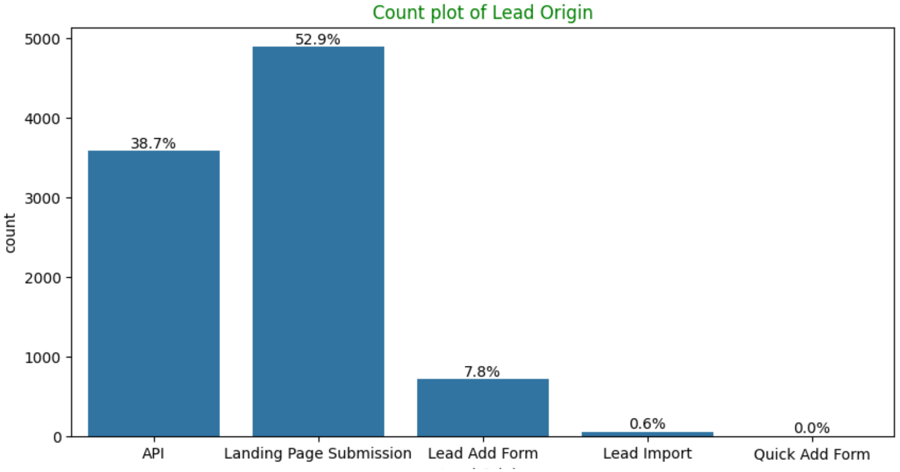
Categorical data filled using **mode values**.

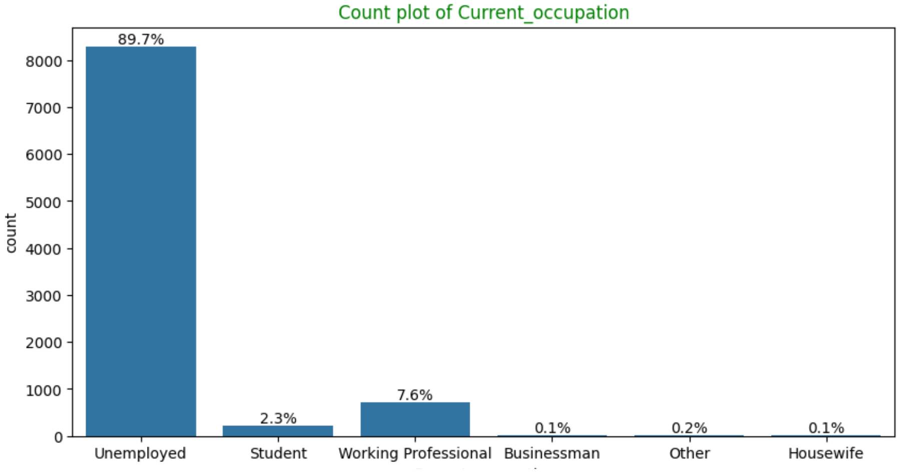
Exploratory Data Analysis (EDA)

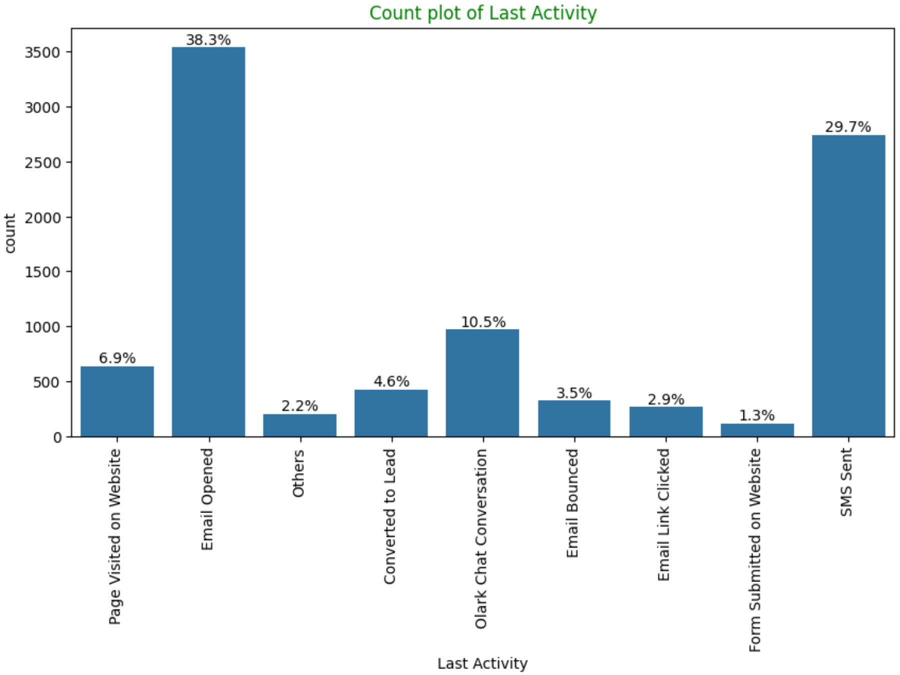
Insights from Univariate Analysis:

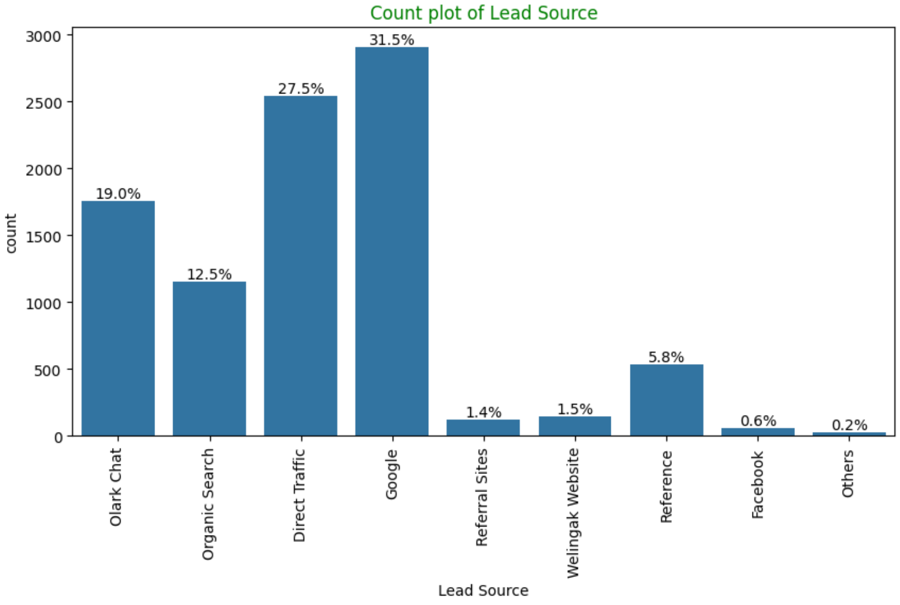
- Leads from **Google and Direct Traffic** exhibit **higher conversion rates**.
- Prospective students specializing in **Finance Management** have the highest conversion probability.
- **Working professionals** show **greater likelihood** of converting.
- Most leads originate from **Mumbai**, with a conversion rate exceeding **50%**.

Univariate Analysis


Count plot of Lead Origin


Count plot of Current_occupation


Count plot of Last Activity

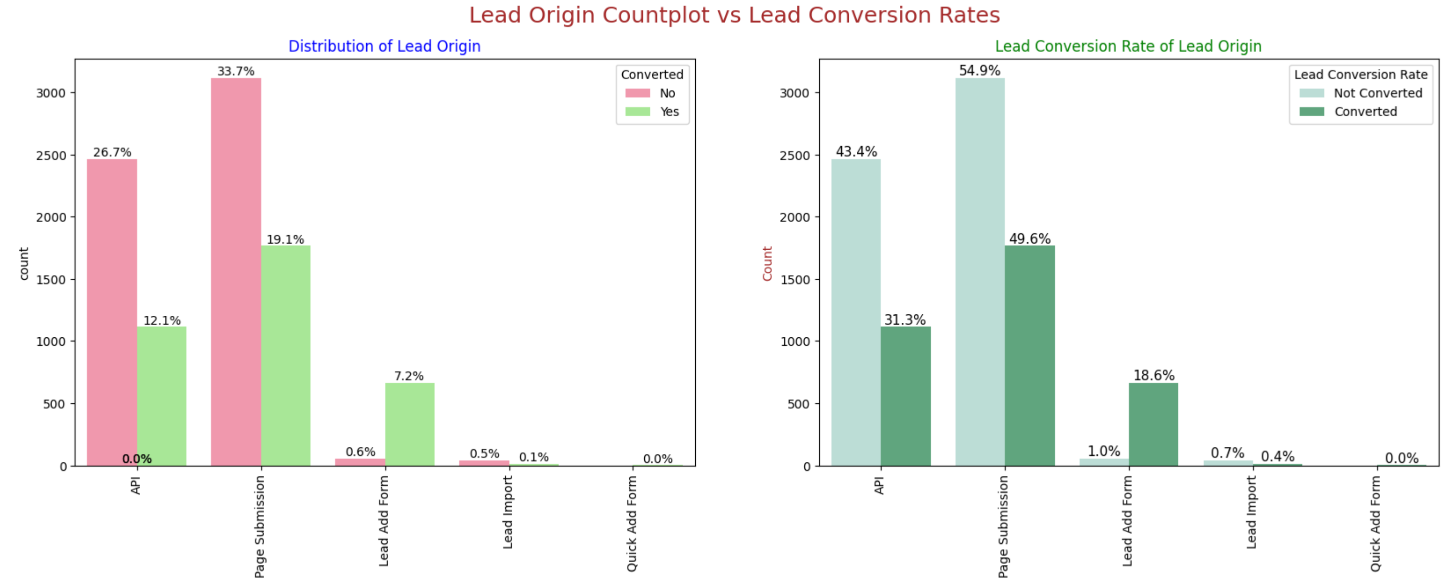

Count plot of Lead Source

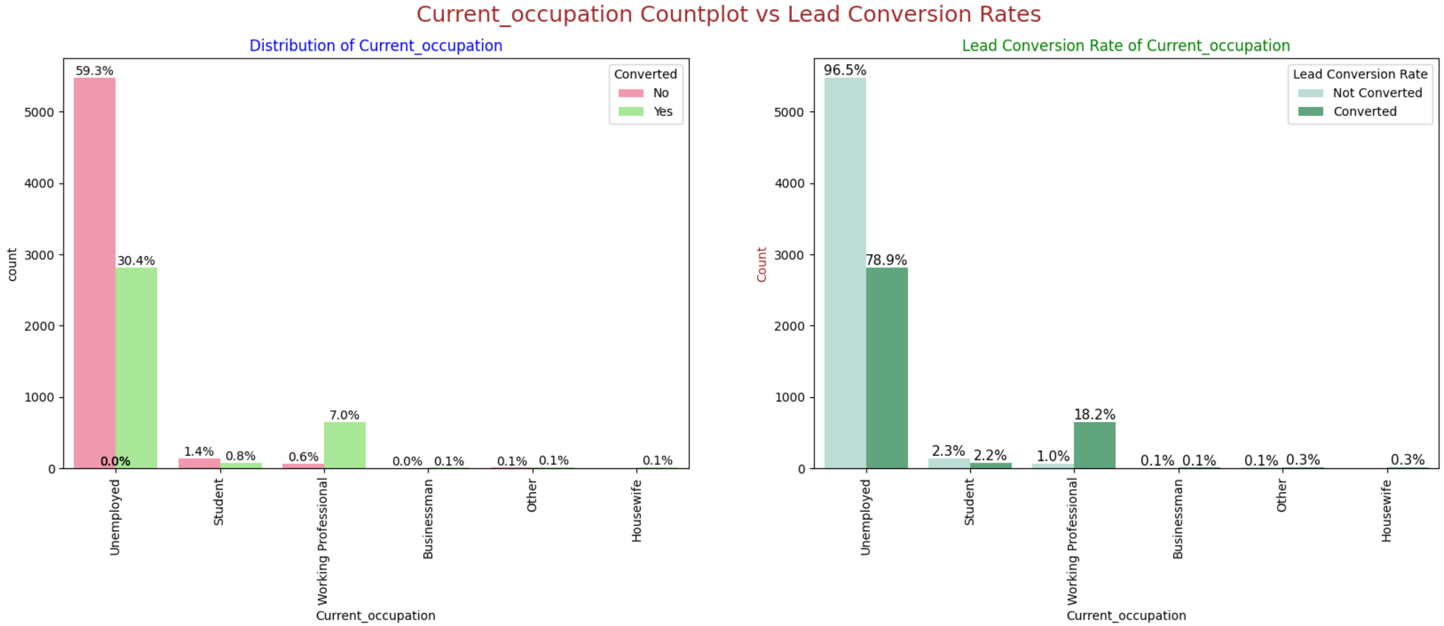
Univariate Analysis - Key Observations

Here is the list of features from variables which are present in majority (Converted and Not Converted included)

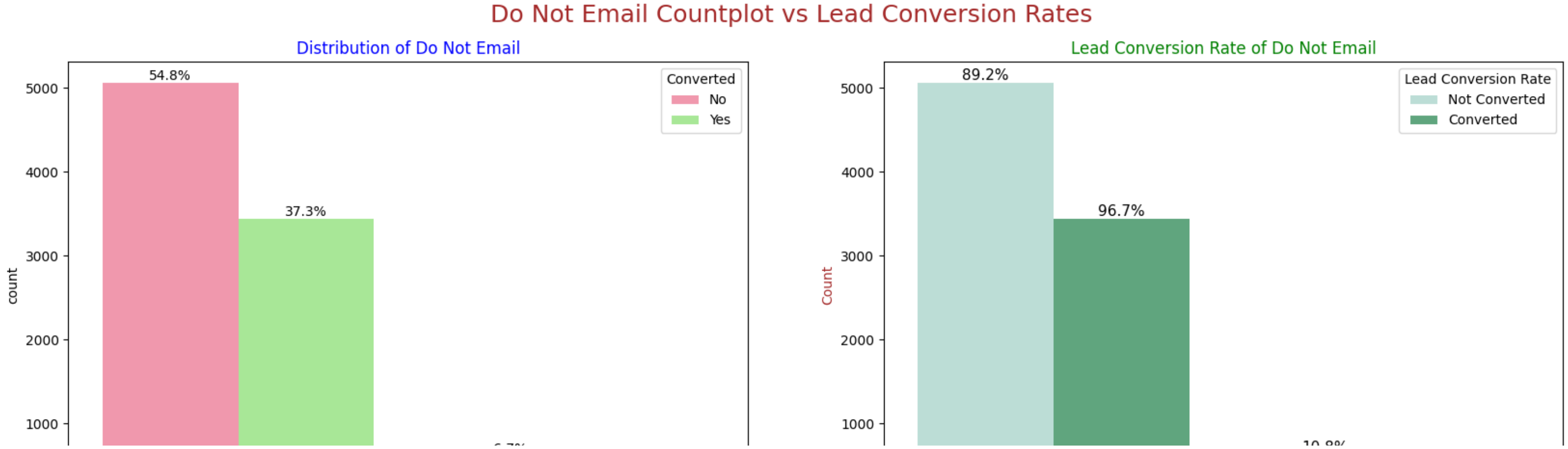
- Lead Origin: "Landing Page Submission" identified 53% customers, "API" identified 39%.
- Current_occupation: It has 90% of the customers as Unemployed
- Do Not Email: 92% of the people has opted that they dont want to be emailed about the course.
- Lead Source: 59% Lead source is from Google & Direct Traffic combined
- Last Activity: 68% of customers contribution in SMS Sent & Email Opened activities

Bi-variate Analysis (1/4)





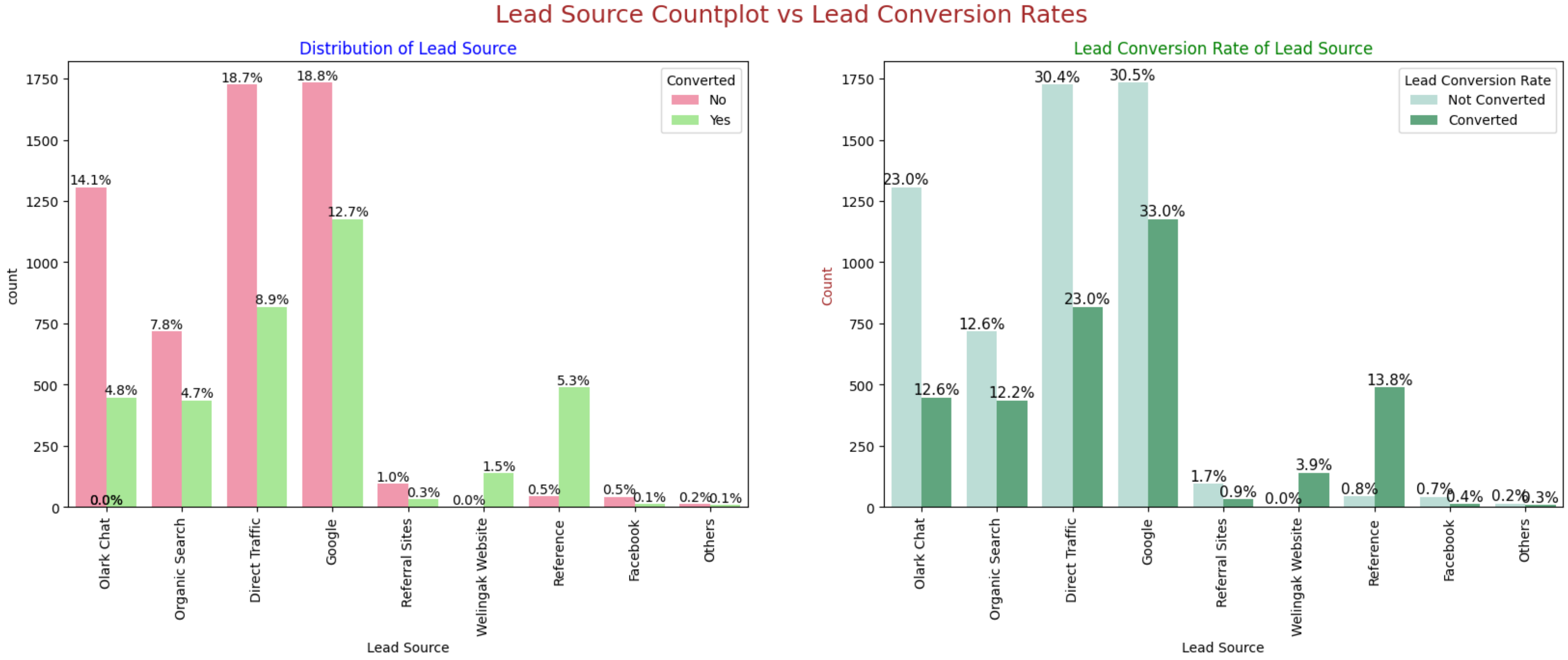
Bi-variate Analysis (2/4)



Do Not Email Countplot vs Lead Conversion Rates

Distribution of Do Not Email

Lead Conversion Rate of Do Not Email



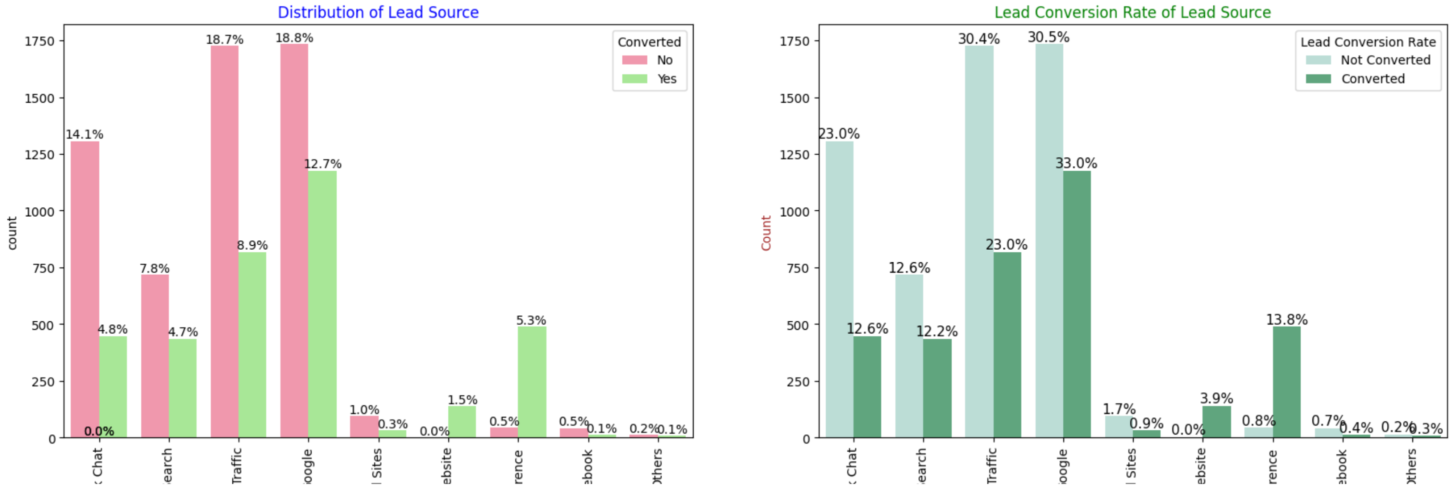
Lead Source Countplot vs Lead Conversion Rates

Distribution of Lead Source

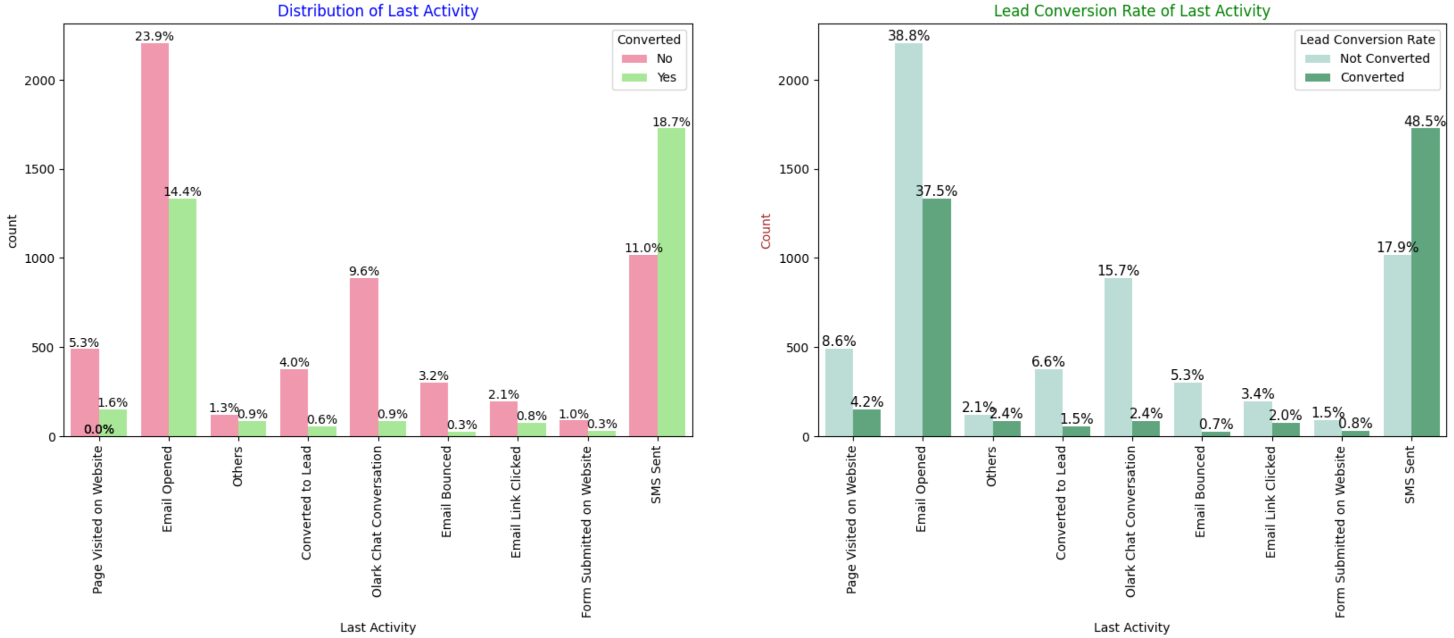
Lead Conversion Rate of Lead Source

Bi-variate Analysis (3/4)

Lead Source Countplot vs Lead Conversion Rates



Last Activity Countplot vs Lead Conversion Rates



Bi-variate Analysis (4/4)

Correlation Analysis:

- **Total Visits and Pages Viewed per Visit** exhibit a **strong positive correlation**.
- **Time Spent on Website** correlates **positively** with:
  - **Lead Conversion**
  - **Page Views**
  - **Total Visits**
- **Page Views and Lead Conversion** show a **negative correlation**.

Bi-variate Analysis - Key Takeaways

Numerical Variables Relationship:

- A **linear correlation** exists between **Total Visits and Pages Viewed per Visit**.

Categorical & Numerical Relationships:

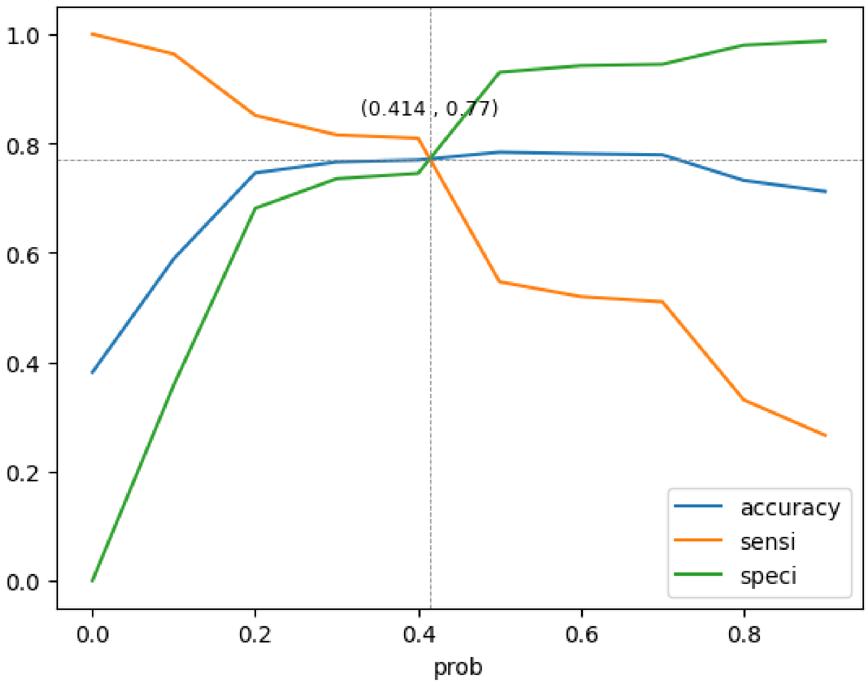
- Leads sourced from **Organic Search** have **higher page views per visit** and **better conversion rates**.
- Those **spending more time** on the website exhibit **higher conversion probabilities**.
- Visitors with **multiple site visits** are **more likely to convert**.
- Individuals enrolling for **career growth** display a **higher likelihood of conversion**.

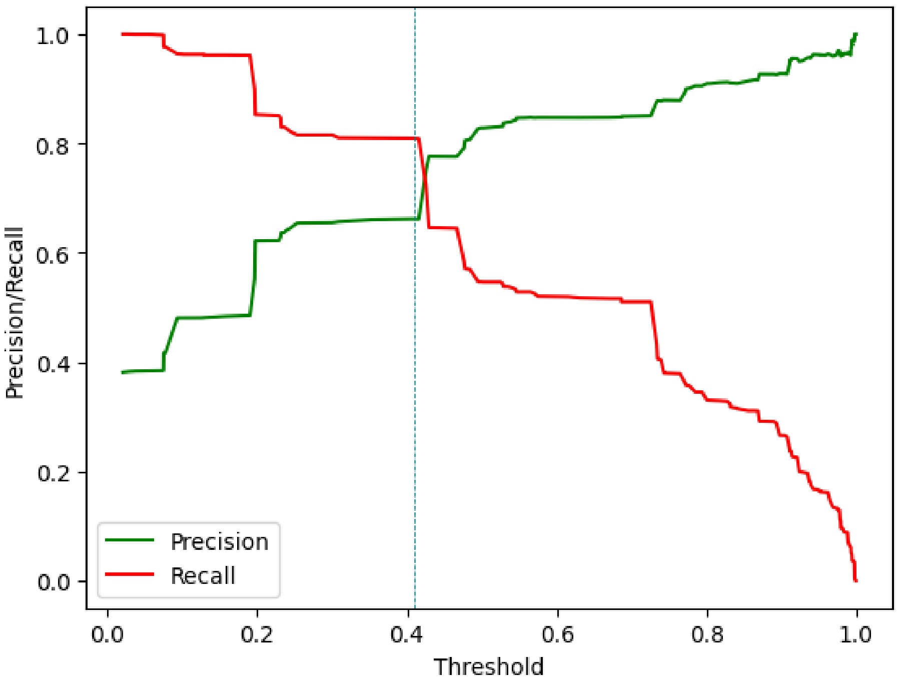
Model Building

Model Overview

- A **logistic regression model** was developed to predict the likelihood of **lead conversion**.
- The model assigns **lead scores**, allowing the sales team to **prioritize high-potential leads**.
- The final model includes **12 features**, with the **top three influential variables** being:
  - **Lead Source – Welingak Website**
  - **Lead Source – Reference**
  - **Current Occupation – Working Professional**

Model Evaluation





The intersection point of the curve is the threshold value where the model achieves a balance between precision and recall. It can be used to optimize the performance of the model based on business requirement, Here our probability threshold is 0.41 approx. from above curve.

Model Evaluation

Training Data

- Optimal Cut-off Probability: 0.345
- Accuracy: 80.51%
- Sensitivity (Recall): 65.69%
- Specificity: 89.65%
- Positive Predictive Value (Precision): 79.64%
- Negative Predictive Value: 80.92%

Model Performance - Test Data

- Accuracy: 80.34%
- Sensitivity (Recall): 79.82%
- Specificity: 80.68%
- Precision: 72.95%
- True Positive Rate (TPR): 79.82%
- False Positive Rate (FPR): 19.32%
- The evaluation metrics remain consistent across training and test datasets, confirming the model's reliability and effectiveness

Confusion Matrix (Test Data)

True Negatives (TN): 1,353	False Positives (FP): 324
False Negatives (FN): 221	True Positives (TP): 874

Key Recommendations & Conclusion

- **Prioritize High-Scoring Leads:**
  - Leads with **higher lead scores** should be given more attention for **enhanced conversion rates**.
- **Strengthen Google Marketing Efforts:**
  - Since **Google-driven traffic** has a **strong conversion performance**, additional marketing efforts should be made here.
- **Encourage Referrals:**
  - Offer **incentives for existing customers** to refer new prospects.
- **Expand Geographic Reach:**
  - Since most leads come from **Mumbai**, **marketing strategies** should be expanded to **other major cities**.
- **Target Unemployed Individuals & Finance Professionals:**
  - Since unemployed individuals and those with a **Finance Management specialization** have **higher conversion rates**, targeted engagement is recommended.
- **Reduce Focus on Students:**
  - Conversion rates among **students** are **significantly lower**, so sales efforts in this segment should be minimized.