LEAD SCORING - CASE STUDY

AMIT JADHAV

TABLE OF CONTENT

- Background Of X Education Company
- Problem Statement & Objective Of The Study
- Suggested Ideas For Lead Conversion
- Analysis Approach
- Data Cleaning
- Eda
- Data Preparation
- Model Building (RFE & Manual Fine Tuning)
- Model Evaluation
- Recommendations

BACKGROUND OF X EDUCATION COMPANY

- **Targeted Audience:** X Education Attracts Industry Professionals Interested In Enhancing Their Skills Through Online Courses.
- **Multi-channel Marketing:** The Company Leverages Various Marketing Channels, Including Partner Websites And Search Engine Optimization (SEO) On Platforms Like Google, To Reach Its Target Audience.
- **Website Engagement:** Website Visitors Can Explore Course Offerings, Request Additional Information Through Forms, Or Engage With Educational Content (E.G., Videos).
- **Lead Capture:** Filling Out A Form With Contact Details (Email Or Phone Number) Qualifies A Website Visitor As A Lead.
- **Lead Nurturing:** Upon Acquiring Leads, The Sales Team Initiates Outreach Through Personalized Calls And Email Communication.
- **Conversion Funnel Analysis:** X Education's Current Lead Conversion Rate Sits At 30%. Optimizing This Funnel Presents An Opportunity To Improve Sales Efficiency.

PROBLEM STATEMENT & OBJECTIVE OF THE STUDY

Problem Statement

- **X Education Seeks To Improve Lead Conversion Efficiency:** While The Company Generates A Substantial Volume Of Leads, The Current Conversion Rate Of 30% Presents An Opportunity For Optimization.
- **Identifying High-value Leads:** X Education Aims To Implement A System For Prioritizing Leads Based On Their Conversion Potential. This Approach, Known As Lead Scoring, Will Allow The Sales Team To Focus Their Efforts On The Most Qualified Leads (Hot Leads).
- **Targeted Sales Outreach:** By Prioritizing Hot Leads, X Education's Sales Team Can Allocate Their Resources More Effectively, Focusing Communication Efforts On The Leads With The Highest Likelihood Of Conversion.

Objective Of The Study

- **Lead Scoring Model Development:** Our Objective Is To Assist X Education In Implementing A Lead Scoring Model. This Model Will Assign A Numerical Value (Score) To Each Lead, Reflecting Their Likelihood Of Conversion Into Paying Customers. Leads With Higher Scores Will Be Identified As Having A Greater Conversion Potential.
- **Target Conversion Rate:** In Alignment With X Education's Goals, The Lead Scoring Model Will Be Designed To Target A Lead Conversion Rate Of Approximately 80%.

SUGGESTED IDEAS FOR LEAD CONVERSION

LEADS GROUPING

- The Lead Scoring Model Will Categorize Leads Based On Their Conversion Probability. This Allows For The Identification Of A Targeted Segment With A High Likelihood Of Conversion, Often Referred To As "Hot Leads."

BETTER COMMUNICATION

- Implementing A Lead Scoring Model Will Allow Us To Prioritize Outreach Efforts. By Focusing On High-potential Leads, We Can Optimize The Utilization Of Our Communication Resources, Leading To A More Impactful Sales Strategy.

BOOST CONVERSION

- By Prioritizing Leads With A Higher Conversion Potential (Hot Leads) Identified Through The Lead Scoring Model, We Can Expect A Significant Increase In Its Conversion Rate. This Targeted Approach Offers A Realistic Path Towards Achieving The Ambitious 80% Conversion Rate Objective.

Given The Target Conversion Rate Of 80%, Achieving A High Level Of Sensitivity In Our Lead Scoring Model Is Crucial.

ANALYSIS APPROACH

Data Cleaning:

Loading Data Set, Understanding & Cleaning Data

EDA:

Check Imbalance, Univariate & Bivariate Analysis

Data Preparation:

Dummy Variables, Test-train Split, Feature Scaling

Model Building:

RFE For Top 15 Feature, Manual Feature Reduction & Finalizing Model

Model Evaluation:

Confusion Matrix, Cutoff Selection, Assigning Lead Score

Predictions On Test Data:

Compare Train Vs Test Metrics, Assign Lead Score And Get Top Features

Recommendation:

Suggest Top 3 Features To Focus For Higher Conversion & Areas For Improvement

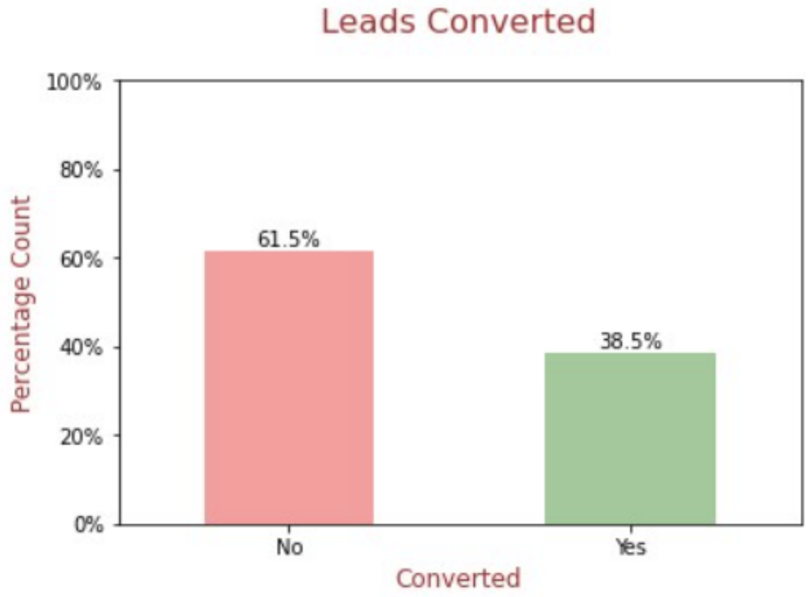
DATA CLEANING

- A Special Category, "Select," Was Created To Represent Null Values Where Customers Did Not Choose An Option From A List.
- Columns With A High Proportion Of Missing Values (Over 40%) Were Removed As They Offered Limited Information.
- For Remaining Categorical Variables, Missing Values Were Imputed Based On Value Frequency And Other Relevant Factors.
- In Some Cases, Additional Categories Were Created To Improve Data Representation.
- Columns Containing No Relevant Information For The Analysis Objectives (E.G., Tags, Country) Were Dropped.
- Columns With Only One Unique Response Category Were Excluded As They Provided No Differentiation For Modeling.
- Missing Values Were Imputed Using The Mode (Most Frequent Value) After Examining The Data Distribution To Ensure Suitability.

DATA CLEANING

- To Mitigate Potential Bias In The Logistic Regression Model, Categorical Features With Significant Skew Were Identified And Removed.
- Low-frequency Categories Within Categorical Features Were Grouped Into An "Other" Category To Improve Modelgeneralizability.
- Binary Categorical Variables Were Efficiently Encoded For Modeling Purposes.
- Potential Outliers In "Totalvisits" And "Page Views Per Visit" Were Addressed Through Capping To Reduce Their Undue Influence On The Model.
- Data Quality Was Ensured Through Various Techniques:
    - ❑ Invalid Values Were Corrected (E.G., Standardizing Lead Source Entries Like "Google" And "Google").
    - ❑ Data Inconsistencies Were Addressed, Such As Checking For And Fixing Casing Inconsistencies (E.G., Uppercase Vs Lowercase).
- Additional Data Cleaning Activities Were Conducted To Guarantee The Accuracy And Integrity Of The Data For Analysis.

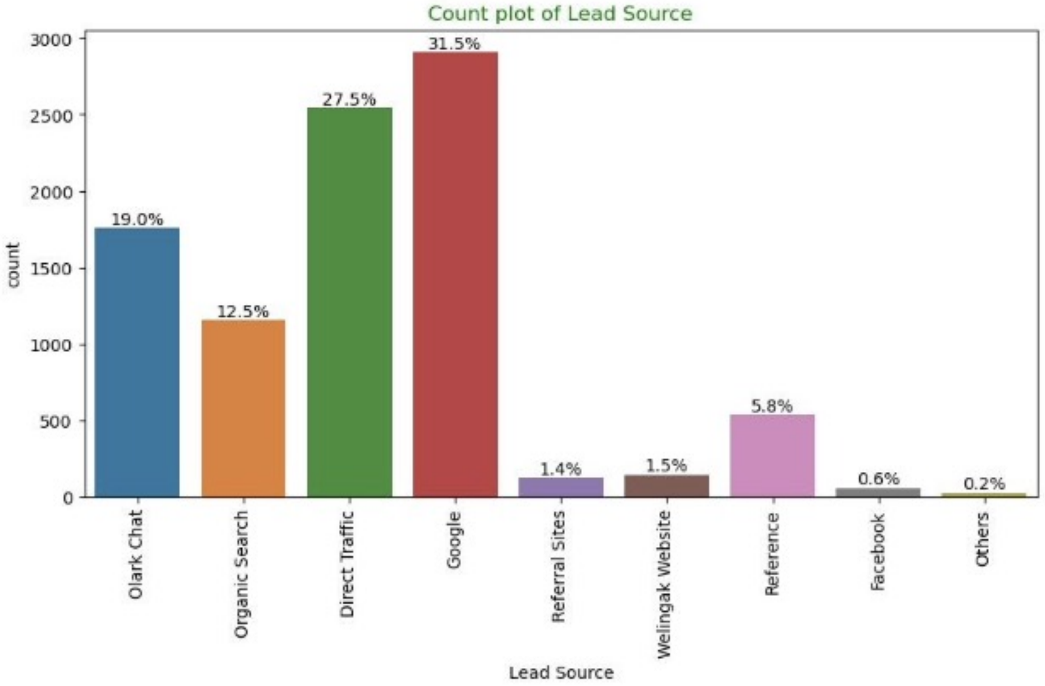
EDA

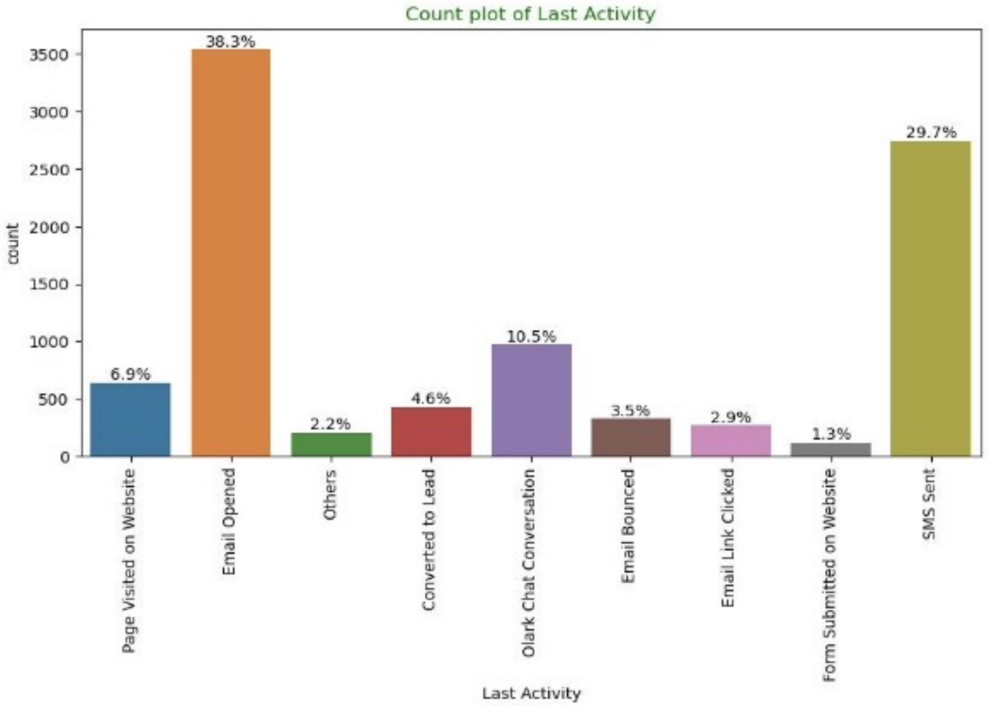


- ❖ Conversion Rate Is Of 38.5%, Meaning Only 38.5% Of The People Have Converted To Leads.(Minority)

- ❖ While 61.5% Of The People Didn't Convert To Leads. (Majority)

EDA
UNIVARIATE ANALYSIS – CATEGORICAL VARIABLES

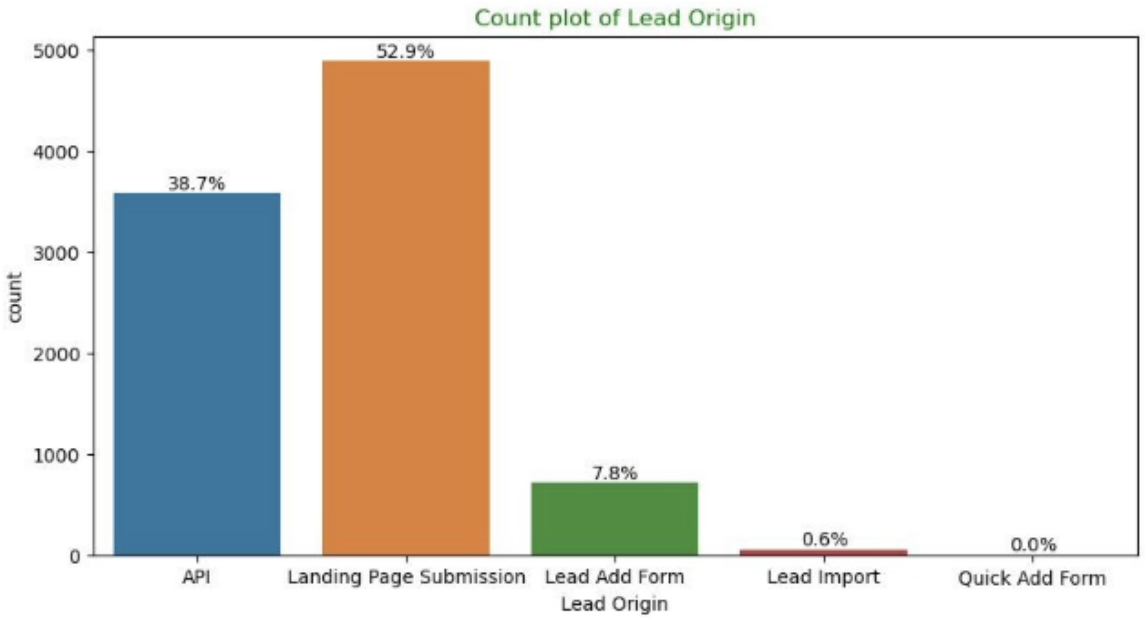


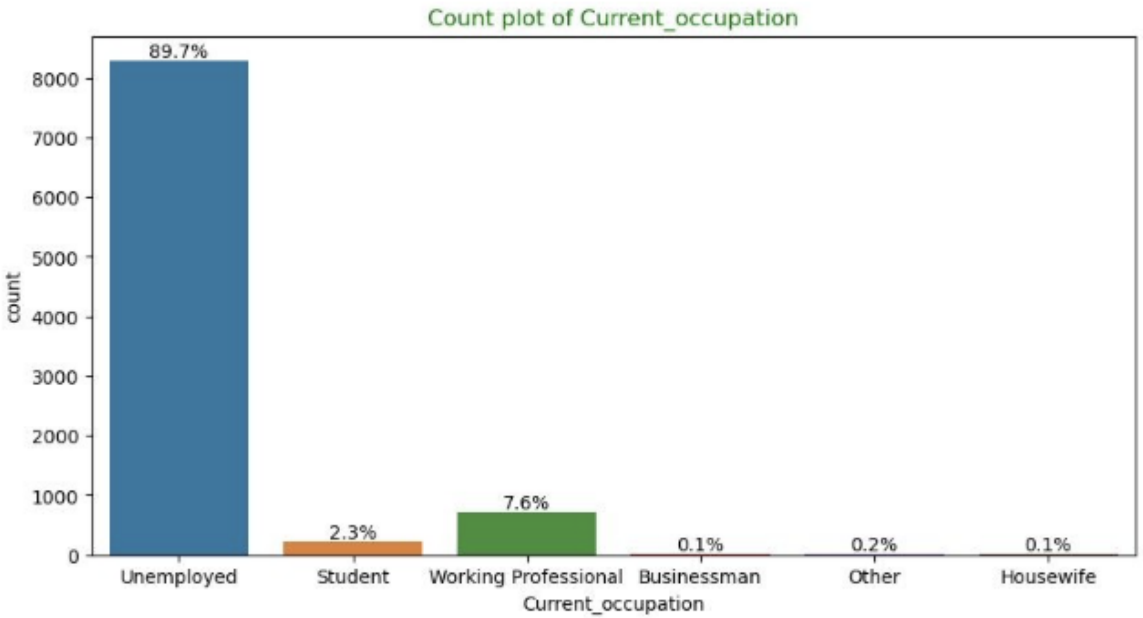


Lead Source: 58% Lead Source Is From Google & Direct Traffic Combined.

Last Activity: 68%of Customers Contribution In & SMS Sent & Email Opened Activities.

EDA
UNIVARIATE ANALYSIS – CATEGORICAL VARIABLES

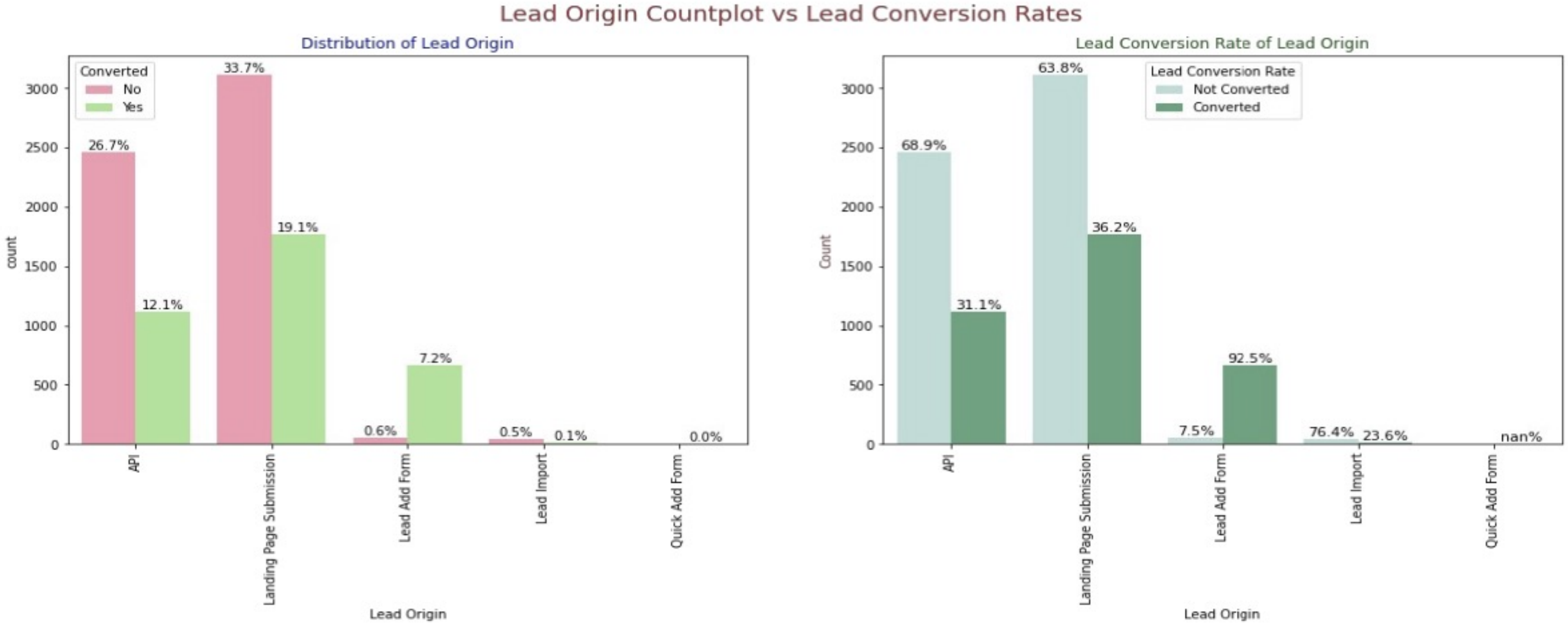




Lead Origin: "Landing Page Submission" Identified 53% Of Customers, "API" Identified 39%.

Current_occupation: It Has 90% Of The Customers Unemployed.

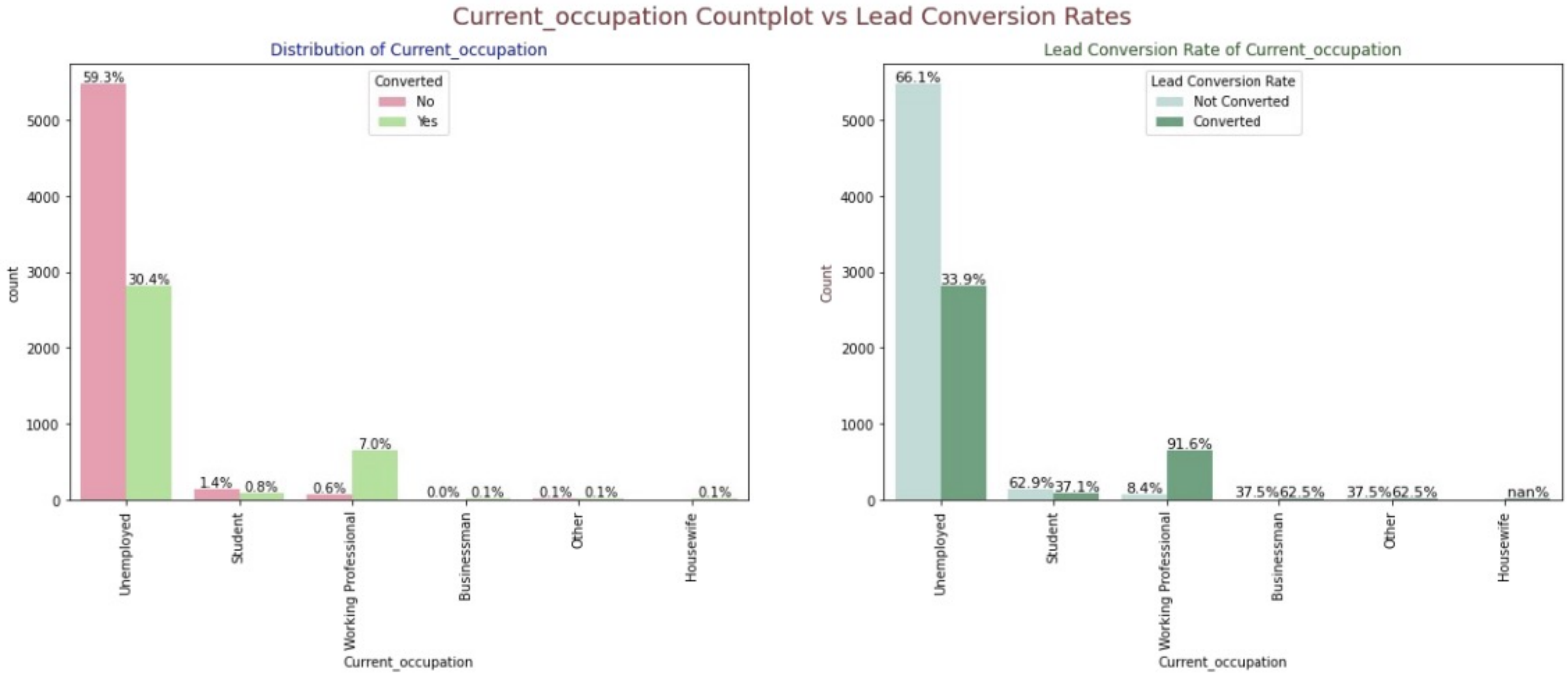
EDA
BIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES



Lead Origin:

- Around 52% Of All Leads Originated From *"Landing Page Submission"* With A **Lead Conversion Rate (LCR) Of 36%**. The *"API"* Identified Approximately 39% Of Customers With A **Lead Conversion Rate (LCR) Of 31%**.

EDA
BIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES



Current_occupation Countplot vs Lead Conversion Rates

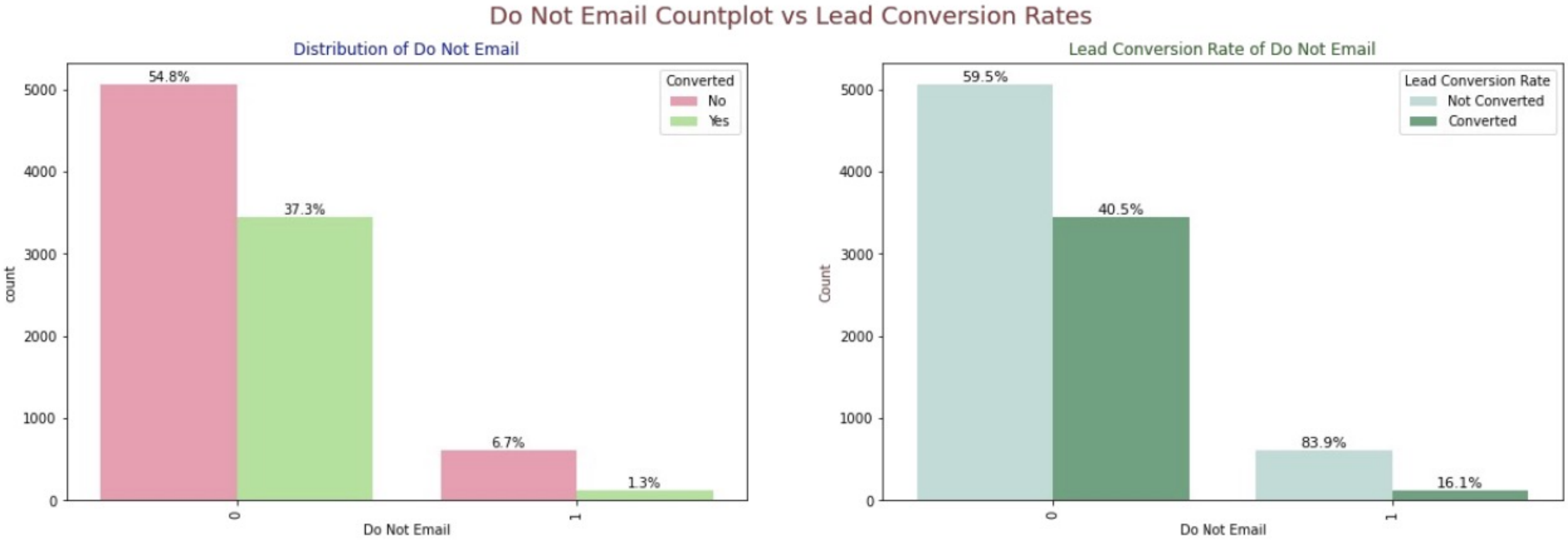
Distribution of Current_occupation

Lead Conversion Rate of Current_occupation

Current_occupation:

- Around 90% Of The Customers Are *Unemployed,* With **Lead Conversion Rate (LCR) Of 34%**. While *Working Professional* Contribute Only 7.6% Of Total Customers With Almost **92% Lead Conversion Rate (LCR)**.

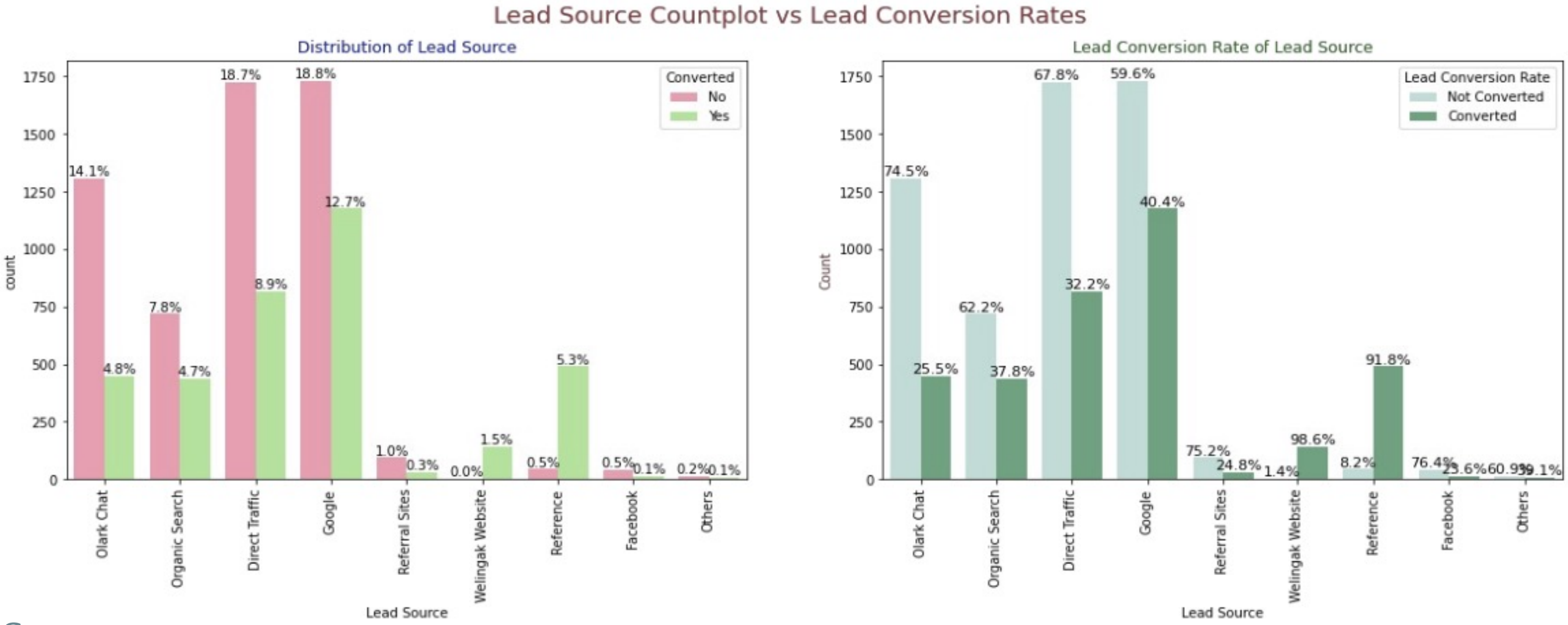
EDA
BIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES



Do Not Email:

- 92% Of The People Has Opted That They Don’t Want To Be Emailed About The Course & 40% Of Them Are Converted To Leads.

EDA
BIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES

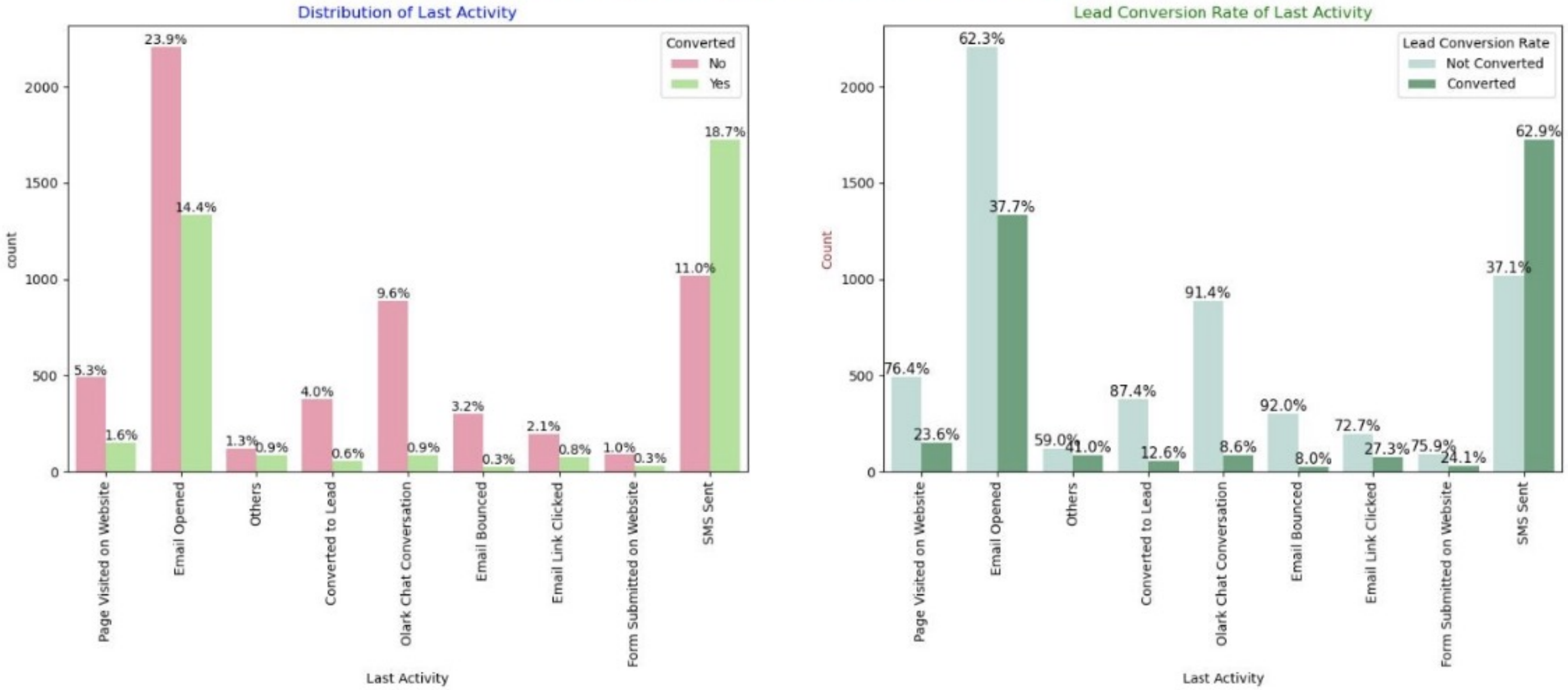


Lead Source:

- Google Has **LCR Of 40%** Out Of 31% Customers.
- Direct Traffic Contributes **32% LCR** With 27% Customers, Which Is Lower Than Google,
- Organic Search Also Gives **37.8% Of LCR**, But The Contribution Is By Only 12.5% Of Customers,
- Reference Has **LCR Of 91%**, But There Are Only Around 6% Of Customers Through This Lead Source.

EDA
BIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES

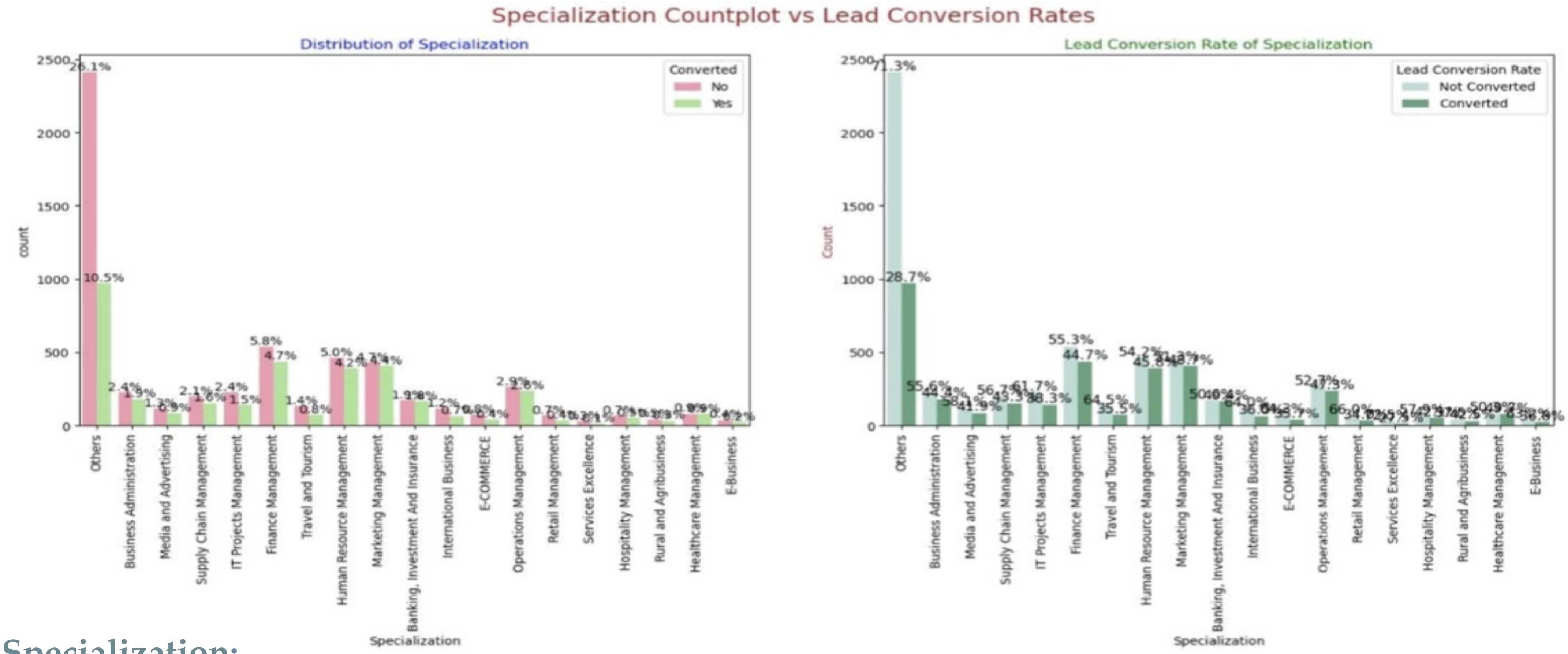
Last Activity Countplot vs Lead Conversion Rates



Last Activity:

- 'SMS Sent' Has **High Lead Conversion Rate Of 63%** With 30% Contribution From Last Activities,
- 'Email Opened' Activity Contributed 38% Of Last Activities Performed By The Customers, With **37% Lead Conversion Rate.**

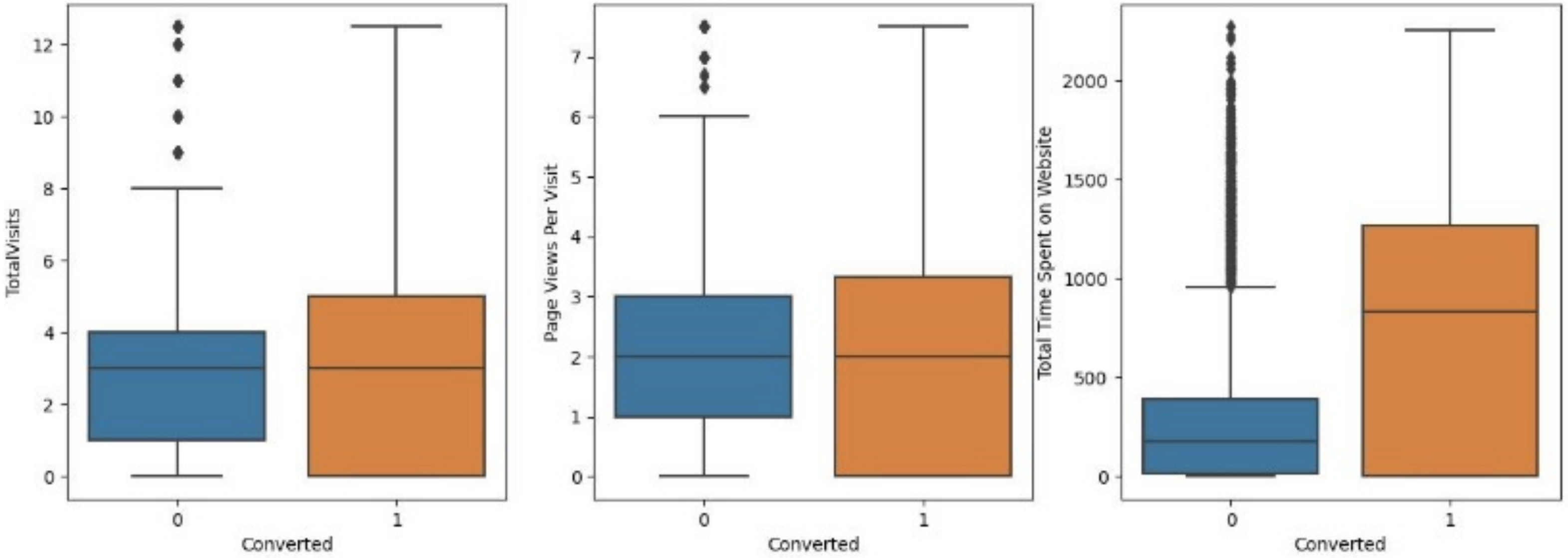
EDA
BIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES

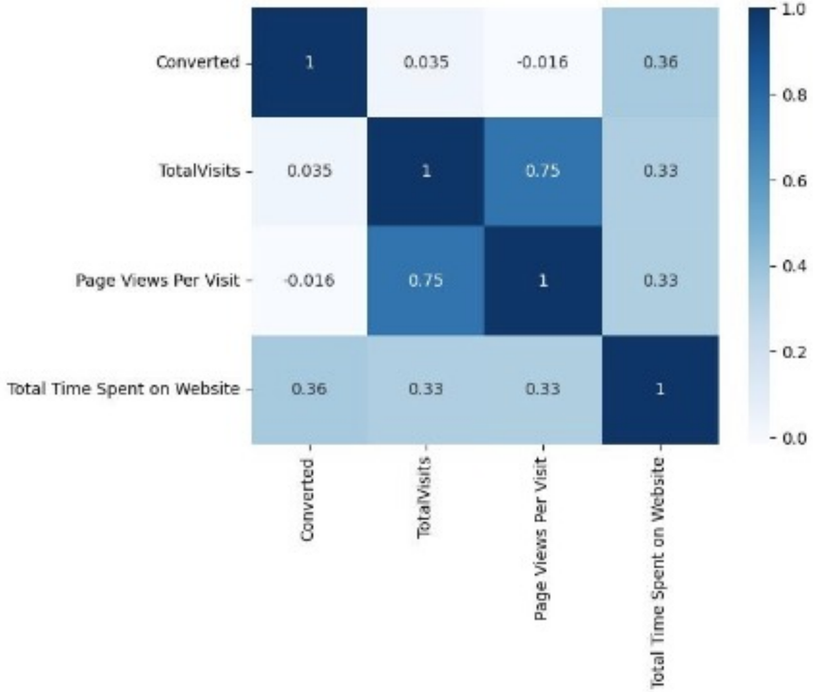


Specialization:

- Marketing Management, HR Management, Finance Management Shows Good Contribution In Leads Conversion Than Other Specialization.

EDA
BIVARIATE ANALYSIS FOR NUMERICAL VARIABLES





Past Leads Who **Spends More Time On The Website** Have A Higher Chance Of Getting Successfully Converted Than Those Who Spends Less Time As Seen In The **Box-plot.**

DATA PREPARATION BEFORE MODEL BUILDING

- Building Upon The Prior Binary Encoding, Dummy Variables (One-hot Encoding) Were Created For The Following Categorical Features: Lead Origin, Lead Source, Last Activity, Specialization, And Current_occupation. This Approach Allows The Model To Capture The Relationships Between These Features And The Target Variable More Effectively.
- The Data Was Divided Into Training And Testing Sets Using A 70:30 Ratio. The Training Set Will Be Used To Build The Logistic Regression Model, And The Testing Set Will Be Used To Evaluate Its Performance On Unseen Data.
- Standardization Was Employed To Scale The Features. This Ensures All Features Contribute Equally To The Model And Prevents Features With Larger Scales From Dominating The Model.
- To Address Potential Multicollinearity, Correlations Between Predictor Variables Were Analyzed. Highly Correlated Features (E.G., Lead Origin_lead Import And Lead Origin_lead Add Form) Were Removed To Avoid Redundancy And Improve Model Stability.

MODEL BUILDING

- The Initial Dataset Possessed A High Dimensionality With A Large Number Of Features. This Can Negatively Impact Model Performance By Increasing Training Time And Potentially Leading To Overfitting.
- To Address This Challenge, Recursive Feature Elimination (RFE) Was Employed To Select A More Parsimonious Feature Set. RFE Iteratively Removes The Least Important Features Until A Desired Number Of Features Is Reached. In This Case, RFE Successfully Reduced The Feature Space From 48 Dimensions To 15, Resulting In A More Efficient And Potentially More Accurate Model.
- Following The Dimensionality Reduction, Manual Fine-tuning Of The Model Can Be Performed For Further Optimization.

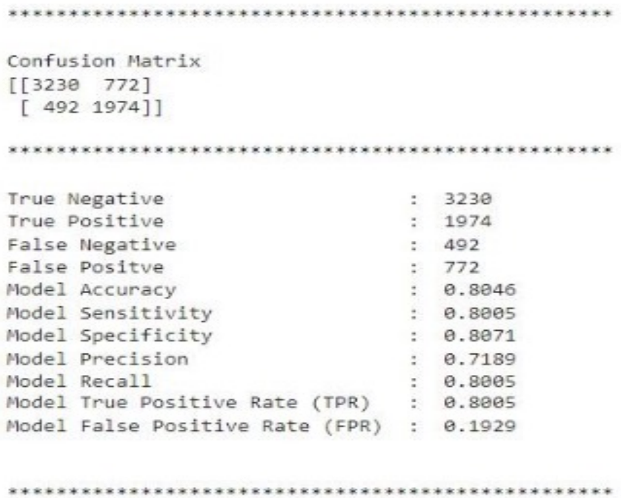
MODEL BUILDING

- A Manual Feature Reduction Technique Was Implemented To Refine The Model. This Involved Removing Variables With P-values Greater Than 0.05. This Threshold Indicates A Lack Of Statistical Significance For The Variable's Association With The Target Variable And Suggests Its Potential Exclusion May Not Significantly Impact The Model's Performance.
- After Four Iterations Of This Process, Model 4 Emerged As The Most Promising Candidate. This Selection Is Based On Two Key Criteria:
  - ❑ Statistically Significant Variables: Model 4 Incorporates Variables With P-values Less Than 0.05, Indicating A Robust Relationship With The Target Variable.
  - ❑ Low Multicollinearity: Vifs (Variance Inflation Factors) Remain Below 5, Suggesting No Significant Multicollinearity Is Present. Multicollinearity Can Negatively Impact Model Interpretability And Stability.
- Consequently, Model 4 Will Be Chosen For Further Evaluation And Prediction. This Model Offers A Balance Between Retaining Informative Features And Mitigating Multicollinearity.

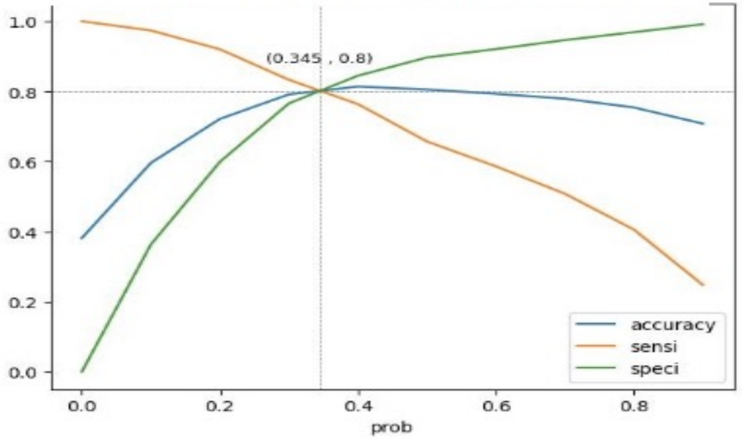
MODEL EVALUATION
TRAIN DATA SET

It Was Decided To Go Ahead With 0.345 As Cut-off After Checking Evaluation Metrics Coming From Both Plots

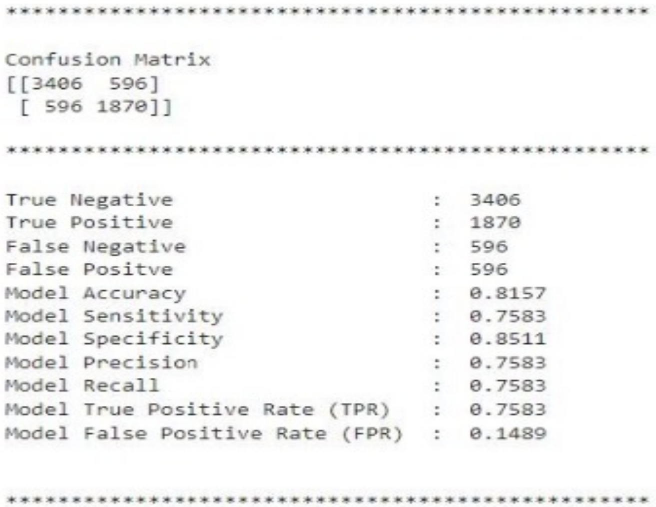
Confusion Matrix & Evaluation Metrics With 0.345 As Cut-off



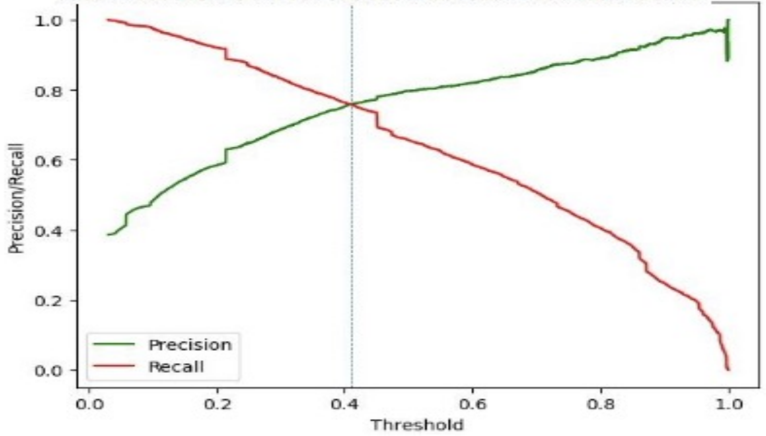
```
**************************************************

Confusion Matrix
[[3230  772]
 [ 492 1974]]

**************************************************

True Negative                      :  3230
True Positive                      :  1974
False Negative                     :  492
False Positve                      :  772
Model Accuracy                     :  0.8046
Model Sensitivity                  :  0.8005
Model Specificity                  :  0.8071
Model Precision                    :  0.7189
Model Recall                       :  0.8005
Model True Positive Rate (TPR)     :  0.8005
Model False Positive Rate (FPR)    :  0.1929

**************************************************
```



Confusion Matrix & Evaluation Metrics with 0.41 as Cut-off



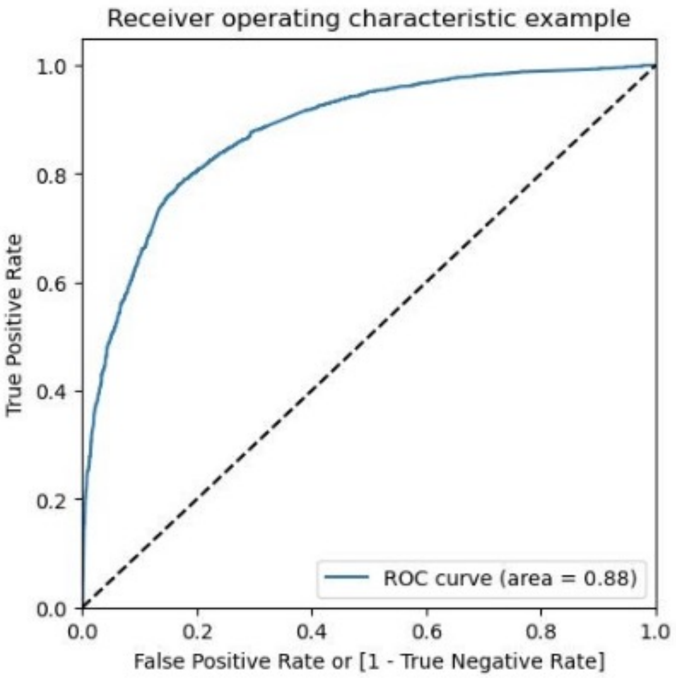
```
**************************************************

Confusion Matrix
[[3406  596]
 [ 596 1870]]

**************************************************

True Negative                      :  3406
True Positive                      :  1870
False Negative                     :  596
False Positve                      :  596
Model Accuracy                     :  0.8157
Model Sensitivity                  :  0.7583
Model Specificity                  :  0.8511
Model Precision                    :  0.7583
Model Recall                       :  0.7583
Model True Positive Rate (TPR)     :  0.7583
Model False Positive Rate (FPR)    :  0.1489

**************************************************
```



MODEL EVALUATION

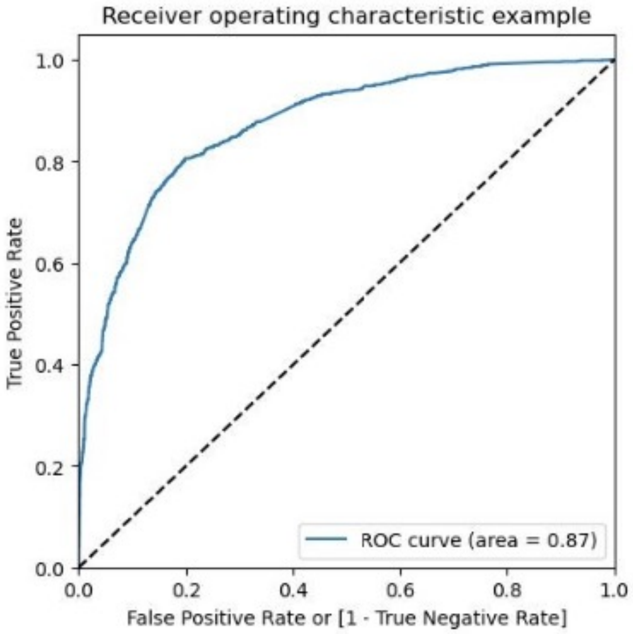
ROC CURVE – TRAIN DATA SET

The Model Achieved An Area Under The ROC Curve (AUC) Of 0.88. This Value Indicates A Good Level Of Discrimination Between Positive And Negative Cases. In Simpler Terms, The Model Can Effectively Distinguish Between Those Who Will Experience The Event Of Interest And Those Who Will Not.



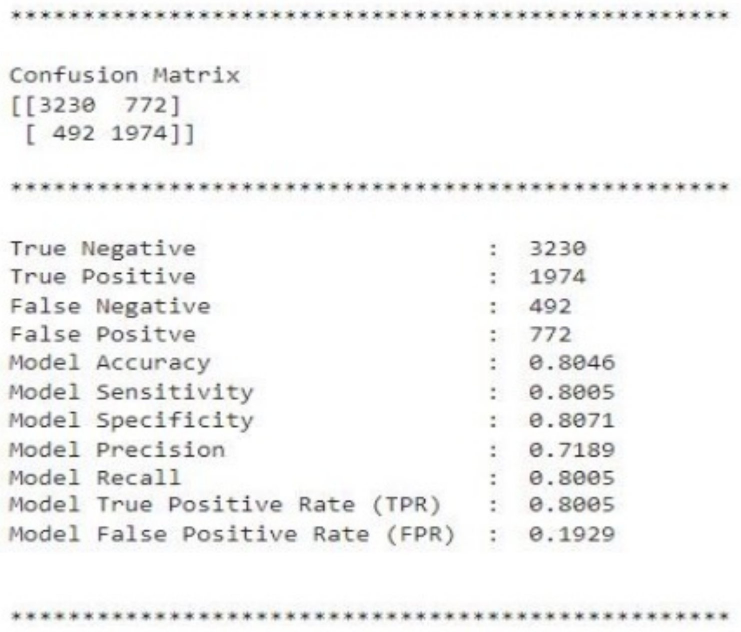
ROC CURVE – TEST DATA SET

The Area Under The ROC Curve (AUC) Is 0.87, Indicating A Strong Ability To Discriminate Between Positive And Negative Cases. In Other Words, The Model Effectively Distinguishes Between Those Who Will Experience The Event Of Interest And Those Who Won't.



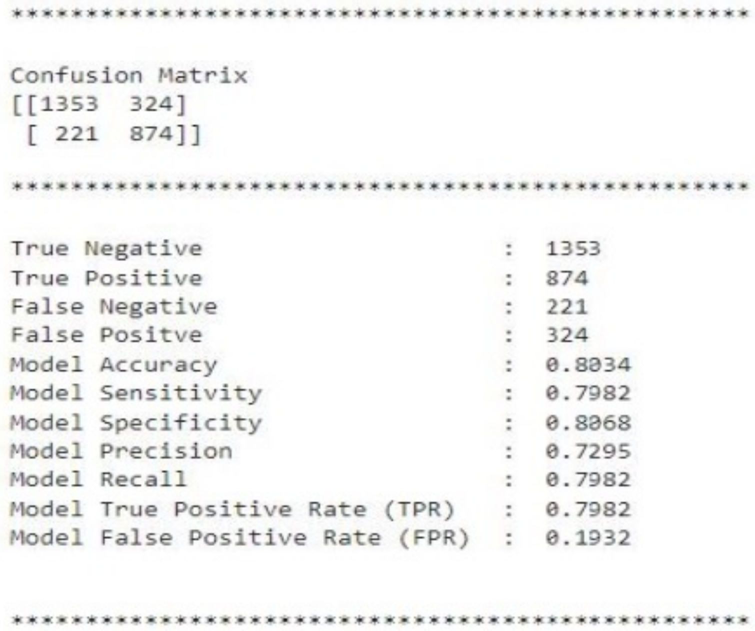
MODEL EVALUATION
CONFUSION MATRIX & METRICS

TRAIN DATA SET



```
**************************************************

Confusion Matrix
[[3230  772]
 [ 492 1974]]

**************************************************

True Negative                     :  3230
True Positive                     :  1974
False Negative                    :  492
False Positve                     :  772
Model Accuracy                    :  0.8046
Model Sensitivity                 :  0.8005
Model Specificity                 :  0.8071
Model Precision                   :  0.7189
Model Recall                      :  0.8005
Model True Positive Rate (TPR)    :  0.8005
Model False Positive Rate (FPR)   :  0.1929

**************************************************
```

TEST DATA SET



```
**************************************************

Confusion Matrix
[[1353  324]
 [ 221  874]]

**************************************************

True Negative                     :  1353
True Positive                     :  874
False Negative                    :  221
False Positve                     :  324
Model Accuracy                    :  0.8034
Model Sensitivity                 :  0.7982
Model Specificity                 :  0.8068
Model Precision                   :  0.7295
Model Recall                      :  0.7982
Model True Positive Rate (TPR)    :  0.7982
Model False Positive Rate (FPR)   :  0.1932

**************************************************
```

- A Sensitivity Of 80.05% (Train Set) And 79.82% (Test Set) Was Achieved Using A Cut-off Value Of 0.345. Sensitivity, In This Context, Refers To The Model's Ability To Correctly Identify True Positives, Which Translates To The Proportion Of Converting Leads The Model Correctly Classified.
- The CEO Of X Education Prioritized Achieving A High Level Of Sensitivity, Ideally Around 80%.
- The Developed Model Achieved An Accuracy Of 80.46%, Demonstrating Successful Alignment With The Study's Objectives Of Prioritizing Sensitivity.

RECOMMENDATION

- Aligning With The Problem Statement's Emphasis On Lead Conversion As A Critical Driver For X Education's Growth, A Regression Model Was Developed. This Model Helps Identify The Most Significant Factors Impacting A Lead's Likelihood Of Converting.
- Analysis Of The Model's Coefficients Revealed The Following Features With The Highest Positive Influence On Conversion:
    - Lead Source_welingak Website: 5.39
    - Lead Source_reference: 2.93
    - Current_occupation_working Professional: 2.67
    - Last Activity_sms Sent: 2.05
    - Last Activity_others: 1.25
    - Total Time Spent On Website: 1.05
    - Last Activity_email Opened: 0.94
    - Lead Source_olark Chat: 0.91
- The Model Also Identified Features With Negative Coefficients, Potentially Indicating Areas For Improvement:
    - Specialization: Hospitality Management: -1.09
    - Specialization: Others: -1.20
    - Lead Origin: Landing Page Submission: -1.26

RECOMMENDATION

❑ **Optimizing Lead Conversion Strategies:**

- The Model's Positive Coefficient Analysis Highlights Key Features For Targeted Marketing Strategies.
- Efforts Should Prioritize Attracting High-quality Leads Through Top-performing Sources Like The Welingak Website And Referrals.
- Increased Advertising Or Promotional Efforts On The Welingak Website Can Be Considered Based On Its Strong Association With Lead Conversion.
- Communication Channels, Such As SMS And Email, Can Be Optimized Based On Their Positive Impact On Lead Engagement.
- Tailored Messaging Targeting Working Professionals Can Be Developed, Considering Their Higher Conversion Rates And Potential For Higher Program Enrollment Fees.
- Implementing Incentive Programs For Referrals That Convert Can Be Explored To Encourage Existing Satisfied Students Or Contacts To Provide More References.

❑ **Addressing Potential Areas For Improvement:**

- Specializations With Negative Coefficients, Like Hospitality Management And "Others," Require Further Analysis To Identify Potential Reasons For Lower Conversion Rates.
- The Landing Page Submission Process Should Be Reviewed For Areas Of Improvement To Enhance Its Effectiveness In Converting Visitors Into Leads.

THANK YOU!