Lead Scoring Case Study

- Neha Tayal
- Raksha Deshmukh
- Vivek Rawat

TABLE OF CONTENTS

Company insights

- X Education focuses on providing online courses tailored for industrial professionals.

- The company utilizes its website as a platform for showcasing its courses to a wide audience.

- Digital marketing strategies are employed to promote these courses on various online platforms, including search engines like Google.

Company insights

- Visitors to the website are engaged through interactive content such as videos and forms, aimed at capturing leads.

- Essential lead information, including email addresses and phone numbers, is collected through these forms.

- The sales team then follows up with leads via phone calls, emails, SMS, and other communication channels to facilitate conversions.

- Despite its efforts, the company's average lead conversion rate currently stands at 30%.

Problem Statement

The current lead conversion rate is relatively low at 30%. The company aims to enhance conversion efficiency by pinpointing hot leads. Insights are required for the sales department to prioritize nurturing specific leads for maximizing conversions. Developing a model to assign lead scores based on conversion probabilities is imperative. Higher scores would be allocated to leads with greater conversion potential, while lower scores would be assigned to those with lower chances. The CEO has set a target lead conversion rate of approximately 80%.

Ideas of lead conversion

Segmentation of Leads:
- Leads categorized based on their likelihood to convert.
- Facilitates targeted nurturing of leads with higher conversion potential.

Optimization of Communication Channels:
- Prioritizing communication channels for targeted leads.
- Maximizes investment of time on leads with higher conversion probability.

Increasing Conversion Rates:
- Enhanced focus on targeted groups leading to higher conversion rates.
- Aims to achieve the CEO's set target of 80% conversion rate.

Considering our target of an 80% conversion rate, it's crucial to recognize that hot leads are highly responsive to variable changes.

Steps taken in Data Analysis

1. Data Cleaning:
- Addressed null values and outliers in the dataset.

2. Exploratory Data Analysis (EDA):
- Managed imbalances and conducted univariate and bivariate analyses.

3. Data Preparation:
- Created dummy variables, performed test-train split, and applied feature scaling.

Steps taken in Data Analysis

4. Model Building:
- Employed feature selection techniques like Recursive Feature Elimination (RFE) and reduction methods.

5. Model Evaluation:
- Utilized the confusion matrix, selected cutoffs, and calculated lead scoring.

6. Predictions:
- Compared test and train results, assigned lead scores, and identified top features.

7. Recommendations:
- Proposed new ideas and areas for improvement.

Data Cleaning

- - The default option "Select" is treated as null for certain categorical variables, as it indicates that no option was chosen.
- - Columns containing more than 40% null values are removed from the dataset to maintain data integrity.
- - Missing values in categorical variable columns are handled based on various factors, including value counts and other considerations.

Data Cleaning

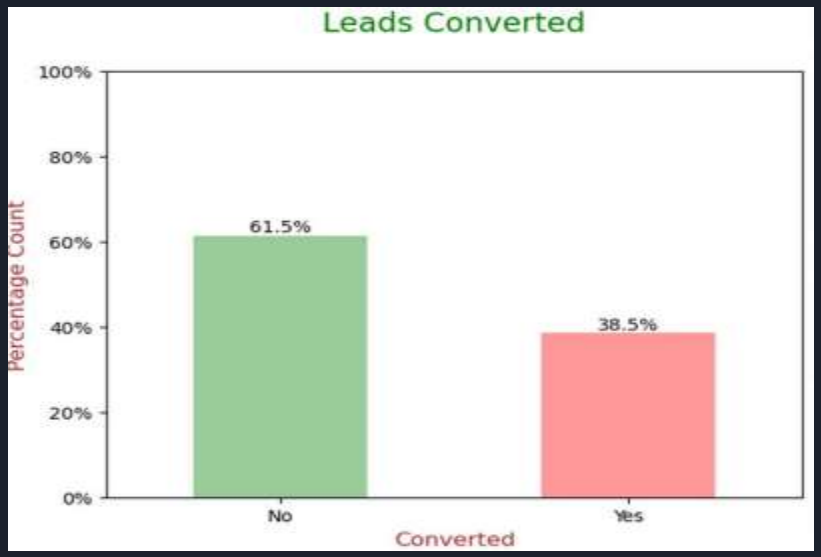
- - Columns that do not significantly contribute to the study's objective are dropped to streamline the dataset.
- - Imputation techniques are applied to fill in missing values for certain categorical variables.
- - New categories are created for specific variables to enhance the granularity of the data.
- - Columns deemed less significant for modeling purposes, such as Prospect ID and Lead Number, are eliminated.
- - Numerical data is imputed using the mode after assessing the distribution of values.

Data Cleaning

- - Skewed categorical variables are identified and dropped from the dataset to minimize bias in logistic regression models.
- - Outliers in variables such as Total Visits and Page views Per Visit are identified and treated by capping them at a certain threshold.
- - Invalid values in certain columns are defaulted and standardized to ensure consistency and reliability in the dataset.
- - Values with low frequency across categorical variables are grouped together under the category "Others" to simplify the data and improve model performance.
- - Binary categorical variables are mapped together to streamline the data representation.
- - Further cleaning processes are implemented to enhance data quality and accuracy, including fixing invalid values and standardizing casing styles for consistency.

Exploratory Data Analysis (EDA)

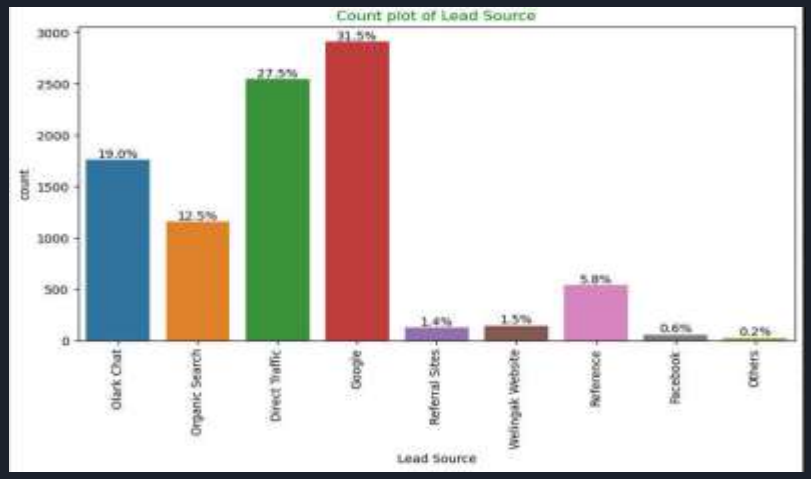
Data imbalance in target variable analysis



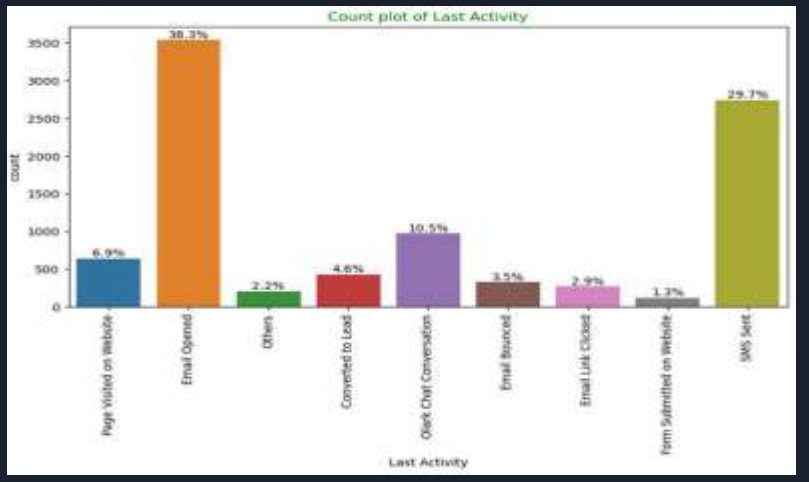
- Conversion rate is 38.5%, meaning only 38.5% of the leads are buying customers.
- 61.5% of the leads didn’t buy anything.

Exploratory Data Analysis (EDA)

Univariate Analysis of Categorical Variables



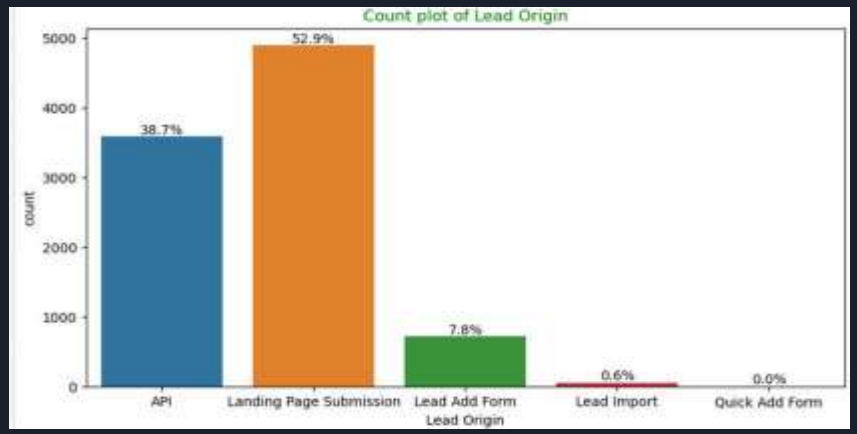
Lead Source: 31.5% of the leads are from Google, followed by 27.5% leads from direct traffic



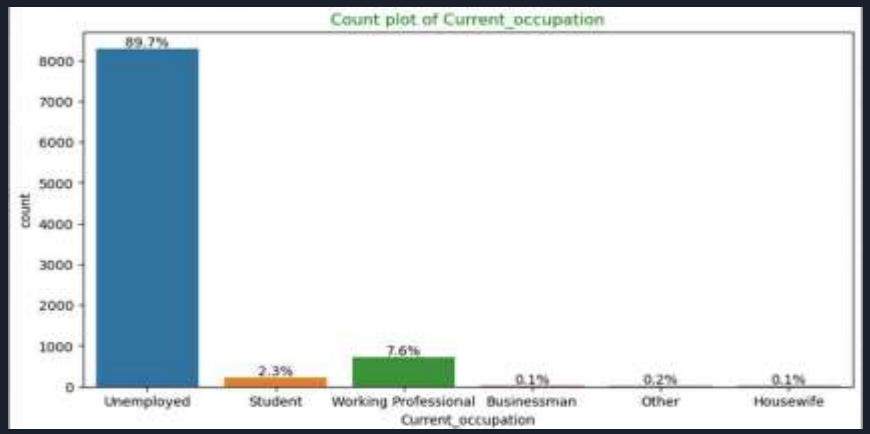
Last Activity: “Email Opened” and “SMS” sent are the major last activity of the leads at 38.3% and 29.7% respectively.

Exploratory Data Analysis (EDA)

Univariate Analysis of Categorical Variables



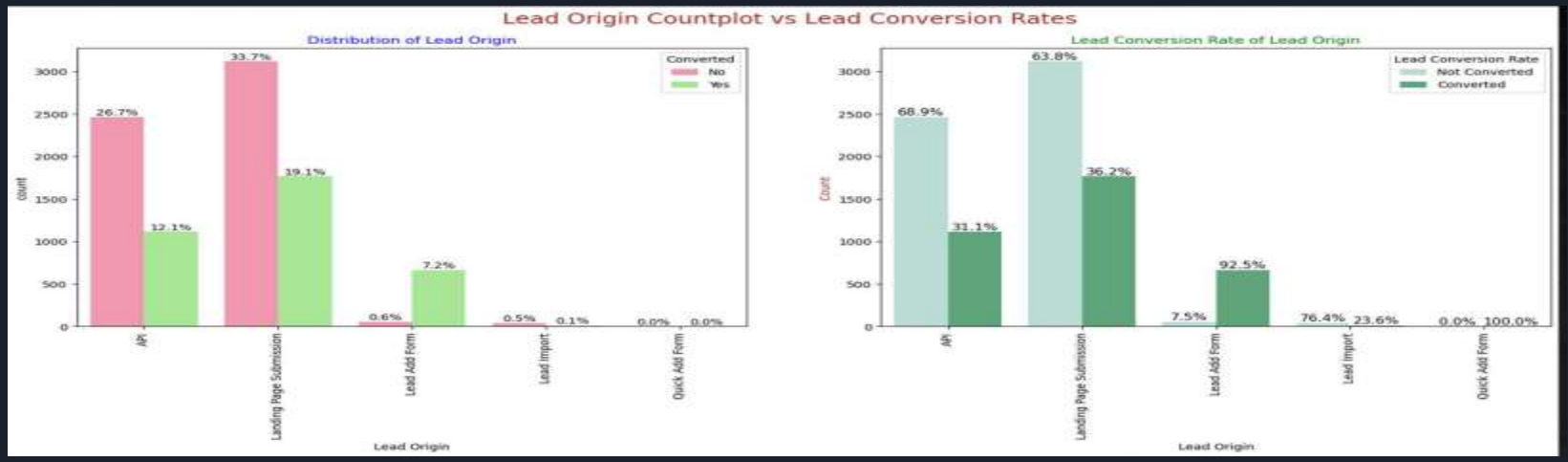
Lead Origin: Majority of the leads come from "Landing Page Submission" with 53% of customers, followed by "API" and "Lead Add Form" with 38.7% and 7.8% respectively.



Current_occupation: 90% of the leads are unemployed, and only 7.6% are working professionals.

EDA- Bivariate analysis

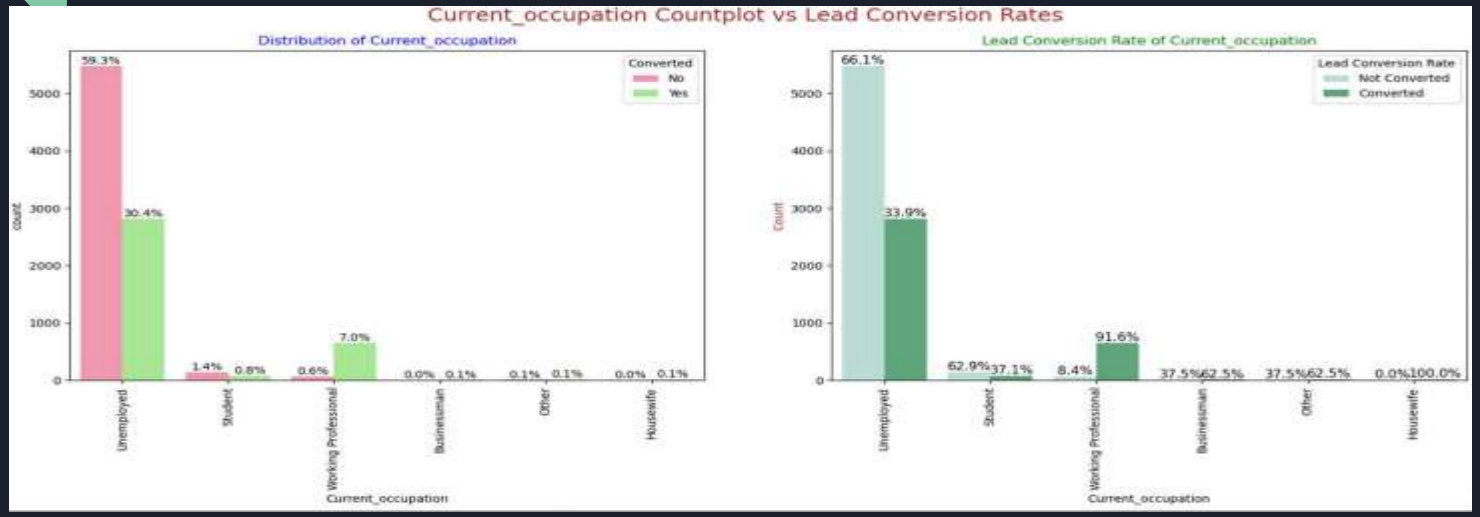
EDA Bivariate Analysis



Lead Origin:

- 52.8% of all leads come from “Landing Page Submission”, and have a Lead conversion rate (LCR) of 36.2%
- 38.8% of all leads come from “API”, and have an LCR of 31.1%

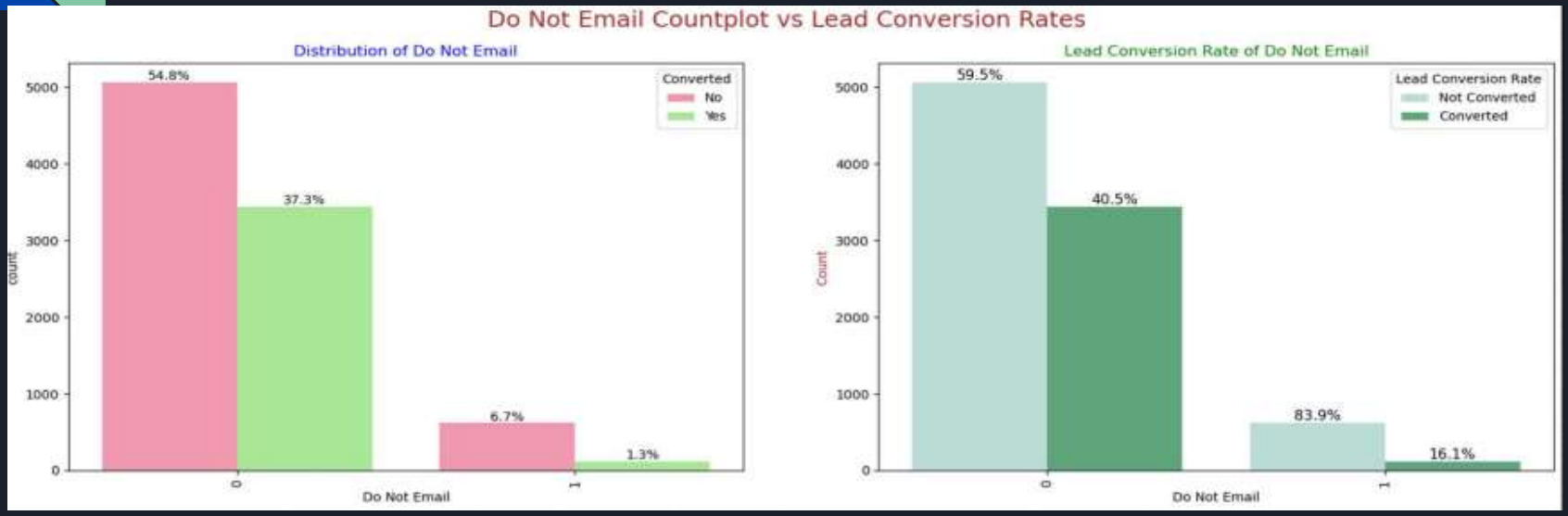
EDA- Bivariate Analysis



Current Occupation:

- 89.7% of the leads are unemployed, and they have a lead conversion rate (LCR) of only 33.9%
- Only 7.6% of the leads are working professionals, but they contribute to 91.6% of the conversions

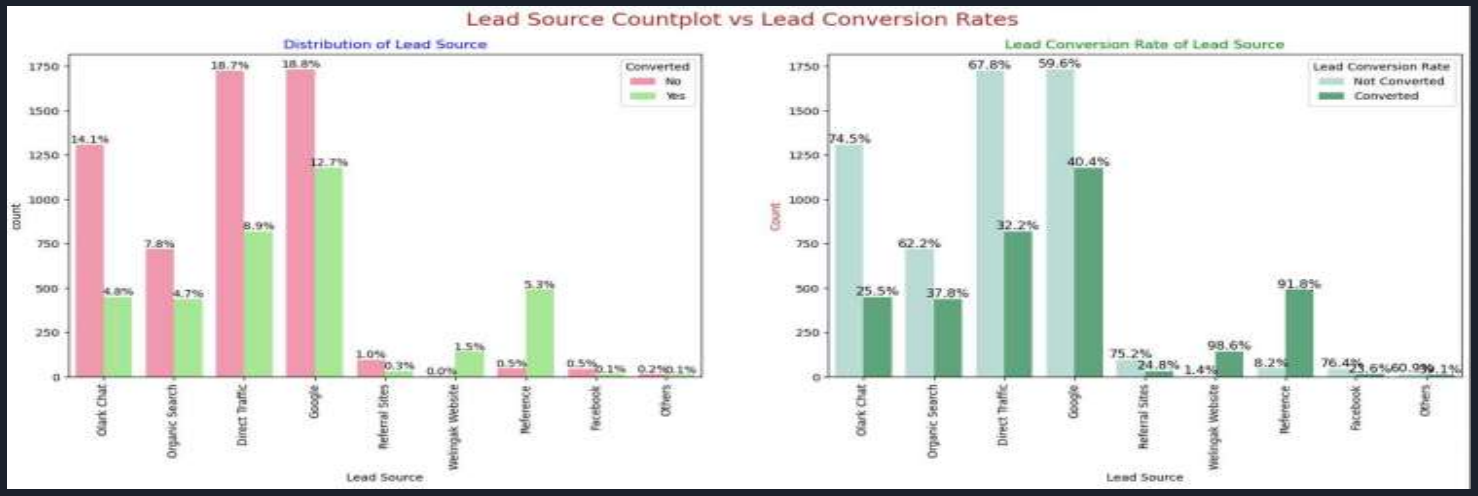
EDA-Bivariate Analysis



Do not Email:

- 91.9% of the leads did not select the option “Do not email”, and their LCR stands out to be 40.5%
- This goes to show that engaging the leads through emails is an effective way to nurture them

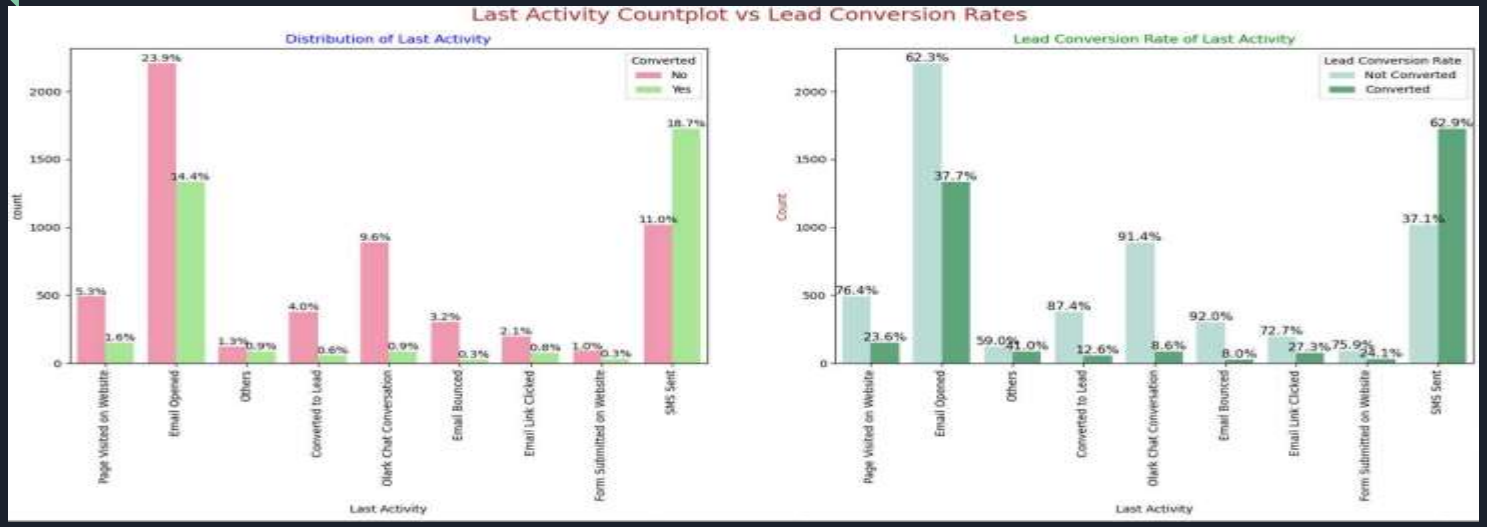
EDA-Bivariate Analysis



Lead Source:

- 31.5% of the leads are sourced from Google out of which we obtain a 40.4% LCR
  Direct traffic accounts for 27.6% of the leads and yields an LCR of 32.2%
- References contribute to only 5.8% of all leads, yet 91.8% of them convert

EDA-Bivariate Analysis

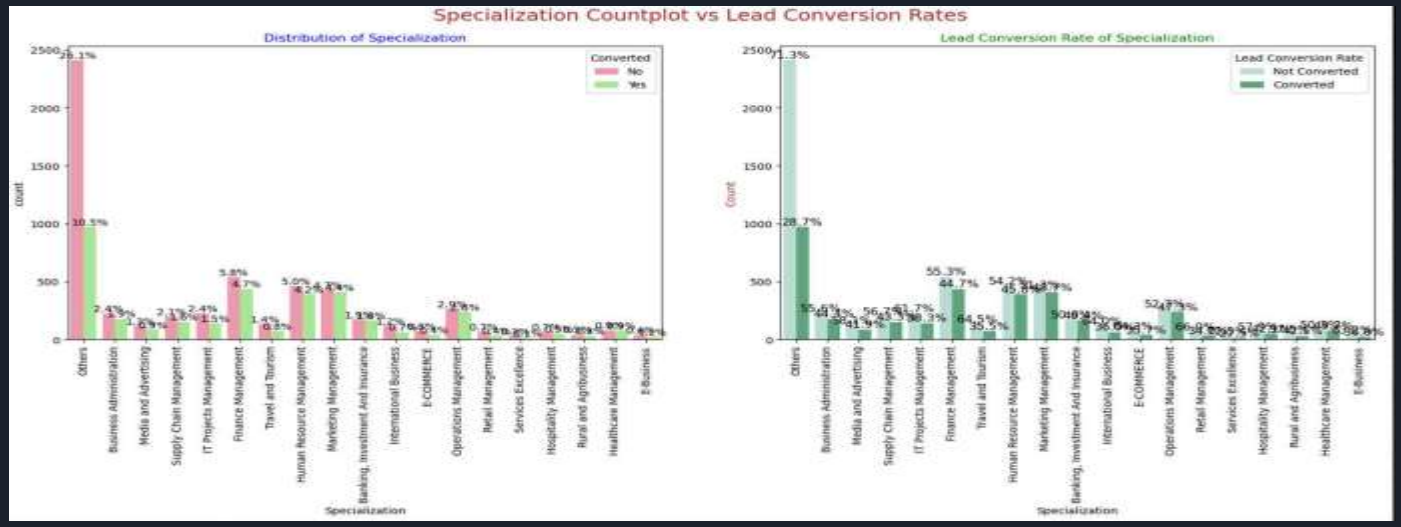


Last Activity:

"Email opened" is the most common last activity (38.3%), in which 37.7% leads are converted

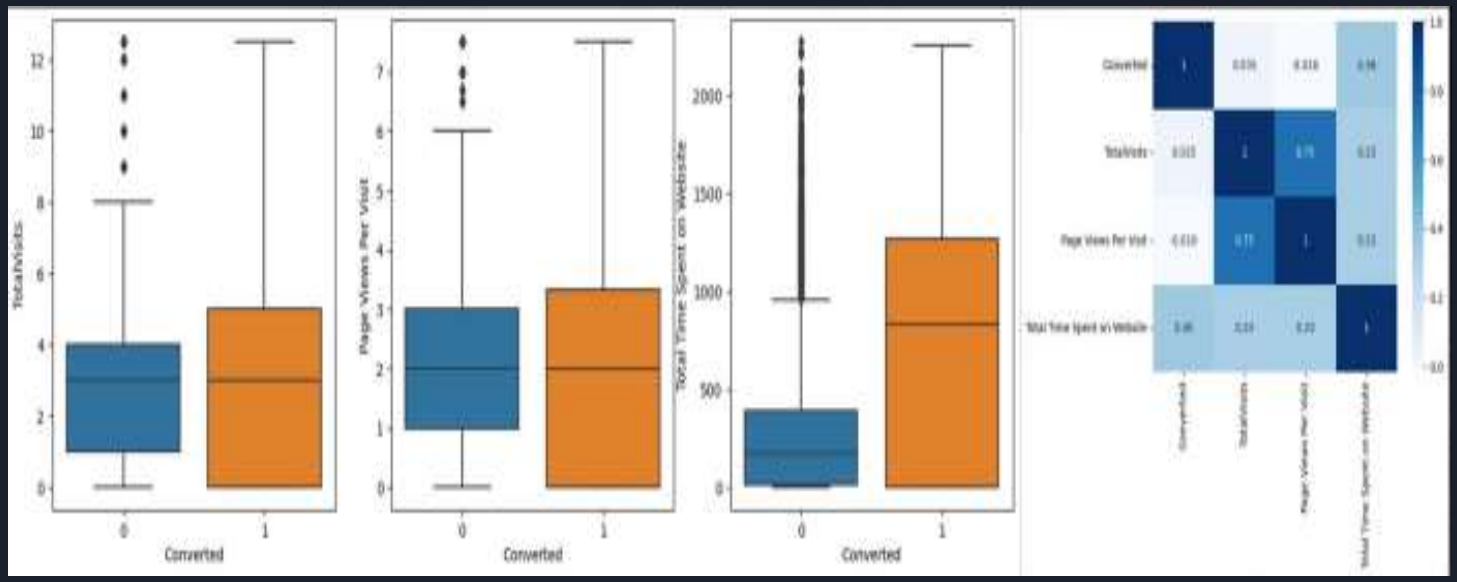
"*SMS sent" is the second most common one* (29.7%), that yields an LCR of 62.9

EDA-Bivariate Analysis



Specialization:
Management contribute to the highest lead pool and convers
Marketing, Human Resources, and Finance ion rate after “Other” specialization

EDA-Bivariate Numerical Variables



Based on the data, it can be inferred that leads that spend more time on the website are more likely to convert than those who spend less time on the same.

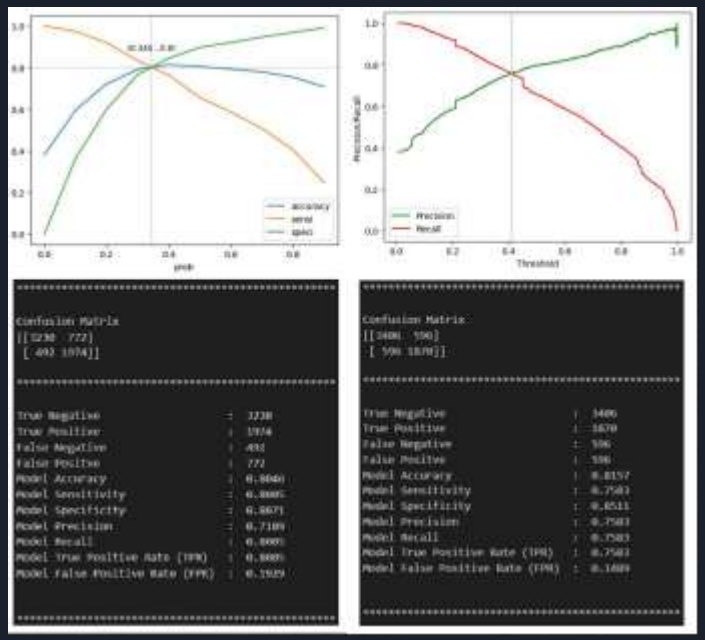
Preparing the Data

- Categorical binary variables were previously encoded as 1s and 0s.
- Dummy features were generated for categorical variables including Lead Origin, Lead Source, Last Activity, Specialization, and Current_occupation.
- The dataset was split into training and testing sets using a 70:30 ratio.
- Feature scaling was applied to standardize the features using the standardization method.
- Correlation analysis was conducted to identify highly correlated predictor variables, which were subsequently removed to mitigate multicollinearity issues.

Model Building

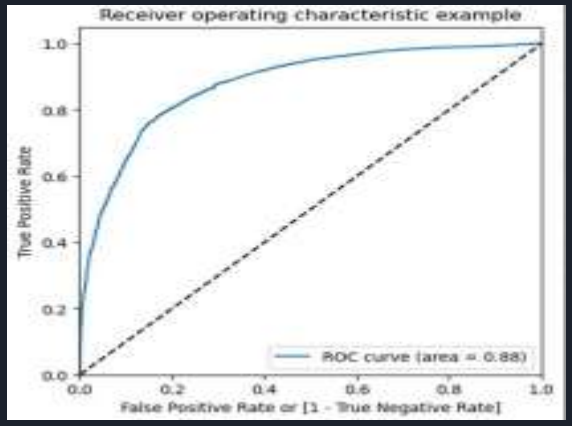
- **-** Given the dataset's high dimensionality and numerous features, including all features in the model may negatively impact performance and computational efficiency.
- - Therefore, Recursive Feature Elimination (RFE) is performed to identify and select only the most important columns.
- -The outcome of RFE reveals that prior to the process, the dataset contained 48 columns.
- - Manual feature reduction was employed by systematically dropping variables with p-values greater than 0.05, ensuring statistical significance in the model.
- - Through four iterations, Model 4 emerged as the most appropriate for addressing the study's objectives.
- - Consequently, Model 4 is selected as the final model for further analysis and interpretation.

Evaluating the model (train data set)

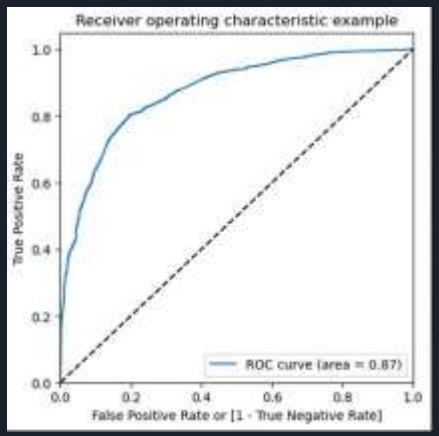


- A cutoff threshold of 0.345 was selected based on the evaluation metrics obtained from both plots.
- The confusion matrix and evaluation metrics using the 0.345 cutoff threshold are displayed in the image on the left.
- Additionally, the image on the right showcases the confusion matrix and evaluation metrics using a cutoff threshold of 0.41.

Evaluating the model (train data set)



ROC train data set



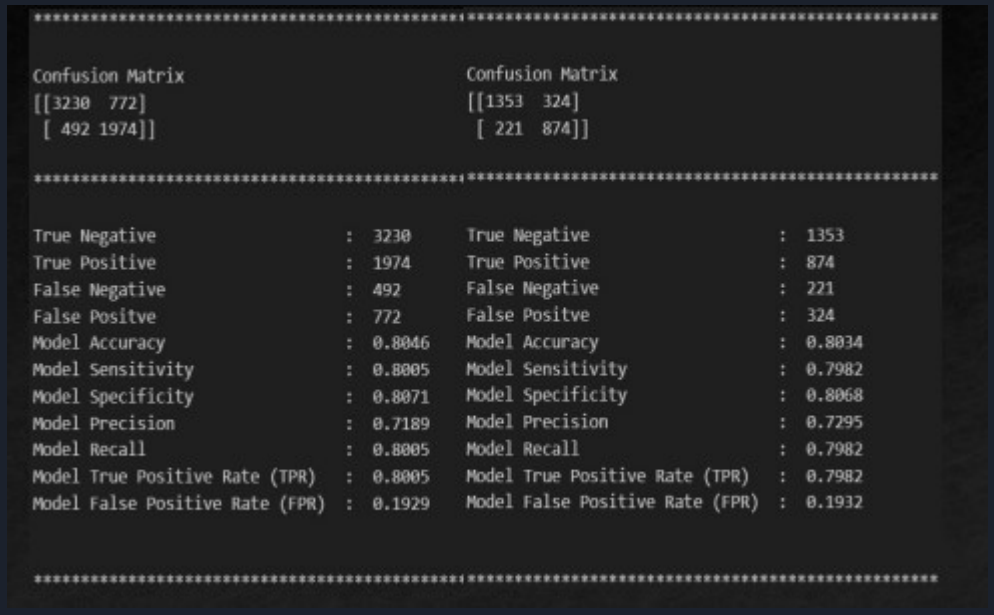
ROC test data set

Evaluating the model (train data set)

The ROC (Receiver Operating Characteristic) curve covers an area of 0.88 and 0.87 (train and test), respectively, for the two cutoff thresholds. This indicates that the model's prediction performance is good. A higher area under the ROC curve suggests that the model has better discrimination ability and is more effective at distinguishing between the positive and negative classes. Therefore, with ROC areas of 0.88 and 0.87, the model demonstrates strong predictive capability.

The ROC curve being closest to the top-left corner indicates a high true positive rate and a low false positive rate. This positioning signifies that the model achieves a high level of accuracy in correctly identifying positive cases while minimizing the number of false alarms (false positives). Therefore, the model demonstrates strong predictive performance with a favorable balance between true positives and false positives.

Evaluating the model Confusion matrix and matrices



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

Confusion Matrix                                          Confusion Matrix
[[3230  772]                                              [[1353  324]
 [ 492 1974]]                                              [ 221  874]]

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

True Negative                  :  3230     True Negative                  :  1353
True Positive                  :  1974     True Positive                  :  874
False Negative                 :  492      False Negative                 :  221
False Positve                  :  772      False Positve                  :  324
Model Accuracy                 :  0.8046   Model Accuracy                 :  0.8034
Model Sensitivity              :  0.8005   Model Sensitivity              :  0.7982
Model Specificity              :  0.8071   Model Specificity              :  0.8068
Model Precision                :  0.7189   Model Precision                :  0.7295
Model Recall                   :  0.8005   Model Recall                   :  0.7982
Model True Positive Rate (TPR) :  0.8005   Model True Positive Rate (TPR) :  0.7982
Model False Positive Rate (FPR):  0.1929   Model False Positive Rate (FPR):  0.1932

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

Evaluating the model (train data set)

- The model achieved a sensitivity of 80.05% in the train set and 79.82% in the test set, using a cut-off value of 0.345.
- Sensitivity in this case indicates the accuracy of the model for each customer.
- The model also achieved an accuracy of 80.46%, which is in line with the study's objectives.

Recomendations

To increase Lead Conversion Rates:

- Allocate more budget for advertisement expenditure on Welingak website.
- Come up with a referral program for existing customers and incentivize them
- Optimize the product to suit the requirements of working professionals. Engage them with tailored messaging.
- Optimize communication channels such as SMS and email.

Recommendations

- Make sure customers spend more time on the website. This can be done by optimizing the UI/UX.
- Leverage email marketing.

To identify areas of improvement:

- Optimize content for specializations since not many leads are converting in specific specialization programs.
- Use Olark Chat only as the initial touch point. After that, shift the primary communication channel to SMS or WhatsApp (since it is easier to send multimedia messages, and make customer see the company logo regularly).

THANK YOU