X Education - Lead Scoring Case Study

- Tilak Shah

Problem Statement

1. X Education's lead conversion rate is suboptimal despite acquiring a significant number of leads daily.
2. Identifying potential leads with a high likelihood of conversion remains a challenge.
3. The company aims to increase its lead conversion rate to 80% by implementing targeted strategies for potential leads.

Objectives

1. Develop a logistic regression model to assign lead scores, prioritizing potential leads based on their likelihood of conversion.
2. Improve lead conversion rates by focusing sales efforts on high-scoring leads identified by the model.
3. Ensure the model's adaptability to future challenges and evolving business requirements for sustained effectiveness.

Approach

Data Cleaning

Load data, Underdtanding Data and Cleaning

EDA

Check imbalance, Univariate & Bivariate analysis

Data Preparation

Get a modern PowerPoint Presentation that is beautifully designed.

Model Building

RFE for top 15 feature, Manual Feature Reduction & finalizing model

Model Evaluation

Confusion matrix, ROC Curve
Finding Optimal Cutoff Point

Predictions on Test Data

Compare train vs test metrics

Recommendation

Suggest features to focus for higher conversion & areas for improvement

Data Cleaning

- Checking for Select Values and replacing with Nulll.

- Droping Columns with more than 40% NuLL Values.

- Imputing categorical columns missing values
  Imputing the following columns
  -'Specialization' with 'Others'
  -'Lead Source' with 'Google'
  - 'Last Activity’ with 'Email Opened'
  - 'What is your current occupation' with 'Unemployed‘.
- Imputing numeric columns missing values.

- Removing Columns with only one Values.

- Outliers Check.

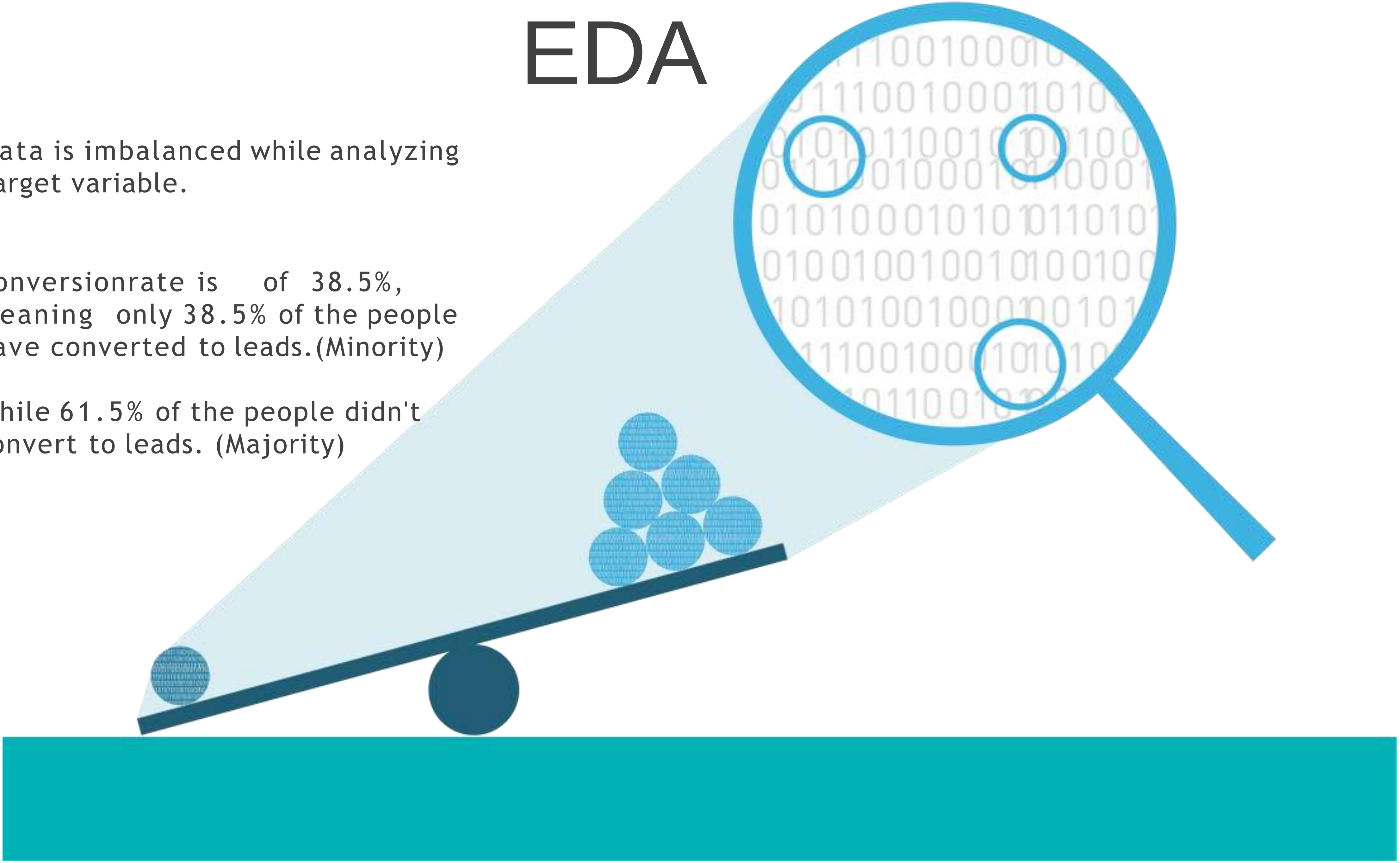
- Grouping Low frequency values to ‘Others’.

- Mapping Binary categorical variables (Yes to 1/No to 0).

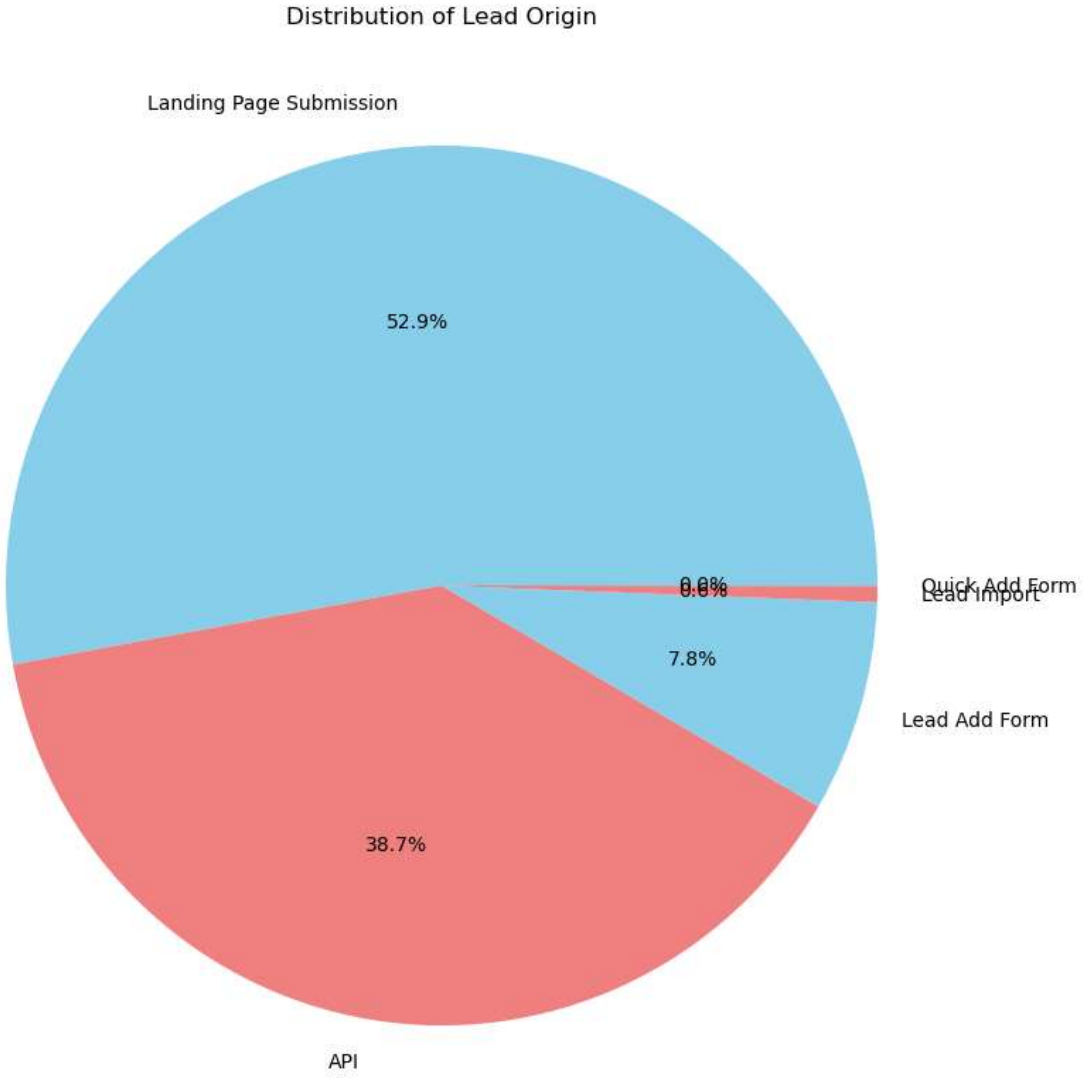
EDA

1. Data is imbalanced while analyzing target variable.

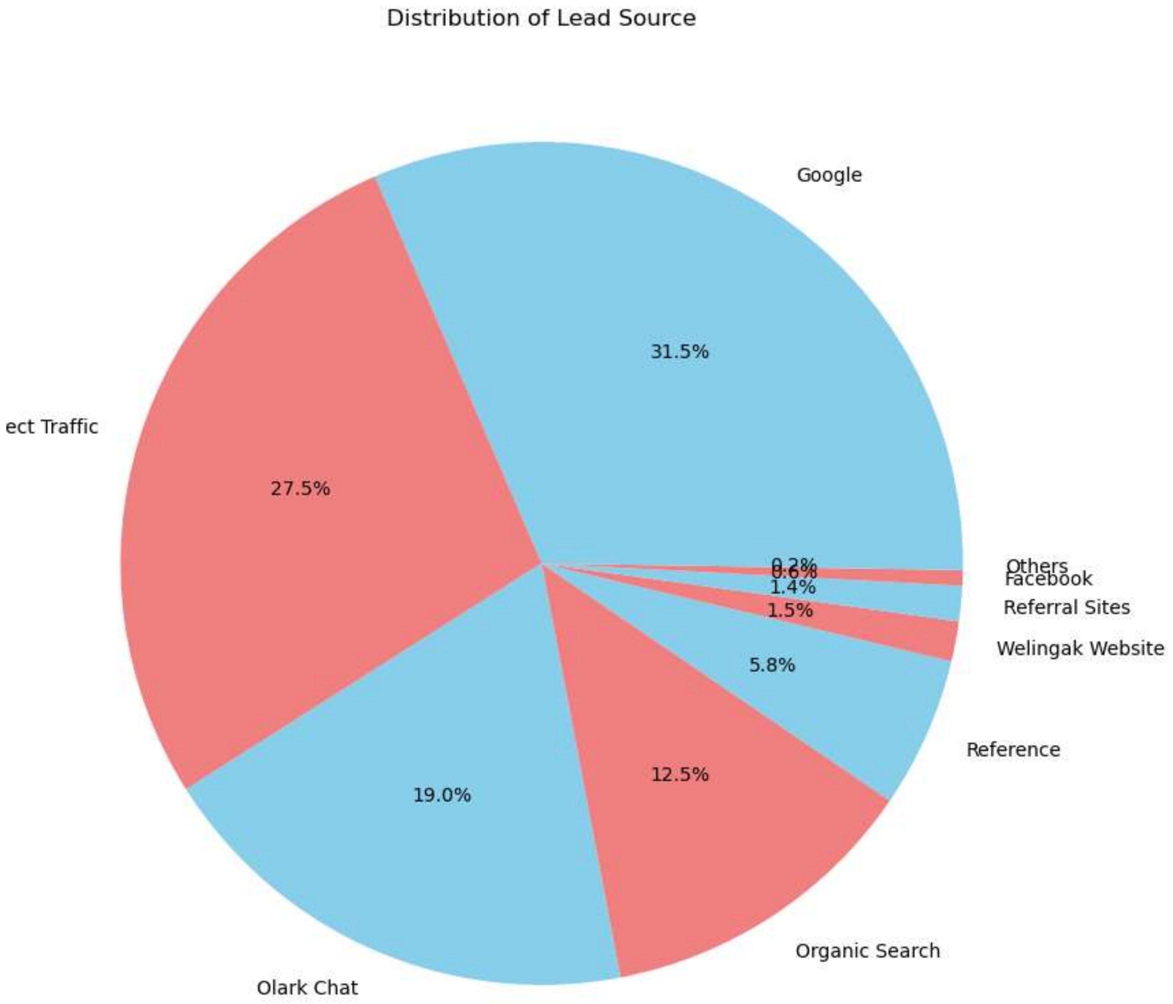
- Conversionrate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)
- While 61.5% of the people didn't convert to leads. (Majority)



EDA **Univariate Analysis**

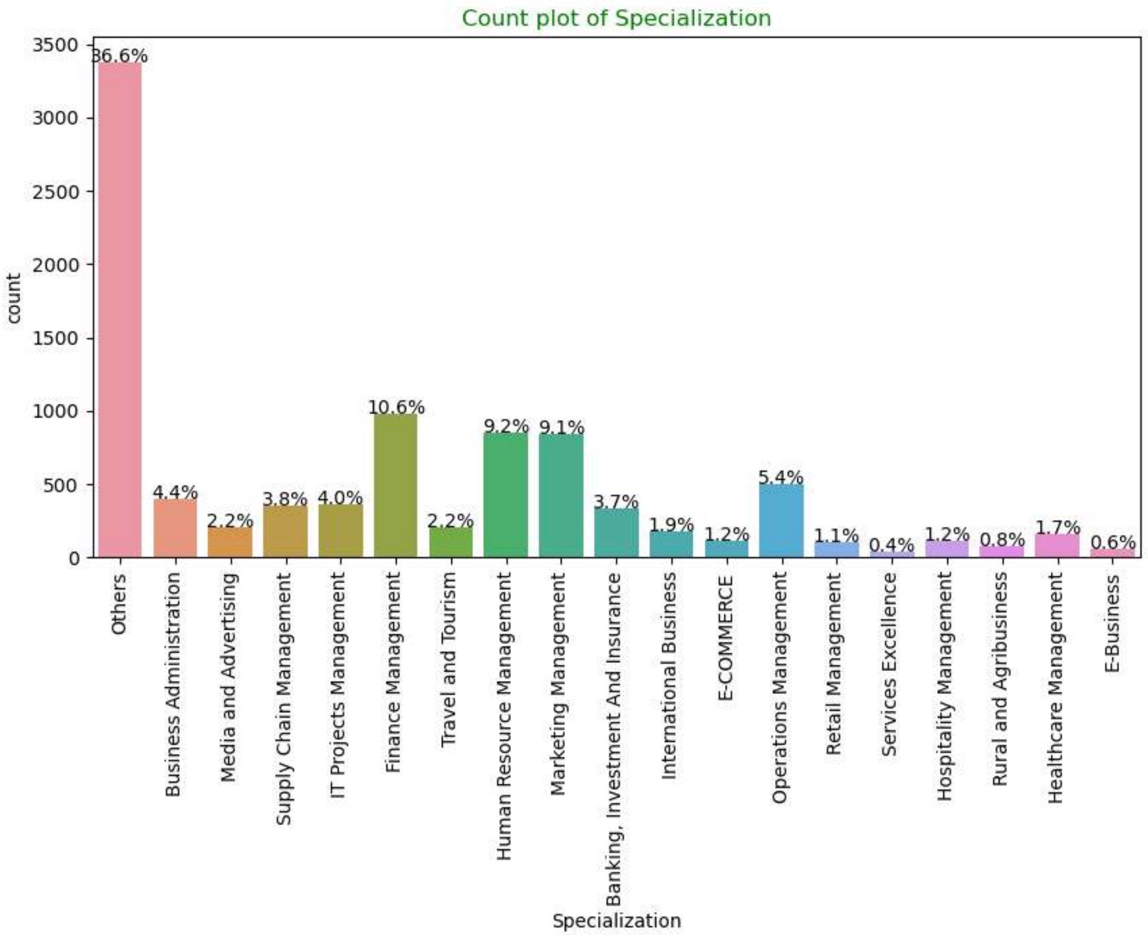


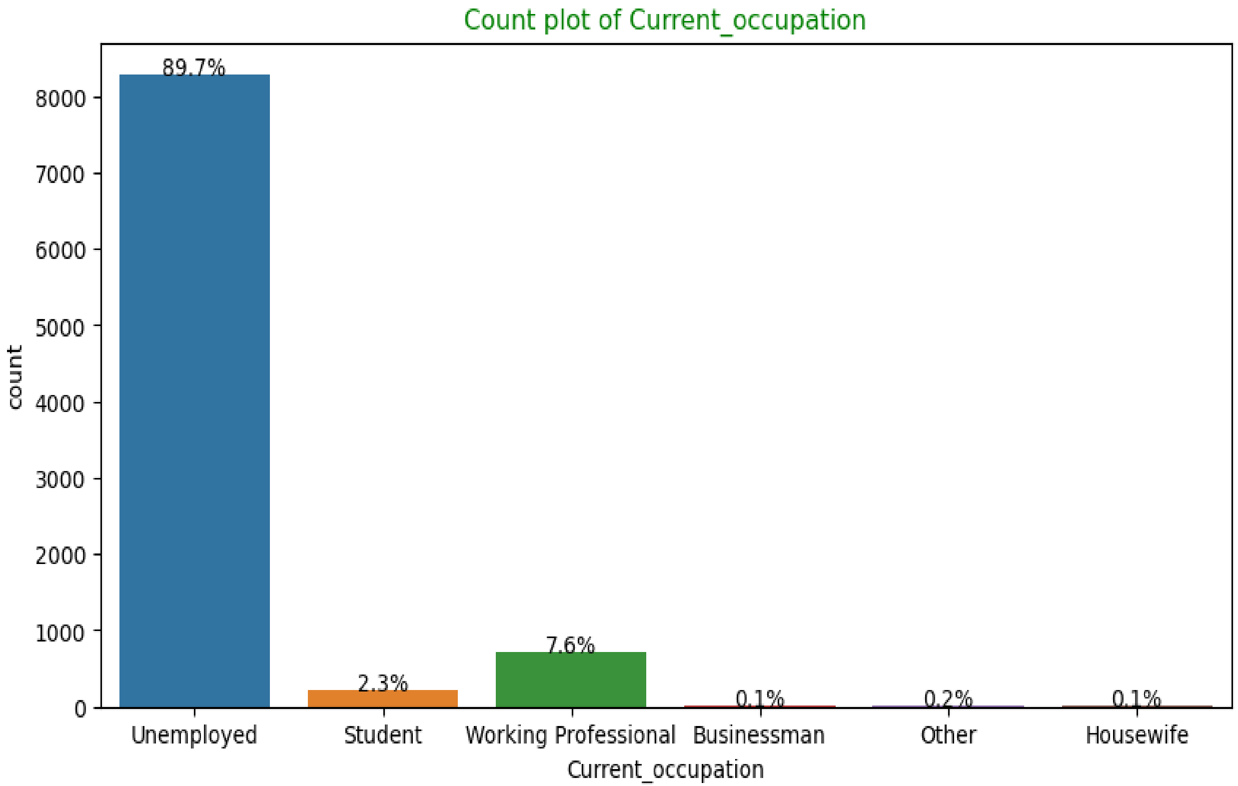
Lead Origin: "Landing Page Submission" identified 53% customers, "API" identified 39%.



Lead Source: 31.5% Lead source is from Google & 27.5% from Direct Traffic followed by Olark Chat which is 19.0%

EDA **Univariate Analysis**

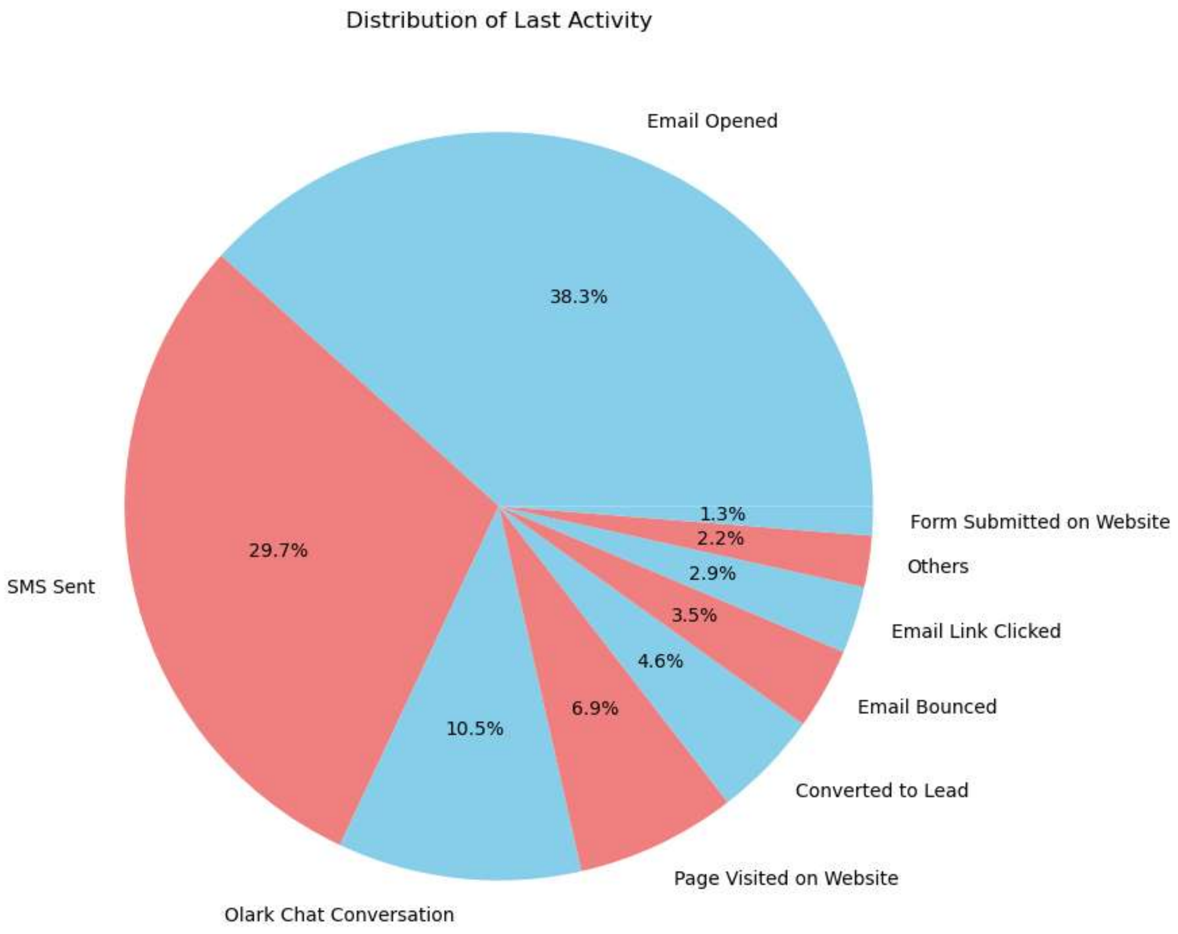


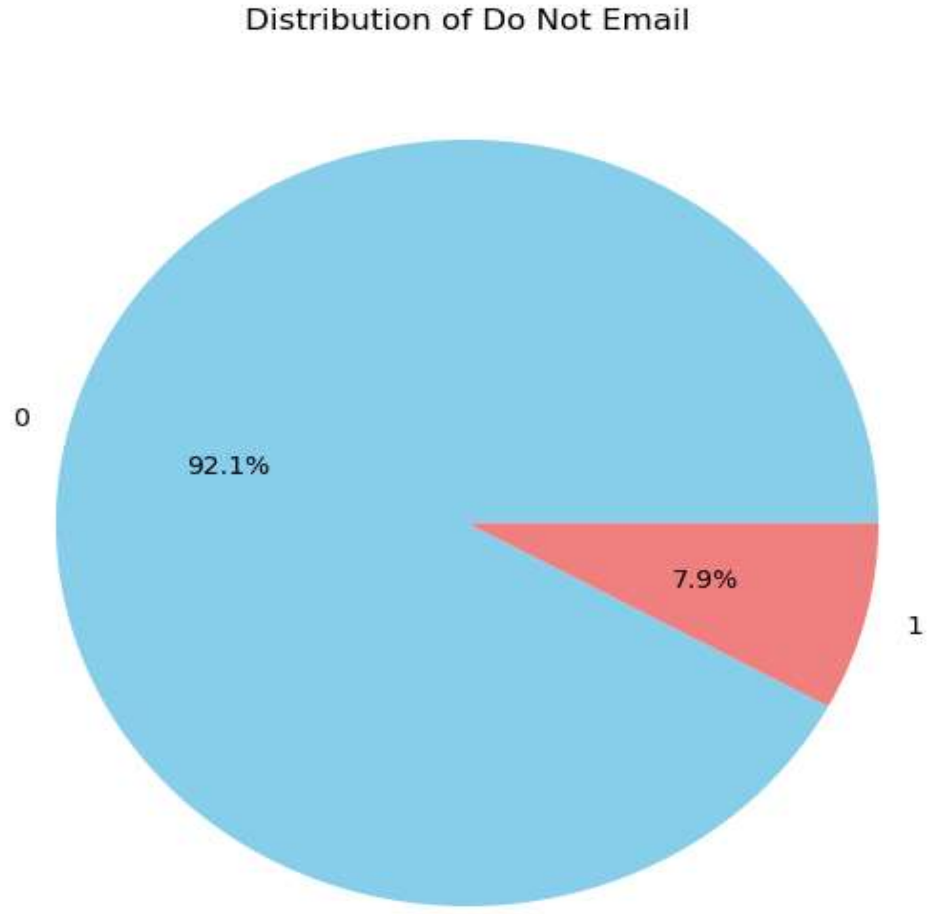


Specialization: Apart from Finance Managemant, HR management and Marketing management we 36% customers from ‘Others’.

Current_occupation: It has 90% of the customers as Unemployed

EDA **Univariate Analysis**

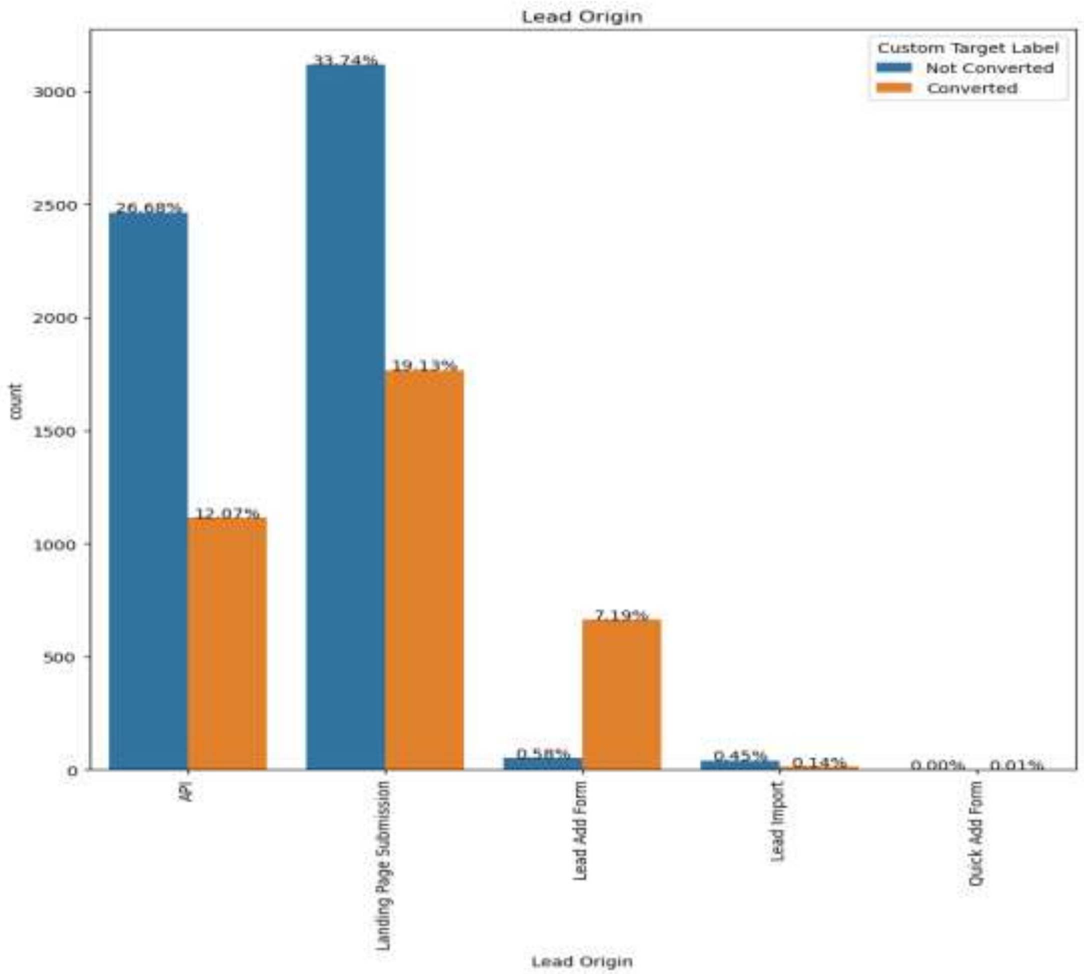


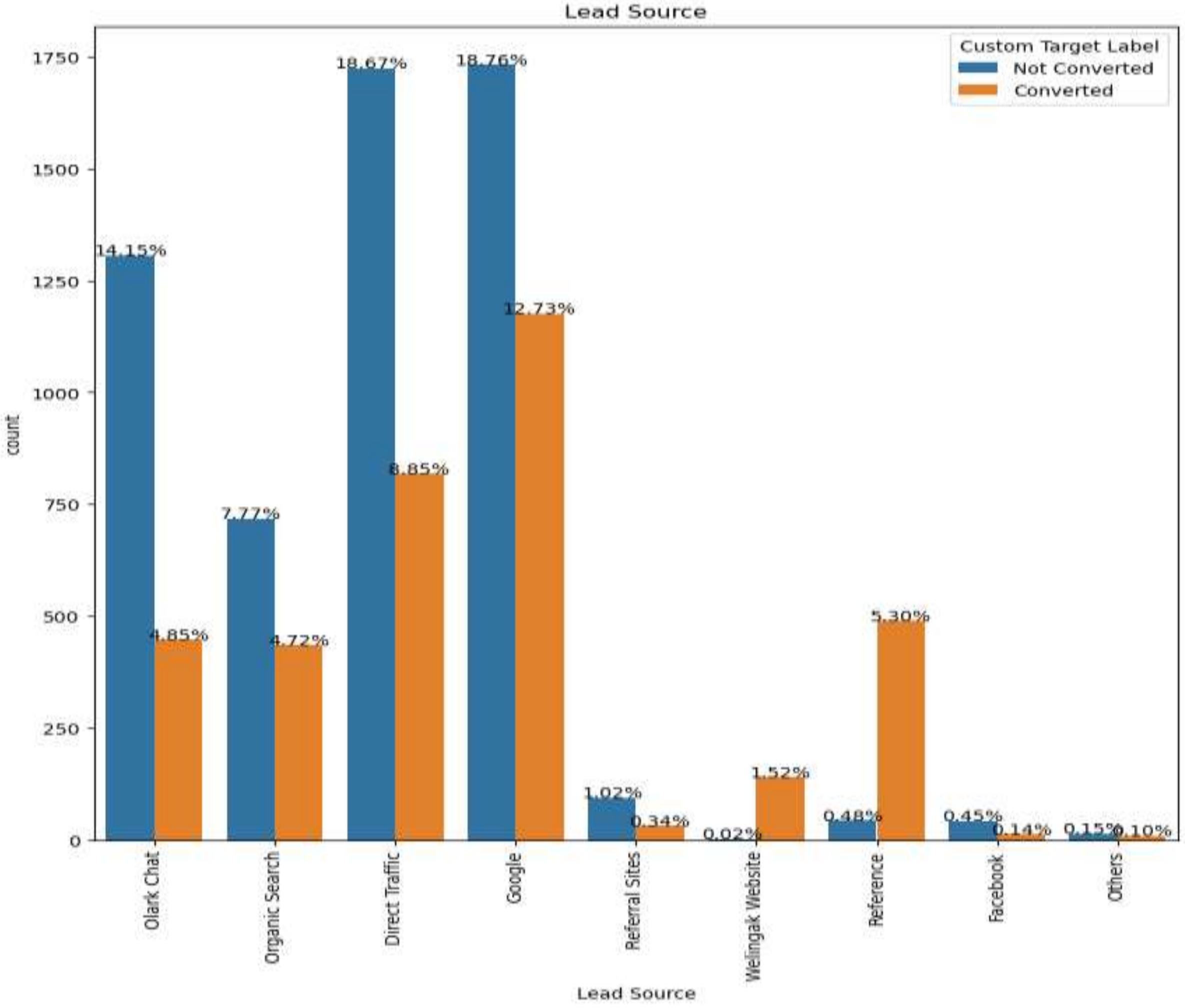


Last Activity: 68% of customers contribution in SMS Sent & Email Opened activities

Do Not Email: 92% of the people dont want to be emailed about the course.

EDA **Bivariate Analysis**

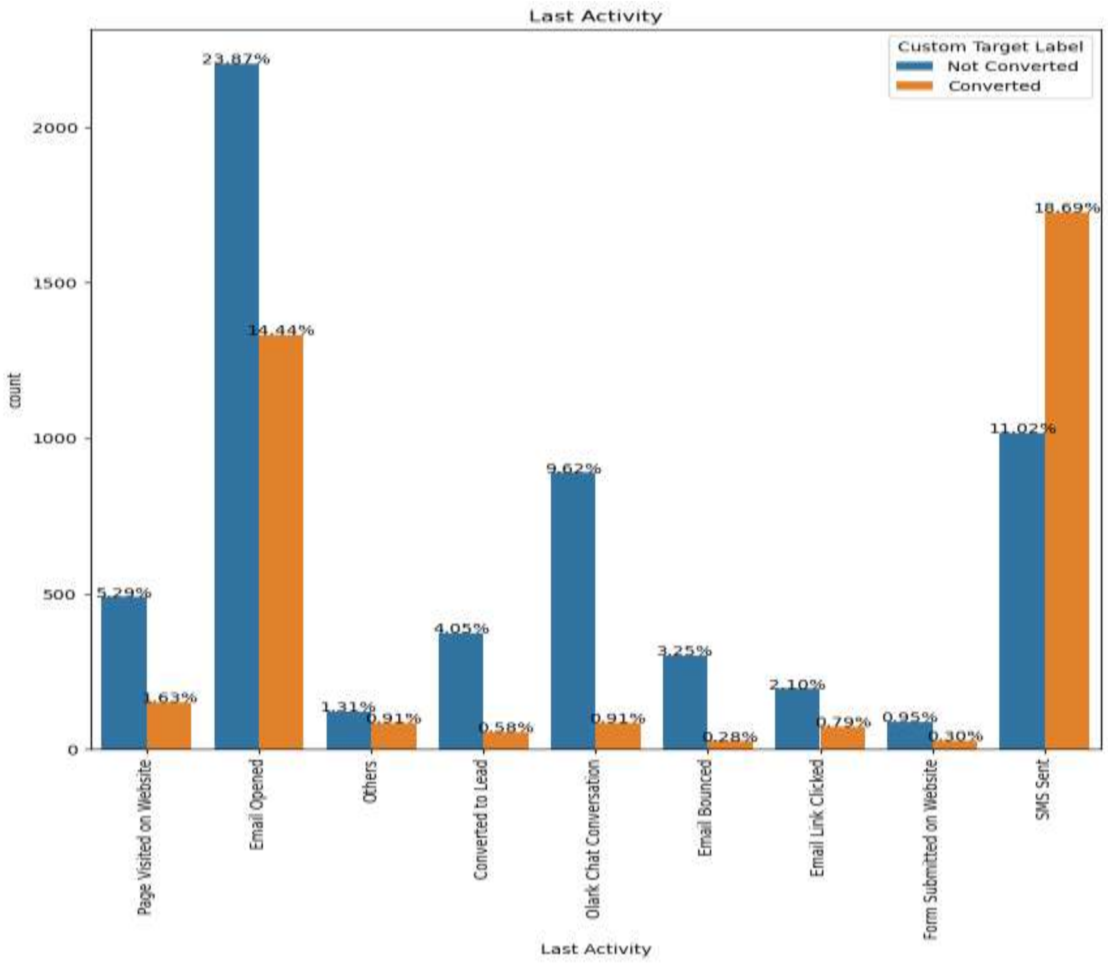


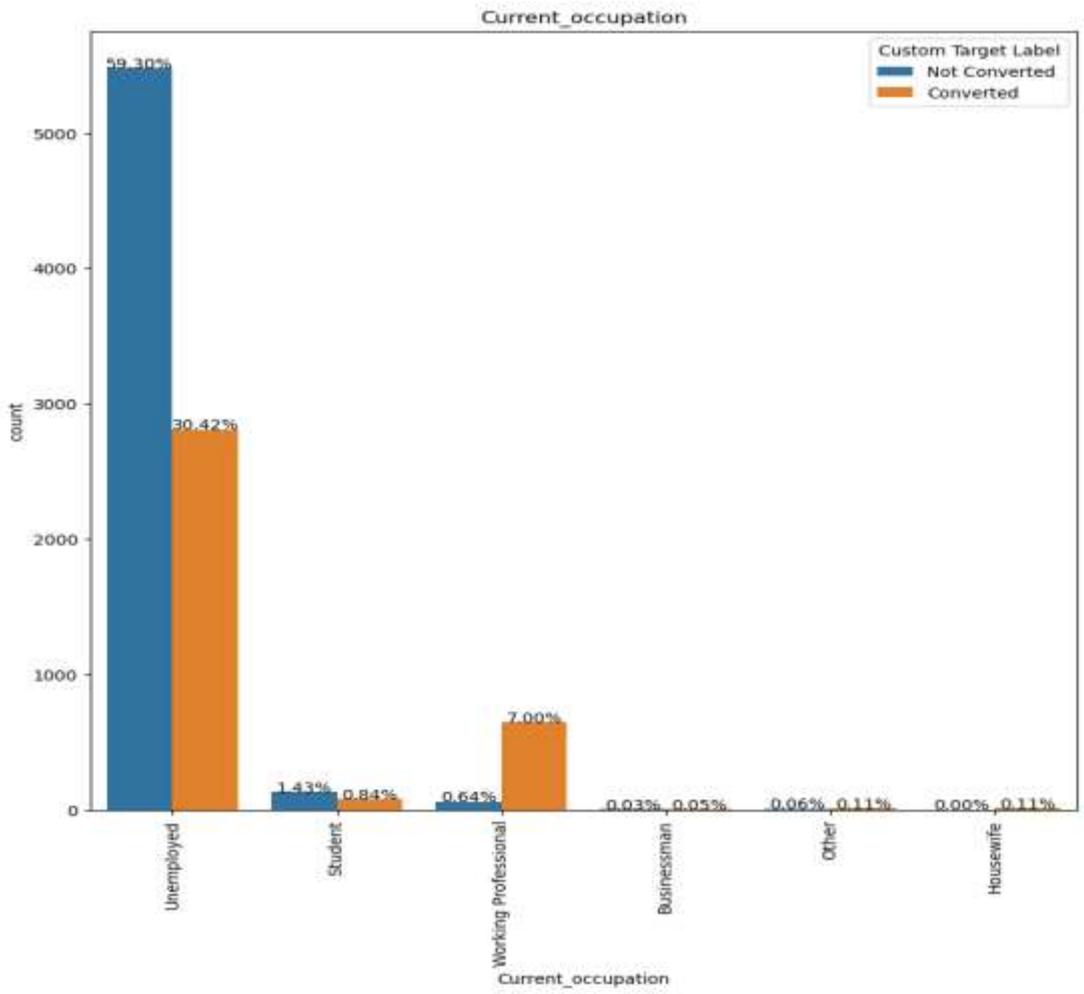


Lead Origin: We can see that leads originated from "Landing Page Submission is highest which is 52.87 (33.74 + 19.13) which has Lead conversion 36% The "API" has approximately 39% of Leads with a lead conversion rate (LCR) of 31% approximately.

Lead Source: Google has approx 30 % of Leads from which 40% of leads gets converted also Direct Traffic has approx 27% Leads from which has 32% of LCR. Reference Share highest LCR which is 91.6 %

EDA **Bivariate Analysis**

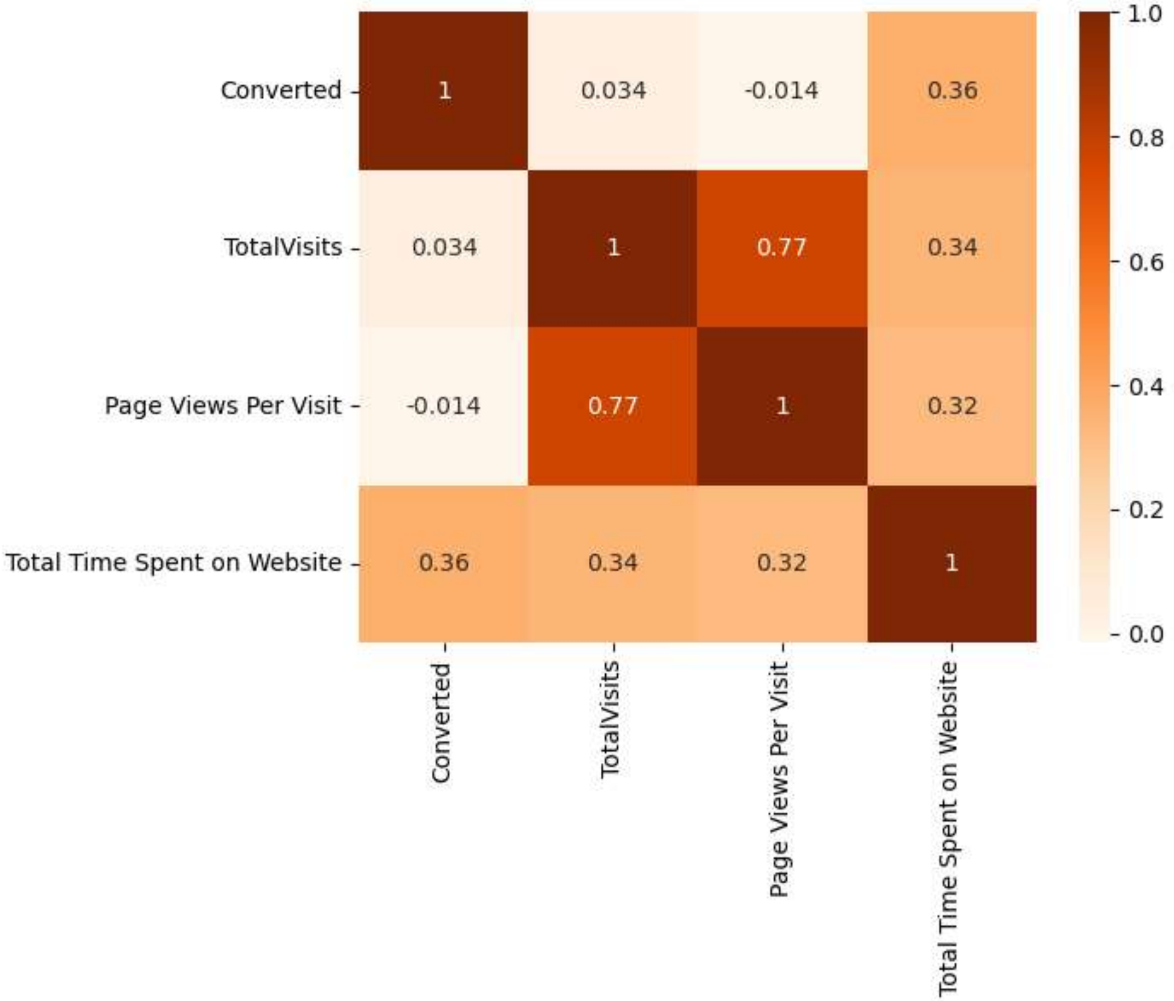


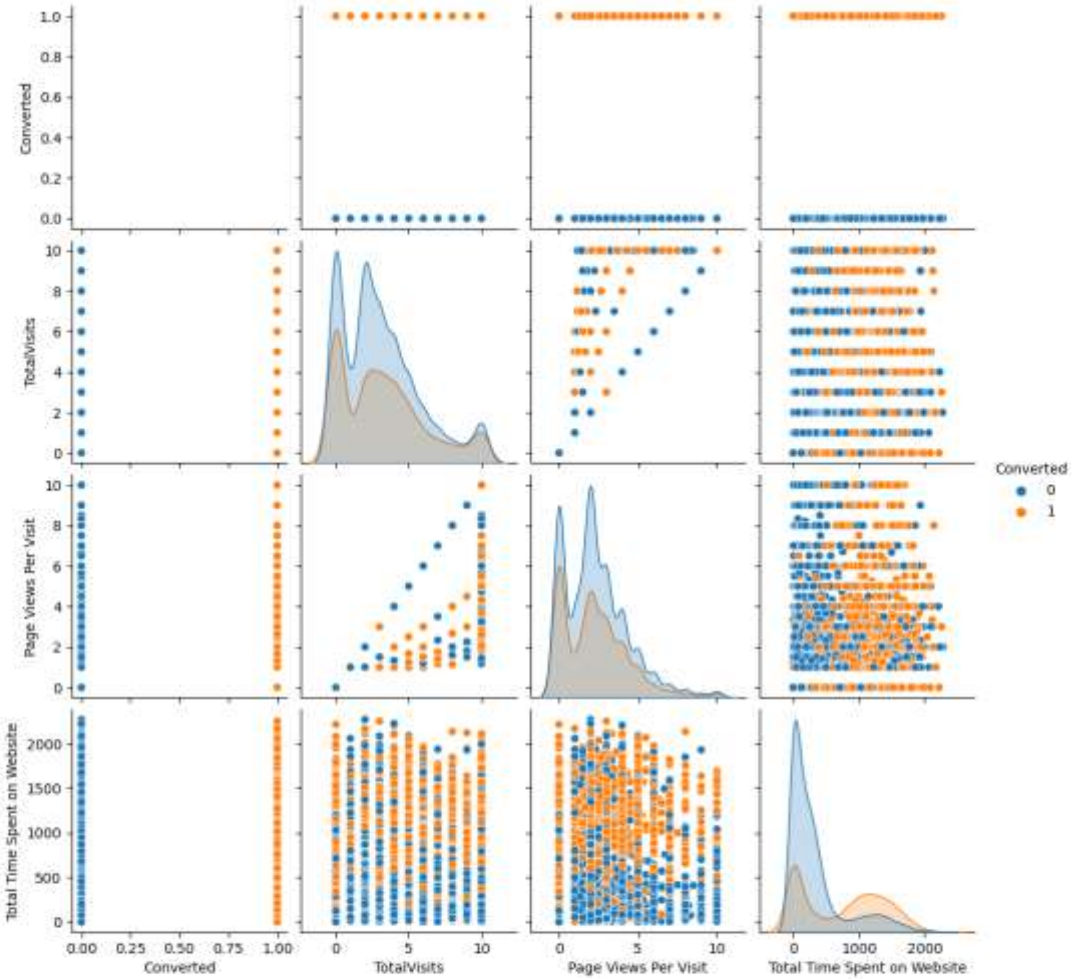


Last Activity: ‘SMS Sent’ has highest LCR (61%), which has generated 29% of leads, where 'Email Opened' activity contributed 38% of last activities performed by the customers with 37% lead conversion rate.

Current_occupation: We can see that 90% of the customers are Unemployed with lead conversion rate (LCR) of 34%. While Working Professional has 7.6% of total customers and has 92% lead conversion rate (LCR).

EDA **Bivariate Analysis**





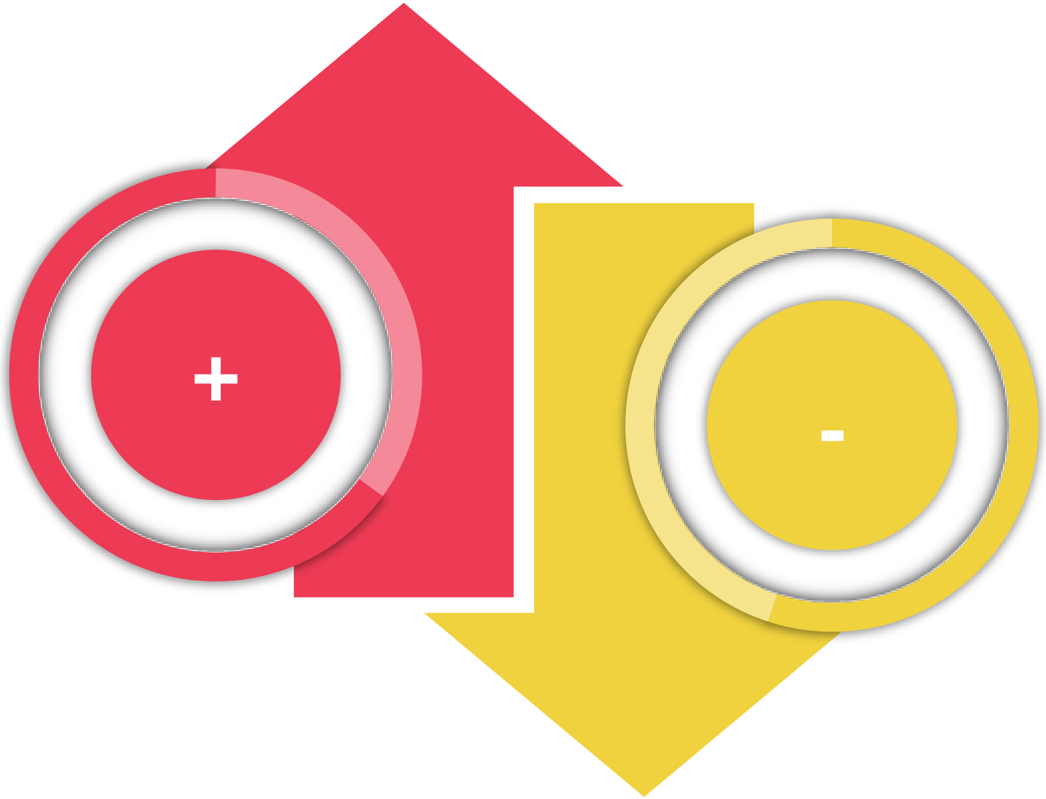
There is a strong positive correlation between 'Total Visits' and 'Page Views per Visit', indicating that customers who visit the website more frequently tend to view more pages per visit.

Data Preparation before Model building

- Categorical Encoding: Binary-level categorical columns converted to 1s and 0s, Which was ‘Yes’ and ‘No’ previously.
- Data Split: Dataset split into 70% training and 30% testing sets.
- Feature Scaling: Standardization used to scale features for consistent magnitude.
- Correlation Analysis: Identified and removed highly correlated features, like Lead Origin_Lead Import and Lead Origin_Lead Add Form.

Model Building

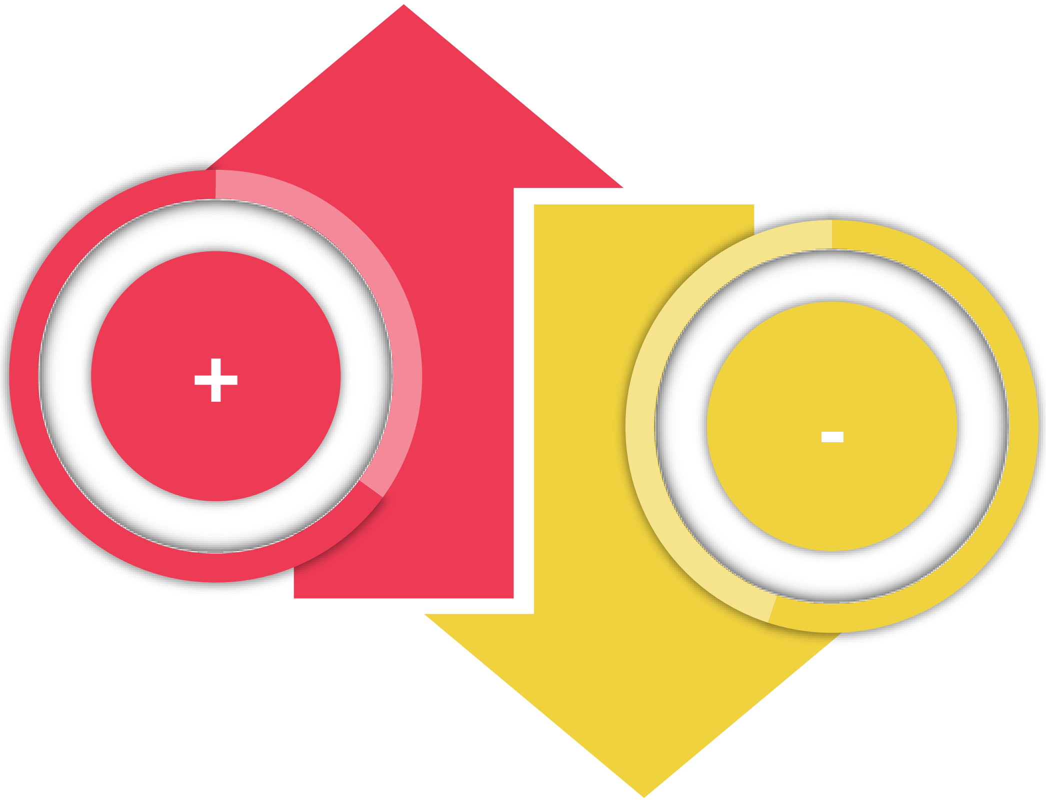
- Feature Selection Importance: Due to high dimensionality and numerous features, performing Recursive Feature Elimination (RFE) is crucial to enhance model performance and reduce computation time.
- RFE Outcome: Initially, the dataset had 55 columns, which were reduced to 15 columns post RFE, focusing on the most important features.
- Manual Feature Reduction: Variables with p-values greater than 0.05 were dropped through manual feature reduction to refine the model.
- Model Stability: Model 4 demonstrated stability after four iterations, exhibiting significant p-values (<0.05) and no multicollinearity issues (VIFs < 5).
- Final Model Selection: Based on stability criteria, "logm4" was chosen as the final model for Model Evaluation and prediction purposes.



Model Equation

```
Equation : =

-0.855087 x const
- 1.110290 x Do Not Email
+ 1.043936 x Total Time Spent on Website
- 1.225732 x Lead Origin_Landing Page Submission
+ 0.916386 x Lead Source_Olark Chat
+ 2.949361 x Lead Source_Reference
+ 5.476113 x Lead Source_Welingak Website
+ 0.752973 x Last Activity_Email Opened
- 0.717945 x Last Activity_Olark Chat Conversation
+ 1.408670 x Last Activity_Others
+ 1.929261 x Last Activity_SMS Sent
- 1.071914 x Specialization_Hospitality Management
- 1.194564 x Specialization_Others
+ 2.632282 x Current Occupation_Working Professional
```



Model Evaluation

Train Data Set

Confusion Matrix & Evaluation Metrics with 0.35 as cutoff

In [151]:
```
# Finding Confusion metrics for 'y_train_pred_final' df
confusion_matrix = metrics.confusion_matrix(y_train_pred_final['Converted'], y_train_pred_final['final_predicted'])
print("Confusion Matrix")
print(confusion_matrix,"\n")
```

```
Confusion Matrix
[[3223  779]
 [ 489 1977]]
```

In [153]:
```
print("accuracy:",round((TN+TP)/(TN+TP+FN+FP),5))
```

```
accuracy: 0.80396
```

In [154]:
```
print("Sensitivity:",round(TP/(TP+FN),5))
```

```
Sensitivity: 0.8017
```

In [155]:
```
print("Specificity:",round(TN/(TN+FP),5))
```

```
Specificity: 0.80535
```

In [156]:
```
print("Precision:",round(TP/(TP+FP),5))
```

```
Precision: 0.71734
```

In [157]:
```
print("Recall:",round(TP/(TP+FN),5))
```

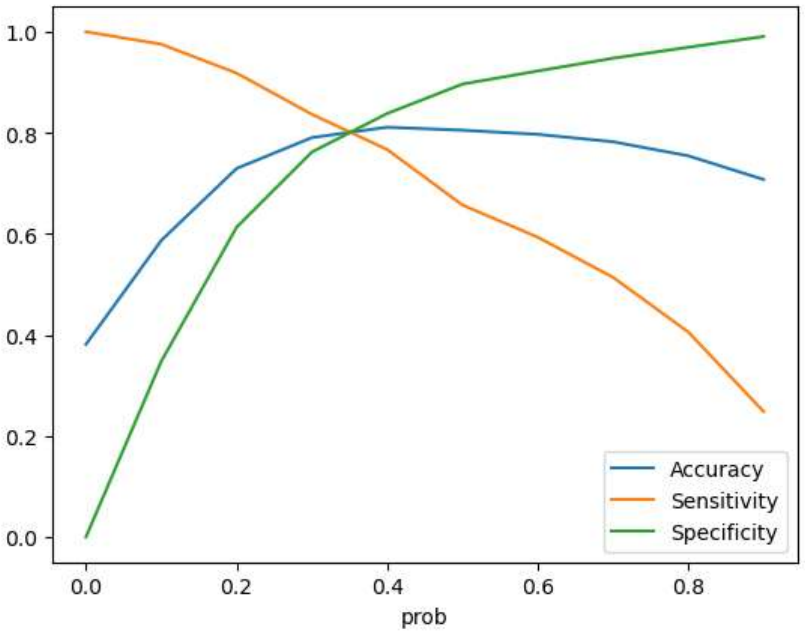
```
Recall: 0.8017
```

In [158]:
```
print("Model True Positive Rate (TPR):",round(TP/(TP + FN),5))
```

```
Model True Positive Rate (TPR): 0.8017
```

In [159]:
```
print("Model False Positive Rate (FPR):",round(FP/(FP + TN),5))
```

```
Model False Positive Rate (FPR): 0.19465
```



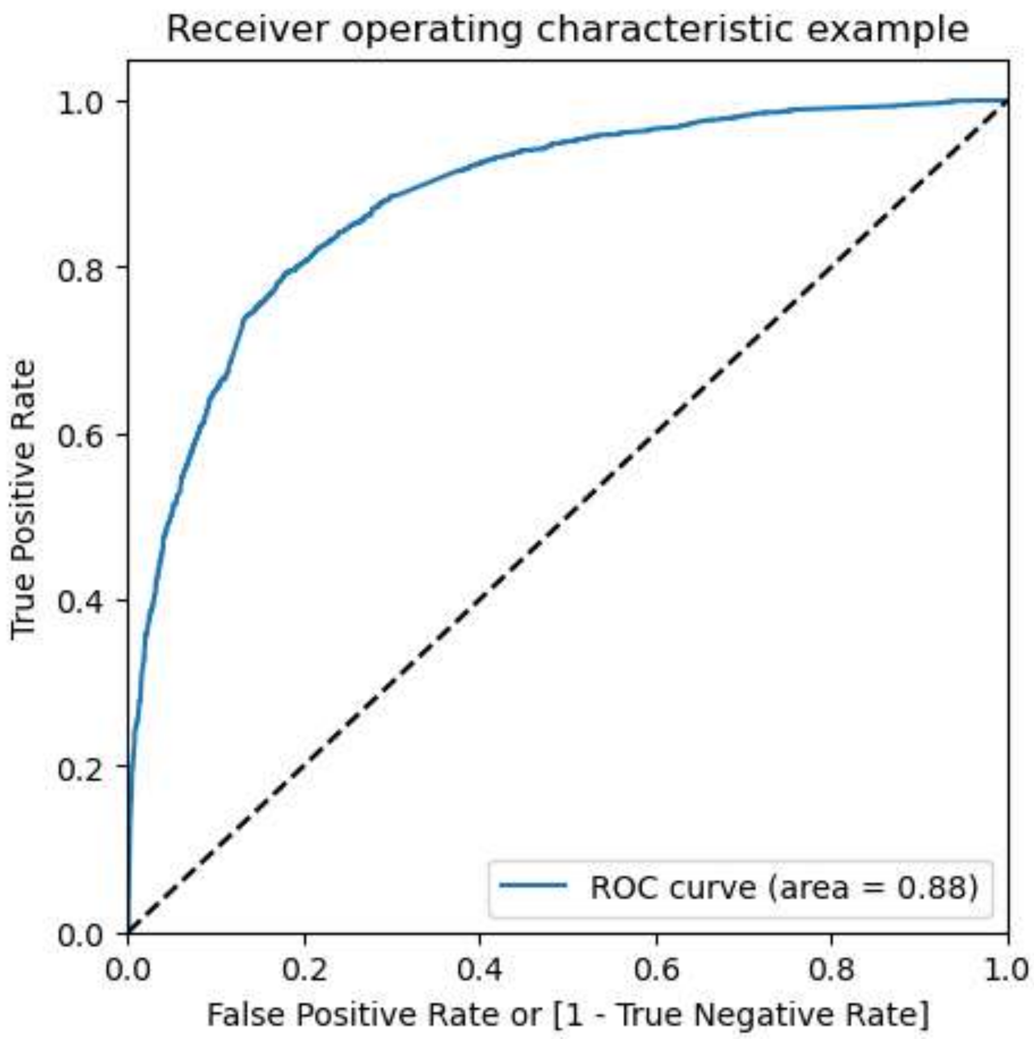
Model Evaluation

Train Data Set

ROC Curve

- ROC Curve Assessment: The Area under the ROC curve (AUC) is 0.88 out of a maximum value of 1, indicating strong predictive capability within the model.

- Curve Interpretation: The ROC curve closely approaches the top-left corner of the plot, symbolizing high true positive rates and low false positive rates across all threshold values, further affirming the model's effectiveness.



Model Evaluation

Test Data Set

Confusion Matrix & Evaluation Metrics

In [173]:
```
confusion_matrix = metrics.confusion_matrix(y_pred_final['Converted'], y_pred_final['final_predicted'])
print("Confusion Matrix")
print(confusion_matrix,"\n")
```

```
Confusion Matrix
[[1363  314]
 [ 219  876]]
```

In [175]:
```
print("accuracy:",round((TN+TP)/(TN+TP+FN+FP),5))
```

```
accuracy: 0.80772
```

In [176]:
```
print("Sensitivity:",round(TP/(TP+FN),5))
```

```
Sensitivity: 0.8
```

In [177]:
```
print("Specificity:",round(TN/(TN+FP),5))
```

```
Specificity: 0.81276
```

In [178]:
```
print("Precision:",round(TP/(TP+FP),5))
```

```
Precision: 0.73613
```

In [179]:
```
print("Recall:",round(TP/(TP+FN),5))
```

```
Recall: 0.8
```

In [180]:
```
print("Model True Positive Rate (TPR):",round(TP/(TP + FN),5))
```

```
Model True Positive Rate (TPR): 0.8
```

In [181]:
```
print("Model False Positive Rate (FPR):",round(FP/(FP + TN),5))
```

```
Model False Positive Rate (FPR): 0.18724
```

Recommendation

Boost Lead Conversion:

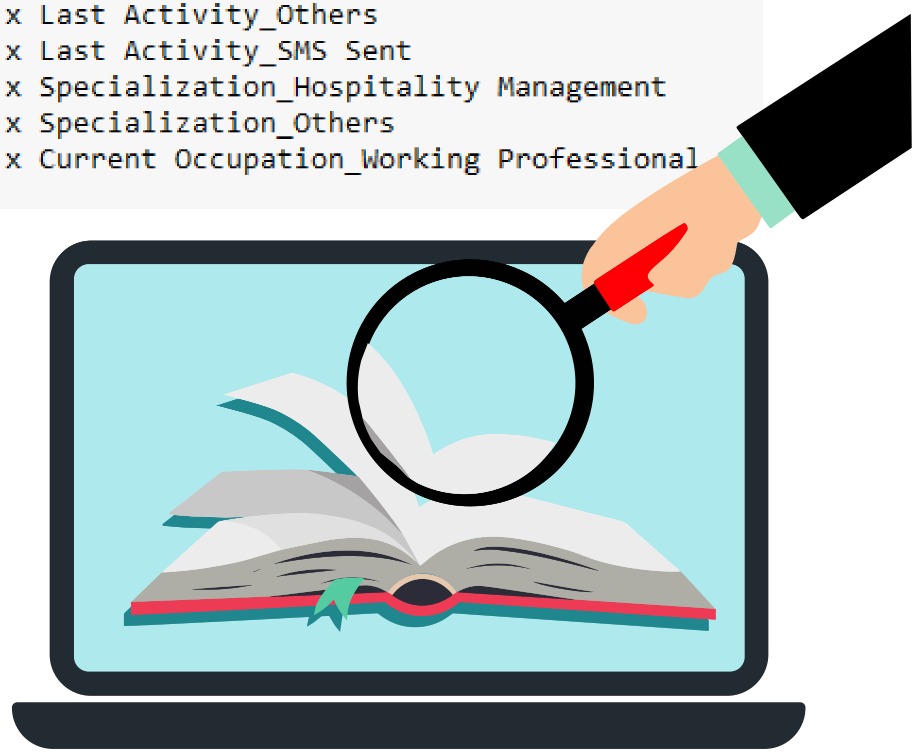
- Focus more on **Welingak Website** advertising.
- Aggressively target **working professionals** due to high conversion rates and better financial capabilities.
- Prioritize features with positive coefficients for targeted marketing.
- Customize messaging to effectively engage working professionals.
- Offer incentives for successful **referrals** to increase leads.
- Attract top-quality leads from high-performing sources.

```
Equation : =

-0.855087 x const
- 1.110290 x Do Not Email
+ 1.043936 x Total Time Spent on Website
- 1.225732 x Lead Origin_Landing Page Submission
+ 0.916386 x Lead Source_Olark Chat
+ 2.949361 x Lead Source_Reference
+ 5.476113 x Lead Source_Welingak Website
+ 0.752973 x Last Activity_Email Opened
- 0.717945 x Last Activity_Olark Chat Conversation
+ 1.408670 x Last Activity_Others
+ 1.929261 x Last Activity_SMS Sent
- 1.071914 x Specialization_Hospitality Management
- 1.194564 x Specialization_Others
+ 2.632282 x Current Occupation_Working Professional
```

Identify Improvement Areas:

- Analyze negative coefficients in specialization offerings.
- Review landing page submission process for enhancements.



Thank You!


Thank You