Lead Scoring Case Study

Done By :Sushree Sangita sahoo

Content

- Problem Statement
- Solution Approach
- Data inspection and cleaning
- Exploratory Data Analysis
- Data Preparation
- Feature Scaling
- Model building
- Plotting ROC curve
- Making predictions on the test dataset

Problem Statement

X Education markets its courses on various websites and search engines like Google, generating leads when potential customers fill out forms with their contact information or through referrals. Currently, the lead conversion rate is about 30%. The company wants to improve this by selecting the most promising leads and assigning a lead score, enabling higher-scoring leads to have a greater chance of conversion. The CEO has set a target conversion rate of around 80%.

Solution Approach

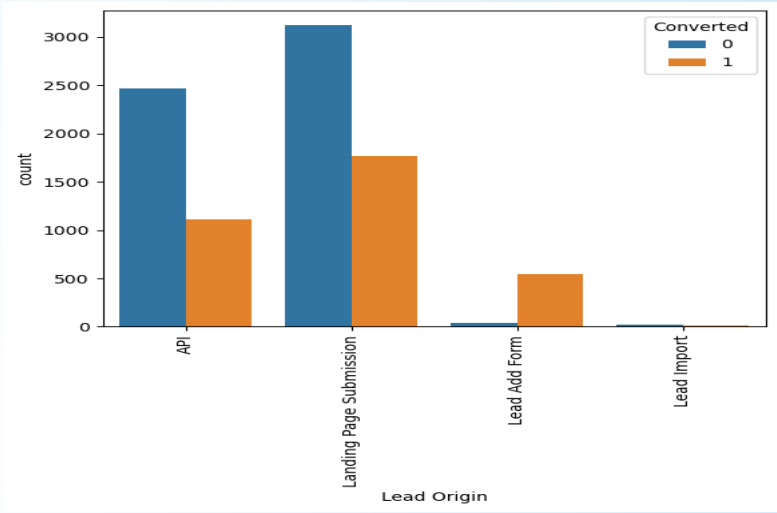
- Reading and Inspecting the Data frame
- Data Cleaning
- Exploratory Data Analysis
- Data Preparation
- Splitting Test-Train Data
- Feature Scaling
- Model Building
- Plotting the ROC Curve
- Model Evaluation
- Precision and Recall
- Making predictions on the test set

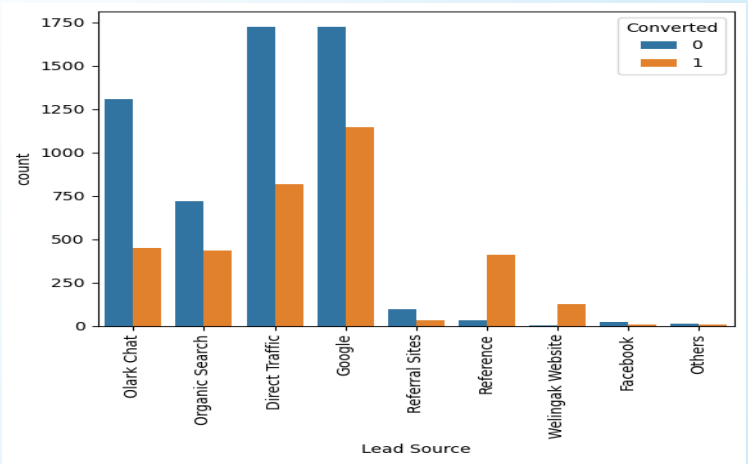
❑ Data inspection and cleaning

- we start by reviewing the overall structure of the dataset, looking at the number of rows and columns to get a sense of its size. Initially we had 9240 rows and 37.
- Check the data types of each column to ensure they are appropriate for analysis. Generate statistical analysis to get an overview of the data.
- Identify and handling missing values in the dataset. Many columns contain 'Select' values, likely indicating that customers didn’t choose an option. Since 'Select' values functionally represent missing data, we replaced them with null values. Many columns had too many missing values, so we removed those with over 3500 missing entries from our 9000 data points. The Specialization column had 37% missing values, so we added a new category called 'Others' to handle these entries. "Tag" column has 36% null values. SO we imputed it with majority of the entries.
- Checking Outliers and handling it. Using box plots, we identified outliers in columns like "Total Visits" and "Views Per Visit," and we handled them by capping their values.

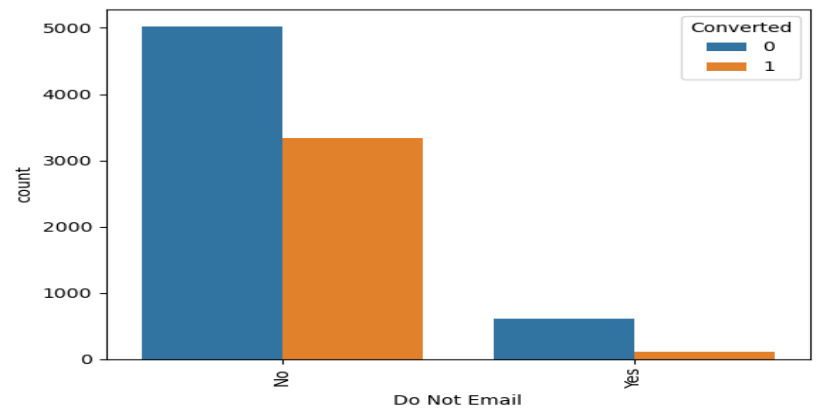
❑ Exploratory Data Analysis

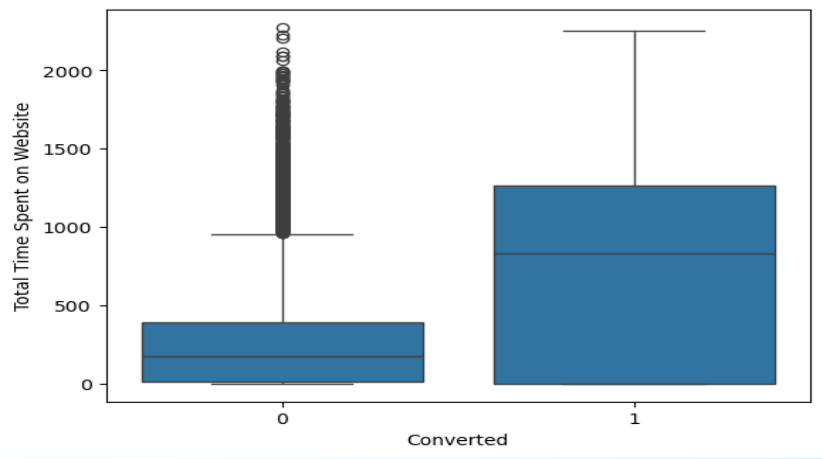
- We conducted univariate and bivariate analyses to gain a better understanding of the data and draw meaningful inferences.
- In “Lead Origin” API and Landing Page Submission generated the most leads, with a conversion rate of around 30-40%. The Lead Add Form had the highest conversion rate, but resulted in fewer leads, while Lead Imports had a very low count.
- In “Lead source” Google and Direct traffic generate the most leads. Reference leads have a high conversion rate, while leads from the Welingak website also show a strong conversion rate but have fewer leads.

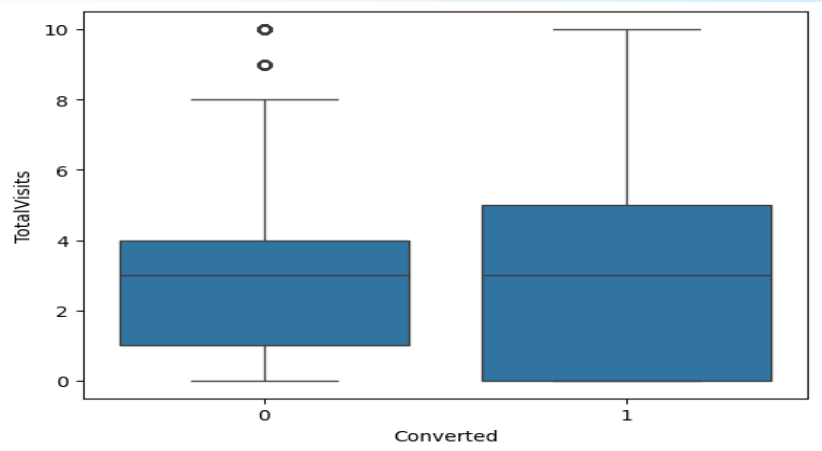




- In “Do Not Email” The number of “No” responses is much higher than “Yes”.
- In” Total Time Spent on Website” Leads who spend more time on the website are more likely to convert.
- In “TotalVisit”, While we can't make strong conclusions, the number of visits is an important variable that can enhance the likelihood of conversion. Interested leads tend to visit multiple times, further increasing their chances of converting. the median values for both converted and not converted leads are nearly the same.

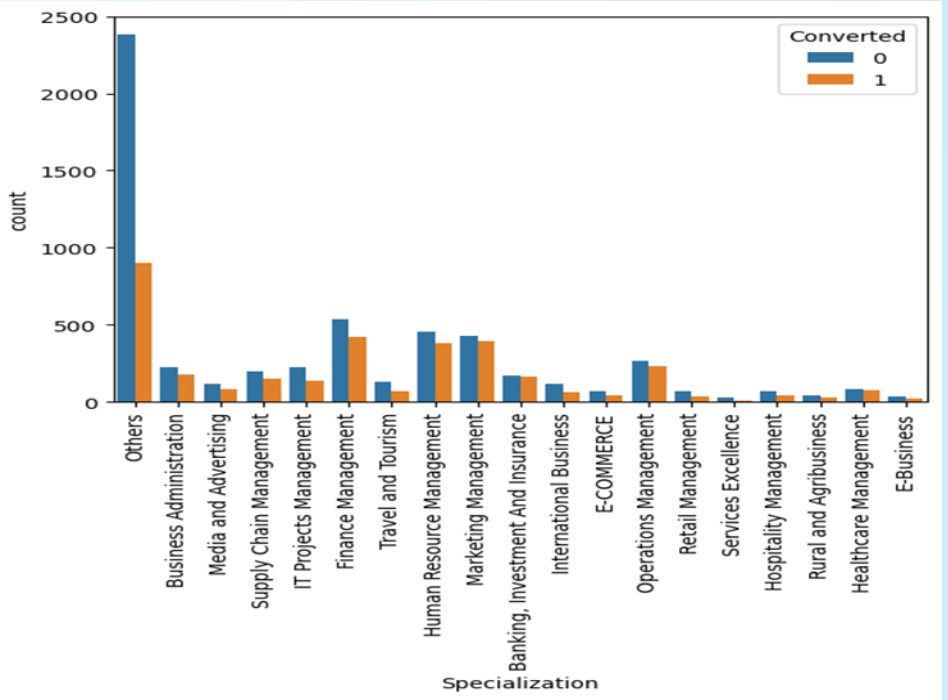


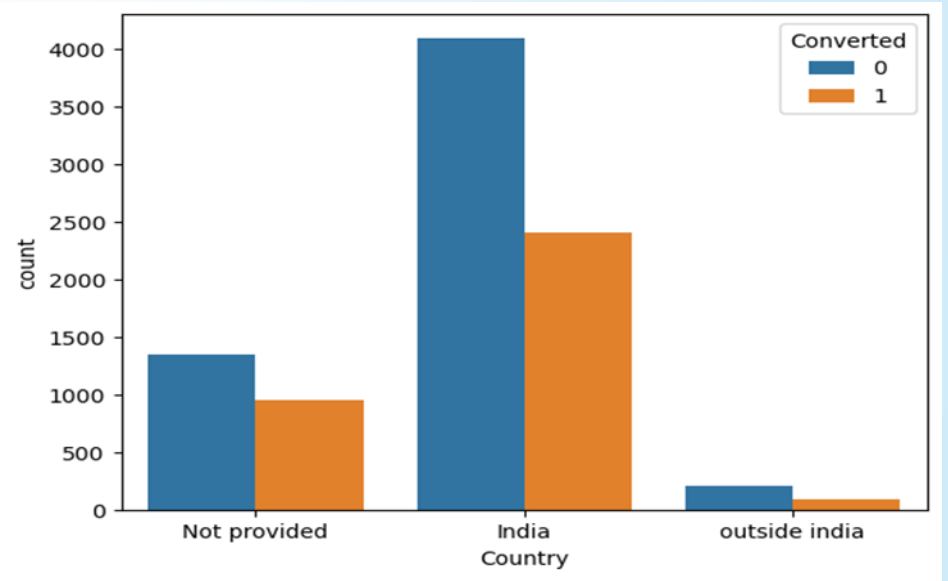




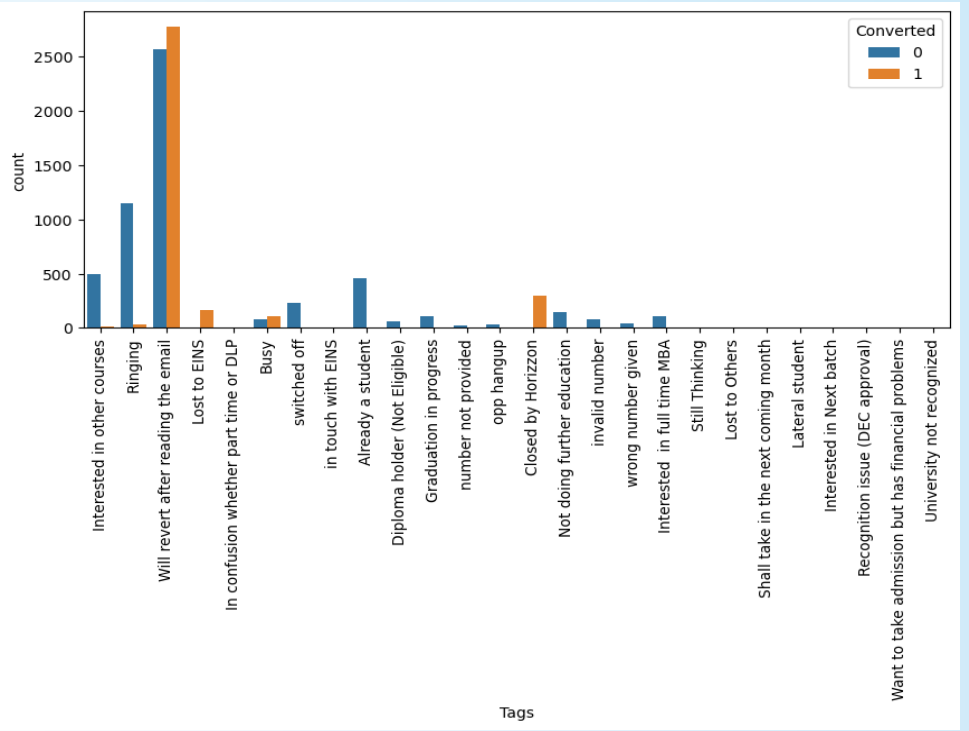
- In “Specialization ”, Human Resource Management, Marketing Management, and Operations Management show high conversion rates. Although Health Care Management, Banking, and Investment and Insurance also have high conversion rates, their total number of leads is quite small,the Others category, which includes those who did not provide a specialization, has the highest number of leads.

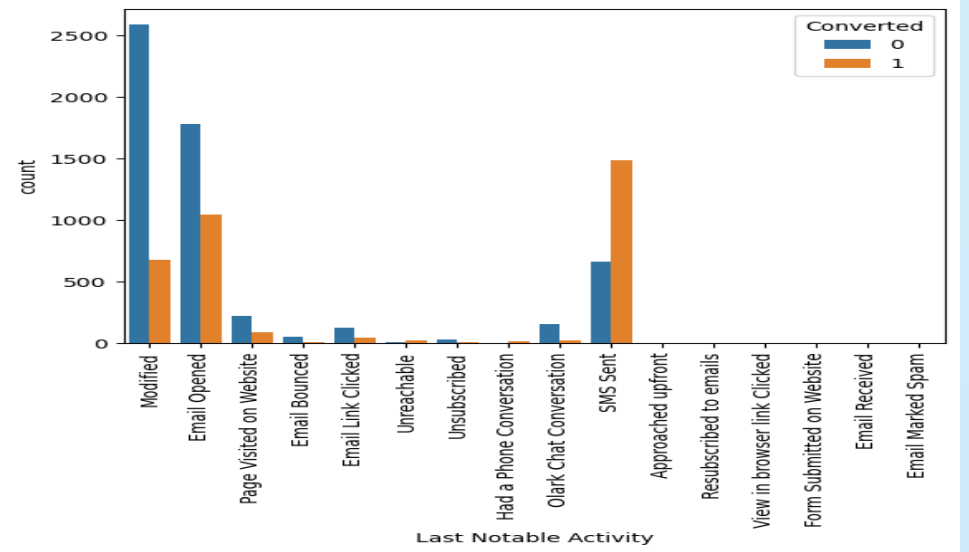
- In “Country” India has the highest number of leads, while the count from outside India is very low. Additionally, a significant number of leads did not provide their country name.





- “Will revert after reading the email” has the highest number of leads and conversion rate. Although “Lost to EINS” and “Closed by Horizzon” also have high conversion rates, they have very few leads. According to the data dictionary, this column indicates the current status of the lead assigned to them.
- In”Last Notable Activity ” the “SMS Sent” category has the highest conversion rate, while “Modified” and “Email Opened” have the highest number of leads.





Data Preparation:

- Created Dummy variables for the categorical features.
- Dropped the repeated variables for which dummies are created.
- Converting some binary variables (Yes/No) to 1/0.

Splitting Test-Train Data

- Splitting the dataset into training and test sets to train the model and evaluate its performance on unseen data.

Feature Scaling

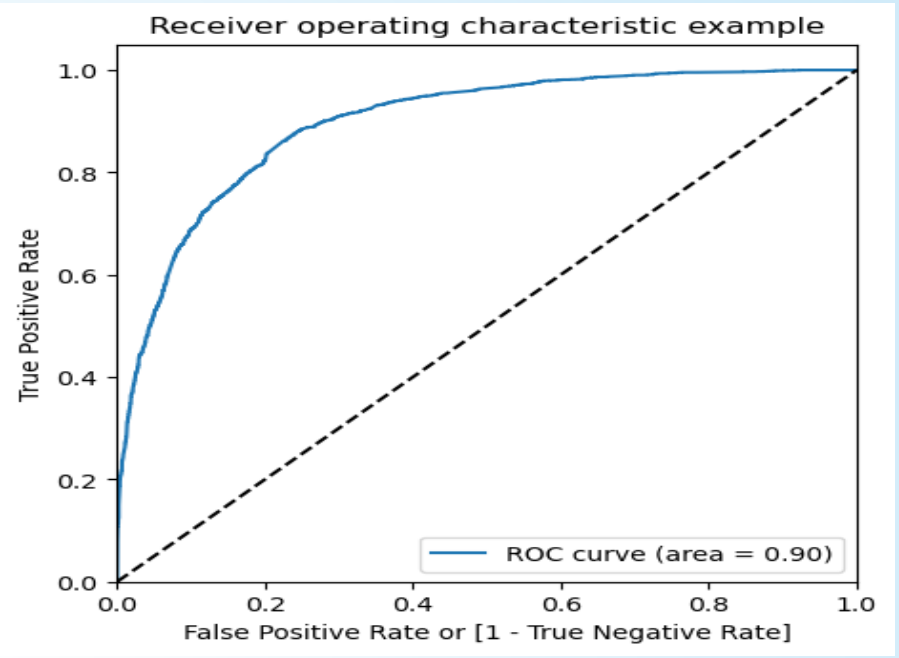
- Performed feature scaling in my logistic regression model to ensure that all features were on a similar scale, which helps improve the model's accuracy and convergence during training.

❑ Model Building

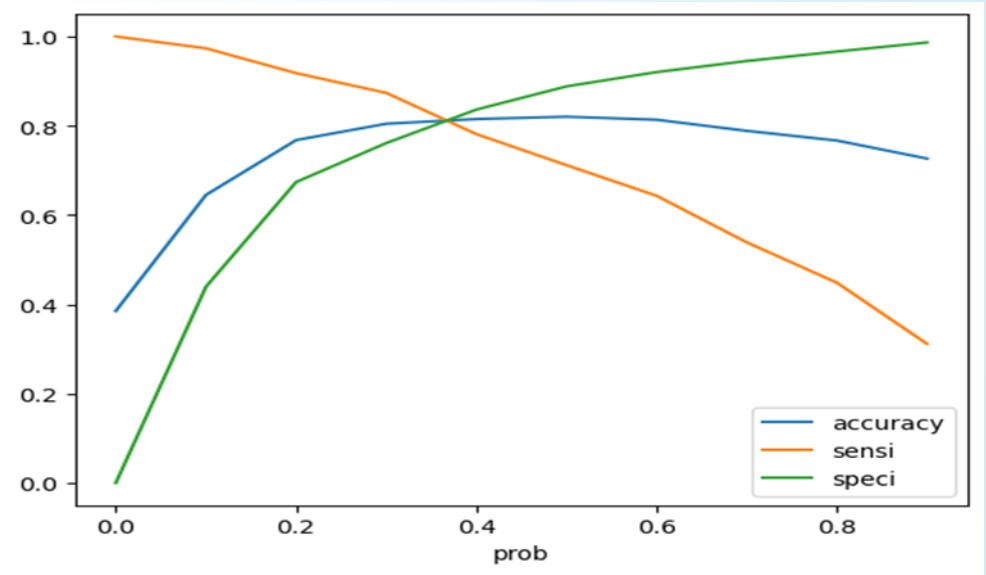
- **Feature Selection Using RFE:**Selected 20 variables using Recursive Feature Elimination (RFE) to focus on the most important features for my model.
- Evaluated the model with StatsModels by iteratively building it and reviewing the summary for p-values and VIFs, removing variables until I achieved a model with p-values below 0.05 and VIFs under 4.
- The final model included 16 variables after the evaluation and selection process.
- made Prediction on the Train set, Created a dataframe with the actual Converted flag and the predicted probabilities.
- An arbitrary cut-off probability point was chosen to determine the predicted labels.

❑ Plotting the ROC Curve

A high area under the ROC curve (0.90) indicates that the model performs well.



This plot is used to visualize and compare accuracy, sensitivity, and specificity across different probability cut-offs.
From the plot 0.4 is the optimum point to take it as a cutoff probability.



❑ Model Evaluation

The metrics at optimal cutoff point 0.4 are :

Accuracy:81%

Sensitivity:78%

Specificity:84%

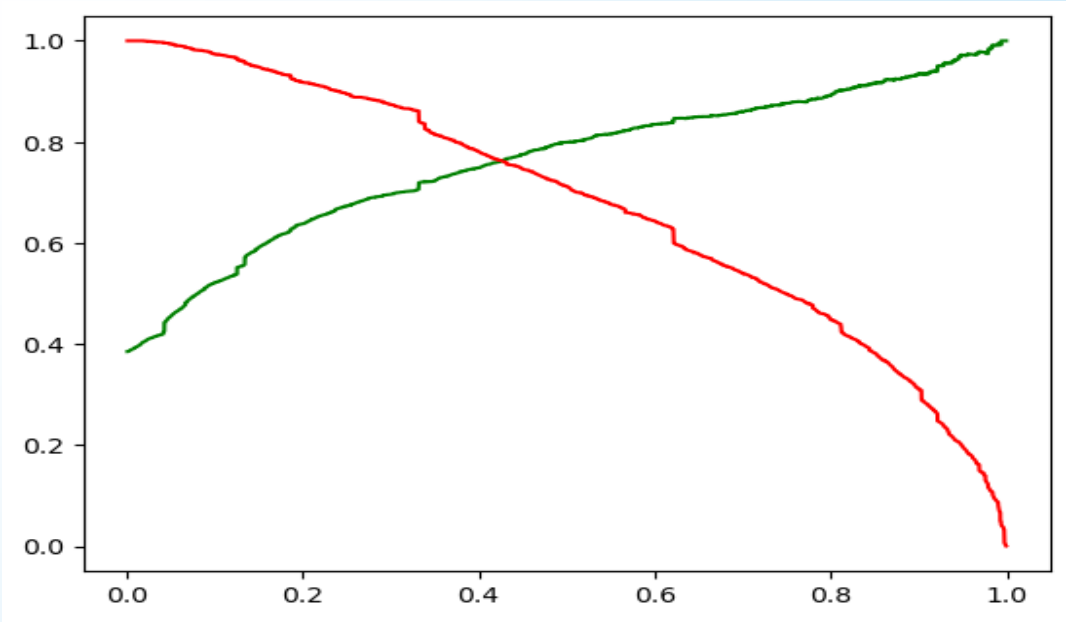
False Positive rate: 16%

❑ Precision and recall

Precision:80%

Recall:71%

In Precision and recall plot 0.4 looks opimal.



❑ Making predictions on the test dataset

The predictions for the test data set are consistent with those from the training set supporting 80% conversion rate with:

Accuracy:81 %

Sensitivity: 77%

Specificity:84%

❑ Recommendations:

- Prioritize calling leads from "Welingak Websites" and "Reference" sources for higher conversion chances.
- Focus on contacting "working professionals" due to their higher likelihood of conversion.
- Prioritize contacting leads from the “Olark Chat” source, as these leads are more likely to convert.
- Engage leads whose last activity was an "SMS sent" as they are more likely to convert.
- Deprioritize calling leads who selected "Do not Email" as they are unlikely to convert.
- Deprioritize leads whose origin is "Landing Page Submission," as they are unlikely to convert.
- Do not call leads whose last activity was "Olark Chat Conversation," as they are unlikely to convert.
- By focusing on high-probability leads and minimizing unnecessary calls, the sales team can utilize their resources more effectively while still engaging with potential customers.

THANK YOU