LEAD SCORE CASE STUDY

Presented by

Tummalacheruvu Baji Babu
Kriti Tiwari
Aysushi Tyagi

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For e.g., they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, *also known as 'Hot Leads'*
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective:

- **X education wants to know most promising lead**
- **Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads.**

ANALYSIS APPROACH

❑ Data Overview

- This dataset contains a variety of attributes related to lead generation and conversion, including:
- **Lead Source:** The origin of the lead.
- **Total Time Spent on Website:** The duration of the user's website visit
- **Total Visits:** The number of times the user has visited the website
- **Last Activity:** The most recent interaction with the lead (e.g., email open, click)

❑ Analysis Methodology

1. Data Preparation

- **Data Cleaning:** Address missing values and inconsistencies to ensure data integrity.
- **Data Transformation:** Convert variables between numerical and categorical formats as needed.
- **Outlier Detection and Handling:** Identify and address outliers that may skew analysis results.

ANALYSIS APPROACH (CONT..)

2. Exploratory Data Analysis (EDA)

- **Univariate Analysis:** Examine individual variables to understand their distributions, central tendencies, and variability.
- **Bivariate Analysis:** Investigate the relationships between pairs of variables to identify correlations and patterns.
- **Visualization:** Use appropriate visualizations.

3. Feature Engineering

- **Feature Scaling:** Standardize numerical variables to ensure comparable scales.
- **Feature Encoding:** Convert categorical variables into numerical representations.

4. Model Selection and Training

5. Model Evaluation

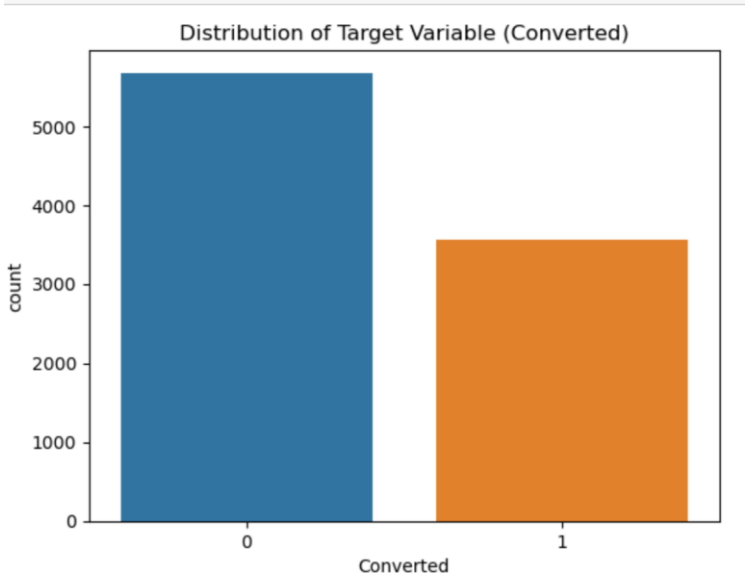
6. Conclusion and Recommendation

DATA MANIPULATION

- Total Number of Rows =9240, Total Number of Columns =37.
- 'Select' value in the feature is consider as Null,
- Dropping the columns having more than 30% as missing value such as 'How did you hear about X Education' and 'Lead Profile'.
- After checking for the value counts for some of the object type variables, we find some of the features having approx. 99% of data of one type only, i.e. no enough variance, we have dropped those features, such that:

"I agree to pay the amount through cheque", "Get updates on DM Content", "Update me on Supply Chain Content", "Receive More Updates About Our Courses", "Through Recommendations", "Newspaper", "Digital Advertisement" etc.

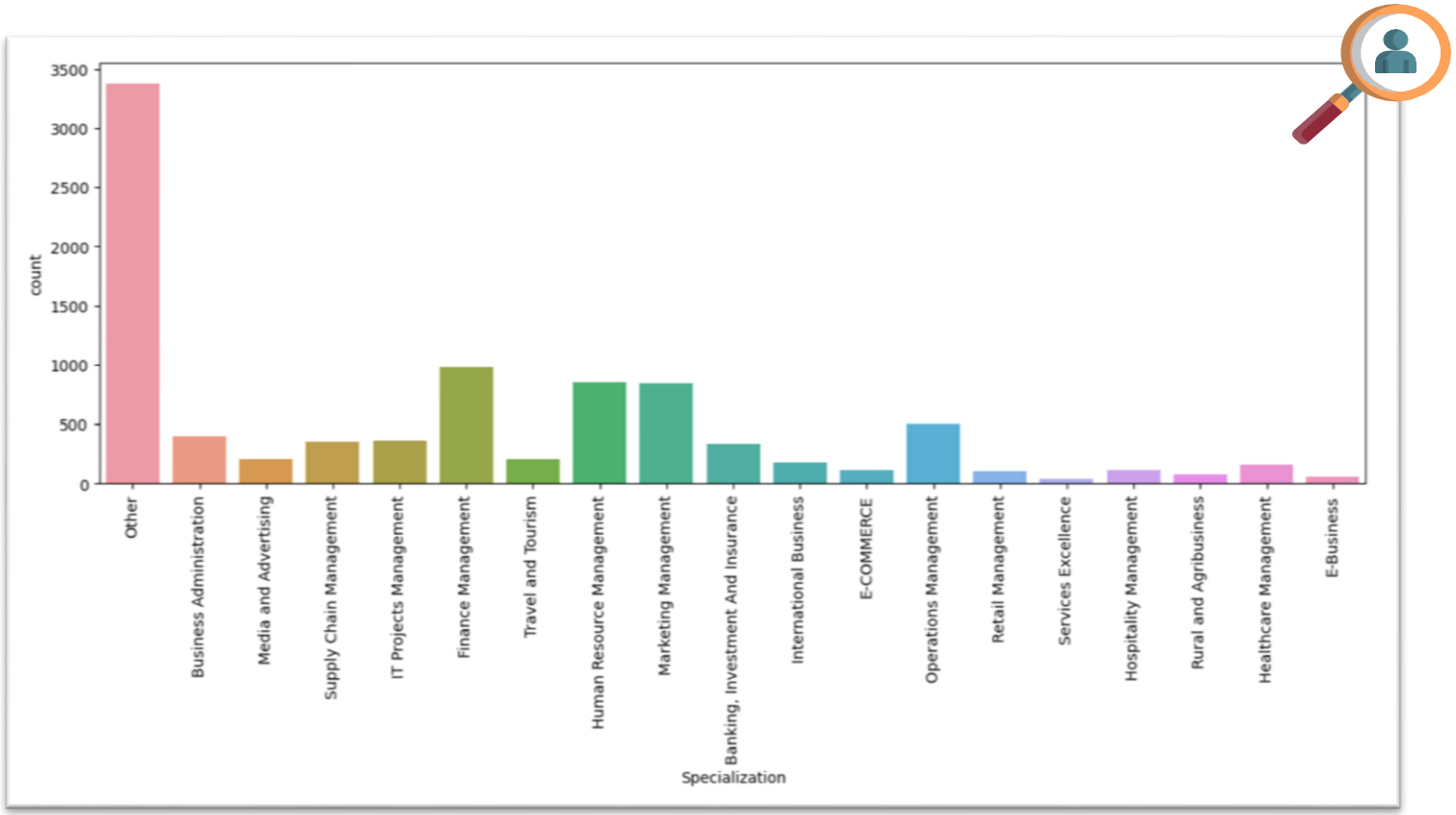
- Removing the "Prospect ID" and "Lead Number" which is not necessary for the analysis
- Features which having less 30% null values, for that we impute with median in case of numerical column and mode in case categorical column.

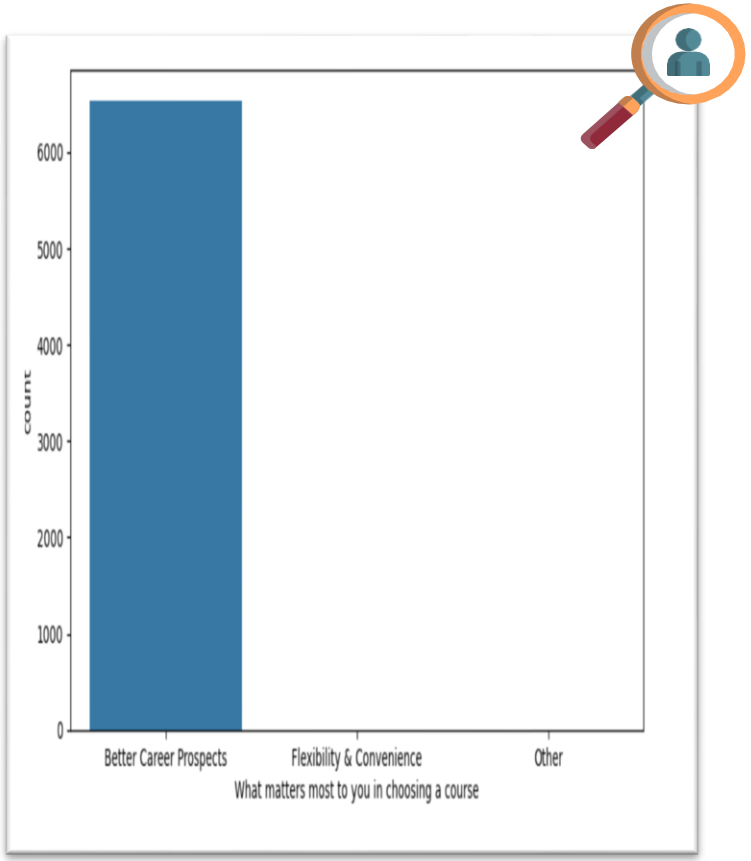


Interpretation:
The taller bar indicates that the majority of observations (around 5000+) fall into this category.
The shorter bar represents fewer observations (around 2500+) that fall into this category.
Conversation rate is low compare to total number leads company acquired.

UNIVARIATE ANALYSIS

ANALYSIS & FINDINGS





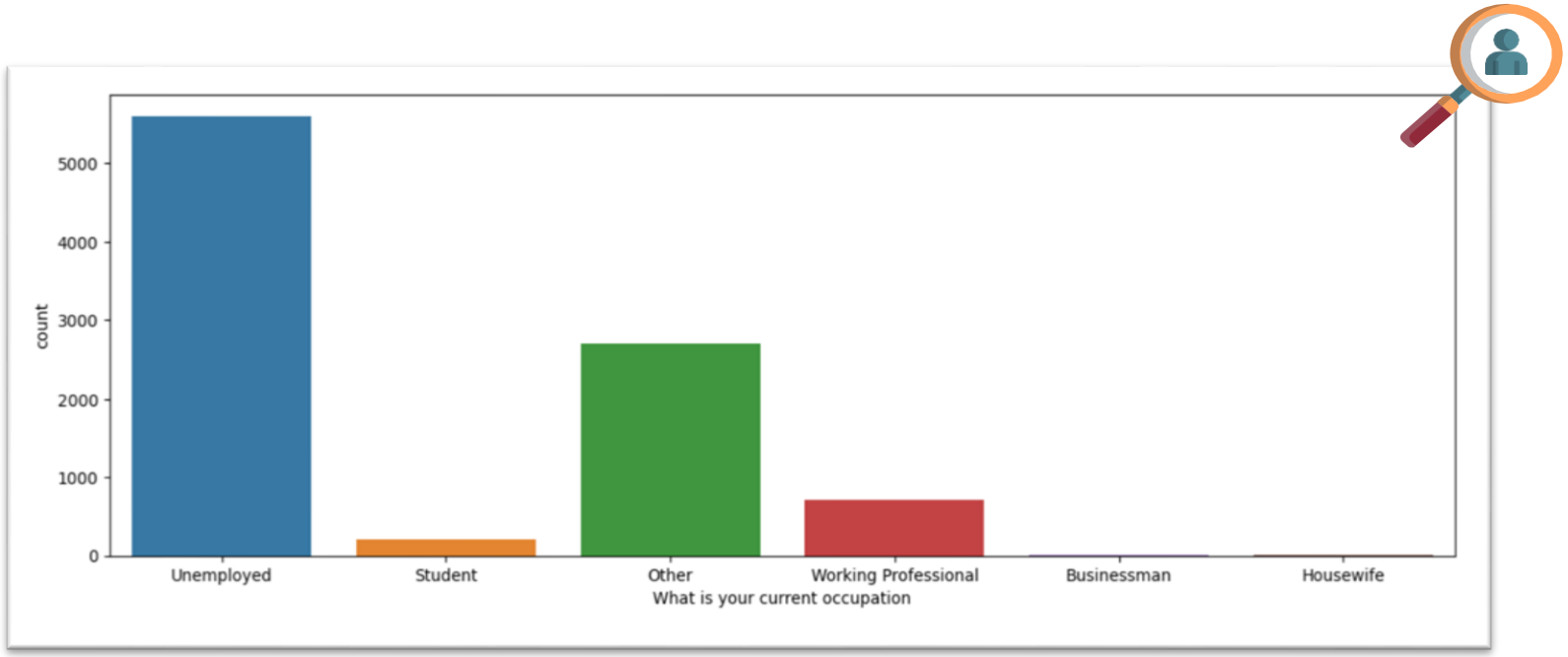
Insights

The bar chart shows a significant concentration in the "Other" specialization, with over 3,500 entries. "Finance Management," "Human Resource Management," and "Marketing Management" follow, each with 500-1,000 occurrences. Most other specializations, like "Business Administration" and "Operations Management," have moderate to low representation.

Insights

“Better Career Prospects" is the most important factor for most people when choosing a course.

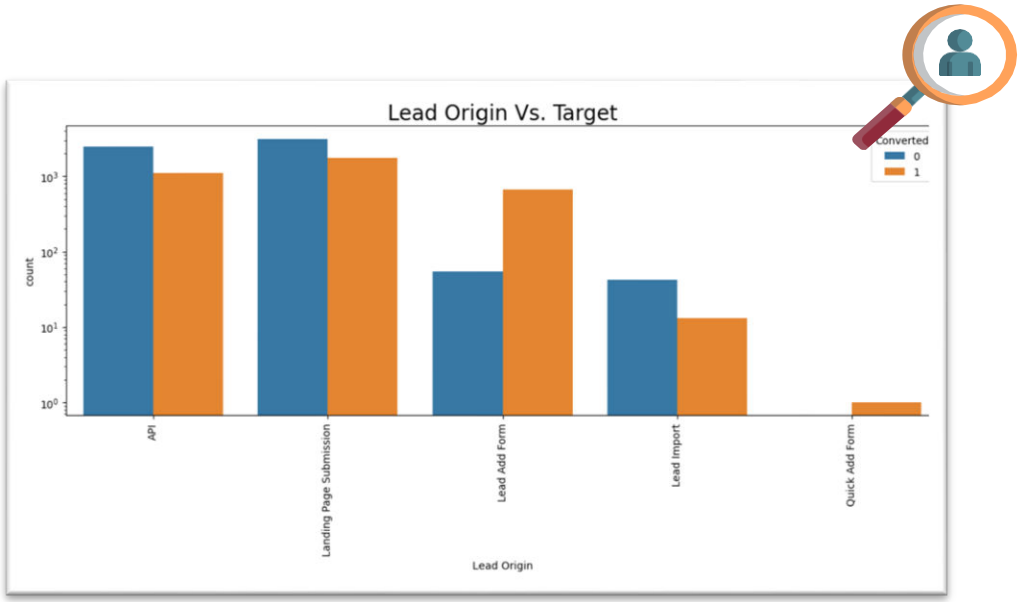
ANALYSIS & FINDINGS

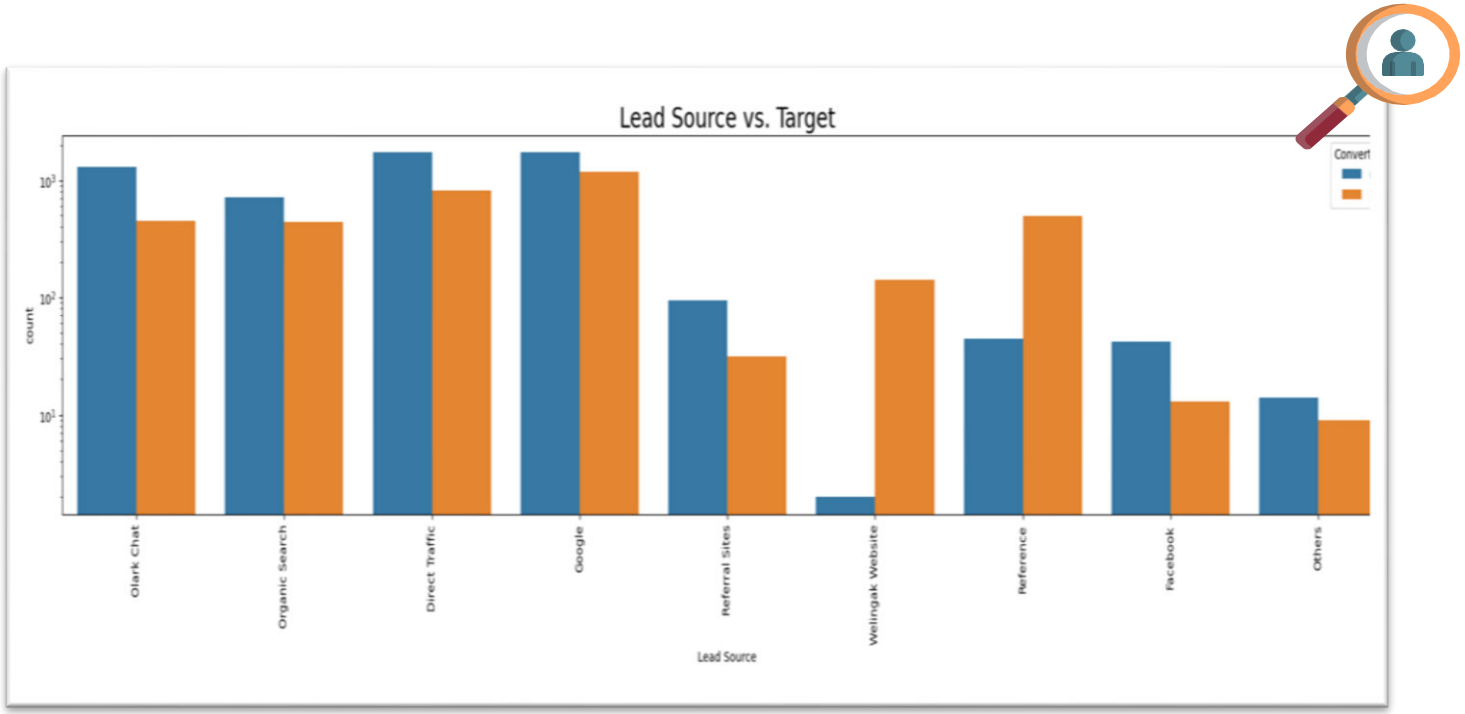


Insights

The graph shows the distribution of respondents' current occupations. The most common occupation is "Unemployed," followed by "Student" and "Other." The remaining categories have relatively low counts.

ANALYSIS WITH TARGET





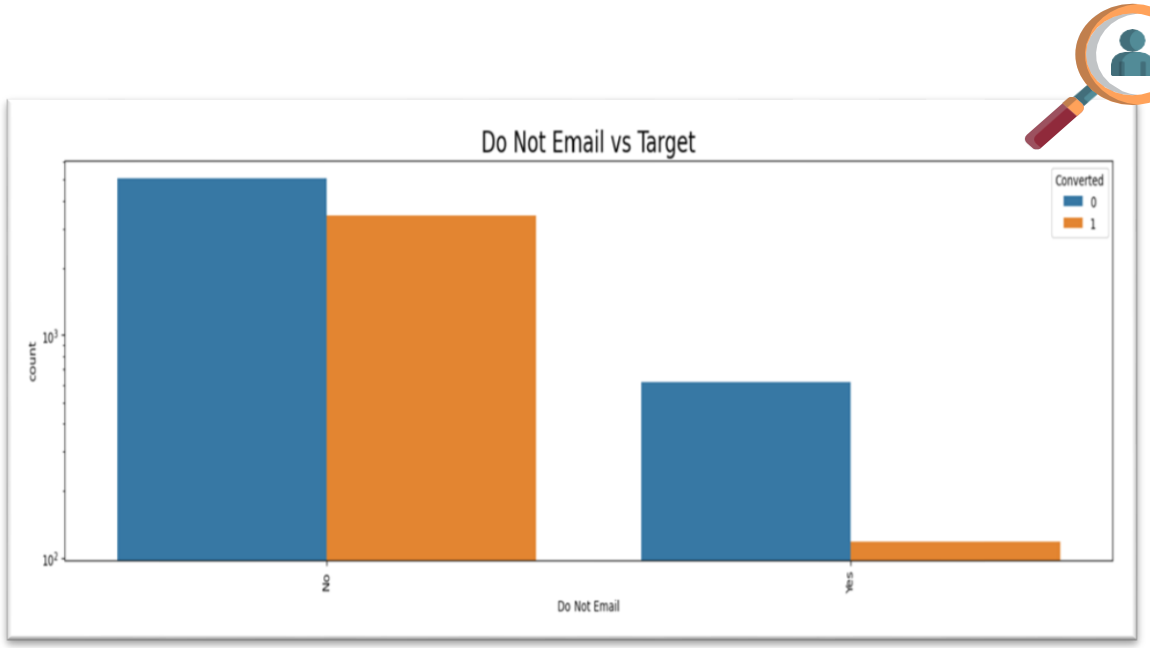
Insights

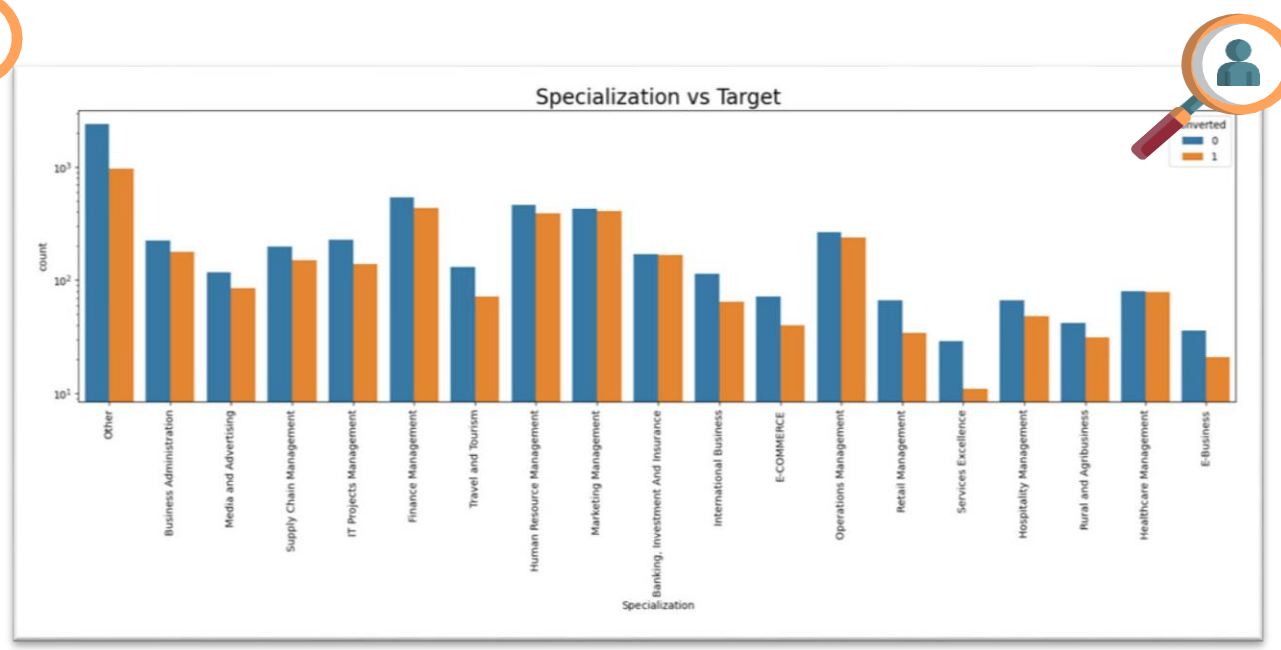
The graph shows the distribution of lead conversions by lead origin. Conversion rate for 'API' and 'Landing Page Submission' having more leads and conversion also. For 'Lead Add Form' number of conversion is more than unsuccessful conversion., conversion rates vary significantly across different lead origins.

Insights

The graph shows lead conversion rates by source. Google and Direct traffic generates maximum number of leads. Conversion rate of 'Reference' and 'Welingak Website' leads is high. Conversion varies significantly across sources.

ANALYSIS WITH TARGET





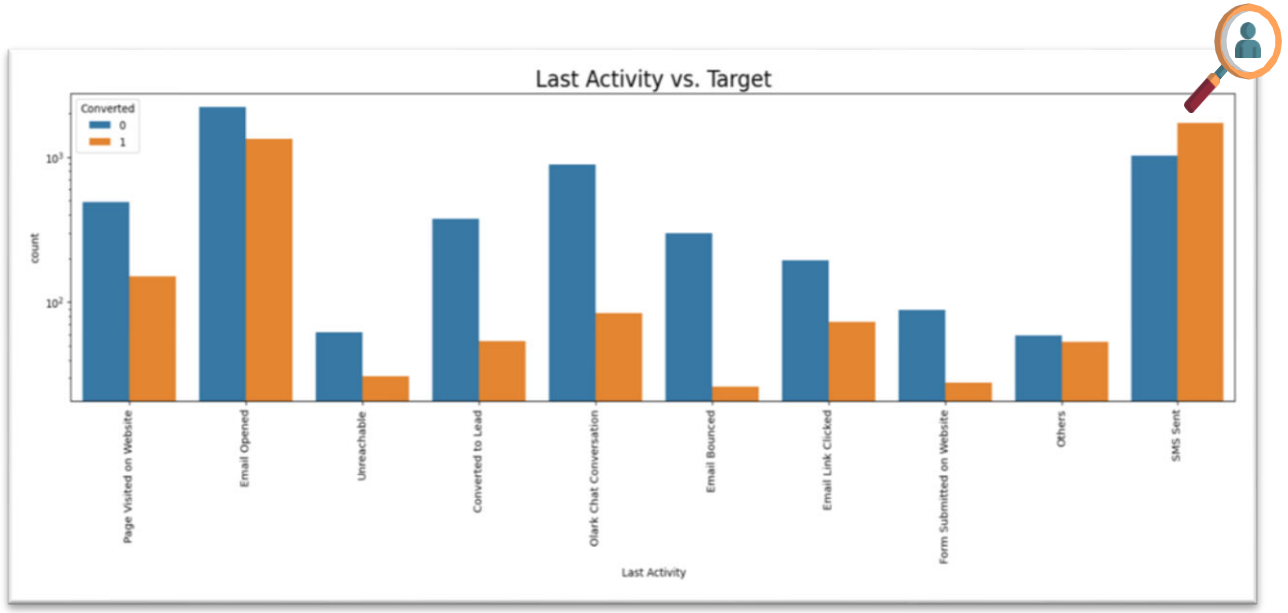
Insights

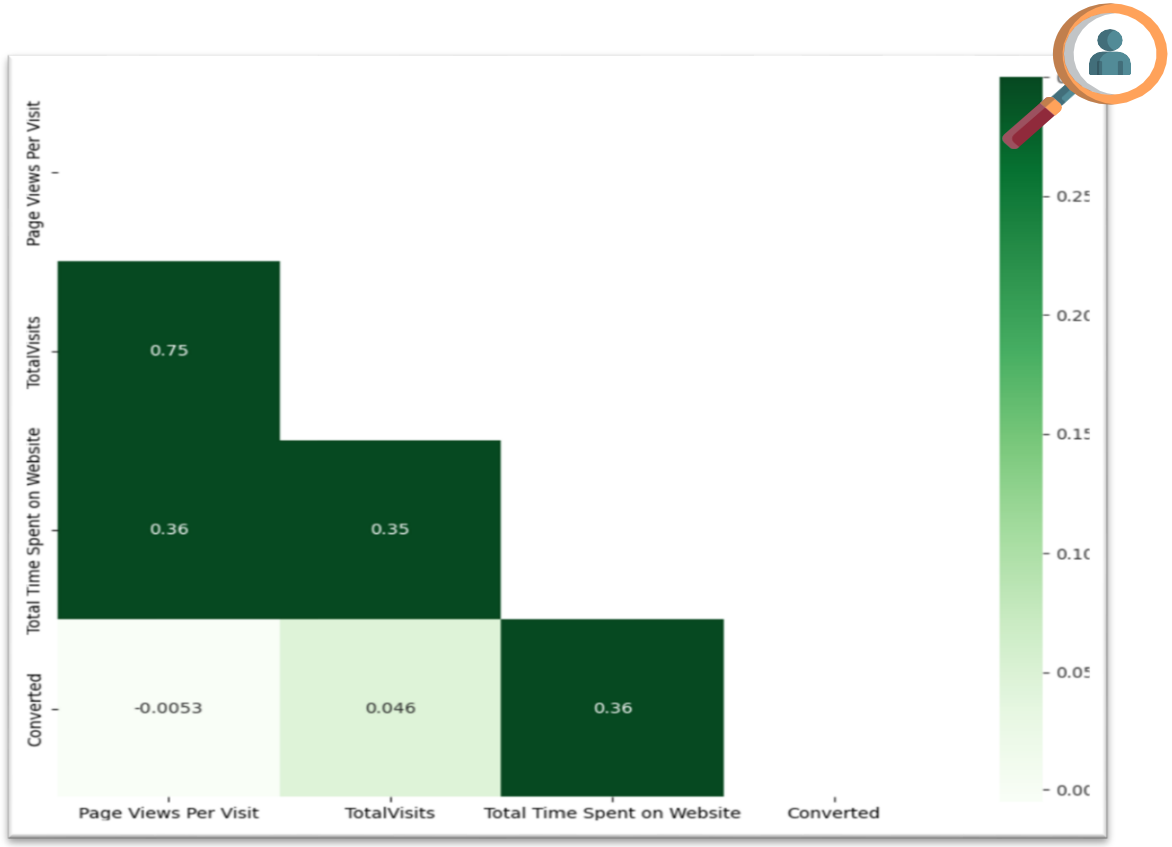
The graph shows the distribution of lead conversions based on whether they were sent an email ("Do Not Email"). The "no" group has a higher conversion rate than the "yes" group. This suggests that sending emails to leads might be negatively impacting conversion rates. However, further analysis is needed to draw definitive conclusions.

Insights

The graph shows lead conversions by specialization. Conversion rates vary significantly. 'Other' has highest leads but low conversion. Retail, Services, and Rural have low leads and conversions. Business and Media have moderate leads and conversions.

ANALYSIS WITH TARGET





Insights

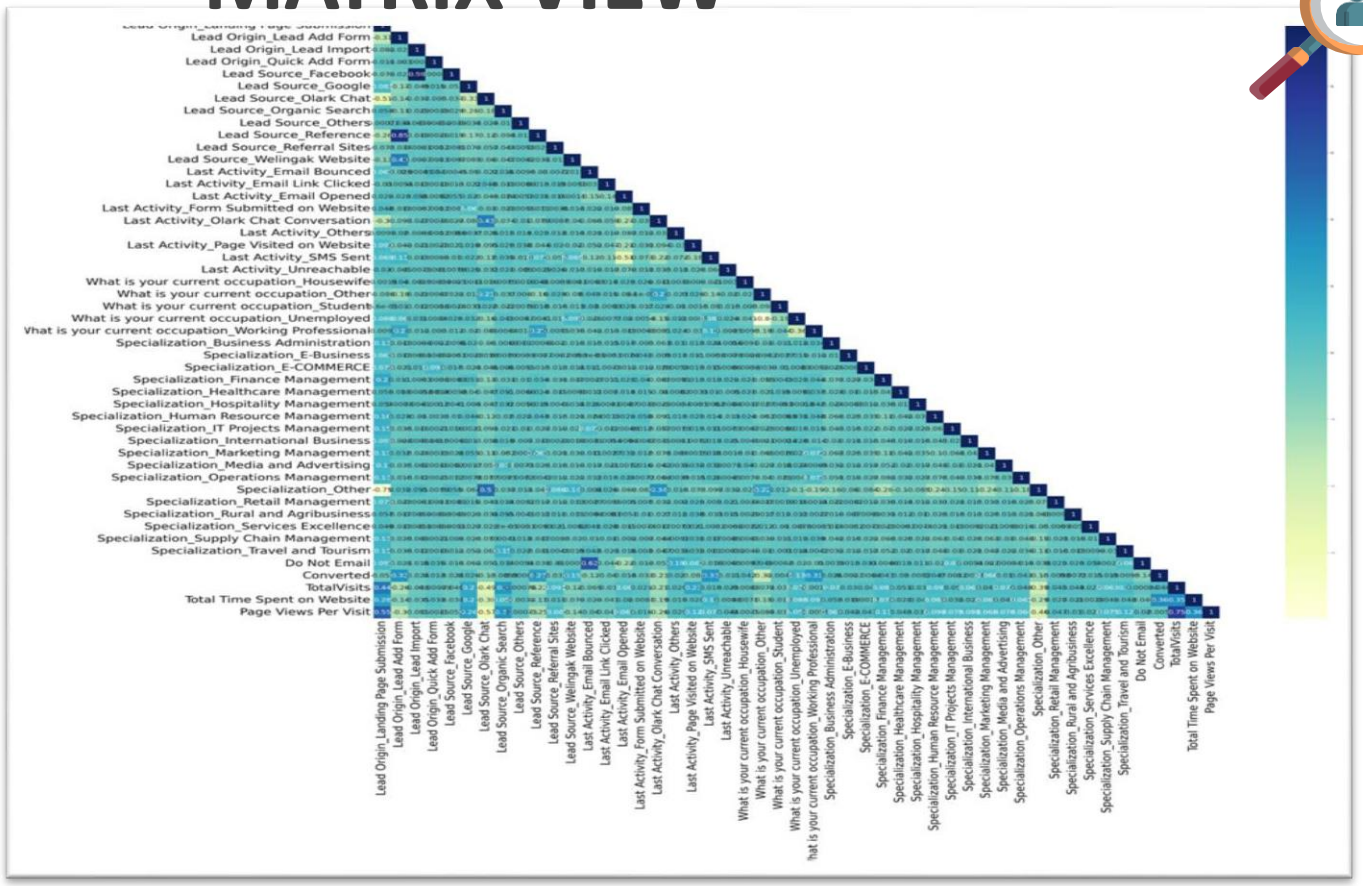
The graph shows lead conversions by last activity. Conversion rate for last activity of 'SMS Sent' higher. Highest last activity for leads is 'Email Opened'.

Insights

The graph shows correlation among features. 'TotalVisits' and 'Page Views per Visit' are highly correlated with correlation of .75. 'Total Time Spent on Website' has correlation of 0.36 with target variable 'Converted'.

CORRELATION MATRIX

MATRIX VIEW



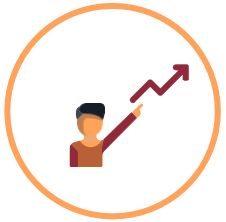
Insights

The graph shows the correlation matrix between different variables.

- 'Lead Source_Facebook' and 'Lead Origin_Lead Import' having higher correlation of 0.98.
- 'Do Not Email' and 'Last Activity_Email Bounced' having higher correlation. approx. .62.
- 'Lead Origin_Lead Add Form' and 'Lead Source_Referance' having higher correlation of 0.85.
- 'TotalVisits' and 'Page Views Per Visit' having correlation of 0.75.
- Lead Origin_Lead Add Form' ,'Lead Source_Welingak Website', 'Last Activity_SMS Sent' and 'What is your current Occupation_Working Professionals' having positive correlation with our target variable 'Converted'.

.

MODEL BUILDING AND EVALUATION



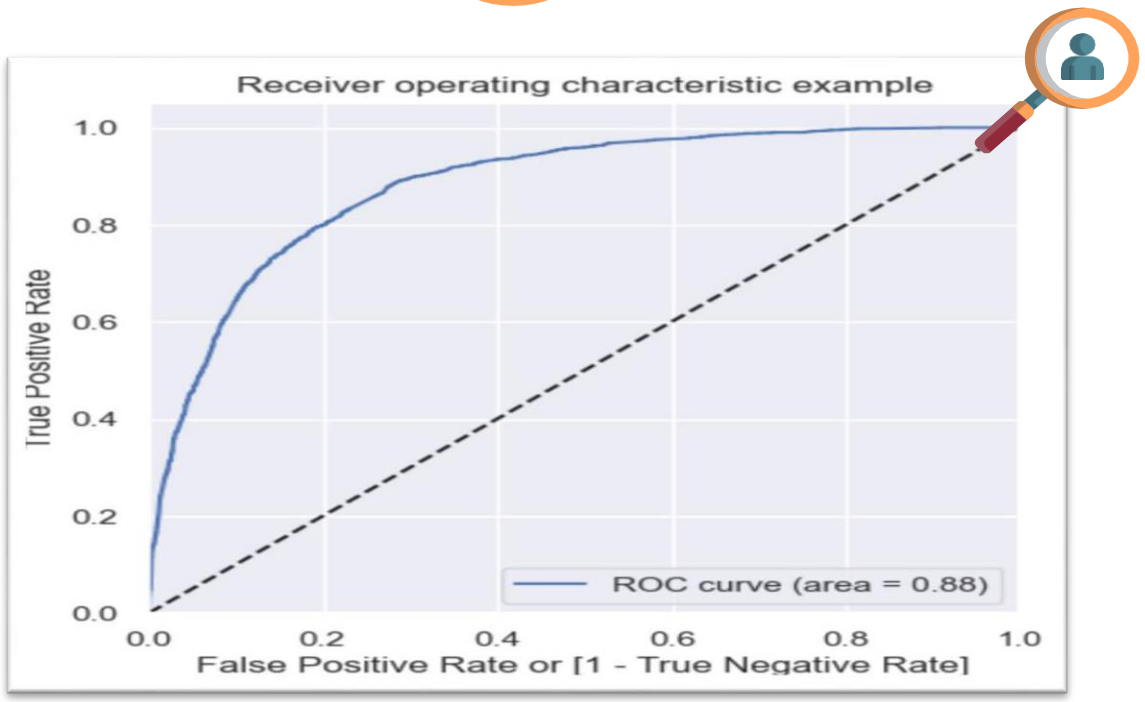
DATA CONVERSION AND DUMMY VARIABLE CREATION

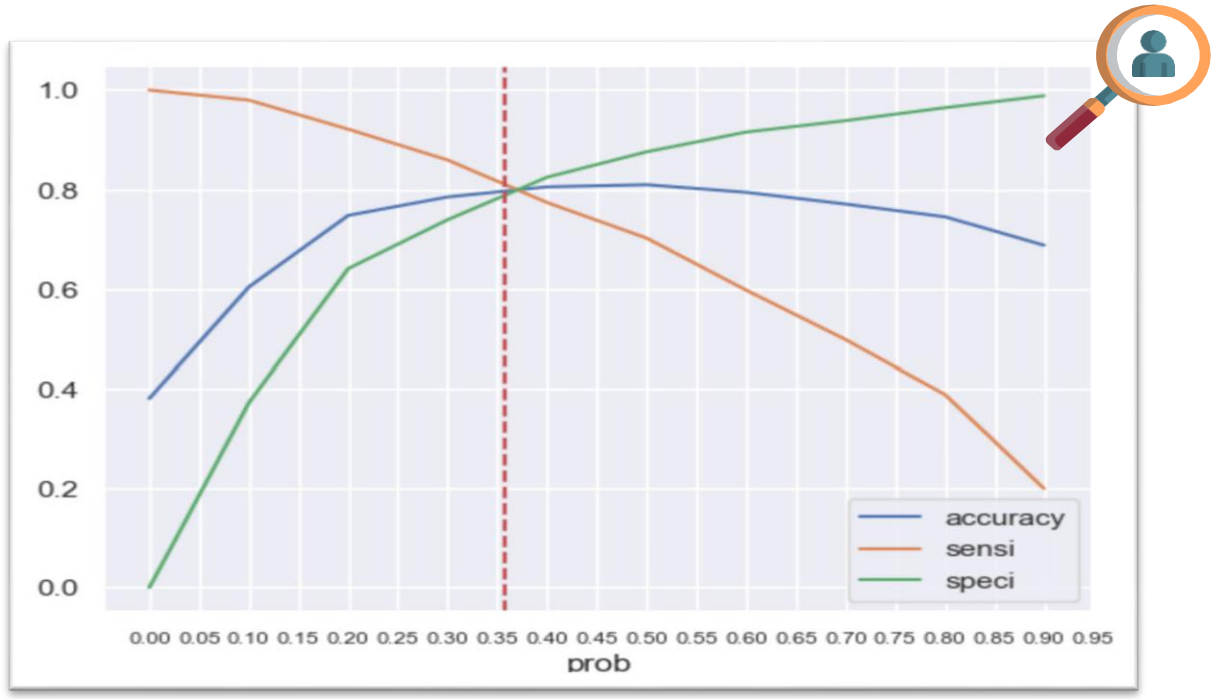
- converting some binary variables (Yes/No) to 0/1 - 'Do Not Email'
- For categorical variables with multiple levels, create dummy features like 'Lead Origin', 'Lead Source', etc.
- Total Rows for Analysis: 9240
- Total Columns for Analysis: 49

MODEL BUILDING

- Splitting the Data into Training and Testing Sets (70% Train dataset, 30% Test Dataset)
- Use RFE for Feature Selection (15 features selected using RFE)
- Building Model by removing the variable whose p- value is greater than 0.05 and VIF > 3
- In Model 10 we have all features with p-value < 0.05 and VIF < 3
- Predictions on test data set
- Evaluation Metrics for the test Dataset:-
- Accuracy : 0.80
- Sensitivity: ~ 0.81
- Specificity: 0.79
- Precision: 0.71
- Recall: 0.81

INSIGHTS





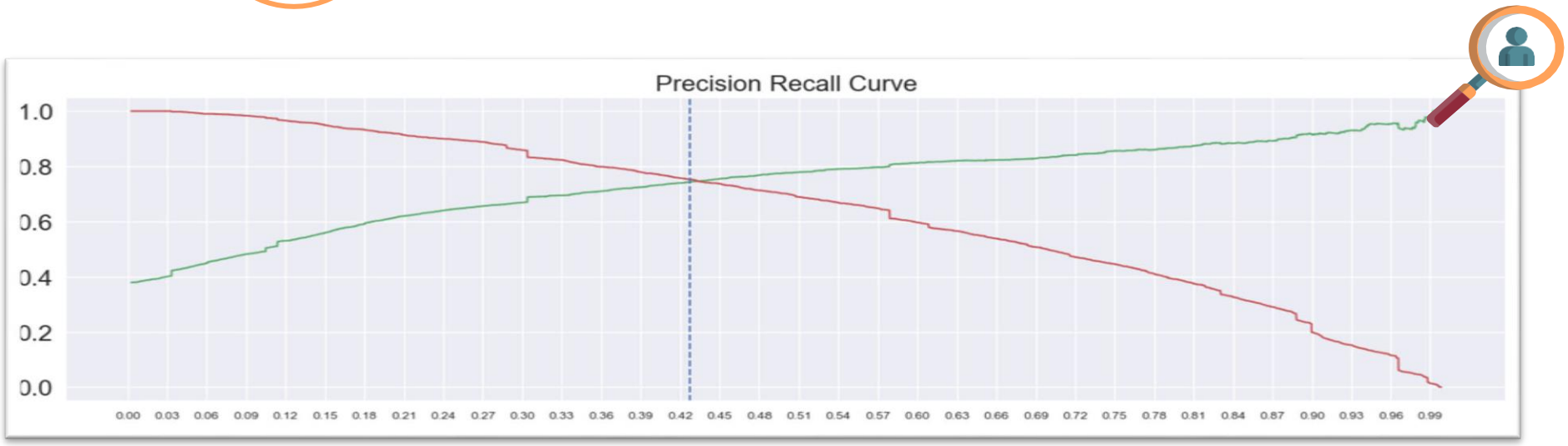
Insights

The graph shows a receiver operating characteristic (ROC) curve with an area under the curve (AUC) of 0.88. This indicates good overall classification performance. The curve is closer to the top-left corner, suggesting a high true positive rate and low false positive rate.

Insights

Graph shows that, optimal cut-off probability is .35 on which we have maximum accuracy, sensitivity and specificity. We choose this probability for prediction.

INSIGHTS



Insights

The graph shows a precision-recall curve. From above 'precision_recall_curve' we can see that cutoff point is 0.42. By using the Precision - Recall trade off curve cut off point, True Positive number has decrease and True Negative number has increase Thus, we cannot use Precision-Recall trade-off method as it reduced True Positive so 'Recall'/'sensitivity' decreased. We have to increase Sensitivity Recall value to increase True Positives. Thus we will use 0.35 as optimal cutoff point.

SUMMARY

- The most influential variables in identifying potential leads are:
- **Lead Origin:** Add Form
- **Lead Source:** Welingak Website
- **Last Activity:** SMS Sent
- **Total Time Spent on Website**
- These factors appear to have the strongest correlation with lead conversion.

THANK YOU