Lead Scoring Case Study

- To Build a Logistic Regression Model to predict & Identify the most potential leads/Hot Leads, which In turn increases the lead conversion rate.

By : Dhaval Gala

Business Objective

- To help X Education to select the most promising leads(Hot Leads) : the leads that are most likely to convert into paying customers.

- To build a logistic regression model to assign a lead score value between 0 and 100 to each of the leads which can be used by the company to target potential leads.

Approach

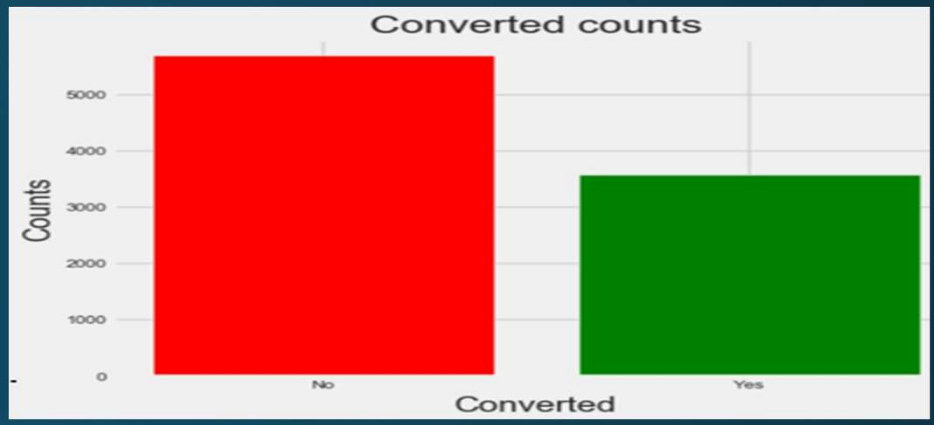
No	Method	Activities
1	Data Cleaning & Data Manipulation	➢ Duplicate data Handling ➢ NA Handling & Missing Values ➢ Drop Unnecessary Columns ➢ Outlier Handling ➢ Imputation of Values
2	EDA	➢ Univariate data analysis: value count, distribution of variable ➢ Bivariate data analysis: correlation coefficients and pattern between the variables
3	Feature Scaling	➢ Dummy Variable treatment
4	Model Building	➢ Logistic Regression Model for model making and Prediction through GLM Model ➢ Validation & Presentation of Model
5	Conclusion	➢ Identify Variables which are influencing model ➢ Recommendation

Data Manipulation

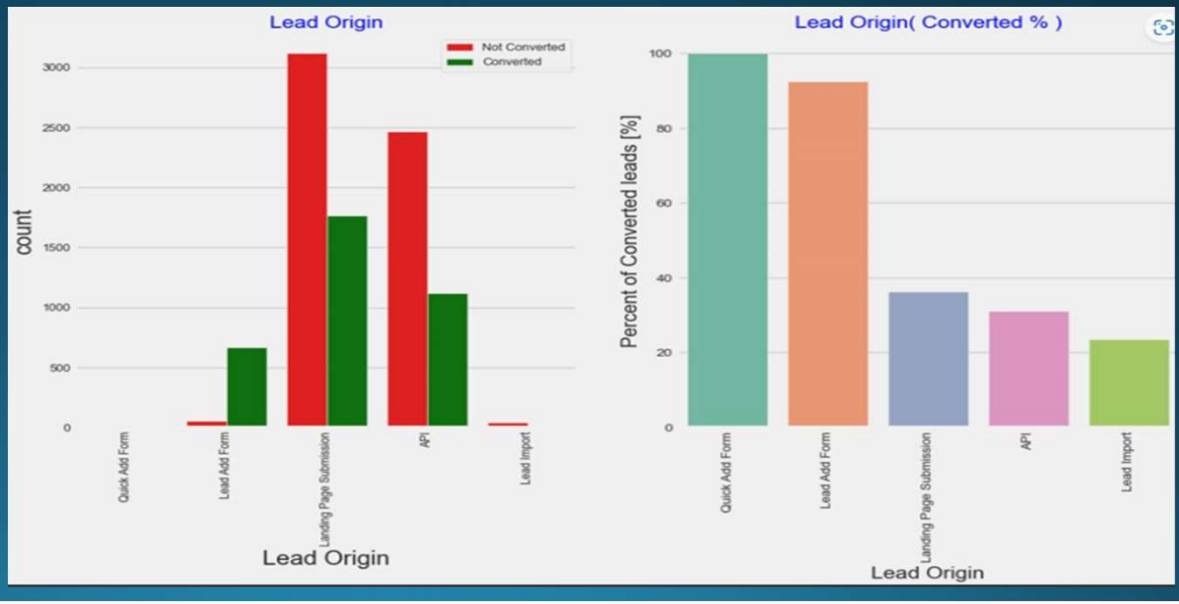
- Checking for percentage of leads converted and not converted in the given data. Dropped Columns which are having least variance of data.
- Drop numerical Columns having high null percentage(>35 %).
- Removing the "Prospect ID" and "Lead Number" which is not necessary for the analysis
- As per the data 'select' which means the customer had not selected this option while filling the form Is as good as NaN . So we replaced NaN values with Select for " lead profile ", "specialization" and "city" without dropping these columns. If we drop these columns, we will loose 40% of data.
- Applied Outlier treatment.

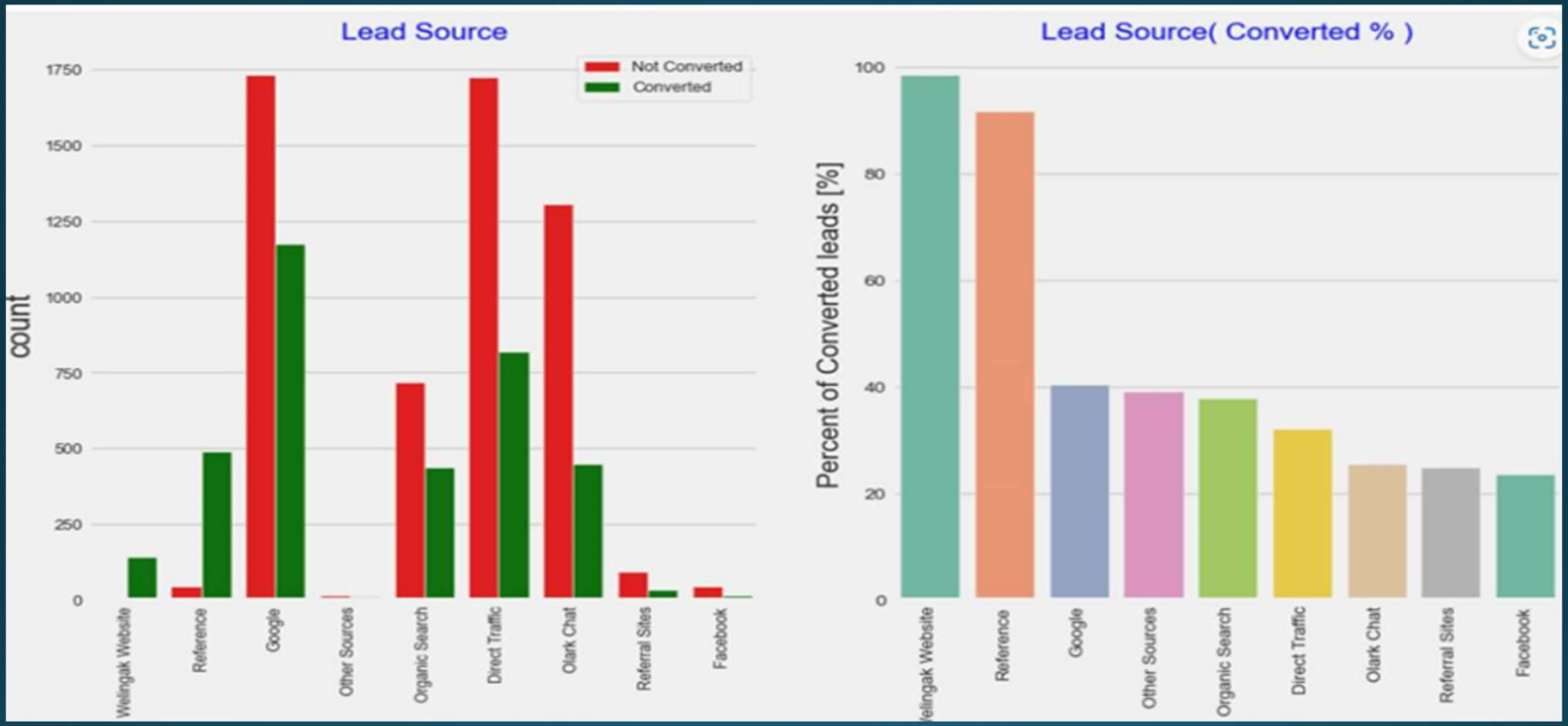
Exploratory Data Analysis

In the lead conversion ratio, 38.5% of visitors turned to leads, whereas 61.5% did not. As a result, It appears to be a well-balanced dataset

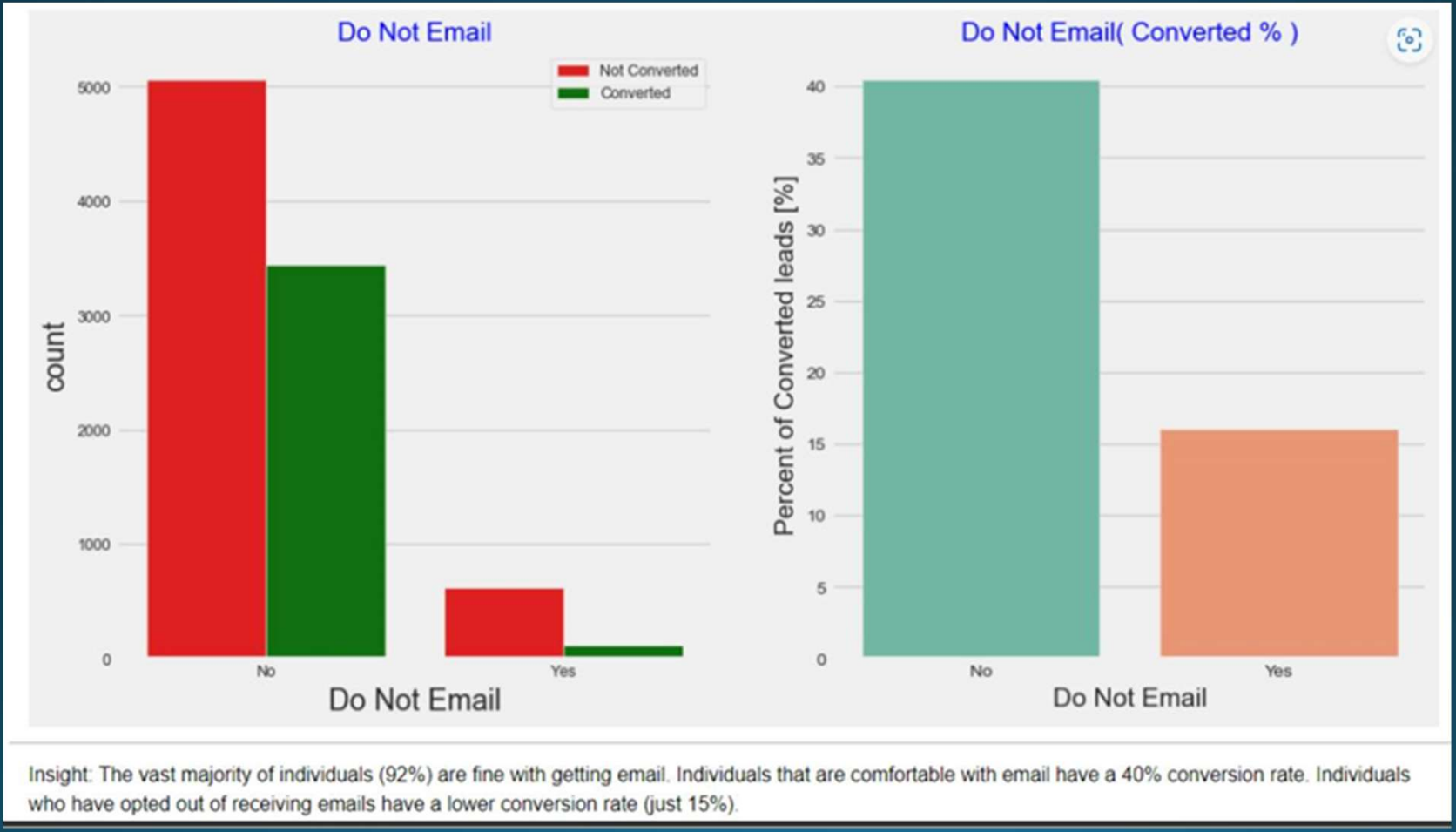


In Lead Origin, maximum conversion happened from Landing Page Submission

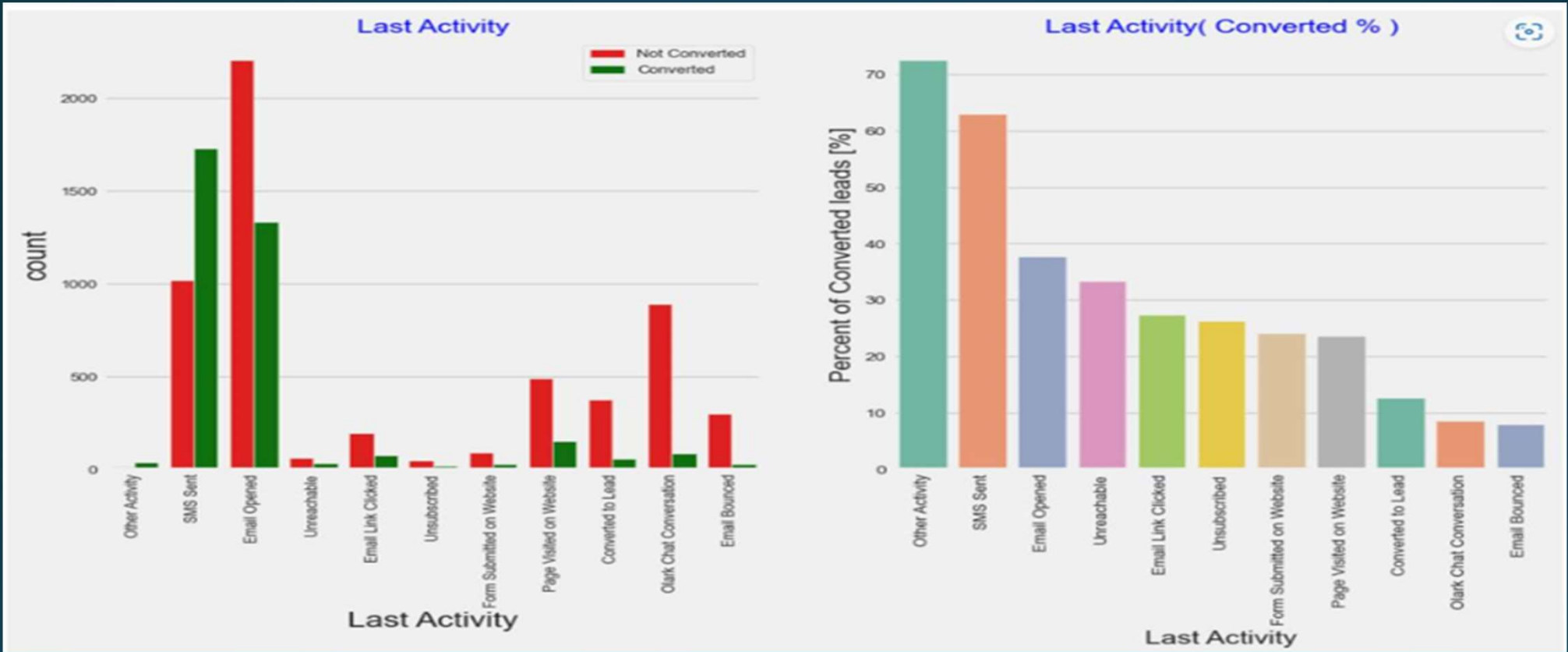




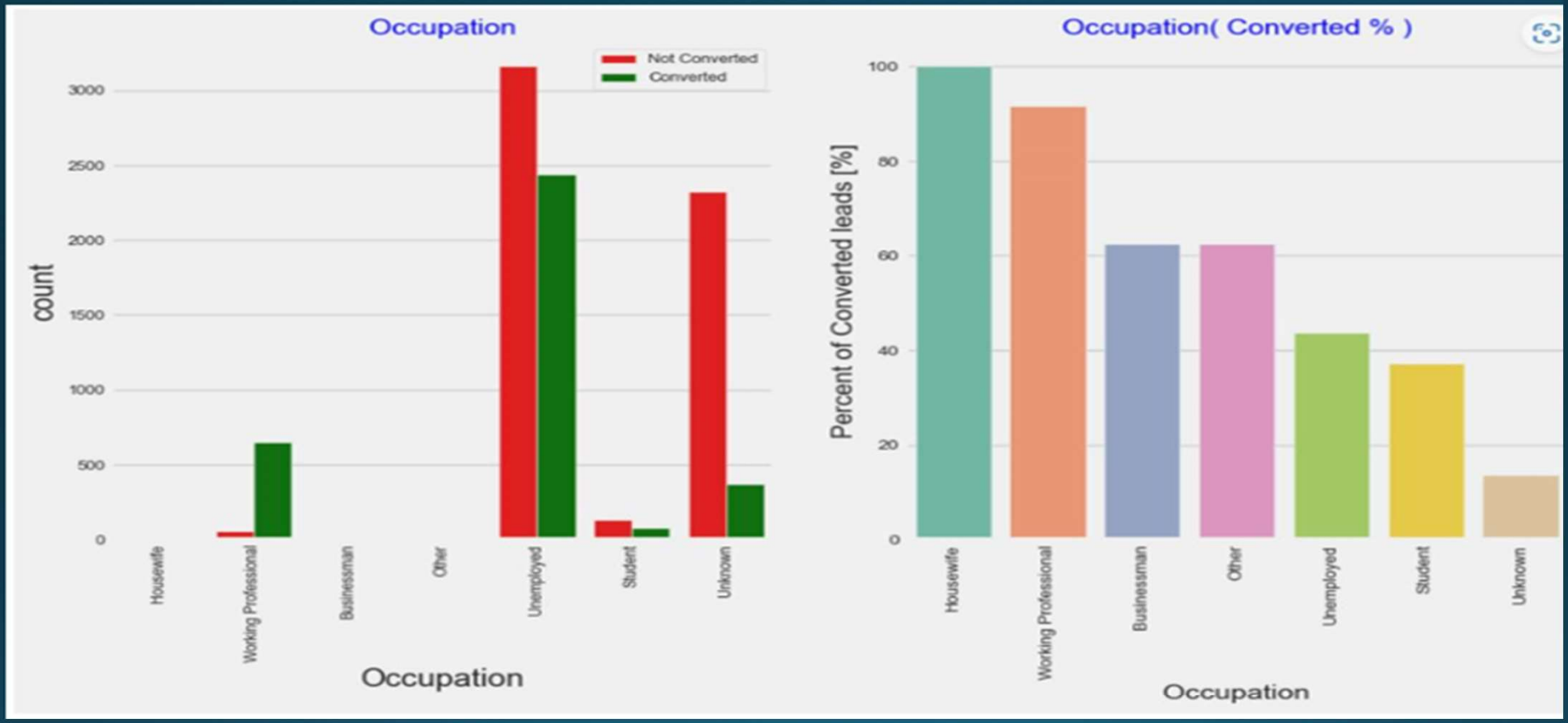
Lead Source: The majority of leads came from Google, with 40% of them converting. Direct traffic, organic search, and olark chat were the next most popular lead sources, with 35%, 38%, and 30% conversion rates, respectively.



Do Not Email: Major conversion has happened from email sent



Last Activity : Not much impact on conversion rates through Search, digital advertisements and through recommendations. Strategy- We will combine smaller Last Activity values as 'Other Activity'. Last Activity value of SMS Sent' had more conversion



Occupation: More conversion happened with people who are unemployed

Variables Impacting the Conversion Rate

Total Time Spent On Website

Lead Origin Lead Add Form

Last Source Selinga Website

Do Not Email_yes

Last Activity converted to lead

Last Activity Email Bounced

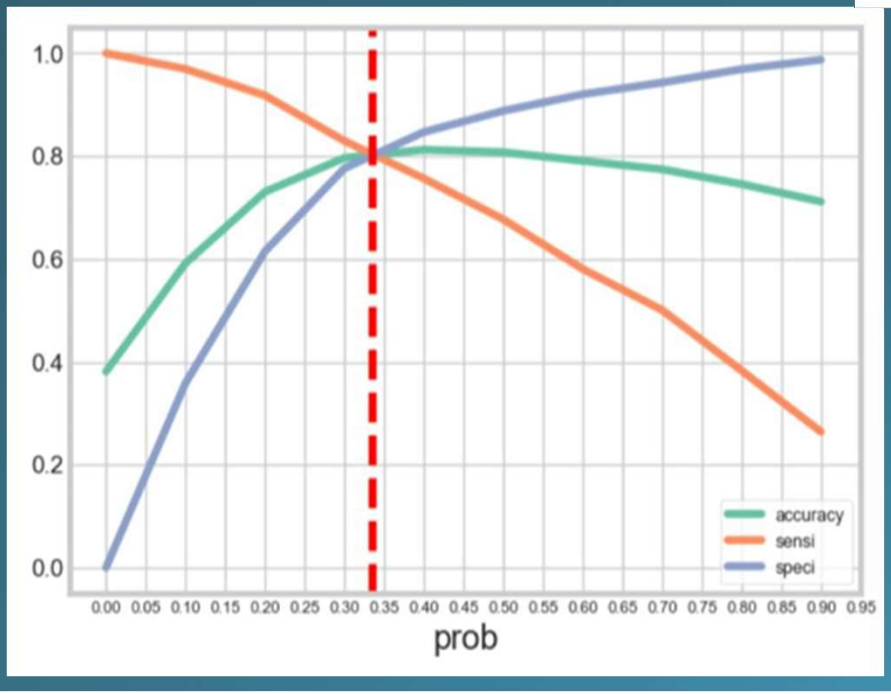
What is your current occupation_housewife

Other Impacting Variables :

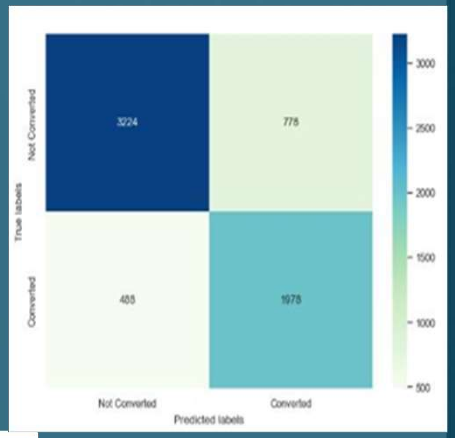
- ✓ What is your current occupation_student
- ✓ What is your current occupation_unemployed
- ✓ What is your current occupation_working professional
- ✓ Last Notable Activity_email link clicked
- ✓ Last Notable Activity_email opened
- ✓ Last Notable Activity_modified
- ✓ Last Notable Activity_olark chat conversation
- ✓ Last Notable Activity_page visited on website

Model Evaluation - Sensitivity and Specificity on Train Data Set

The graph depicts an optimal cut off of 0.37 based on Accuracy, Sensitivity and Specificity



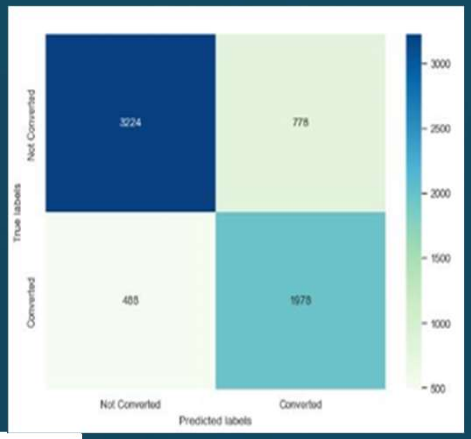
Confusion Matrix



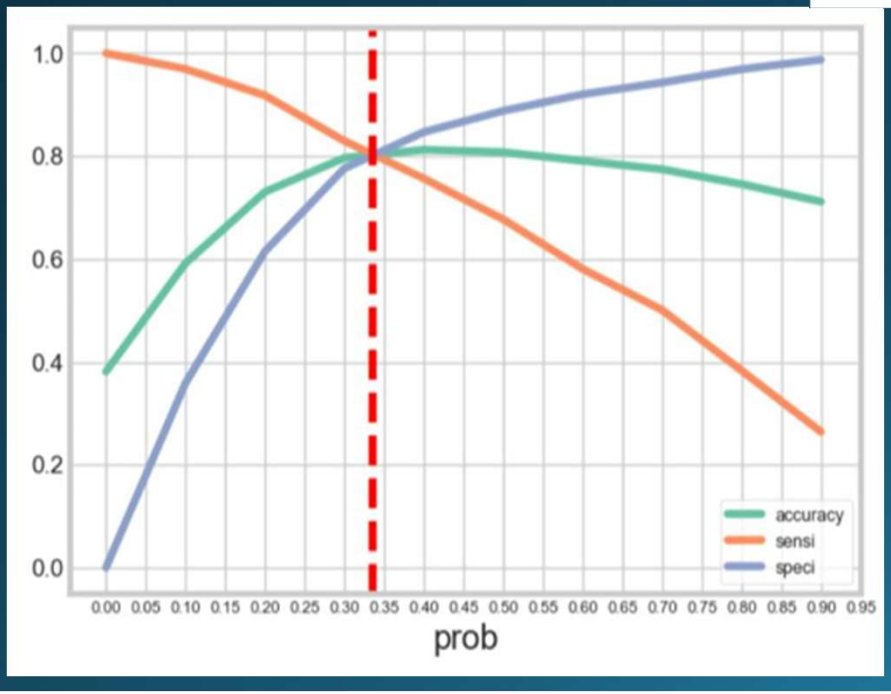
- Accuracy- 80%
- Sensiitvity - 80%
- Specifycity - 81%
- FalsePositiveRate- 16%
- PositivePredictiveValue - 74%

Model Evaluation Precision and Recall on Train Data Set

Confusion Matrix

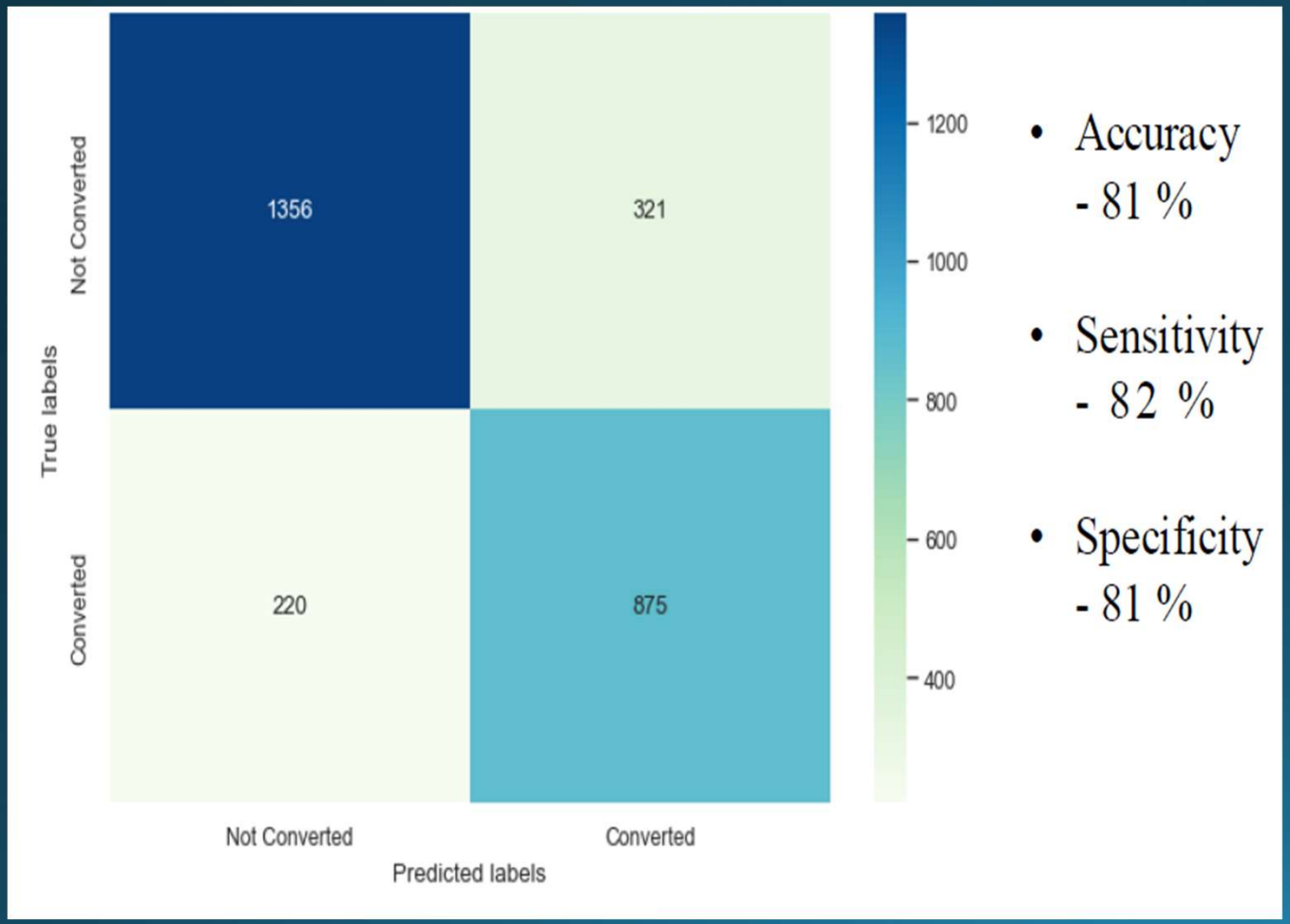


The graph depicts an optimal cut off of 0.42 based on Precision and Recall



Precision - 73%
Recall - 78%

Confusion Matrix



Model Evaluation Sensitivity and Specificity on Test Dataset

- Accuracy - 81 %
- Sensitivity - 82 %
- Specificity - 81 %

Interpretation Several predictor variables In a logistic regression model

- In general, numerous predictor variables can be included in a logistic regression model, as shown below:

- $\text{logit}(p) = \log(p/(1-p)) = 0 + 1 * X1 + \ldots + n * Xn$

- Using our example dataset as a basis, each estimated coefficient is the predicted change in the log odds of being a possible lead for a unit increase in the relevant predictor variable while keeping the other predictor variables constant at a specific value. An exponentiated coefficient Is the ratio of two probabilities, or the change in odds on a multiplicative scale given a unit Increase in the related predictor variable while maintaining the other variables constant.

The magnitude and sign of the coefficients loaded In the logic function:

- logit(p) = log(p/(1p)) = (3.42 * Lead Origin Lead Add Form) + (2.84 *Occupation_Working Professional) + (1.99 * Lead Source_WelingakWebsite) + (1.78 * Last Activity_SMS Sent) + (1.25 * LastActivity_Unsubscribed) + (1.09 * Total Time Spent on Website) + (0.98 *Lead Source_Olark Chat) + (0.84 * Last Activity_Unreachable) + (0.66 *Last Activity_Email Opened) --(0.25 * Lead Origin_Landing PageSubmission) --(0.87 * Last Activity_Olark Chat Conversation) --(1.26 * DoNot Email) 1.77
- The estimations allow us to make forecasts. This is accomplished byestimating the effects of all predictors for a specific scenario, adding themup, and applying a logistic transformation. Consider the case of a workingprofessional who was recognised via the Selinga website, spoke on OlarkChat, spent little Itme on the website, and requested to be contacted through email.
- We can then compute his conversion probability as 3.42 * 0 + 2.84 * 1 +1.99 * 1 + 1.78 * 0 + 1.25 * 0 + 1.09 * 0 + 0.98 * 0 + 0.84 * 0 + 0.66 * 0 0.25* 0 0.87 * 1 1.26 * 0 1.77 = 2.84 + 1.99 0.87 1.77 = 2.19 log(p/(1 p)).
- The logistic transformation Is: Probability = 1 / (1 + exp(x)) = 1 /(1 + exp(2.19)) = 1 /(1 + exp(2.2)) = 0.10 = 10%

Probability Prediction

- The estimations allow us to make forecasts. This is accomplished by estimating the effects of all predictors for a specific scenario, adding them up, and applying a logistic transformation.
- Consider the case of a working professional who was recognised via the Selinga website, spoke on Olark Chat, spent little time on the website, and requested to be contacted through email.
- Then we can calculate his conversion probability as 3.41 * 0 + 2.82 *1 + 2.34 * 0 + 2.01 * 1 + 1.86 * 0 + 1.32 * 0 + 1.09 * 0 + 0.97 * 0 +0.93 * 0 + 0.76 * 0 0.26 * 0 0.77 * 1 1.24 * 0 1.86 which Is 2.82 +2.01 0.77 1.86 = 2.2 which is log(p/(1 p))
- The logistic transformation is: Probability = 1 / (1 + exp(x)) = 1 /(1 +exp(2.2)) = 1 /(1 + exp(2.2)) = 0.143 = 14.3%

Probability ratios

- Because the idea of odds ratios is more social than rational, the marketing team may need to get odds rather than probabilities at times.
- To understand odds ratios, we must first define odds, which is defined as the ratio of the probability of two mutually incompatible occurrences. Consider our prior forecast of a 10% lead conversion chance in the section on probabilities. Because the lead conversion chance is 10%, the no conversion probability is 100% 10% = 90%, and hence the odds are 10%vs 90%. When we divide both sides by 90%, we get 0.11 versus 1, which wean just write as 0.11. Thus, 0.11 odds Is merely another way of describing chance of lead conversion of 10%.
- Similarly, leaving other categorical and numerical factors constant, the odds of a lead being converted for a Working Professional (Working Professional = 1) over the odds of a lead being converted for non working professionals (Working Professional = 0) Is exp(.2.84) = 17.11.
- When all other variables are set to zero, log(p/(1p)) = 17.11.
- We may utilize odds ratios to detect possible lead conversions by comparing people' profiles.

Conclusion

- While we have checked both Sensitivity Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction. As per our Logistic Regression Model, we can conclude that the model would help X Education to identify the leads that are most likely to convert into paying customers
- Accuracy, Sensitivity and Specificity values of test set are around 81%, 82% and 81% which are approximately closer to the respective values calculated using trained set. Since the model has an Accuracy and Precision of about 80% it would also help meet the CEO's ballpark target of lead conversion rate to be around 80%.
- Also the lead score calculated shows the conversion rate on the final predicted model Is around 79 %(in train set) and 78% In test set.
- The top 3 variables that contribute for lead getting converted in the model are
    - Total Time Spent on Website
    - Lead Add Form (from Lead Origin)
    - What is your current occupation_working professional