Lead Scoring Case Study

To Build a Logistic Regression Model To predict whether a lead for online courses for an X education company would be successfully converted or not

By- 1. Ujwal Kesharwani

2. Harleen kaur

Problem Statement

- X Education cells online courses to industry professionals
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, 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

- To help X Education to select the most promising leads(Hot Leads), i.e. 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.

Solution Methodology

- **Data cleaning and data manipulation.**
  1. Check and handle duplicate data.
  2. Check and handle NA values and missing values.
  3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
  4. Imputation of the values, if necessary.
  5. Check and handle outliers in data.
- **EDA**
  1. Univariate data analysis: value count, distribution of variable etc.
  2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- **Feature Scaling & Dummy Variables and encoding of the data.**
- **Classification technique: logistic regression used for the model making and prediction.**
- **Validation of the model.**
- **Model presentation.**
- **Conclusions and recommendations.**

Data Manipulation

- Total Number of Rows =37, Total Number of Columns =9240.

❖Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply

❖Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped

❖Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis

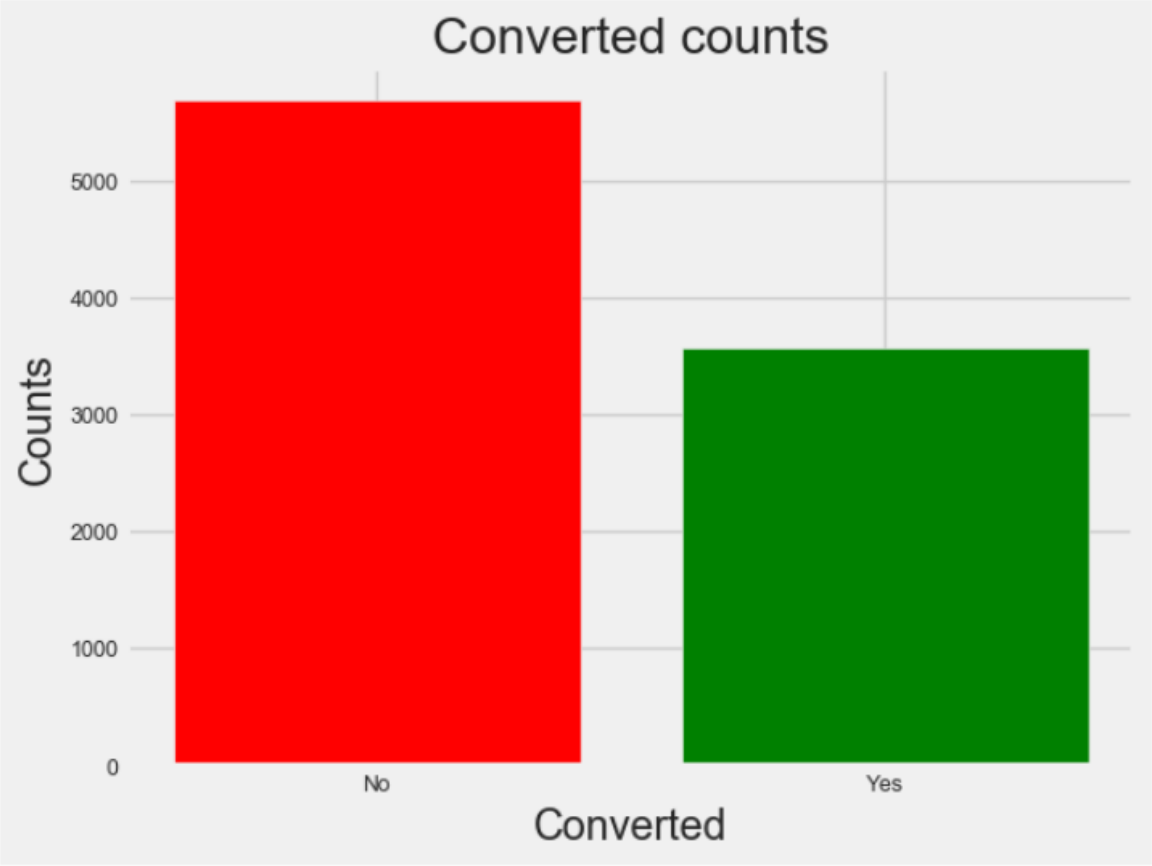
❖After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper”

❖Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.

❖Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

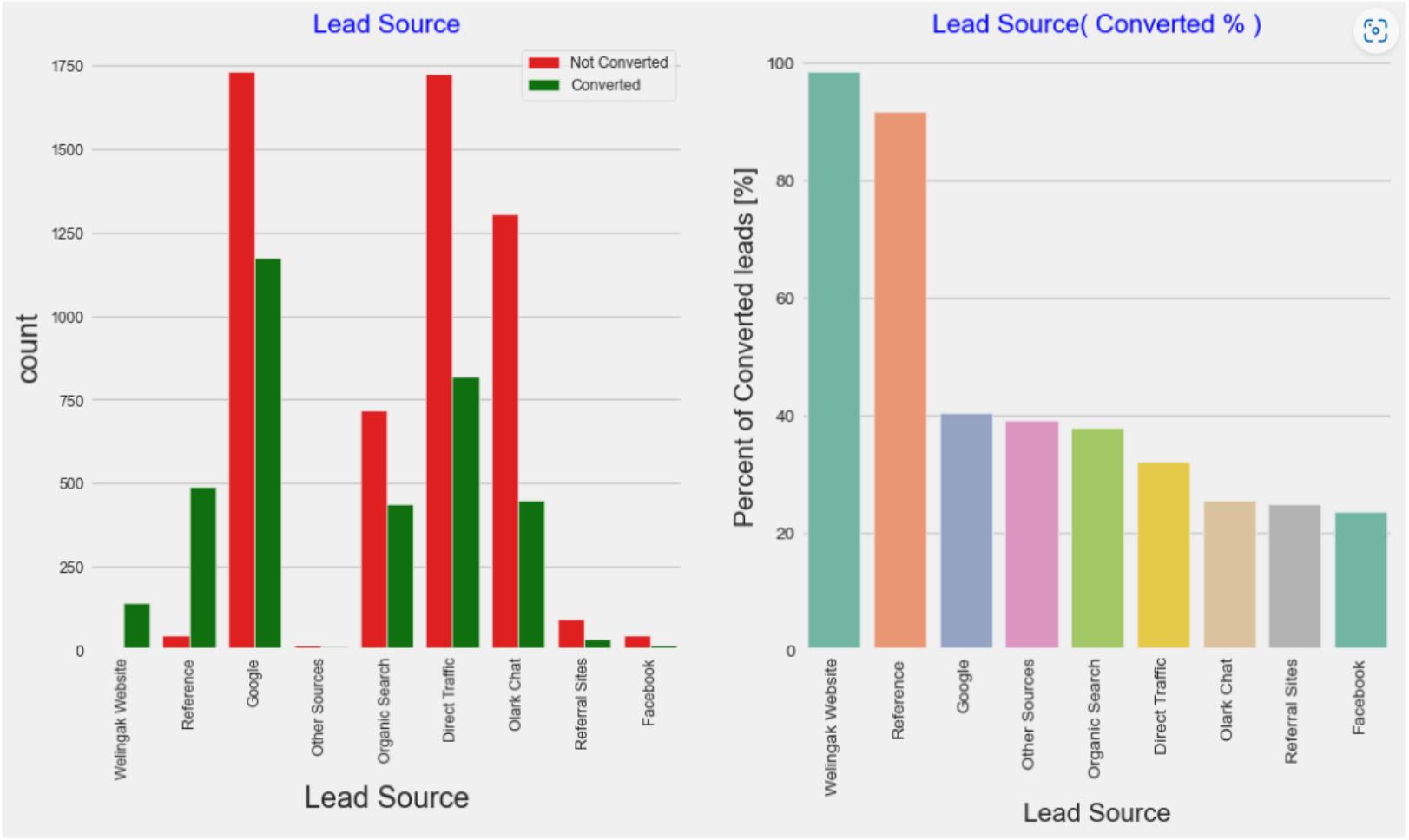
Exploratory Data Analysis

We have around 39% Conversion rate in Total

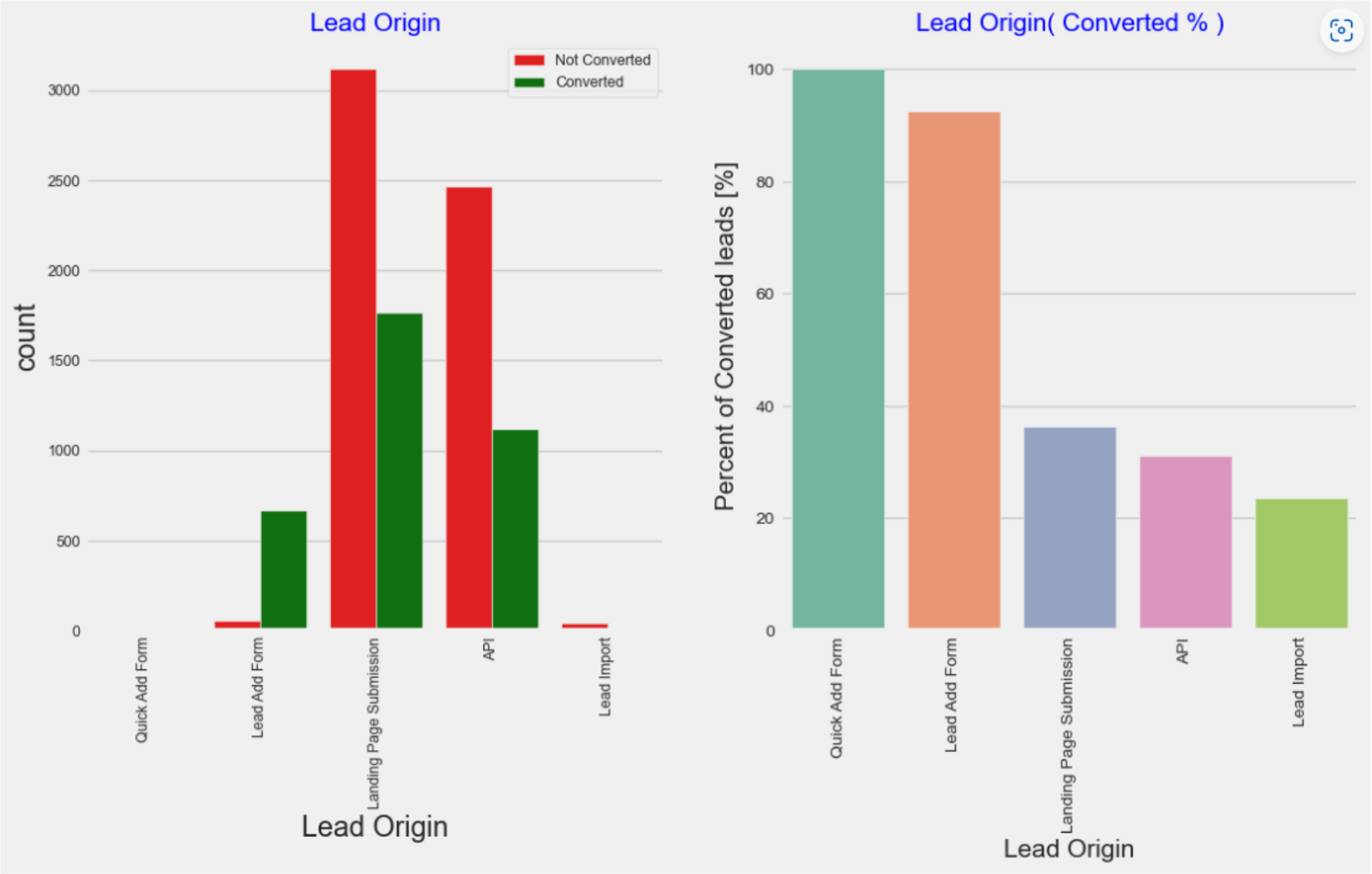


Insight: In the lead conversion ratio, 38.5% of visitors turned to leads, whereas 61.5% did not. As a result, it appears to be a well-balanced dataset.

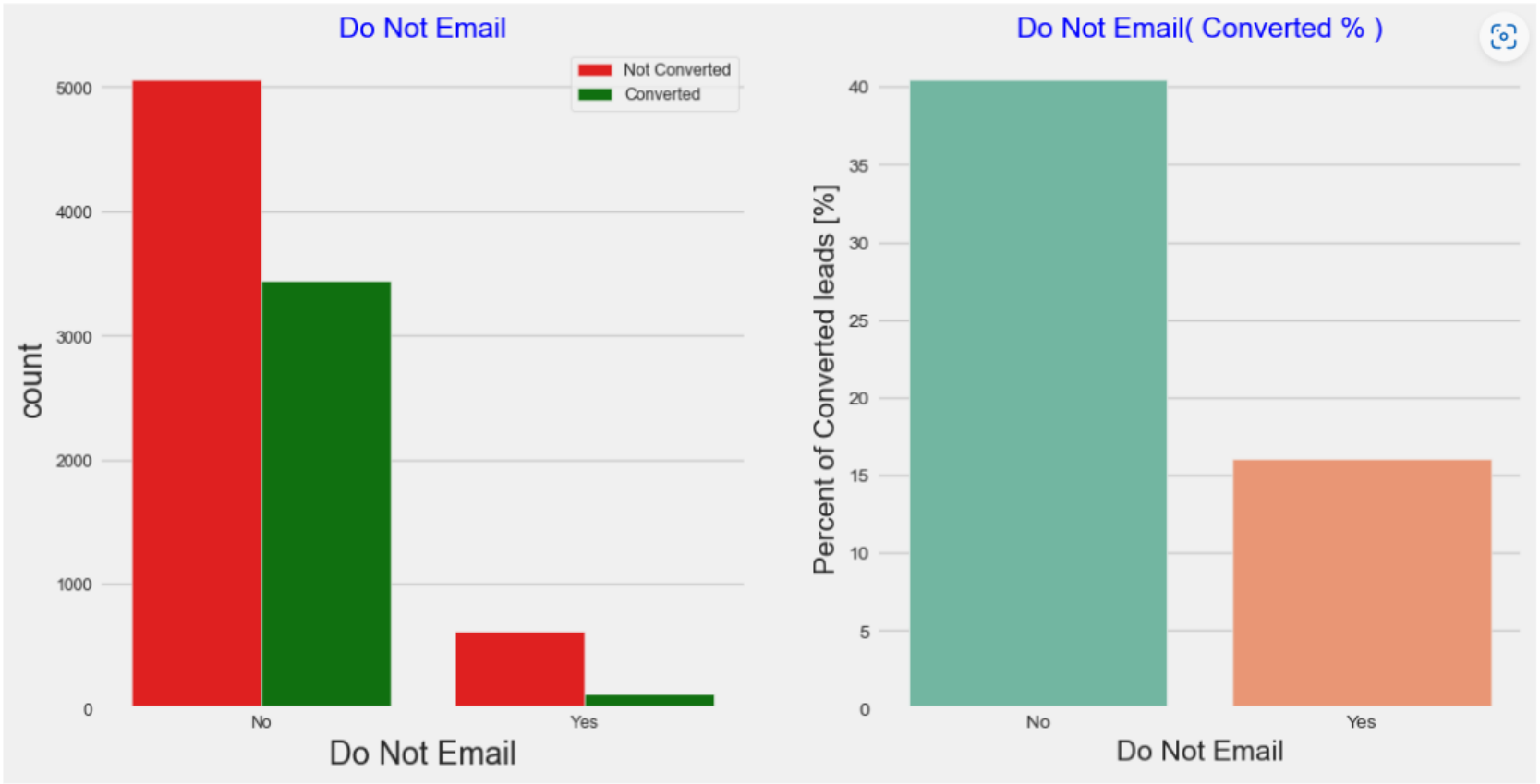
Lead Source


Lead Source
Not Converted
Converted
count
1750
1500
1250
1000
750
500
250
0
Welingak Website
Reference
Google
Other Sources
Organic Search
Direct Traffic
Olark Chat
Referral Sites
Facebook
Lead Source
Lead Source(Converted %)
Percent of Converted leads [%]
100
80
60
40
20
0
Welingak Website
Reference
Google
Other Sources
Organic Search
Direct Traffic
Olark Chat
Referral Sites
Facebook
Lead Source

In Lead Origin, maximum conversion happened from Landing Page Submission

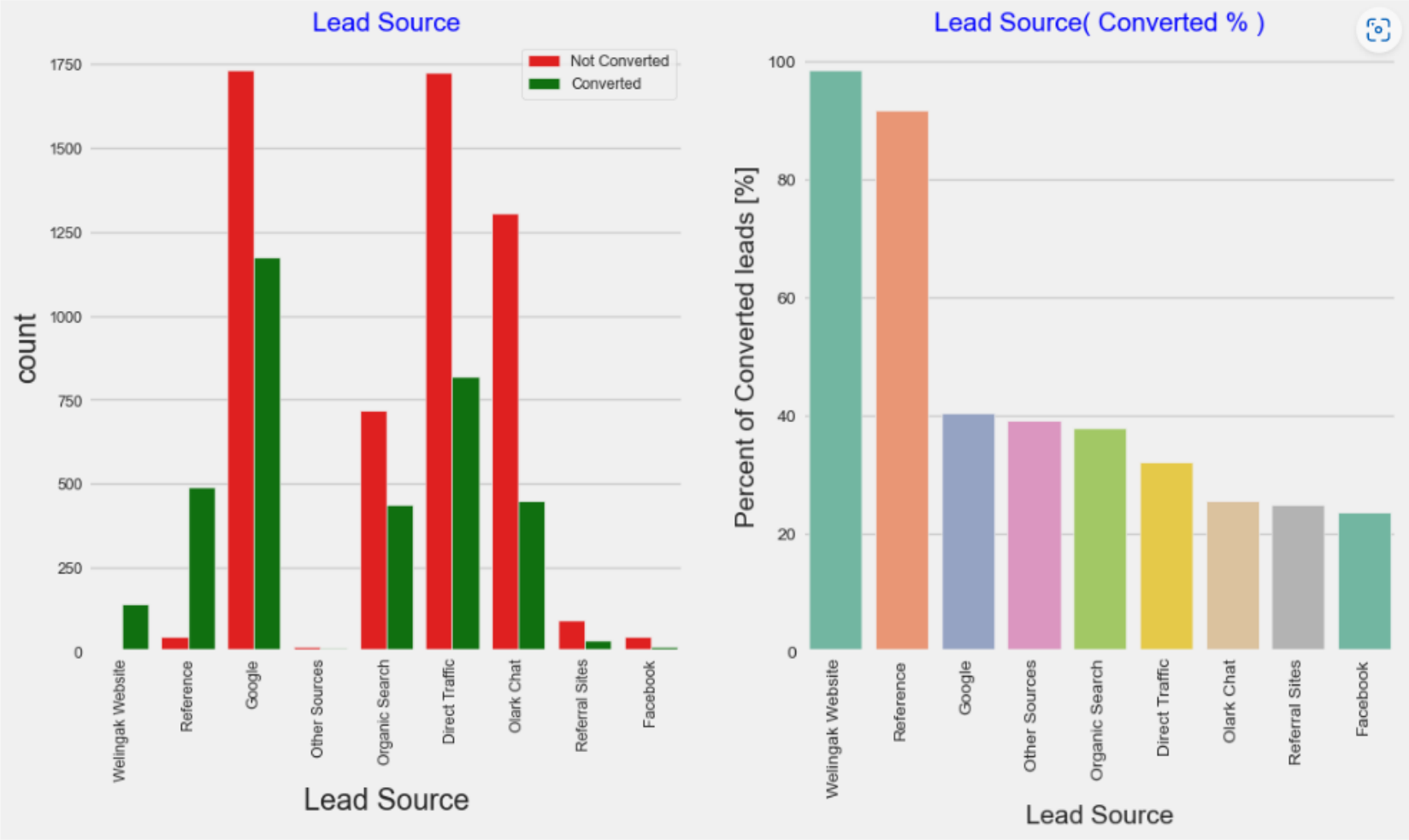


Major conversion has happened from Emails sent



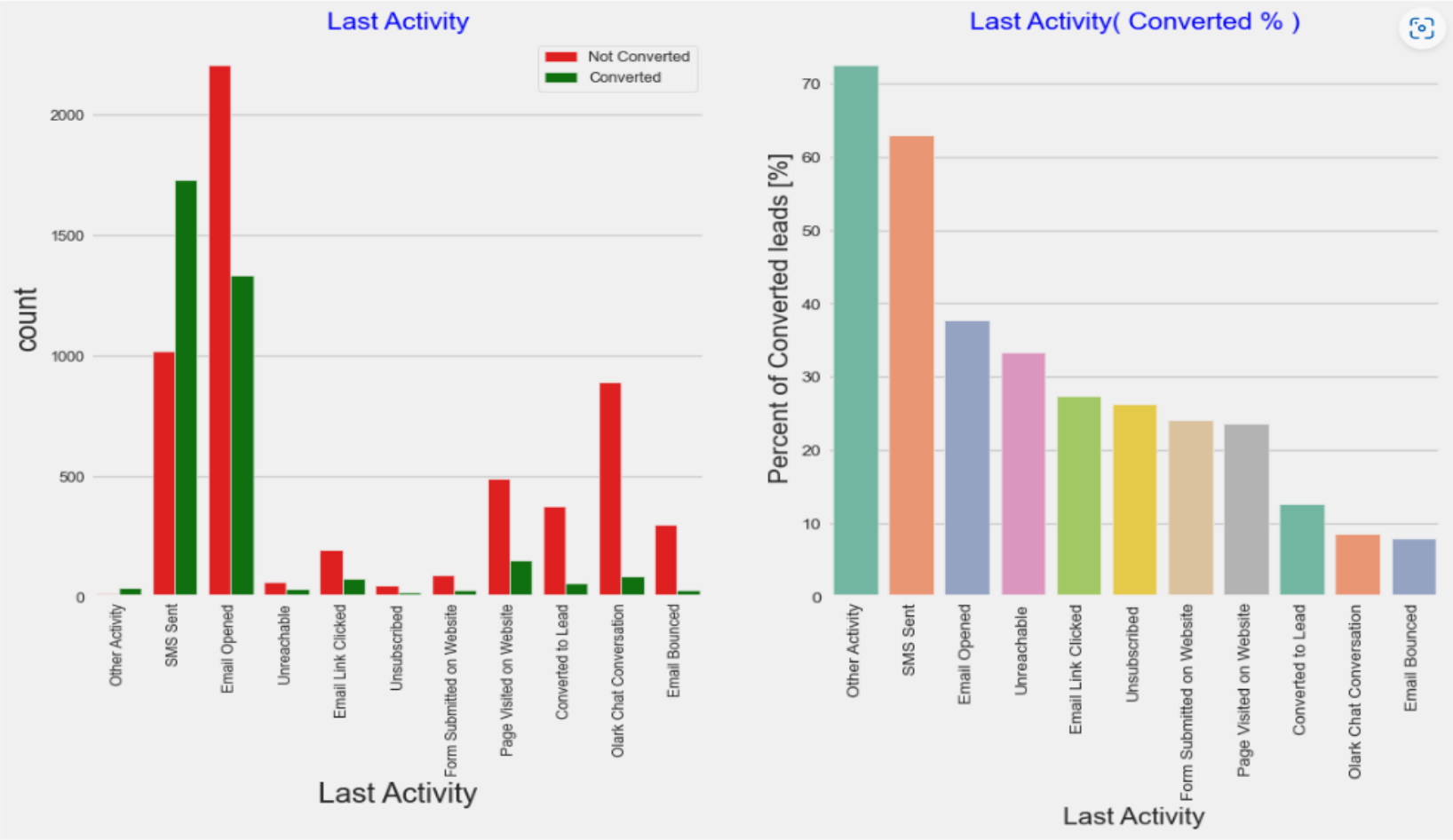
Insight: The vast majority of individuals (92%) are fine with getting email. Individuals that are comfortable with email have a 40% conversion rate. Individuals who have opted out of receiving emails have a lower conversion rate (just 15%).

Major conversion in the lead source is from Google



Insight: Google was the source of the most leads, with 40% of them converting, followed by Direct Traffic, Organic Search, and Olark Chat, with 35%, 38%, and 30% converting, respectively. A lead generated by a referral has a conversion rate of more than 90% out of a total of 534. The lead conversion rate on the Welingak website is around 100%. This option should be investigated further in order to boost lead conversion. To improve lead count, measures should be implemented to encourage existing members to refer more people.

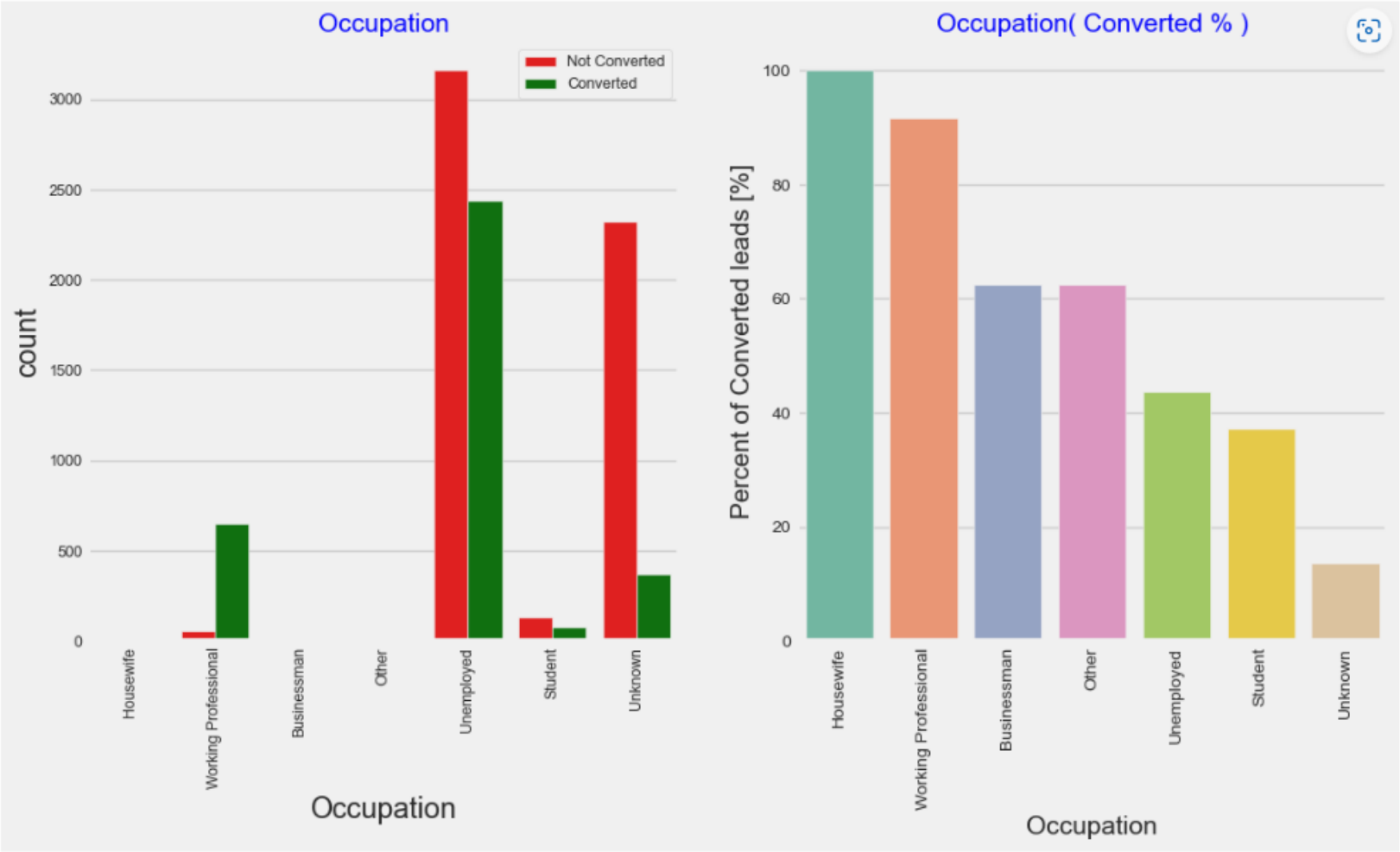
Not much impact on conversion rates through Search, digital advertisements and through recommendations



Insight:

The majority of leads have their Email open as their most recent action. Lead conversion is quite high (70%) after merging smaller Last Activity kinds like Other Activity. The conversion rate for leads with the most recent action as an SMS sent is around 60%.

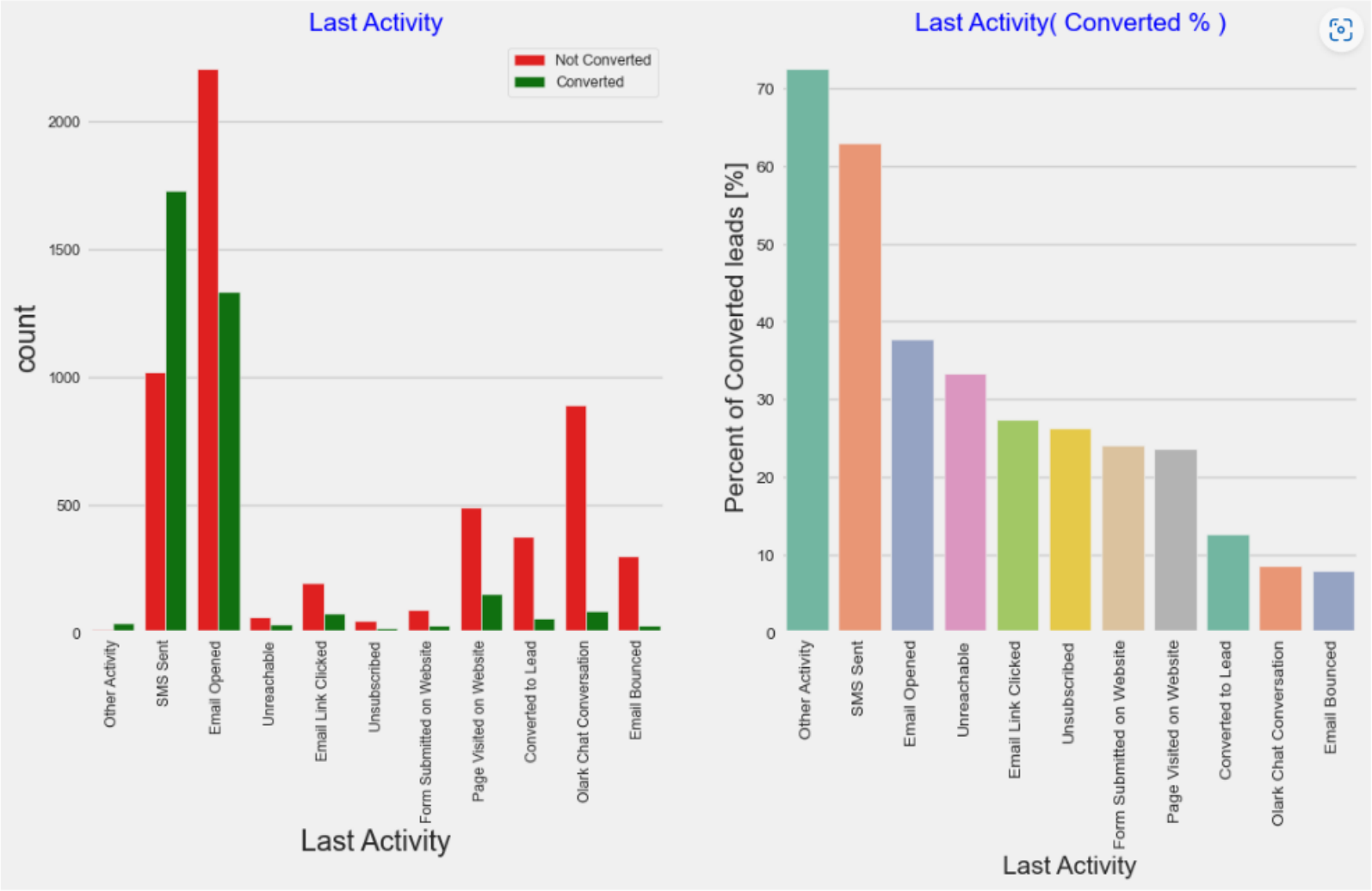
More conversion happened with people who are unemployed



Insight

Despite there are fewer housewives, they have a 100% conversion rate. Working professionals, business owners, and others have a high conversion rate. Despite the fact that the most individuals have been reached, the conversion rate (40%) is poor. We cannot mix smalled value categories since their conversion rates differ greatly. Combining them may result in incorrect forecasts.

Last Activity value of SMS Sent' had more conversion



Insight:

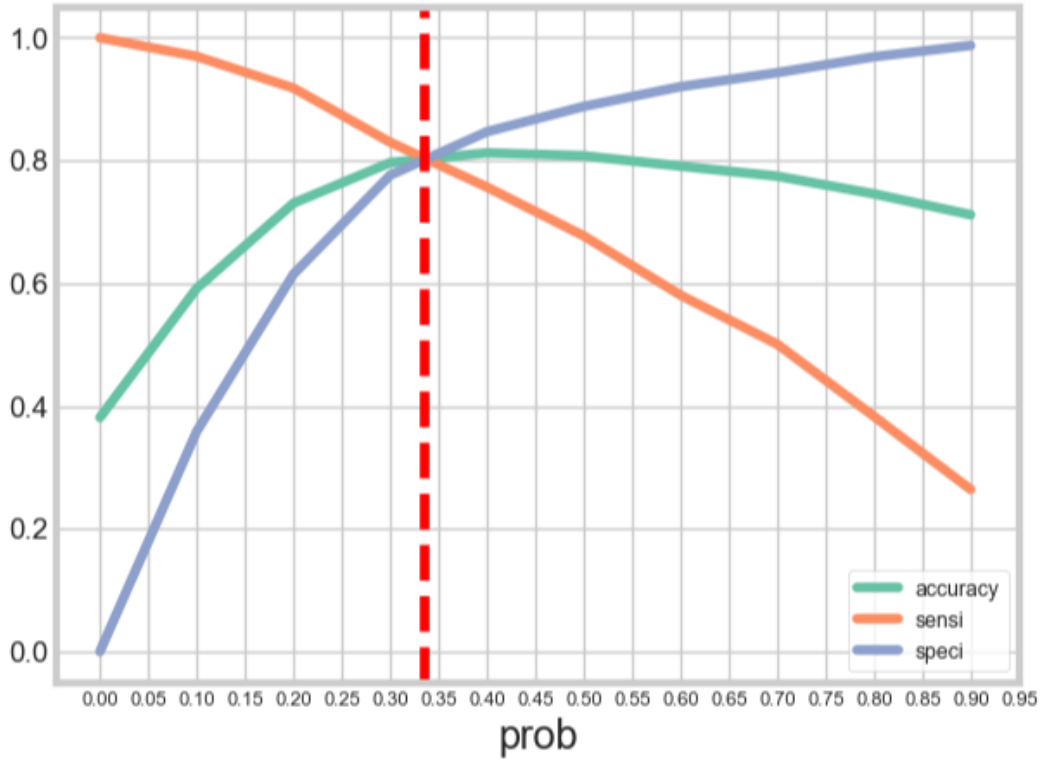
The majority of leads have their Email open as their most recent action. Lead conversion is quite high (70%) after merging smaller Last Activity kinds like Other Activity. The conversion rate for leads with the most recent action as an SMS sent is around 60%.

Variables Impacting the Conversion Rate

- Total Time Spent On Website
- Lead Origin – Lead Add Form
- Last Source – Welingak Website
- Do Not Email_yes
- Last Activity_converted to lead
- Last Activity – Email Bounced
- What is your current occupation_housewife
- 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

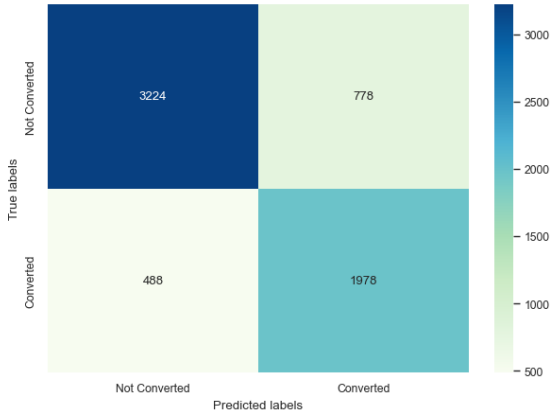
The graph depicts an optimal cut off of 0.37 based on Accuracy, Sensitivity and Specificity



Inferences:

From the above graph, 0.335 seems to be ideal cut-off points

Confusion Matrix

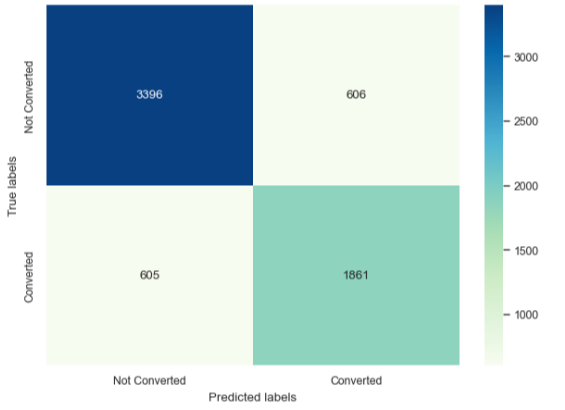


- Accuracy - 80%
- Sensitivity - 80 %
- Specificity - 81 %
- False Positive Rate - 16 %
- Positive Predictive Value - 74 %
- Negative Predictive Value – 87%

Model Evaluation – Precision and Recall on Train Data Set

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

Confusion Matrix

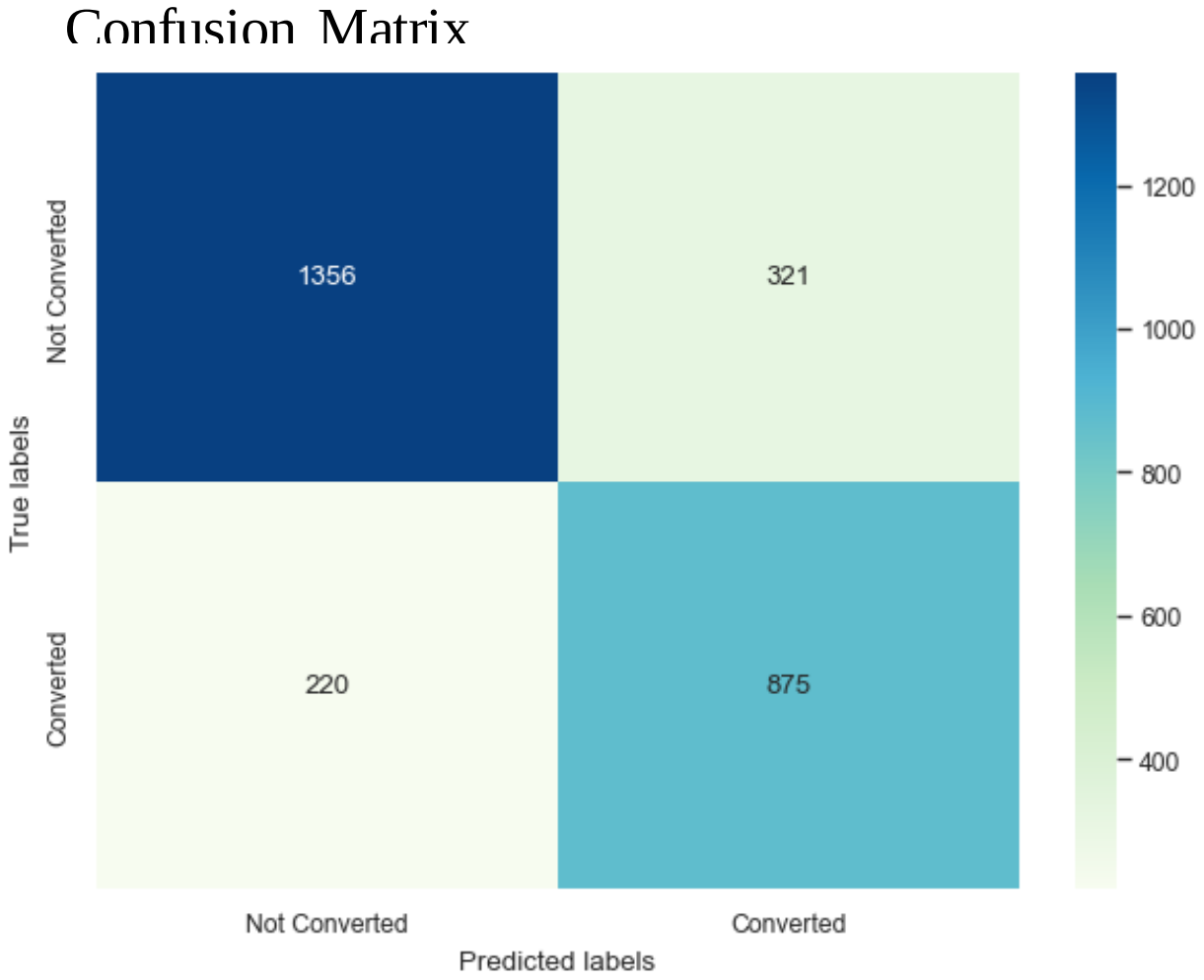


Precision - 73%
Recall - 78 %

Inferences:

The cutoff point appears to be 0.404 based on the Precision-Recall Trade off curve. This threshold value will be used to evaluate test data.

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 logit function:

- logit(p) = log(p/(1-p)) = (3.42 * Lead Origin_Lead Add Form) + (2.84 * Occupation_Working Professional) + (1.99 * Lead Source_Welingak Website) + (1.78 * Last Activity_SMS Sent) + (1.25 * Last Activity_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 Page Submission) -(0.87 * Last Activity_Olark Chat Conversation) - (1.26 * Do Not Email) -1.77
- 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 Welingak website, spoke on Olark Chat, spent little time 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 Welingak 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 non-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 we can just write as 0.11. Thus, 0.11 odds is merely another way of describing a 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/(1-p)) = 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. –
- 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.
- 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
- Hence overall this model seems to be good.