

The image is a digital collage on a light blue background. It features a central purple rectangular card with horizontal lines and the text 'HELLO! WELCOME TO MY PRESENTATION'. To the left is a yellow notepad with horizontal lines. To the right is a yellow spiral-bound notebook with a grid pattern. A green triangular warning sign with a white exclamation mark is placed on the purple card. In the bottom left corner, there is a green rectangular label with a barcode and the alphanumeric string '0001Q2315204L900Q82900'.

HELLO!

WELCOME TO MY
PRESENTATION








LEAD SCORING CASE STUDY



PRESENTATION'S AGENDA

- 
- 
- 
1. Problem statement
 2. Business Goal
 3. Strategy
 4. Data visualisation
 5. Model Building
 6. Model Evaluation
 7. Model summary

PROBLEM STATEMENT

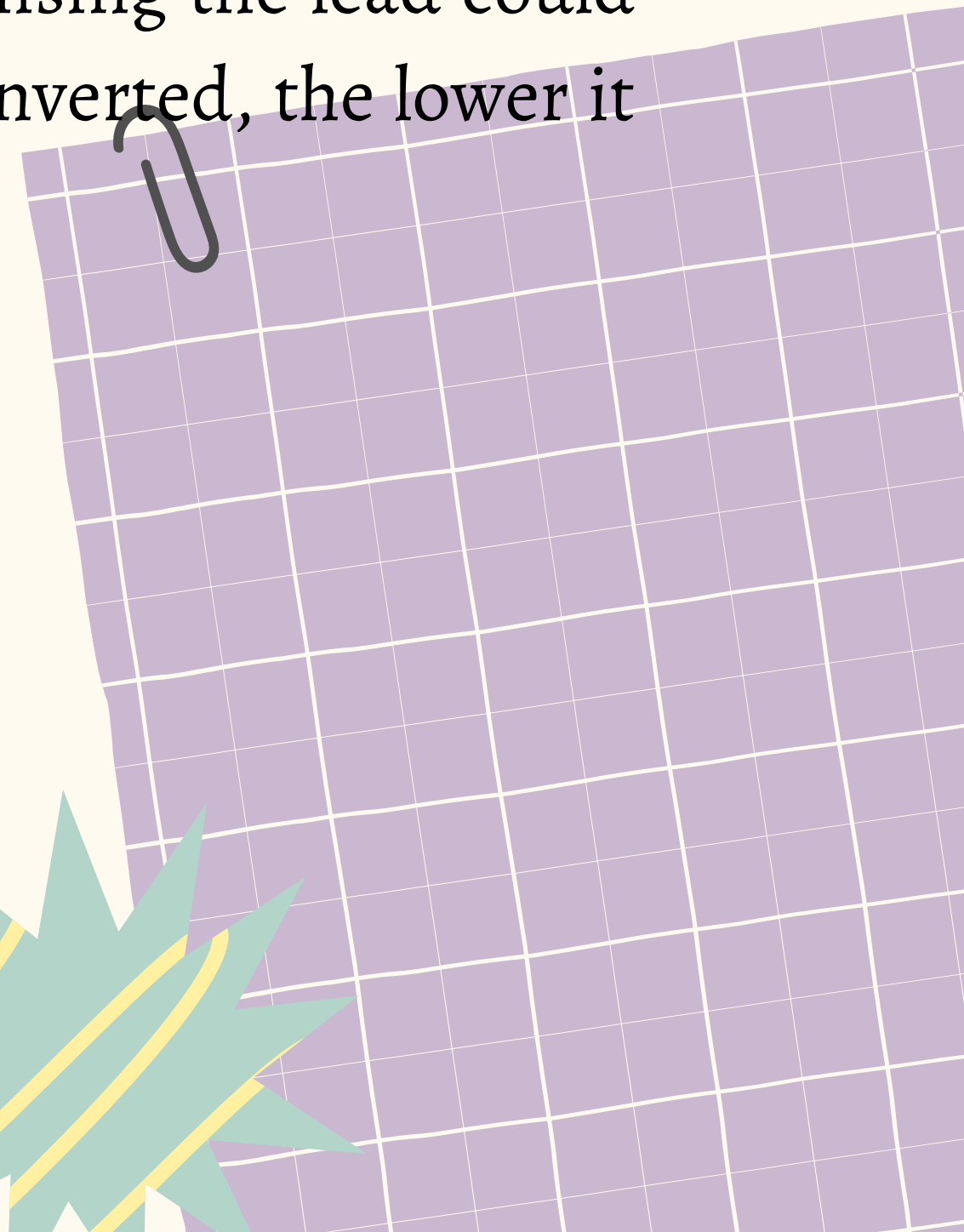
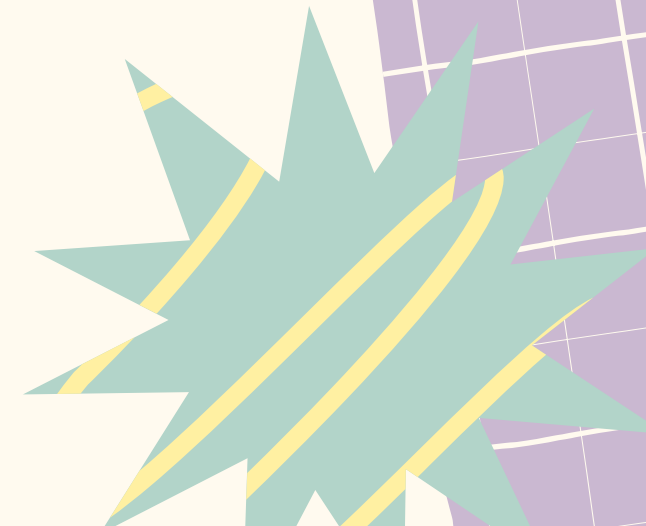
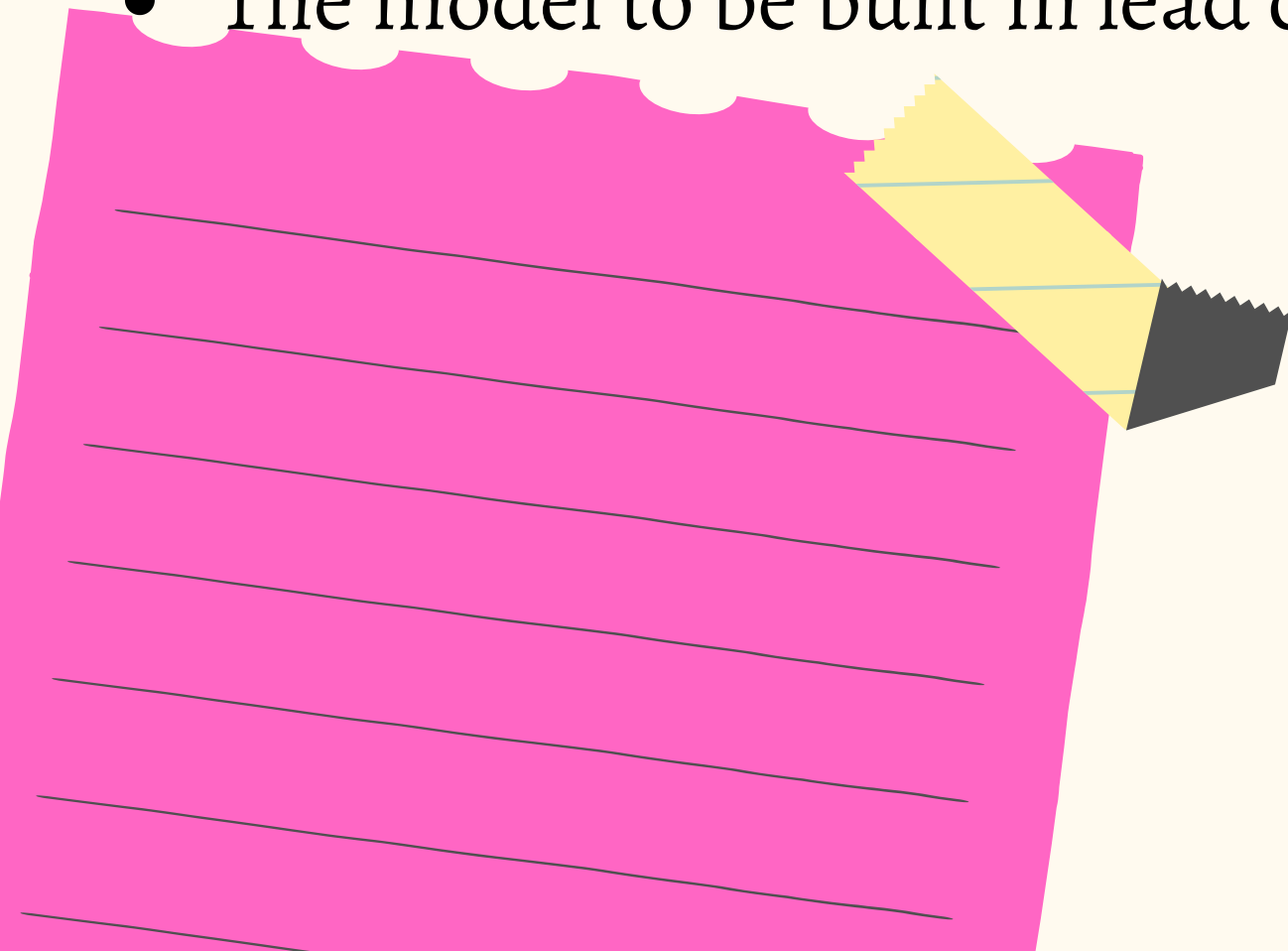
An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although 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 GOAL

- The company requires a model to be built for selecting most prominent leads.
- Lead score to be given to each lead such that it indicates how promising the lead could be. The higher the lead score the more promising the lead to get converted, the lower it is the lesser the chances of conversion.
- The model to be built in lead conversion rate around 80% or more



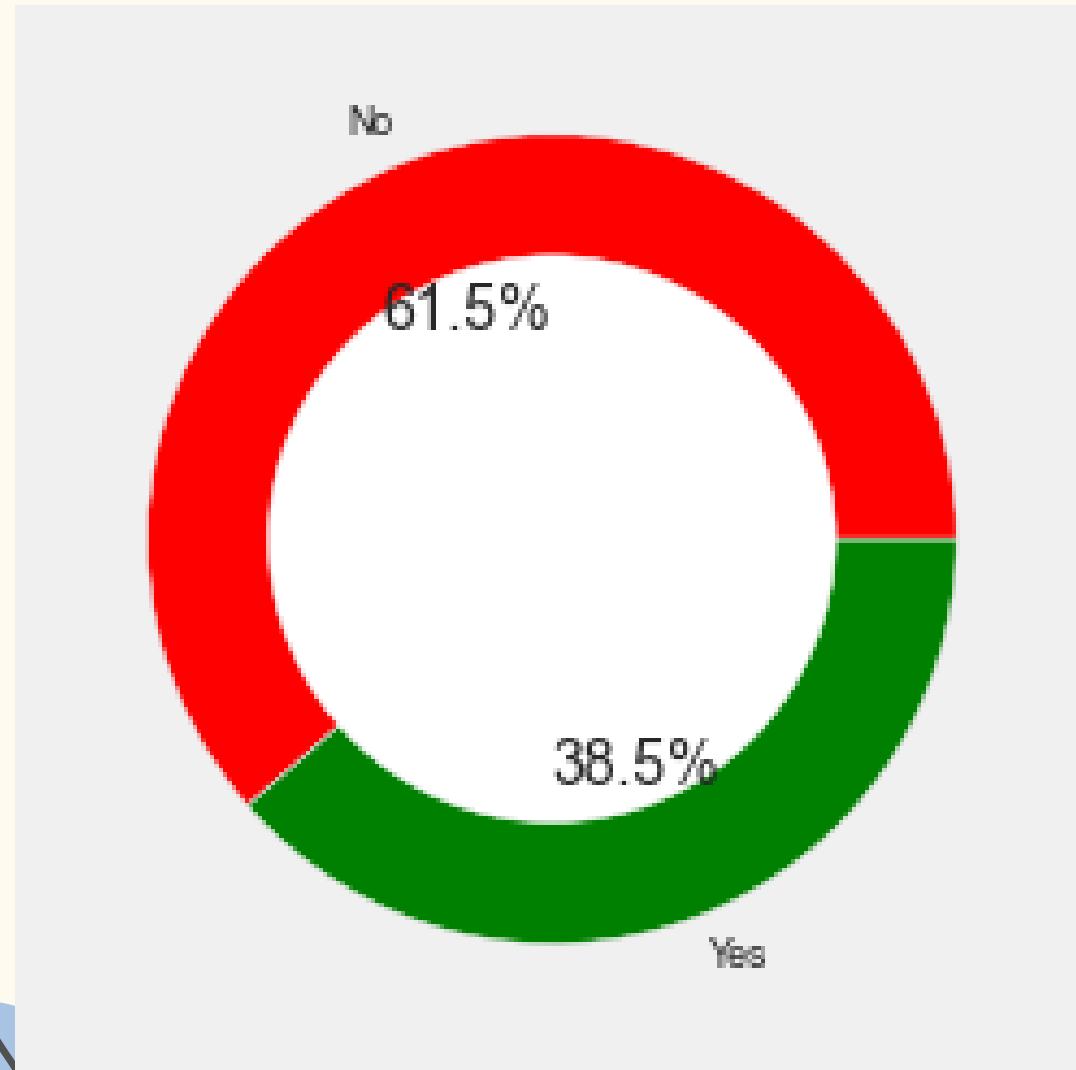
STRATEGY

- Import data.
- clean and prepare the acquired data for further analysis.
- Exploratory data analysis for figuring out most helpful attributes for conversion.
- Scaling features.
- Prepare the data for model building.
- Build a logistic regression model.
- Assign a lead score for each leads.
- Test the model on train set.
- Evaluate model by different measures and metrics.
- Test the model on test set.
- Measure the accuracy of the model and other



EXPLORATORY DATA ANALYSIS

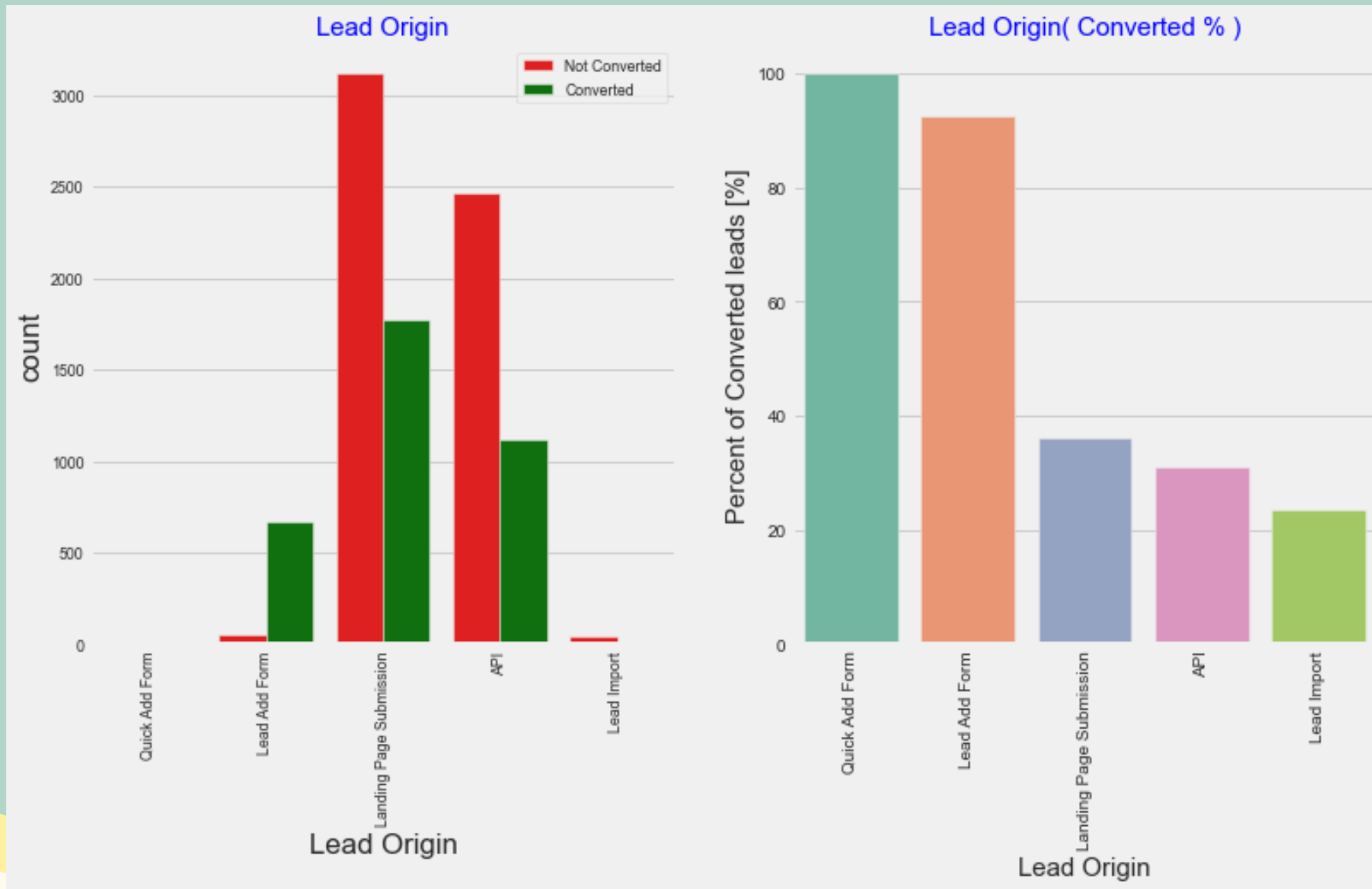
- Data Imbalance



Insight:

In the lead conversion ratio, 38.5% have converted to leads whereas 61.5% did not convert to a lead. So it seems like a balanced dataset.

- **Univariate Analysis - Categorical**
1. Lead Origin

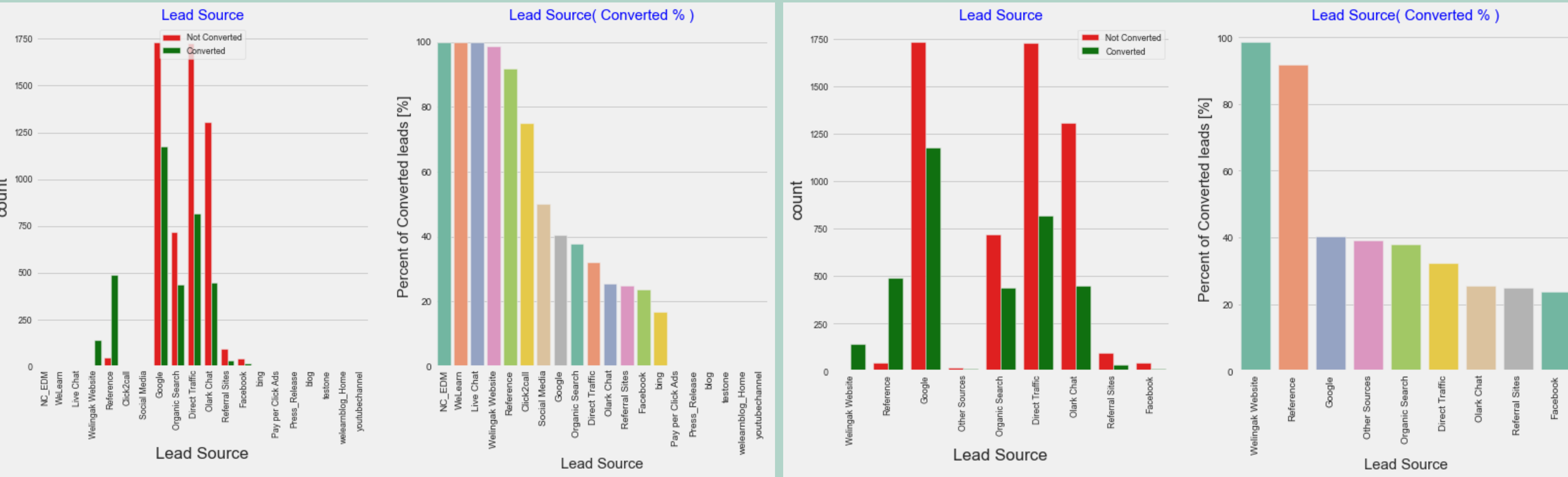


Insight:

- Most Leads originated from submissions on the landing page and around 38% of those are converted followed by API, where around 30% are converted.
- Even though Lead Origins from Quick Add Form are 100% Converted, there was just 1 lead from that category. Leads from the Lead Add Form are the next highest conversions in this category at around 90% of 718 leads.
- Lead Import are very less in count and conversion rate is also the lowest

To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

2. Lead Source

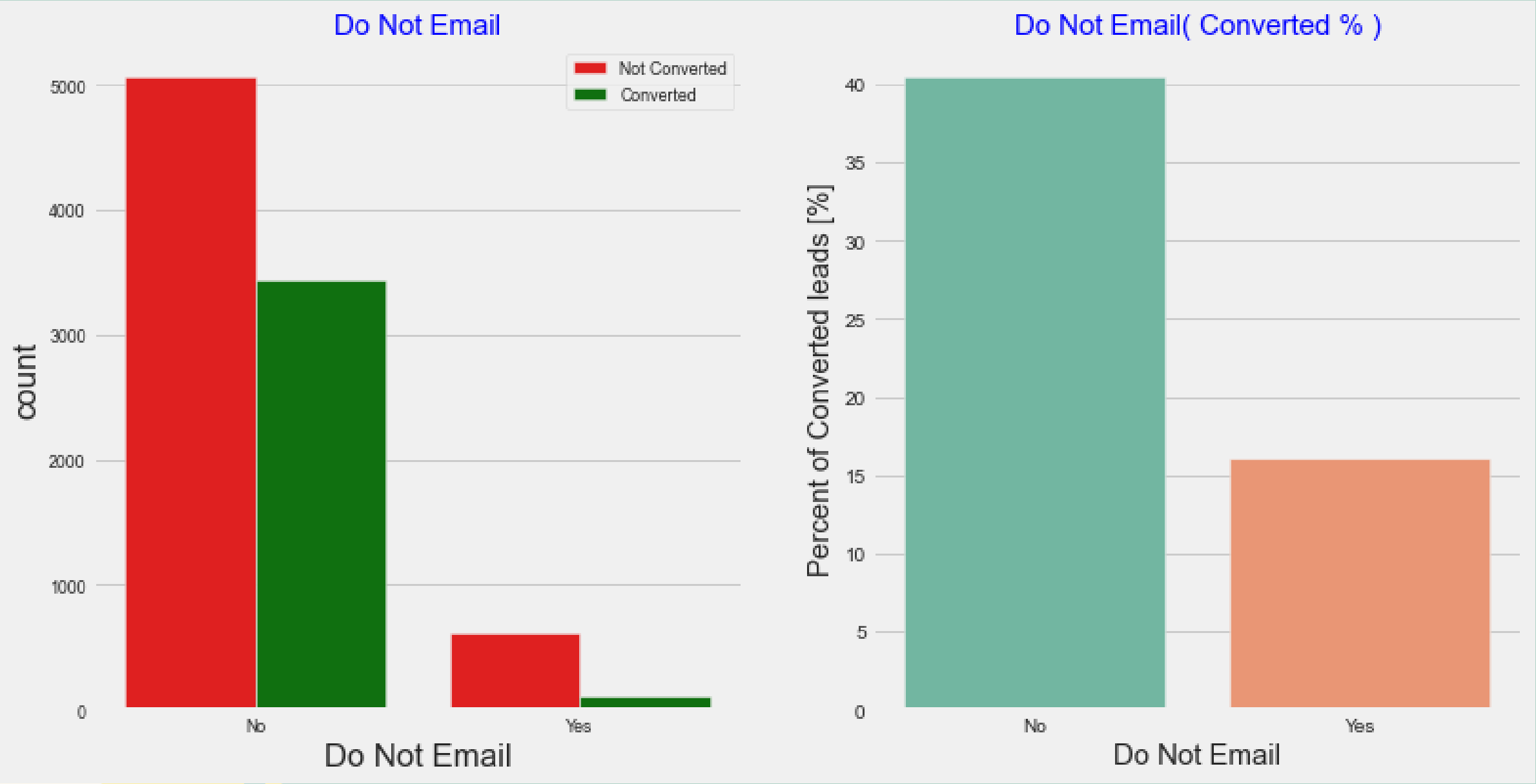


Insight:

- The source of most leads was Google, and 40% of the leads converted, followed by Direct Traffic, Organic search, and Olark chat where around 35%, 38%, and 30% converted respectively.
- A lead from a reference has over 90% conversion from the total of 534.
- Welingak's Website has an almost 100% lead conversion rate. This option should be explored more to increase lead conversion.

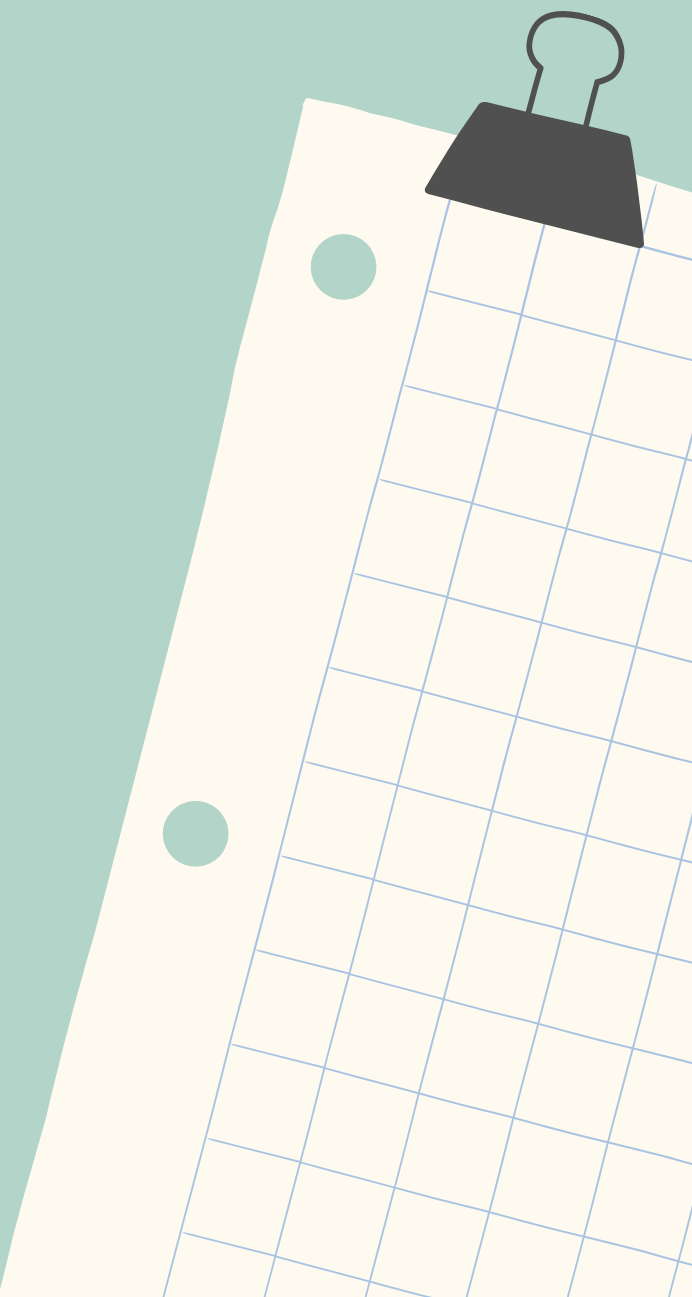
To increase lead count, initiatives should be taken so already exitsing members increase their referrals.

3. Do not Email

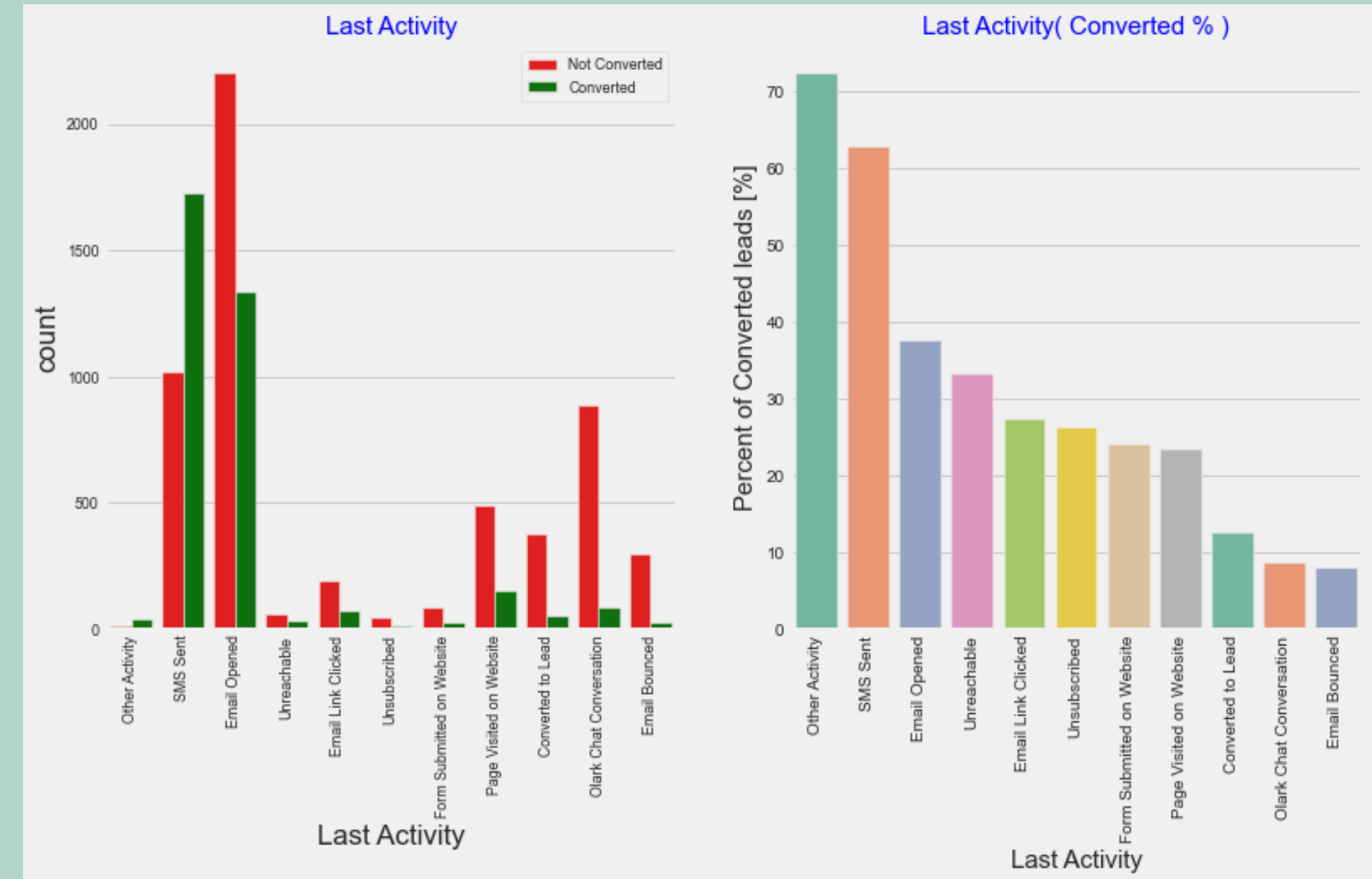
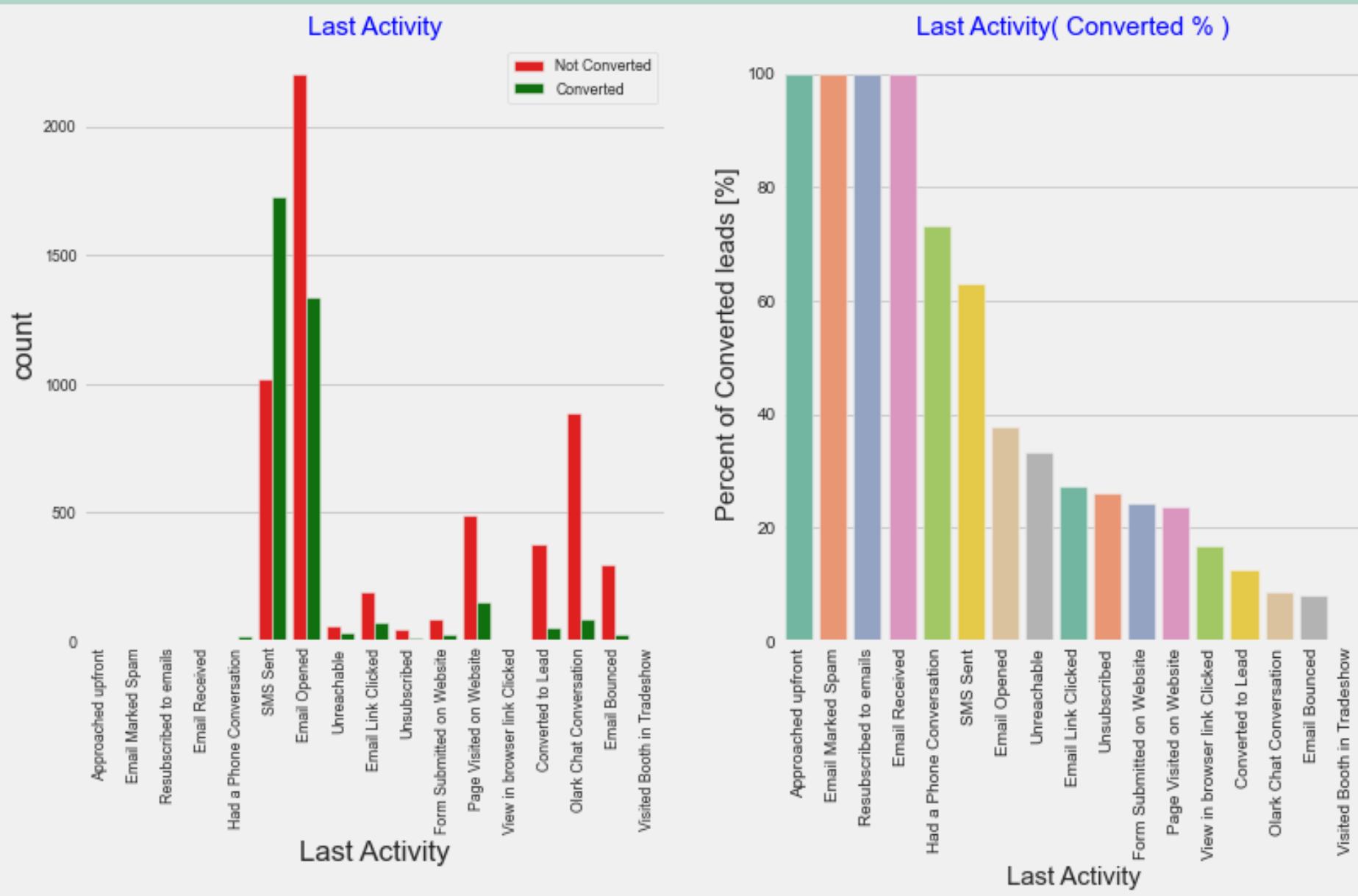


Insight:

- Majority of the people are ok with receiving email (~92%)
- People who are ok with email has conversion rate of 40%
- People who have opted out of receive email has lower rate of conversion (only 15%)



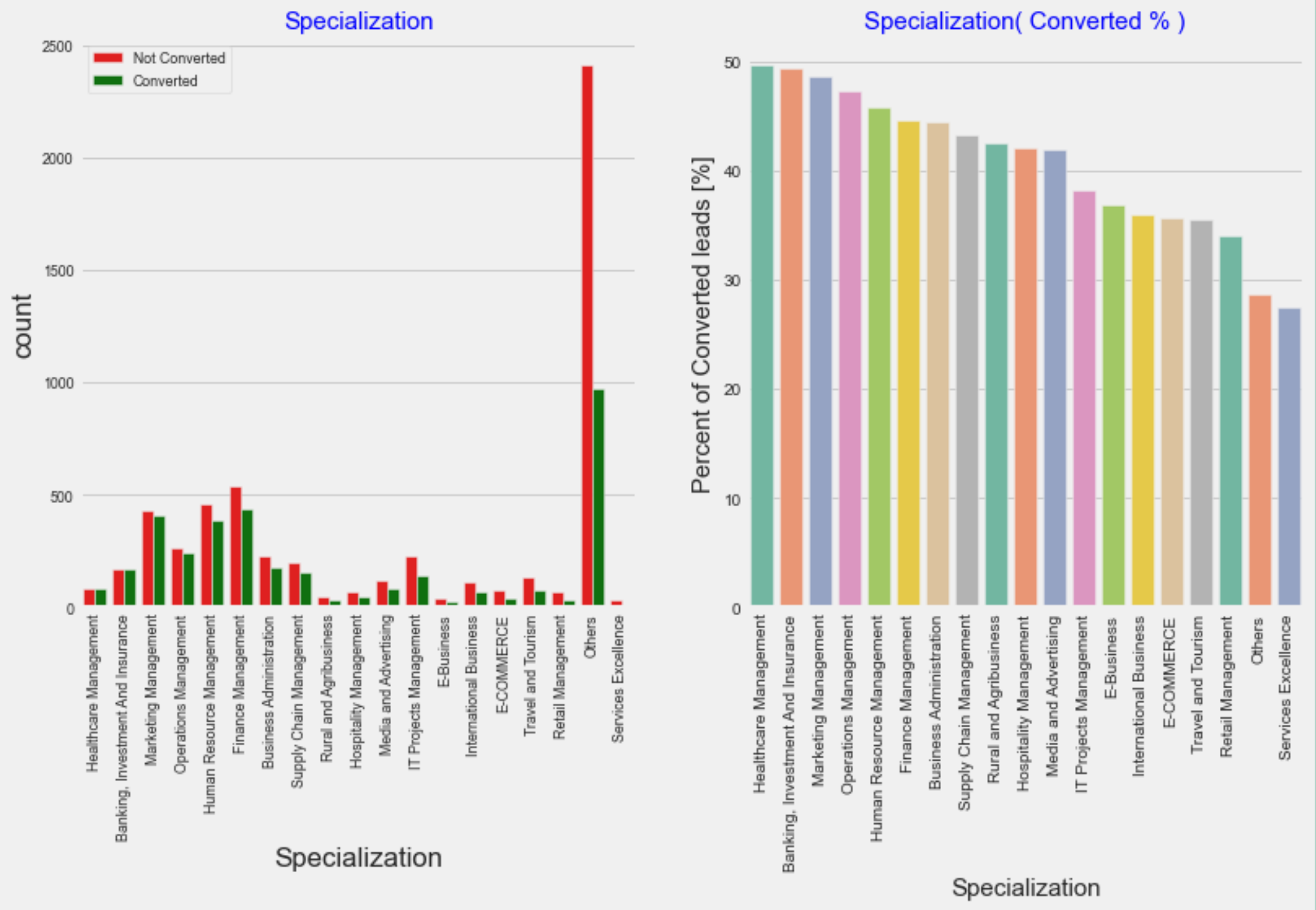
4. Last Activity



Insight:

- Most of the leads have their Email opened as their last activity
- After combining smaller Last Activity types as Other Activity, the lead conversion is very high (~70%)
- Conversion rate for leads with last activity as SMS Sent is almost 60%

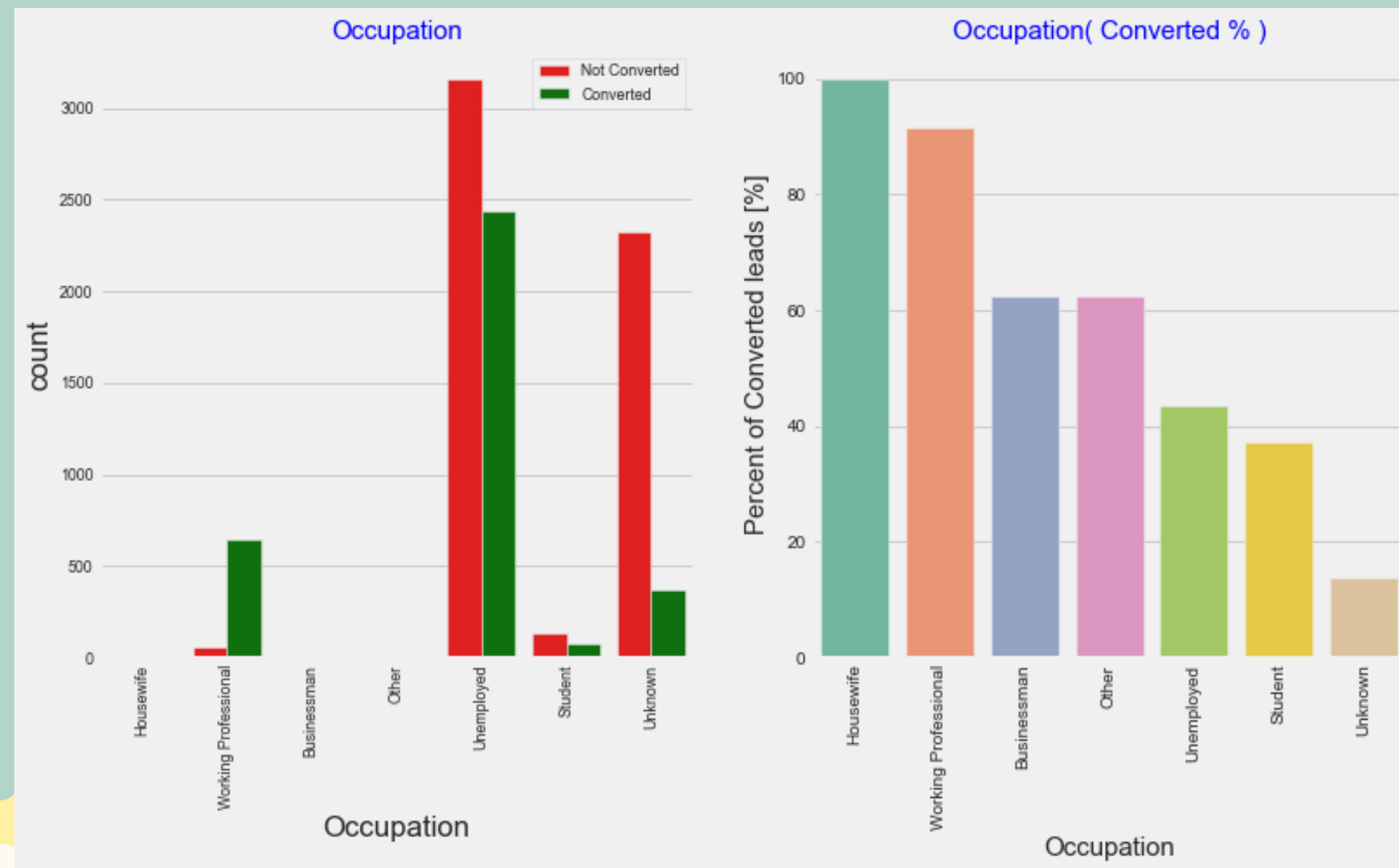
5 .Specialization



Insight:

- Most of the leads have not mentioned a specialization and around 28% of those converted
- Leads with Finance management and Marketing Management - Over 45% Converted

6. Occupation

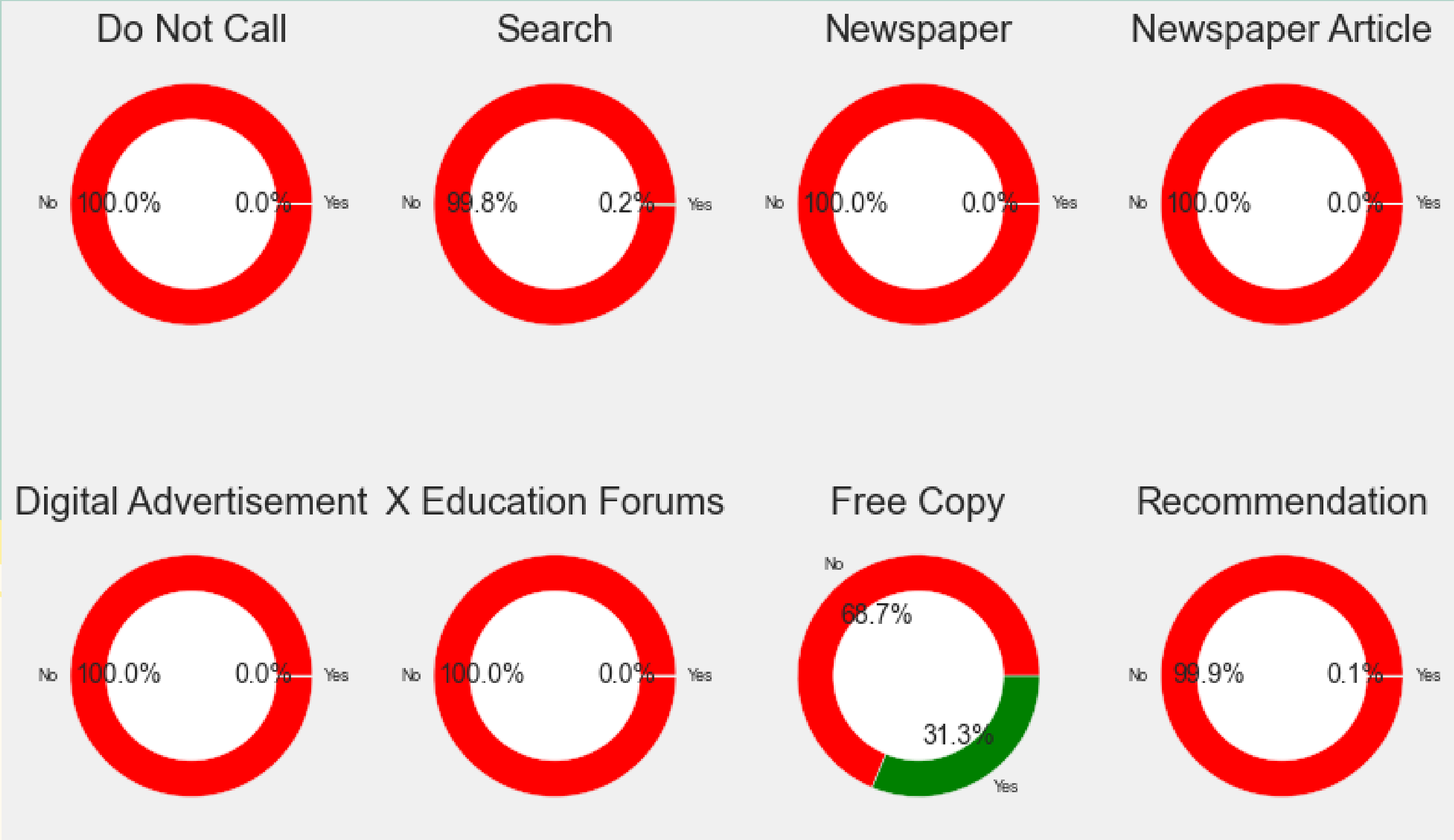


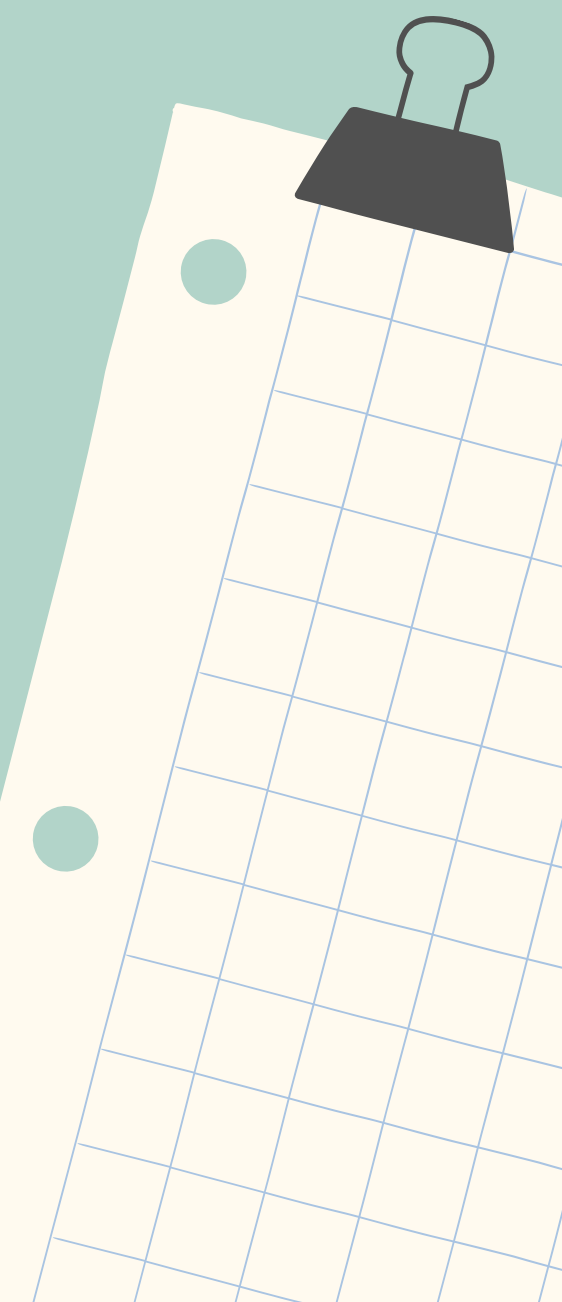
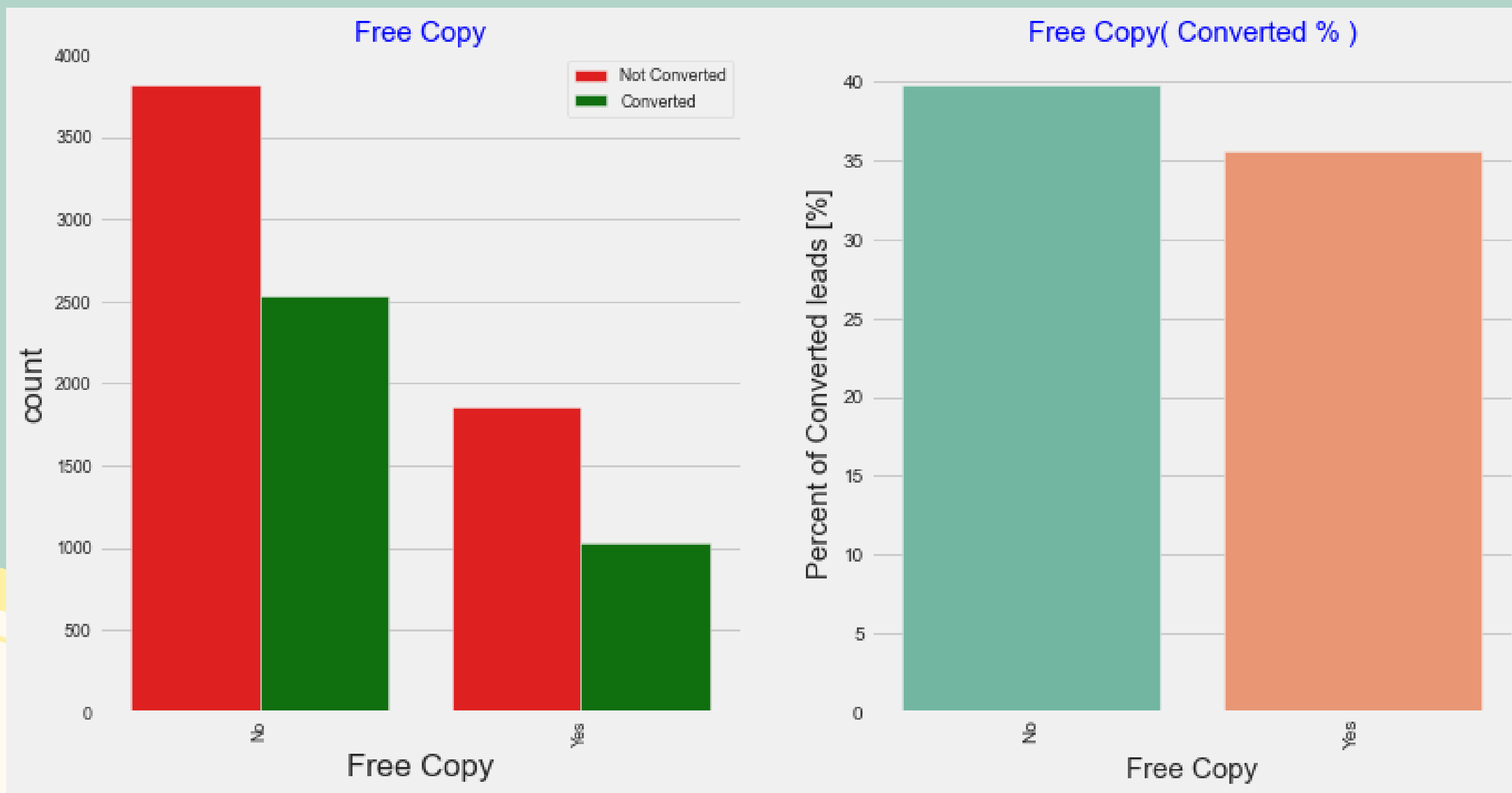
Insight:

- Though Housewives are less in numbers, they have 100% conversion rate
- Working professionals, Businessmen and Other category have high conversion rate
- Though Unemployed people have been contacted in the highest number, the conversion rate is low (~40%)
- We cannot combine small value categories as their conversion rate is very different. Combining them may provide wrong predictions.

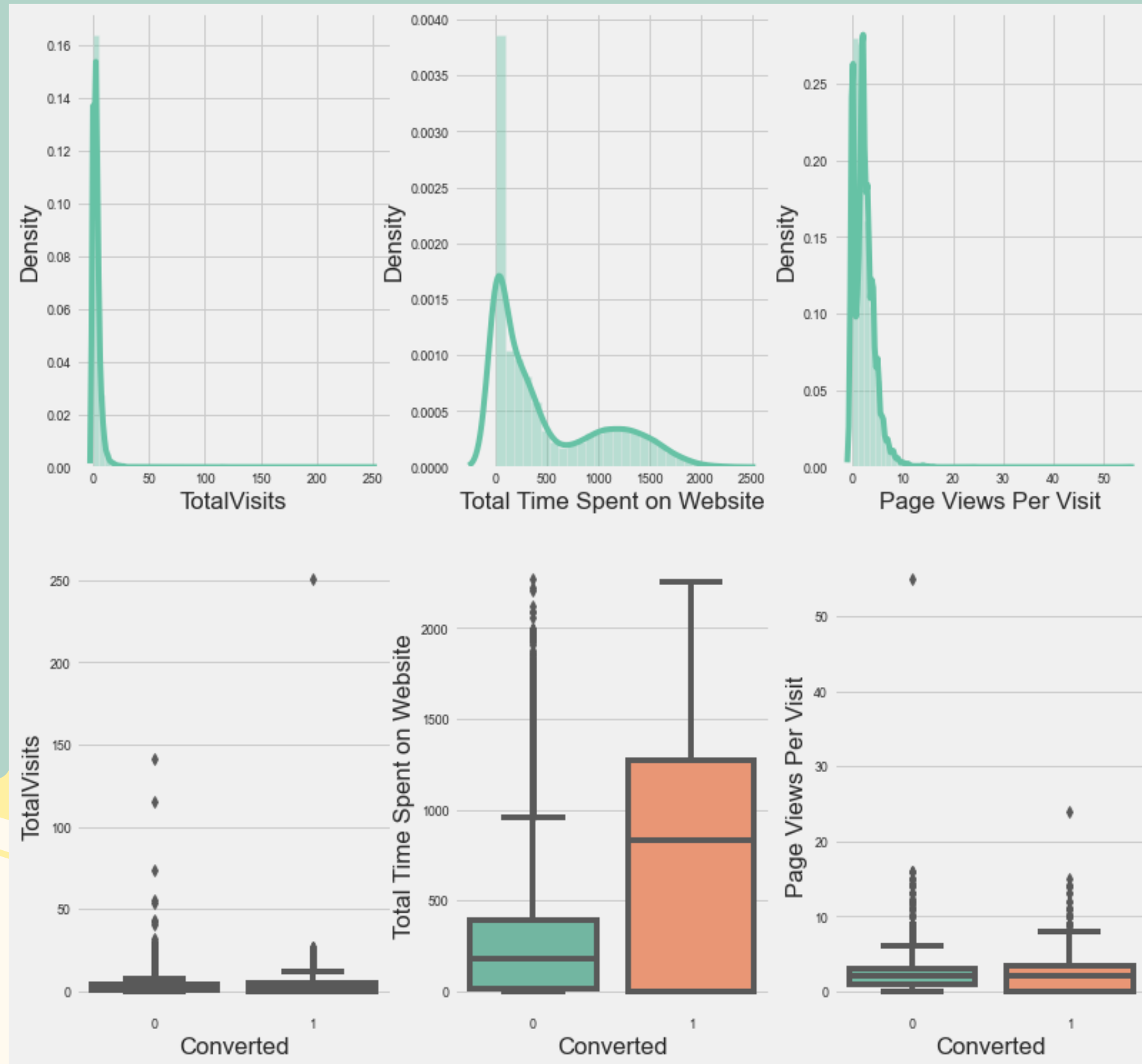
7 .Search, Newspaper, Newspaper Article, Digital Advertisement, ,X Education Forums, Free Copy

The following features have two categories only. We are going to evaluate the skewness of the data and decide whether to exclude them from model building.





- **Univariate Analysis - Numerical**

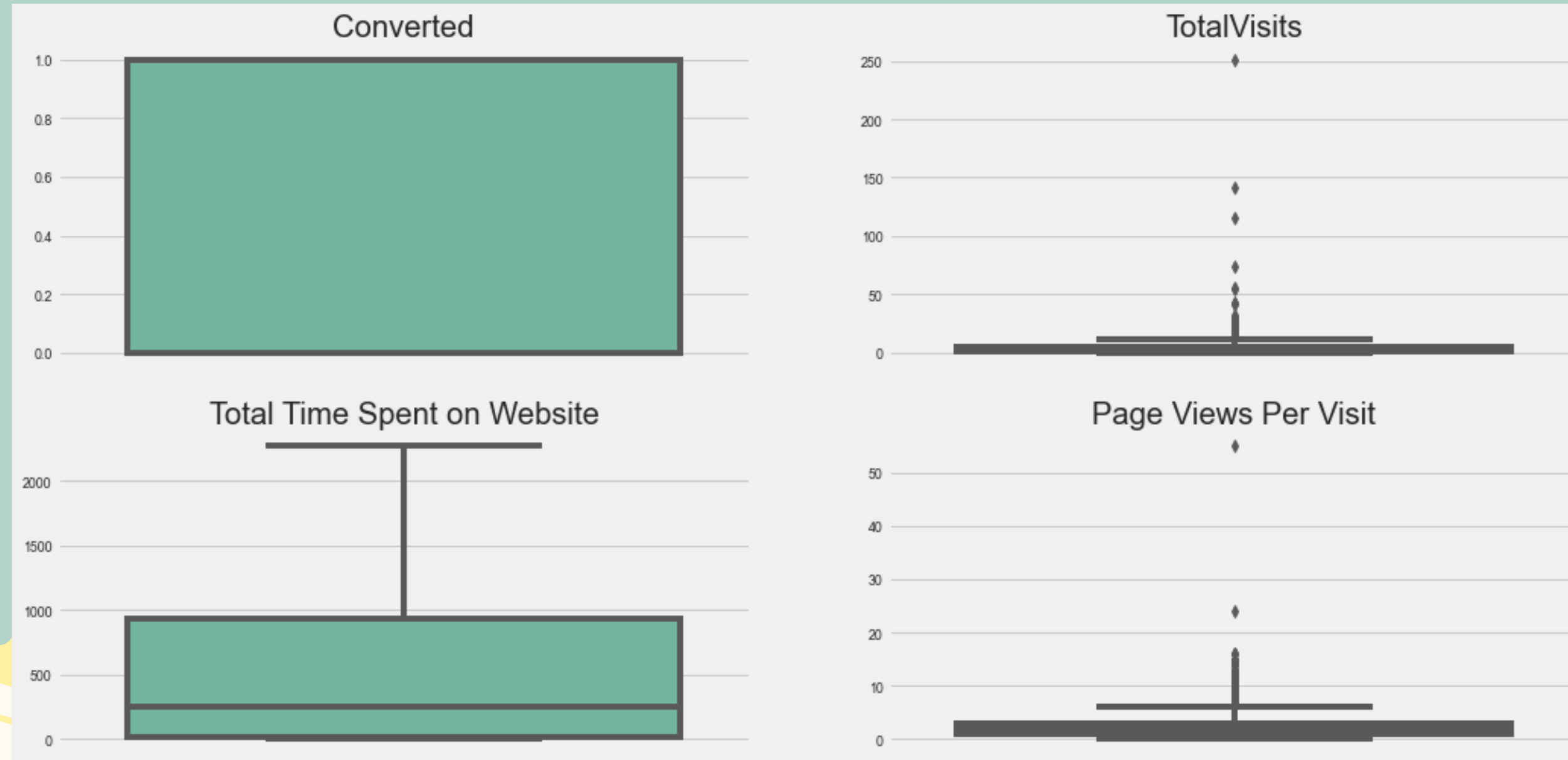


Insight:

TotalVisits and Page Views per Visit has some outliers which needs to be treated.

• Data Preparation

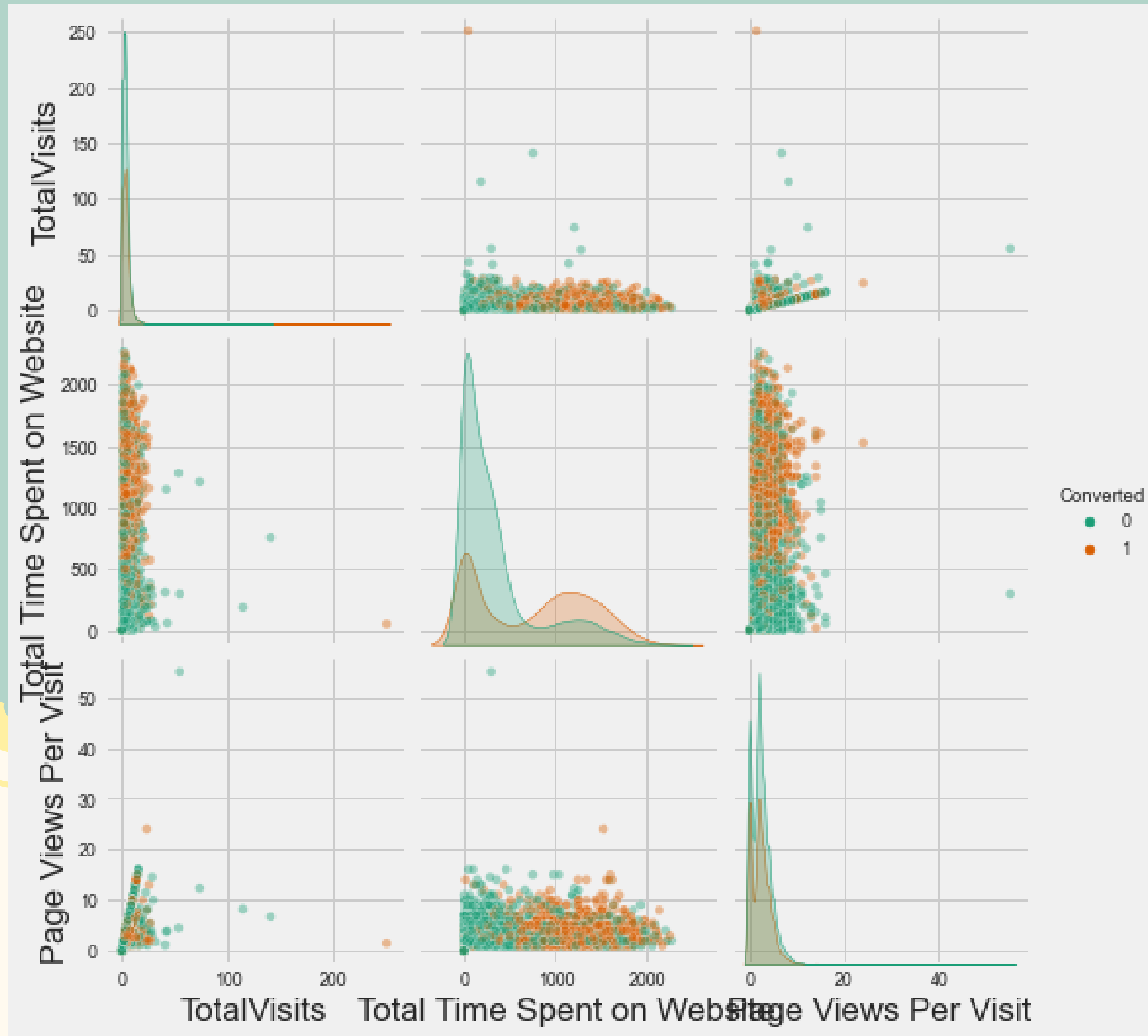
1.Outliers Treatment



Insight:

- Though outliers in TotalVisits and Page Views Per Visit shows valid values, this will misclassify the outcomes and consequently create problems when making inferences with the wrong model. Logistic Regression is heavily influenced by outliers. So lets cap the TotalVisits and Page Views Per Visit to their 95 th percentile due to following reasons:
- Data set is fairly high number
- 95th percentile and 99th percentile of these columns are very close and hence impact of capping to 95th or 99th percentile will be the same

- **Bivariate Analysis**



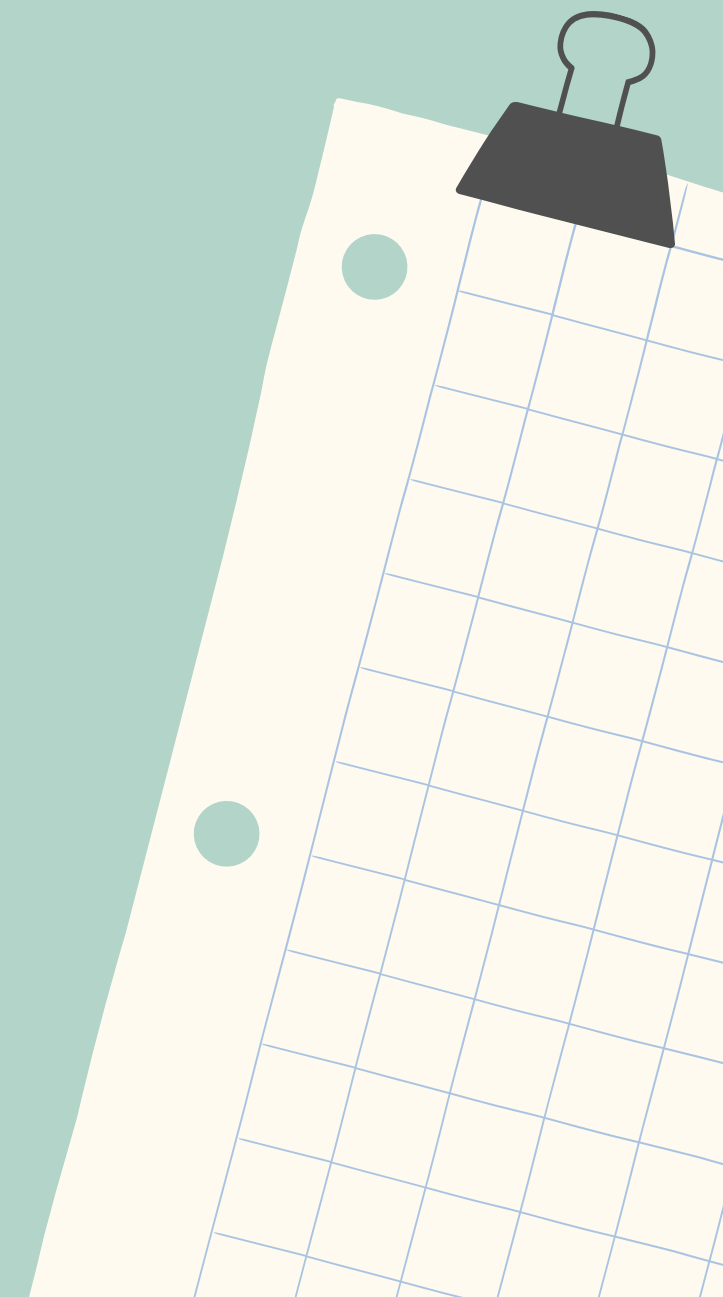
Insight:

Data is not normally distributed.

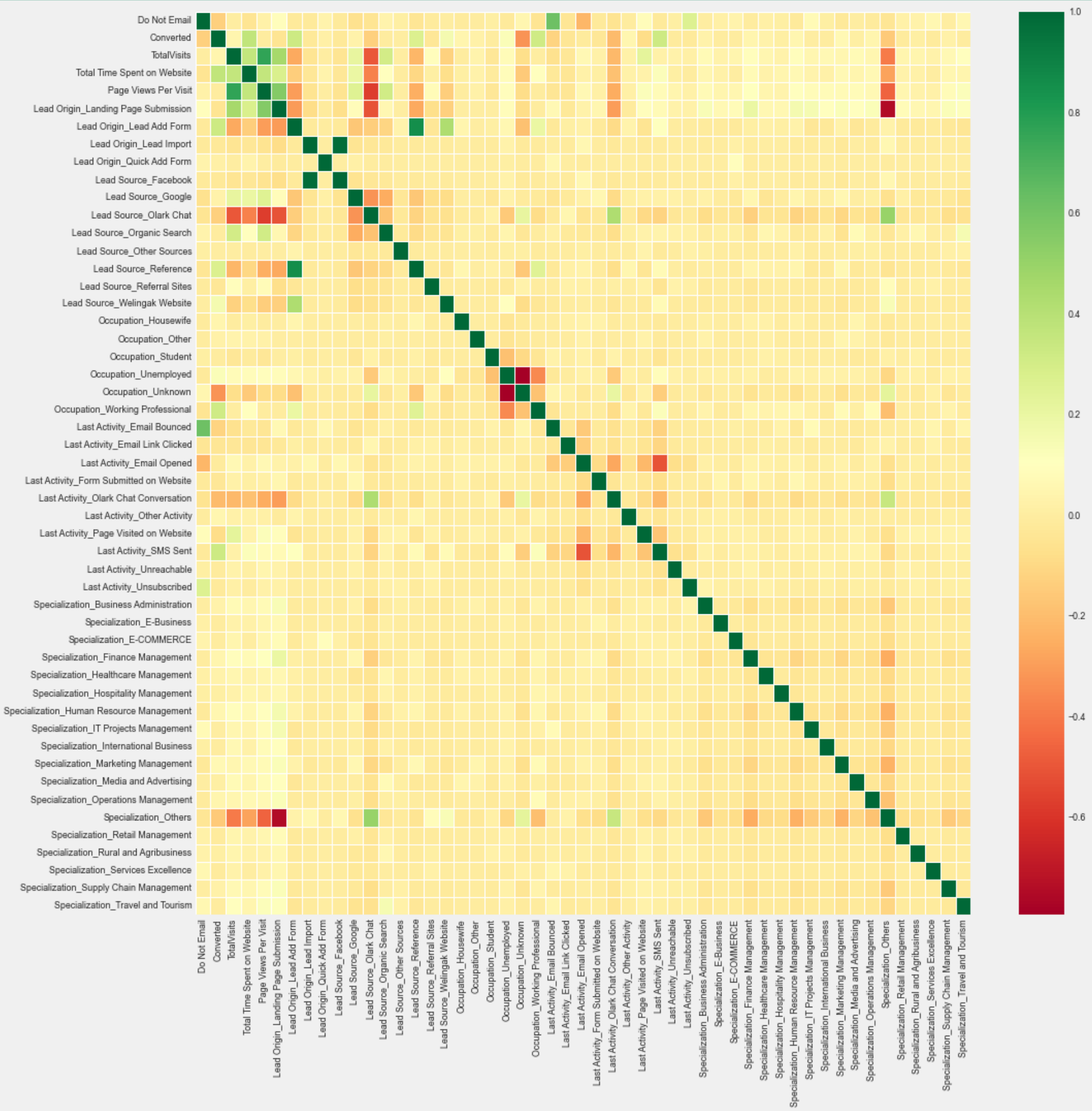


Insight:

Now that we have capped the outliers, let's proceed to data preparation for model building.

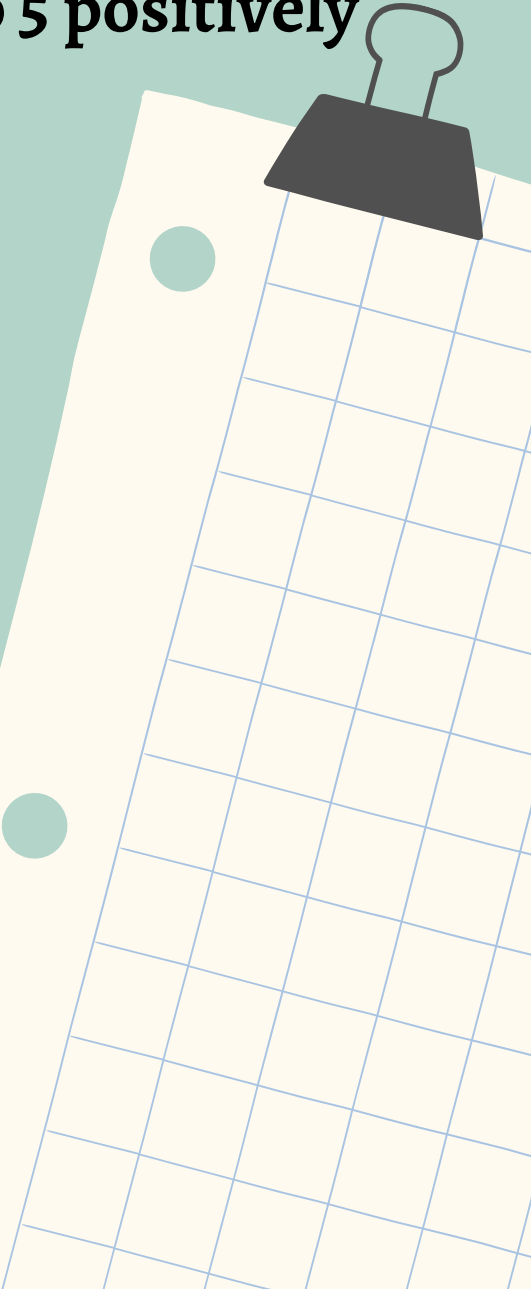


• Correlation heatmap



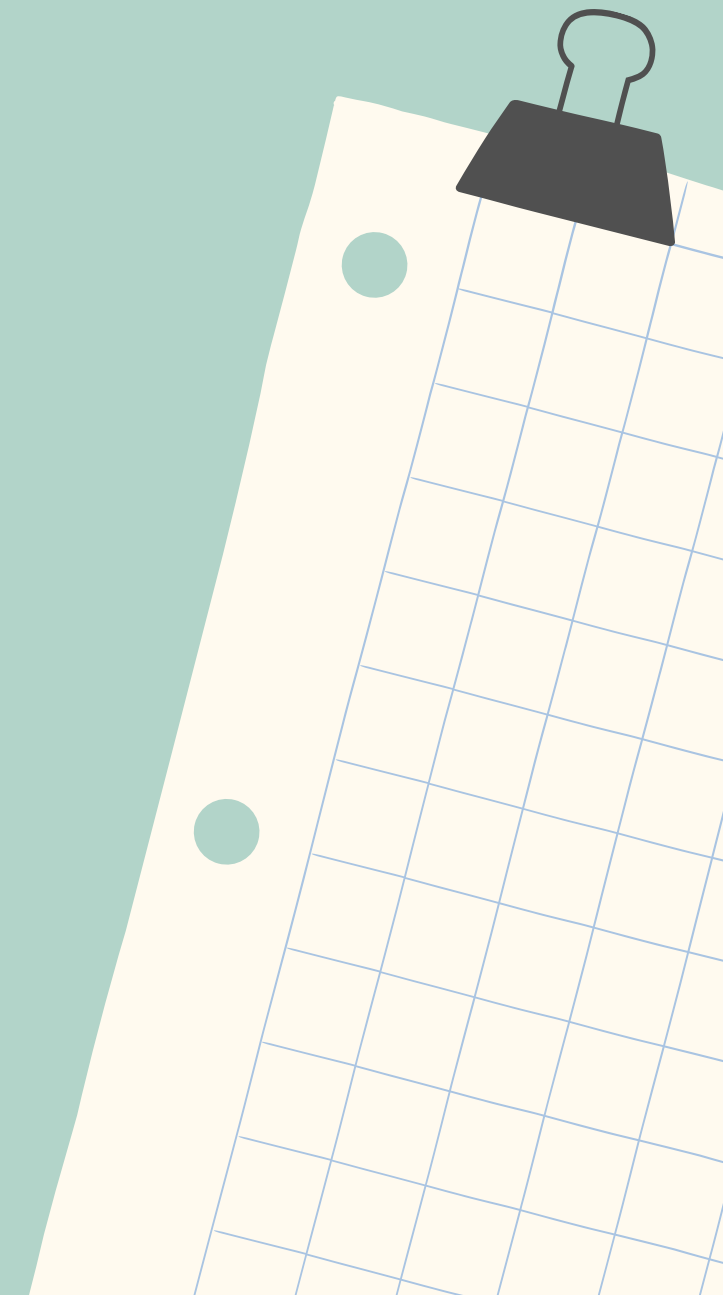
Insight:

There are 51 columns in Heatmap which makes it difficult to interpret. Let's review top 5 positively and negatively correlated features.

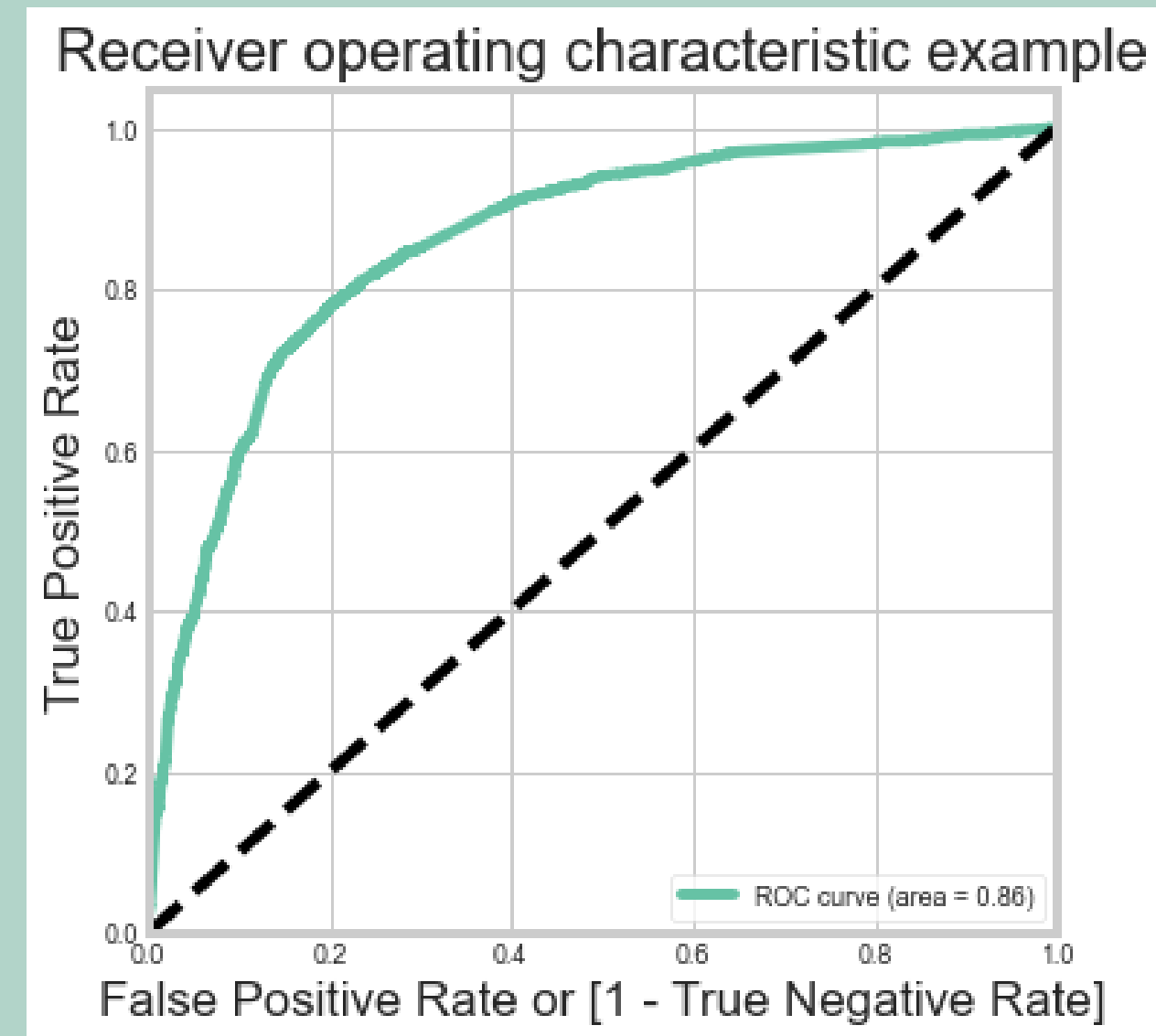
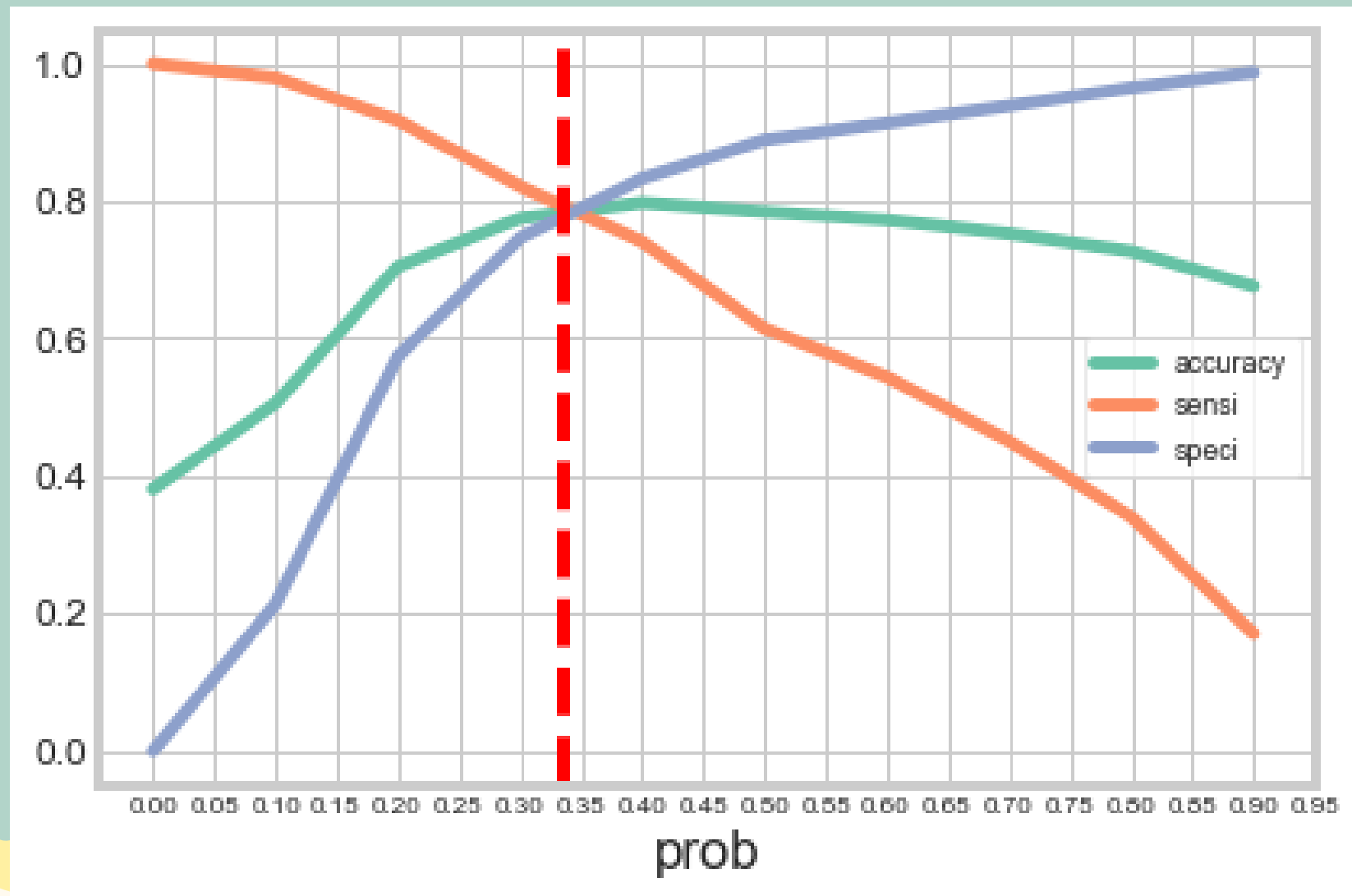


Model Building

- Splitting into train and test set.
- Scale variables in train set.
- Build the first model.
- Use RFE to eliminate less relevant variables.
- Build the next model.
- Eliminate variables based on high p-values.
- Check the VIF value for all the existing columns.
- Predict using train set.
- Evaluate the accuracy and another metric.
- Predict using a test set.
- Precision and recall analysis on test predictions



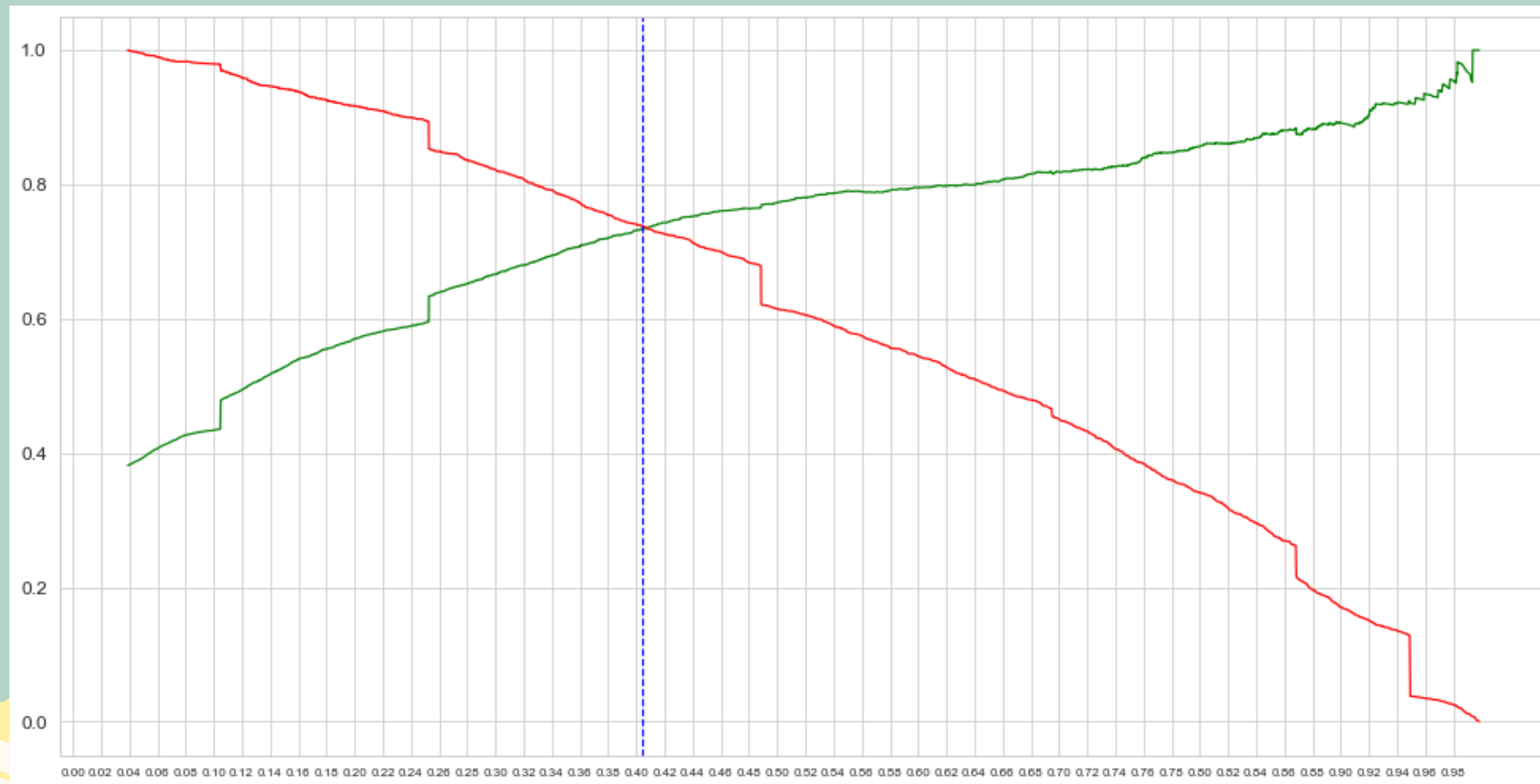
Model Evaluation



Inferences:

ROC Curve aread is o.88, which indicates that the model is good.

Precision - Recall Trade off



Inferences:

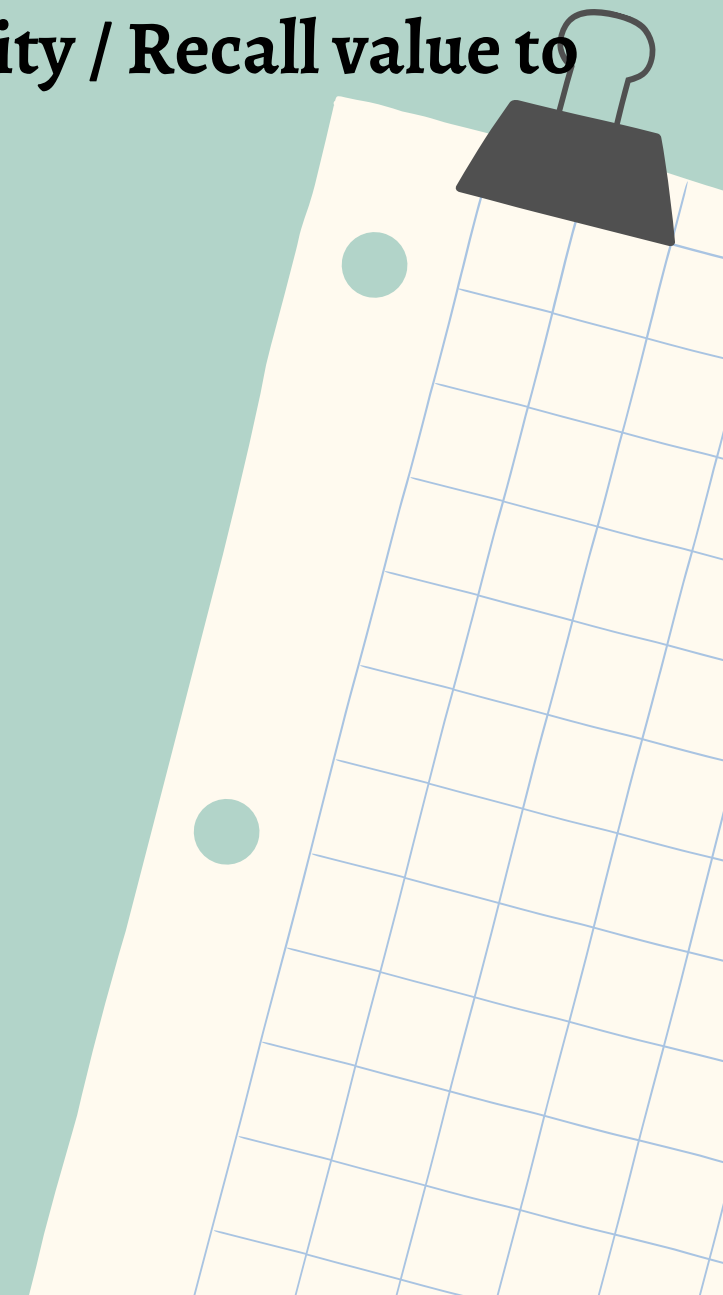
Based on Precision- Recall Trade off curve, the cutoff point seems to 0.404. We will use this threshold value for Test Data Evaluation

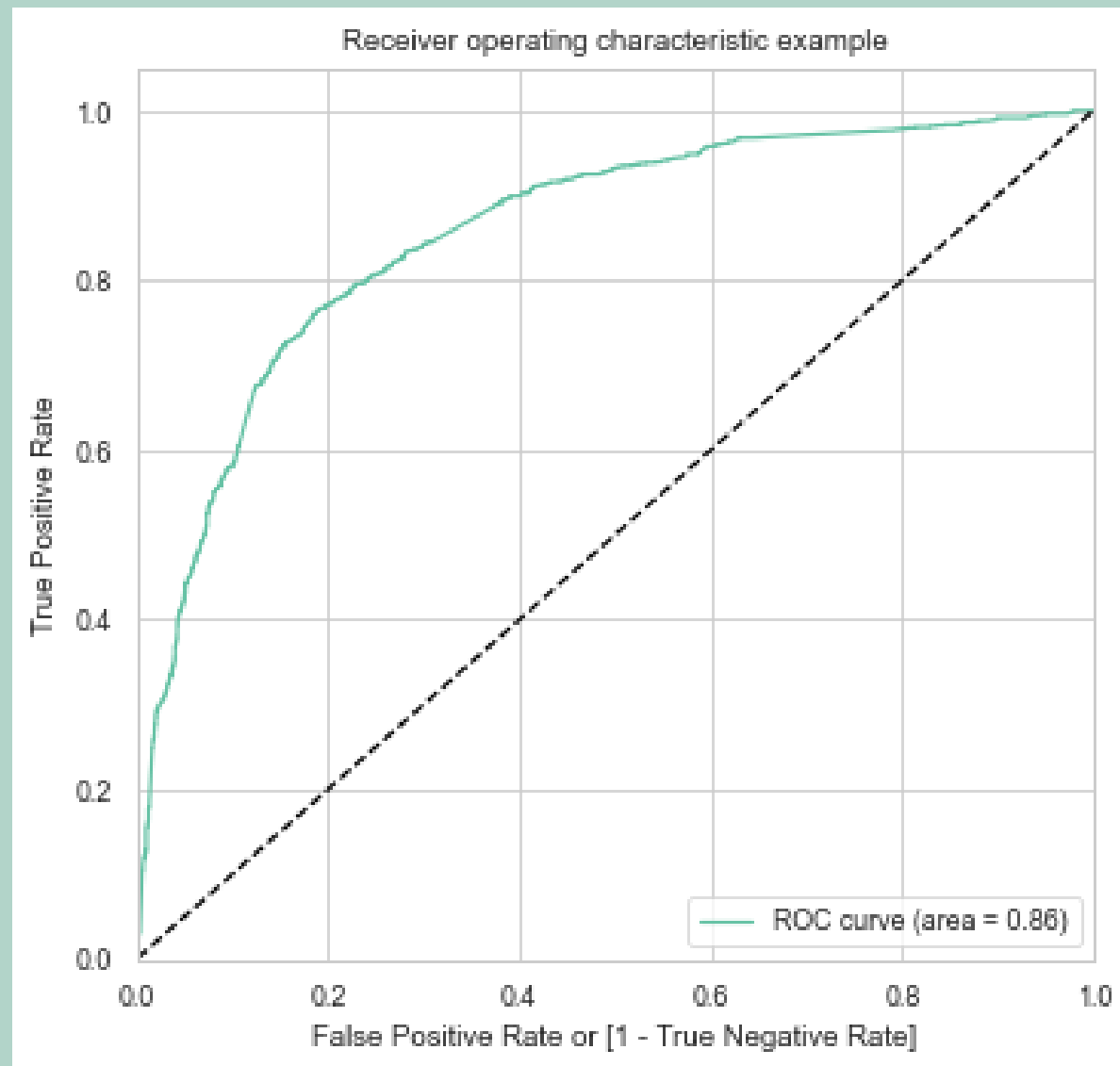
Inferences:

By using the Precision-Recall trade off chart cut-off points, the model output has changed in the following way :

- **The true Positive number has decreased.**
- **True Negative number has increase**
- **False Negative number has increase**
- **False Positive number has decreased**

For our purpose CEO wants to identify the people correctly who will convert to leads. Thus, we cannot use the Precision-Recall trade-off method as it reduced True Positive. We have to increase the Sensitivity / Recall value to increase True Positives. Thus we will use 0.335 as cutoff point.





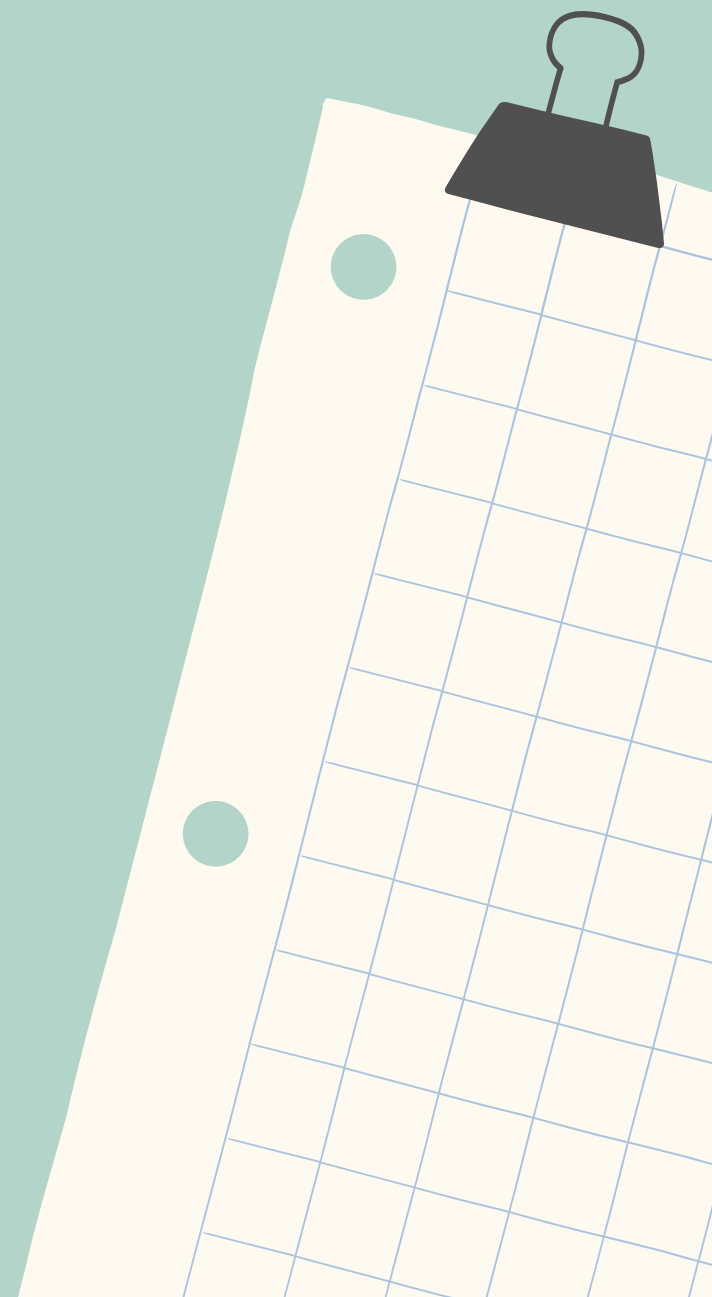
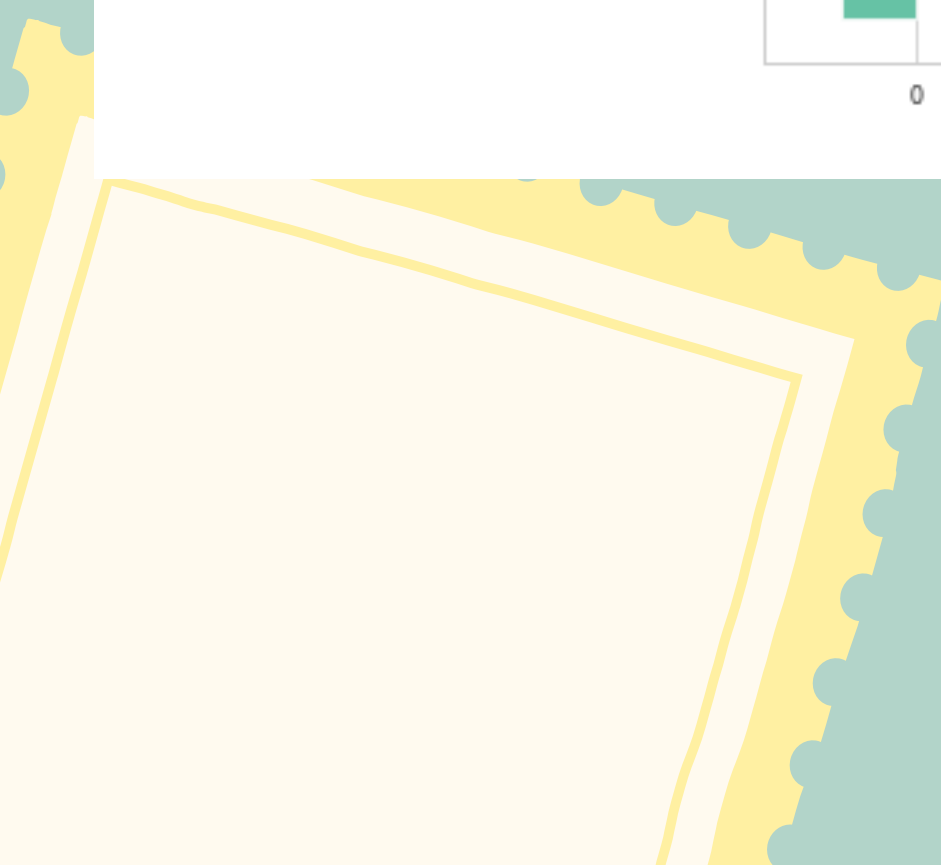
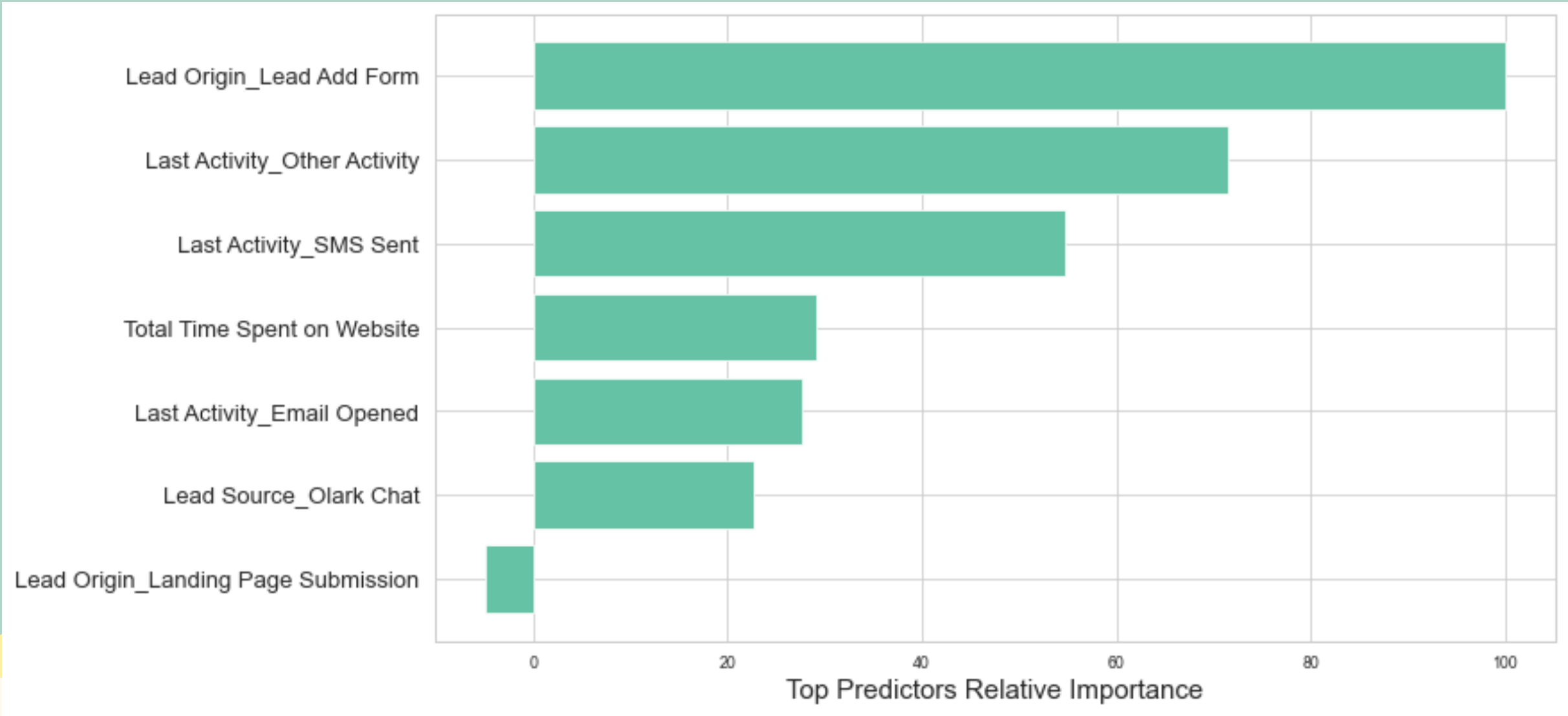
Inferences:

ROC value of 0.88 shows the model is performing well in Test dataset.

We can use the lead_score column to identify which potential leads to prioritize first. The higher the score, the higher chances are there for the lead to convert. If there are limited sales representatives, then score cut-off should be higher to ensure a higher conversion probability people are contacted further to turn them into a potential customer. It is the same as increasing the precision value of the model by adjusting the cut-off point to a higher value. In case there are more resources available in the sales team (i.e., interns, etc.), then the score cut-off can be lowered. As there are more human resources, the company can afford a higher rate of False positives as it will increase the customer outreach and, in turn, increase the potential customer who will take the online courses.

Conclusion:

1 Model Features / Predictors



Model Summary

Interpretation Logistic regression model with multiple predictor variables.

In general, we can have multiple predictor variables in a logistic regression model as below:

$$\text{logit}(p) = \log(p/(1-p)) = \beta_0 + \beta_1 * X_1 + \dots + \beta_n * X_n$$

Applying such a model to our example dataset, each estimated coefficient is the expected change in the log odds of being a potential lead for a unit increase in the corresponding predictor variable holding the other predictor variables constant at a certain value. Each exponentiated coefficient is the ratio of two odds, or the change in odds in the multiplicative scale for a unit increase in the corresponding predictor variable holding other variables at a certain value.

The magnitude and sign of the coefficients loaded in the logit function:

$$\text{logit}(p) = \log(p/(1-p)) = (3.42 * \text{Lead Origin_Lead Add Form}) + (2.84 * \text{Occupation_Working Professional}) + (1.99 * \text{Lead Source_Welingak Website}) + (1.78 * \text{Last Activity_SMS Sent}) + (1.25 * \text{Last Activity_Unsubscribed}) + (1.09 * \text{Total Time Spent on Website}) + (0.98 * \text{Lead Source_Olark Chat}) + (0.84 * \text{Last Activity_Unreachable}) + (0.66 * \text{Last Activity_Email Opened}) - (0.25 * \text{Lead Origin_Landing Page Submission}) - (0.87 * \text{Last Activity_Olark Chat Conversation}) - (1.26 * \text{Do Not Email}) - 1.77$$

We can make predictions from the estimates. We do this by computing the effects for all of the predictors for a particular scenario, adding them up, and applying a logistic transformation.
Consider the scenario of a lead who is a working professional and who was identified from Welingak website and who had chatted on Olark Chat and who spent no time on the website and wanted to be contacted by E-mail.

Then we can calculate his conversion probability as $3.42 * 0 + 2.84 * 1 + 1.99 * 1 + 1.78 * 0 + 1.25 * 0 + 1.09 * 0 + 0.98 * 0 + 0.84 * 0 + 0.66 * 0 - 0.25 * 0 - 0.87 * 1 - 1.26 * 0 - 1.77 = 2.84 + 1.99 - 0.87 - 1.77 = 2.19$ which is $\log(p/(1-p))$.
The logistic transformation is:
 $\text{Probability} = 1 / (1 + \exp(-x)) = 1 / (1 + \exp(-2.19)) = 1 / (1 + \exp(2.2)) = 0.10 = 10\%$

Predicting Probabilities

We can make predictions from the estimates. We do this by computing the effects for all of the predictors for a particular scenario, adding them up, and applying a logistic transformation.

Consider the scenario of a lead who is a working professional and who was identified from Welingak website and who had chatted on Olark Chat and who spent no time on the website and wanted to be contacted by E-mail.

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 = $1 / (1 + \exp(-x)) = 1 / (1 + \exp(-2.2)) = 1 / (1 + \exp(2.2)) = 0.143 = 14.3\%$.

Odds ratios

Sometimes, marketing team may need to get odds rather than probabilities as the concept of odds ratios is of sociological rather than logical importance.

To understand odds ratios we first need a definition of odds, which is the ratio of the probabilities of two mutually exclusive outcomes. Consider our prediction of the probability of lead conversion of 10% from the earlier section on probabilities. As the probability of lead conversion is 10%, the probability of non-conversion is $100\% - 10\% = 90\%$, and thus the odds are 10% versus 90%. Dividing both sides by 90% gives us 0.11 versus 1, which we can just write as 0.11. So, the odds of 0.11 is just a different way of saying a probability of lead conversion of 10%.

Similarly We can interpret from the model that, holding all categorical and numerical variables at a fixed value, the odds of a lead being converted for a Working Professional (Working Professional = 1) over the odds of lead being converted for non-working professionals (Working Professional = 0) is $\exp(.2.84) = 17.11$

This means $\log(p/(1-p)) = 17.11$ when all other variables are at fixed value

We can use this odds ratios method to identify the potential lead conversions on comparing the individuals profile.

THANK YOU FOR
LISTENING!

The image is a digital collage on a teal background. It features a central blue notepad with a white spiral binding at the top, displaying the text 'THANK YOU FOR LISTENING!' in a bold, black, sans-serif font. To the left of the notepad is a stack of papers: a cream-colored sheet with horizontal lines, a purple sheet, and a yellow sheet with black diagonal stripes. To the right is another cream-colored sheet with horizontal lines, featuring a purple oval outline and a small blue rectangular sticker at the top. In the bottom left corner, there is a yellow CD with a black barcode and the alphanumeric string '0001Q2315204L900' printed below it. The bottom right corner shows a portion of a cream-colored sheet with horizontal lines and two black dots on the left margin.