LEAD SCORING CASE STUDY

SUMMARY

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers. The company needs a model wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The basic data provided gave us a lot of information about the potential customers visit the website, the time they spend there, how they reached the website and the conversion rate. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

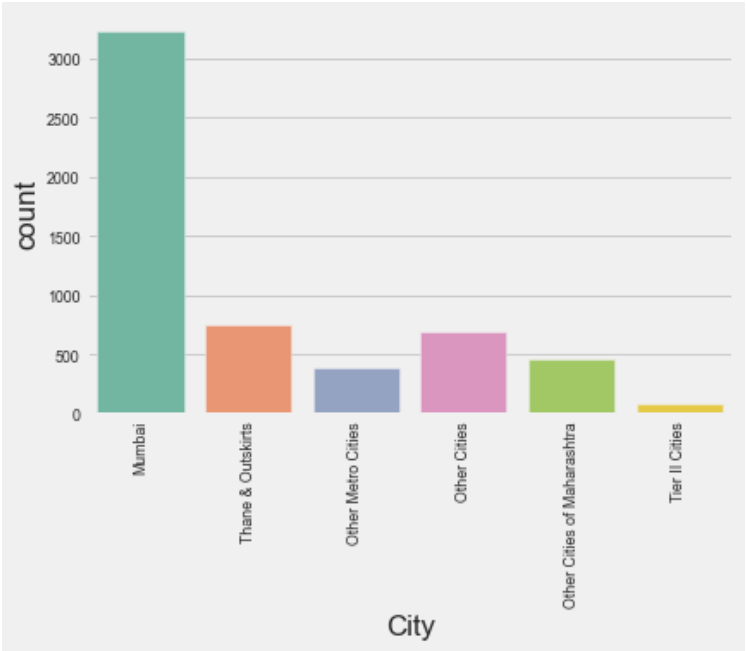
The following are the steps used:

1. Cleaning data:

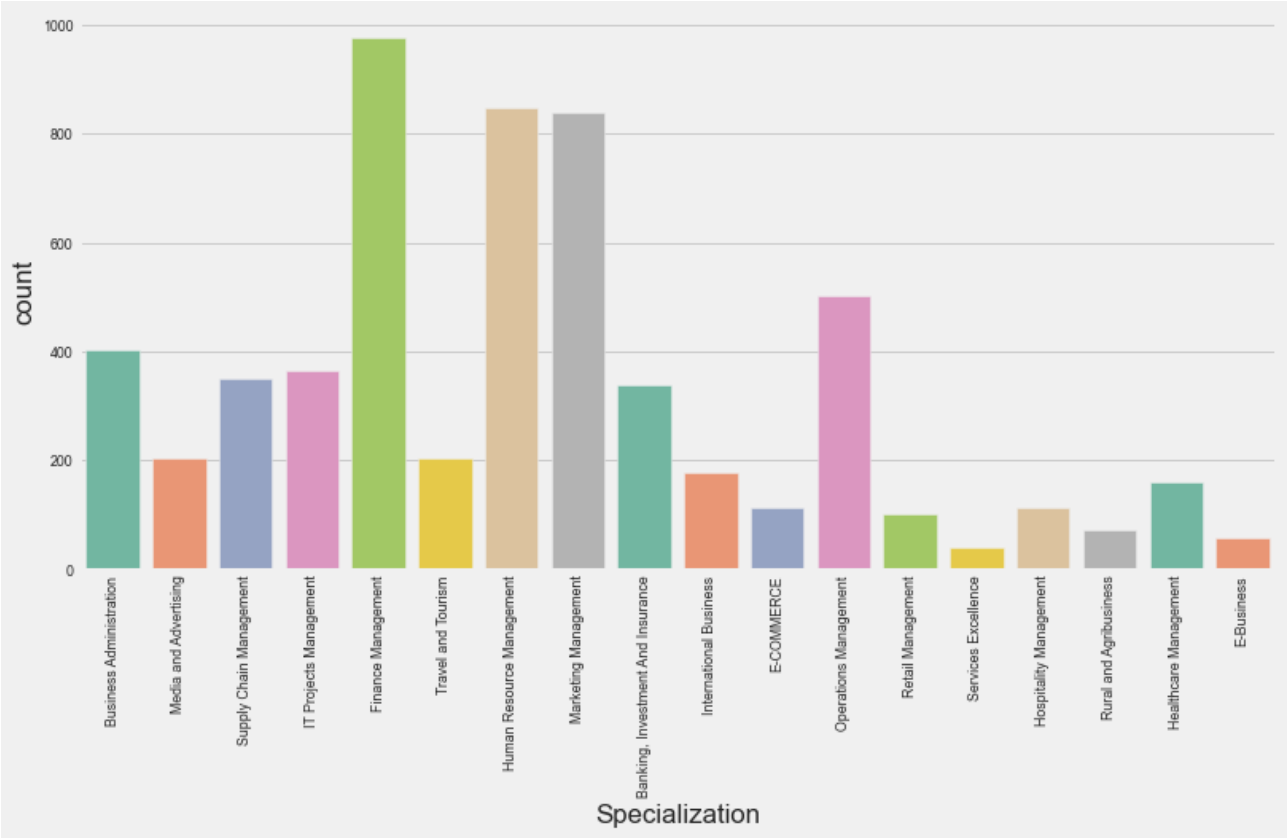
The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changes to 'not provided' so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside the elements were changed to "India", "Outside India" and "not provided".

2. Exploratory Data Analysis:

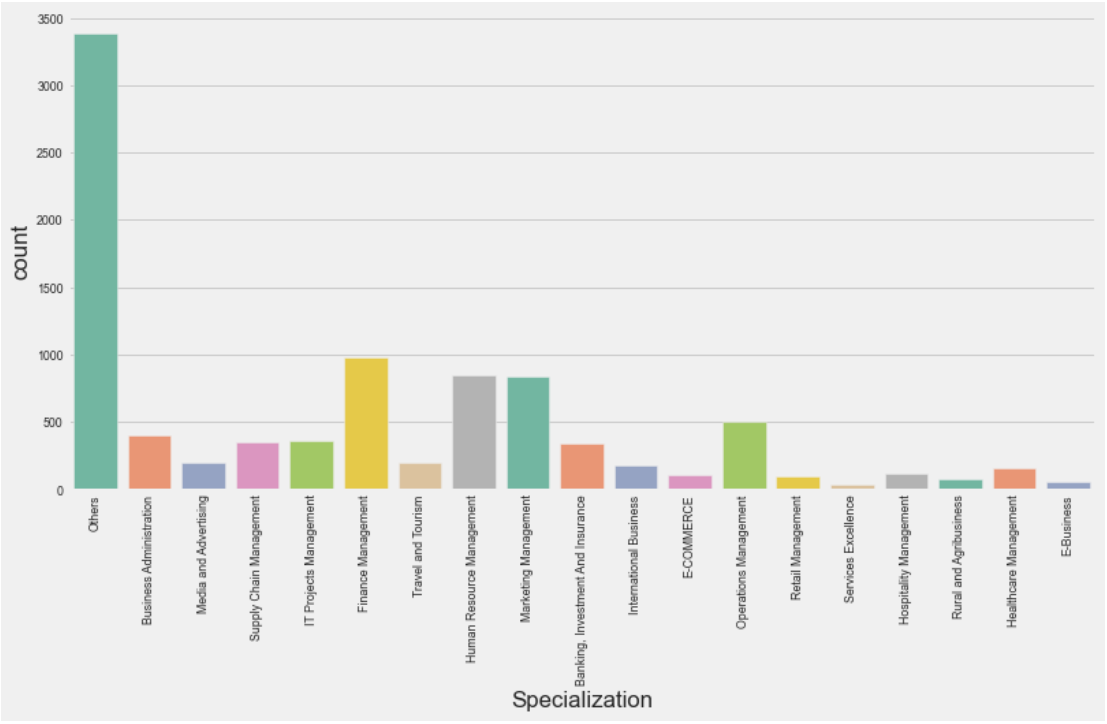
A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good and no outliers were found.

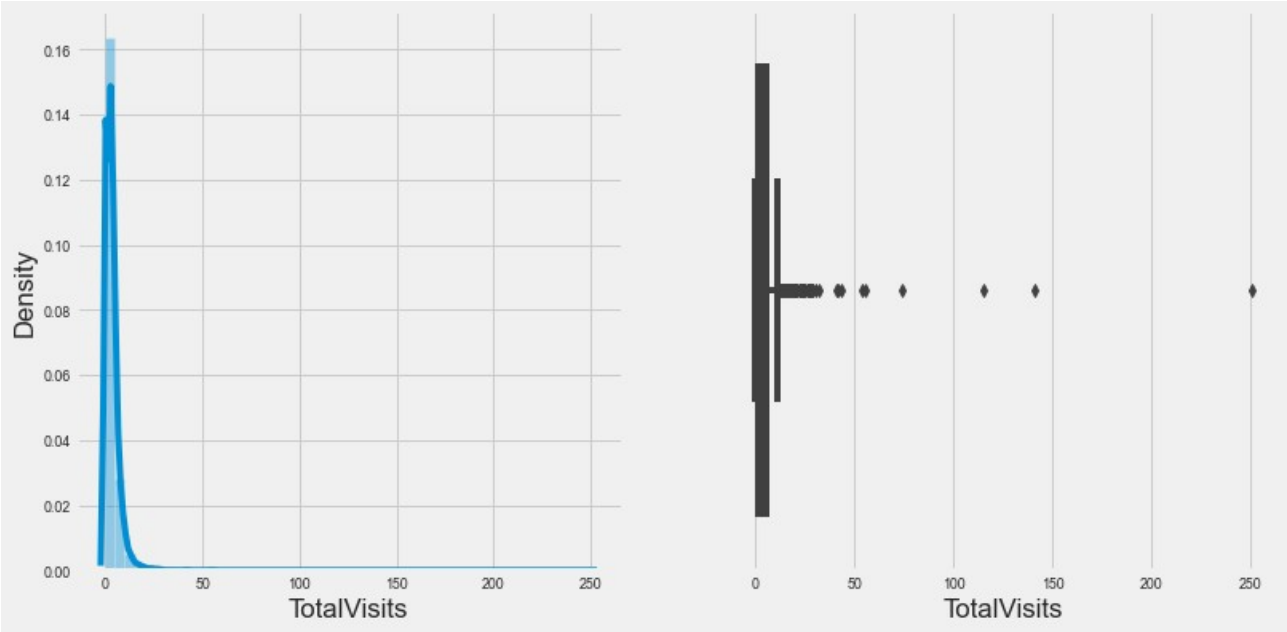


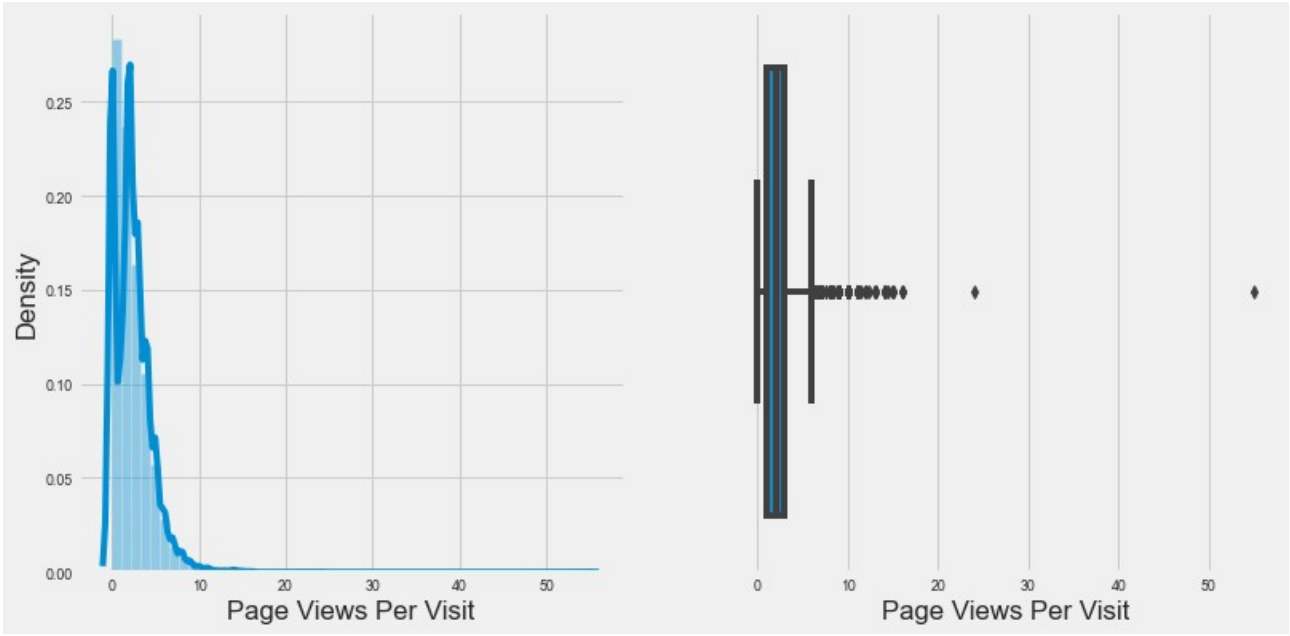
As there is almost 40% unknown values, we cannot impute with mode as it is make the whole data skewed. Also, X-Education is online teaching platform. The city information will not be much useful as potential students can available any courses online despite their city. We will drop the column from analysis.



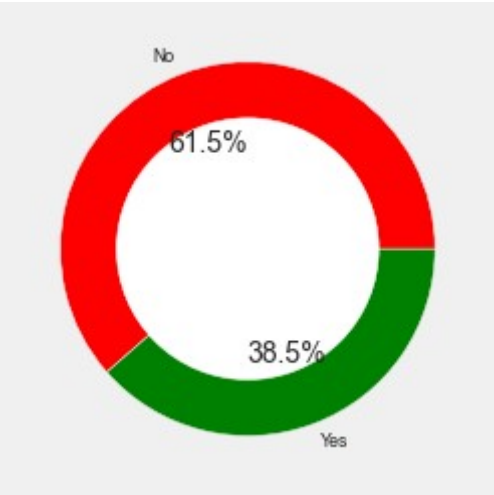
It may be possible that the lead has no specialization or may be a student and has no work experience yet , thus he/she has not entered any value. We will create a new category called 'Others' to replace the null values.

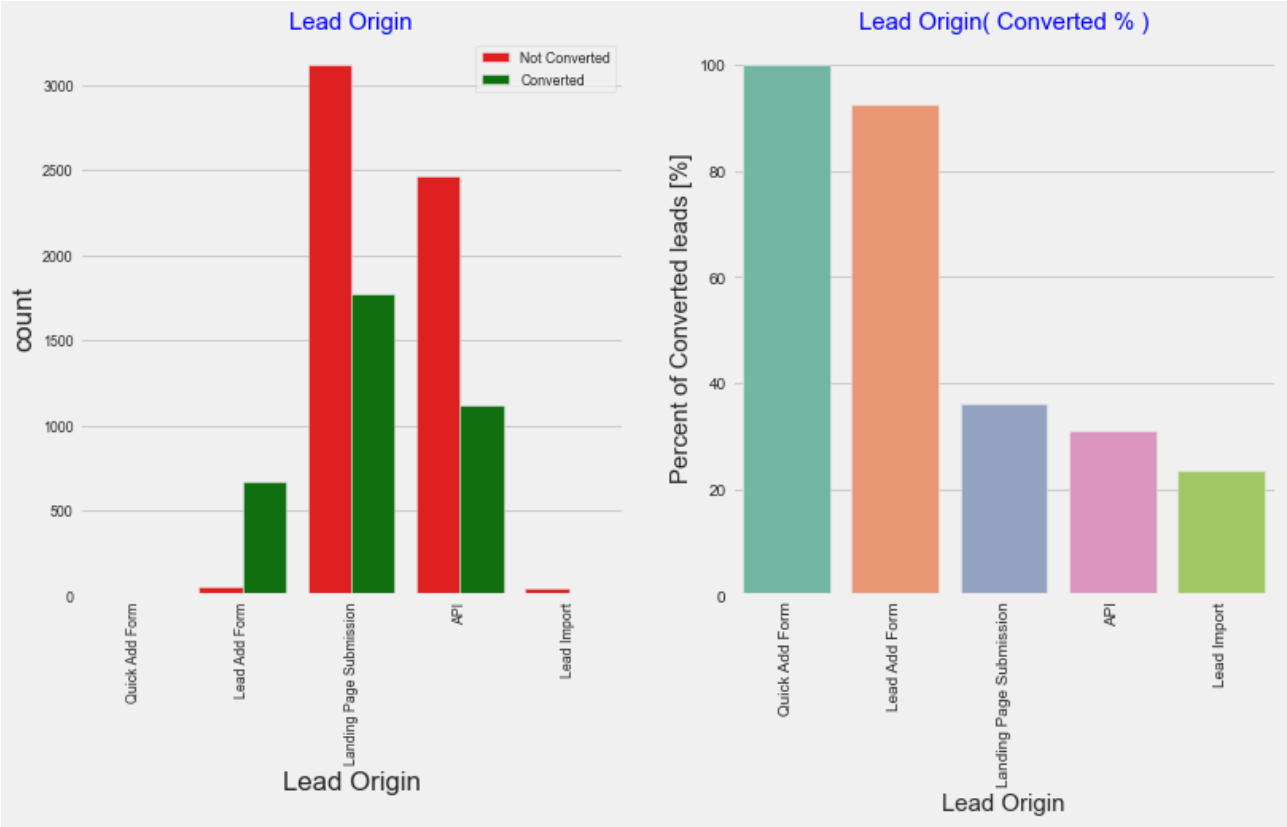






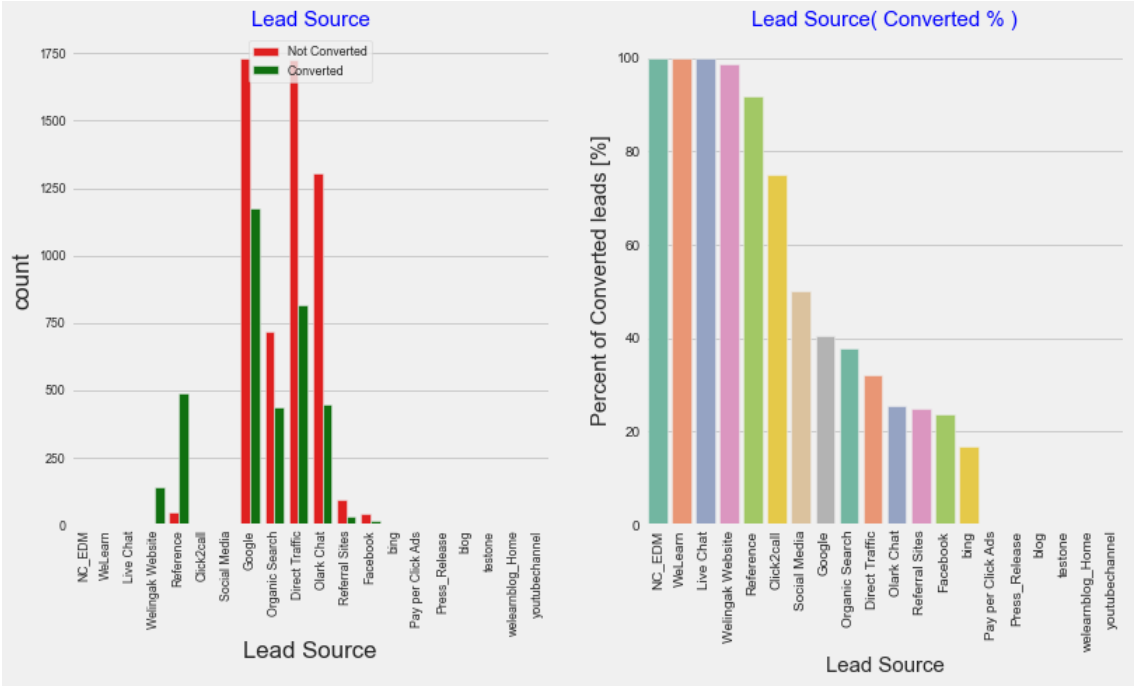
As we see there are some outliers in the data, we will impute with median and not mean value.

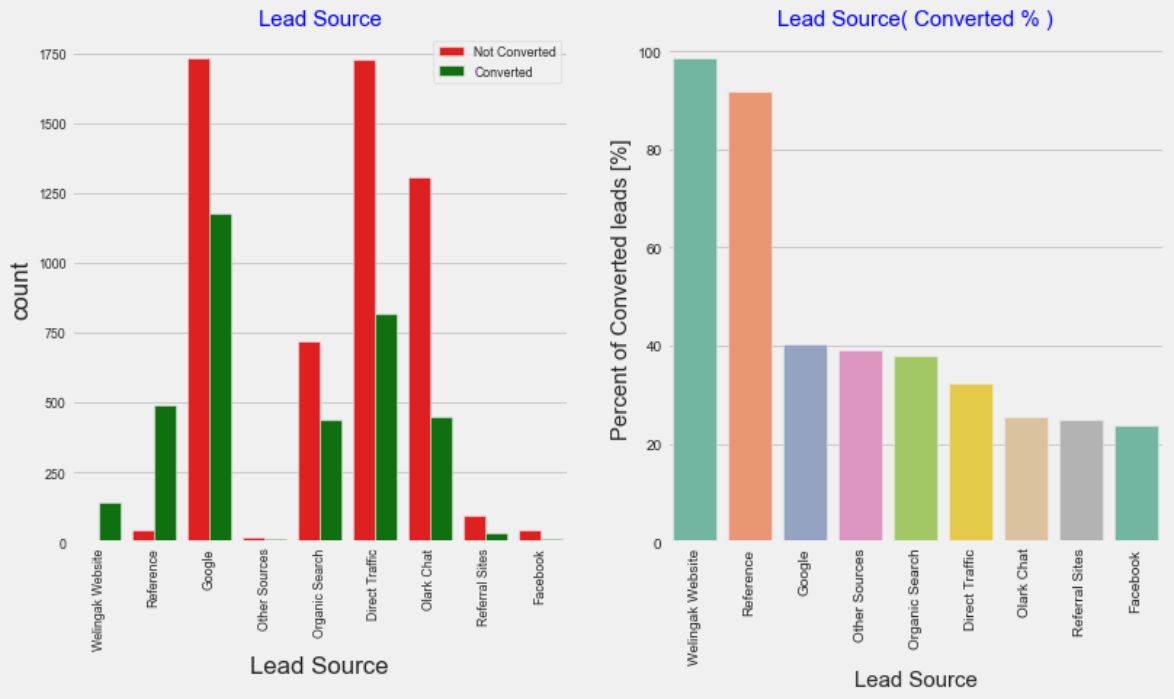




- **Most Leads originated from submissions on the landing page and around 38% of those are converted followed by API, where around 30% are converted.**
- **Even though Lead Origins from Quick Add Form are 100% Converted, there was just 1 lead from that category. Leads from the Lead Add Form are the next 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**

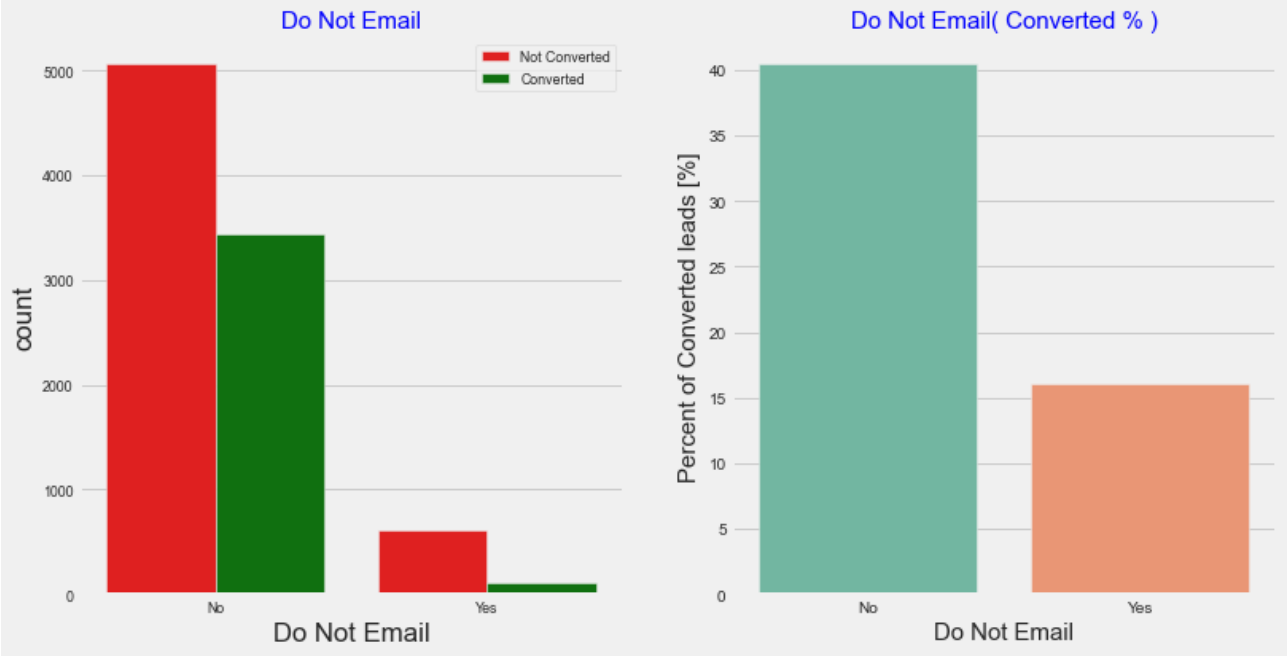
To improve overall lead conversion rate, we need to focus more on improving lead converion of API and Landing Page Submission origin and generate more leads from Lead Add Form.





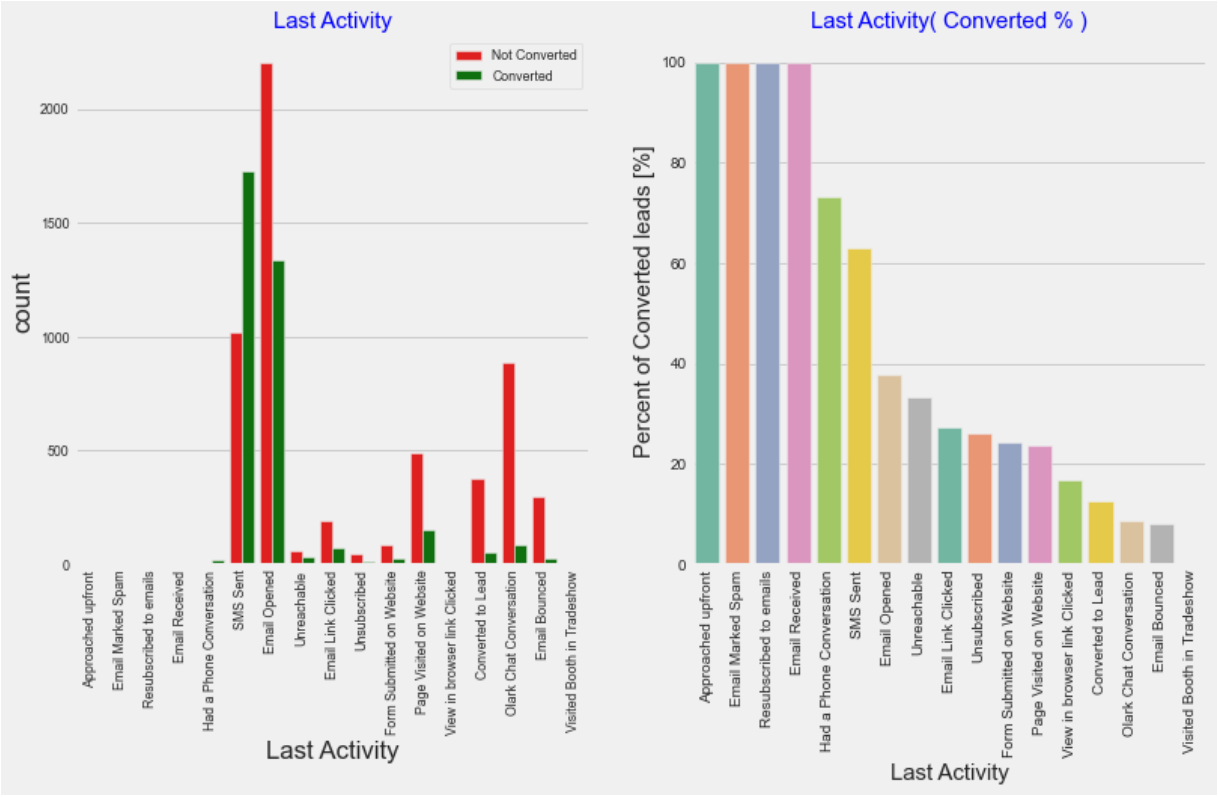
- **The source of most leads was Google, and 40% of the leads converted, followed by Direct Traffic,Organic search and Olark chat where around 35%, 38% and 30% converted respectively.**
- **A lead that came from a reference has over 90% conversion from the total of 534.**
- **Welingak Website has almost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

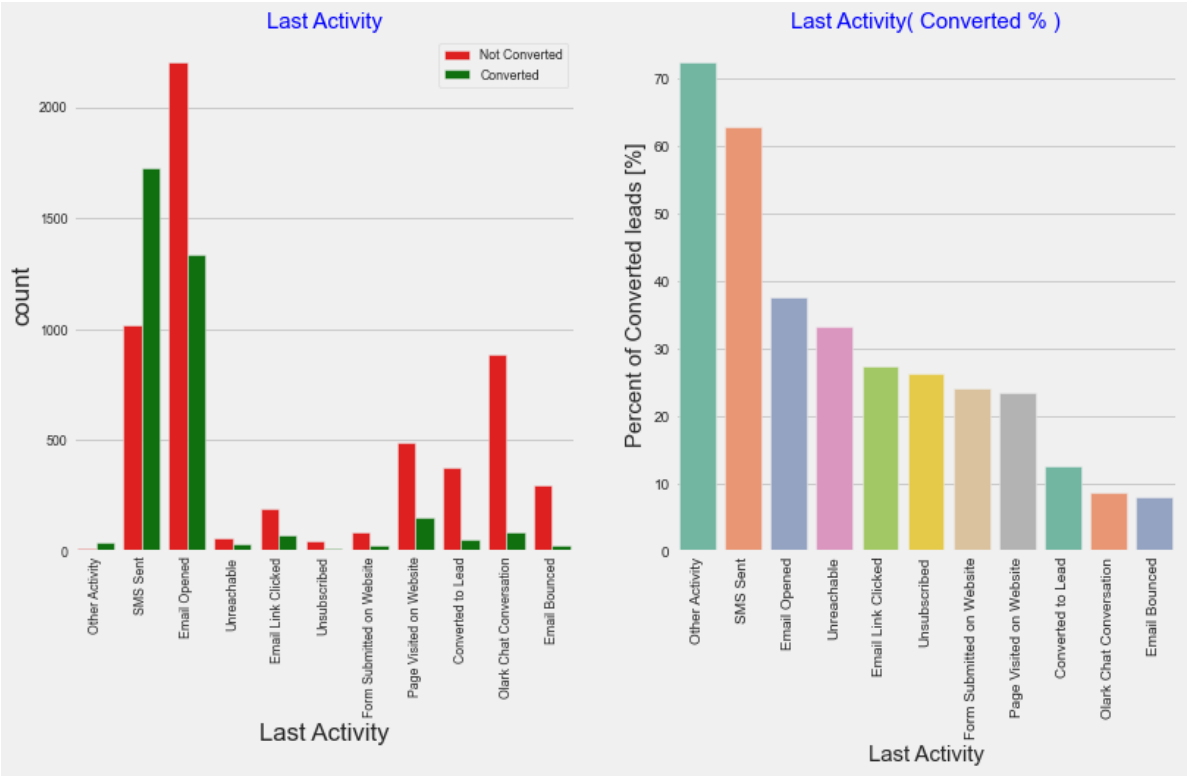
To increase lead count, initiatives should be taken so already exitsing members increase their referrals.



- **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%)**

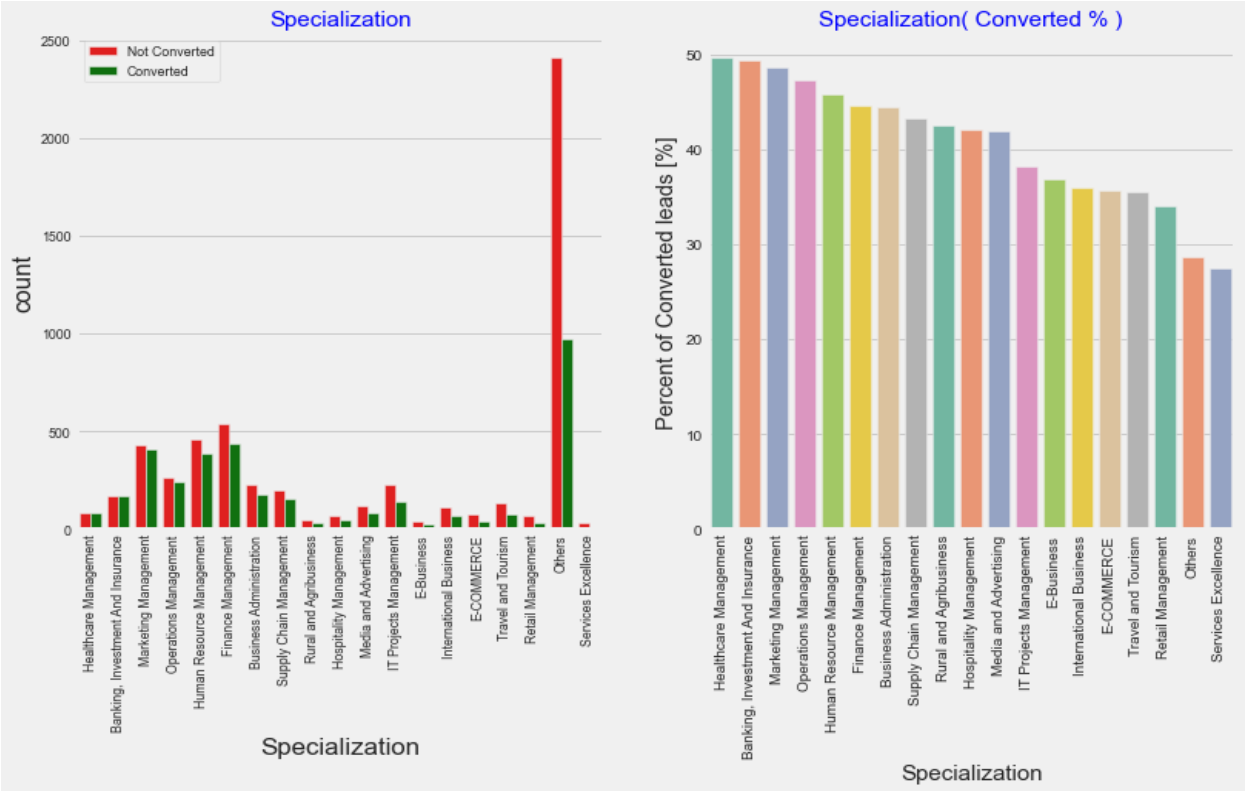




●**Most of the lead have their Email opened as their last activity**

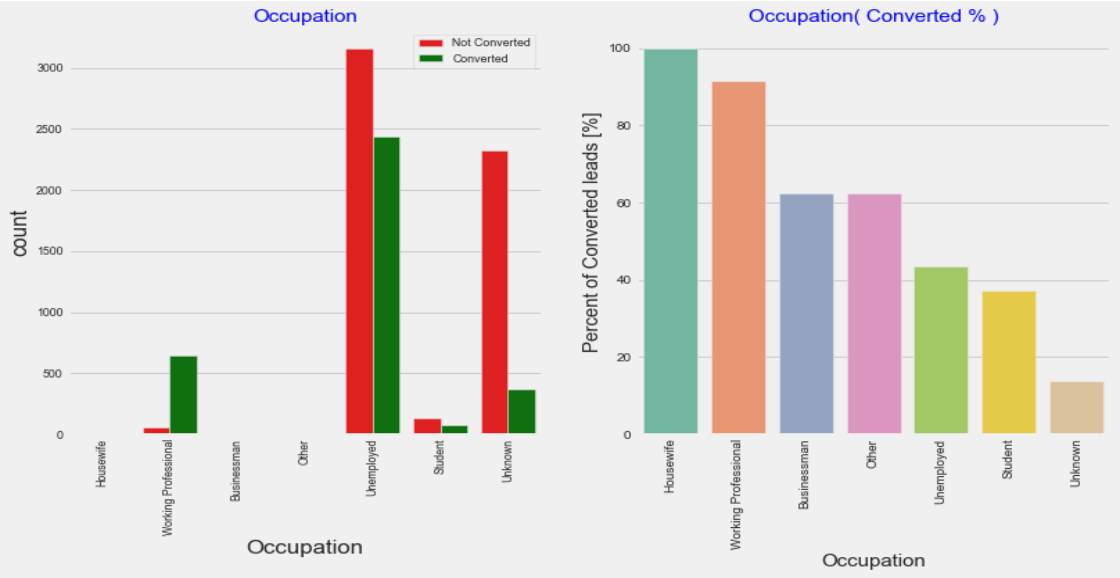
●**After combining smaller Last Activity types as Other Activity, the lead conversion is very high (~70%)**

●**Conversion rate for leads with last activity as SMS Sent is almost 60%**



●**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**

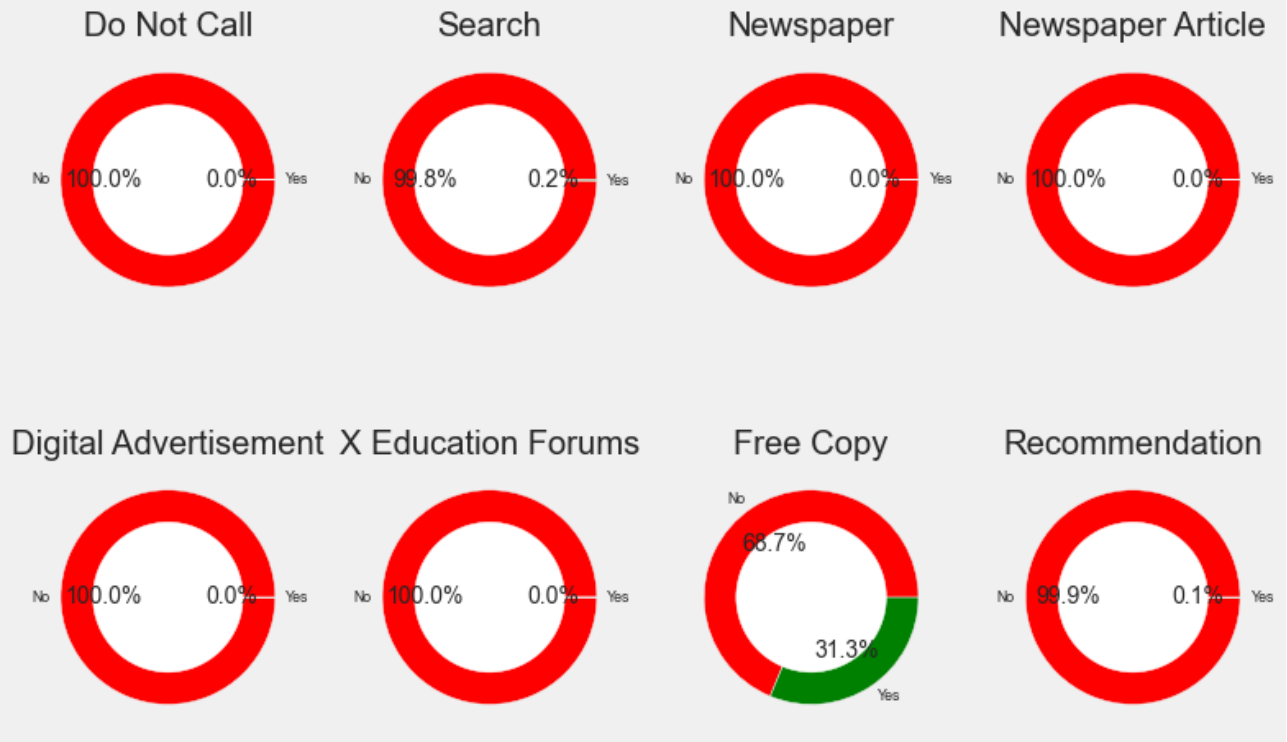


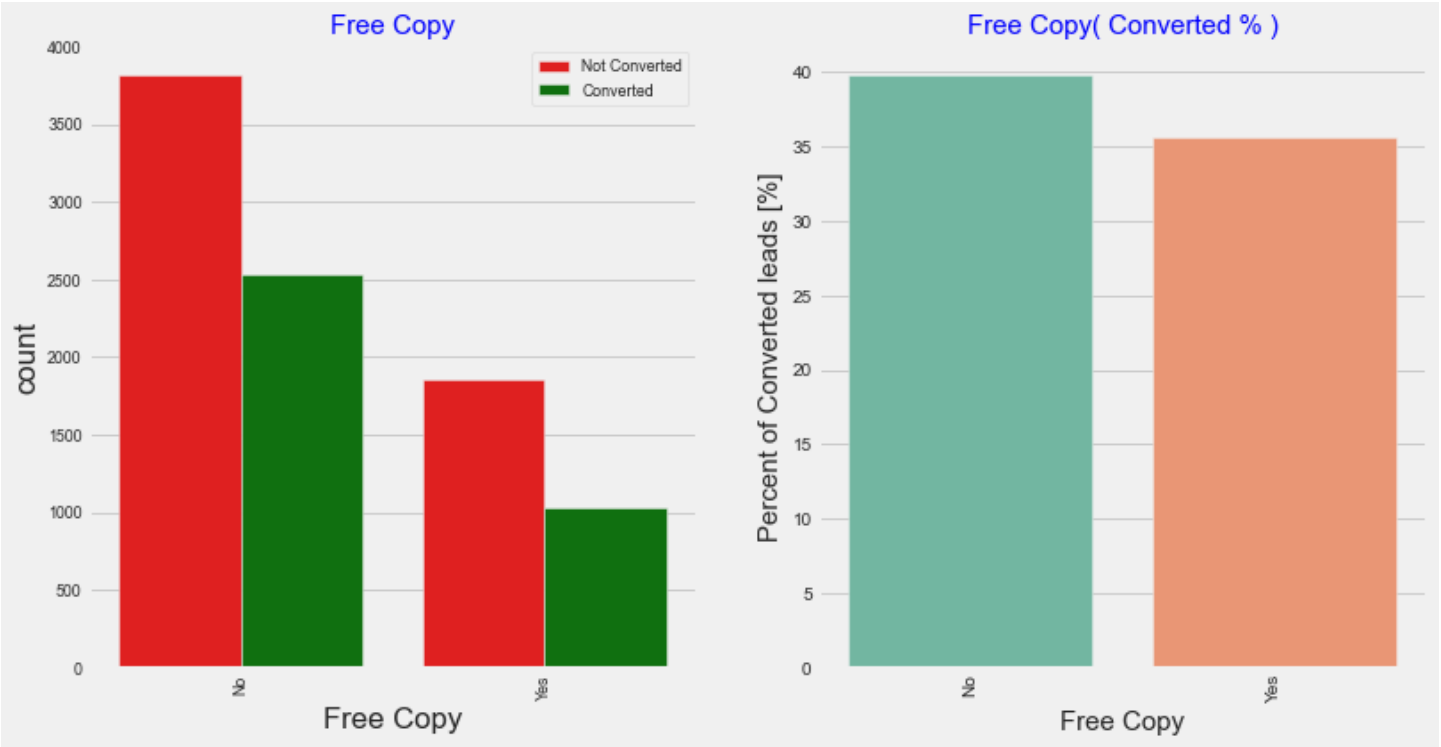
●**Though Housewives are less in numbers, they have 100% conversion rate**

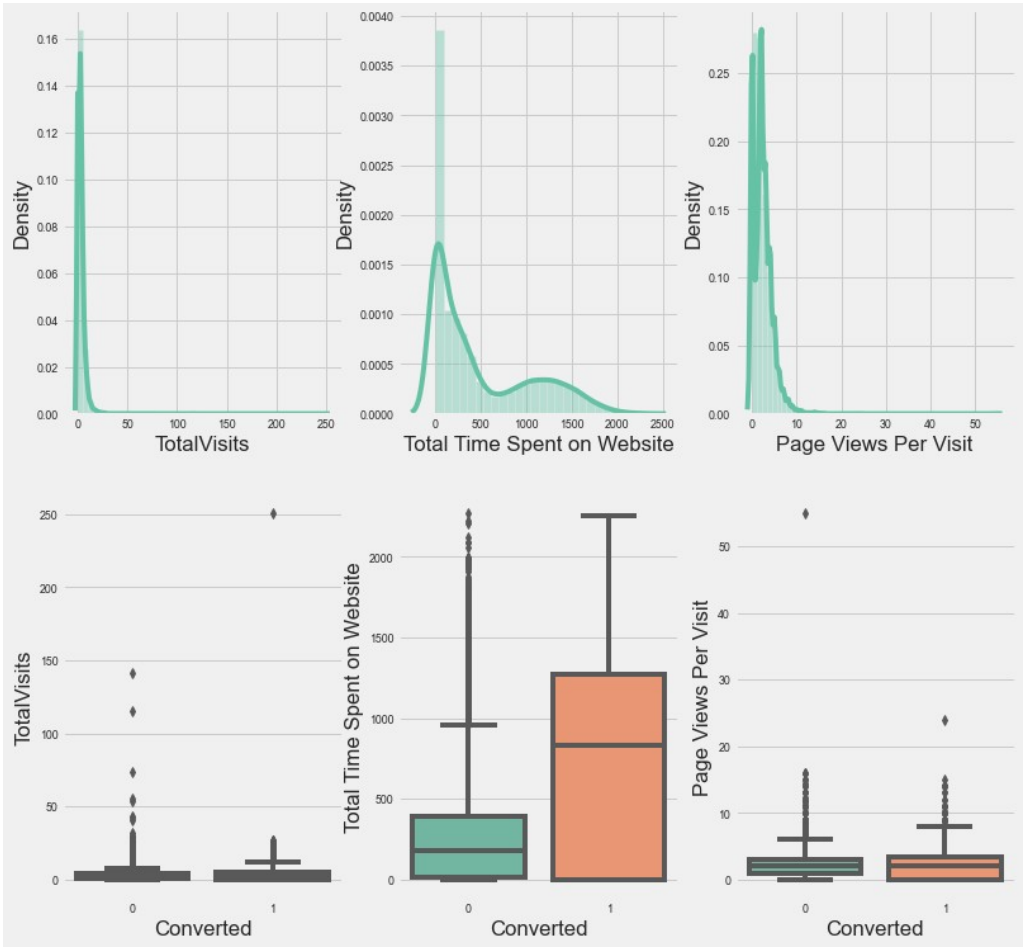
●**Working professionals, Businessmen and Other category have high conversion rate**

- **Though Unemployed people have been contacted in the highest number, the conversion rate is low (~40%)**

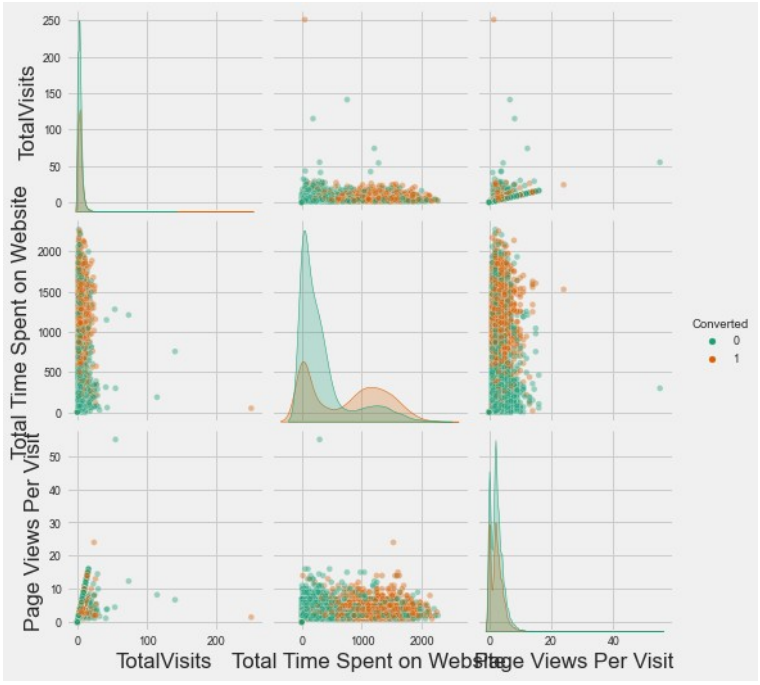
We cannot combine smalled value categories as their conversion rate is very different. Combing them may provide wrong predictions.



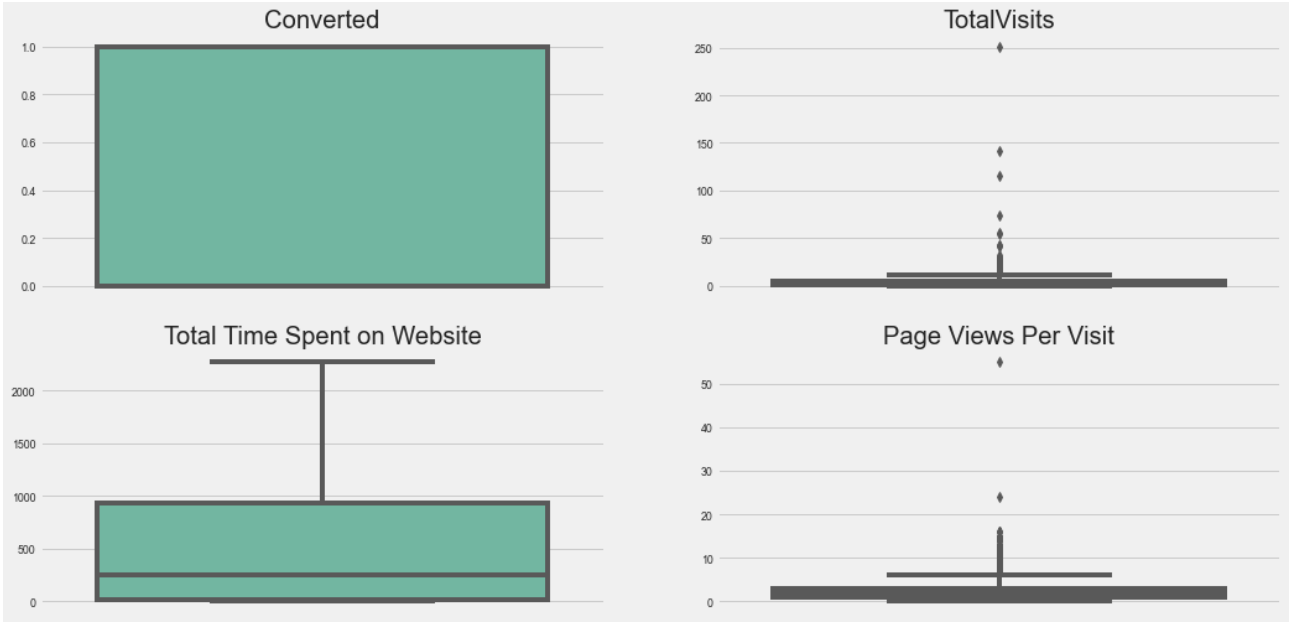




TotalVisits and Page Views per Visit has some outliers which needs to be treated.



Data is not normally distributed.



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**

Now that we have capped the outliers.

3. Dummy Variables:

The dummy variable were created and later on the dummies with “not provided” elements were removed . For numeric values we used the MinMax Scaler.

4. Train- Test split:

The split was done at 70% and 30% for train and test data respectively

5. Model Building:

Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p- value(The variables with VIF<5 and p-value<0.05 were kept).For Manual Feature Reduction, the following methods will be followed in order to reduce the features until we reach reasonable amount of feature count and maintain Sensitivity of the modeal =>80%

1. High P-Value
2. High VIF
3. High negative GLM coeeficient

Low Information Value (IV) generated based on WoE (Weight of Evidence)

6. Model Evalution : Train Dataset

Confusion Matrix :

	Predicted Negative(0)	Predicted Positive(1)
Actual Negative(0)	True Negative (TN)	False Postive (FP)
Actual Positive(1)	False Negative (FN)	True Positive (TP)

Accuracy =

$$\frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Sensitivity =

$$\frac{TP}{(TP + FN)}$$

Specificity =

$$\frac{TN}{(TN + FP)}$$

Precision =

$$\frac{TP}{(TP + FP)}$$

Recall =

$$\frac{TP}{(TP + FN)}$$

F Measure (F_1) = 2 *

$$\frac{\text{Precision} * \text{Recall}}{}$$

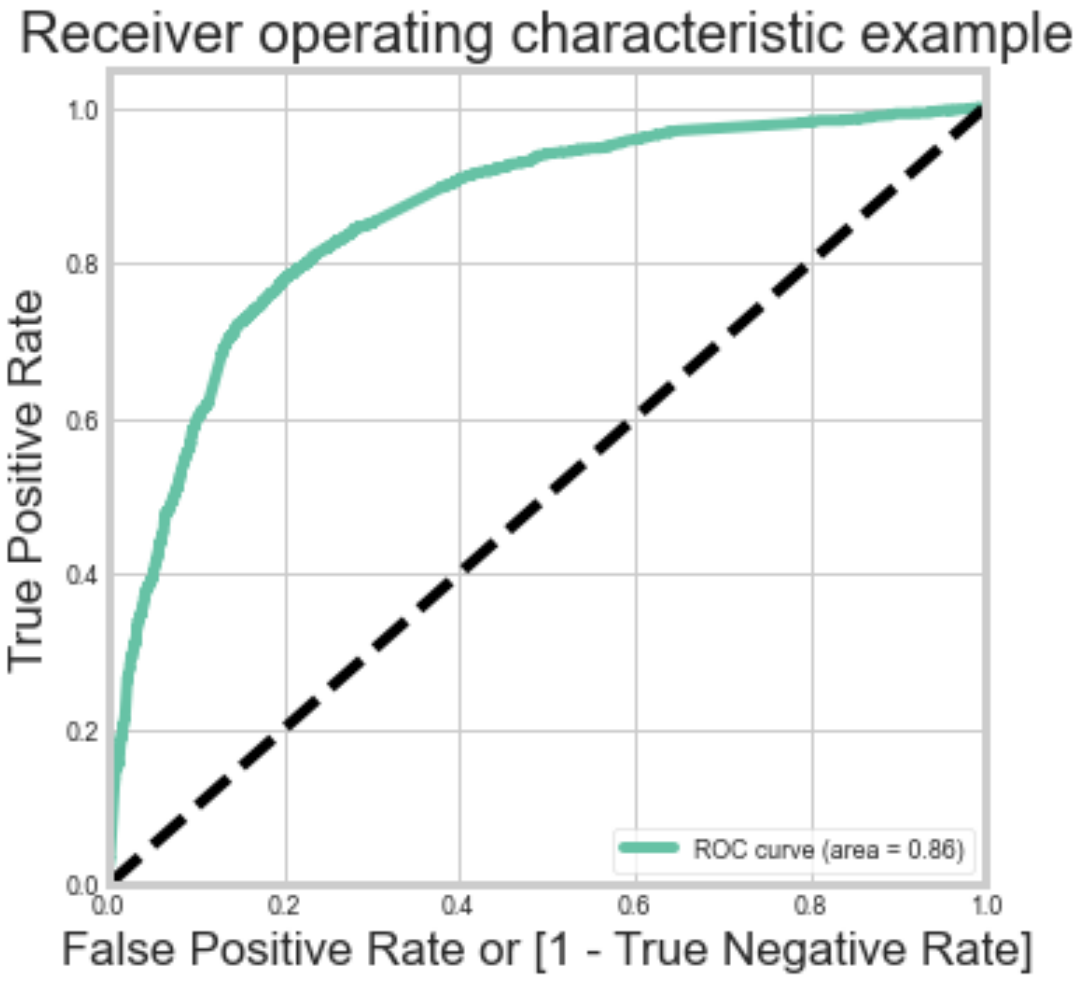
(Precision + Recall)

- **TPR (True Positive Rate)**= TP/(TP + FN)
- **TNR (True Negative Rate)**= TN/(TN + FP)
- **FPR (False Positive Rate)**= FP/(TN + FP)
- **FNR (False Negative Rate)**= FN/(TP + FN)

ROC Curve

An ROC curve demonstrates several things:

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity)
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

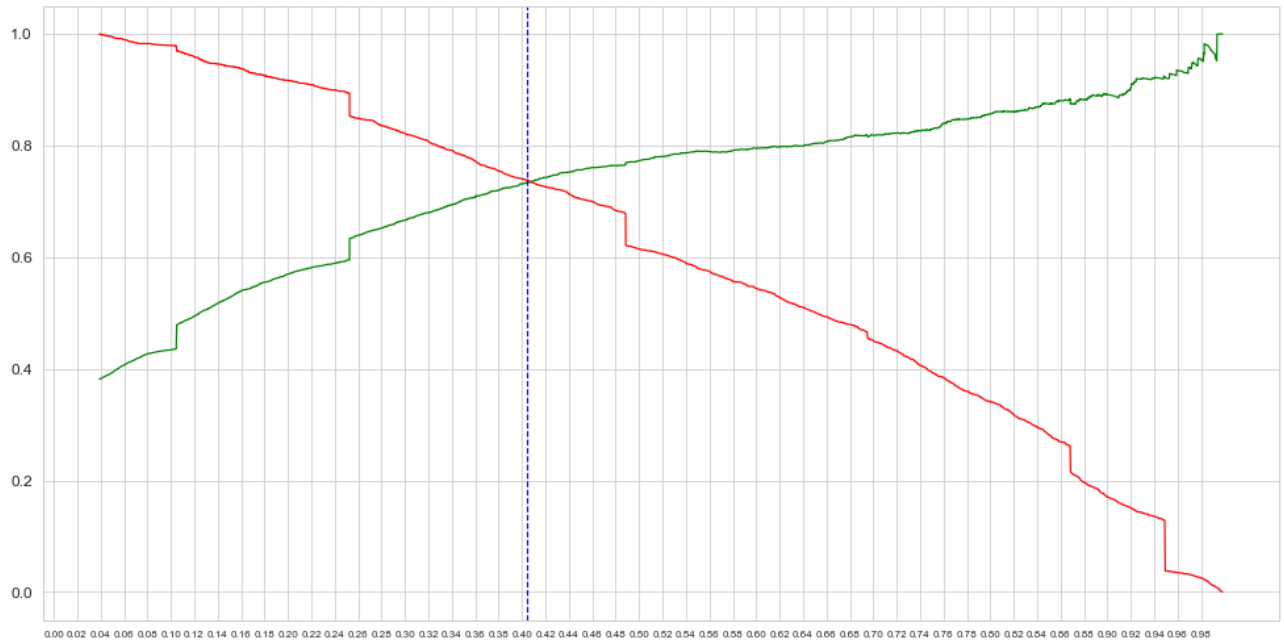


ROC Curve aread is 0.88, which indicates that the model is good.

Precision - Recall Trade off

Next we will look into Precision- Recall trade off to see if balancing these values provides better output.
Precision means out of all leads which are predicted at 1, how many have truly converted.
Recall means out of all leads that have converted, how many of them were correctly identifies as 1. This is the same value as sensitivity.

Precision-Recall trade-off point is used to decide the cut-off point especially when there is huge imbalance in data. In our case the data distribution is 62% vs 38%. So imbalance of data is not a big factor.



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

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

- 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 Precision-Recall trade-off method as it reduced True Positive. We have to increase Sensitivity / Recall value to increase True Positives. Thus we will use 0.335 as cutoff point.

Conclusion:

Interpretation Logistic regression model with multiple predictor variables

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

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

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

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:
Probability = 1 / (1 + exp(-x)) = 1 /(1 + exp(- -2.19)) = 1 /(1 + exp(2.2)) = 0.10 = 10%

<u>Predicting Probabilities:</u>

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.