LEAD SCORING PRESENTATION

Shraavan Sridhar & Shruthi Manjrekar

Approach/Steps

01 Data Cleaning

This includes observing missing values, imputing them with alternate values and removing unwanted columns.

02 EDA

This includes analysing the categorical and numerical variables and comparing them with target variable 'Converted'

03 Model Building

This includes building a logistic regression model, evaluating the train and test data separately, measuring accuracy, plotting the ROC curve etc.

01 Data Cleaning

→ **Removing columns with no unique values.**

There were certain columns with no unique values, these columns were removed as nothing can be inferred from these columns.

→ **Replacing 'select' column.**

We noticed many entires with 'select' option in our data frame, this is simply because the users have not selected the attributes through the drop down menu. We replace this by Null values (np.nan)

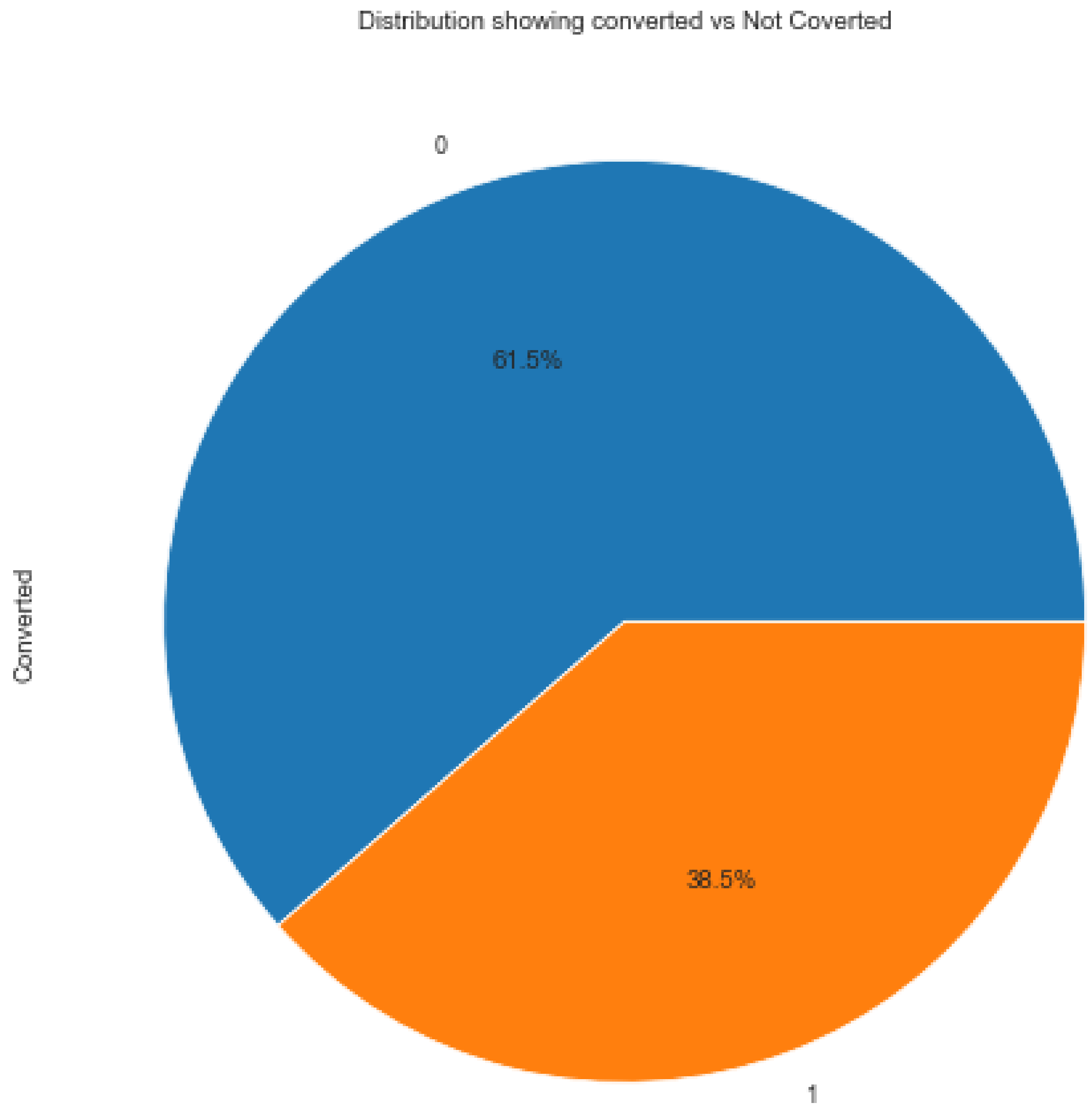
→ **Drop columns with null values.**

Next, we drop columns with missing value percentage>40%. We replace the rest of the missing categorical variables with 'Uknown' and numerical columns with median.

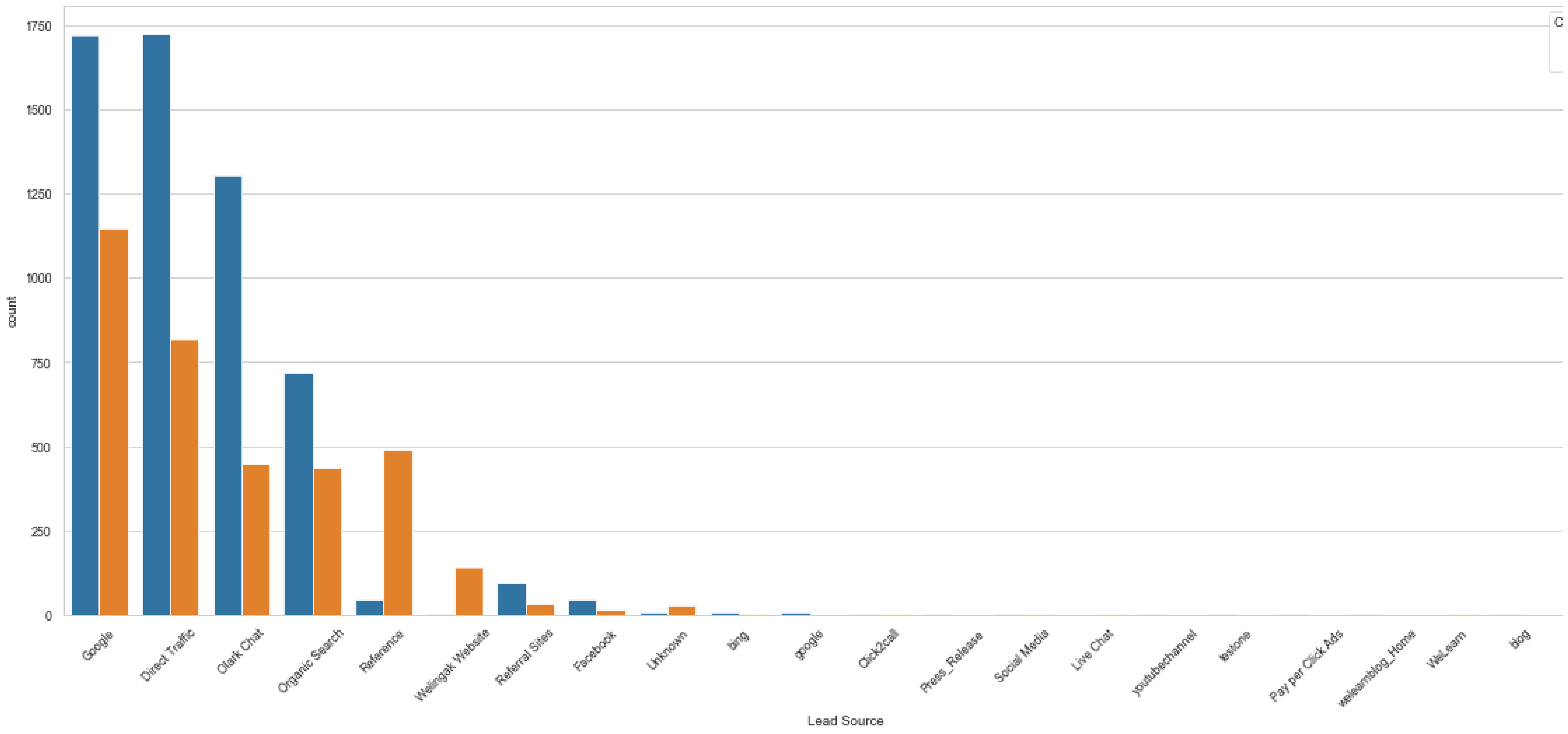
02 EDA

→ Analysing 'COVERTED' target variable.

- Performed data imbalance check on target variable- “Converted”. It had lower converted (38.5%) records as compared to those which were not converted(61.5%).
- X Education company has too many non-conversions which need to be improved.



02 EDA



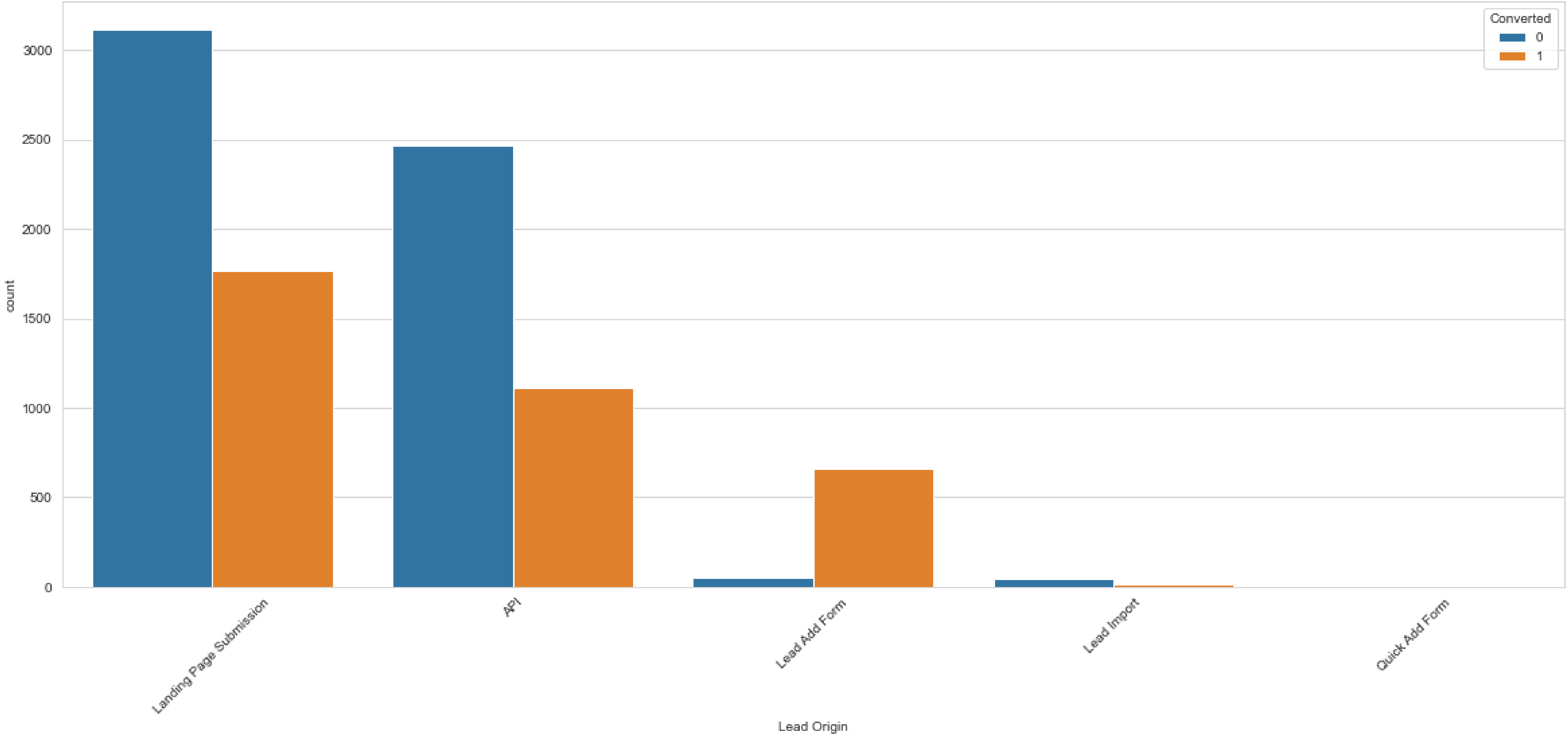
Lead Source

- Most traffic comes from Google and Direct Traffic.

- For leads through references, the conversion is more than non-conversion.

- X Education company can target conversions through References and Welligak Website as they see higher conversions.

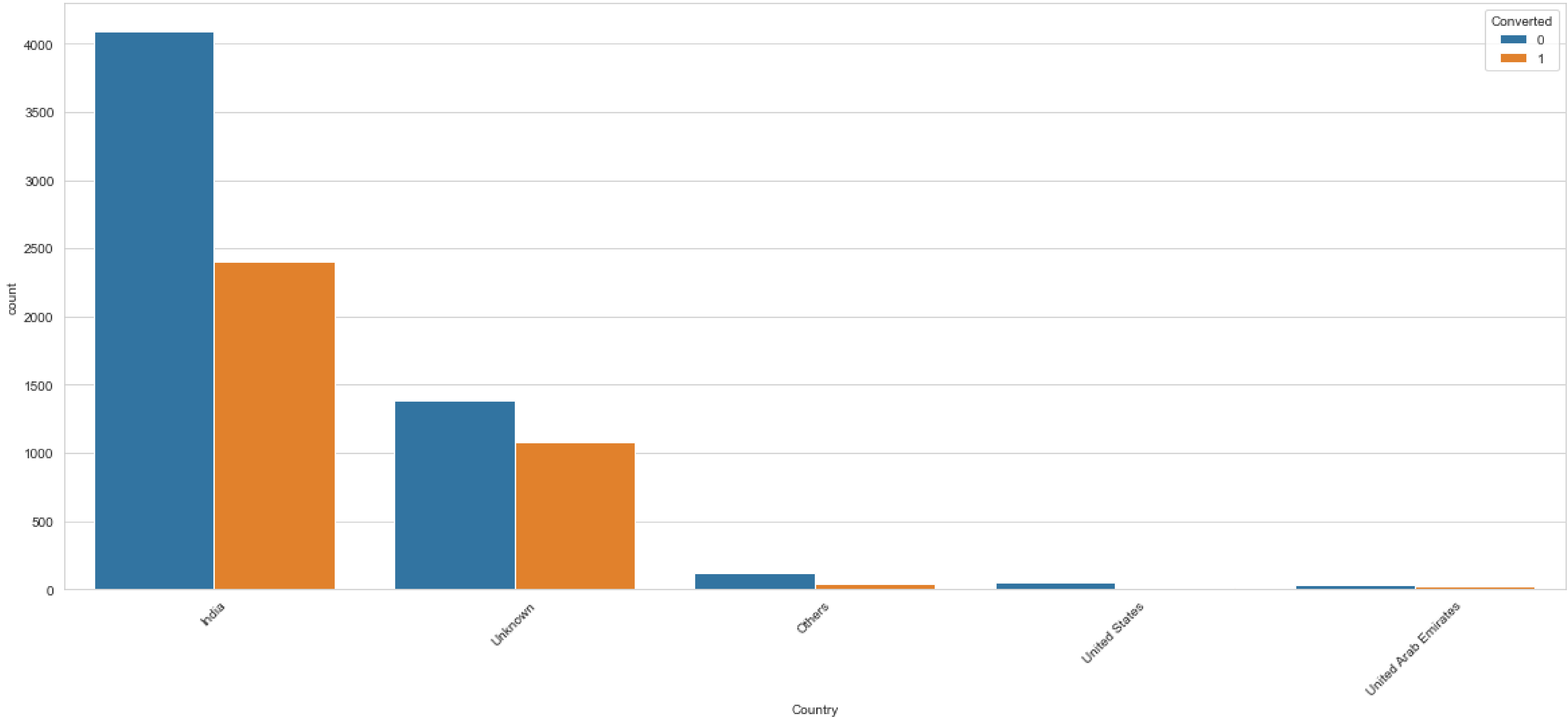
02 EDA



→ **Lead Origin**

- APIs and landing page leads yield more leads.
- Lead Add Form seems to have higher rate of conversion. One can pursue leads from this source actively.

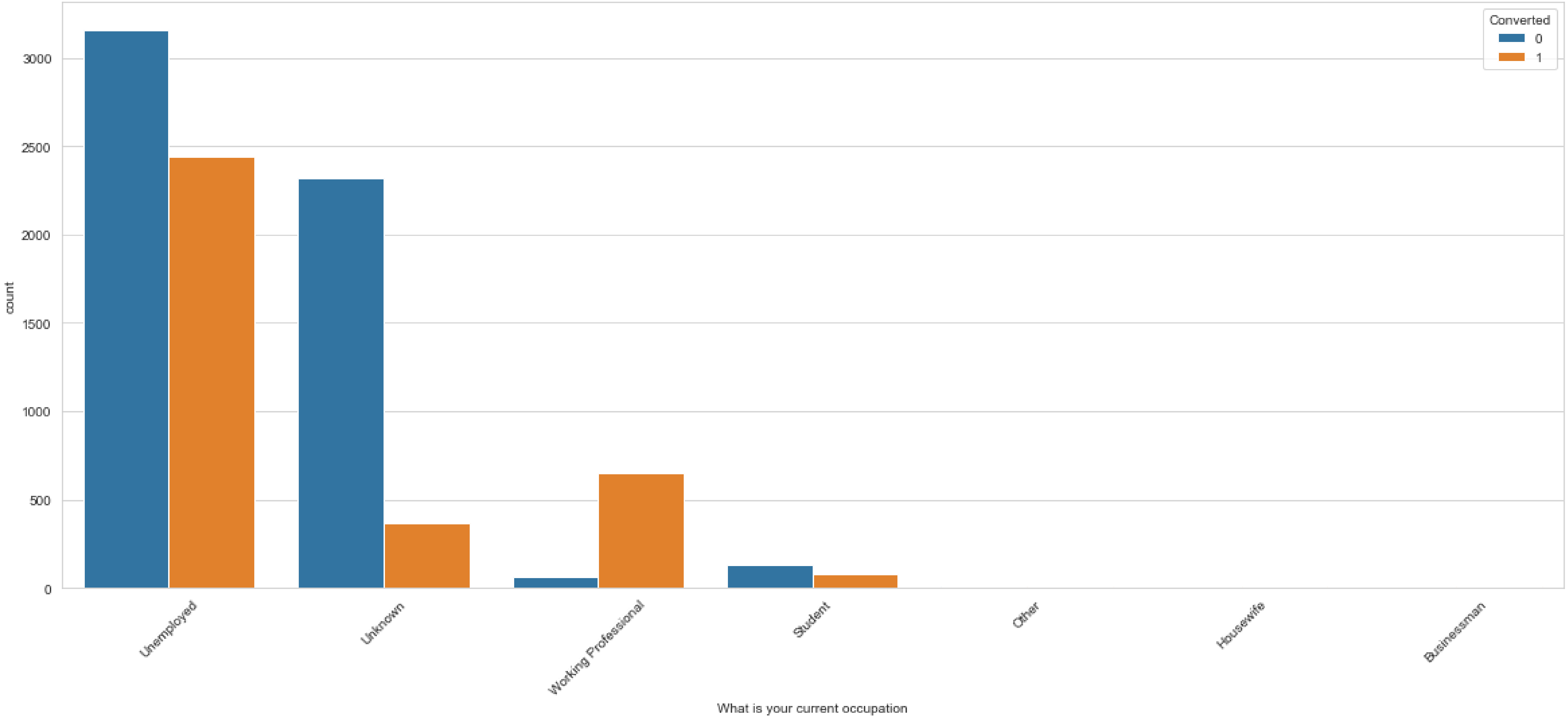
02 EDA



Country

- Maximum leads are from India.
- The leads from United States hardly see any conversion.

02 EDA

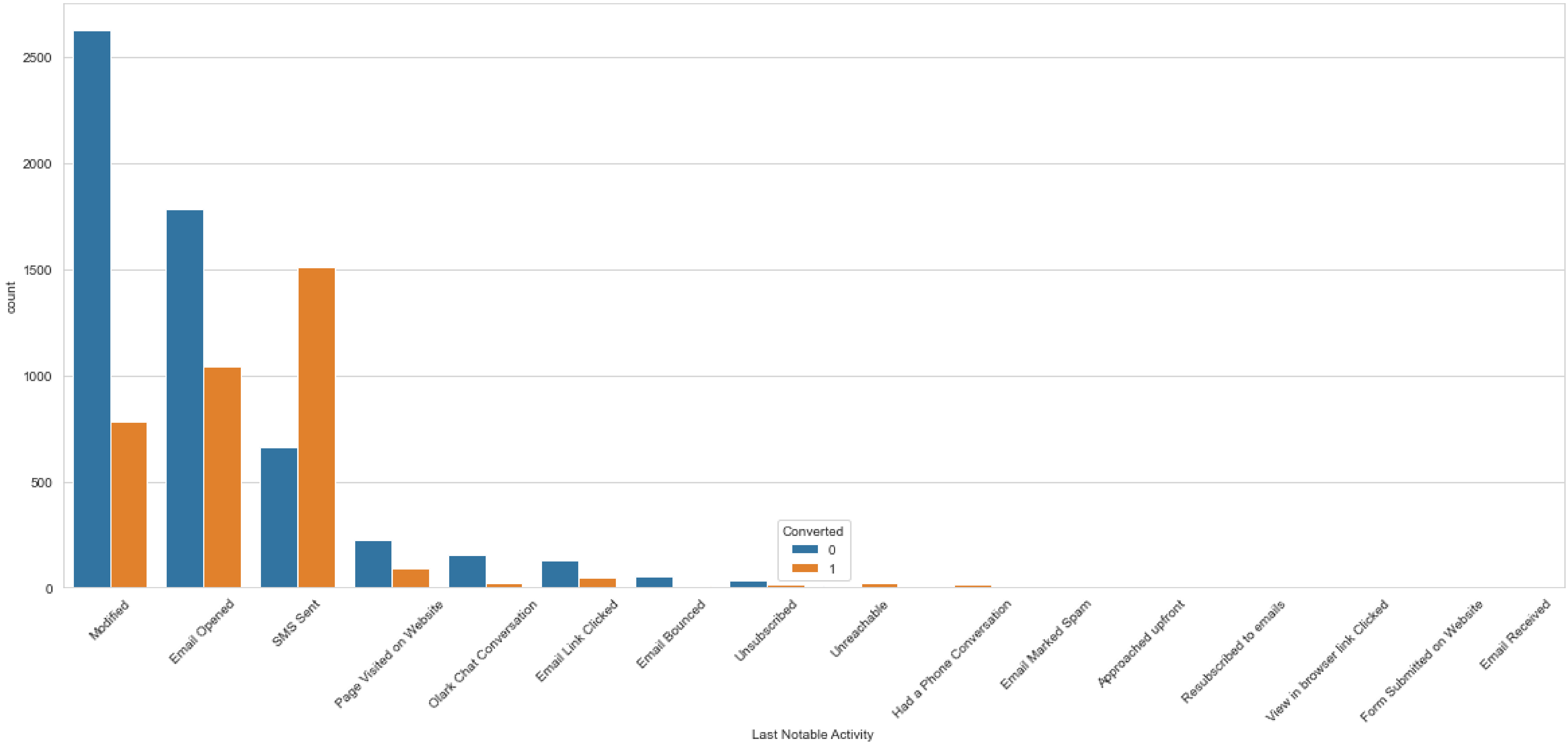


Occupation

- It is interesting to note that max no of leads have come from 'Unemployed'. Most leads are looking to get employed through this course.

- The conversion among working professionals is more than non conversions. The company can target working professionals more.

02 EDA



→ **Last Notable Activity**

- Leads to whom SMS was sent get converted easily. The conversions through this are more than non conversions.

02 EDA

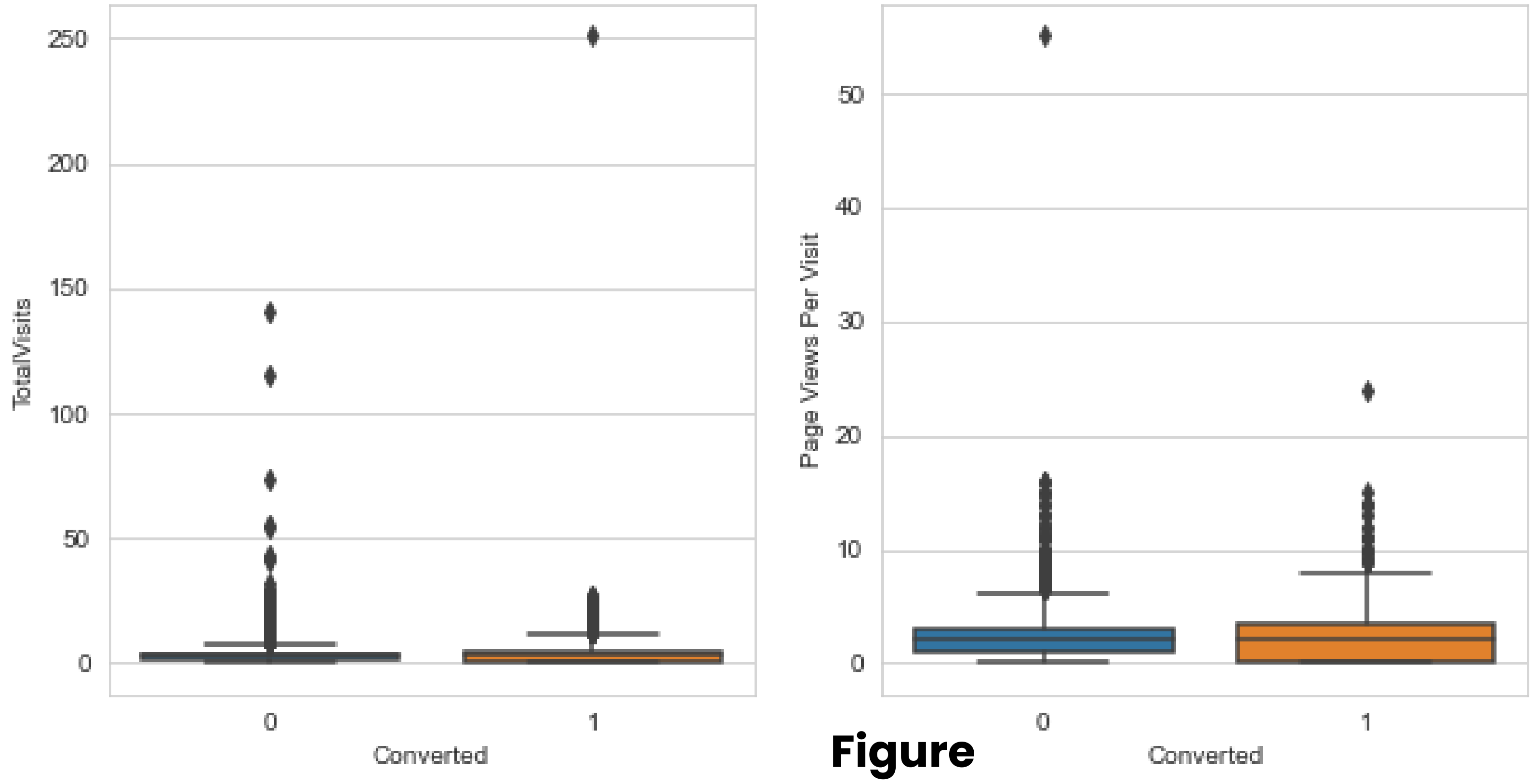


Figure 1

Total Visits vs 'Converted'

- As seen in figure1, there are outliers which need to be treated.

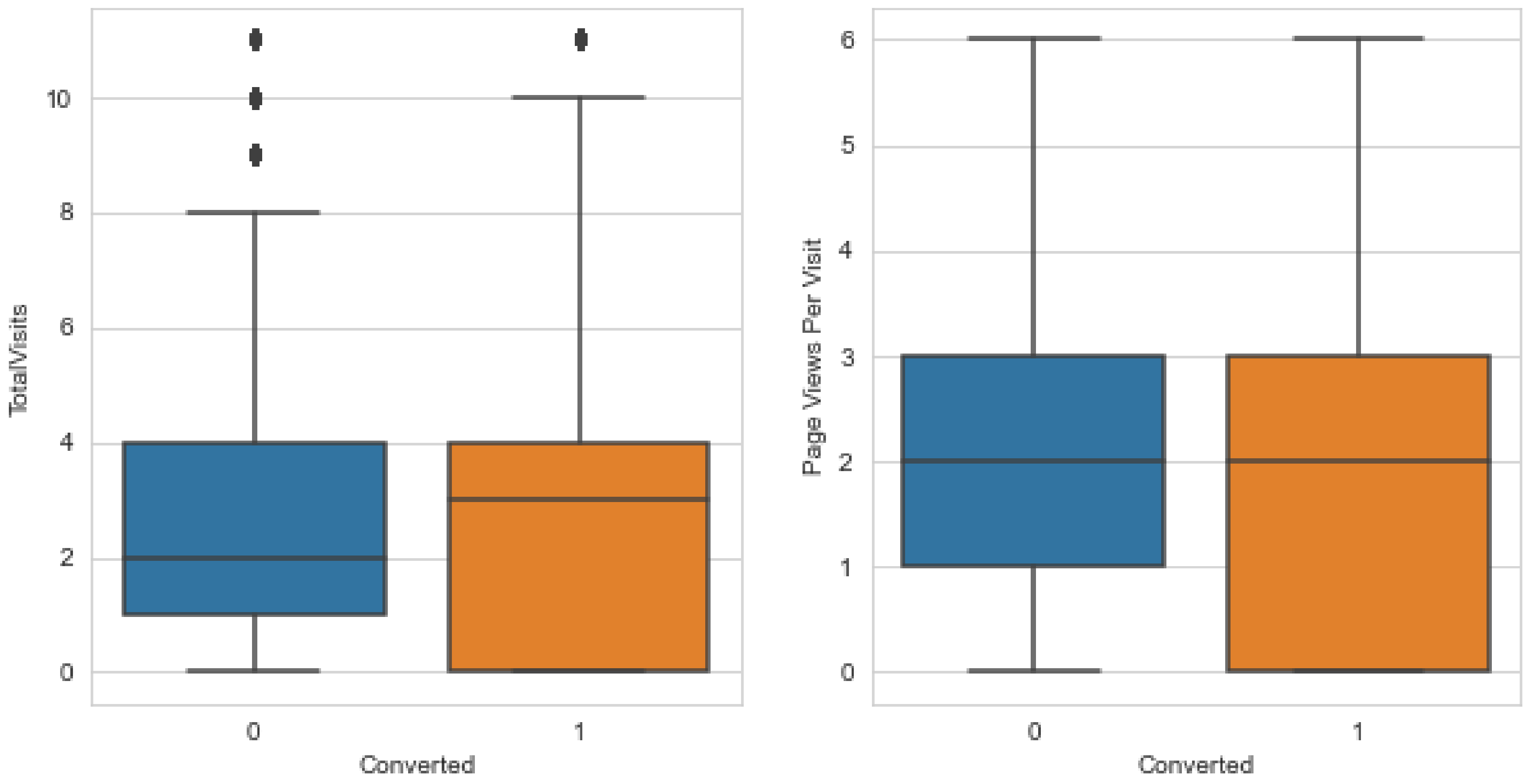
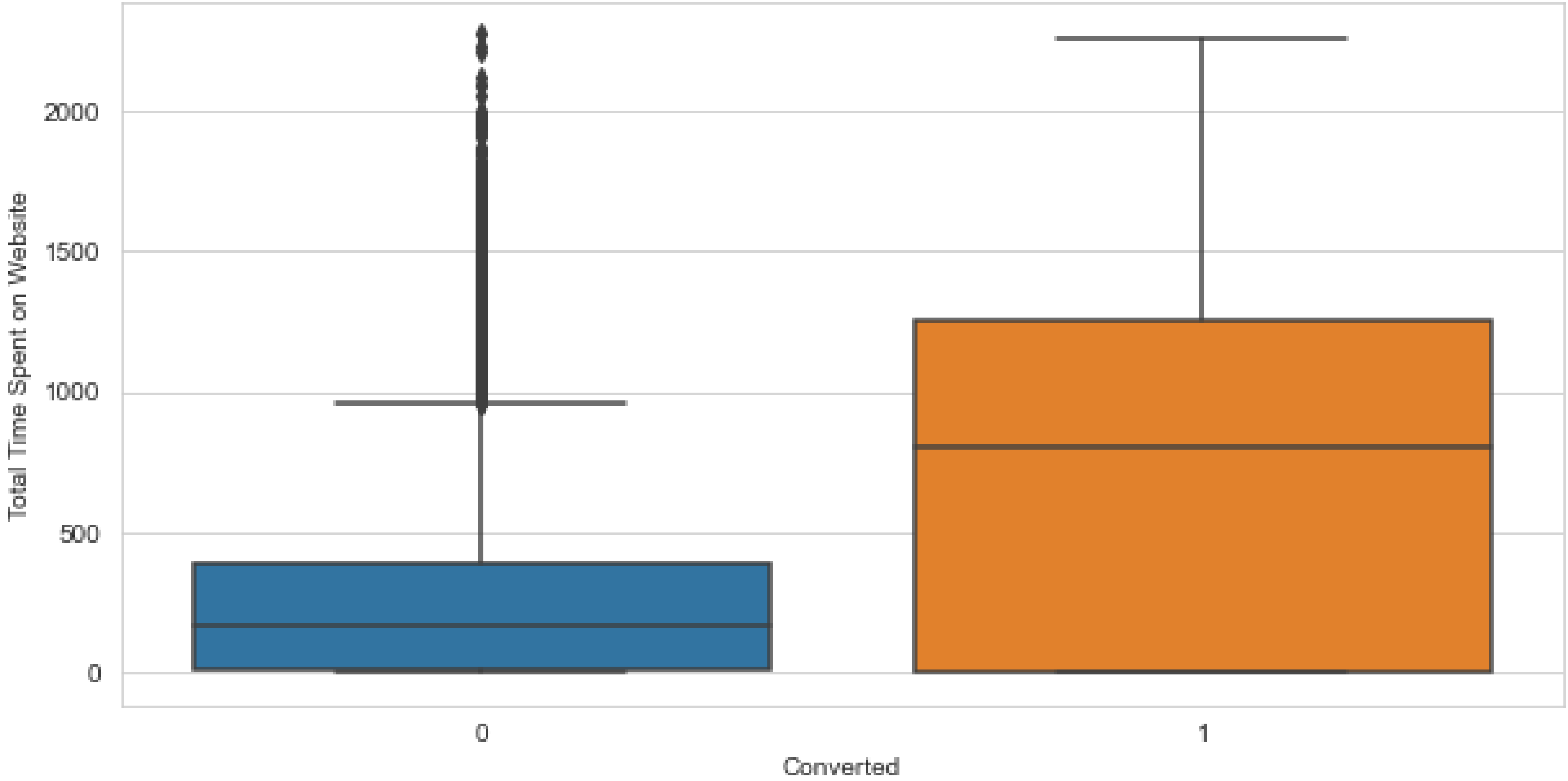


Figure 2

Total Visits vs 'Converted'

- As seen in figure 2, the outliers have been treated
- Nothing significant can be inferred from the figure though.

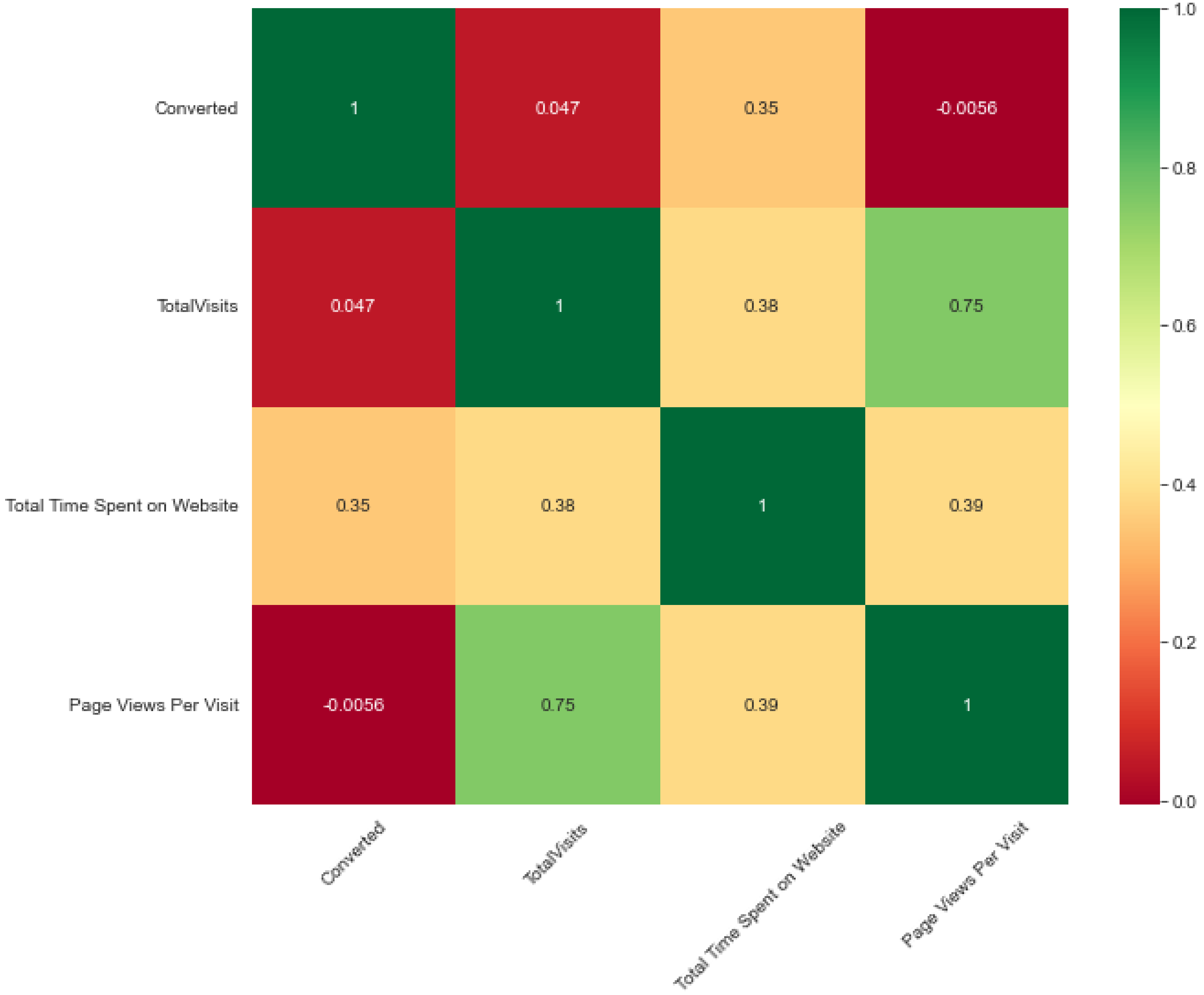
02 EDA



Time spent on Website vs 'Converted'

- Clearly, a person's time spend on a website is somewhat indicative of him/her being more interested to take up the course.
- Target leads which spend more time on the website.

02 EDA



→ Correlation heat map

- Highest Positive correlation between TotalVisits and Page Views per visit at 0.75
- Total time spent on website has some positive corelation with target variable converted at 0.35

03 Model Building

→ **Creating Dummy Variables.**

In this step, we create dummy variables for all categorical variables and drop the duplicate columns

→ **Separate Target Variable**

In this step we separate the target variable and store it in y and the rest of the dataframe becomes X.

→ **Train-Test Split**

We split the data into train and test, by having 70% of data as train and 30% as test with random sampling at 42.

→ **Feature Scaling**

Scaling the numerical variables with the help of MinMax Scaler(). We use scaler.fit_transform for train and only scaler_transform for test.

03 Model Building

Feature Selection using RFE.

We use a logistic regression model and selected 15 columns supported by RFE.

Assessing model with StatsModel.

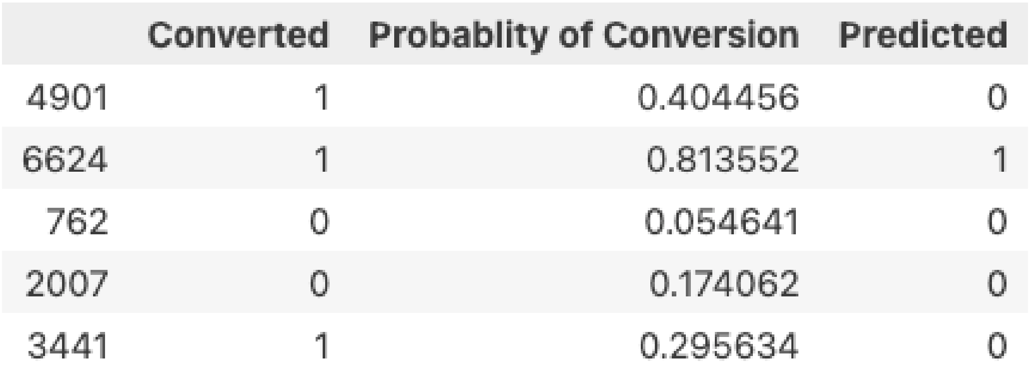
We remove 'Last Notable Activity_Had a Phone Conversation' and 'Lead Source_Welingak Website' due to them having p-values greater than 0.05.

Final Model

Next, we arrive at a model with permissible pvalues and VIF.

Prediction

We create a data-frame with probability of conversions by selecting 0.45 as cut off probability.

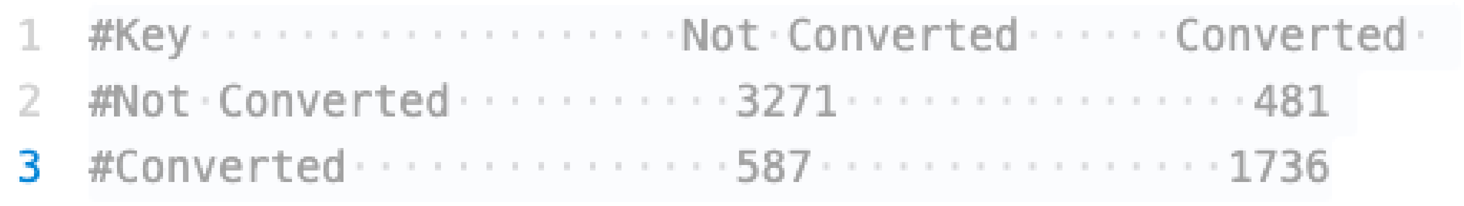


	Converted	Probablity of Conversion	Predicted
4901	1	0.404456	0
6624	1	0.813552	1
762	0	0.054641	0
2007	0	0.174062	0
3441	1	0.295634	0

03 Model Building

Training Model Evaluation.

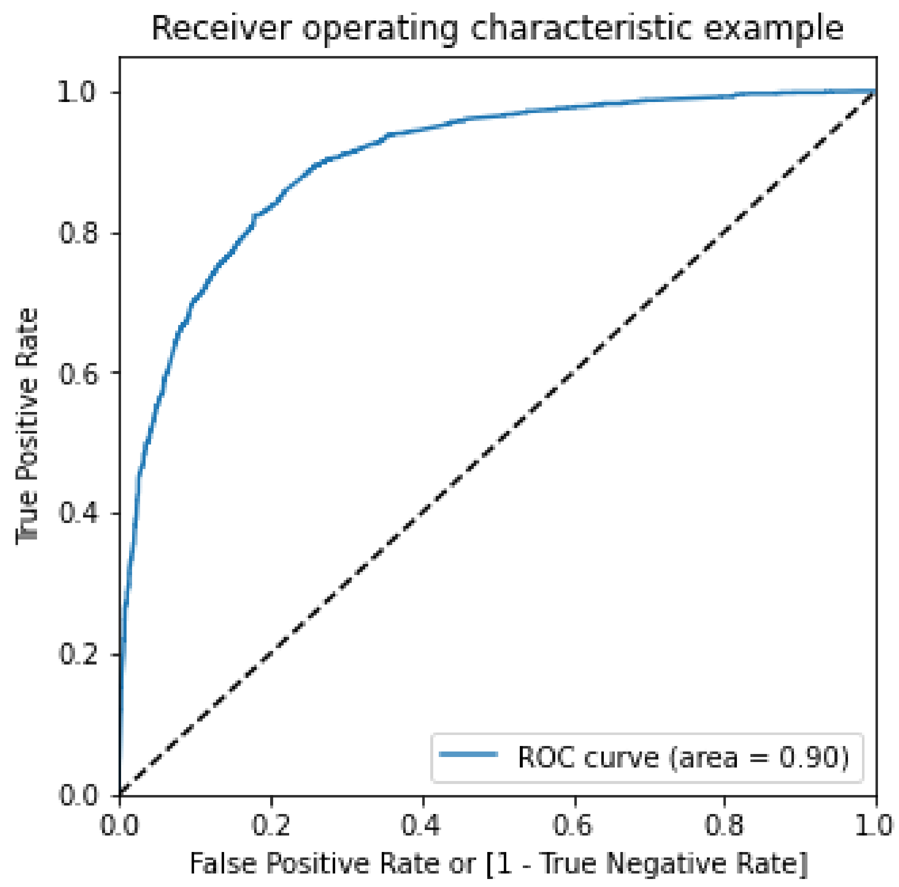
- **Confusion Matrix - As shown the figure.**
- **Accuracy score - 82.4%.**
- **Sensitivity- 74.7 %**
- **Specificity 87.18%**



```
#Key                    Not Converted     Converted
#Not Converted              3271               481
#Converted                   587              1736
```

ROC Curve

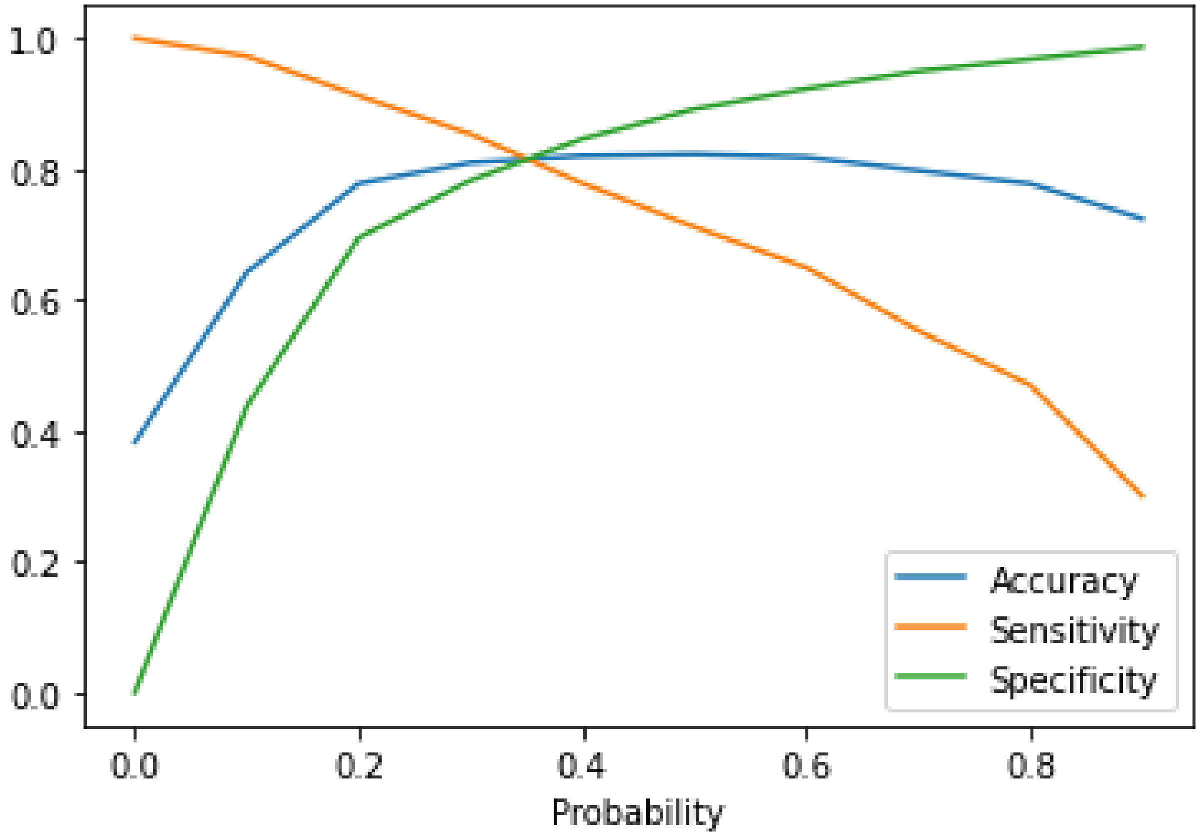
ROC is at 0.90 for training set.



03 Model Building

→ Training Model Evaluation with optimal cutoff.

- From the plot we can conclude that the Optimal Cutoff is at 0.34.
- Next, the optimal cut-off is applied to our final prediction with new predictive values.



→ New Model Metrics at cutoff 0.34

- **Accuracy-81.5%**
- **Sensitivity-82.9%**
- **Specificity-80.65%**
- **Precision- 72.64%**
- **Recall- 82.9%**

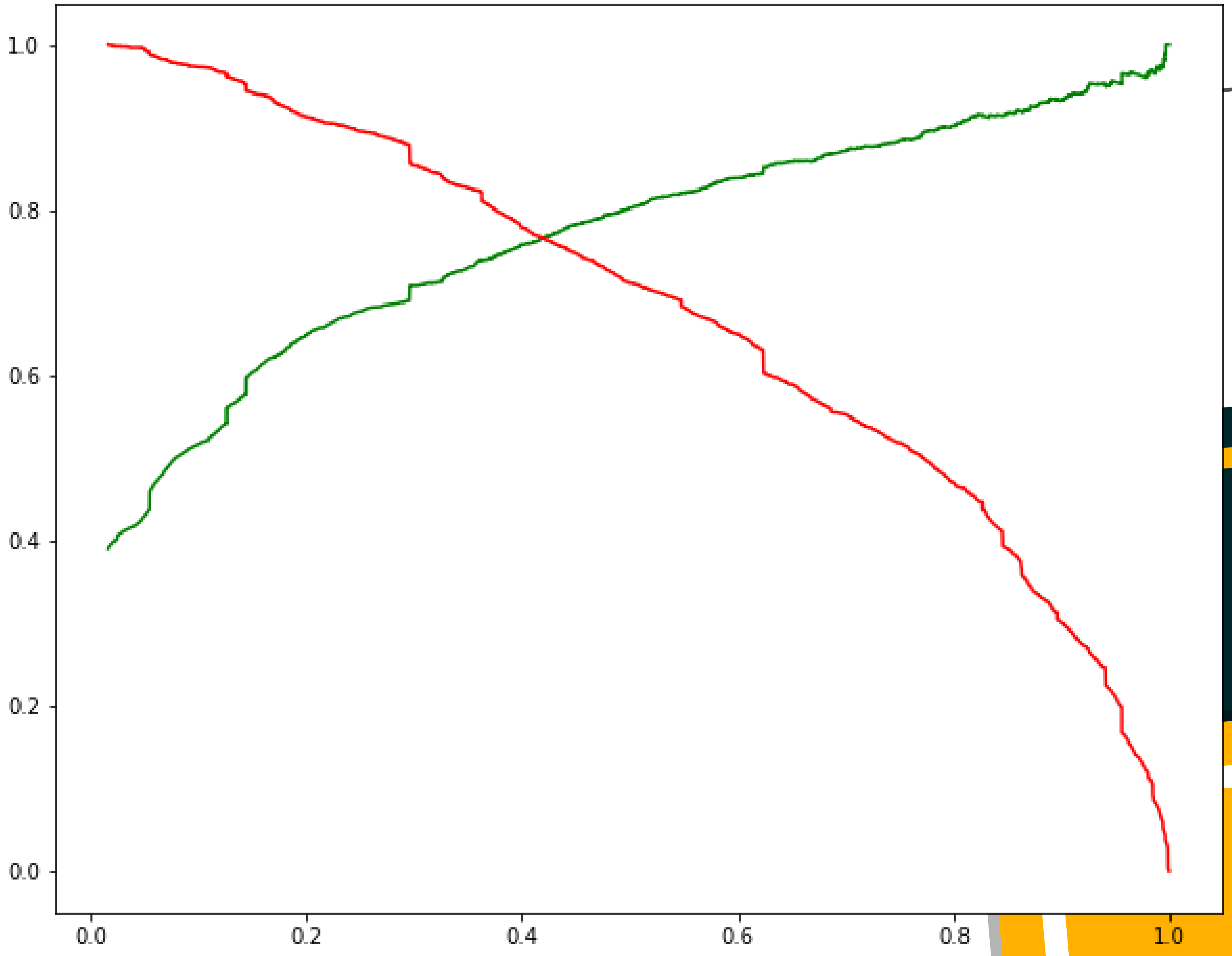
03 Model Building

Precision Recall Tradeoff

- From the precision call trade-off we arrive at a cut off of 0.41. We adjust the model prediction the cutoff of 0.41 and re-evaluate all the model metrics.

New Model Metrics at cutoff 0.41

- **Accuracy-81.9%**
- **Sensitivity-77.09%**
- **Specificity-85.02%**



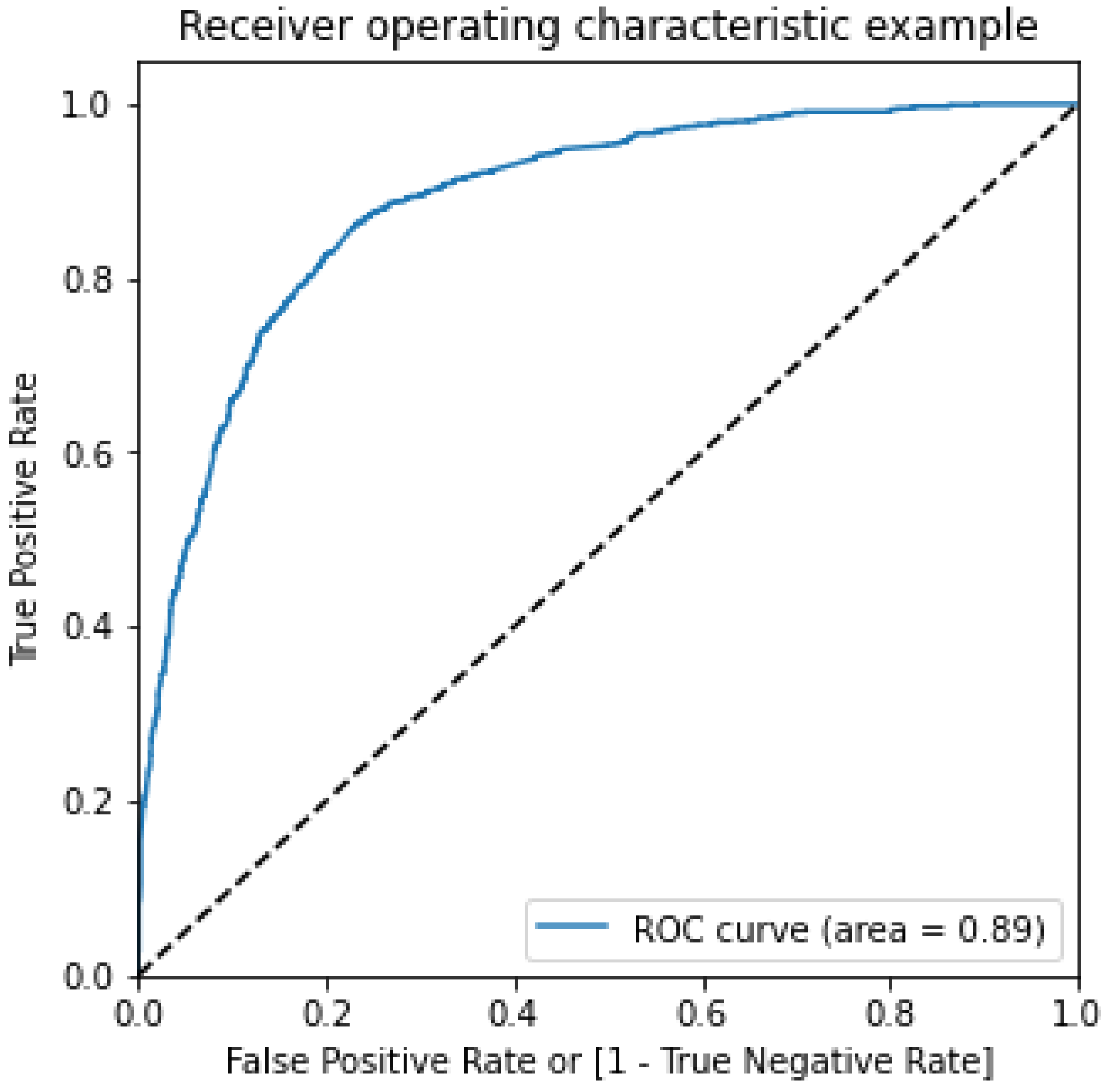
03 Model Building

Test Prediction

- Here we use the test data set and directly use the final columns that inferred from the train model.
- We take the same precision-recall tradeoff cutoff of 0.41 and evaluate the test metrics.

Test Model Metrics at cutoff 0.41

- **Accuracy-81.6%**
- **Sensitivity-75.02%**
- **Specificity-85.83%**
- **Precision-77.08%**
- **Recall-75.02%**



Test Prediction

- The area under ROC curve is 0.89 which is close to roc area for the training data set.

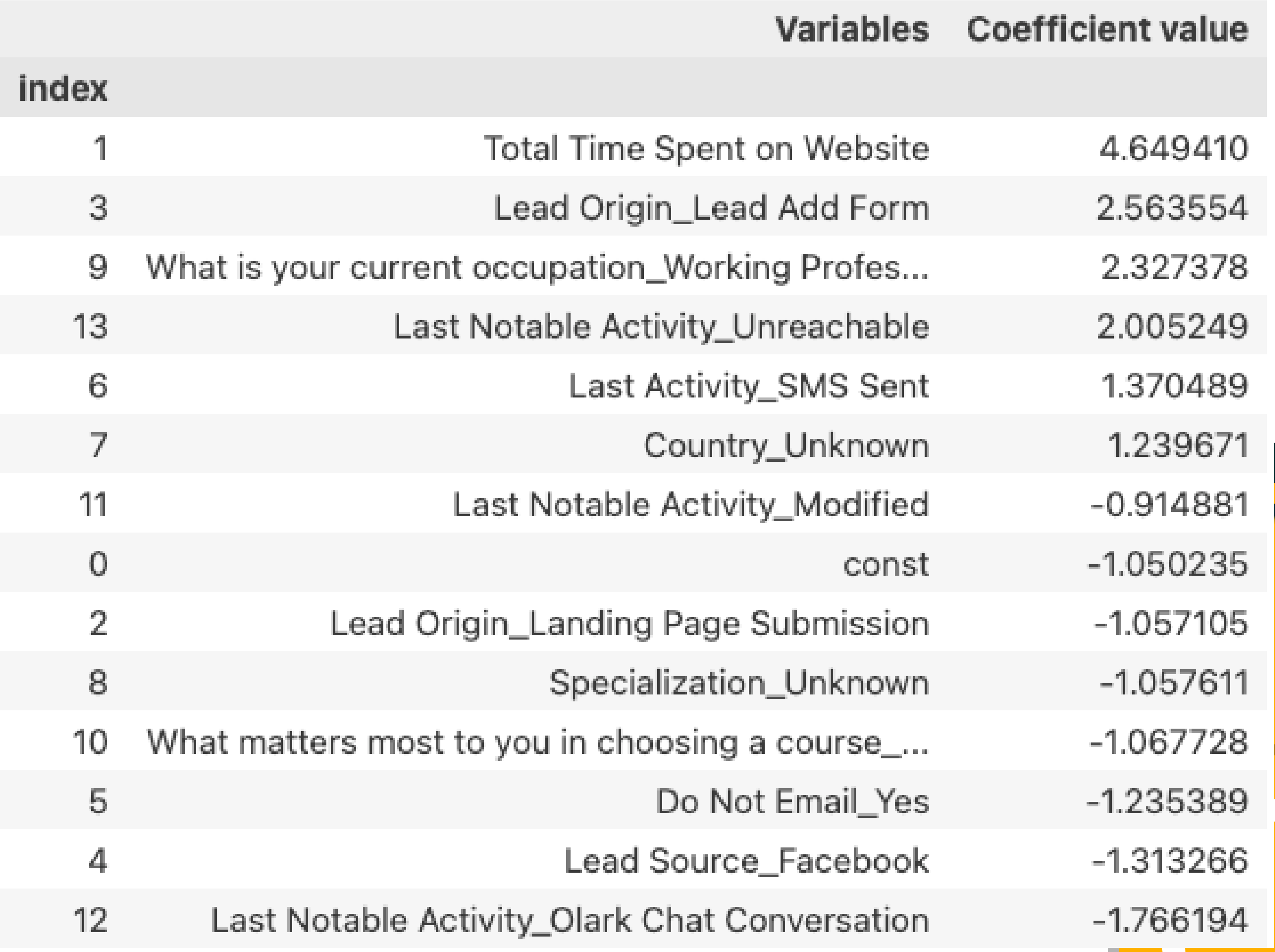
03 Model Building

Assigning Lead Score.

- Finally, a lead score is assigned to each lead by multiplying the probability of conversion with 100.
- The higher the lead score, the more the porbability of conversion.

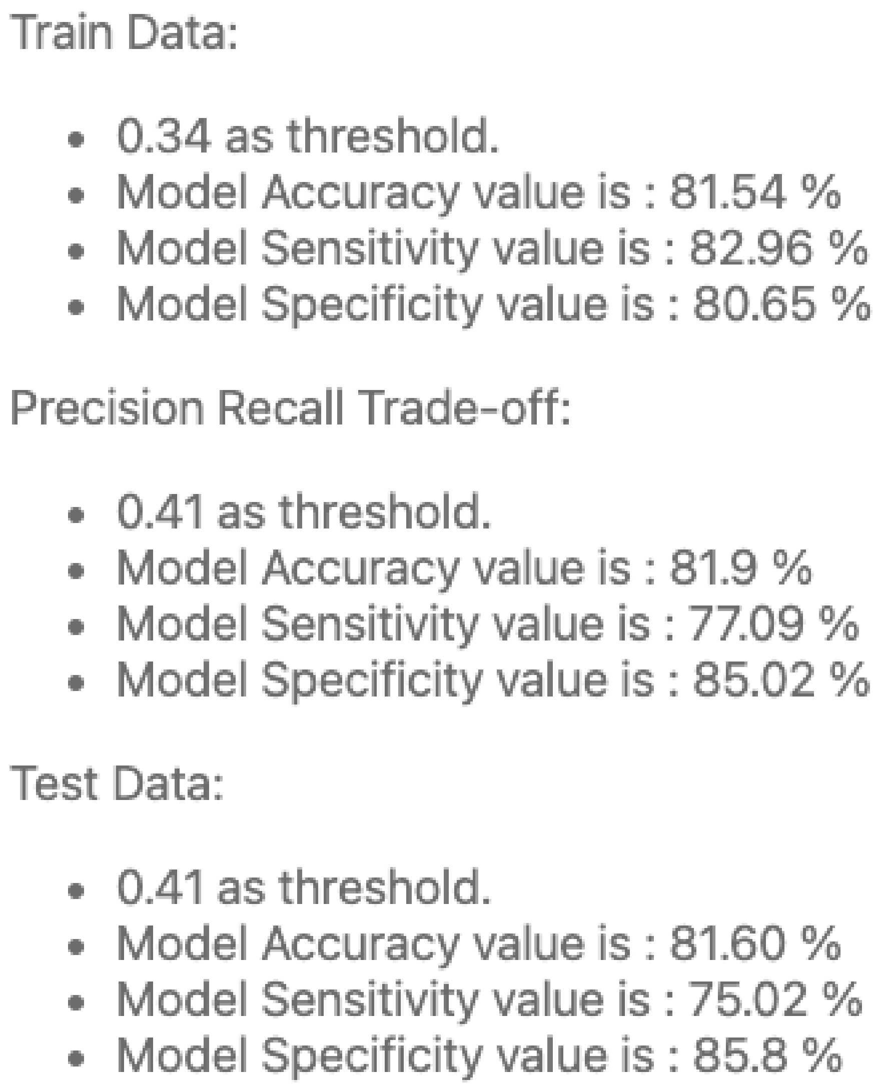
List of Variables

- **The figure on the right has the final model data with 13 variables.**



	Variables	Coefficient value
index		
1	Total Time Spent on Website	4.649410
3	Lead Origin_Lead Add Form	2.563554
9	What is your current occupation_Working Profes...	2.327378
13	Last Notable Activity_Unreachable	2.005249
6	Last Activity_SMS Sent	1.370489
7	Country_Unknown	1.239671
11	Last Notable Activity_Modified	-0.914881
0	const	-1.050235
2	Lead Origin_Landing Page Submission	-1.057105
8	Specialization_Unknown	-1.057611
10	What matters most to you in choosing a course_...	-1.067728
5	Do Not Email_Yes	-1.235389
4	Lead Source_Facebook	-1.313266
12	Last Notable Activity_Olark Chat Conversation	-1.766194

04 Final Metrics



Train Data:

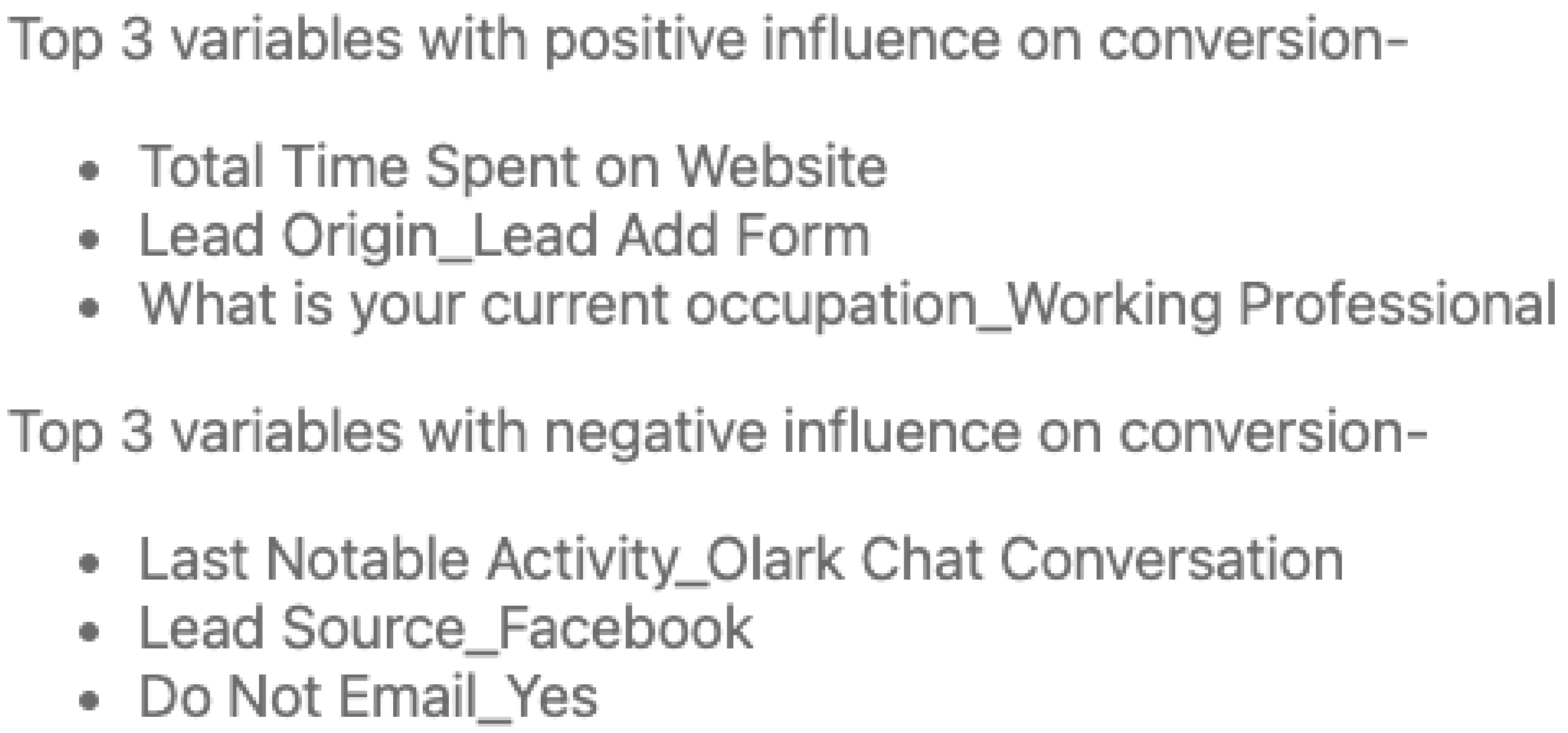
- 0.34 as threshold.
- Model Accuracy value is : 81.54 %
- Model Sensitivity value is : 82.96 %
- Model Specificity value is : 80.65 %

Precision Recall Trade-off:

- 0.41 as threshold.
- Model Accuracy value is : 81.9 %
- Model Sensitivity value is : 77.09 %
- Model Specificity value is : 85.02 %

Test Data:

- 0.41 as threshold.
- Model Accuracy value is : 81.60 %
- Model Sensitivity value is : 75.02 %
- Model Specificity value is : 85.8 %



Top 3 variables with positive influence on conversion-

- Total Time Spent on Website
- Lead Origin_Lead Add Form
- What is your current occupation_Working Professional

Top 3 variables with negative influence on conversion-

- Last Notable Activity_Olark Chat Conversation
- Lead Source_Facebook
- Do Not Email_Yes

05 Suggestions

- The total time spent on the website heavily influences lead conversion. More time the lead spends on website, more likely is the conversion.

- Leads generated through Add_forms will see greater conversions.

- Working professionals tend to take up the course more often than not.

- Olark chat conversations, Facebook leads, the ones who asked not be emailed have a negative influence on conversion and one needs to exercise caution while approaching these leads.

- If the lead engages activity through SMS, he/she is more likely to be converted.

- For leads through references, the conversion is more than non-conversion.

- X Education company can target conversions through References and Welligak Website leads as they see higher conversions.
- Target leads with higher lead score as they have higher probablity to get converted.