LEAD SCORING CASE STUDY

Submitted by : Ashwin Chaudhari
Yash Dudure

Lead Score Case Study for X education

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%.

Business Goal :

X Education want to develop a model to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score 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 CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Goal of the Case Study is :

1.Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

2.There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Problem solving methodology

Data Sourcing , Cleaning and Preparation

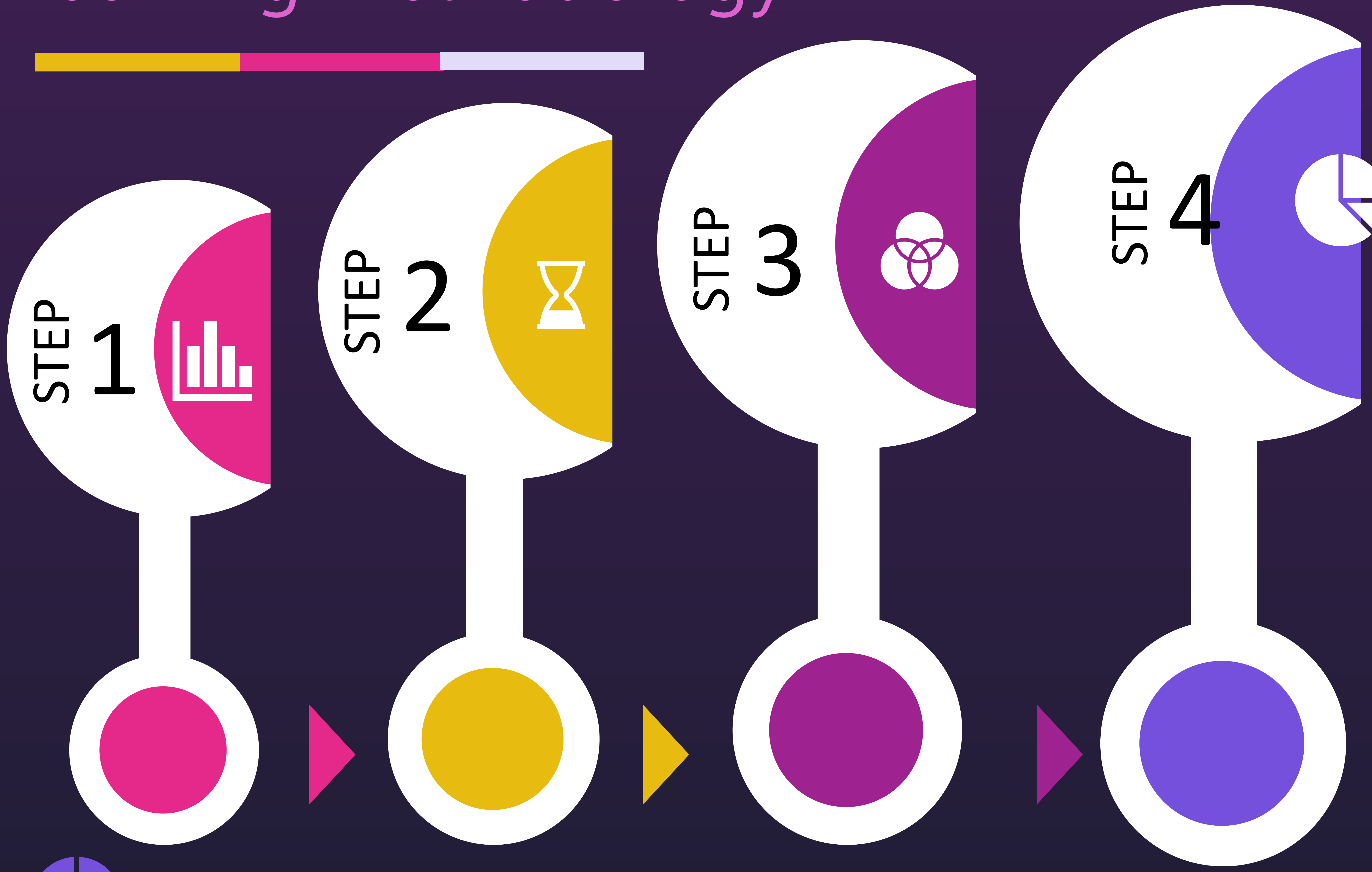
- Read the Data from Source
- Convert data into clean format suitable for analysis
- Remove duplicate data
- Outlier Treatment
- Exploratory Data Analysis
- Feature Standardization

Feature Scaling and Splitting Train and Test Sets

- Feature Scaling of Numeric data
- Splitting data into train and test set.

Model Building

- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.



Result

- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

Data Cleaning and Preparation

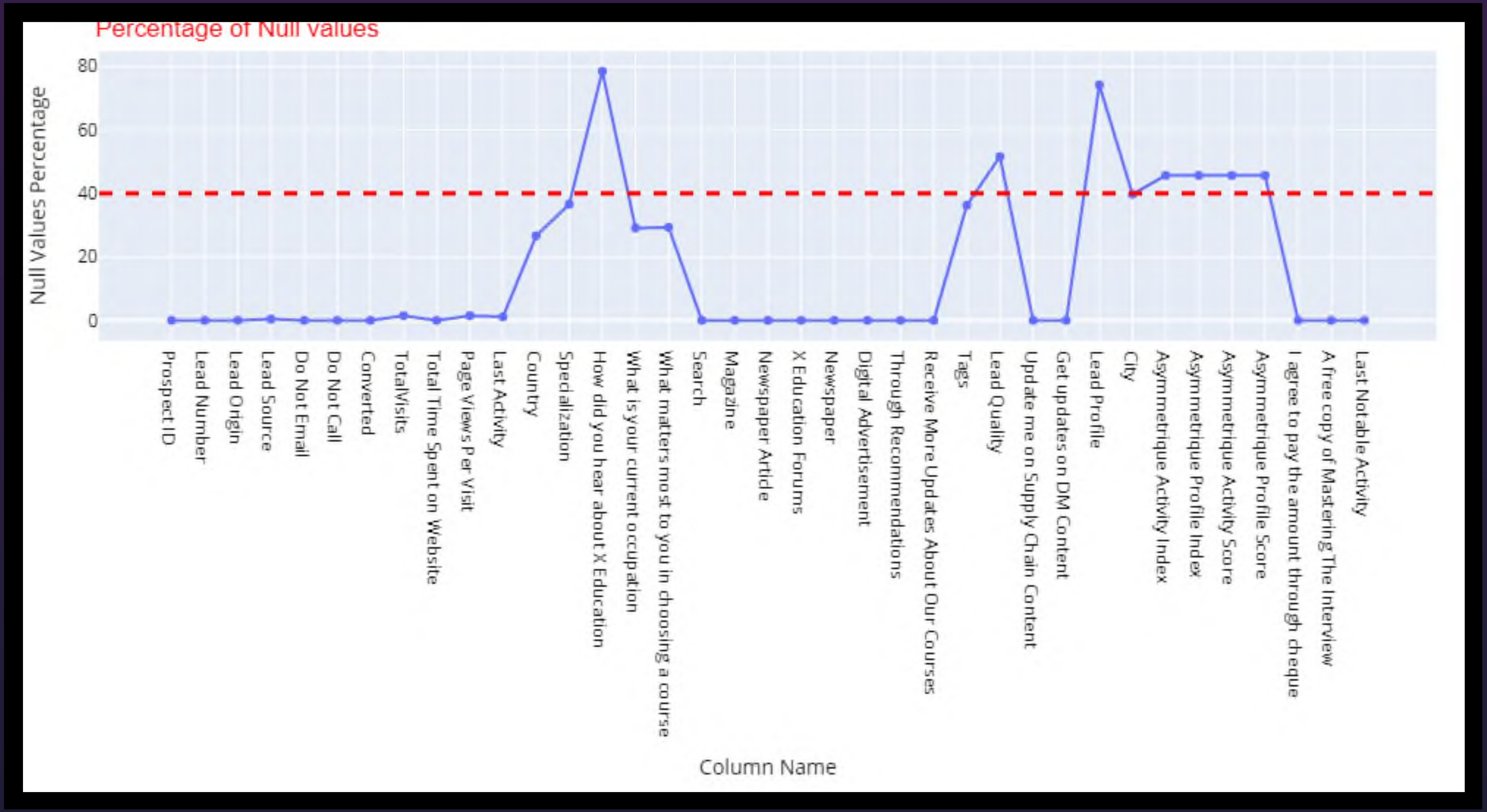
Strategy for Data Cleaning :

a. Check percentage of null values in columns and drop the columns which have more than 45% missing values.

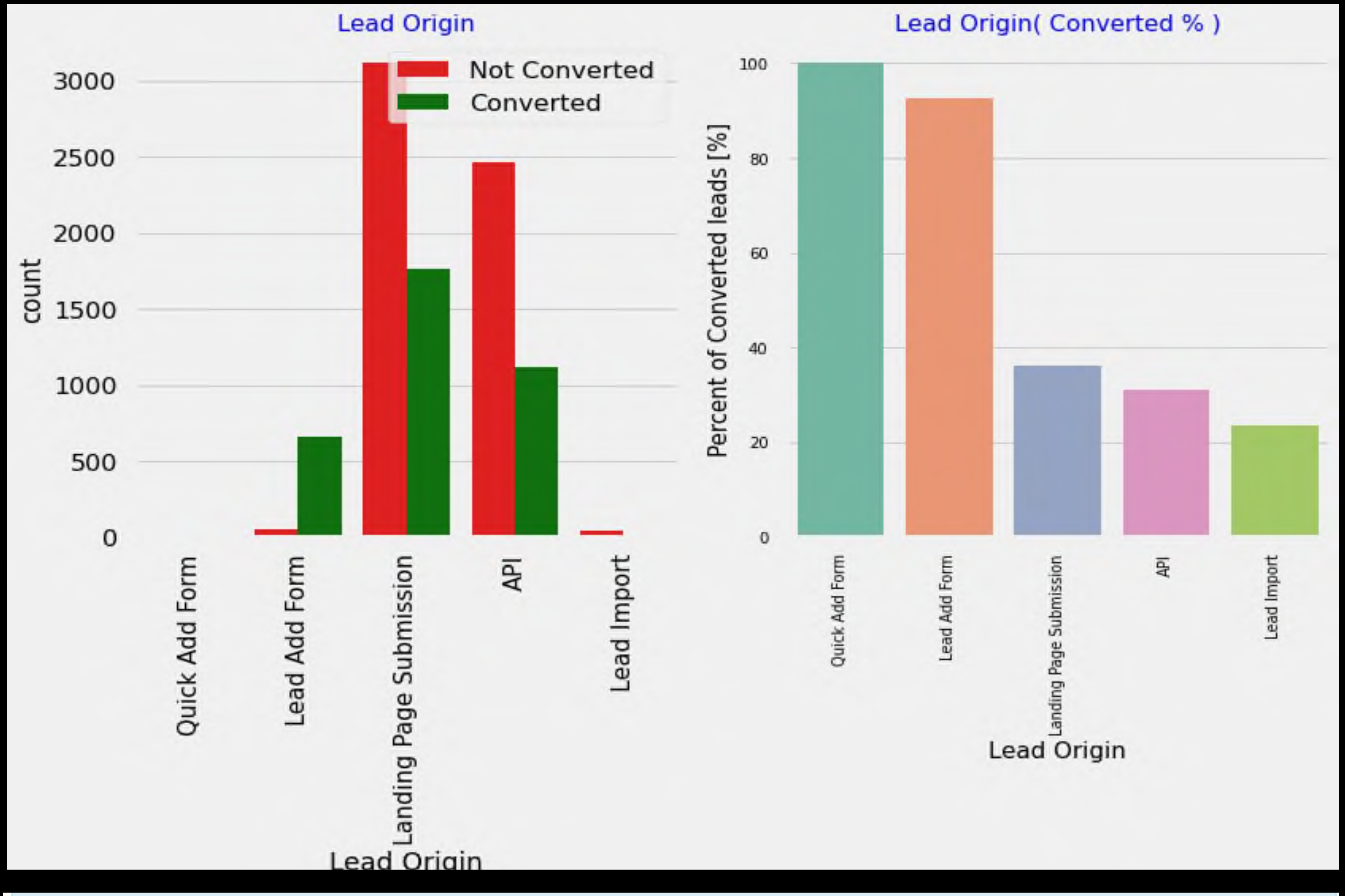
b. Also, some of the variables are created by the sales team once they contact the potential lead. These variables will not be available for the model building as these features would not be available before the lead is being contacted. We will drop these columns too.

c. Some of the columns have only 1 category. These columns will not add any value to the model and can be deleted.

d. Some of the columns have one of the value as "Select" These should be considered as null values. Data Value needs to be updated for these columns

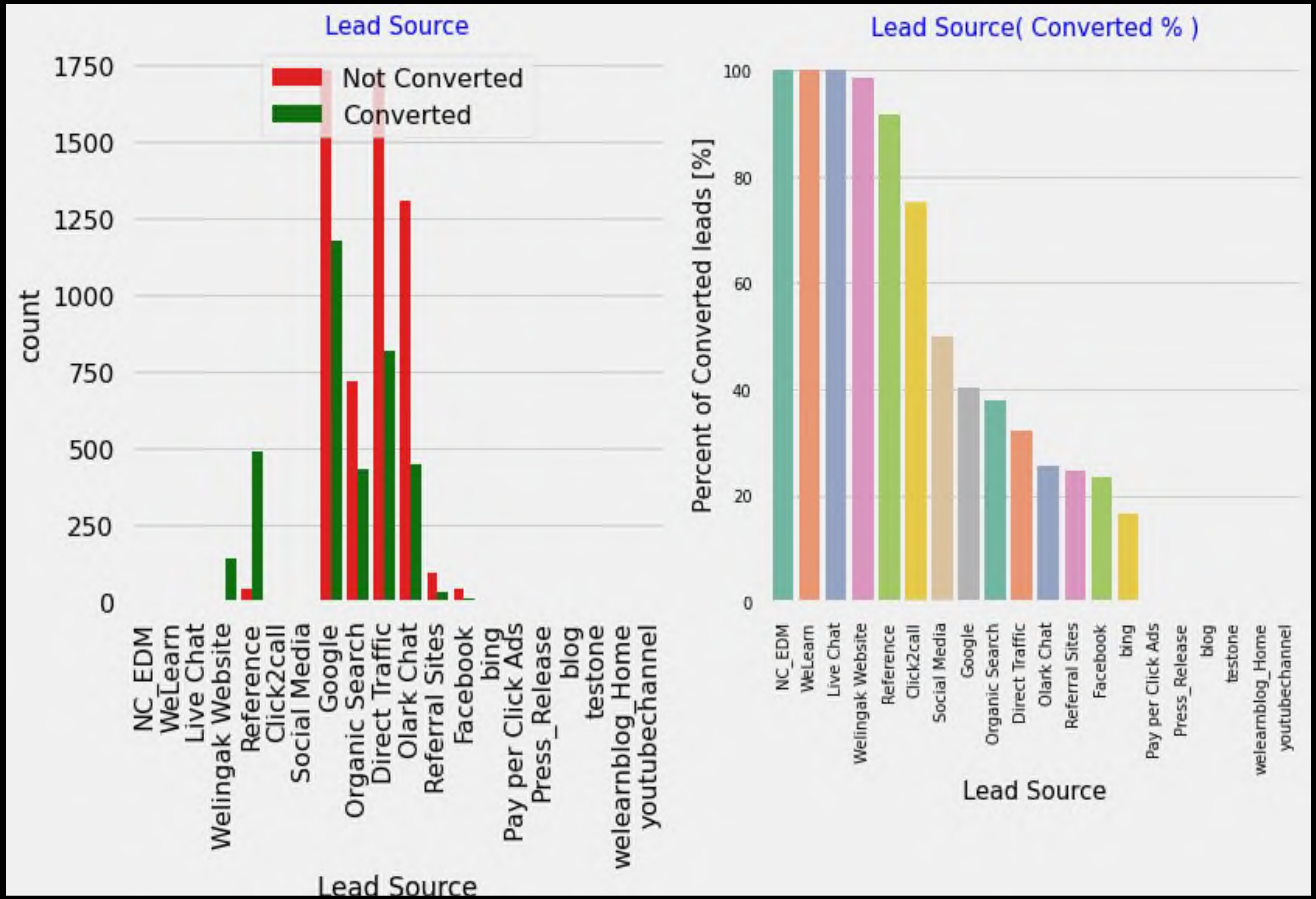


Univariate Analysis - Categorical



Insight:

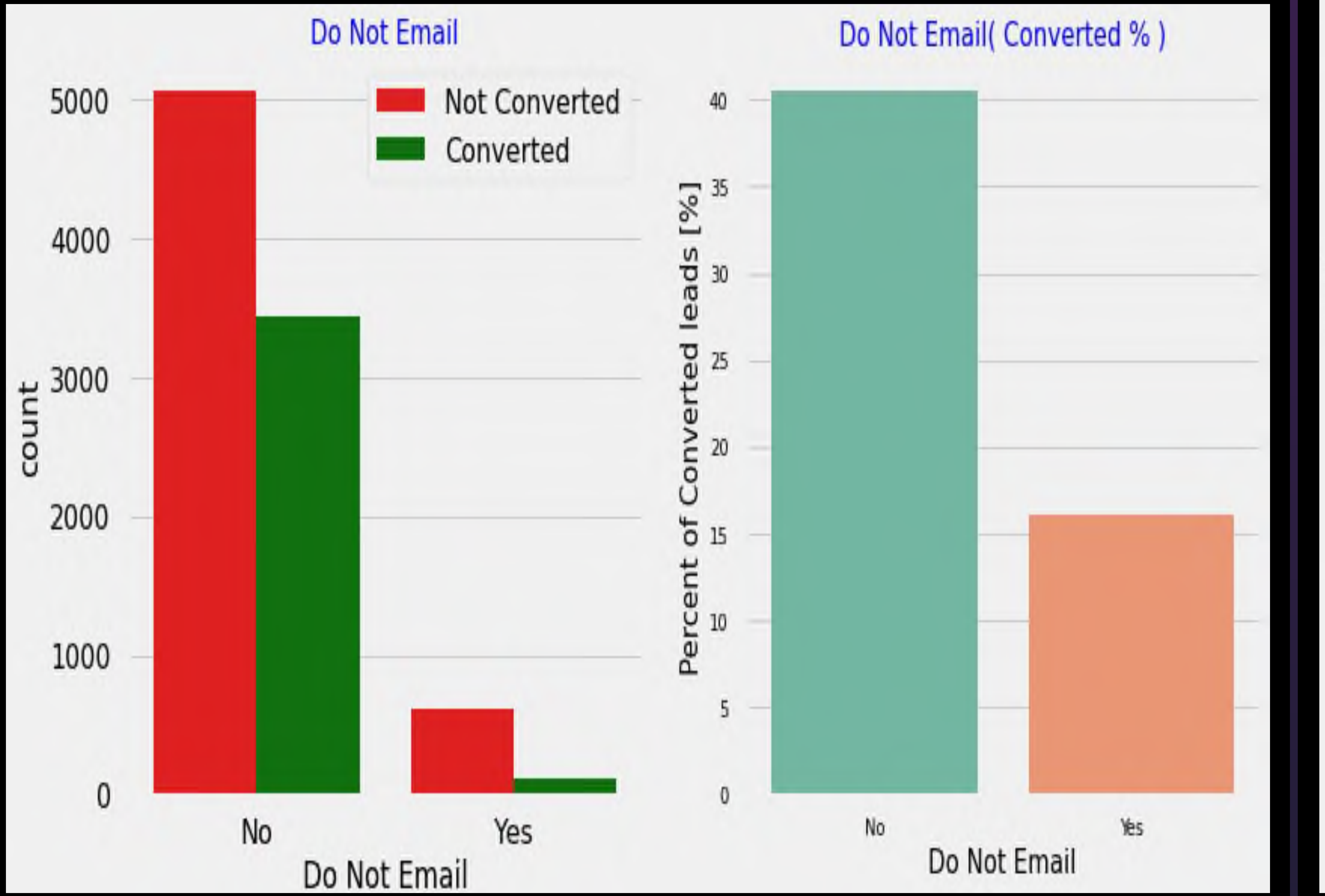
- Most Leads originated from submissions on the landing page and around 38% of those are converted followed by API, where around 30% are converted.
- Even though Lead Origins from Quick Add Form are 100% Converted, there was just 1 lead from that category. Leads from the Lead Add Form are the next highest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



Insight:

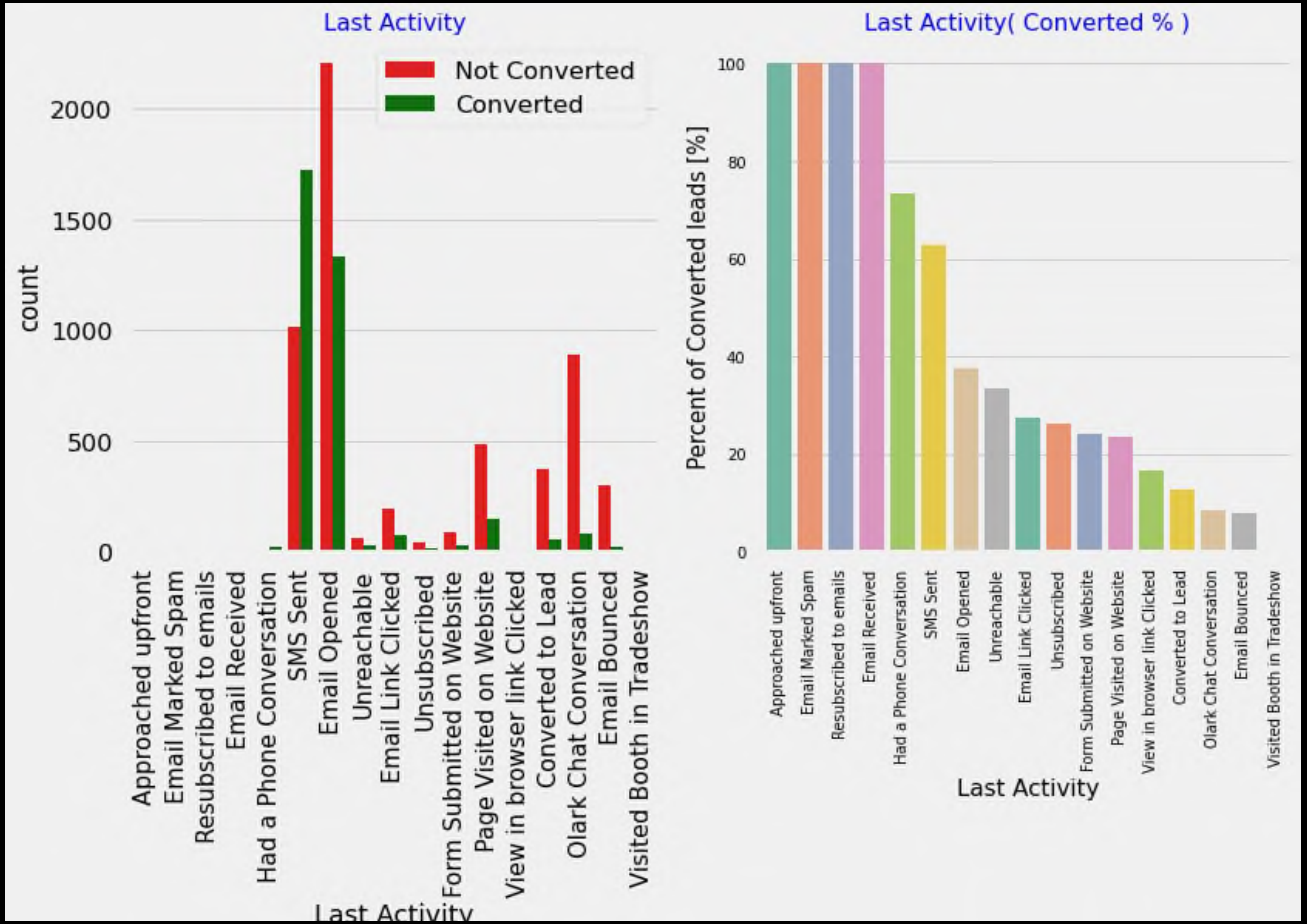
- The source of most leads was Google, and 40% of the leads converted, followed by Direct Traffic,Organic search and Olark chat where around 35%, 38% and 30% converted respectively.
- A lead that came from a reference has over 90% conversion from the total of 534.
- Welingak Website has almost 100% lead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

Univariate Analysis - Categorical



Insight:

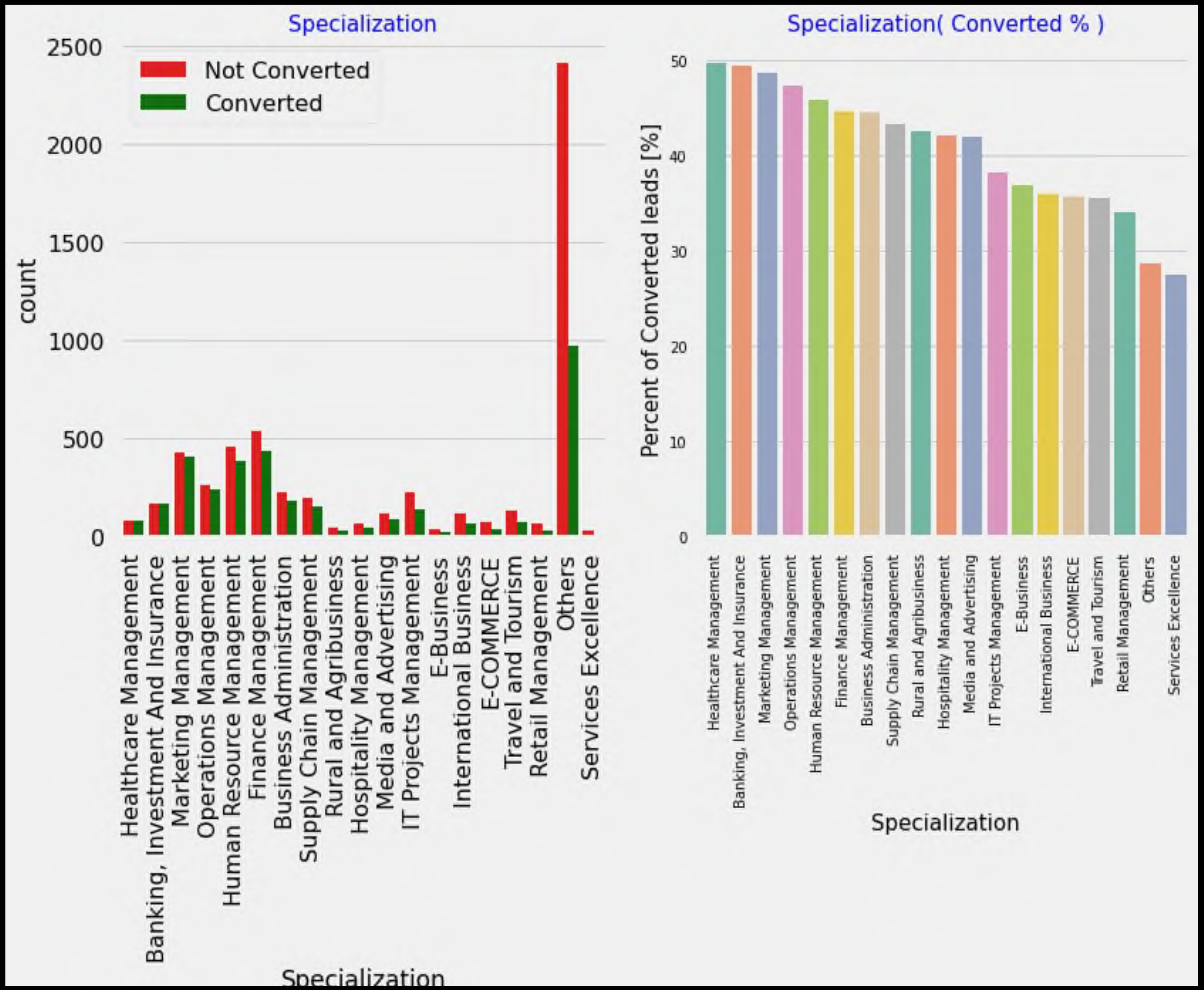
- Majority of the people are ok with receiving email (~92%)
- People who are ok with email has conversion rate of 40%
- People who have opted out of receive email has lower rate of conversion (only 15%)



Insight:

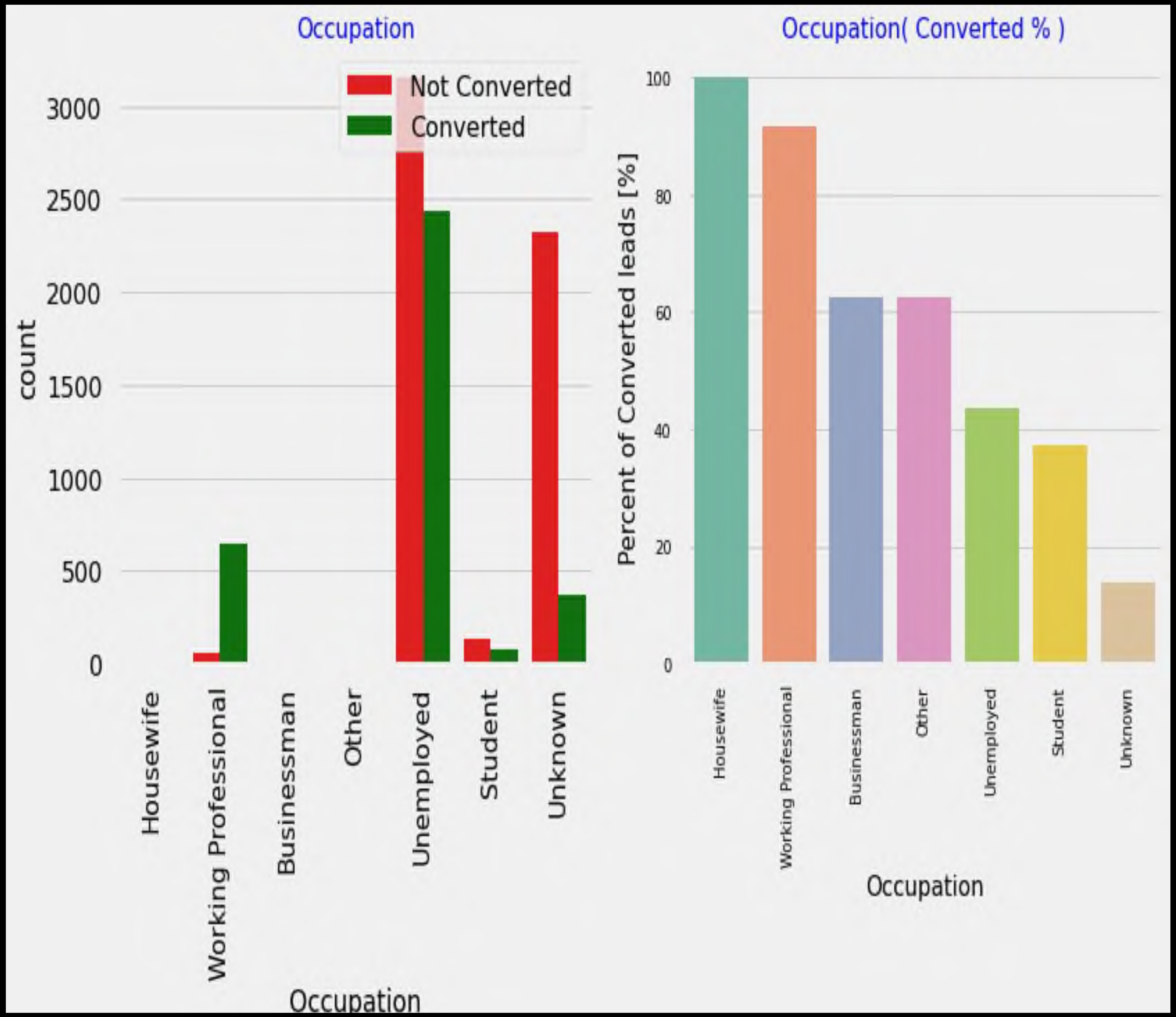
- Most of the lead have their Email opened as their last activity
- After combining smaller Last Activity types as Other Activity, the lead conversion is very high (~70%)
- Conversion rate for leads with last activity as SMS Sent is almost 60%

Univariate Analysis - Categorical



Insight:

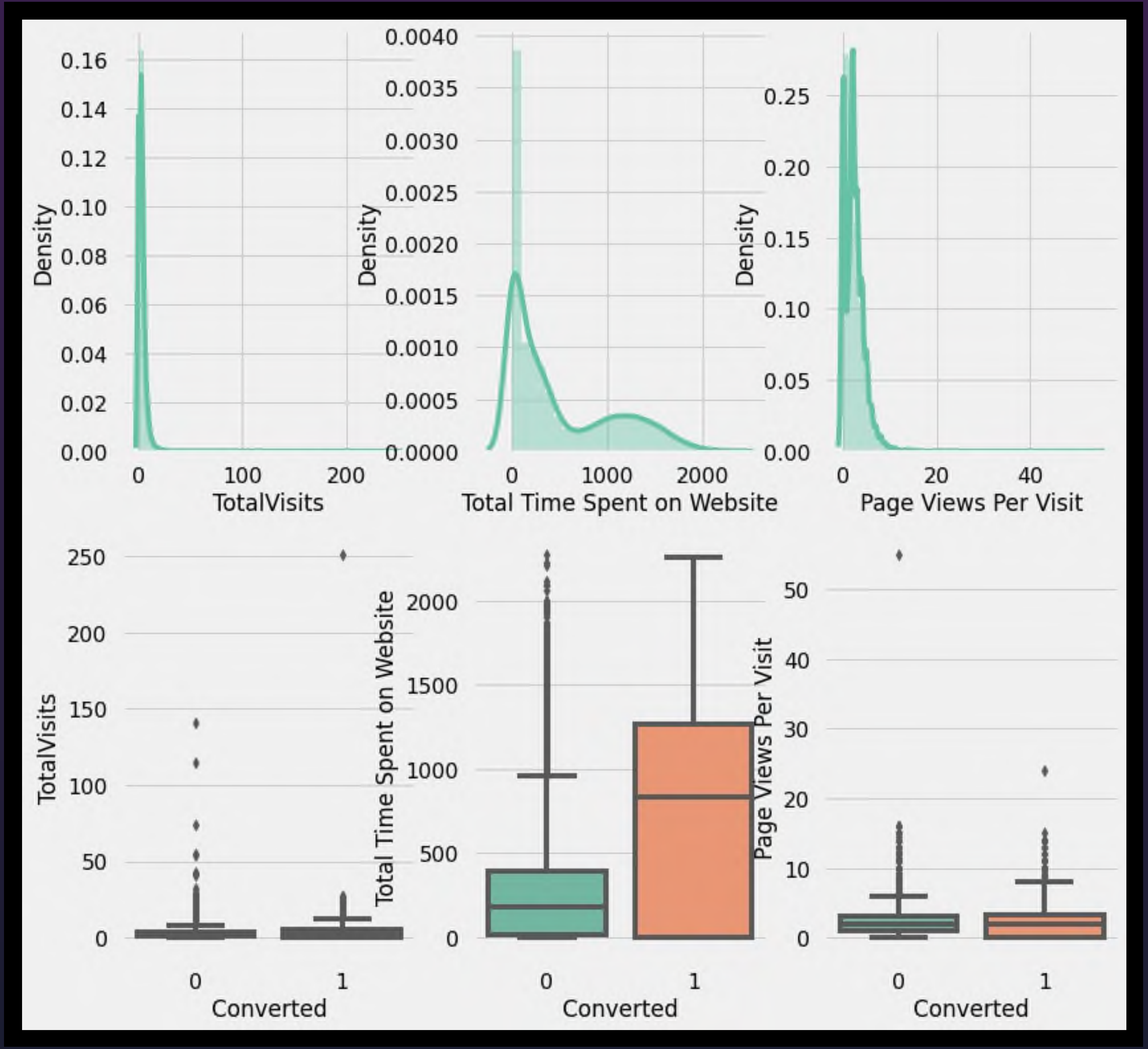
- Most of the leads have not mentioned a specialization and around 28% of those converted
- Leads with Finance management and Marketing Management - Over 45% Converted



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%)

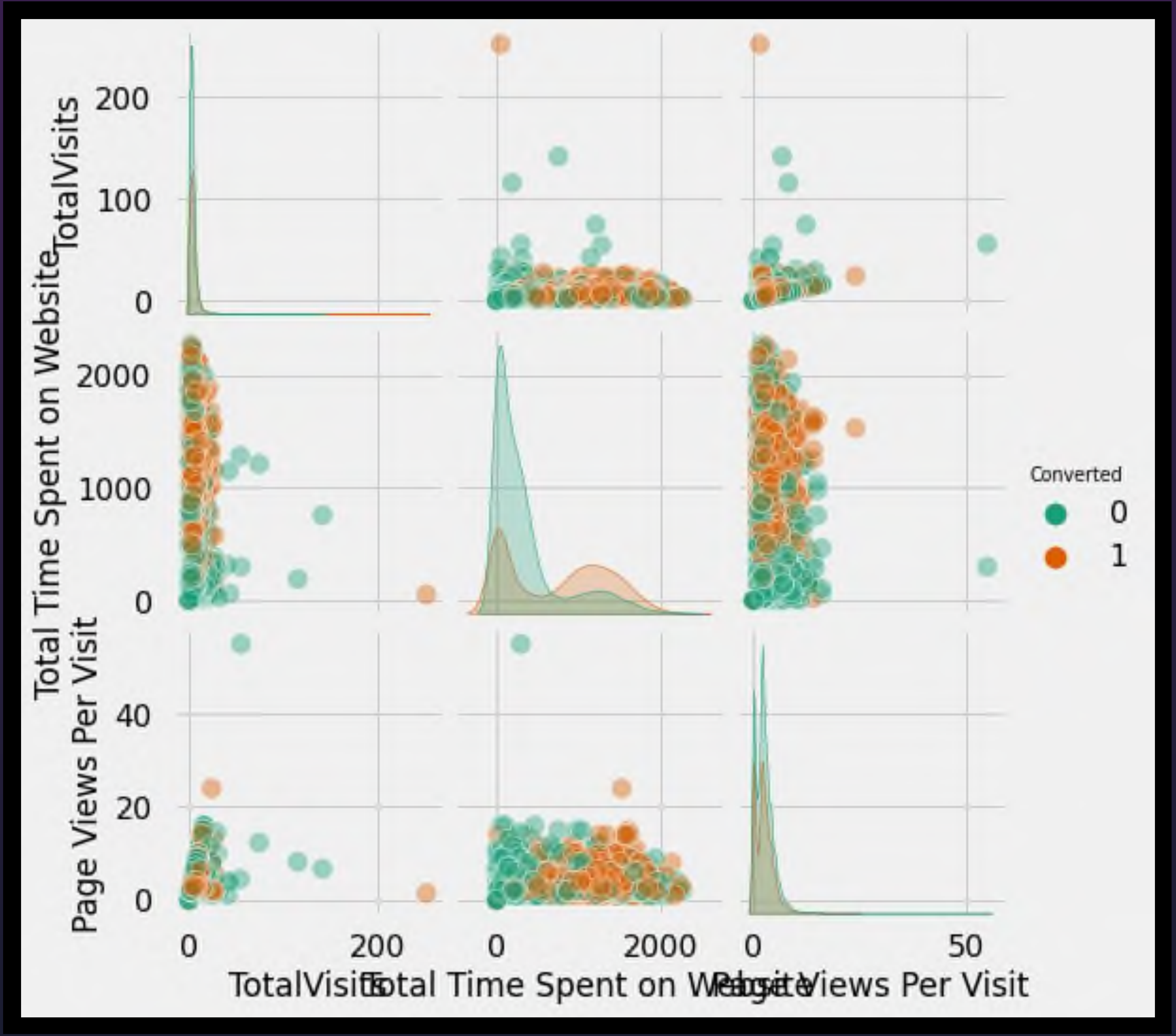
Univariate Analysis - Numerical



Insight:

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

Bivariate Analysis - Numerical



Insight:

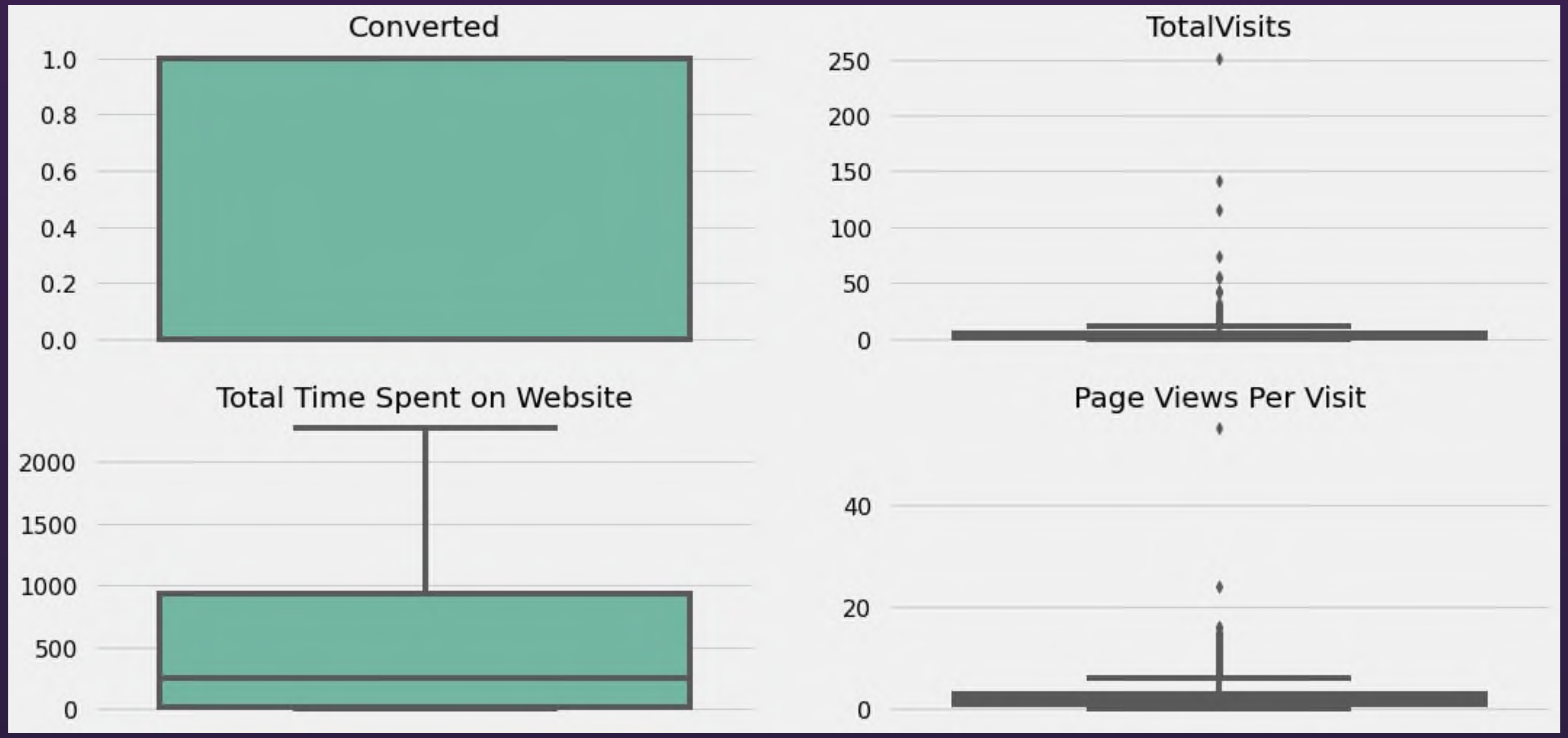
Data is not normally distributed.

Outliers Treatment

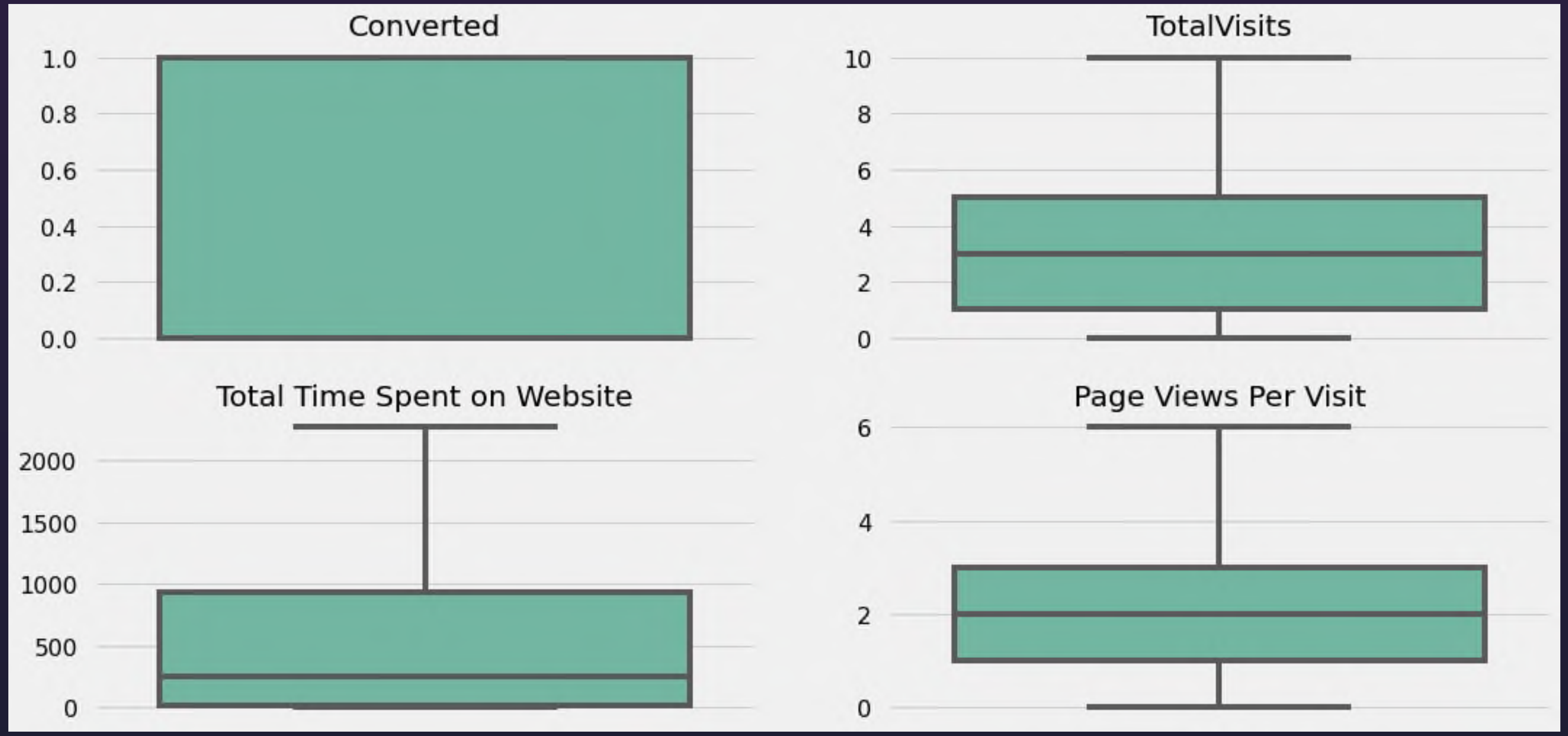
Insight:

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.

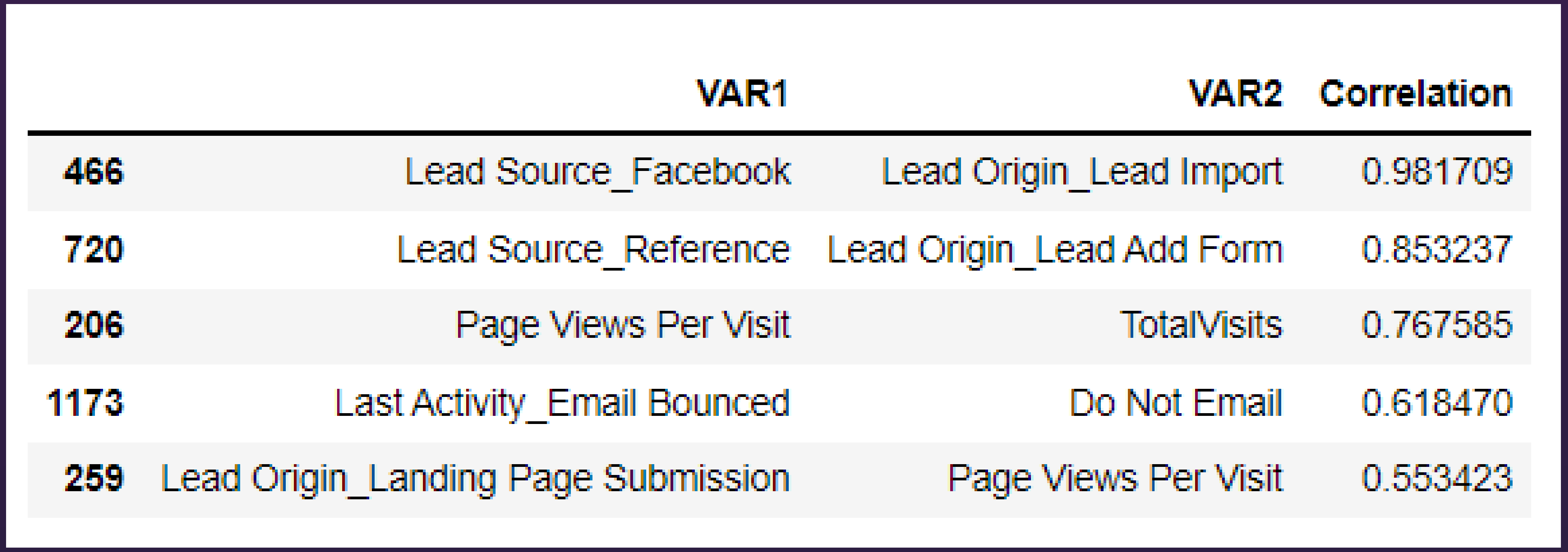


Box Plot before handling outliers



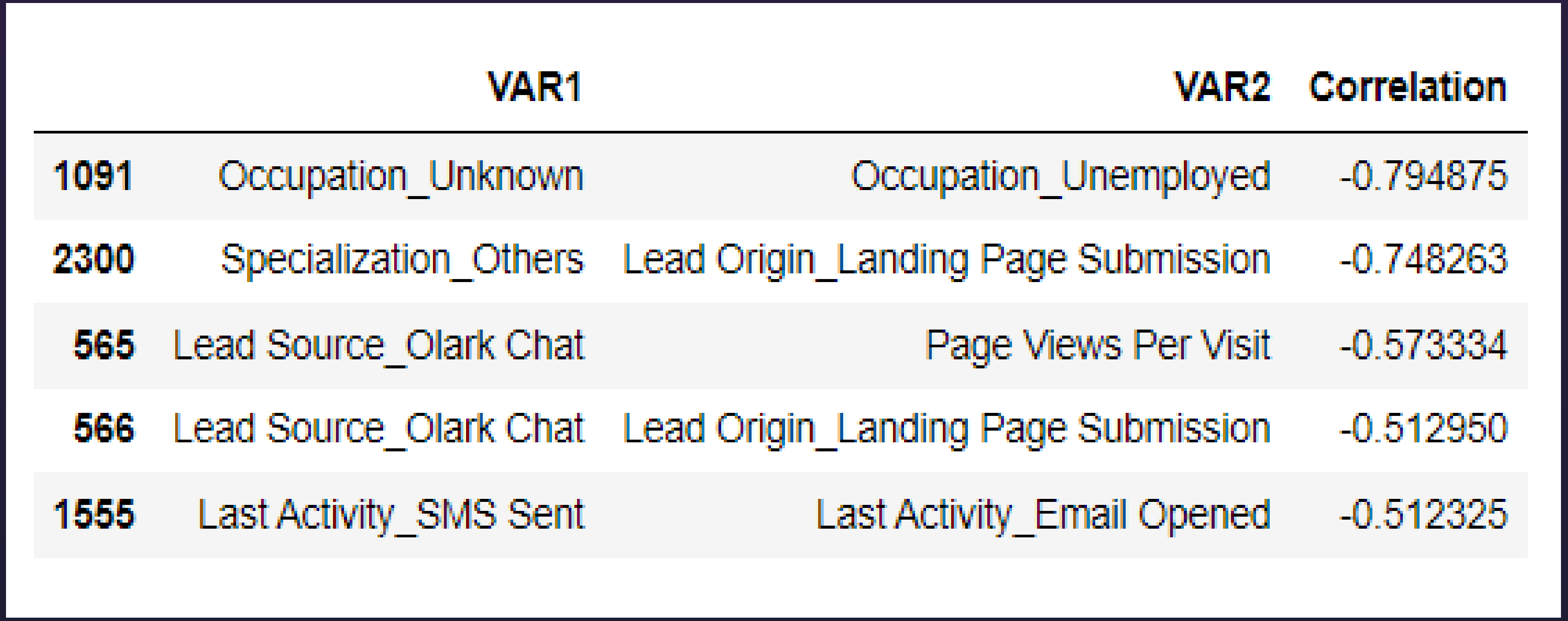
Box Plot after handling outliers

Correlation of variables with Target variable



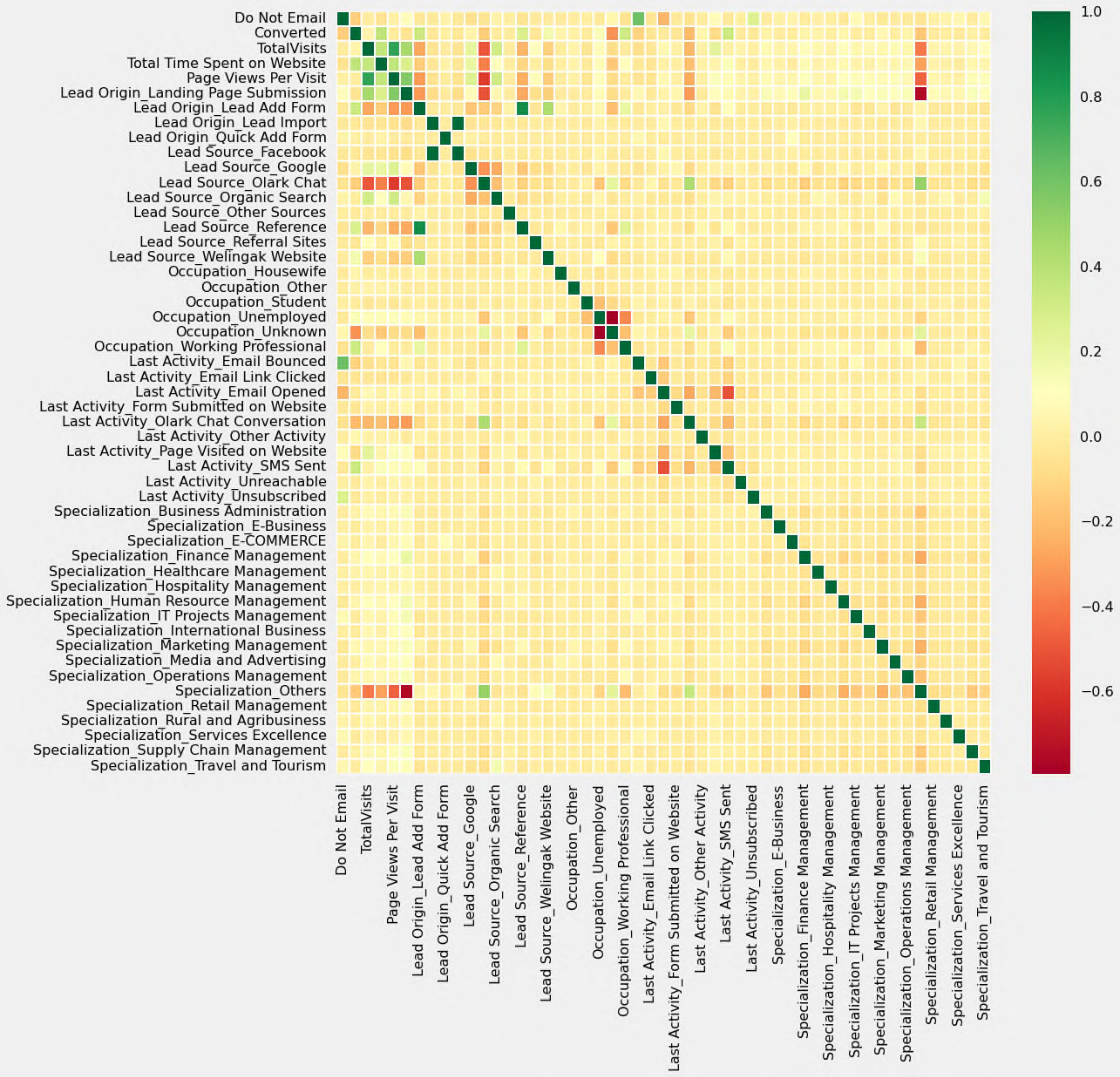
	VAR1	VAR2	Correlation
466	Lead Source_Facebook	Lead Origin_Lead Import	0.981709
720	Lead Source_Reference	Lead Origin_Lead Add Form	0.853237
206	Page Views Per Visit	TotalVisits	0.767585
1173	Last Activity_Email Bounced	Do Not Email	0.618470
259	Lead Origin_Landing Page Submission	Page Views Per Visit	0.553423

Top 5 Positive correlated variables



	VAR1	VAR2	Correlation
1091	Occupation_Unknown	Occupation_Unemployed	-0.794875
2300	Specialization_Others	Lead Origin_Landing Page Submission	-0.748263
565	Lead Source_Olark Chat	Page Views Per Visit	-0.573334
566	Lead Source_Olark Chat	Lead Origin_Landing Page Submission	-0.512950
1555	Last Activity_SMS Sent	Last Activity_Email Opened	-0.512325

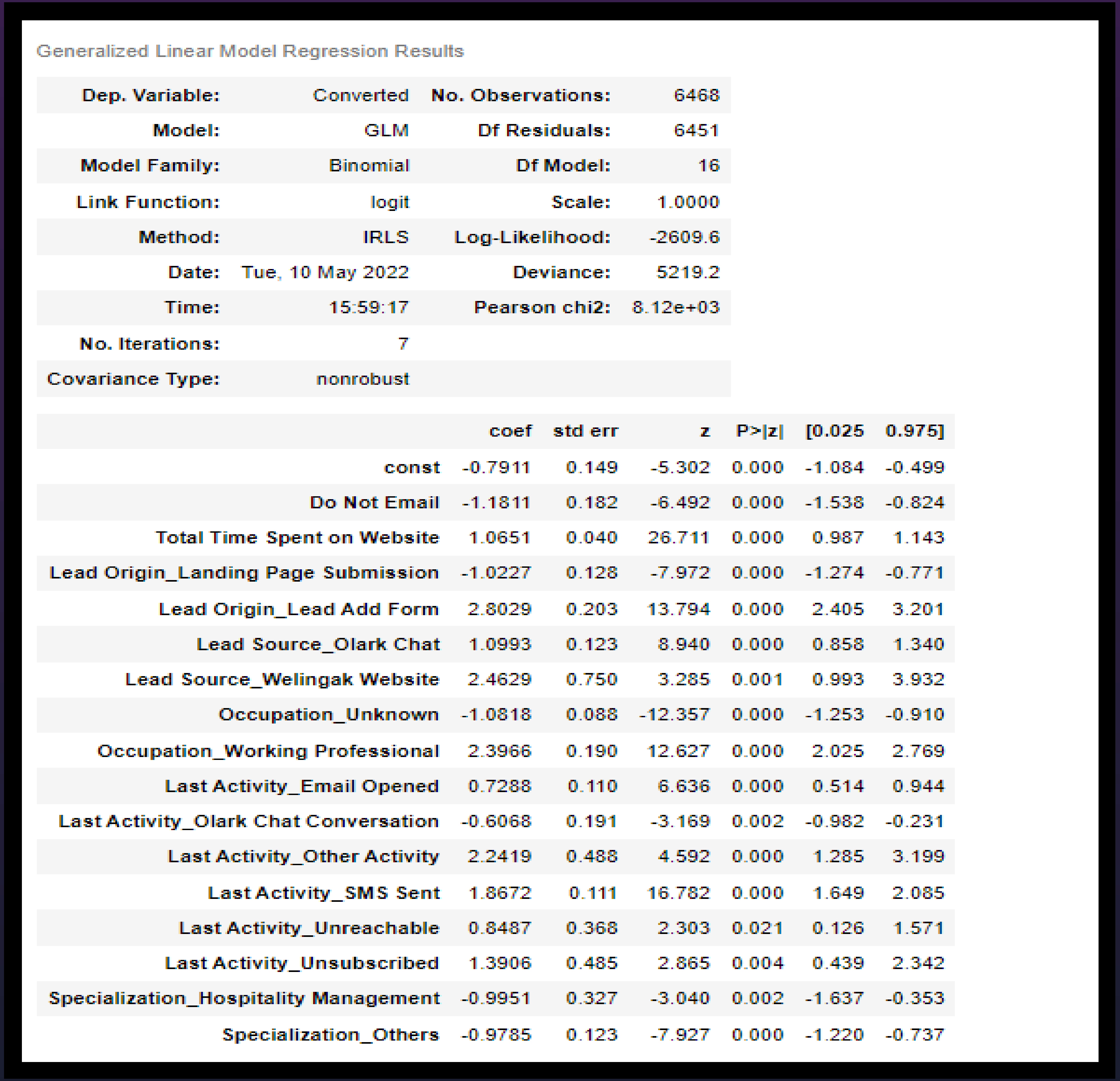
Top 5 Negative correlated variables



Model Building

RFE for Feature Reduction

- So far, we inspected, cleansed, eliminated and visualized the data.
- We also Standardized the continuous variables, one-hot encoded categorical variables and divided the dataset into training and test set
- However, there are still large number of variables, all of which may not be significant, or may have a high multi- collinearity.
- RFE was done to attain the top 20 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).
- The resulting dataset thus consists of features that are significant for the regression modelling



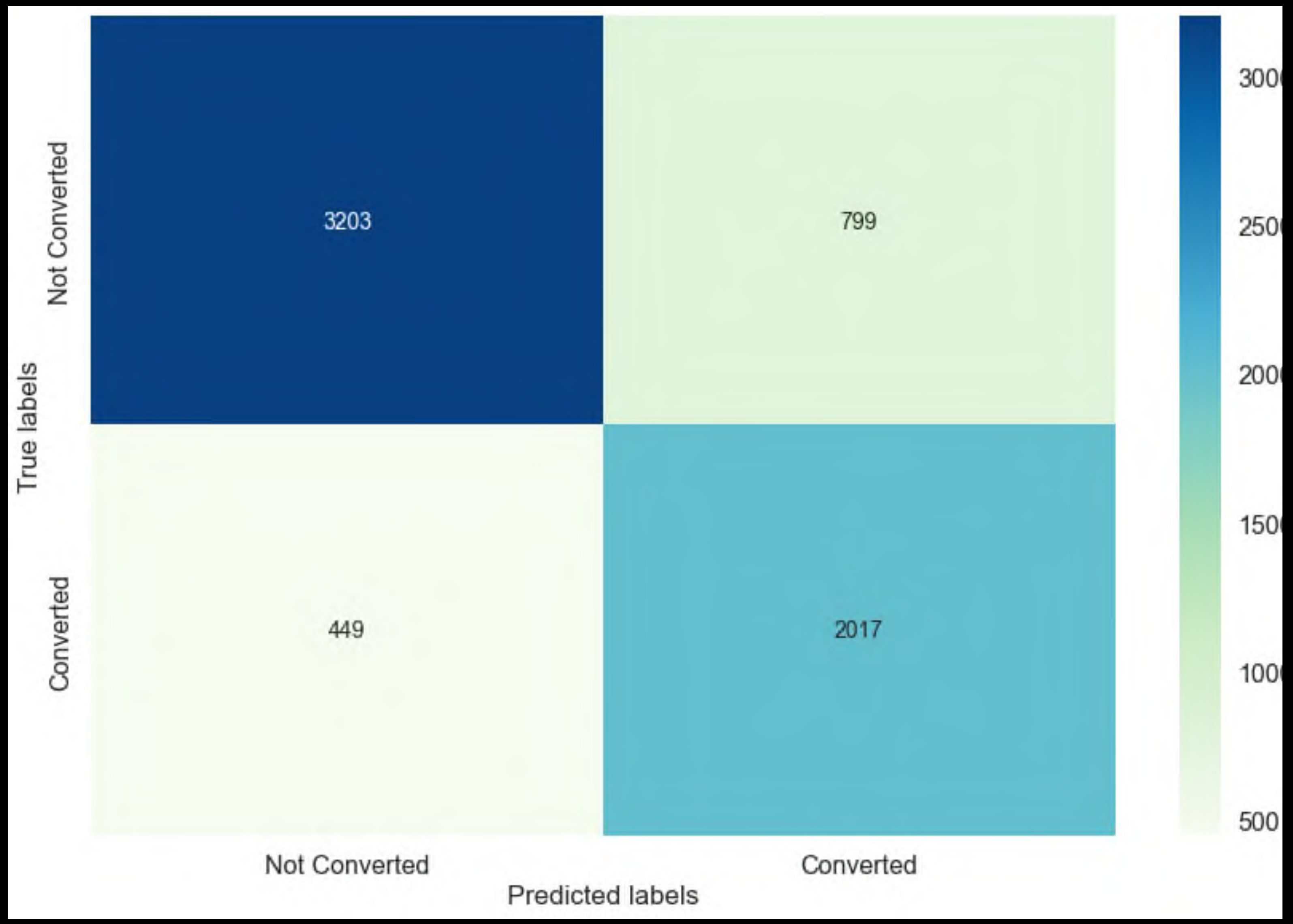
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6451
Model Family:	Binomial	Df Model:	16
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2609.6
Date:	Tue, 10 May 2022	Deviance:	5219.2
Time:	15:59:17	Pearson chi2:	8.12e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

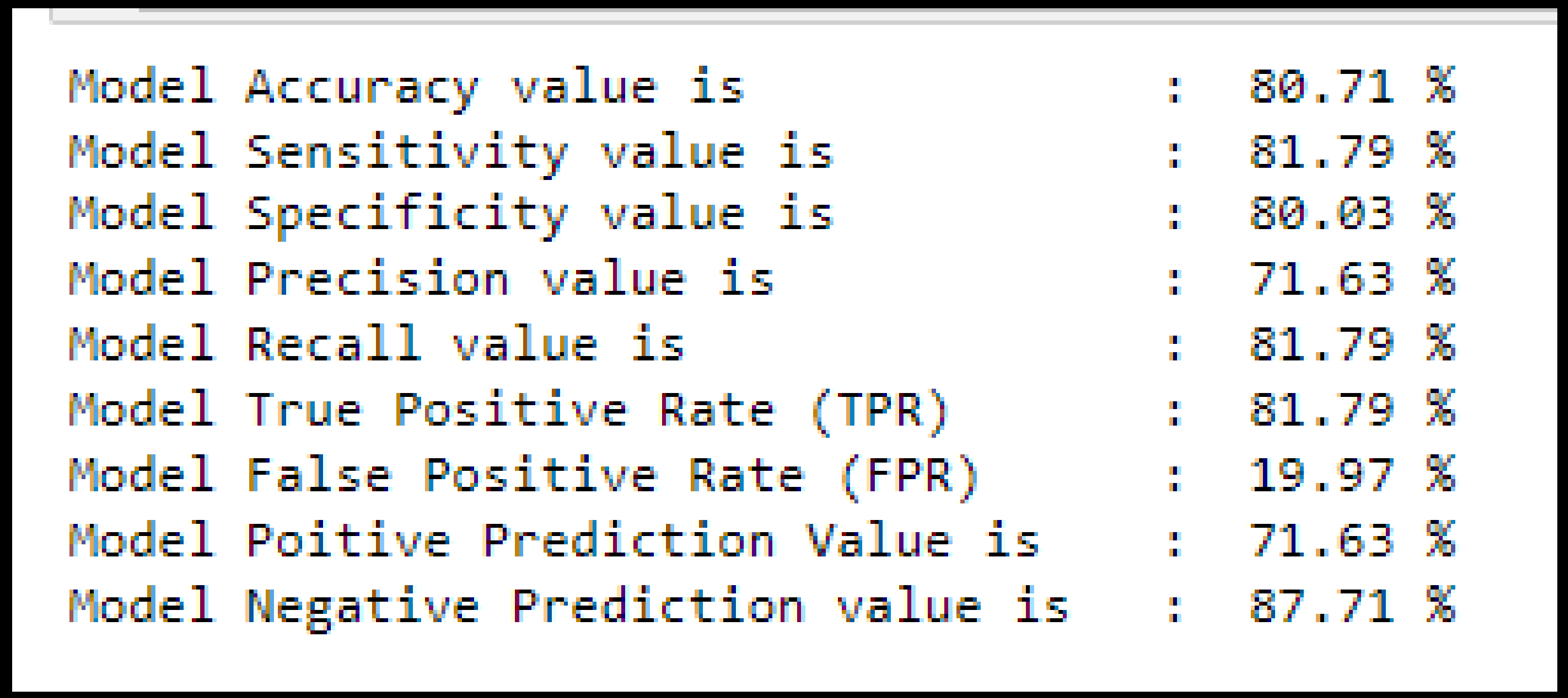
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.7911	0.149	-5.302	0.000	-1.084	-0.499
Do Not Email	-1.1811	0.182	-6.492	0.000	-1.538	-0.824
Total Time Spent on Website	1.0651	0.040	26.711	0.000	0.987	1.143
Lead Origin_Landing Page Submission	-1.0227	0.128	-7.972	0.000	-1.274	-0.771
Lead Origin_Lead Add Form	2.8029	0.203	13.794	0.000	2.405	3.201
Lead Source_Olark Chat	1.0993	0.123	8.940	0.000	0.858	1.340
Lead Source_Welingak Website	2.4629	0.750	3.285	0.001	0.993	3.932
Occupation_Unknown	-1.0818	0.088	-12.357	0.000	-1.253	-0.910
Occupation_Working Professional	2.3966	0.190	12.627	0.000	2.025	2.769
Last Activity_Email Opened	0.7288	0.110	6.636	0.000	0.514	0.944
Last Activity_Olark Chat Conversation	-0.6068	0.191	-3.169	0.002	-0.982	-0.231
Last Activity_Other Activity	2.2419	0.488	4.592	0.000	1.285	3.199
Last Activity_SMS Sent	1.8672	0.111	16.782	0.000	1.649	2.085
Last Activity_Unreachable	0.8487	0.368	2.303	0.021	0.126	1.571
Last Activity_Unsubscribed	1.3906	0.485	2.865	0.004	0.439	2.342
Specialization_Hospitality Management	-0.9951	0.327	-3.040	0.002	-1.637	-0.353
Specialization_Others	-0.9785	0.123	-7.927	0.000	-1.220	-0.737

Model Evaluation : Train Dataset

Confusion Matrix

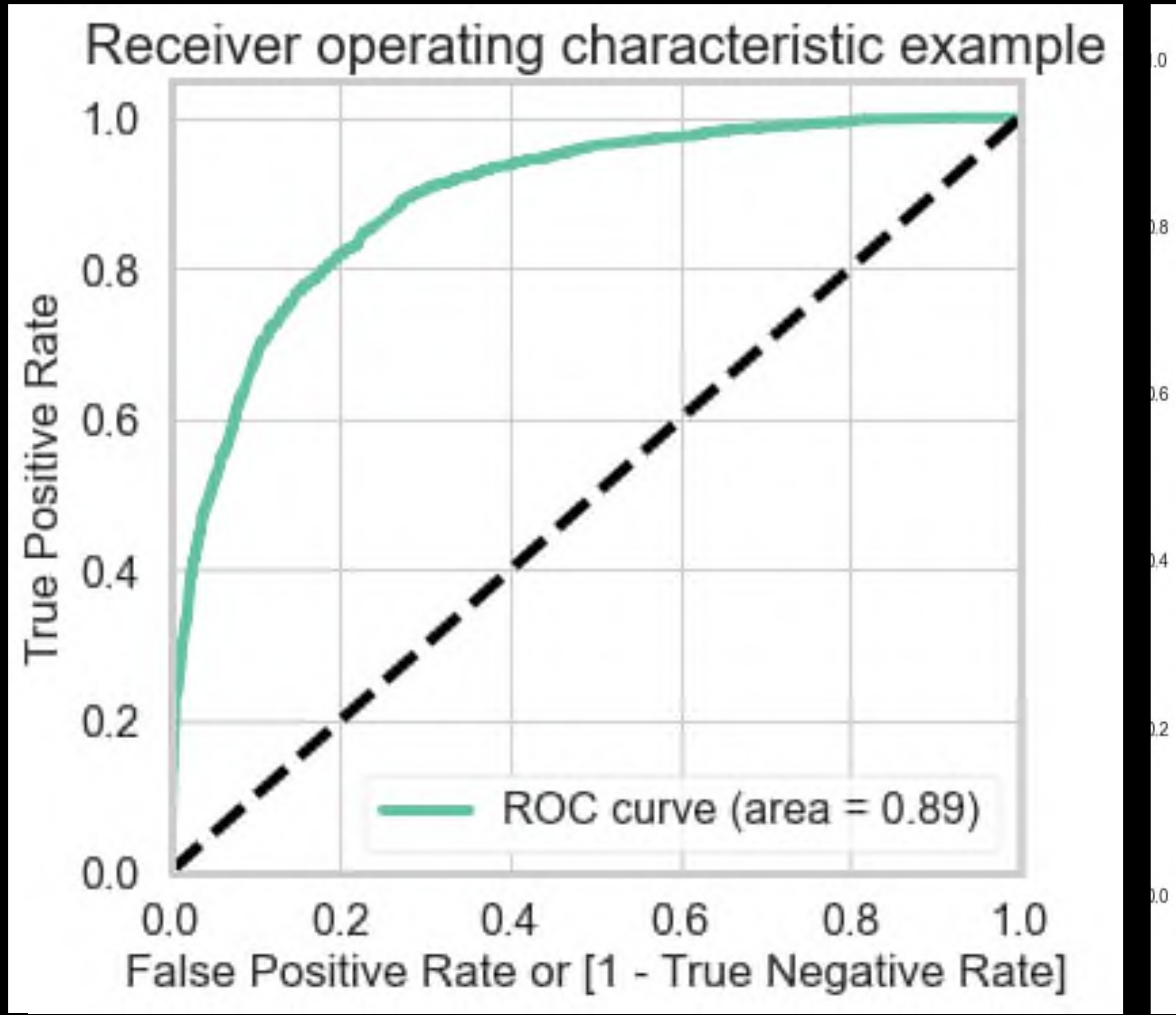


Model Performance



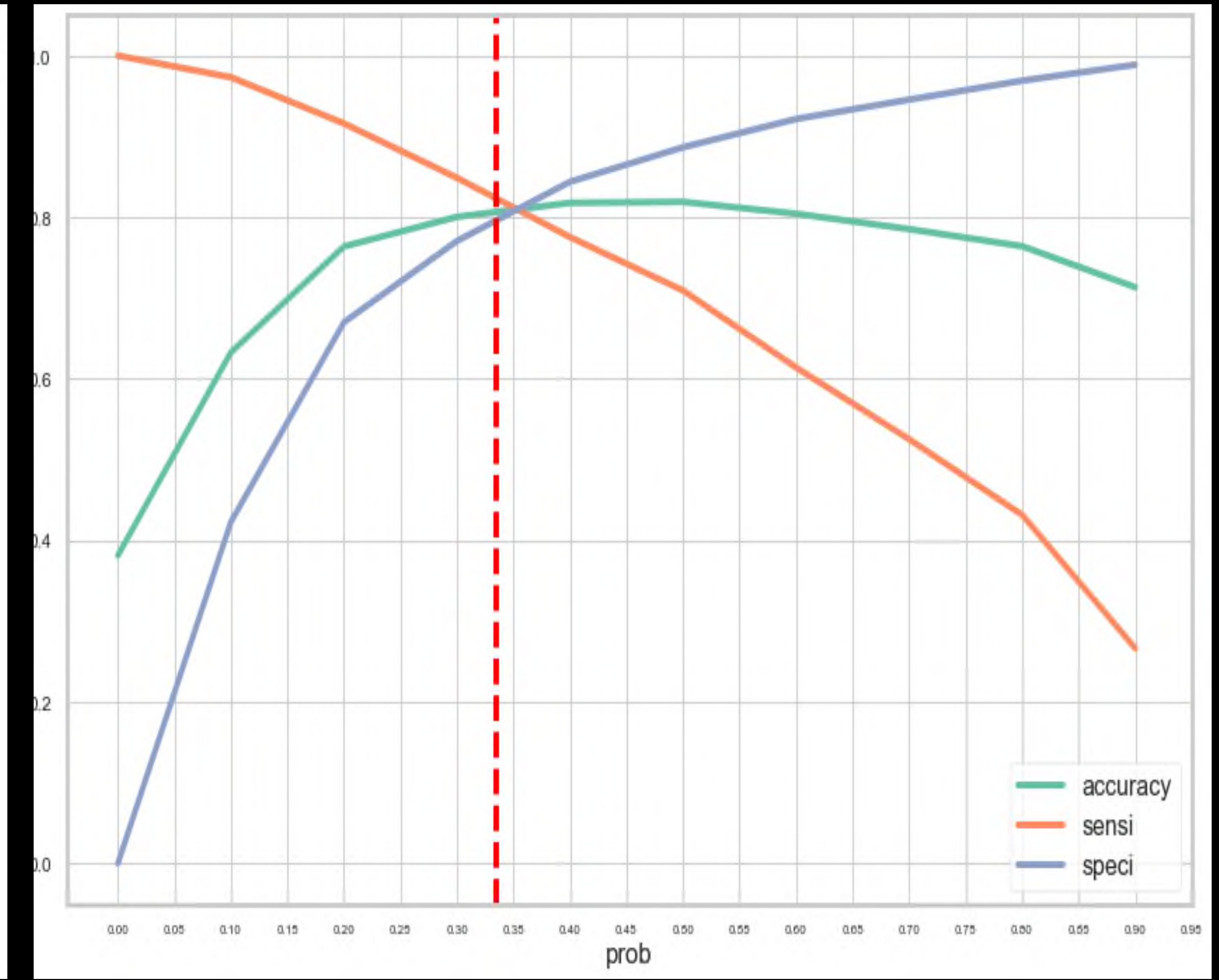
```
Model Accuracy value is              :  80.71 %
Model Sensitivity value is           :  81.79 %
Model Specificity value is           :  80.03 %
Model Precision value is             :  71.63 %
Model Recall value is                :  81.79 %
Model True Positive Rate (TPR)       :  81.79 %
Model False Positive Rate (FPR)      :  19.97 %
Model Poitive Prediction Value is    :  71.63 %
Model Negative Prediction value is   :  87.71 %
```

ROC Curve



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

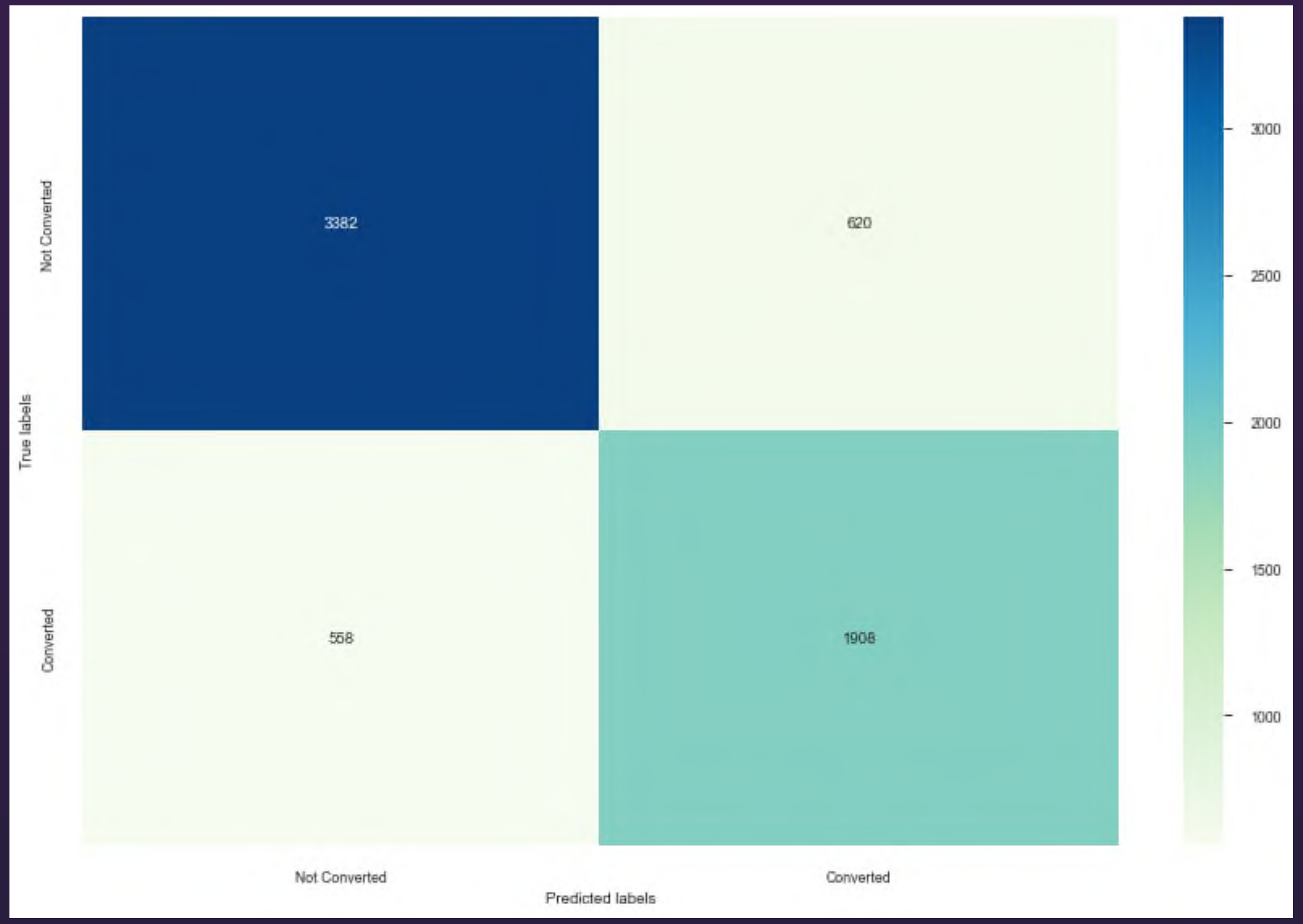
Optimal Cut-off



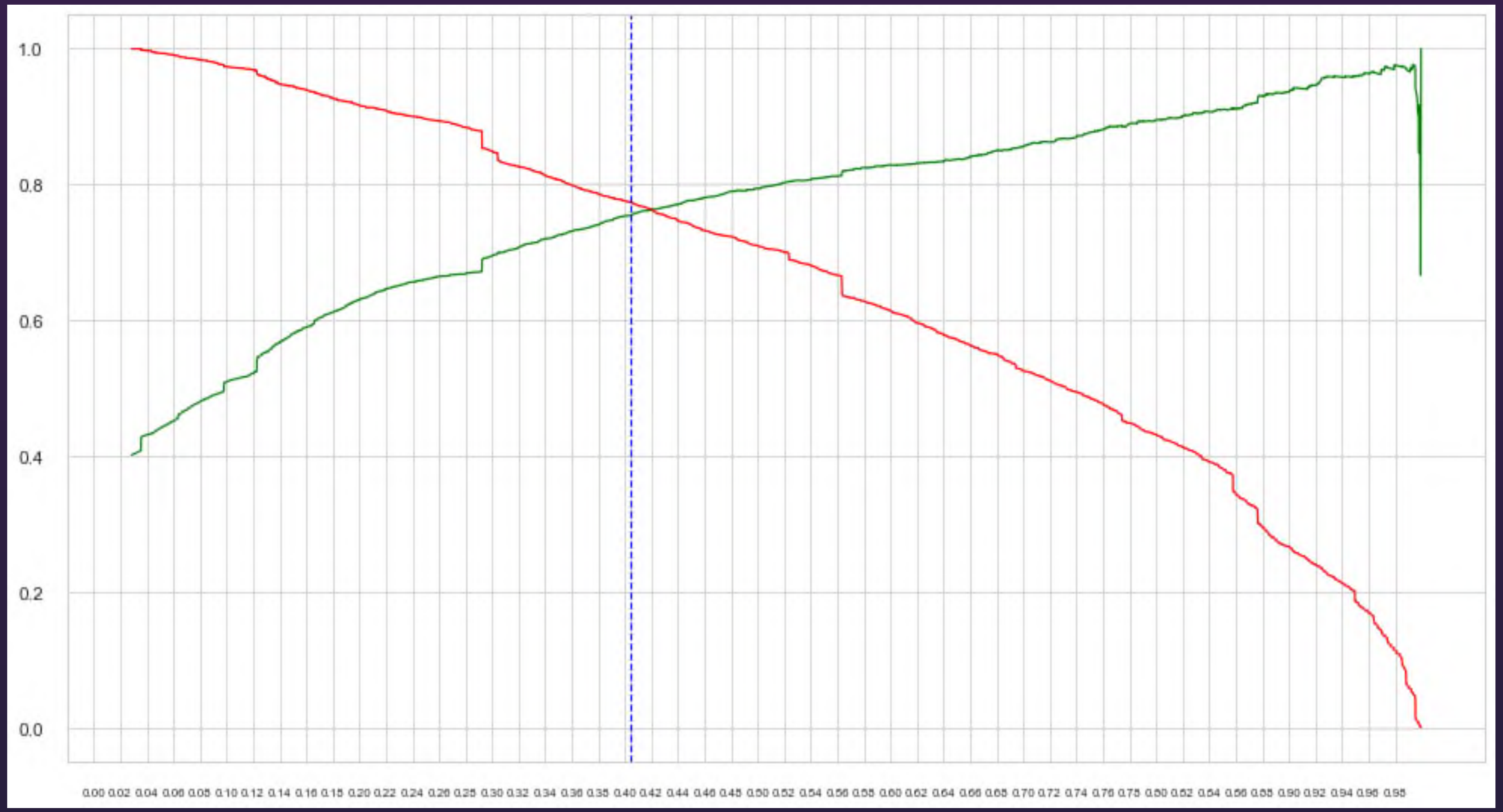
From the above graph, 0.335 seems to be ideal cut-off points

Precision - Recall Trade off

Confusion Matrix



Precision – Recall Curve



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

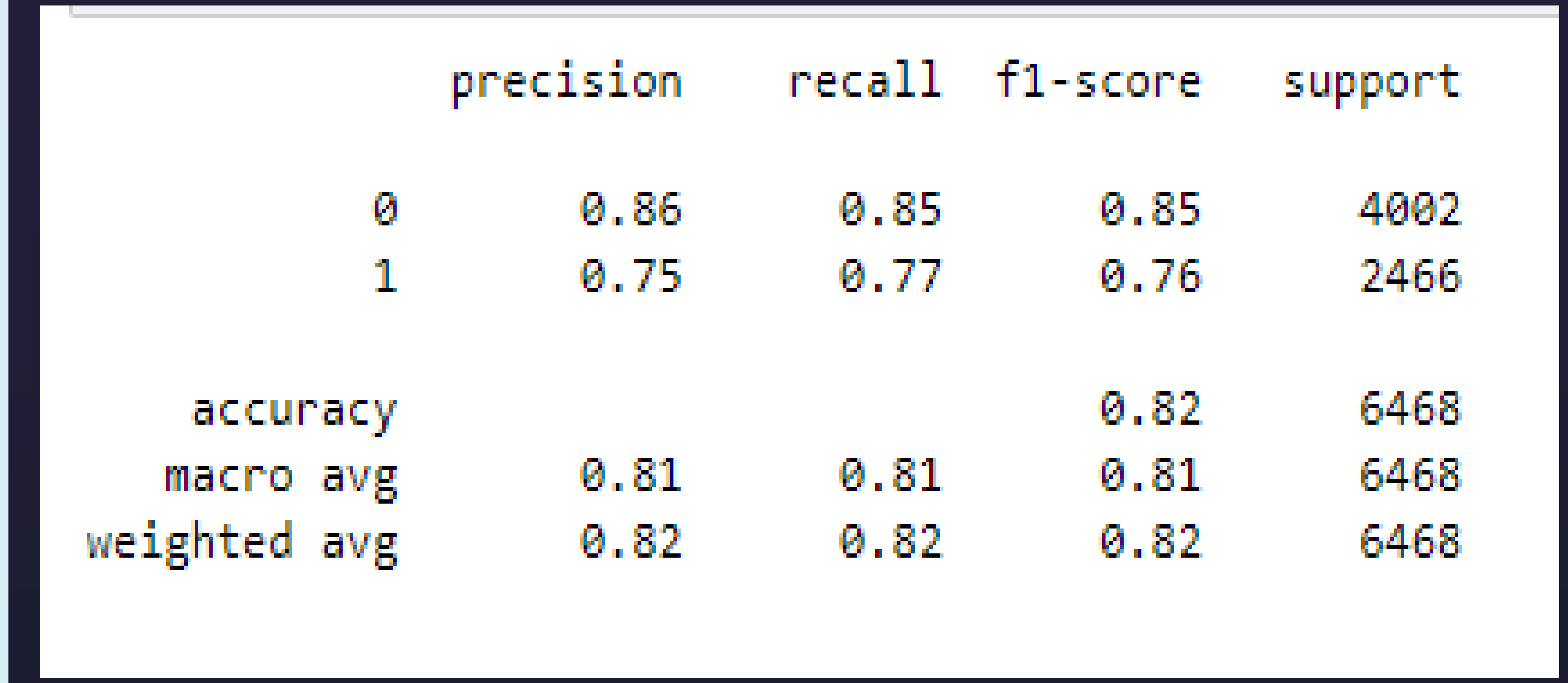
Inferences:

By using the Precision - Recall trade off chart cut-off points, the model output has changed 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 0.34 as cutoff point.

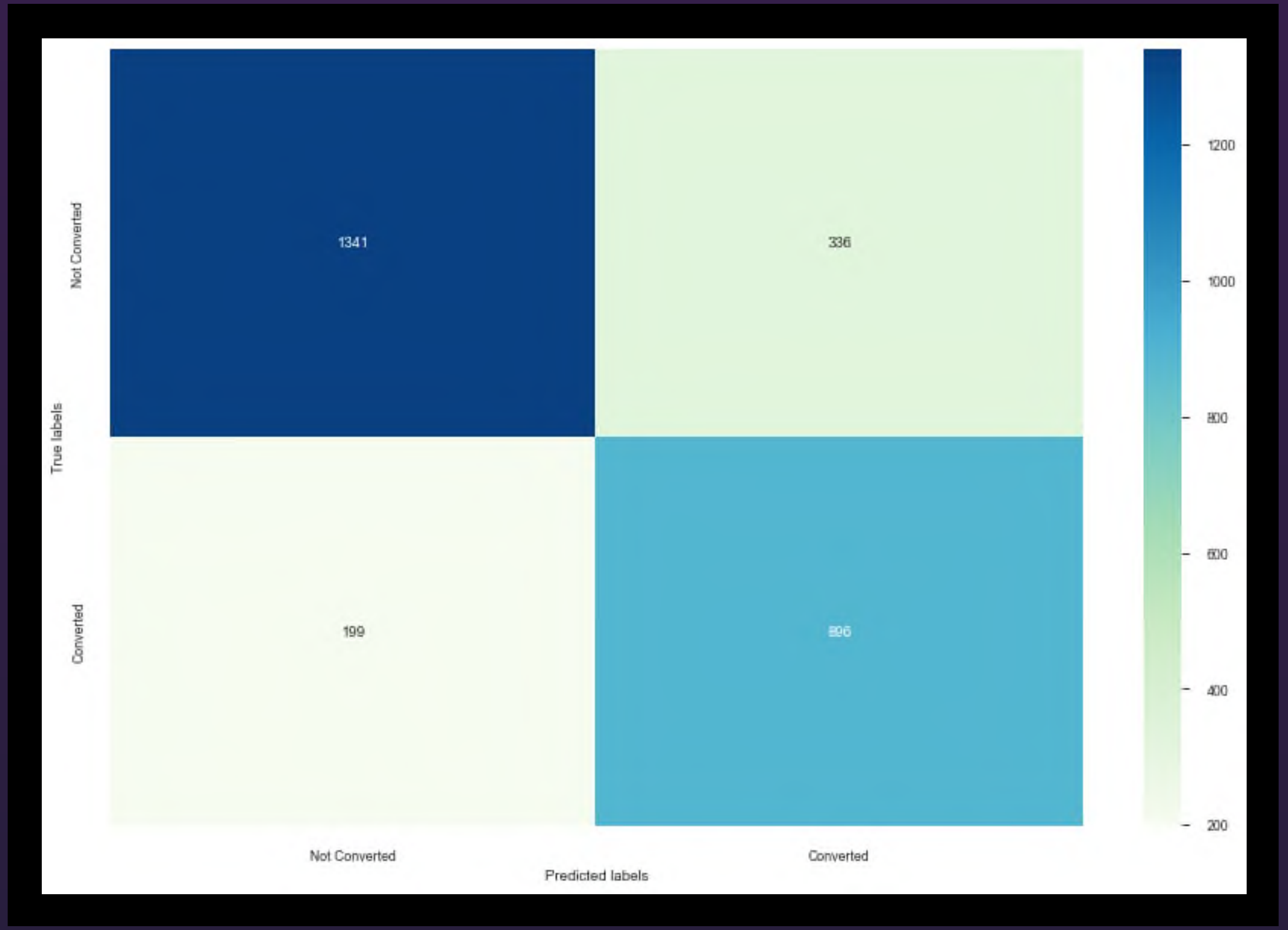
Model Performance



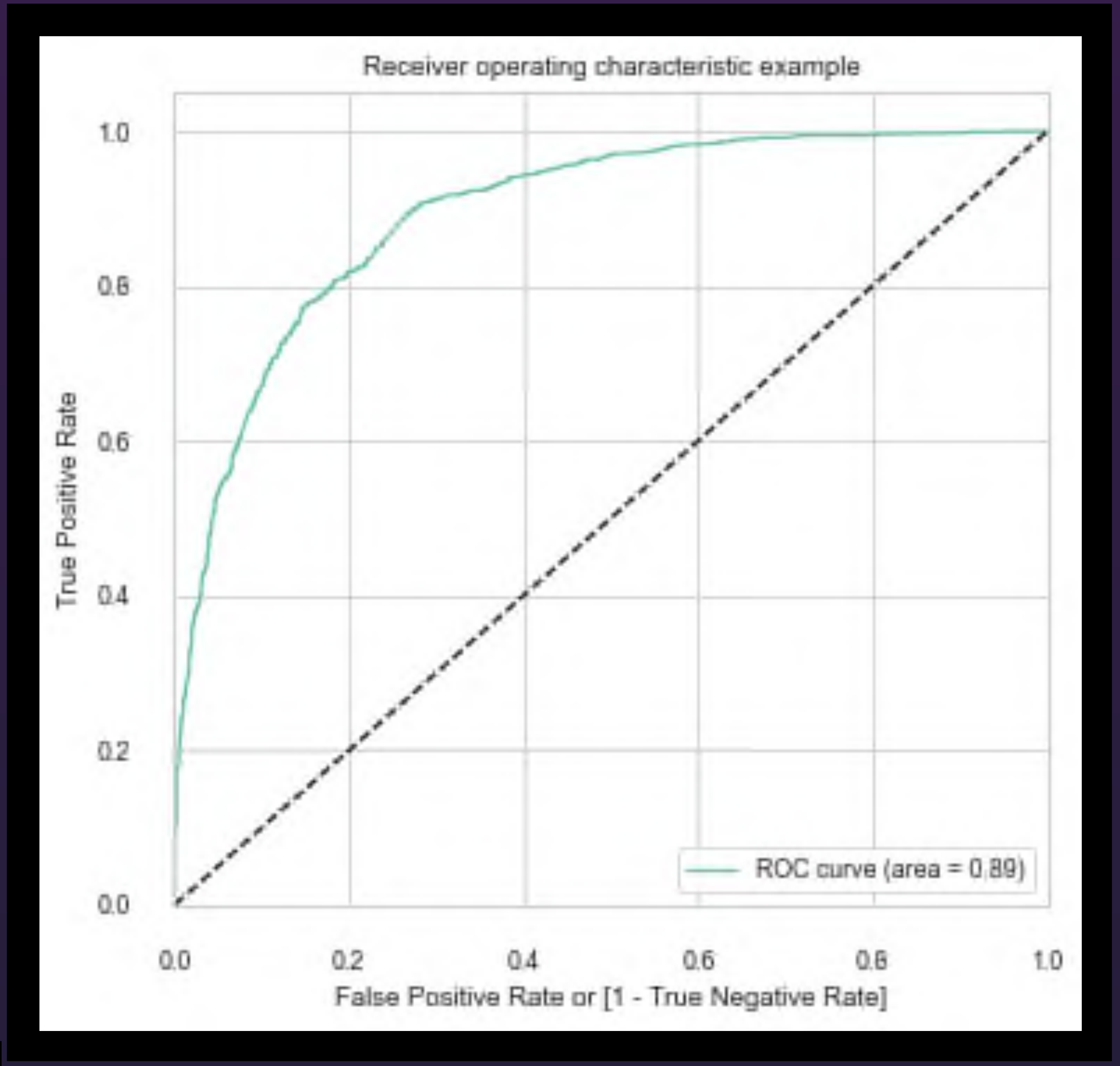
```
              precision    recall  f1-score   support

           0       0.86      0.85      0.85      4002
           1       0.75      0.77      0.76      2466

    accuracy                           0.82      6468
   macro avg       0.81      0.81      0.81      6468
weighted avg       0.82      0.82      0.82      6468
```

Model Evaluation : Test Dataset

Confusion Matrix

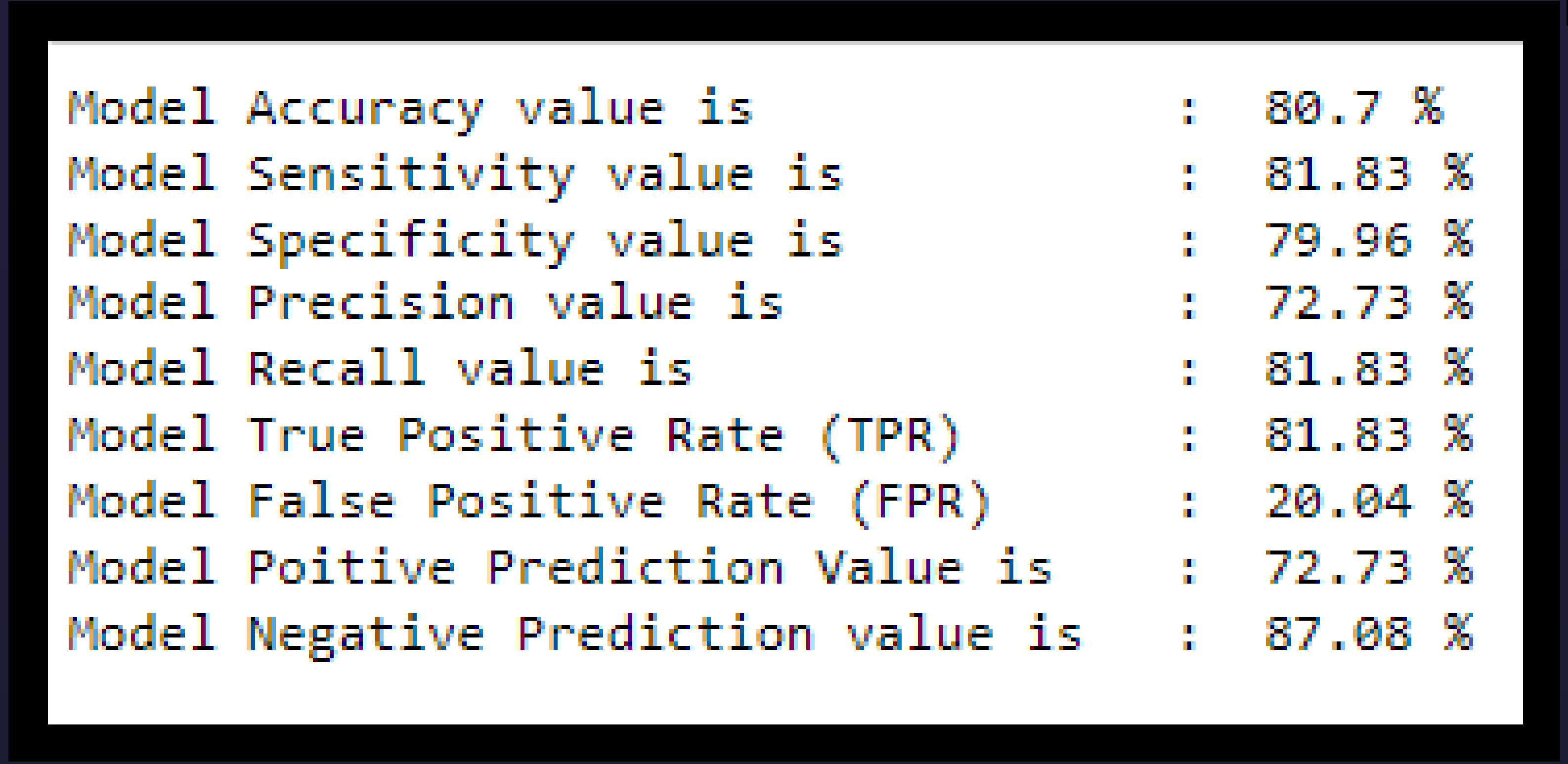


ROC Curve



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

Model Performance



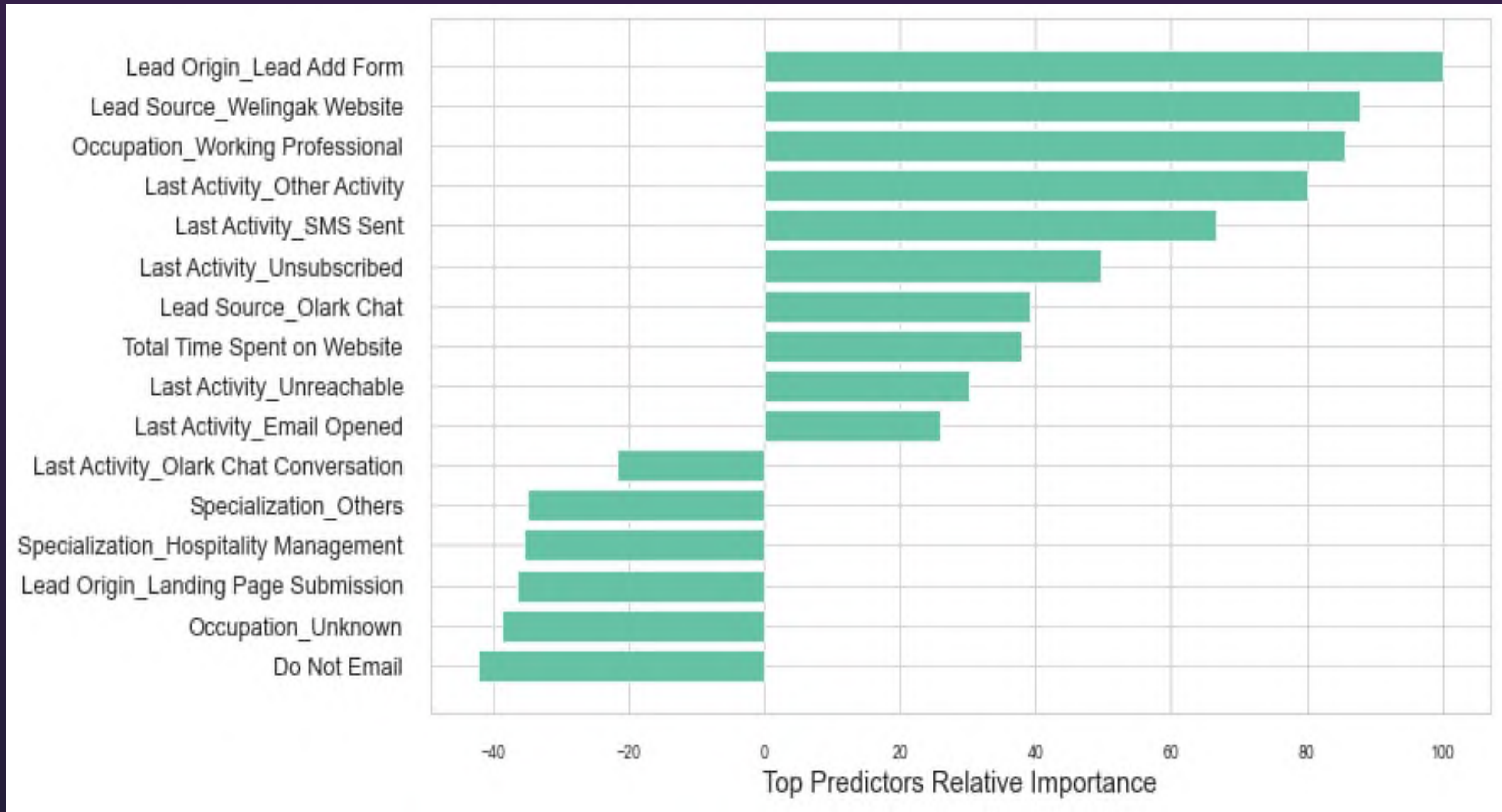
```
Model Accuracy value is                 :   80.7 %
Model Sensitivity value is              :   81.83 %
Model Specificity value is              :   79.96 %
Model Precision value is                :   72.73 %
Model Recall value is                   :   81.83 %
Model True Positive Rate (TPR)          :   81.83 %
Model False Positive Rate (FPR)         :   20.04 %
Model Poitive Prediction Value is       :   72.73 %
Model Negative Prediction value is      :   87.08 %
```

The sensitivity value on Test data is 81.83% vs 80.29% in Train data. The accuracy values is 80.7%. It shows that model is performing well in test data set also and is not over-trained.

Inferences and Recommendations

Major indicators that a lead will get converted to a hot lead:

1. Lead Origin_Lead Add Form : A lead sourced from Lead Origin_Lead Add Form is more likely to get converted.
2. Occupation_Working Professional :- Working professionals are more likely to get converted.
3. Lead_Source_Welingak website : A lead sourced from Welingak Website is more likely to get converted.
4. Last Activity_SMS Sent :A lead having SMS sent previously are more likely to get converted.
5. Lead Source_Olark Chat :A lead sourced from Olark Chat is more likely to get converted



Major indicators that a lead will NOT get converted to a hot lead:

1. Last_Activity_Olark chat conversation : Customer who had olark chat conversion, are less likely to get converted into hot leads.
2. Lead Ongin_Landmg Page Submission : Customer who hadLead Ongin_Landmg Page Submission, are less likely to get converted into hot leads .
3. Do Not Email :Customer who choose Do Not Email, are less likely to get converted into hot leads .

Recommendations:

The company should use a leads score threshold of 34 to identify "Hot Leads" as at this threshold, Sensitivity Score of the model is around 81% which is as good as CEO's target of 80%.