Lead Scoring Case Study

Problem Statement

- An education company called X education which sells online courses to industry professionals is suffering from poor conversion rates to their courses. So, **they want us to build a model which helps them identify their most potential leads ('Hot leads')**, so that they can be targeted easily by the Sales team to increase their conversion rate.

- Herein, **while building the model, we 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%.

Business Goal and Objectives

- 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.

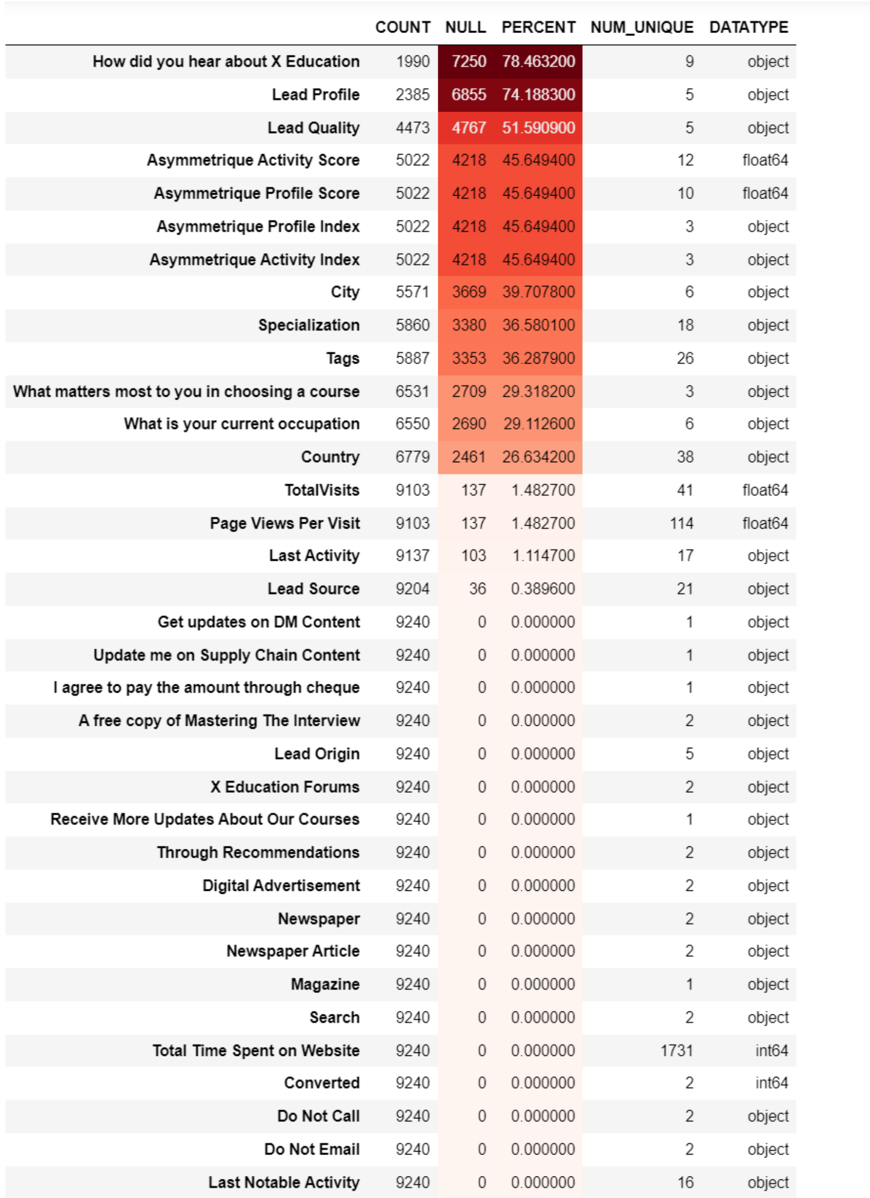
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you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

Strategy for Building Model

- Reading and understanding the data.
- Cleaning the data for null values and skewed data.
- Performing exploratory data analysis (EDA) and try to get insights from the available data.
- Performing **Logistic Regression** on the given data.
- Finding the optimal threshold point to enhance the model metrics.
- Calculating the lead scores using the predicted conversion probabilities.
- Evaluating the model by comparing the model metrics of the predictions on train and test dataset.

Data Cleaning

Removing Null values and Imputing column values

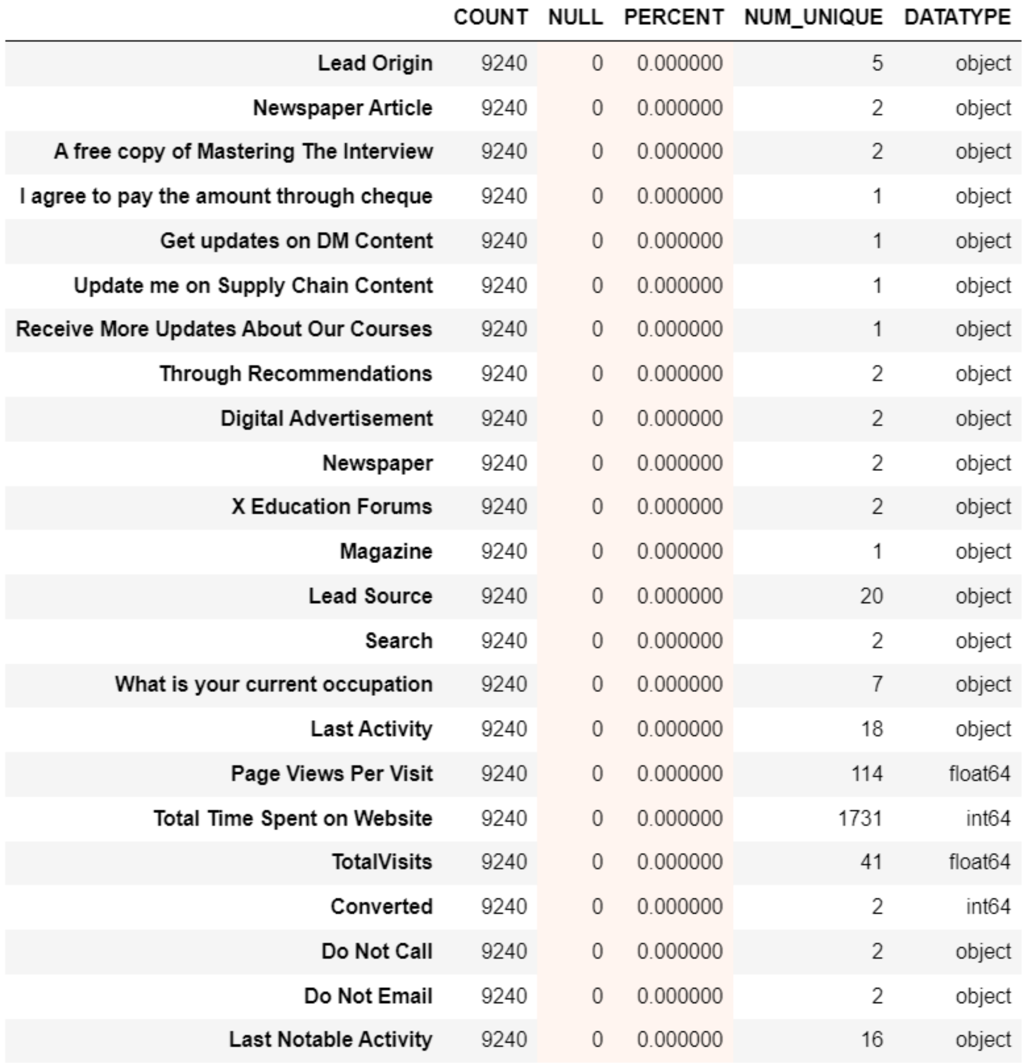


	COUNT	NULL	PERCENT	NUM_UNIQUE	DATATYPE
How did you hear about X Education	1990	7250	78.463200	9	object
Lead Profile	2385	6855	74.188300	5	object
Lead Quality	4473	4767	51.590900	5	object
Asymmetrique Activity Score	5022	4218	45.649400	12	float64
Asymmetrique Profile Score	5022	4218	45.649400	10	float64
Asymmetrique Profile Index	5022	4218	45.649400	3	object
Asymmetrique Activity Index	5022	4218	45.649400	3	object
City	5571	3669	39.707800	6	object
Specialization	5860	3380	36.580100	18	object
Tags	5887	3353	36.287900	26	object
What matters most to you in choosing a course	6531	2709	29.318200	3	object
What is your current occupation	6550	2690	29.112600	6	object
Country	6779	2461	26.634200	38	object
TotalVisits	9103	137	1.482700	41	float64
Page Views Per Visit	9103	137	1.482700	114	float64
Last Activity	9137	103	1.114700	17	object
Lead Source	9204	36	0.389600	21	object
Get updates on DM Content	9240	0	0.000000	1	object
Update me on Supply Chain Content	9240	0	0.000000	1	object
I agree to pay the amount through cheque	9240	0	0.000000	1	object
A free copy of Mastering The Interview	9240	0	0.000000	2	object
Lead Origin	9240	0	0.000000	5	object
X Education Forums	9240	0	0.000000	2	object
Receive More Updates About Our Courses	9240	0	0.000000	1	object
Through Recommendations	9240	0	0.000000	2	object
Digital Advertisement	9240	0	0.000000	2	object
Newspaper	9240	0	0.000000	2	object
Newspaper Article	9240	0	0.000000	2	object
Magazine	9240	0	0.000000	1	object
Search	9240	0	0.000000	2	object
Total Time Spent on Website	9240	0	0.000000	1731	int64
Converted	9240	0	0.000000	2	int64
Do Not Call	9240	0	0.000000	2	object
Do Not Email	9240	0	0.000000	2	object
Last Notable Activity	9240	0	0.000000	16	object

From the chart we can see that these columns have highest number of null percentage:

- Specialization
- How did you hear about X Education
- Tags
- Lead Quality
- Lead Profile
- City
- Asymmetrique Activity Index
- Asymmetrique Profile Index
- Asymmetrique Activity Score
- Asymmetrique Profile Score

The common trend is that these columns have null percentage more than 35%. Hence **we will drop the columns with null percentage more than 35%.**

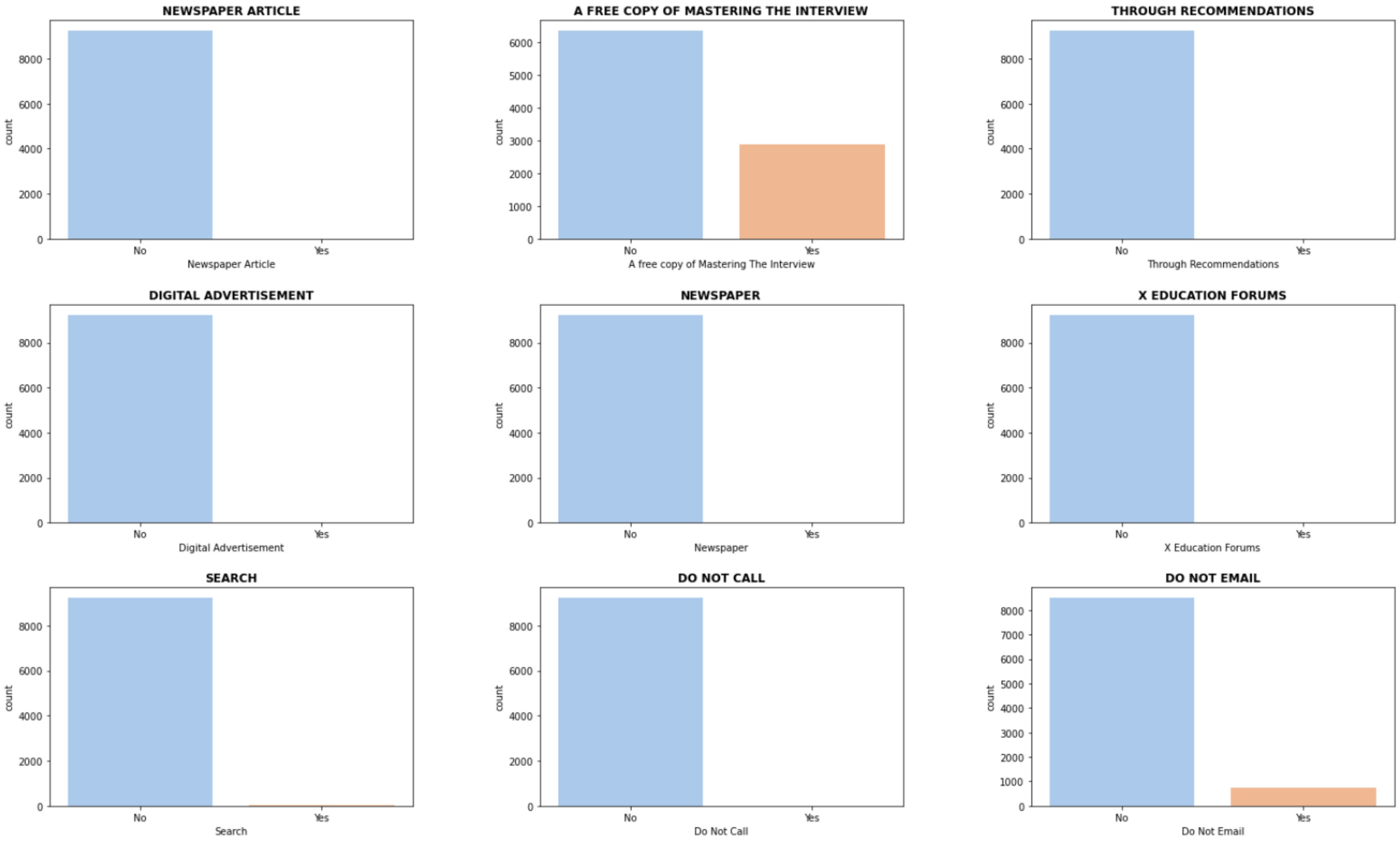


	COUNT	NULL	PERCENT	NUM_UNIQUE	DATATYPE
Lead Origin	9240	0	0.000000	5	object
Newspaper Article	9240	0	0.000000	2	object
A free copy of Mastering The Interview	9240	0	0.000000	2	object
I agree to pay the amount through cheque	9240	0	0.000000	1	object
Get updates on DM Content	9240	0	0.000000	1	object
Update me on Supply Chain Content	9240	0	0.000000	1	object
Receive More Updates About Our Courses	9240	0	0.000000	1	object
Through Recommendations	9240	0	0.000000	2	object
Digital Advertisement	9240	0	0.000000	2	object
Newspaper	9240	0	0.000000	2	object
X Education Forums	9240	0	0.000000	2	object
Magazine	9240	0	0.000000	1	object
Lead Source	9240	0	0.000000	20	object
Search	9240	0	0.000000	2	object
What is your current occupation	9240	0	0.000000	7	object
Last Activity	9240	0	0.000000	18	object
Page Views Per Visit	9240	0	0.000000	114	float64
Total Time Spent on Website	9240	0	0.000000	1731	int64
TotalVisits	9240	0	0.000000	41	float64
Converted	9240	0	0.000000	2	int64
Do Not Call	9240	0	0.000000	2	object
Do Not Email	9240	0	0.000000	2	object
Last Notable Activity	9240	0	0.000000	16	object

- Post imputation and data cleaning, here the stats which tells columns with no NULL values.

- **All the null values in the columns now have either been imputed or we have dropped the columns which have more than 70% data concentrated towards one value.**

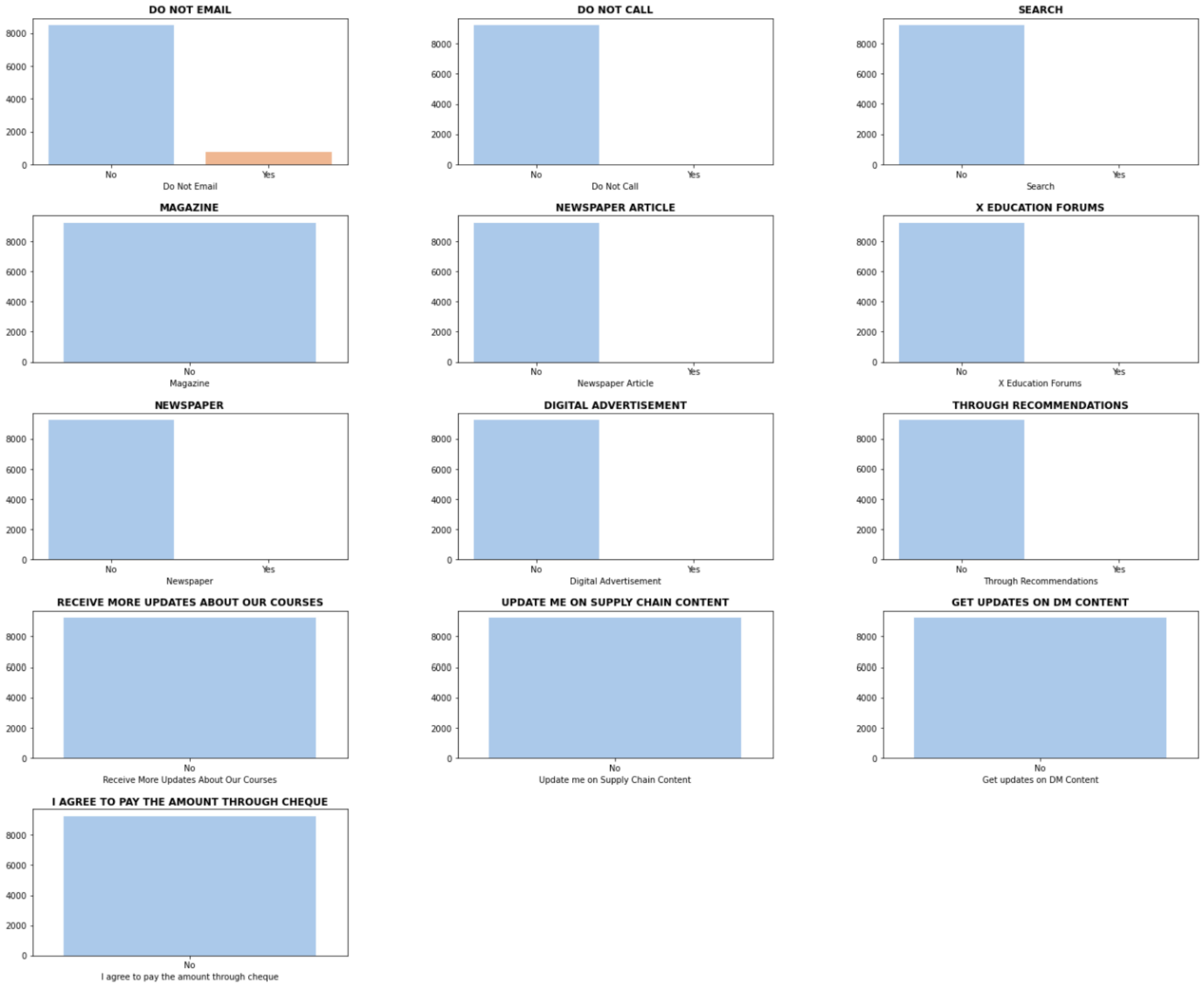
Handling the columns with highly skewed data



- From these plots, we can see that nearly all the columns except A free copy of Mastering The Interview has more than 90% values as 'No'. These columns will not be helpful in our analysis and hence we need to remove them.

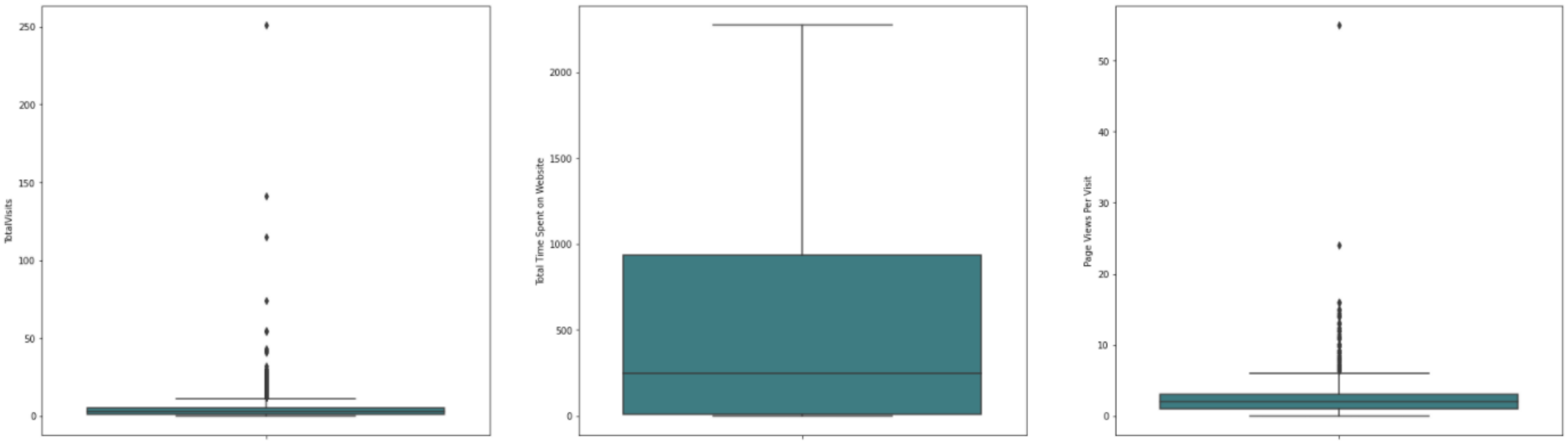
- The following columns are highly skewed:

1. Do Not Email
2. Do Not Call
3. Search
4. Magazine
5. Newspaper Article
6. X Education Forums
7. Newspaper
8. Digital Advertisement
9. Through Recommendations
10. Receive More Updates About Our Courses
11. Update me on Supply Chain Content
12. Get updates on DM Content
13. I agree to pay the amount through cheque



- Plotting the distribution of all columns which are highly skewed
- From these plots, we can see that these variables have nearly 100% of their values in a single category.

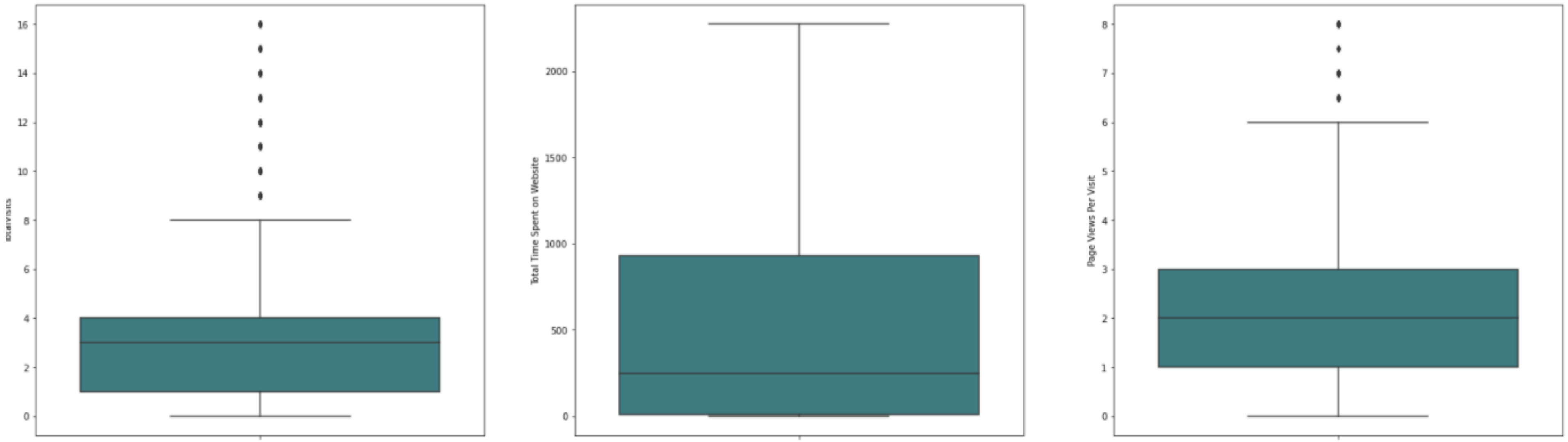
Numerical Columns and their Outliers Treatment



- It can be seen that outlier exists in the columns TotalVisits and Page Views Per Visit columns.
- We will perform outlier treatment on these numerical columns –

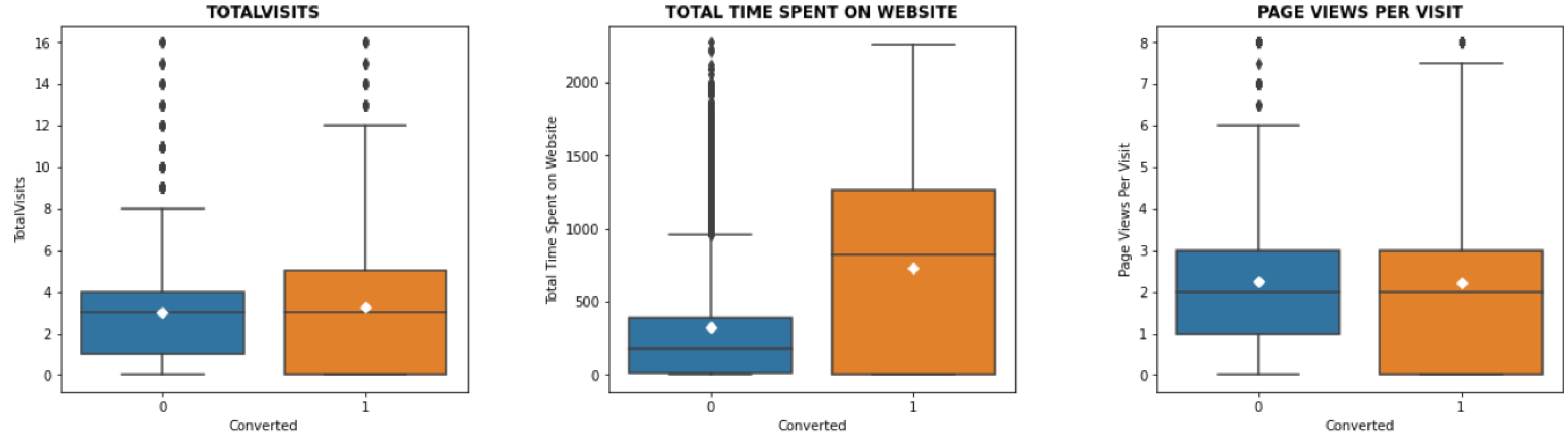
1. TotalVisits
2. Total Time Spent on Website
3. Page Views Per Visit

Numerical Columns post Outlier Treatment



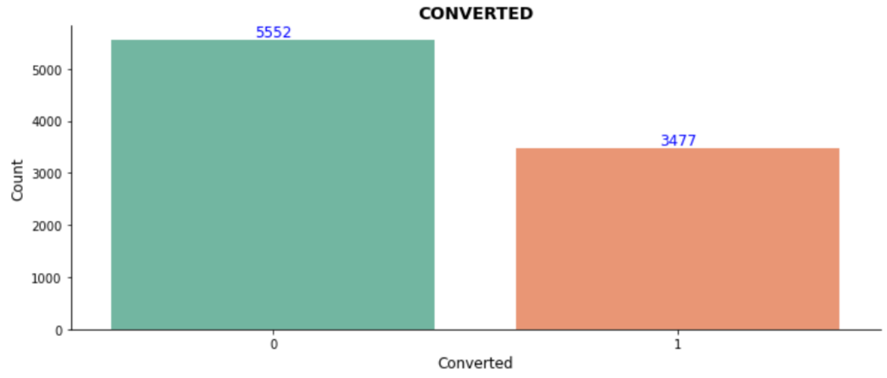
All the outliers have been removed now.

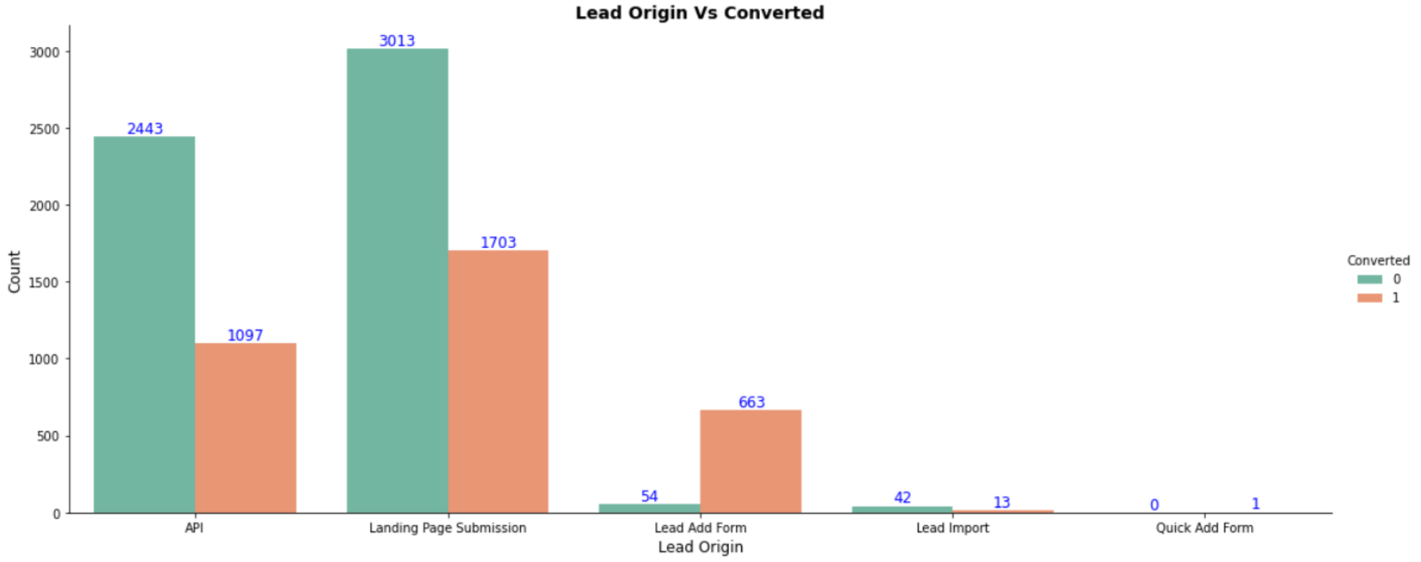
Comparing numerical variables with that of 'Converted'

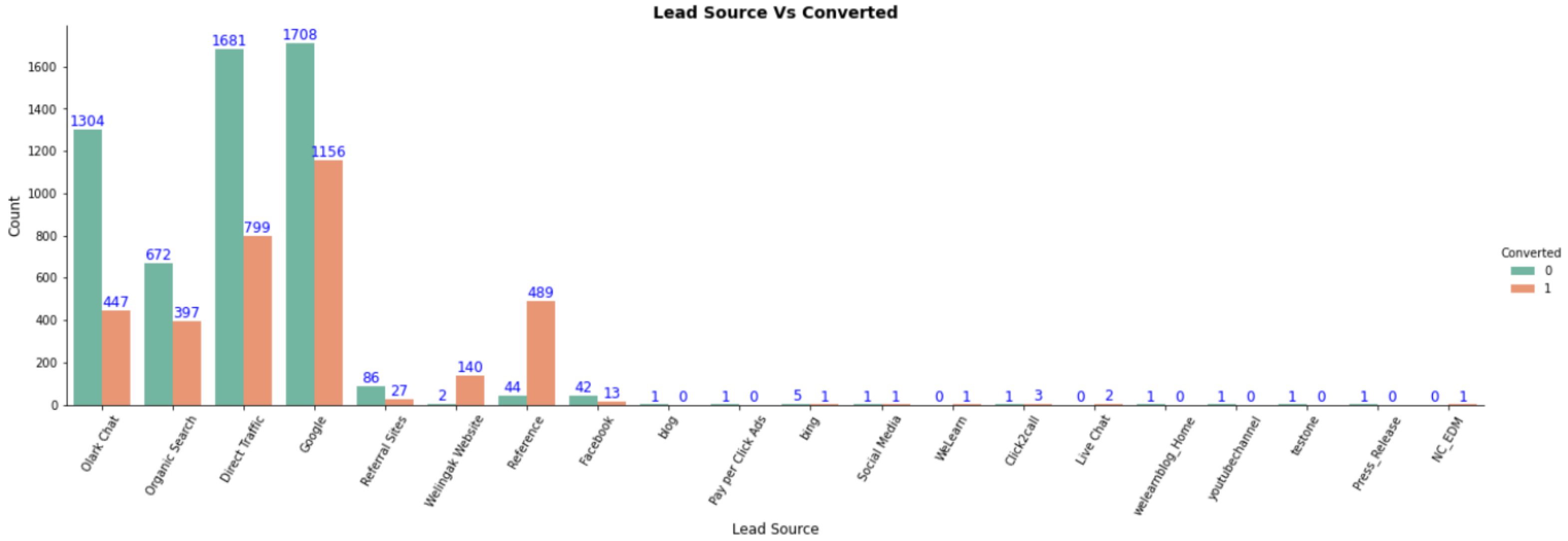


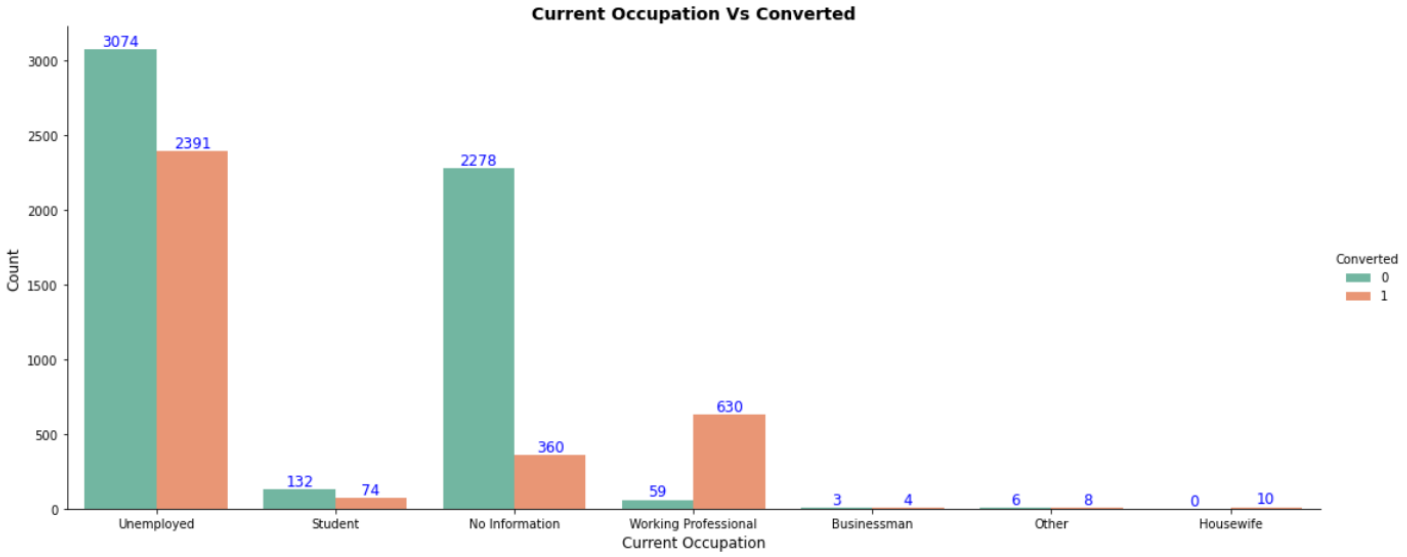
- From the plot we can see that the avg value of TotalVisits and PageViewsPerVisit remains almost the same for both converted and non converted

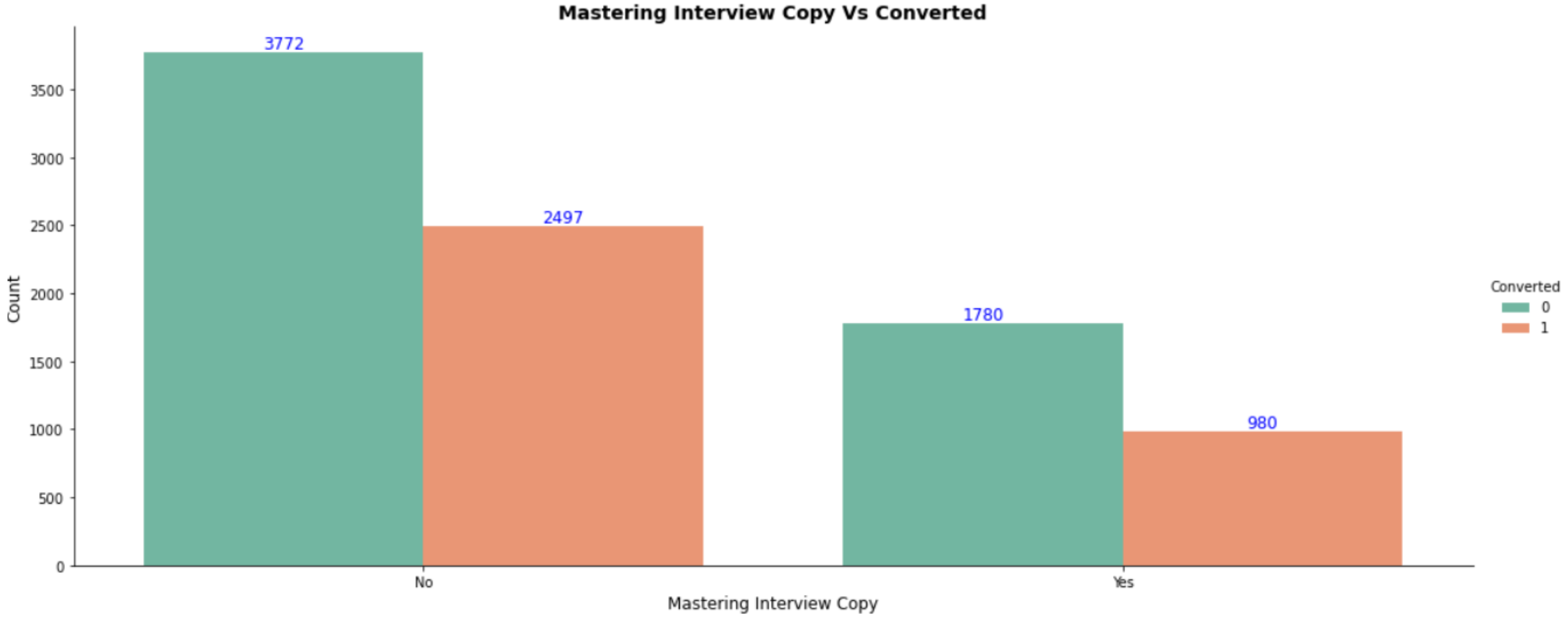
Categorical Columns and their Analysis

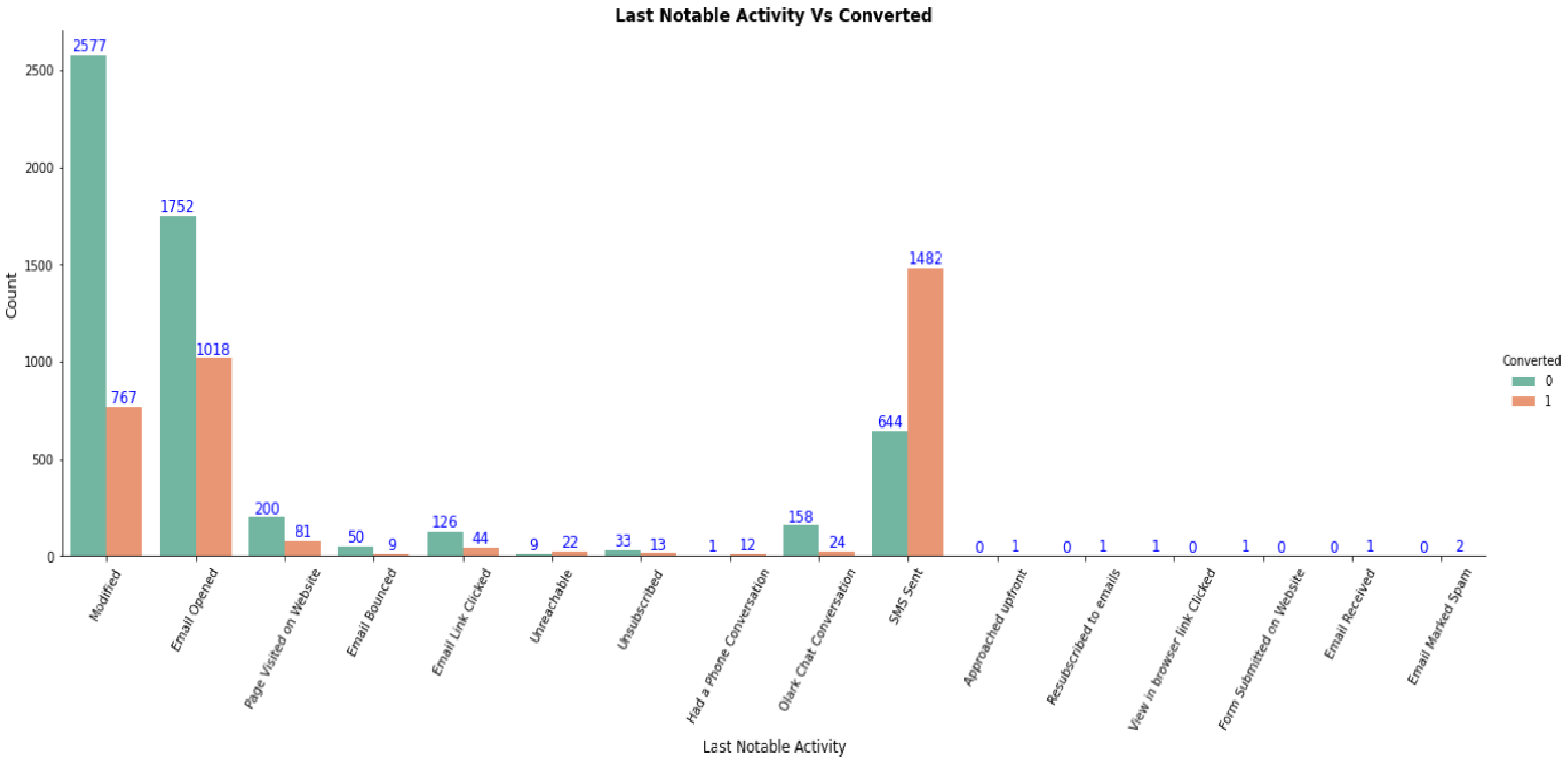






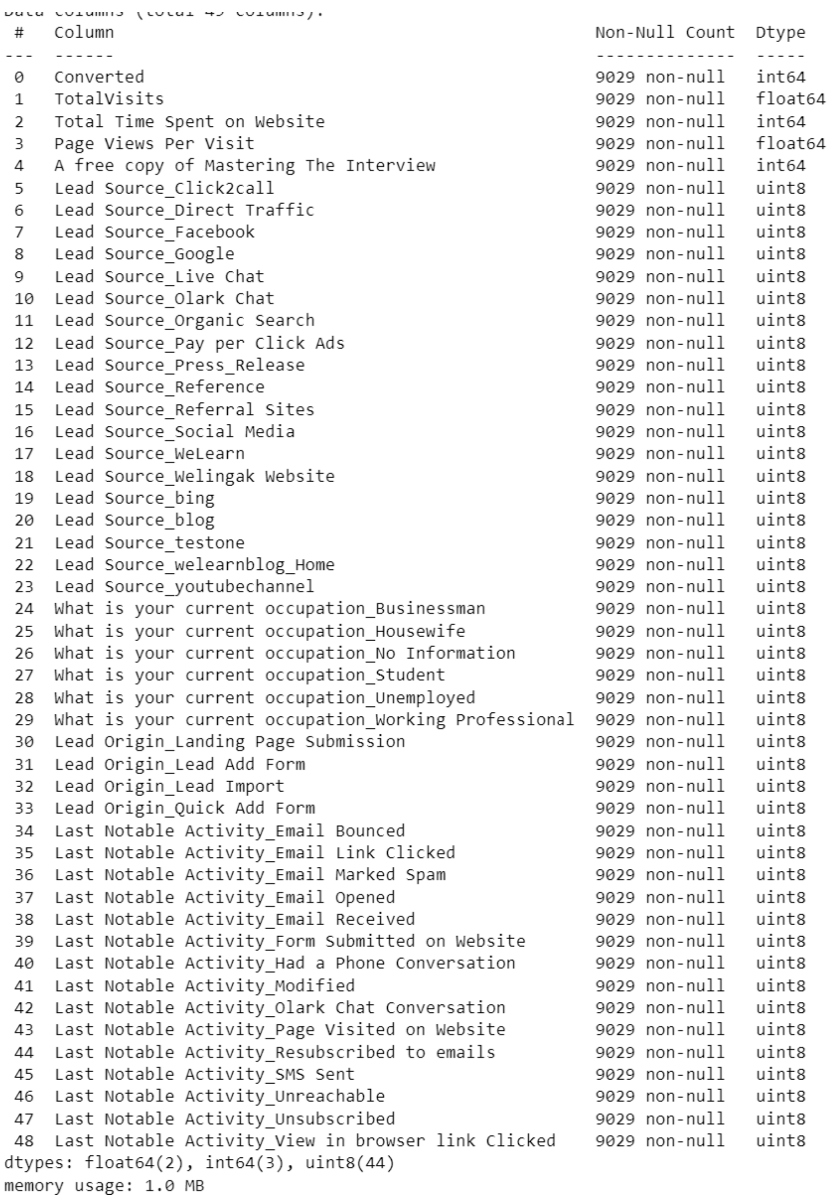

Current Occupation Vs Converted


Mastering Interview Copy Vs Converted


Last Notable Activity Vs Converted
Count
0
500
1000
1500
2000
2500
2577
767
1752
1018
200
81
50
9
126
44
9
22
33
13
1
12
158
24
644
1482
0
1
0
1
1
0
1
0
0
1
0
2
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Approached upfront
Resubscribed to emails
View in browser link Clicked
Form Submitted on Website
Email Received
Email Marked Spam
Last Notable Activity
Converted
0
1

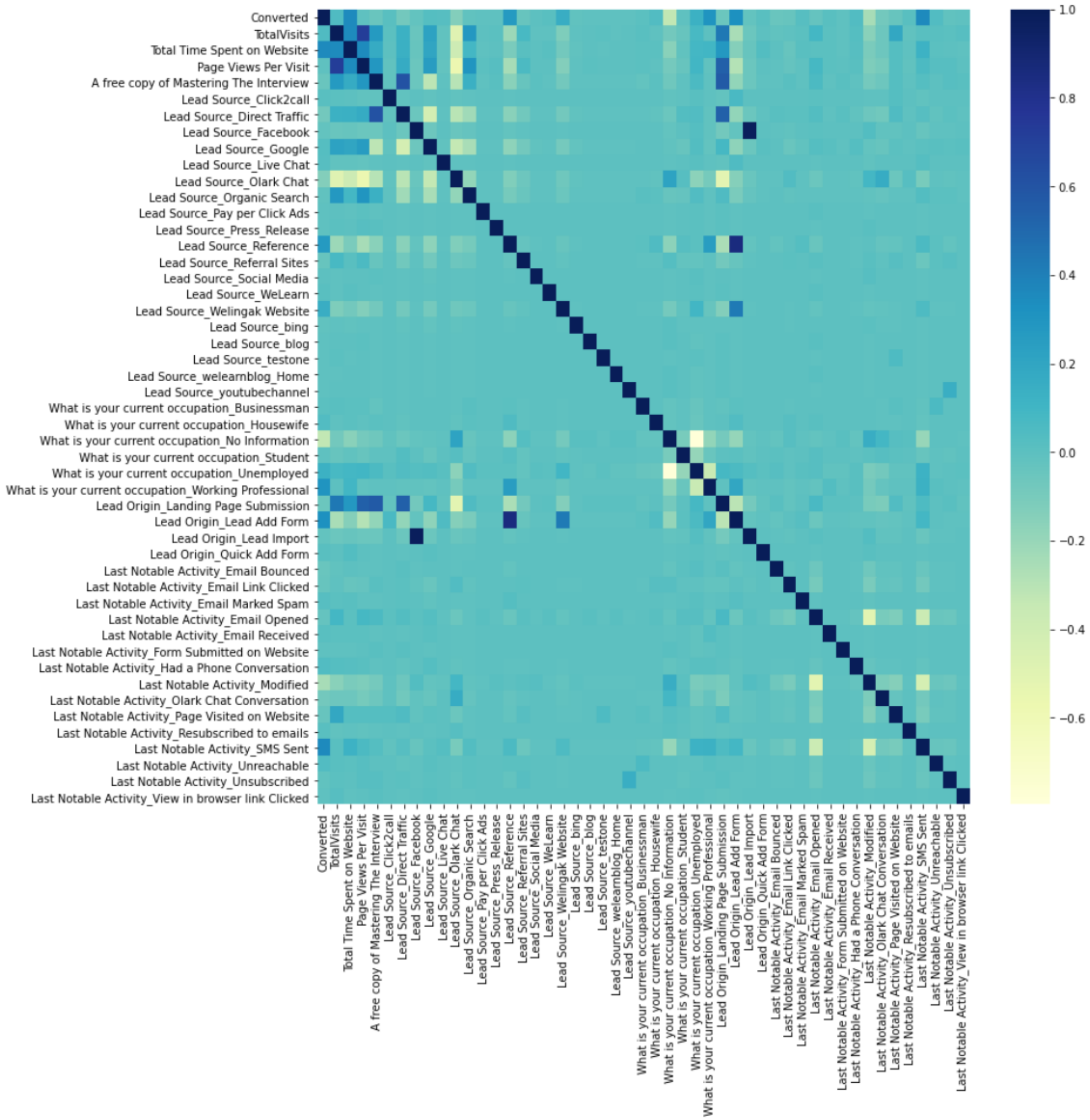
Data Preparation

Columns after creating dummies



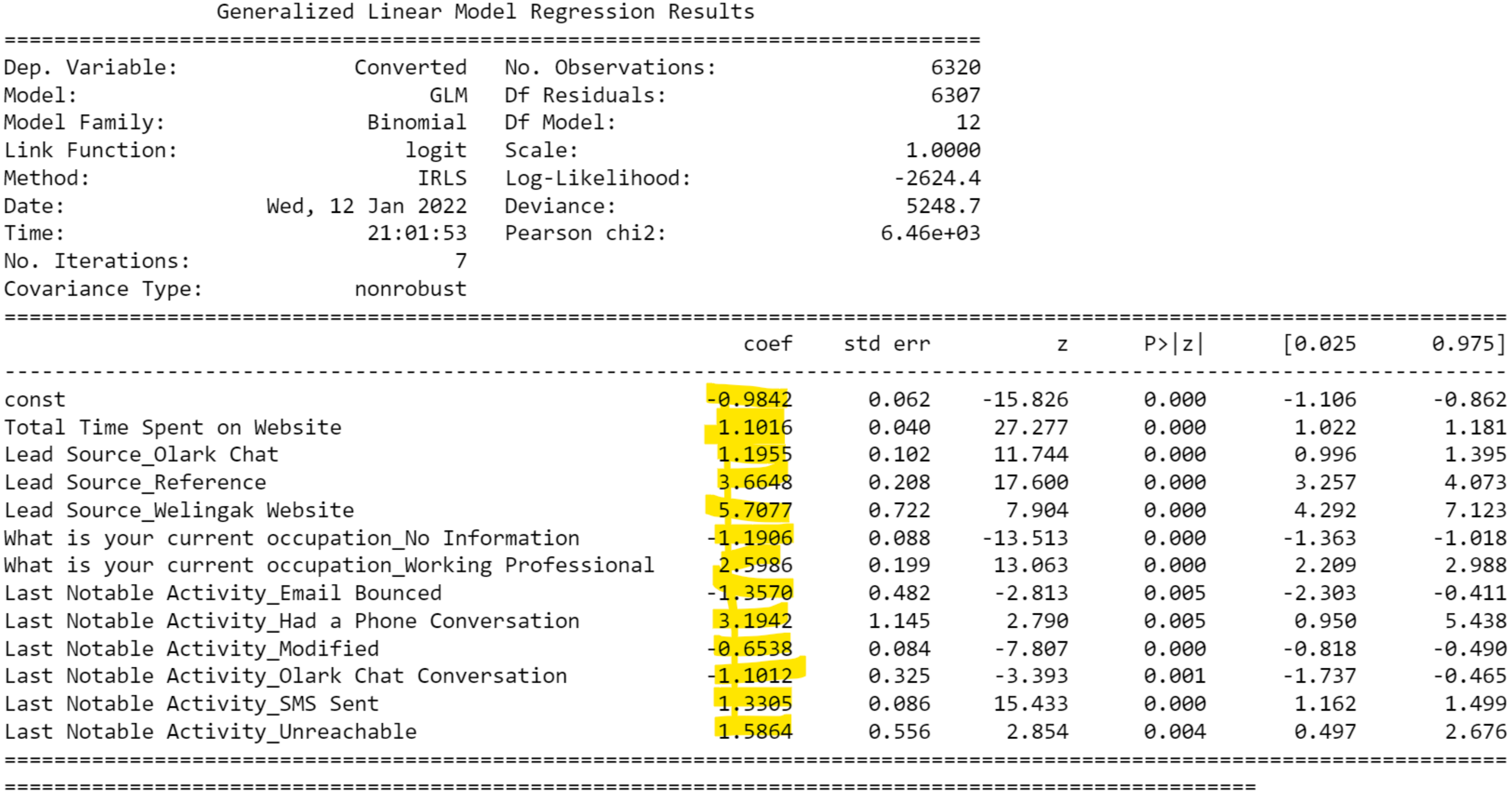
```
Data columns (total 49 columns):
 #   Column                                                Non-Null Count  Dtype  
---  ------                                                --------------  -----  
 0   Converted                                             9029 non-null   int64  
 1   TotalVisits                                           9029 non-null   float64
 2   Total Time Spent on Website                           9029 non-null   int64  
 3   Page Views Per Visit                                  9029 non-null   float64
 4   A free copy of Mastering The Interview                9029 non-null   int64  
 5   Lead Source_Click2call                                9029 non-null   uint8  
 6   Lead Source_Direct Traffic                            9029 non-null   uint8  
 7   Lead Source_Facebook                                  9029 non-null   uint8  
 8   Lead Source_Google                                    9029 non-null   uint8  
 9   Lead Source_Live Chat                                 9029 non-null   uint8  
 10  Lead Source_Olark Chat                                9029 non-null   uint8  
 11  Lead Source_Organic Search                            9029 non-null   uint8  
 12  Lead Source_Pay per Click Ads                         9029 non-null   uint8  
 13  Lead Source_Press_Release                             9029 non-null   uint8  
 14  Lead Source_Reference                                 9029 non-null   uint8  
 15  Lead Source_Referral Sites                            9029 non-null   uint8  
 16  Lead Source_Social Media                              9029 non-null   uint8  
 17  Lead Source_WeLearn                                   9029 non-null   uint8  
 18  Lead Source_Welingak Website                          9029 non-null   uint8  
 19  Lead Source_bing                                      9029 non-null   uint8  
 20  Lead Source_blog                                      9029 non-null   uint8  
 21  Lead Source_testone                                   9029 non-null   uint8  
 22  Lead Source_welearnblog_Home                          9029 non-null   uint8  
 23  Lead Source_youtubechannel                            9029 non-null   uint8  
 24  What is your current occupation_Businessman           9029 non-null   uint8  
 25  What is your current occupation_Housewife             9029 non-null   uint8  
 26  What is your current occupation_No Information        9029 non-null   uint8  
 27  What is your current occupation_Student               9029 non-null   uint8  
 28  What is your current occupation_Unemployed            9029 non-null   uint8  
 29  What is your current occupation_Working Professional  9029 non-null   uint8  
 30  Lead Origin_Landing Page Submission                   9029 non-null   uint8  
 31  Lead Origin_Lead Add Form                             9029 non-null   uint8  
 32  Lead Origin_Lead Import                               9029 non-null   uint8  
 33  Lead Origin_Quick Add Form                            9029 non-null   uint8  
 34  Last Notable Activity_Email Bounced                   9029 non-null   uint8  
 35  Last Notable Activity_Email Link Clicked              9029 non-null   uint8  
 36  Last Notable Activity_Email Marked Spam               9029 non-null   uint8  
 37  Last Notable Activity_Email Opened                    9029 non-null   uint8  
 38  Last Notable Activity_Email Received                  9029 non-null   uint8  
 39  Last Notable Activity_Form Submitted on Website       9029 non-null   uint8  
 40  Last Notable Activity_Had a Phone Conversation        9029 non-null   uint8  
 41  Last Notable Activity_Modified                        9029 non-null   uint8  
 42  Last Notable Activity_Olark Chat Conversation         9029 non-null   uint8  
 43  Last Notable Activity_Page Visited on Website         9029 non-null   uint8  
 44  Last Notable Activity_Resubscribed to emails          9029 non-null   uint8  
 45  Last Notable Activity_SMS Sent                        9029 non-null   uint8  
 46  Last Notable Activity_Unreachable                     9029 non-null   uint8  
 47  Last Notable Activity_Unsubscribed                    9029 non-null   uint8  
 48  Last Notable Activity_View in browser link Clicked    9029 non-null   uint8  
dtypes: float64(2), int64(3), uint8(44)
memory usage: 1.0 MB
```

Finding the correlation of all the variables



Final Model

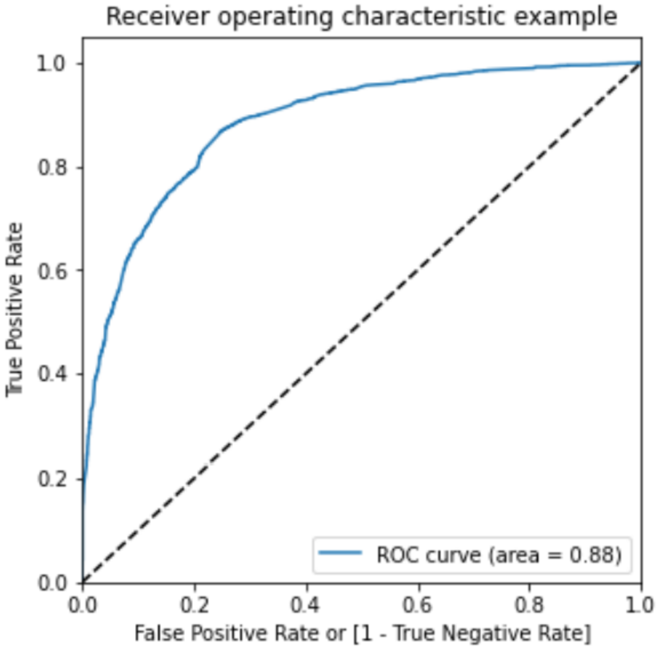
Finding the correlation of all the variables



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6320
Model:                            GLM   Df Residuals:                     6307
Model Family:                 Binomial   Df Model:                           12
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2624.4
Date:                Wed, 12 Jan 2022   Deviance:                       5248.7
Time:                        21:01:53   Pearson chi2:                 6.46e+03
No. Iterations:                     7
Covariance Type:            nonrobust
========================================================================================================================
                                                           coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------------
const                                               -0.9842      0.062    -15.826      0.000      -1.106      -0.862
Total Time Spent on Website                          1.1016      0.040     27.277      0.000       1.022       1.181
Lead Source_Olark Chat                               1.1955      0.102     11.744      0.000       0.996       1.395
Lead Source_Reference                                3.6648      0.208     17.600      0.000       3.257       4.073
Lead Source_Welingak Website                         5.7077      0.722      7.904      0.000       4.292       7.123
What is your current occupation_No Information      -1.1906      0.088    -13.513      0.000      -1.363      -1.018
What is your current occupation_Working Professional 2.5986      0.199     13.063      0.000       2.209       2.988
Last Notable Activity_Email Bounced                 -1.3570      0.482     -2.813      0.005      -2.303      -0.411
Last Notable Activity_Had a Phone Conversation       3.1942      1.145      2.790      0.005       0.950       5.438
Last Notable Activity_Modified                      -0.6538      0.084     -7.807      0.000      -0.818      -0.490
Last Notable Activity_Olark Chat Conversation       -1.1012      0.325     -3.393      0.001      -1.737      -0.465
Last Notable Activity_SMS Sent                       1.3305      0.086     15.433      0.000       1.162       1.499
Last Notable Activity_Unreachable                    1.5864      0.556      2.854      0.004       0.497       2.676
========================================================================================================================
====================================================================================================
```

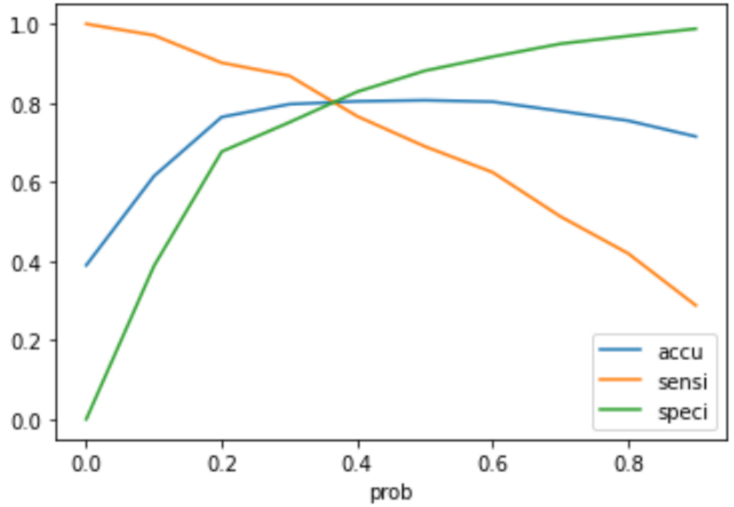
The final model consists of the above 13 variables with all the p values less than 0.05 and the VIF less than 2.

ROC Curve



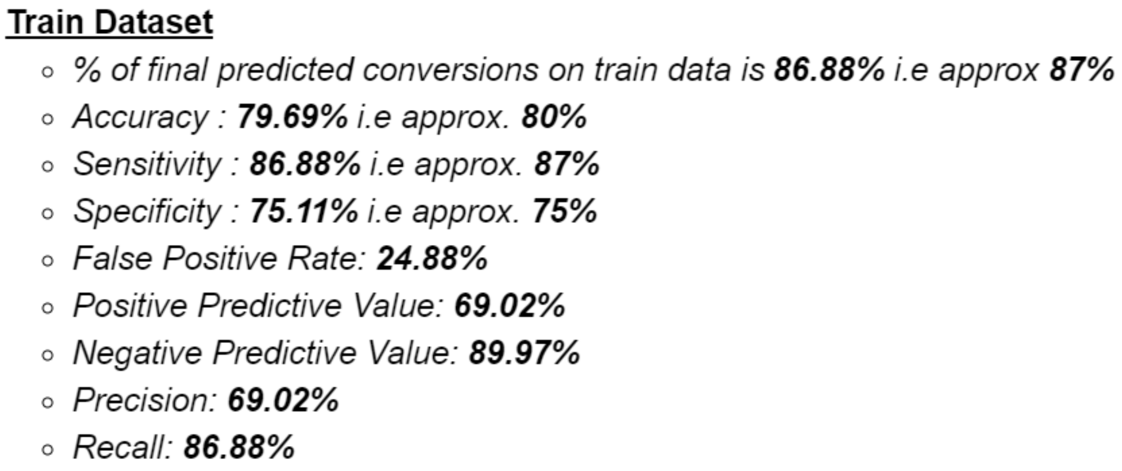
Area under the curve is 0.88
and the curve is hugging towards y
axis, which indicates the model is
performing good.

Optimal Cutoff



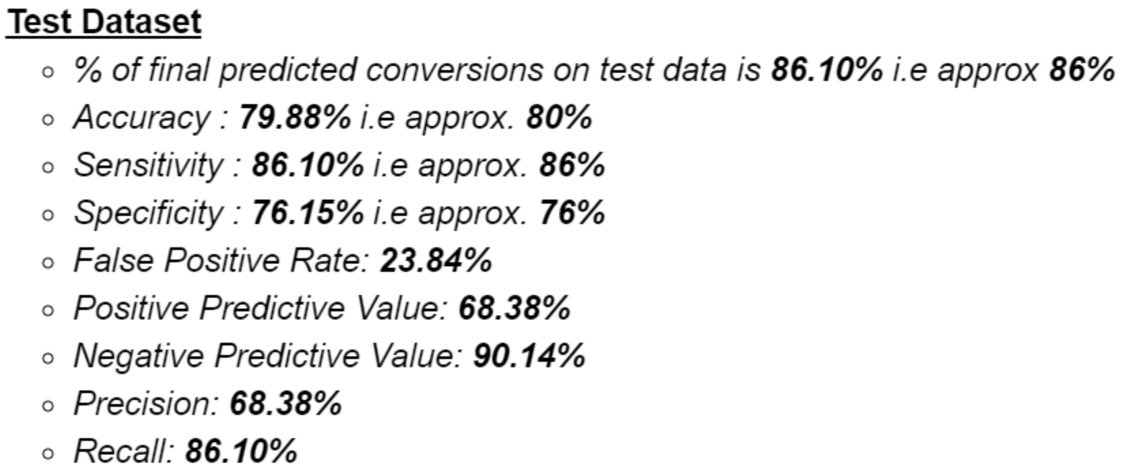
We'll take our optimal cutoff point
as 0.3, as that's the value where
all the above parameters coincide.

Comparing the model metrics



Train Dataset

- *% of final predicted conversions on train data is* ***86.88%*** *i.e approx* ***87%***
- *Accuracy :* ***79.69%*** *i.e approx.* ***80%***
- *Sensitivity :* ***86.88%*** *i.e approx.* ***87%***
- *Specificity :* ***75.11%*** *i.e approx.* ***75%***
- *False Positive Rate:* ***24.88%***
- *Positive Predictive Value:* ***69.02%***
- *Negative Predictive Value:* ***89.97%***
- *Precision:* ***69.02%***
- *Recall:* ***86.88%***

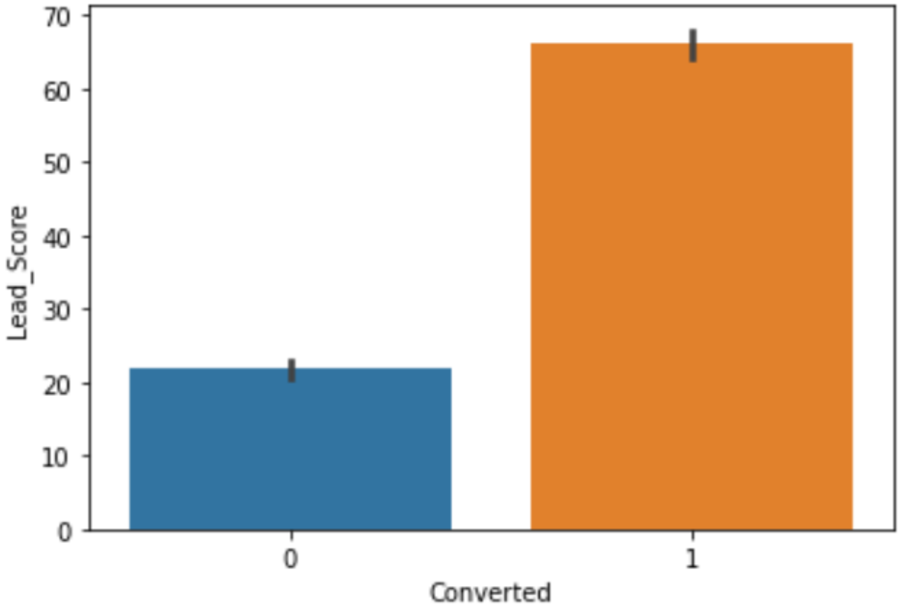


Test Dataset

- *% of final predicted conversions on test data is* ***86.10%*** *i.e approx* ***86%***
- *Accuracy :* ***79.88%*** *i.e approx.* ***80%***
- *Sensitivity :* ***86.10%*** *i.e approx.* ***86%***
- *Specificity :* ***76.15%*** *i.e approx.* ***76%***
- *False Positive Rate:* ***23.84%***
- *Positive Predictive Value:* ***68.38%***
- *Negative Predictive Value:* ***90.14%***
- *Precision:* ***68.38%***
- *Recall:* ***86.10%***

- Our model has a sensitivity of around 0.86 which shows it is able to correctly predict 86% of the converted leads.
- The precision of our model is around 0.69 which shows that the 69% of the leads predicted by the model are truly converted leads.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model as around **87%**

Comparing Lead Scores



From the above plot, we can see that the average Lead Score of the converted is around 60 and that of not converted is around 20. **So, the sales team can focus on leads with Lead Score of around 60 to improve their conversion rate.** i.e., leads with lead score above 60 can be hot leads.