LEAD SCORING CASE STUDY- DSC 61

SUBMITTED BY

VENKATESH G

GUNDLAPALLY SAHITHI

CHARISHMA GONTU

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. They have process of form filling on their website after which the company that individual as a lead.
- 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, this means 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.

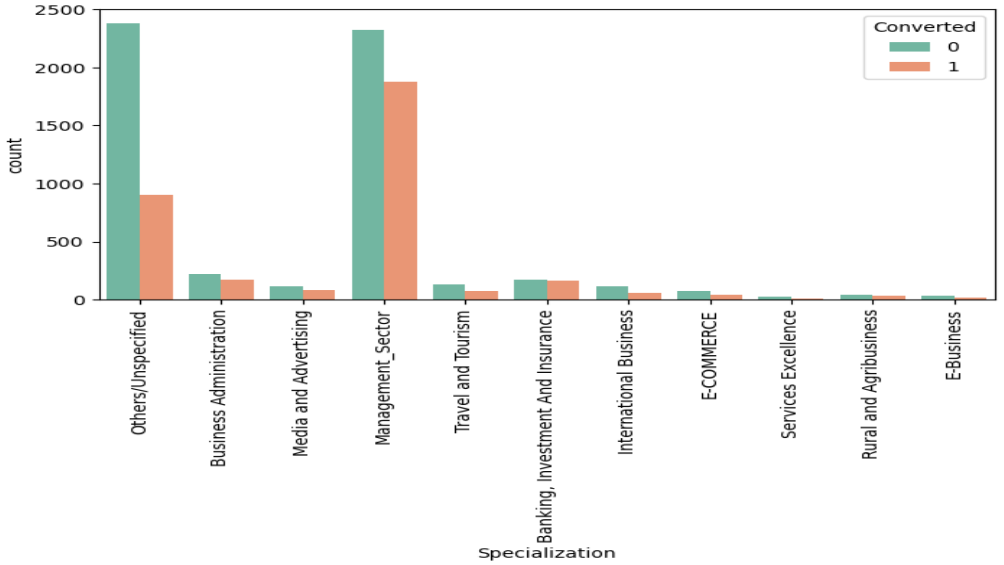
OBJECTIVES

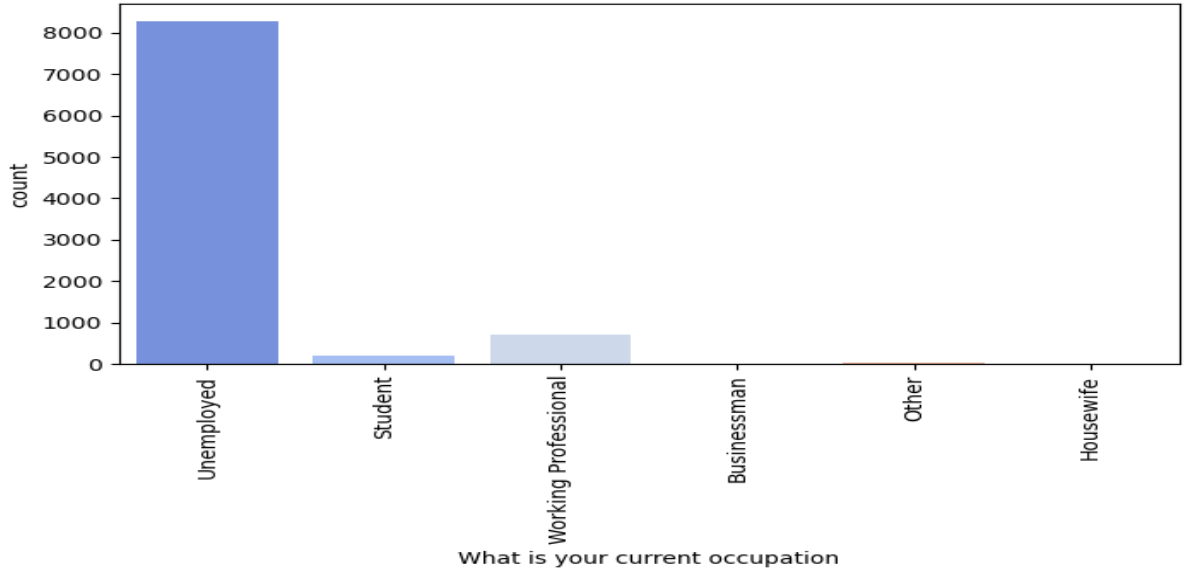
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will 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.
- CEO wants to achieve the lead conversion rate of 80%.
- They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches.

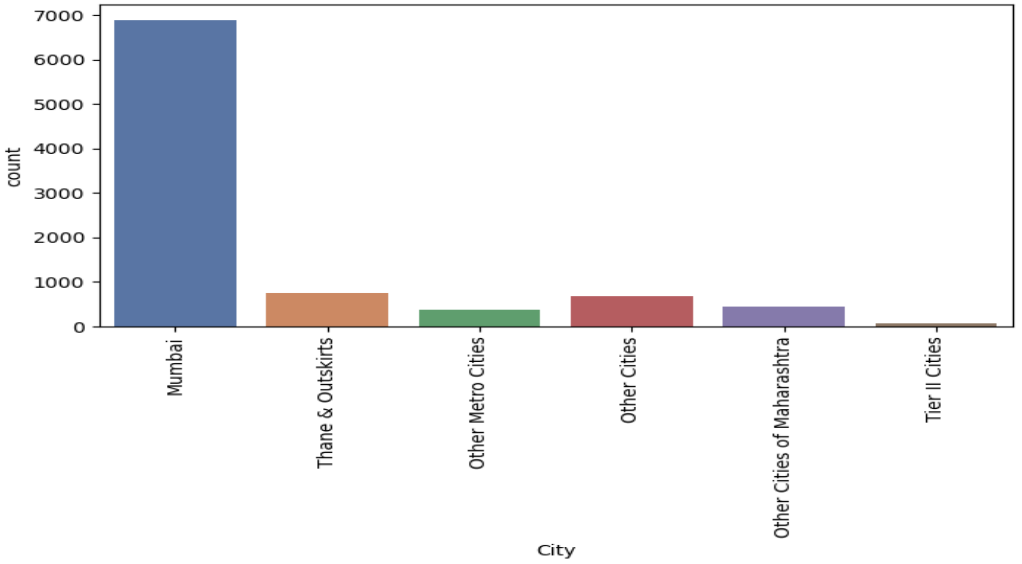
PROBLEM APPROACH

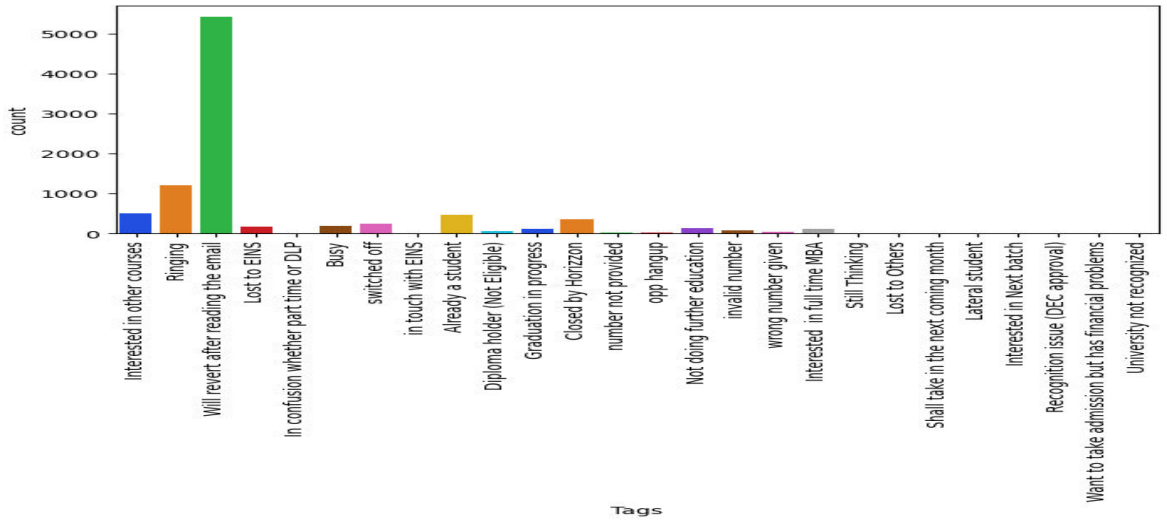
- Reading and loading the data
- Data cleaning – Dropping of missing value variables >40%
- Missing value imputation
- Outlier treatment.
- Feature engineering- Mapping and Creation of dumm
- Train-Test split
- Scaling
- Model Building – Model with stabilised p-value(<0.05) and VIF(<5)
- Predictions on train set , Evaluation metrics like accuracy, sensitivity,specificity
- Plotting ROC Curve and trade-off curve between precision and recall.
- Making predictions on Test dataset.

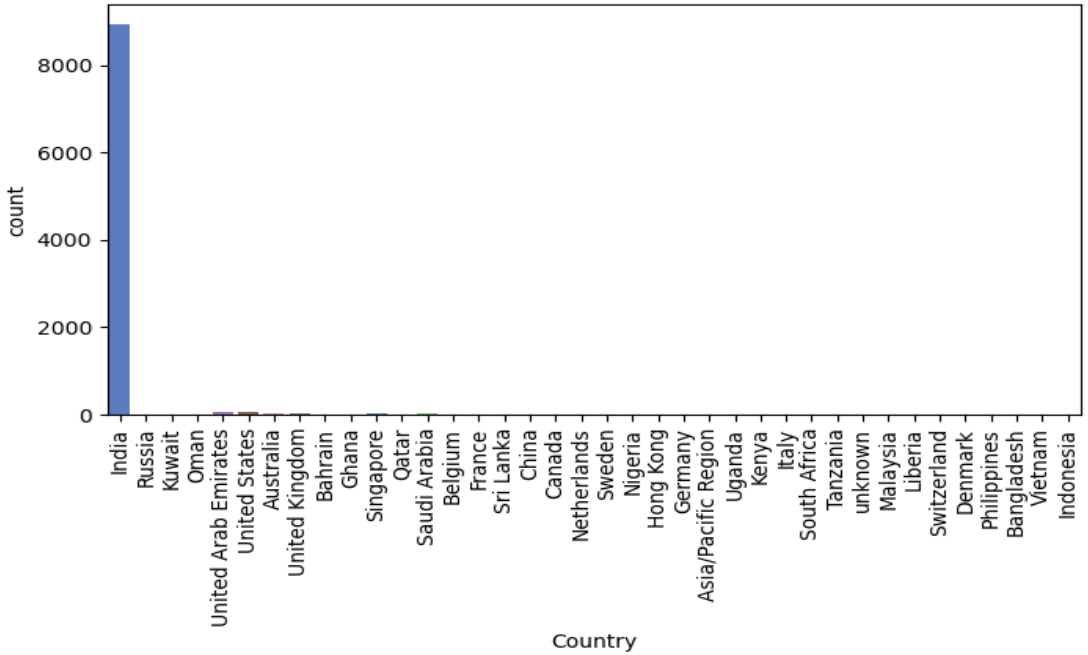
IMPUTED AND MODIFIED VARIABLES

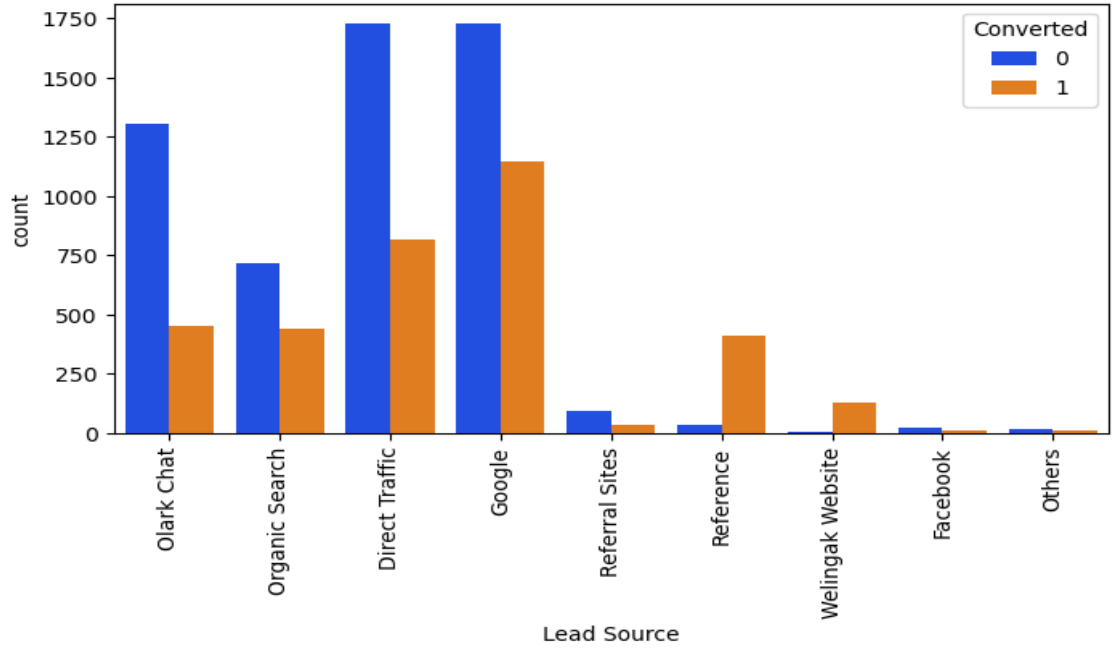


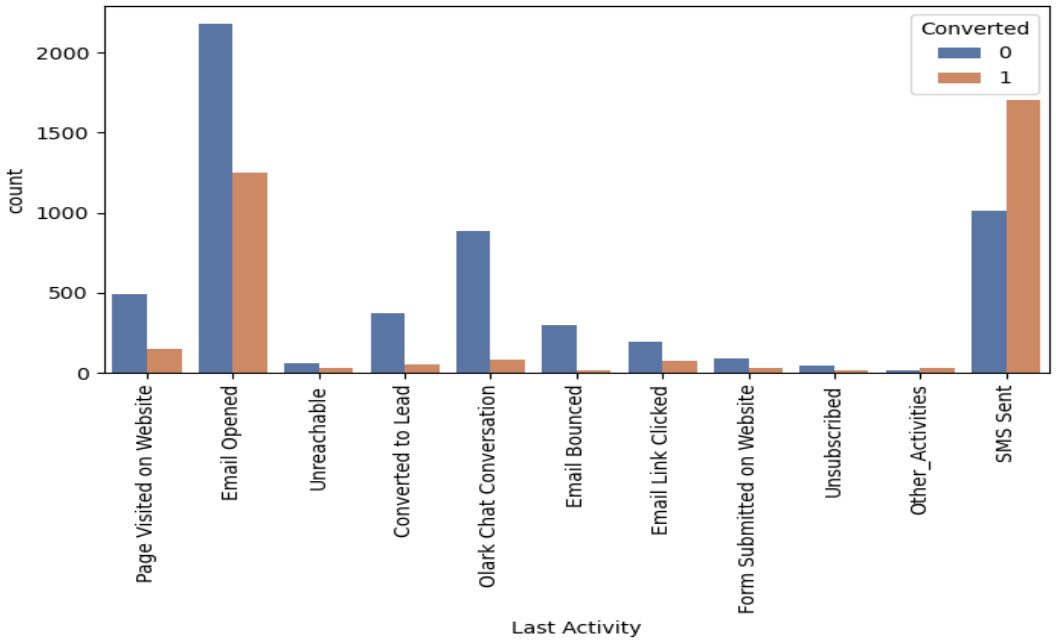


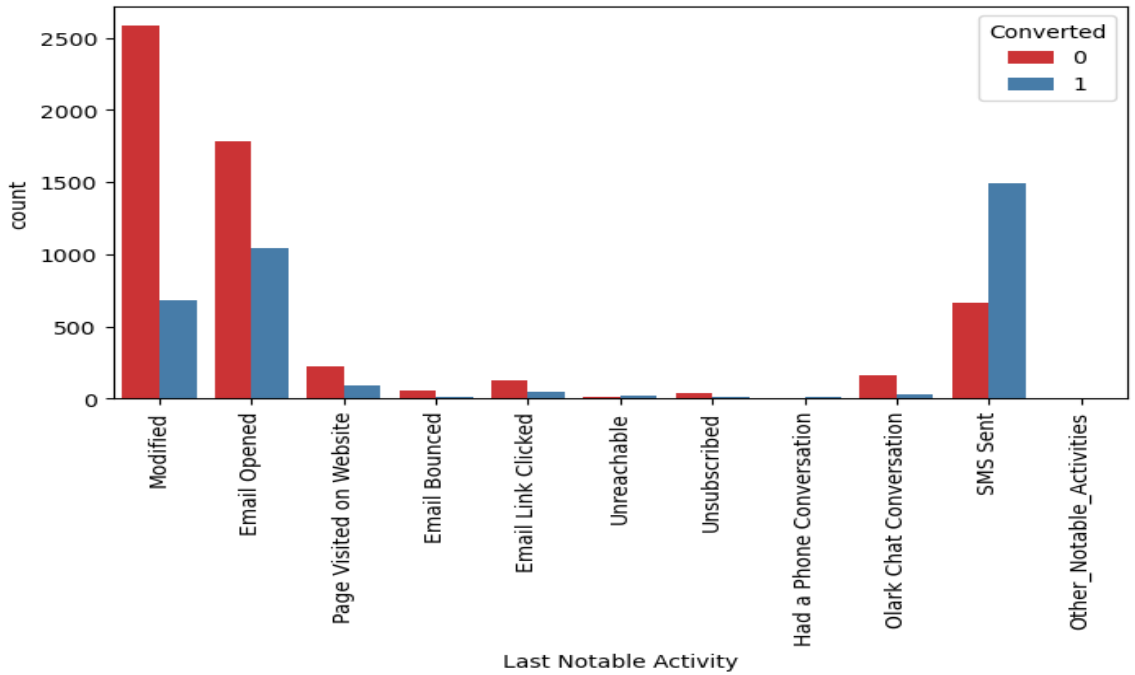




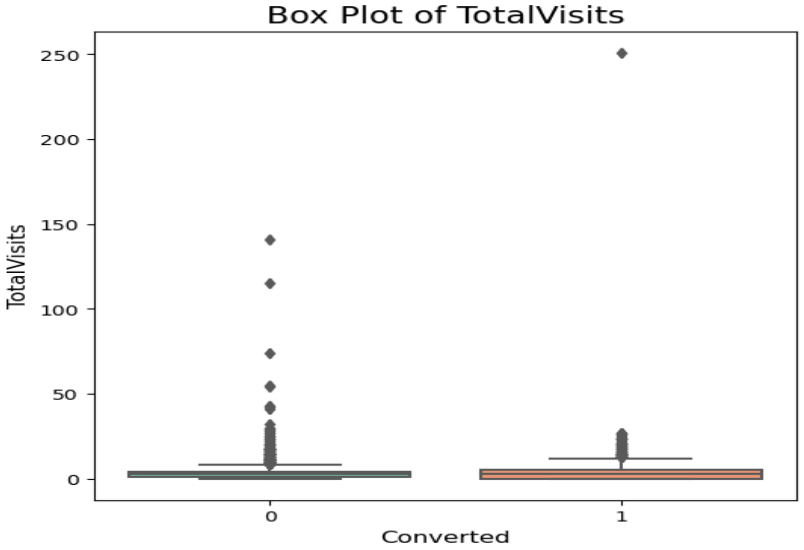

count
8000
6000
4000
2000
0
India
Russia
Kuwait
Oman
United Arab Emirates
United States
Australia
United Kingdom
Bahrain
Ghana
Singapore
Qatar
Saudi Arabia
Belgium
France
Sri Lanka
China
Canada
Netherlands
Sweden
Nigeria
Hong Kong
Germany
Asia/Pacific Region
Uganda
Kenya
Italy
South Africa
Tanzania
unknown
Malaysia
Liberia
Switzerland
Denmark
Philippines
Bangladesh
Vietnam
Indonesia
Country

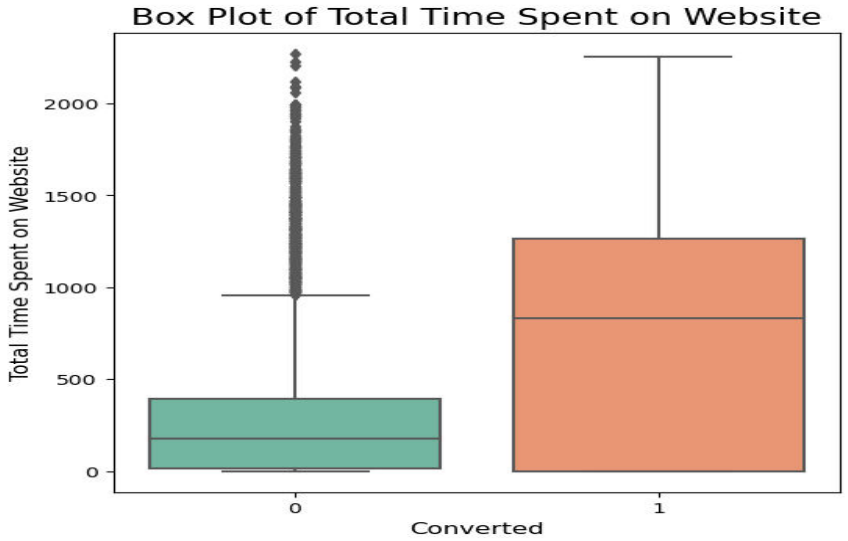

Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
Welingak Website
Facebook
Others
Lead Source

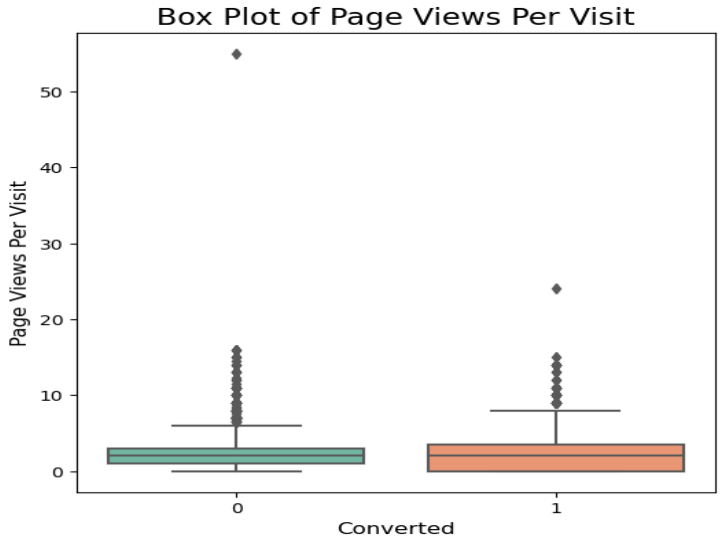

Converted
0
1
count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Other_Activities
SMS Sent
Last Activity

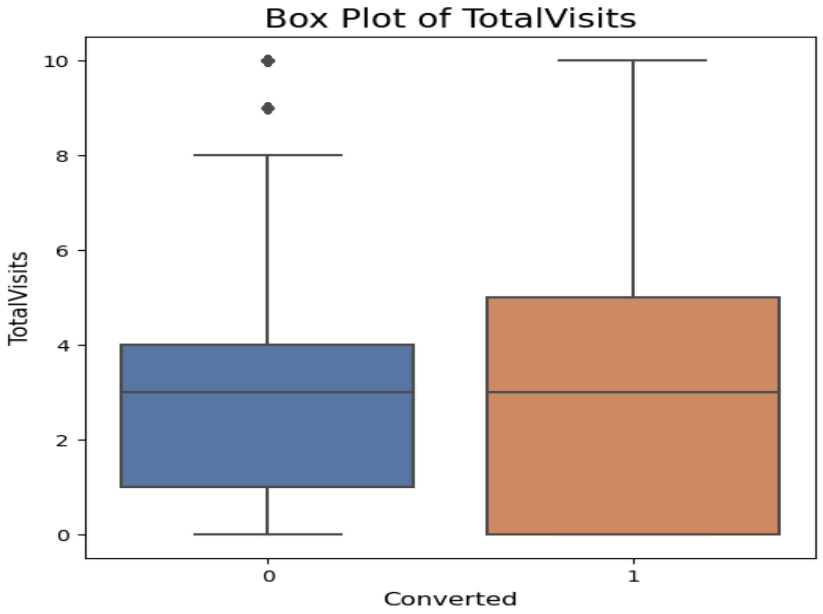

Converted
0
1
count
2500
2000
1500
1000
500
0
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Other_Notable_Activities
Last Notable Activity

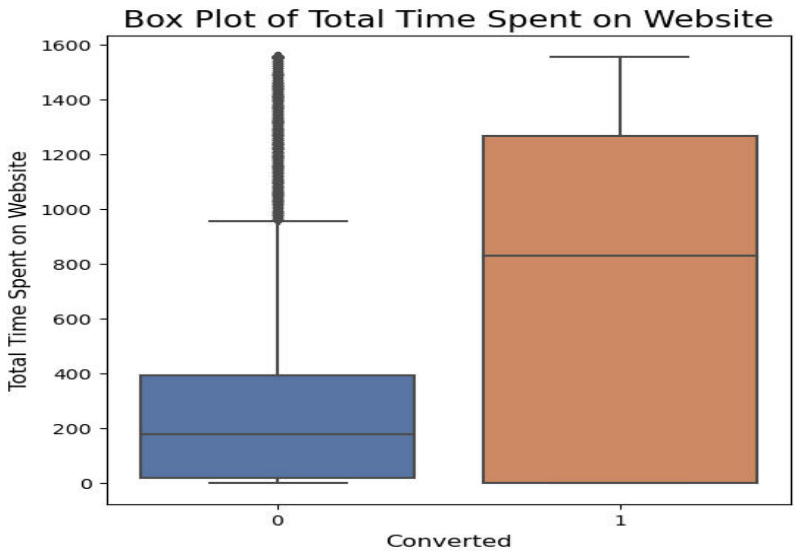
Outlier Treatment- Pre & Post

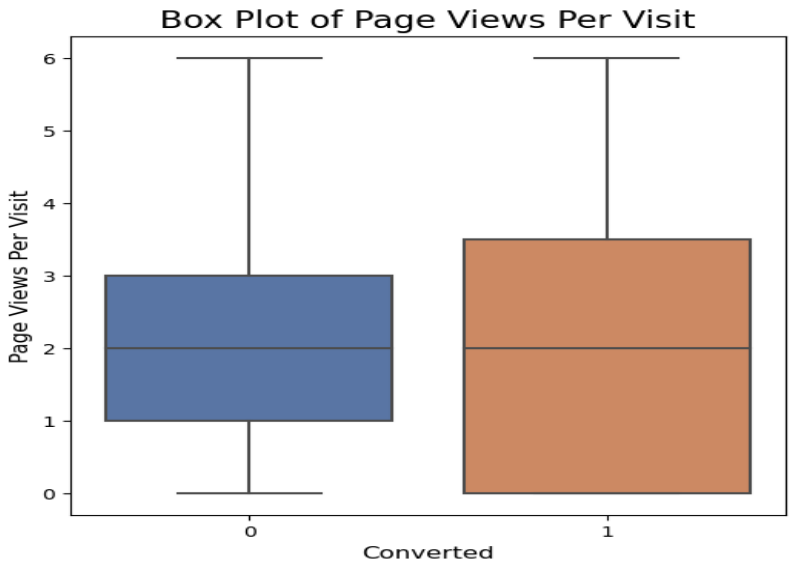




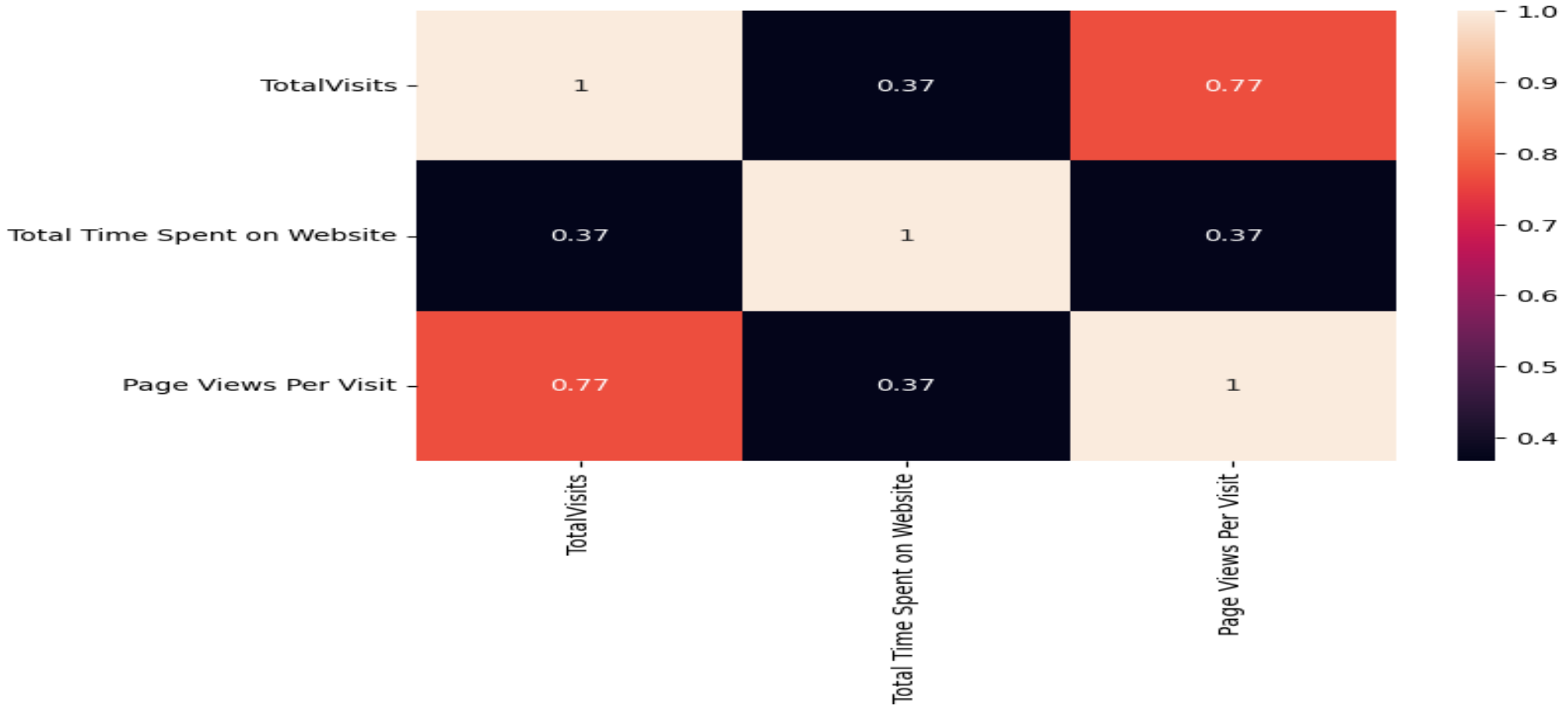








HEATMAP – NUMERICAL VARIABLES



FINAL MODEL

```
In [884]: result_7.summary()
```

Out[884]:

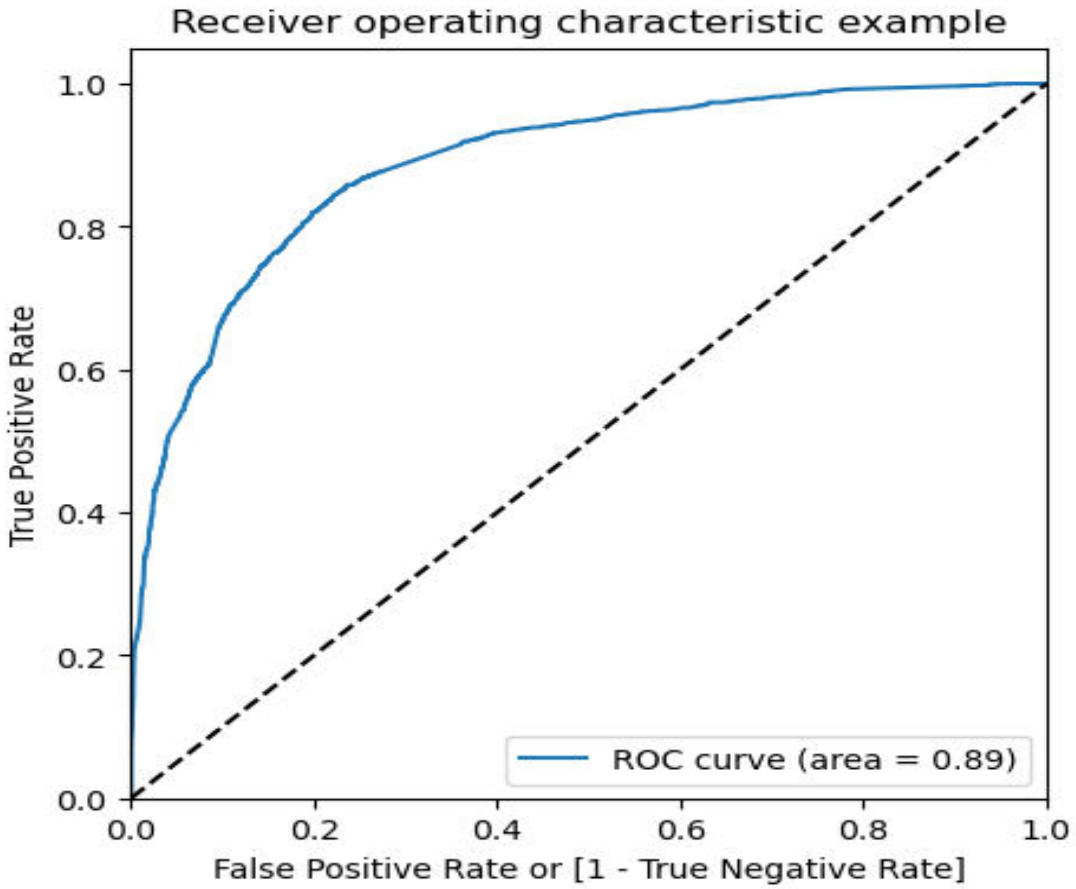
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6336
Model Family:	Binomial	**Df Model:**	14
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2601.1
Date:	Mon, 15 Apr 2024	**Deviance:**	5202.2
Time:	11:11:21	**Pearson chi2:**	6.37e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.4019
Covariance Type:	nonrobust		

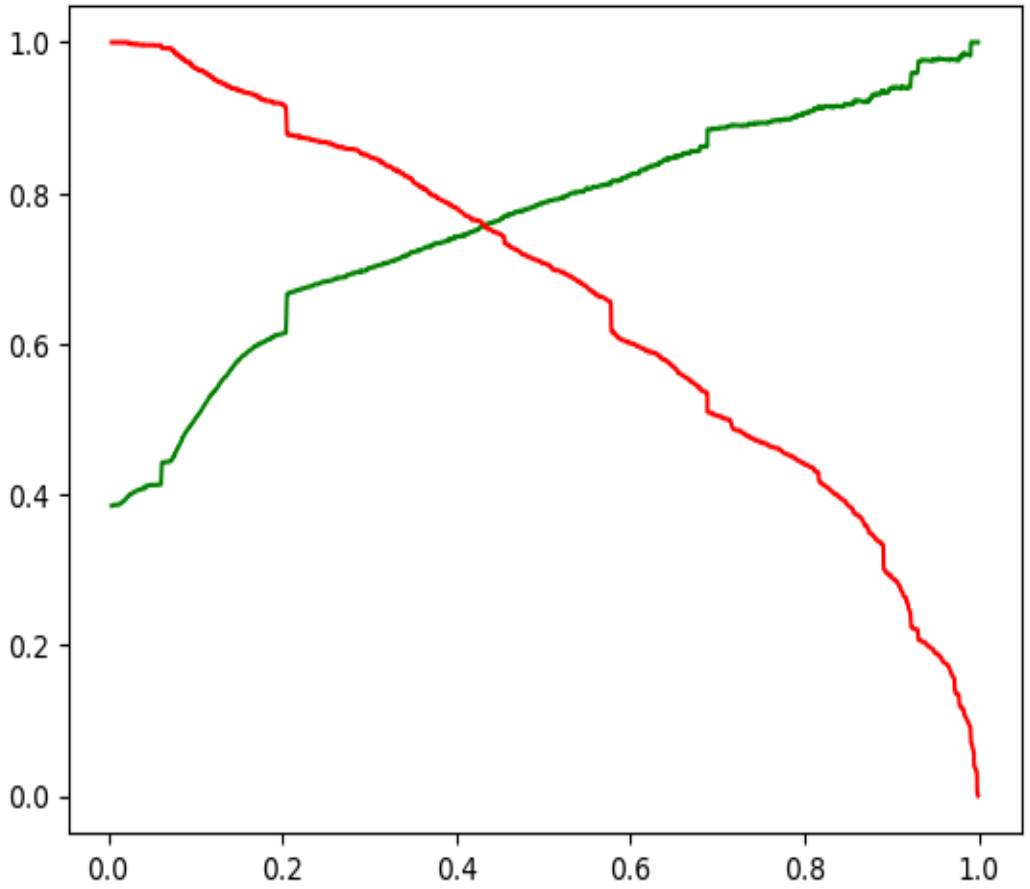
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3820	0.124	-3.080	0.002	-0.625	-0.139
Do Not Email	-1.7696	0.187	-9.462	0.000	-2.136	-1.403
Total Time Spent on Website	1.1269	0.040	27.927	0.000	1.048	1.206
Lead Origin_Landing Page Submission	-1.1397	0.128	-8.885	0.000	-1.391	-0.888
Lead Origin_Lead Import	1.1145	0.476	2.339	0.019	0.181	2.048
Lead Source_Olark Chat	1.2213	0.124	9.858	0.000	0.978	1.464
Lead Source_Reference	3.4987	0.243	14.392	0.000	3.022	3.975
Lead Source_Welingak Website	6.1438	0.735	8.361	0.000	4.704	7.584
Last Activity_Olark Chat Conversation	-1.3775	0.164	-8.377	0.000	-1.700	-1.055
Last Activity_Other_Activities	1.9970	0.454	4.396	0.000	1.107	2.887
Last Activity_Unsubscribed	1.5181	0.472	3.213	0.001	0.592	2.444
Specialization_Others/Unspecified	-1.1794	0.125	-9.438	0.000	-1.424	-0.934
What is your current occupation_Working Professional	2.5841	0.194	13.352	0.000	2.205	2.963
Last Notable Activity_SMS Sent	1.6722	0.081	20.721	0.000	1.514	1.830
Last Notable Activity_Unreachable	1.8093	0.476	3.805	0.000	0.877	2.741

Final Model with predicted features and also with stabilized p-values(<0.05) and VIF values(<5)

ROC CURVE AND TRADE OFF CURVE



ROC Curve



Trade-off curve between Precision and Recall

OBSERVATION

TRAIN DATA

- Accuracy : 80.68 %
- Sensitivity :82.37%
- Specificity :79.56%

TEST DATA

- Accuracy : 80.53 %
- Sensitivity : 81.90 %
- Specificity : 79.75%

CONCLUSION

Lead Source_Welingak Website (coef = 6.1438):

- This feature has the highest positive coefficient, suggesting that leads originating from the Welingak Website have a strong positive impact on the predicted outcome.

Lead Source_Reference (coef = 3.4987):

- Leads from reference sources also have a significant positive impact on the outcome.

What is your current occupation_Working Professional (coef = 2.5841):

- The occupation being a "Working Professional" is positively associated with the predicted outcome.

Last Notable Activity_SMS Sent (coef = 1.6722):

- Sending an SMS as the last notable activity is positively correlated with the predicted outcome.

Last Notable Activity_Unreachable (coef = 1.8093):

- The activity being marked as "Unreachable" in the last notable activity is also positively correlated with the outcome.

Total Time Spent on Website (coef = 1.1269):

- The time spent on the website by the lead has a moderate positive impact on the outcome.

Lead Source_Olark Chat (coef = 1.2213):

- Leads originating from Olark Chat also contribute positively to the predicted outcome.

THANK YOU