LEAD SCORING CASE STUDY

Ganesh Jalakam
Puneet Dadhich

1. 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.
- X-Education team gets a lots of leads through various sources, but their lead conversion is very poor i.e., 30%. For example, let's 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.

Business Objective:

- The company wanted to build a model 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.
- Company wants to achieve lead conversion rate to be around 80%.

2. DATA INSPECTION AND CLEANING

From the Leads dataset we have total of 9240 rows and 37 columns out of which:

- 4 columns with Float datatype
- 3 columns with Int datatype
- 30 columns with Object datatype

1. Removed Unwanted Columns

we dropped the following columns/variables:

- **'Prospect ID', 'Lead Number'**: Indicates unique id of each record and won't be helpful for our analysis
- **'Magazine', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque'** : These variables are having only one category and hence dropped.

2. Handling NULL Values

- From the above results we can consider the missing values cut-off as 36% and drop the columns: **'How did you hear about X Education', 'Lead Profile', 'Lead Quality', 'Asymmetrique Profile Score', 'Asymmetrique Activity Score', 'Asymmetrique Profile Index', 'Asymmetrique Activity Index', 'City', 'Tags'**.
- Although we have high missing percentage for **'Specialization' (36.5%), 'What matters most to you in choosing a course' (29.3%), 'What is your current occupation' (29.1%), 'Country' (26.6%)** columns it can be useful for us in further analysis. So, instead dropping this column we will replace/impute NULL with 'Unknown'.
- For **'Lead Source' and 'Last Activity'** we can impute with 'Unknown' and for **'TotalVisits' and 'Page Views Per Visit'** we can impute with 0.

3. Dropped Duplicate Records

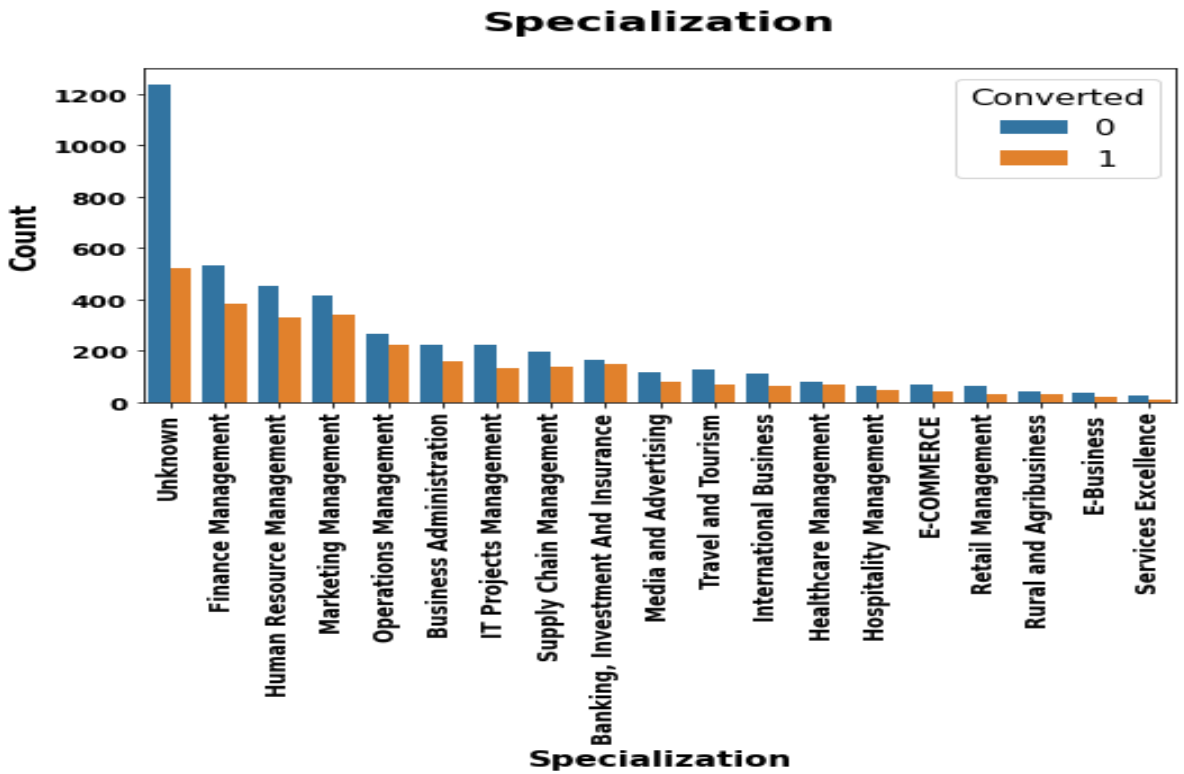
- We found that there are **1847** duplicate records in Leads dataset which may impact our analysis. Hence we dropped these records from the dataset.

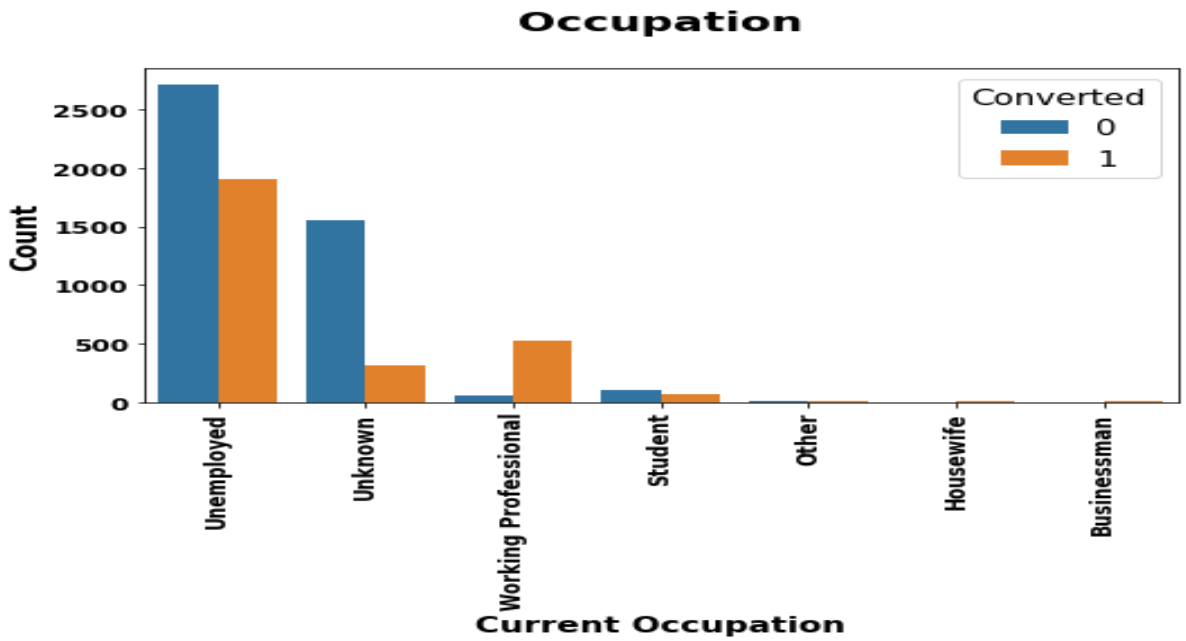
4. Handling Outliers

- Deleted the records for which the values that falls outside of 99th percentile under **'TotalVisits'** and **'Page Views Per Visit'** attributes.
- After performing data cleaning process we have final dataset with **7281 rows** and **21 columns**.

3. DATA VISUALIZATION

Specialization and Occupation



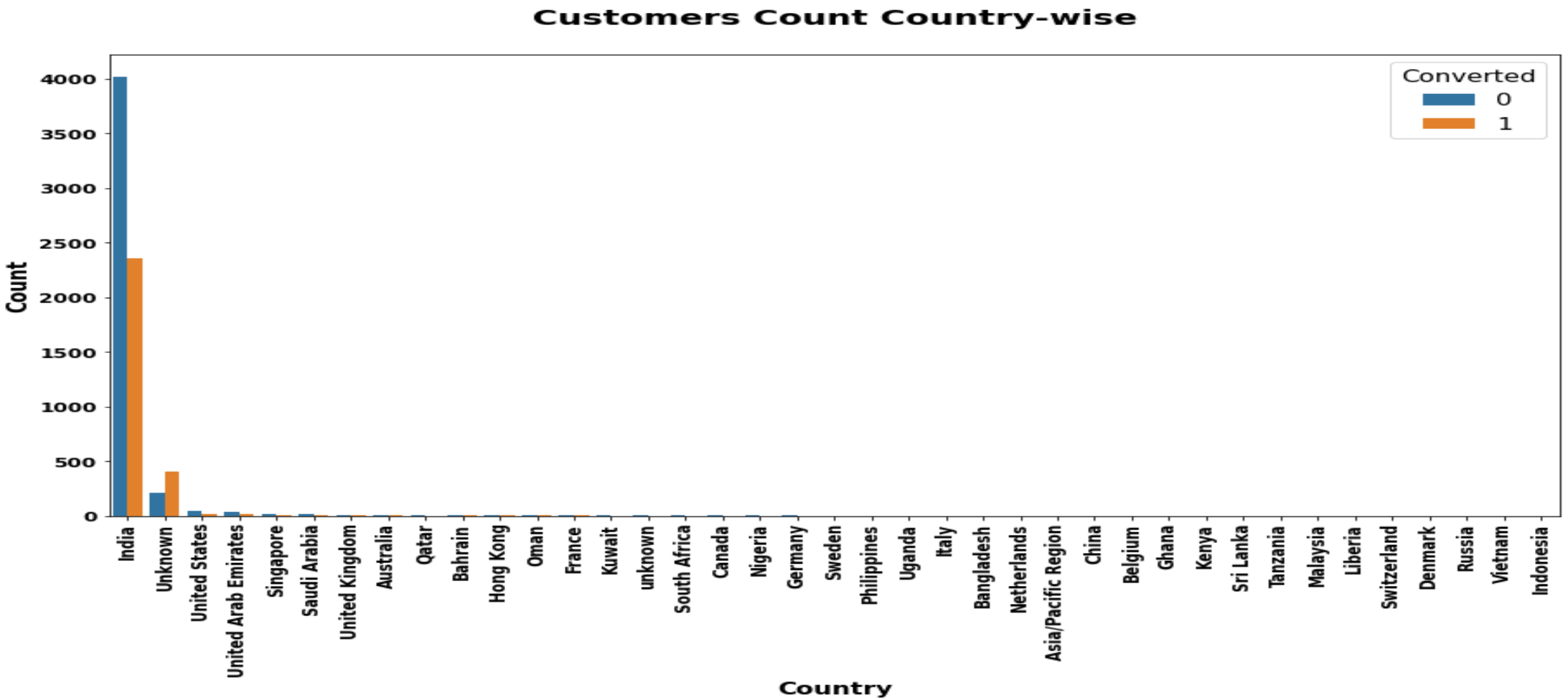


Observation:

From the above plots ignoring Unknown category, we can say that:

- Most of the customers who are working as **Finance Management** has highest conversion rate followed by **Marketing Management** and **Human Resource Management**.
- **Unemployed** Occupation category are higher in terms of population and has a decent amount of conversion rate. Customers who are **working professionals**, although their population is less than **Unemployed** but they are most likely to get converted.

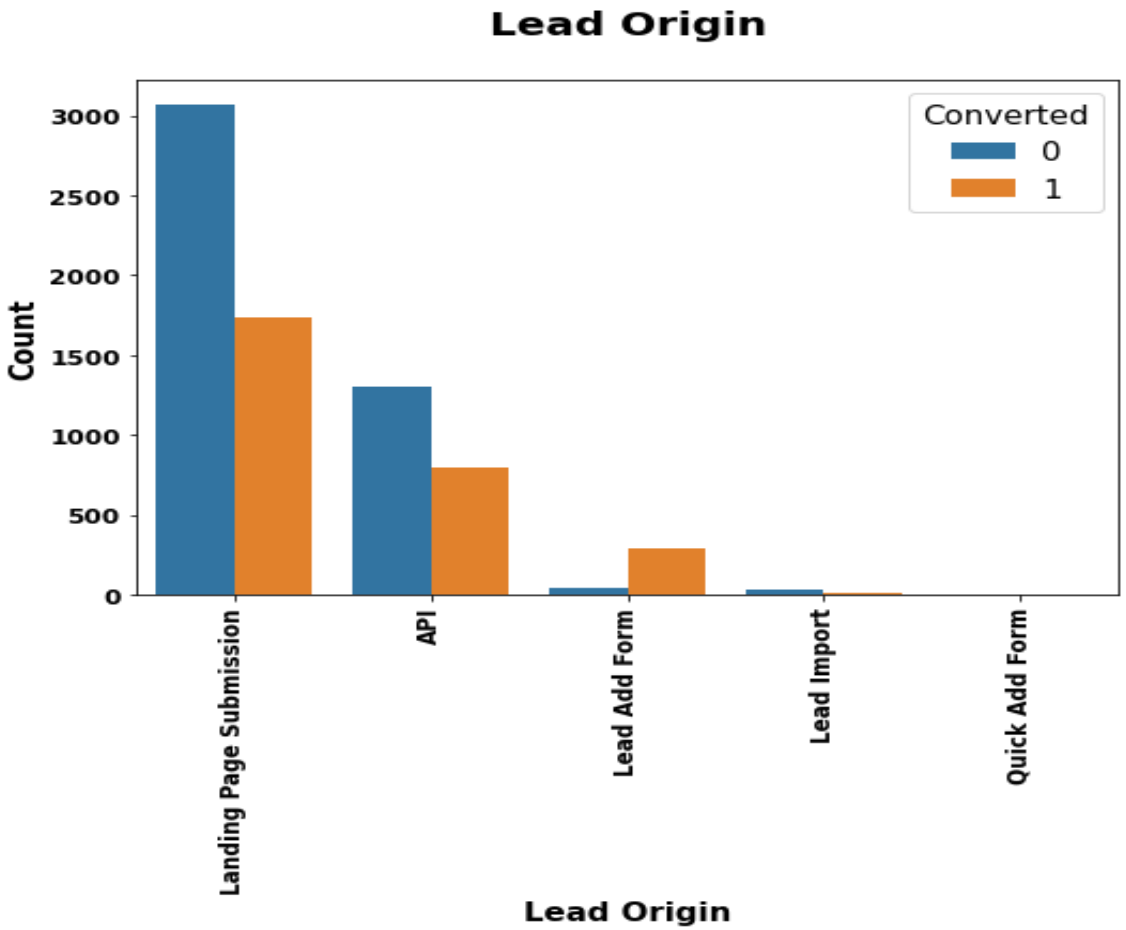
Customer Geographical Location

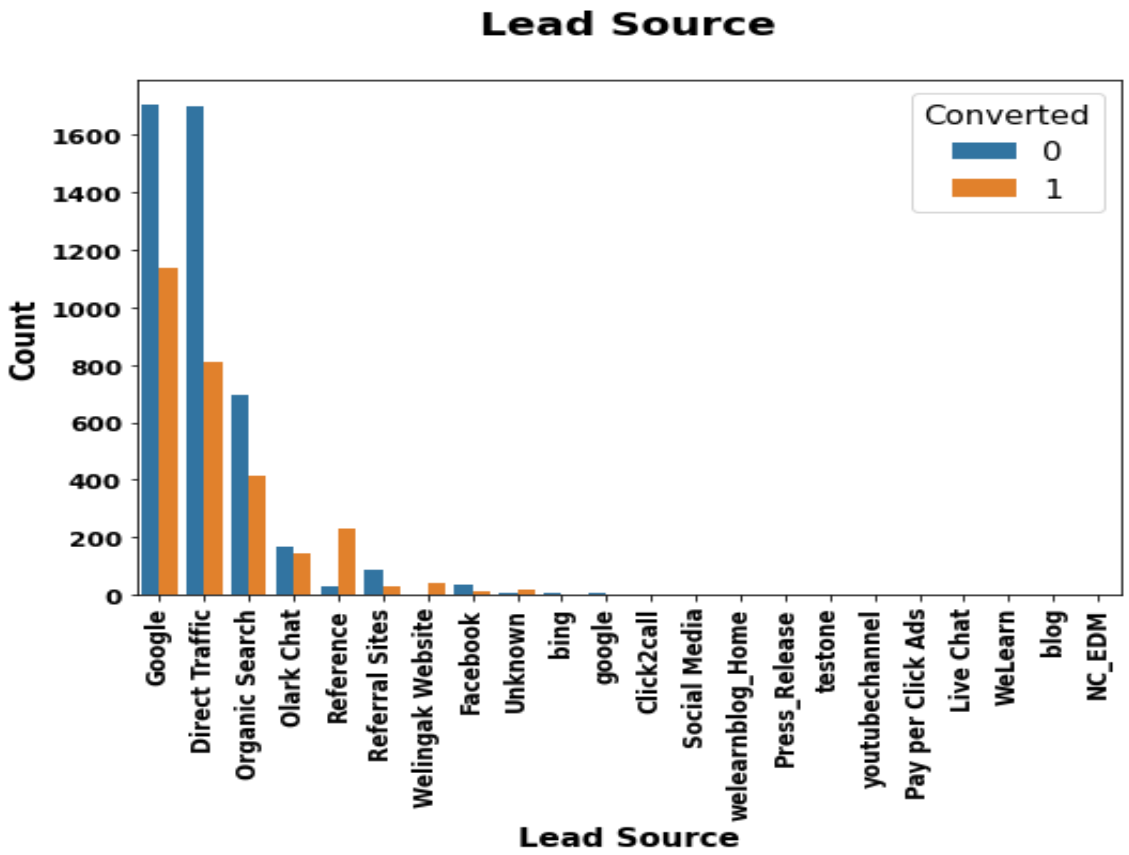


Observation:

- As we can see here that most of the customers belong to **India** and very less in number from other countries.
- Most of the countries (lower in rank) are having only one customer each.

Lead Origin and Lead Source

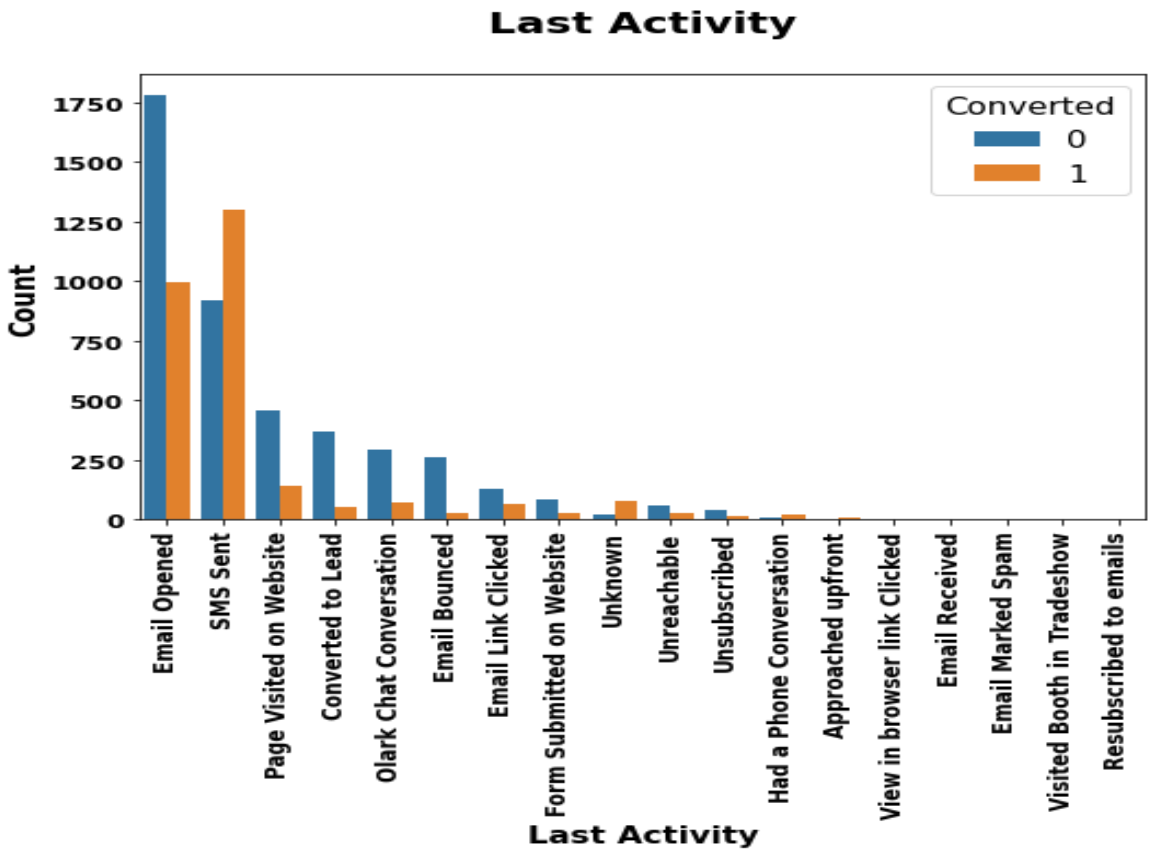


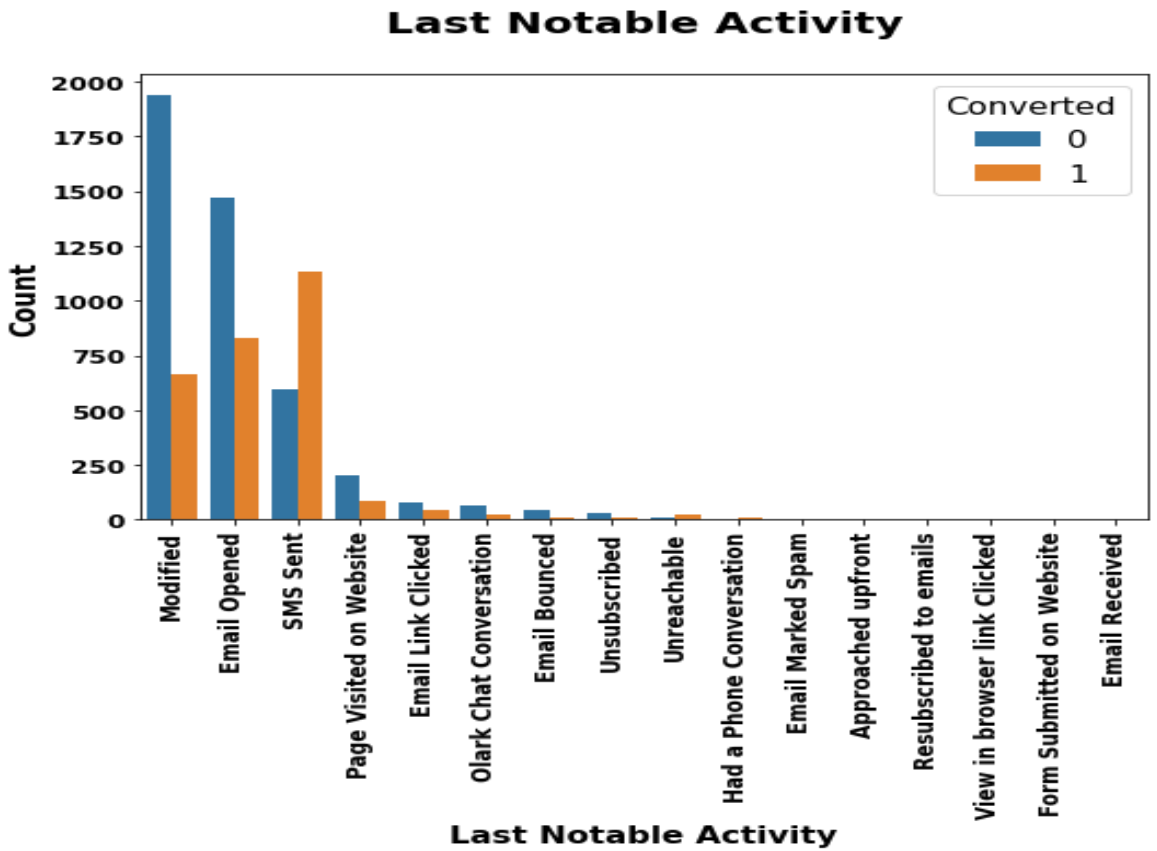


Observation:

- Customers under **'Landing Page Submission'** origin category are higher in total number as well as conversion. Customers who falls under **'Lead Add Form'** category are most likely to get converted.
- Most of the customers got to know the courses through **'Google'** source followed by **'Direct Traffic'** and **'Organic Search'**.

Last Activity and Last Notable Activity

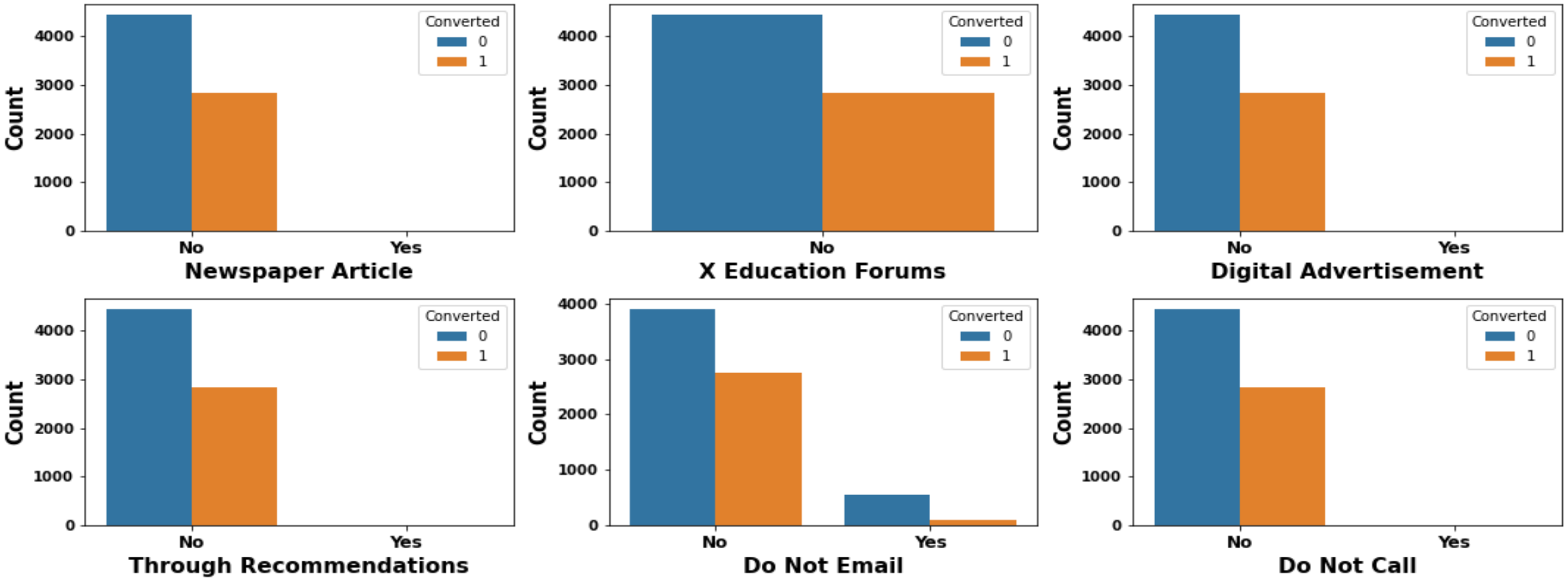




Observation:

- Customers who falls into **'SMS Sent'** and **'Email Opened'** categories are higher in terms of total as well as conversion rate.

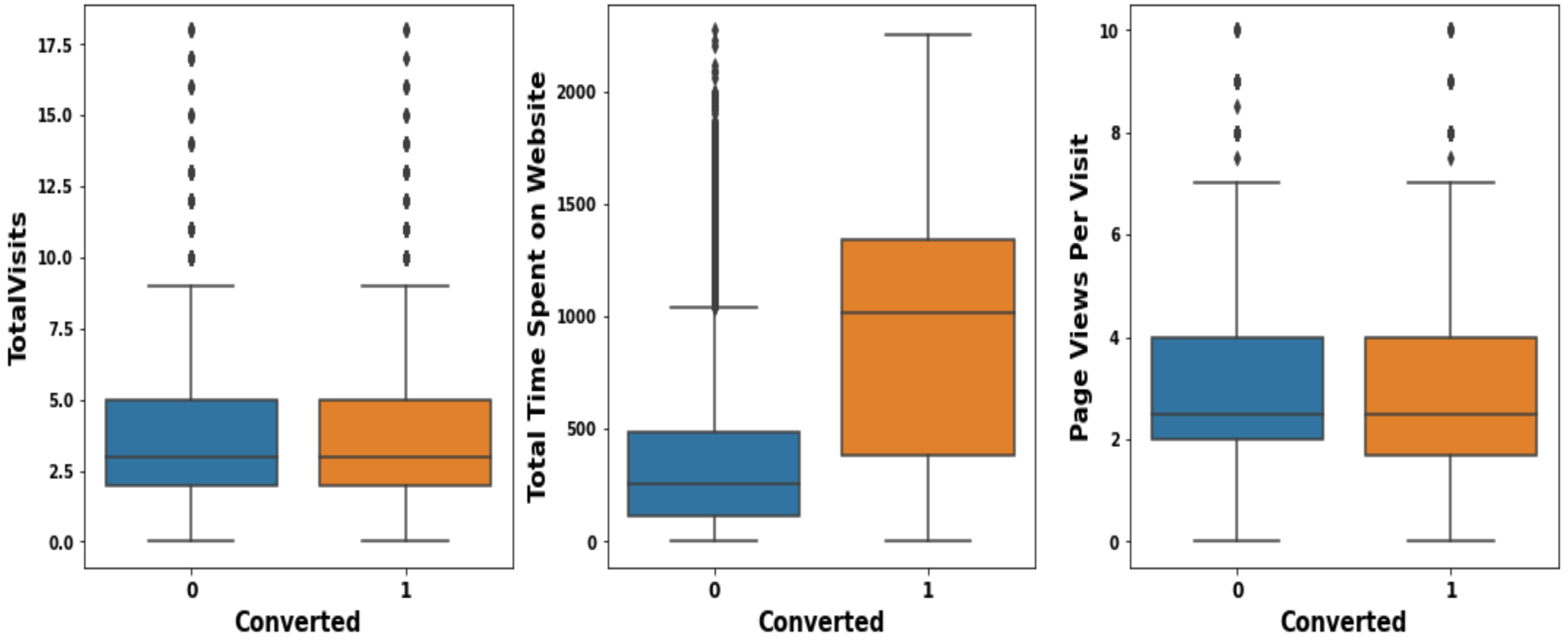
Binary Categorical Variables



Observation

- From the above plots, it seems that very less customers got to know about the courses either through **Newspaper Article** or **Digital Advertisement** or **Recommendations**.
- Very less customers selected/requested to get the course details through **Email** or **Phone Call**.

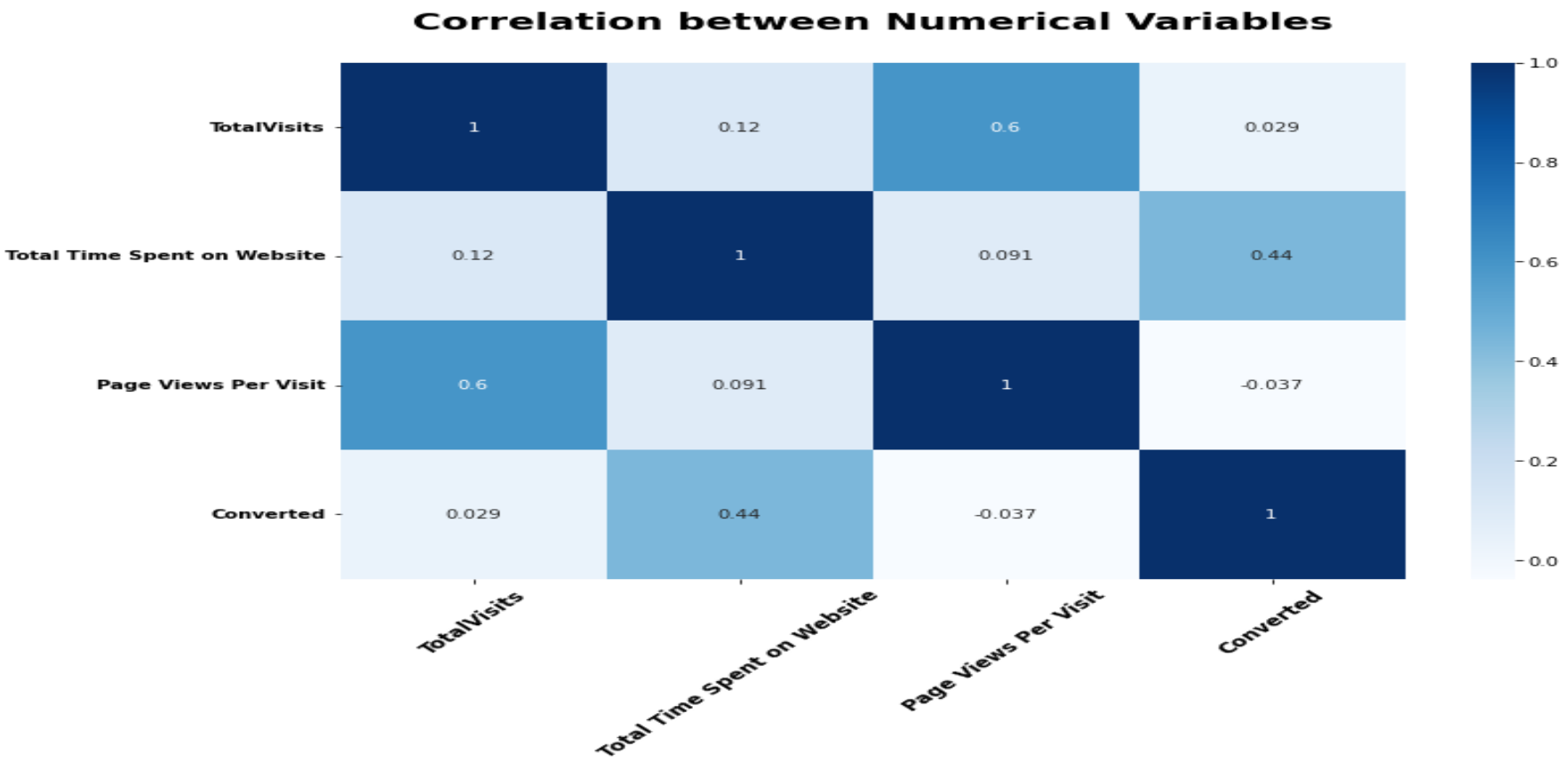
Total Visits, Total Time Spent on Website and Page Views per Visit



Observation:

- We have an average of around 3 units as total number of customers visited the website as well as page views per visit (both converted and not converted).
- Customers who spent more time on the website are likely have a greater chance to get converted.

Correlation between Numerical Variables



Observation:

- Here we can see that **'TotalVisits'** and **'Page Views Per Visit'** variables are highly correlated to each other. So, one of them can be dropped during linear regression model building process.
- There is a good correlation between **'Total Time Spent on Website' (Independent)** and **'Converted' (Target)** variables.

4. DATA PREPARATION FOR BUILDING MODEL

Drop 'X Education Forums' variable

- Dropped ‘X Education Forums’ variable from dataset since it has only one outcome i.e., ‘No’ and doesn’t serve any purpose for further analysis.

Convert binary categorical variables values to 0's and 1's

Converted the following binary categorical variables outcomes of ‘No’ and ‘Yes’ to 0 and 1.

- **'Do Not Email', 'Do Not Call', 'Search', 'Newspaper Article', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'A free copy of Mastering The Interview’ .**

Create Dummy Variables for Multi-level Categorical Variables

Created dummy variables for the following multi-level categorical variables and dropped the original variables.

- **'Lead Origin', 'Lead Source', 'Last Activity', 'Country', 'Specialization', 'What is your current occupation', 'What matters most to you in choosing a course', 'Last Notable Activity’ .**

5. LOGISTIC REGRESSION MODEL BUILD

We have around 133 independent variables in the dataset. We did the following process to build logistic regression model:

- ❖ Split the dataset into Train and Test sets with a proportion of 70% and 30% respectively.
- ❖ Performed MinMaxScaler technique to train set to rescale and bring all the independent variables to same scale.
- ❖ Used RFE (Recursive Feature Elimination) initially to select top 15 variables that are useful for model build.
- ❖ Performed iterative operations of removing the features one by one based on P-Value and VIF (Variance Inflation Factor) until we achieved P-Value < 0.5 and VIF < 5 for all the features.

Final Model Built

Out[85]:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	5096
Model:	GLM	**Df Residuals:**	5084
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2249.4
Date:	Sat, 11 Jun 2022	**Deviance:**	4498.9
Time:	20:57:12	**Pearson chi2:**	5.54e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

Model Summary

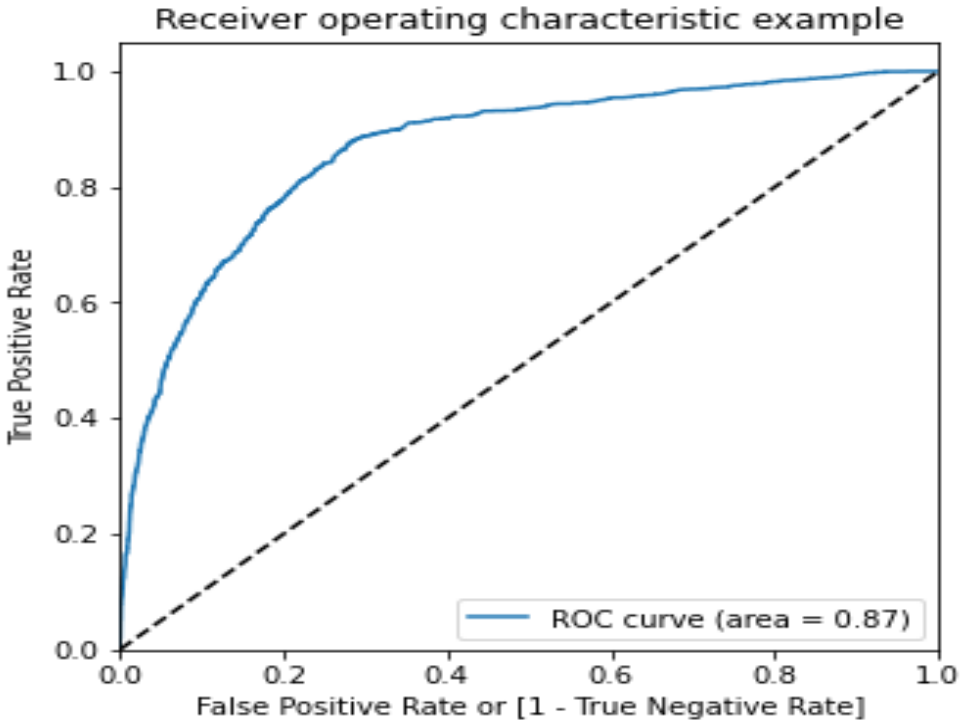
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8252	0.086	-9.650	0.000	-0.993	-0.658
Do Not Email	-1.3448	0.201	-6.705	0.000	-1.738	-0.952
Total Time Spent on Website	4.5017	0.166	27.124	0.000	4.176	4.827
Lead Origin_Lead Add Form	2.3637	0.271	8.737	0.000	1.833	2.894
Last Activity_Email Bounced	-1.3099	0.375	-3.493	0.000	-2.045	-0.575
Country_Unknown	1.3507	0.174	7.773	0.000	1.010	1.691
What is your current occupation_Working Professional	2.5828	0.186	13.915	0.000	2.219	2.947
Last Notable Activity_Email Link Clicked	-1.4098	0.292	-4.824	0.000	-1.983	-0.837
Last Notable Activity_Email Opened	-1.3252	0.096	-13.831	0.000	-1.513	-1.137
Last Notable Activity_Modified	-1.9048	0.100	-18.963	0.000	-2.102	-1.708
Last Notable Activity_Olark Chat Conversation	-2.3691	0.367	-6.450	0.000	-3.089	-1.649
Last Notable Activity_Page Visited on Website	-1.7588	0.208	-8.449	0.000	-2.167	-1.351

VIF

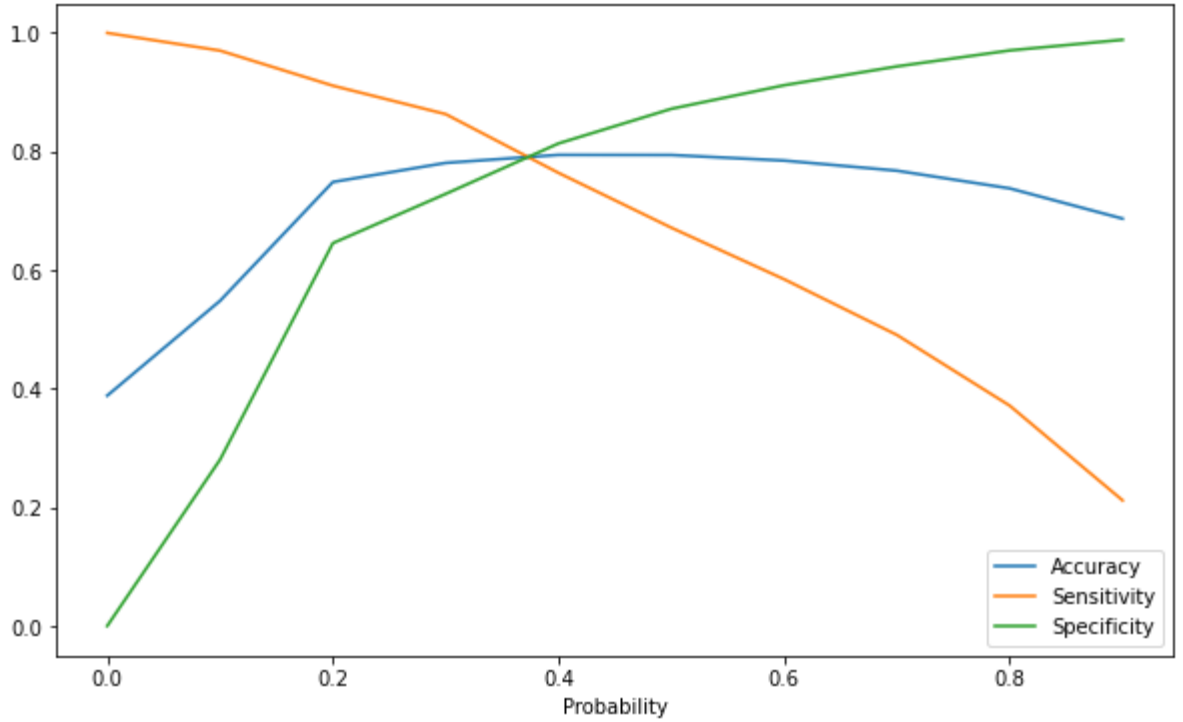
	Features	VIF
4	Country_Unknown	1.99
2	Lead Origin_Lead Add Form	1.90
0	Do Not Email	1.76
3	Last Activity_Email Bounced	1.73
1	Total Time Spent on Website	1.63
8	Last Notable Activity_Modified	1.41
7	Last Notable Activity_Email Opened	1.27
5	What is your current occupation_Working Profes...	1.13
10	Last Notable Activity_Page Visited on Website	1.06
9	Last Notable Activity_Olark Chat Conversation	1.02
6	Last Notable Activity_Email Link Clicked	1.01

6. CALCULATE OPTIMAL CUT-OFF VALUE

ROC Curve



Cut-Off Graph



We evaluated Accuracy, Sensitivity and Specificity metrics for all the probabilities between 0 and 0.9 and decided the cut-off value as **0.35** for prediction as all these three metrics lines are intersecting at that point as shown in the above graph.

7. MODEL PREDICTION AND EVALUATION

- We predicted both train and test sets by considering cut-off predicted probability value as 0.35, did evaluation in both the sets and got good results as shown below.

Train Data

Metrics	Score (%)
Accuracy	79
Sensitivity	81.25
Specificity	77.64

Test Data

Metrics	Score (%)
Accuracy	77.7
Sensitivity	82.02
Specificity	74.9

- As you can see that the evaluation metrics for train and test sets are closer to each other. This indicates good predictive power in real world data.

8. CONCLUSION

- We can conclude that the final logistic regression model built has a very good predictive power which means that the model isn't fit by chance and well generalized for prediction.
- For model prediction, we considered the optimal probability cut-off value as **0.35** based on **Sensitivity** and **Specificity** metrics.
- We got the following model evaluation results for train and test sets:
- The top three variables that contributed for prediction are:
  - **'Total Time Spent on Website'**: Coefficient of **4.5017**.
  - **'What is your Current Occupation' (Working Professional)**: Coefficient of **2.5828**.
  - **'Lead Origin' (Lead Add Form)**: Coefficient of **2.3637**.

Metrics	Train Set	Test Set
Model Accuracy	79%	77.7%
Sensitivity	81.25%	82.02%
Specificity	77.64%	74.9%