Case Study on Lead Score– Classification problem

Venkata Ramana Puppala

Contents:

1. Problem Statement:

X Education company sells online courses to industry professionals. Once these people land on the website, they 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. 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%. "X Education gets a lot of leads, its lead conversion rate is very poor." select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

Business Objective:

X education wants to know the most promising leads. For that they want to build a 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. which identifies the hot leads. Deployment of the model for future use. if the company's requirement changes in the future so you will need to handle these as well.

2. Data Set

- Leads dataset with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.
- Data set has 9240 rows and 37 columns
- The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

3. Problem Approach:

Import all the necessary libraries for Data Wrangling , Building model. Exploratory data analysis on the given data set to derive the Business Insights. Feature Engineering > Model Building > Model Evaluation > Feeding the Model to the Test Data.

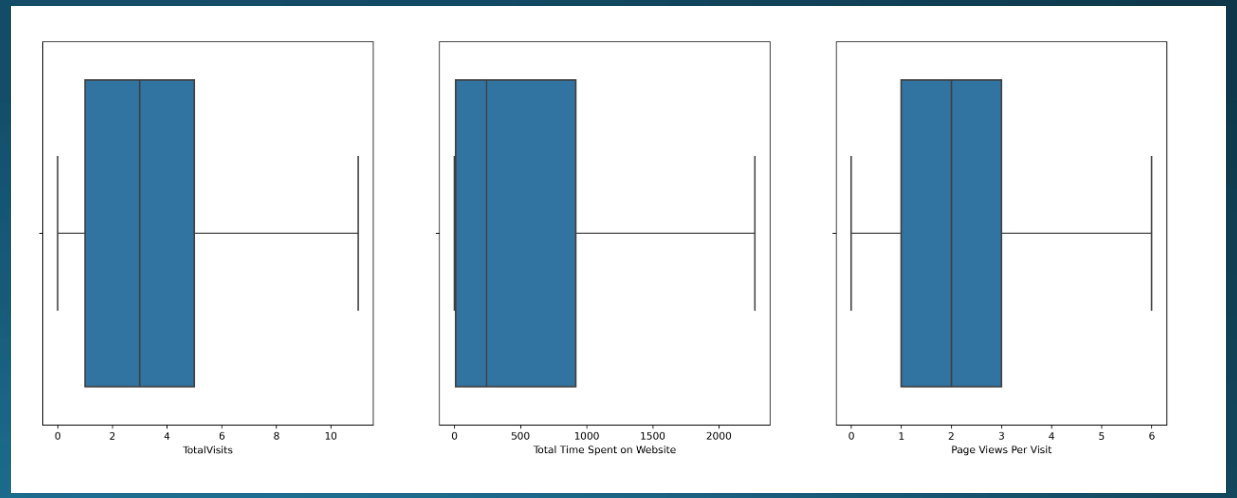
3. Data wrangling:

3.1 Null Value Treatment:

1. Features like 'Lead source','totalvisits','page views per visit','last activity' have very less null value – dropped records
2. Features like 'Country','specialization','how did you hear about X education','what is your current occupation','what matters most to you in choosing a course','tags','lead quality','lead profile','city','asymmetrique activity index','asymmetrique profile index','asymmetrique activity score','asymmetrique profile score' are having null values with high percentage - Dropped columns which are having 40% null values and al the remaining are imputed with missing category.

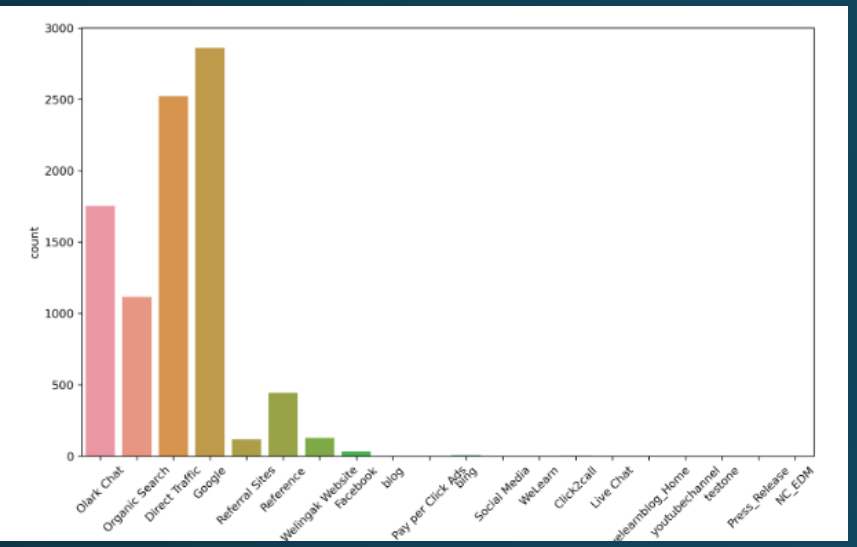
3.2 Outlier Treatment:

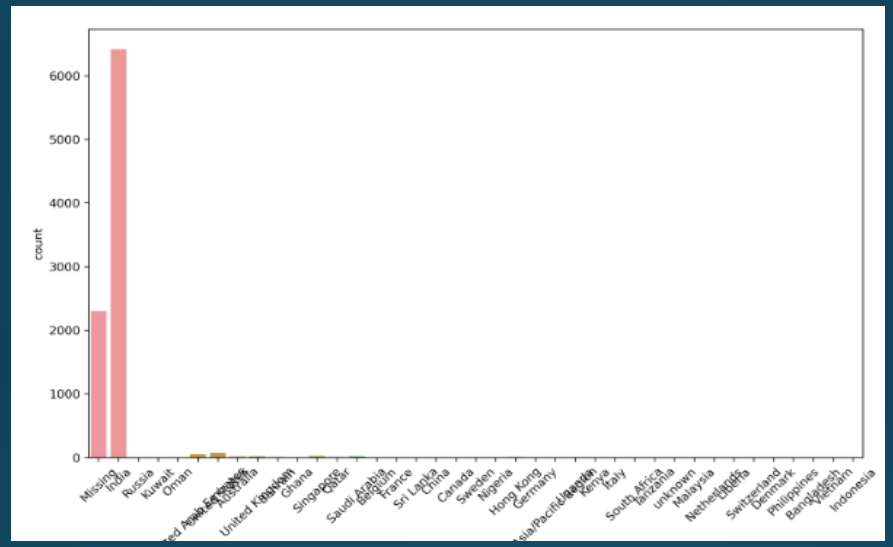
1. check outliers on continuous data ('Total visits','Total time spent on website','page views per visit')
2. As we can see in graphs "Page Views Per Visit","TotalVisits" having outliers.
   after dropping outliers more than 40 in "TotalVisits" and Capping and Flooring done Accordingly
3. After dropping outliers more than 10 in "Page Views Per Visit" and Capping and Flooring done Accordingly.

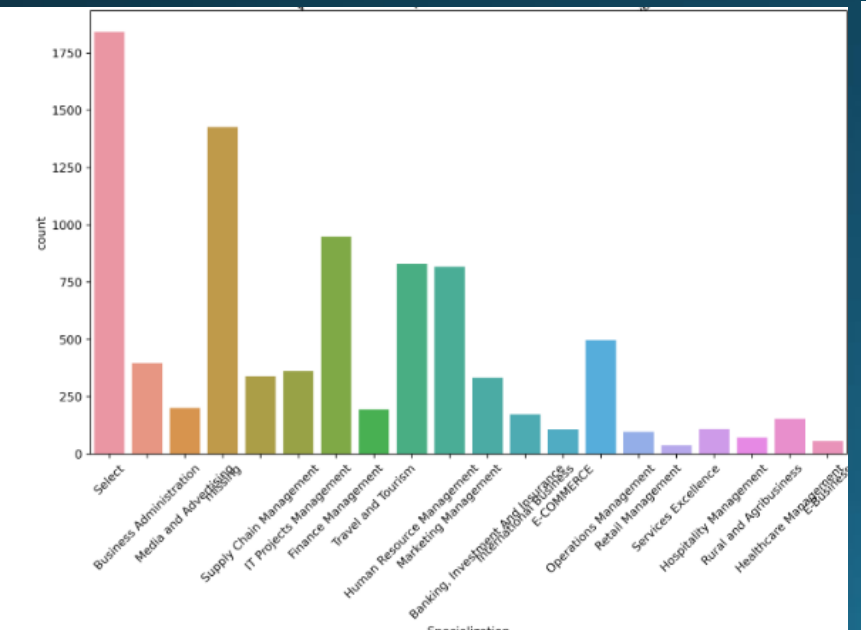


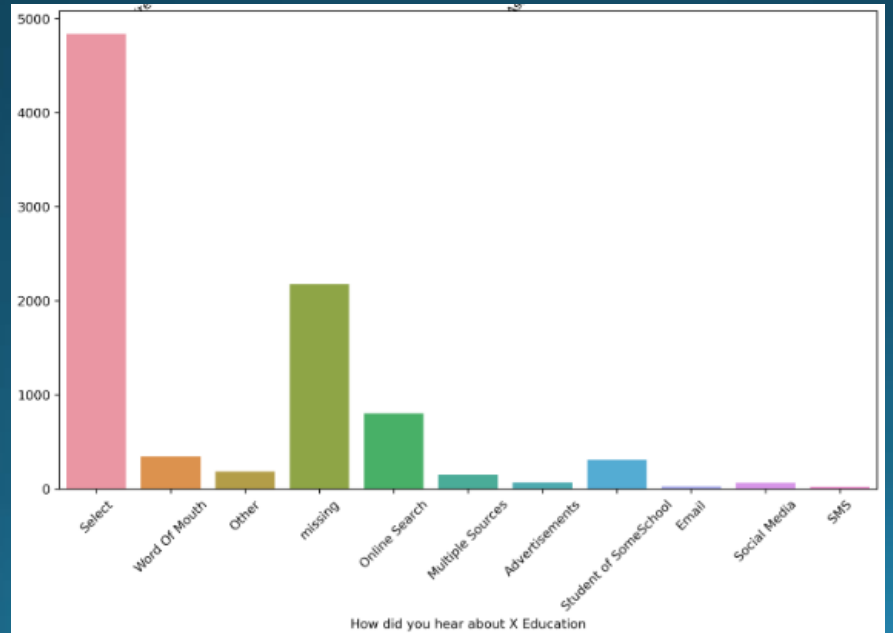
4.Exploratory Data Analysis:

4.1 Univariate Analysis on Categorical Features:



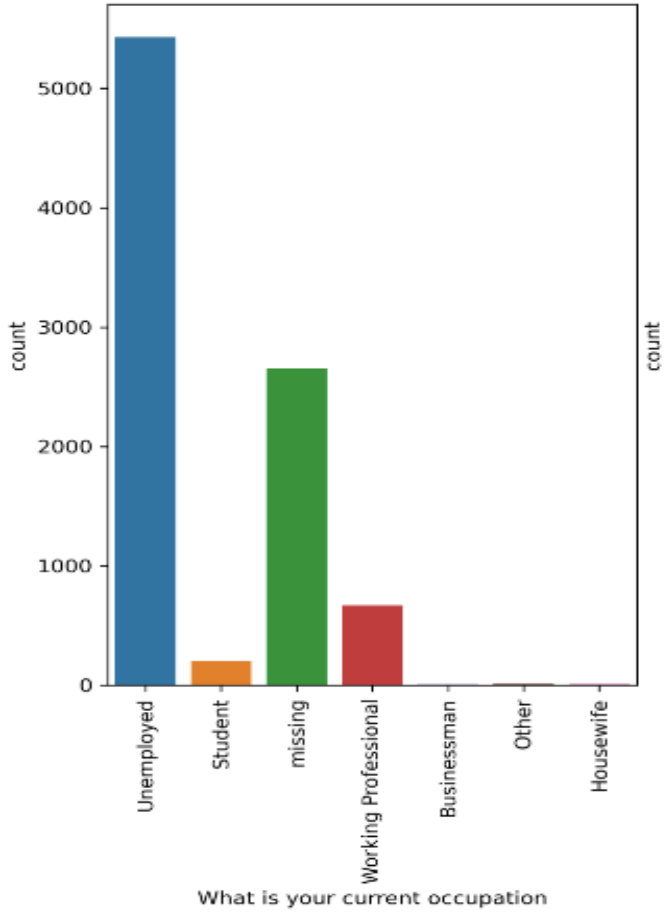


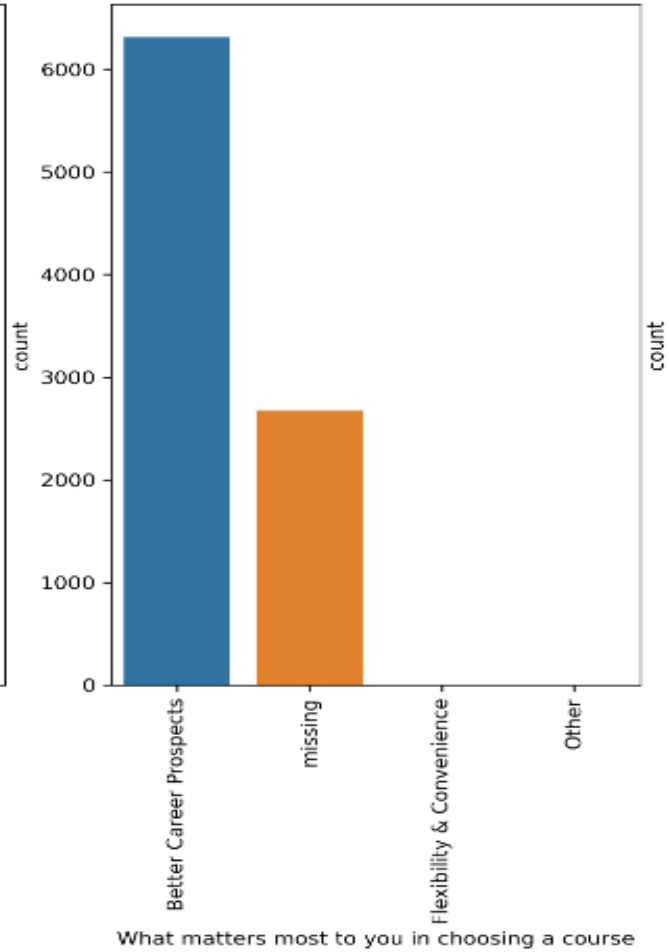


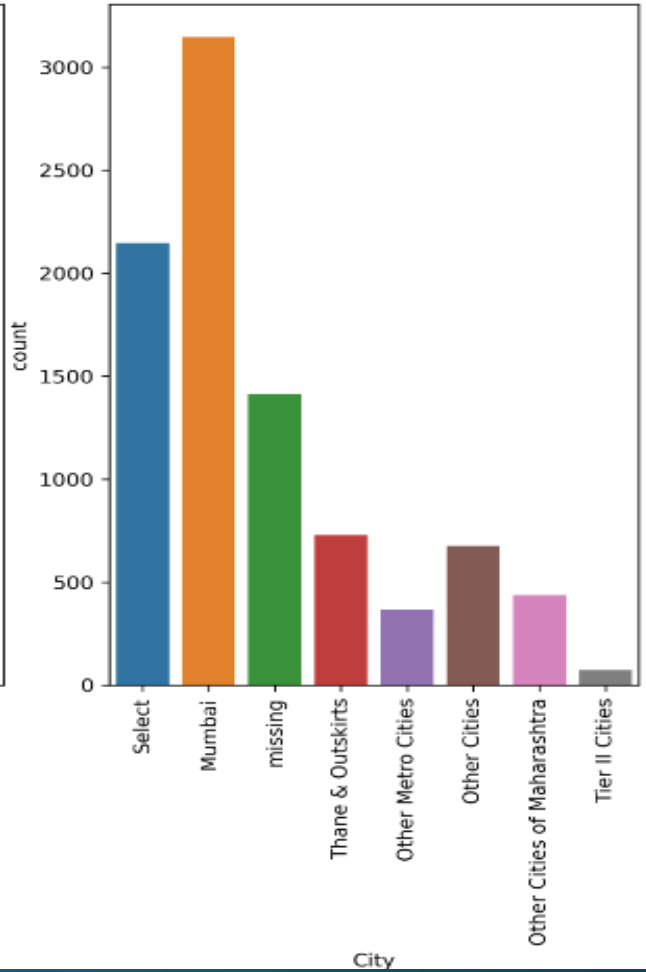


1. It is Observed that Lead Source graph "Google", “direct traffic”, “Organic search”, “olark chat” having higher counts.
2. It is observed that most of the customer is from INDIA.
3. Major conversion in the lead source from google.

4.2 Univariate Analysis on Categorical Features:

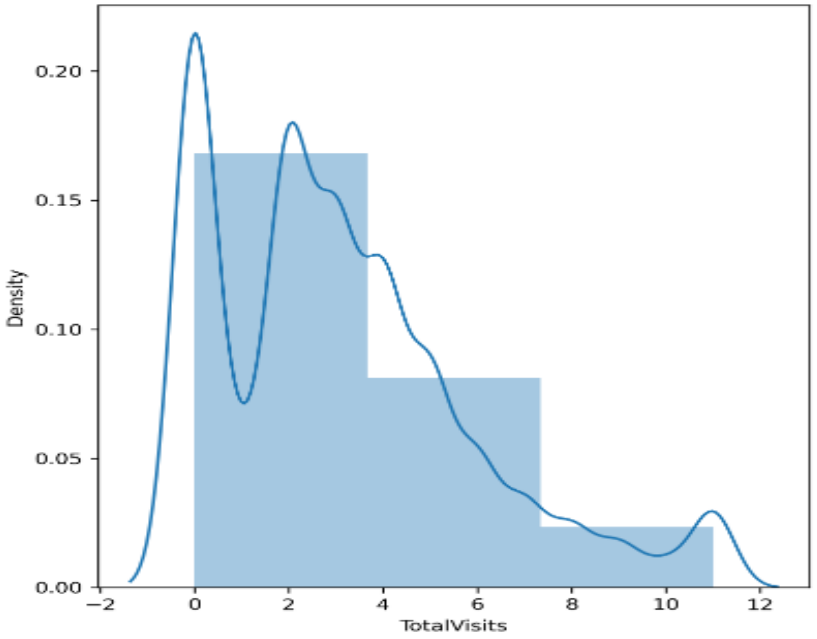


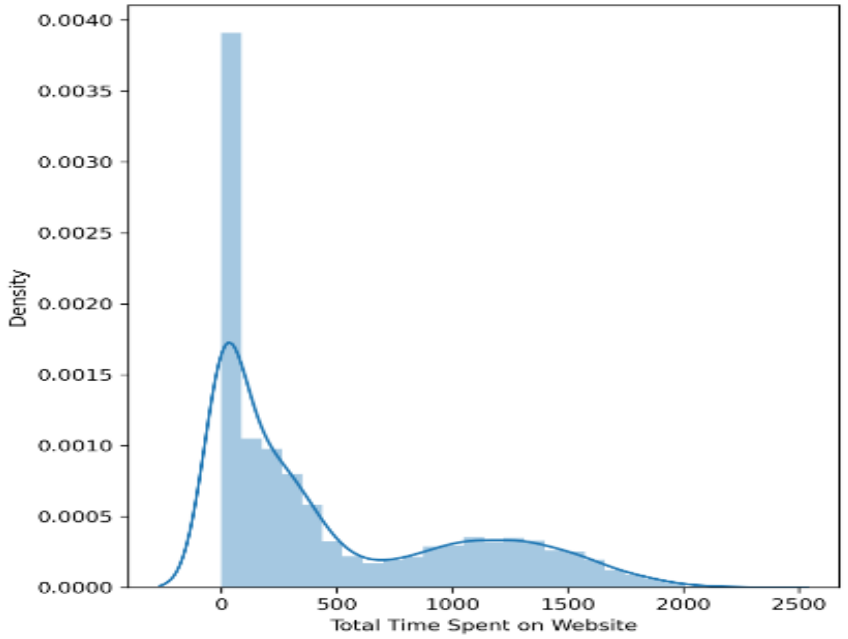


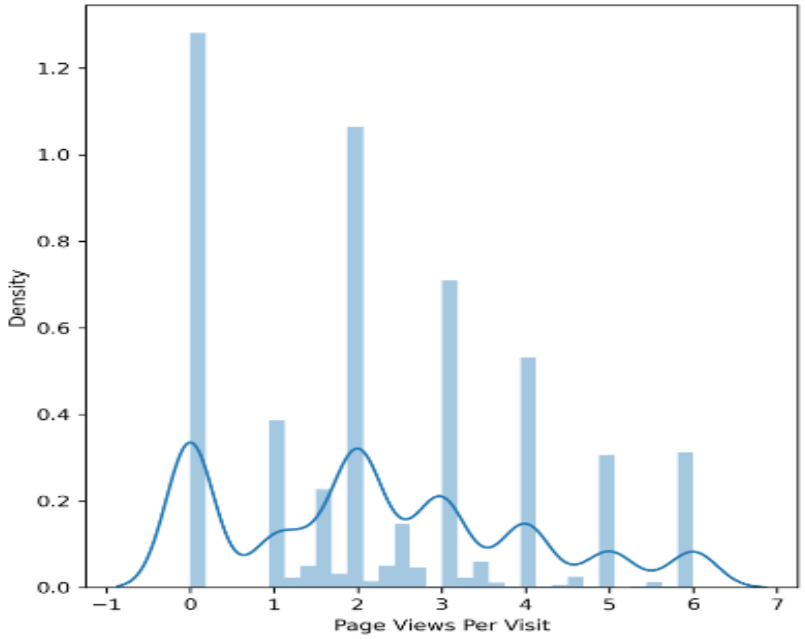


1. Most of the customers are unemployed.
2. Most of the customers choose courses for “Better Career Prospects”
3. Most of the customers are from "Mumbai"

4.3 Univariate Analysis on Continuous Features:



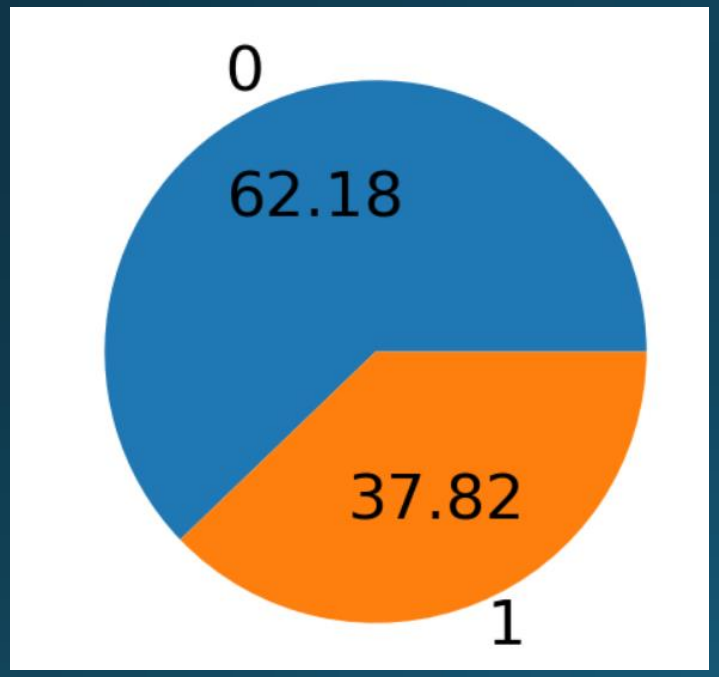


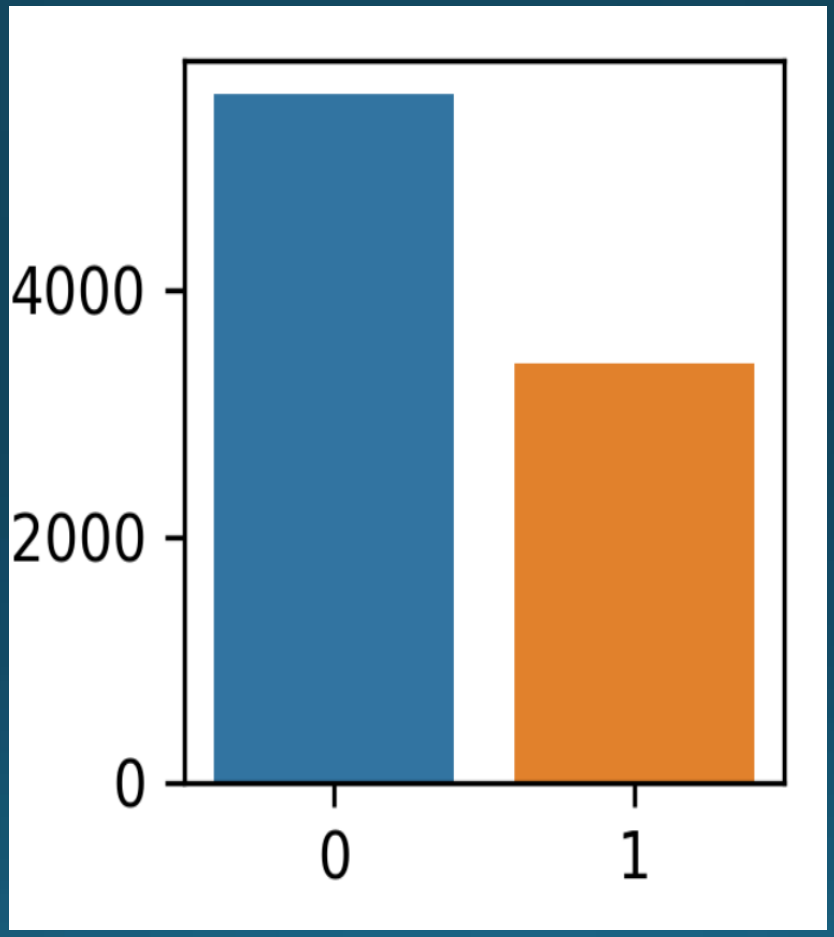


Total Visits and Page Views per Visit are distributed unevenly having peak at o where as Total time Spent on Website following Right skew distribution i.e. Positive Skewness

4.4 Univariate Analysis on Response feature:(Data Imbalance)

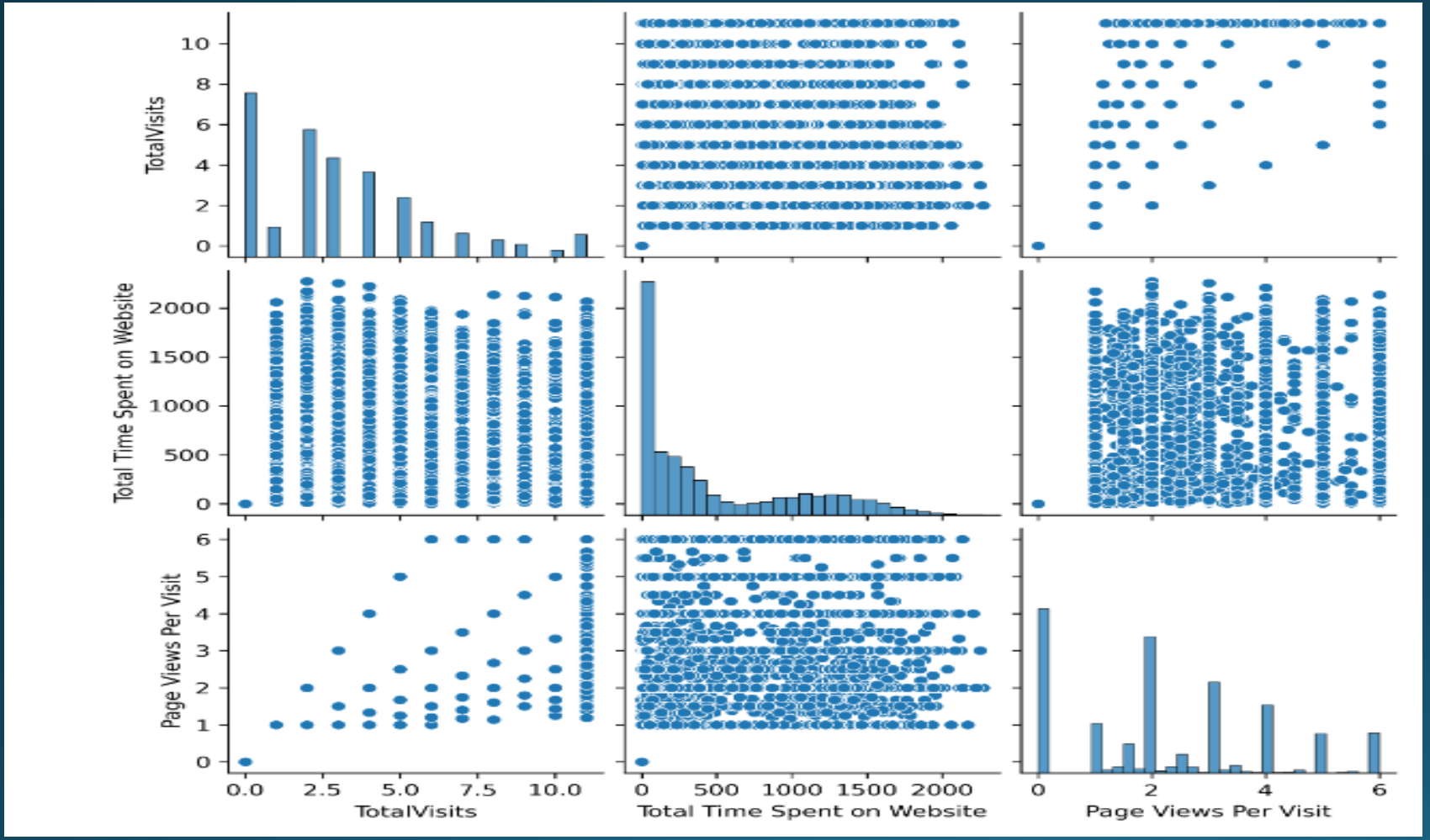
```
0    5591
1    3401
Name: Converted, dtype: int64
```





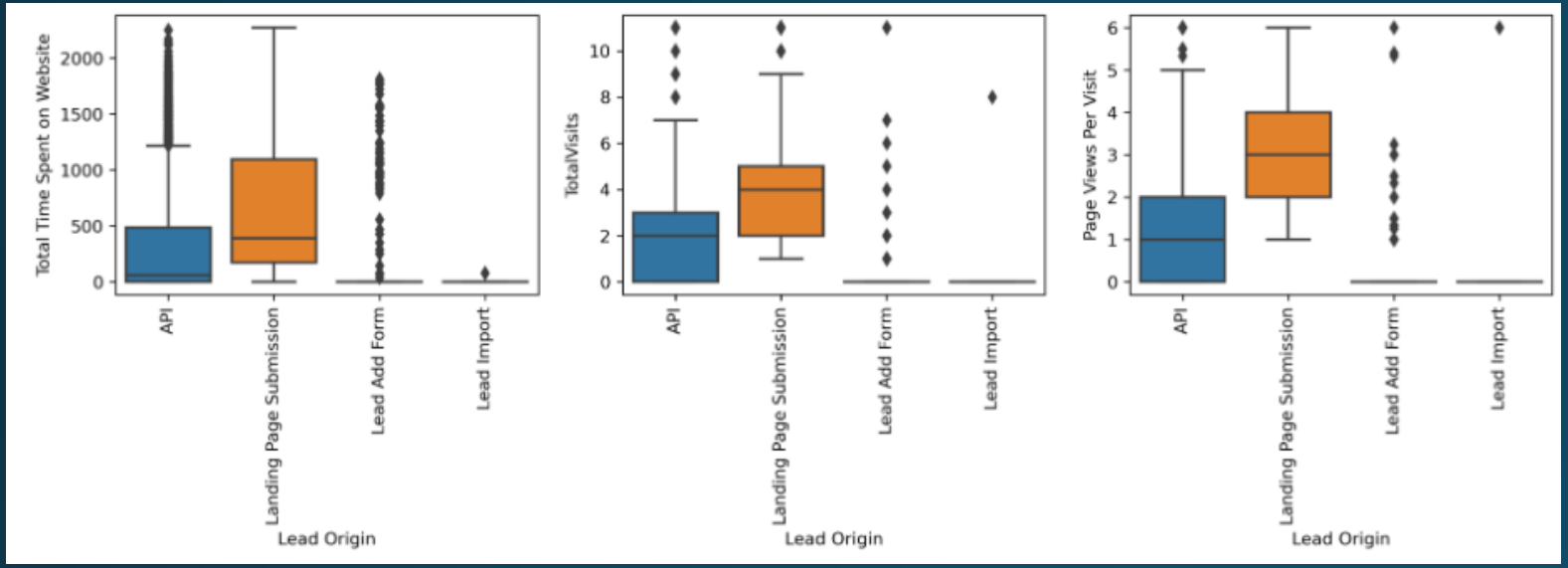
- Total conversion rate is 37.82%

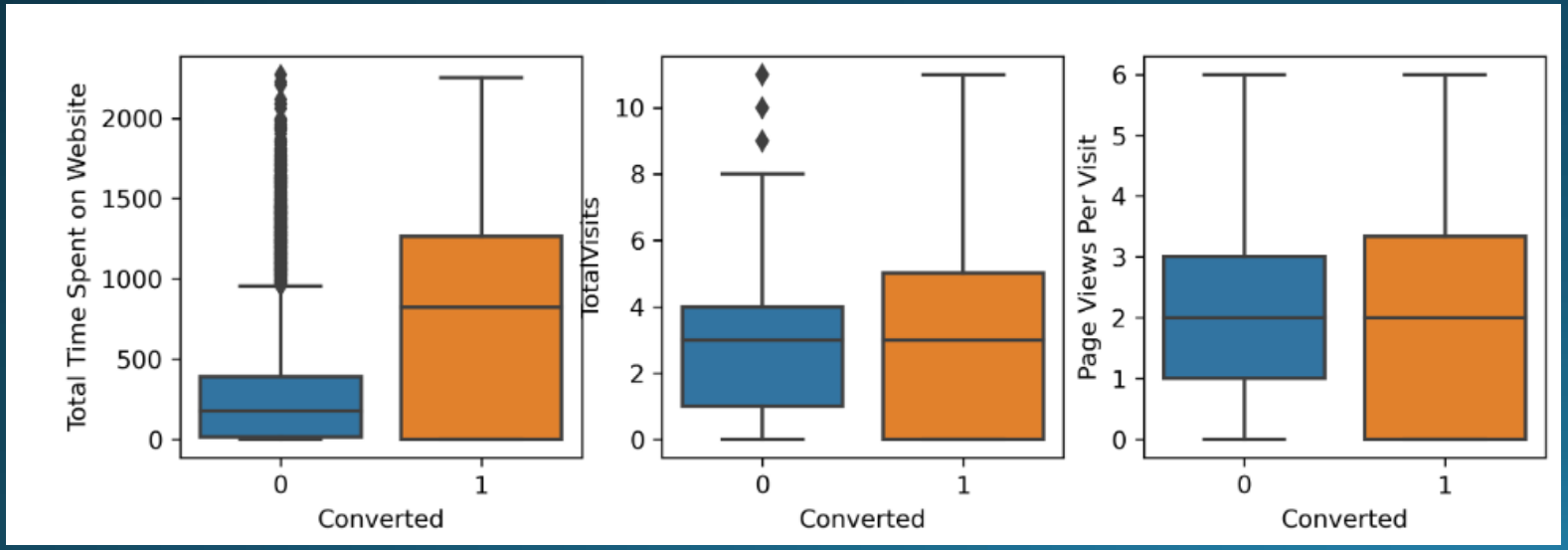
4.5 BI-Variate analysis on Continuous Features:



- It is Observed that Page Views per Visit and Total Visits are Directly proportional to each other up to 10 Total Visits

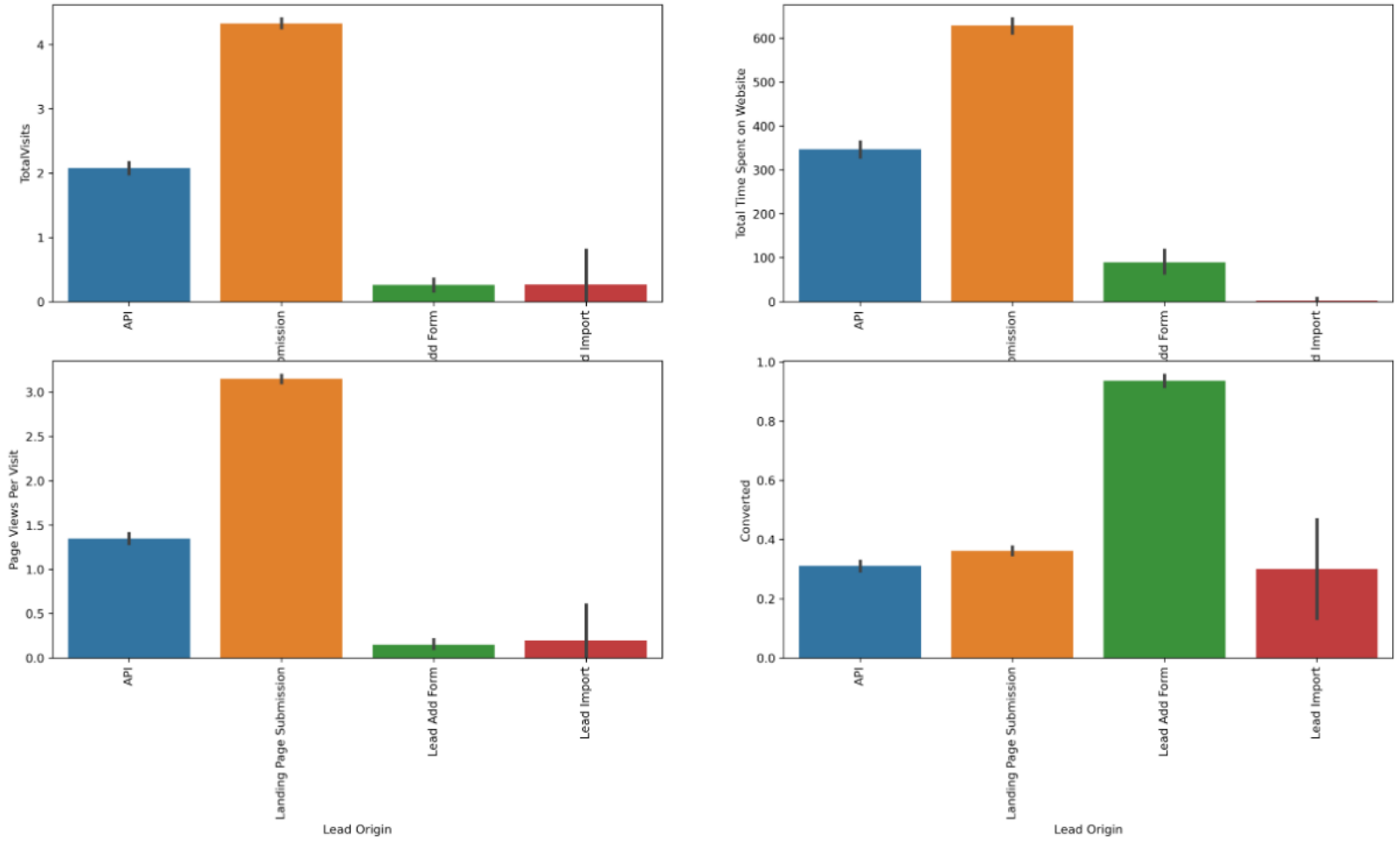
4.6 BI-Variate analysis on Continuous and Categorical features:





1. Leading Page Submission from Lead Origin are Having More Total Time Spent on Website , Total Visits and Page Views Per Visit compared to other Category Levels.
2. Customers who took the Course spent more time on website.

4.7 BI-Variate analysis on Response feature and Continuous features:



1. API and Landing Page Submission bring a higher number of leads as well as conversion.
2. Lead Add Form has a very high conversion rate but the count of leads is not very high.
3. Lead Import and Quick Add Form get very few leads.
4. In order to improve the overall lead conversion rate, we have to improve lead conversion of API and Landing Page Submission
5. origin and generate more leads from Lead Add Form.

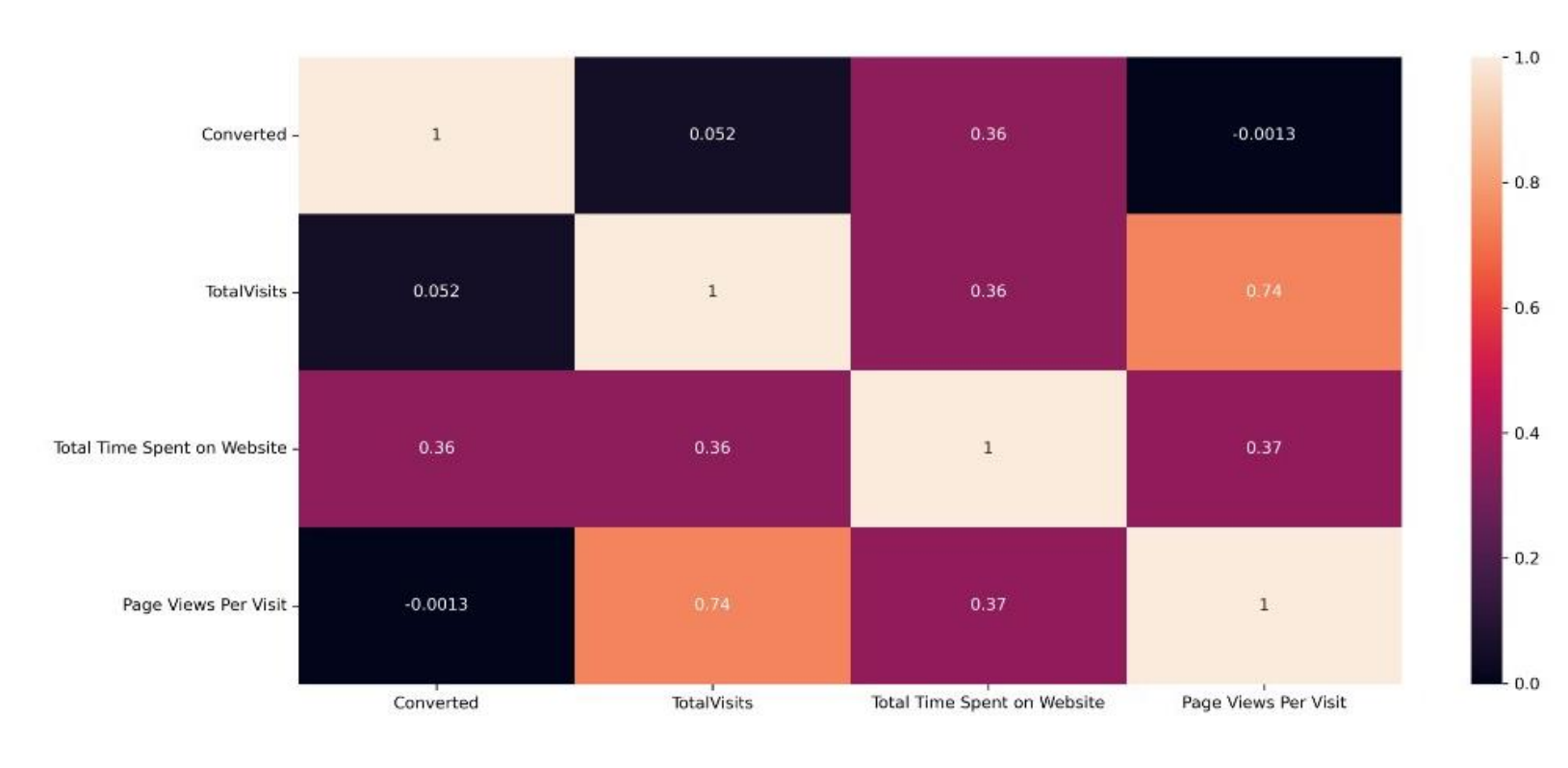
Grouped by City with Sum Aggregation:

City	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Mumbai	1266	13339.0	1931503	9392.49
Other Cities	271	2750.0	398471	1954.56
Other Cities of Maharashtra	190	1711.0	280062	1279.27
Other Metro Cities	149	1511.0	224120	1134.31
Select	1019	3820.0	687379	2704.03
Thane & Outskirts	328	2944.0	453157	2044.66
Tier II Cities	25	282.0	45213	209.40
missing	153	2080.0	303099	1365.60

Grouped by Country with Sum Aggregation:

Country	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Asia/Pacific Region	1	3.0	1667	2.00
Australia	3	60.0	8694	48.00
Bahrain	4	29.0	5294	18.00
Bangladesh	1	8.0	2192	5.00
Belgium	0	9.0	310	6.00
Canada	0	22.0	1079	11.33
China	0	4.0	383	3.00
Denmark	1	4.0	1200	4.00
France	3	34.0	4480	16.73
Germany	1	18.0	2028	13.50
Ghana	0	4.0	431	4.00
Hong Kong	4	30.0	5896	22.00
India	2368	26895.0	4065476	19008.60

4.10 Heat map with Correlation Values:



- It is Observed that Total Visits and Page Views per Visit are highly Correlated where as Converted Customers Don't have any correlation with Page views per visit.

5.Feature Engineering:

1. Splitting the data set into train and test data with proportion of 70 to 30 ratio.
2. Creating dummy features for the categorical variables which are having more than two levels also dropped the first dummy feature in all categorical variables to reduce number of independent features for model building.
3. Feature Scaling : Min max scaler is used to normalise the continuous data. It converts all numerical values into 0 to 1.

6.Model Building:

1. Logistic Regression – Generalised Linear models is used to build a model as it is a classification problem. Both Scikit learn and Stats models are used to build and Evaluate the models
2. Techniques like Recursive Feature Elimination is used to drop the features automatically which are highly related
3. With the help of Variance inflation factor (VIF) and p - values of the features, manual elimination was done having a cut-offs VIF < 3 and P-value < 0.05.

6.1Final Model Summary:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6294
Model:	GLM	**Df Residuals:**	6280
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2514.5
Date:	Sun, 12 Jun 2022	**Deviance:**	5029.0
Time:	11:44:50	**Pearson chi2:**	6.54e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

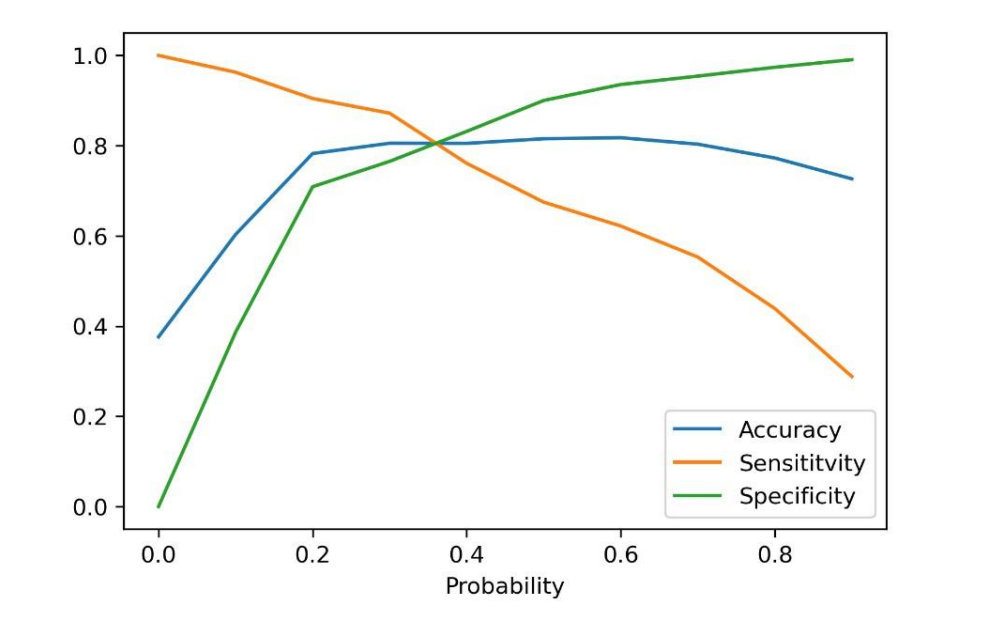
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.6684	0.077	-8.703	0.000	-0.819	-0.518
Do Not Email	-1.7302	0.183	-9.460	0.000	-2.089	-1.372
Total Time Spent on Website	3.8268	0.149	25.688	0.000	3.535	4.119
Lead Origin_Lead Add Form	3.1449	0.234	13.450	0.000	2.687	3.603
Lead Source_Welingak Website	3.3841	1.045	3.239	0.001	1.336	5.432
Last Activity_Had a Phone Conversation	2.4935	0.986	2.529	0.011	0.561	4.426
What is your current occupation_Working Professional	2.3844	0.192	12.440	0.000	2.009	2.760
Lead Profile_Potential Lead	1.7324	0.099	17.553	0.000	1.539	1.926
Lead Profile_Student of SomeSchool	-2.4804	0.557	-4.456	0.000	-3.571	-1.389
Last Notable Activity_Email Link Clicked	-1.6428	0.271	-6.070	0.000	-2.173	-1.112
Last Notable Activity_Email Opened	-1.4463	0.091	-15.936	0.000	-1.624	-1.268
Last Notable Activity_Modified	-2.0373	0.095	-21.526	0.000	-2.223	-1.852
Last Notable Activity_Olark Chat Conversation	-2.4379	0.349	-6.990	0.000	-3.121	-1.754
Last Notable Activity_Page Visited on Website	-1.5777	0.197	-7.999	0.000	-1.964	-1.191

	Probability	Accuracy	Sensititvity	Specificity
0.0	0.0	0.376072	1.000000	0.000000
0.1	0.1	0.603114	0.962822	0.386300
0.2	0.2	0.782491	0.904520	0.708938
0.3	0.3	0.805370	0.871990	0.765215
0.4	0.4	0.805052	0.760879	0.831678
0.5	0.5	0.815380	0.674694	0.900178
0.6	0.6	0.817604	0.621884	0.935574
0.7	0.7	0.803305	0.553021	0.954163
0.8	0.8	0.772641	0.438952	0.973771
0.9	0.9	0.726406	0.288128	0.990578

- Above table gives broad Description about Evaluation metrics like Accuracy, Sensitvity, Specificity for each Probability.
- Final Model with 13 Features
- Features like Total time spent on website, Lead source_welingal Website, Lead origin_lead Add form are positively related to the response variable.
- Features like Do not Email, Lead_profile Student of some school, Last Notable Activiti_Olark Chat Conversation etc. are Negatively Correlated to Response Variable.

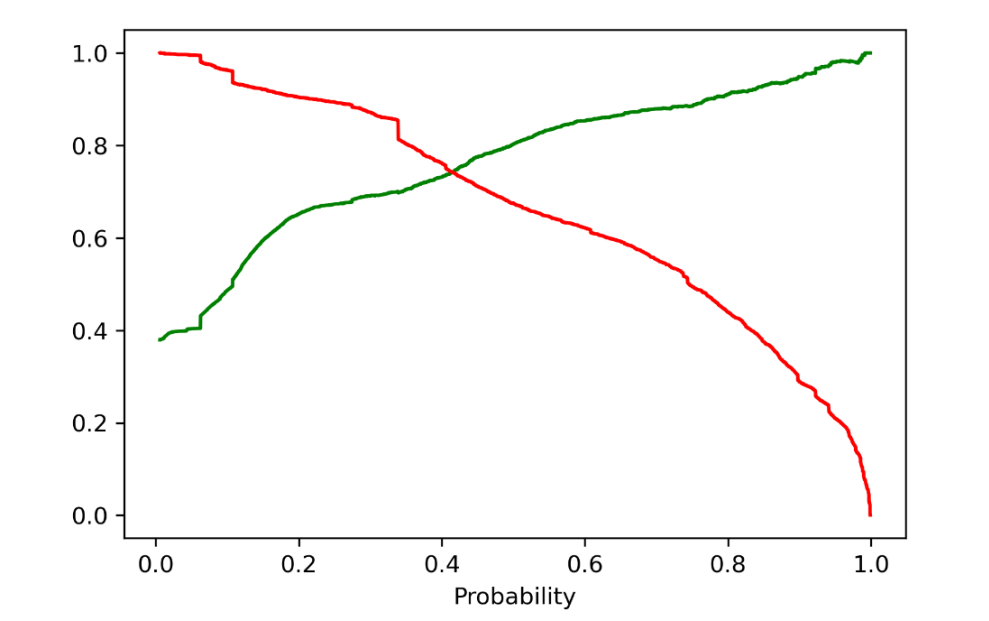
6.2 Optimal Cut-off:

1. Probability vs (Accuracy, Sensitivity and Specificity):



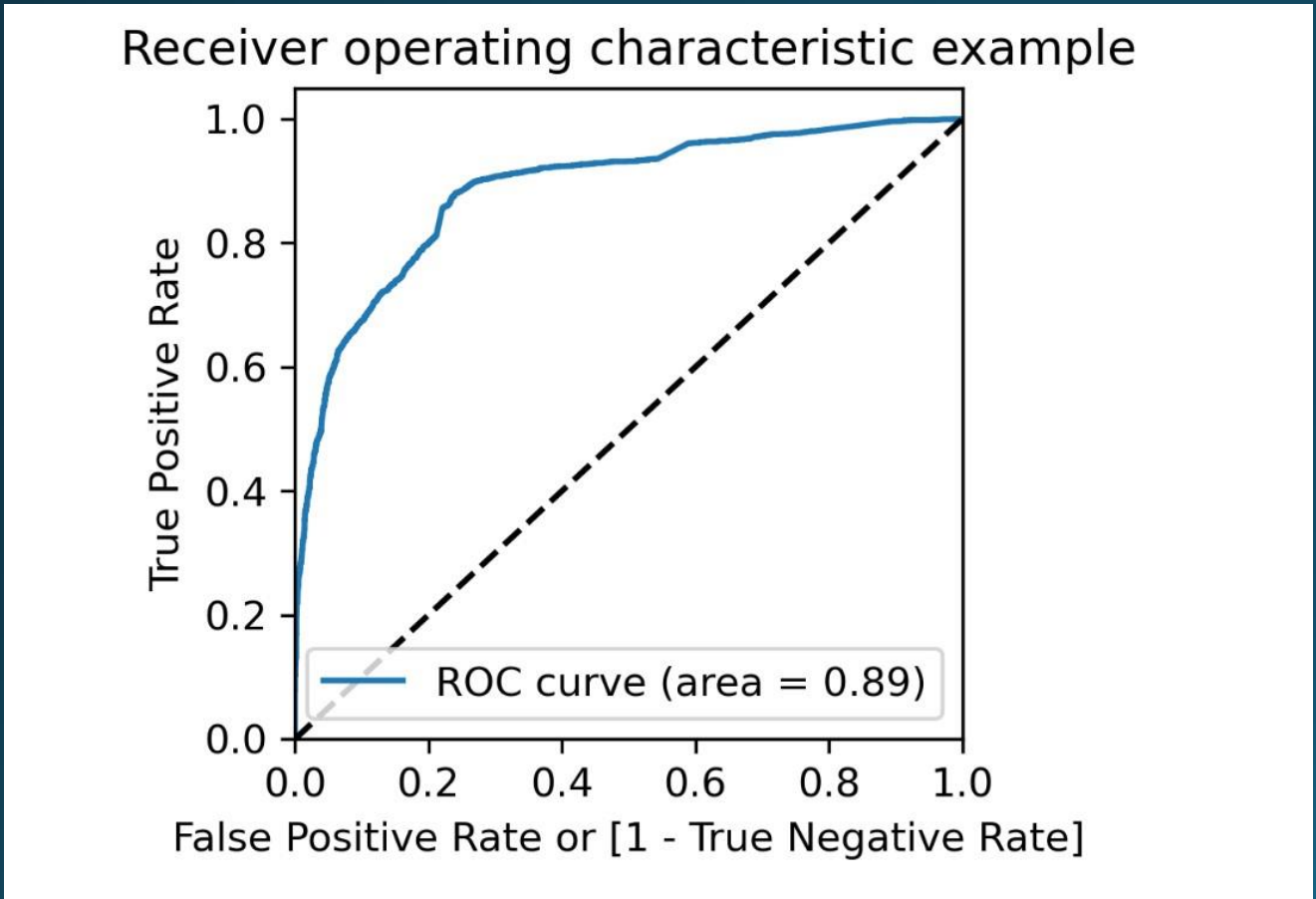
- It is Observed that Sensitivity and Specificity intersected at a probability of 0.4.

2.Precision Recall Trade off:



- The Above between Precision and Recall tells that probability of 0.4 is the Optimal cut-off for the model.

6.3 Receiver operating Characteristic Curve:



- Receiver Operating Characteristic Curve is Plot between True Positive Rate (Sensitivity) and False Positive rate.
- It measures the Efficiency of the model as the Curve Covers more area towards True Positive Rate.
- Area under the Curve (AUC) is 0.89 for the Mentioned plot.

6.4 Confusion Matrix and metrics on Training Data:

Actual/Predicted	Not Converted	Converted
Not Converted	3535	392
Converted	770	1597

1. True Negative = 3535
2. False Positive = 392
3. False Negative = 770
4. True Positive = 1597

Metric	Value
Accuracy	0.81
Sensitivity or Precision or True Positive rate	0.67
Specificity	0.9
False Positive rate	0.09
Positive Predicted value	0.8
Negative Predicted value	0.82

6.5 Confusion Matrix and metrics on Test Data:

Actual/Predicted	Not Converted	Converted
Not Converted	1650	14
Converted	776	258

Metric	Value
Accuracy	0.7
Sensitivity or Precision or True Positive rate	0.24
Specificity	0.99

7. Summary:

➢ Exploratory Data Analysis:

- It is Observed that the Lead Source graph "Google", “direct traffic”, “Organic search”, and “olark chat” have higher Counts. It is observed that most of the customer is from INDIA. Major conversion in the lead source from google.
- Most of the customers are unemployed, choose courses for “Better Career Prospects” and they are from "Mumbai".
- The total conversion rate of customers is 37.82%
- Leading Page Submission from Lead Origin are Having More Total Time Spent on Website, Total Visits and Page Views Per Visit compared to other Category Levels and Customers who took the Course spent more time on the website.

➢ Model Building and Evaluation:

- Final Model with 13 Features. Features like Total time spent on the website, Lead source_welingal Website and Lead origin_lead Add form are positively related to the response variable. Features like Do not Email, Lead_profile Student of some school, Last Notable Activiti_Olark Chat Conversation, etc. are Negatively Correlated to Response Variable.
- It is Observed that Sensitivity and Specificity intersected at a probability of 0.4 and Precision and Recall also tell that a probability of 0.4 is the Optimal cut-off for the model.
- The area under the Curve (AUC) is 0.89 which is Obtained from Receiver Operating Characteristic Curve.