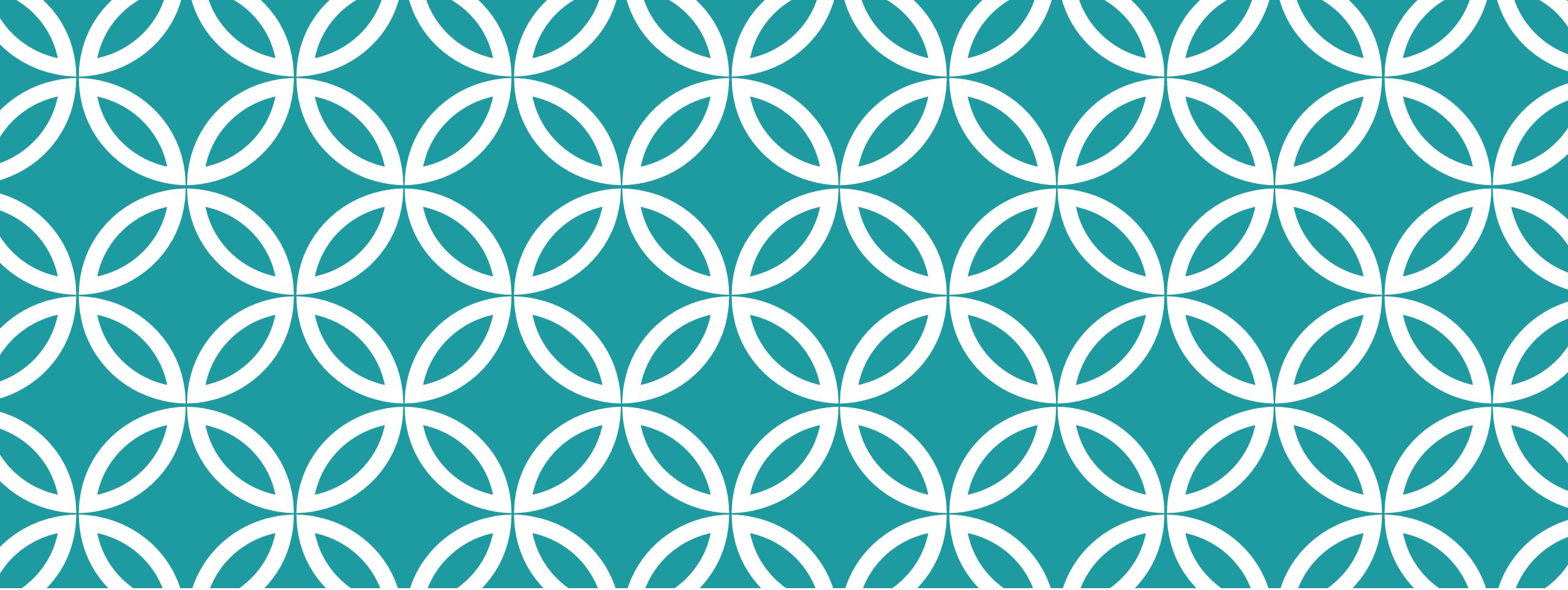
LEAD SCORING CASE STUDY FOR X EDUCATION

Prepared by –
Myron Cardoso
Baharika Sopori
Ayan Pramanik



BUSINESS PROBLEM AND GOAL

BUSINESS PROBLEM

X Education sells online courses to industry professionals.

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.

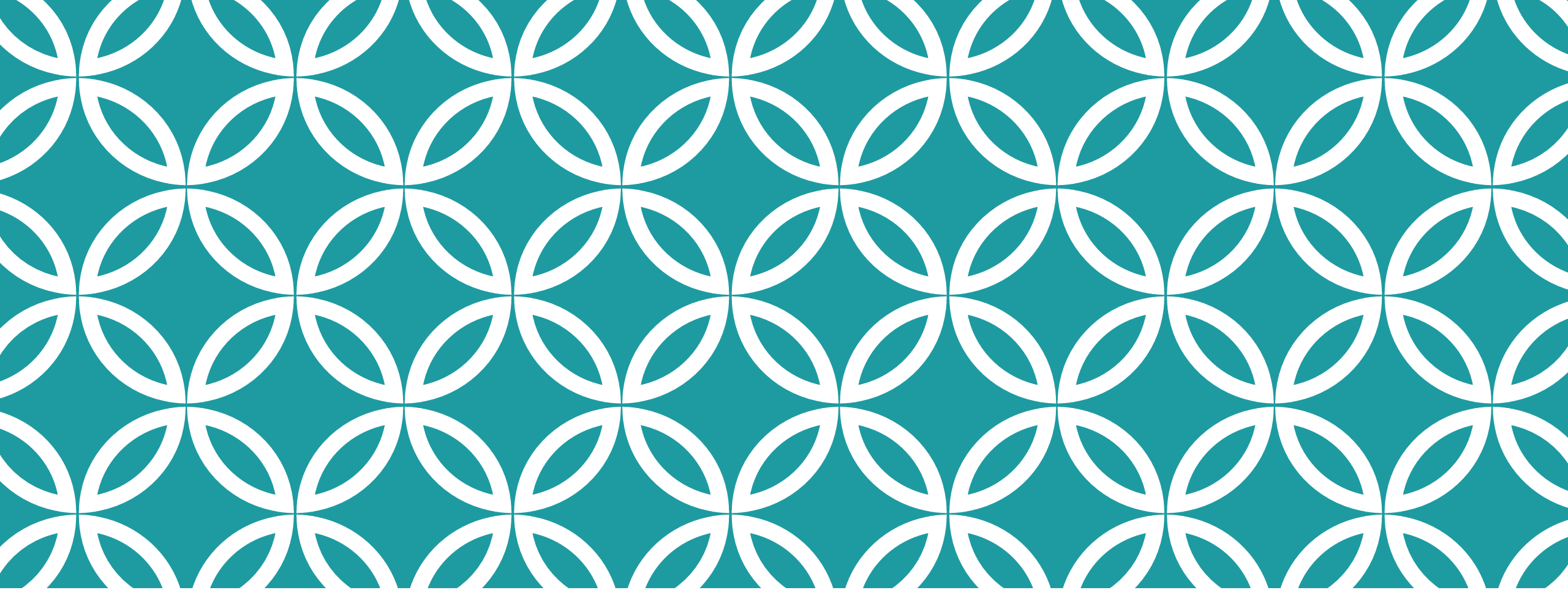
Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

BUSINESS GOAL

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company needs a model wherein you a lead score is assigned 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%.



PROBLEM SOLVING MINDMAP AND STRATEGY

STEPS TAKEN

1. Import necessary libraries.
2. Loading and Analyzing the data.
3. Data Cleaning –
    1. Checking Datatypes
    2. Handling Outliers
    3. Handling Missing Values
4. Exploratory Data Analysis
    1. Univariate Analysis
    2. Bivariate Analysis
    3. Multivariate Analysis
5. Data Preparation for Feature Scaling
6. Train – Test Split
7. Model Building (Linear Regression)
8. Model Evaluation

Data Sourcing, Cleaning and Preparation

- Read Data from the source.
- Clean Data to make it suitable for analysis
- Exploratory Data Analysis
- Feature Standardization

Feature Scaling & Train-Test Split

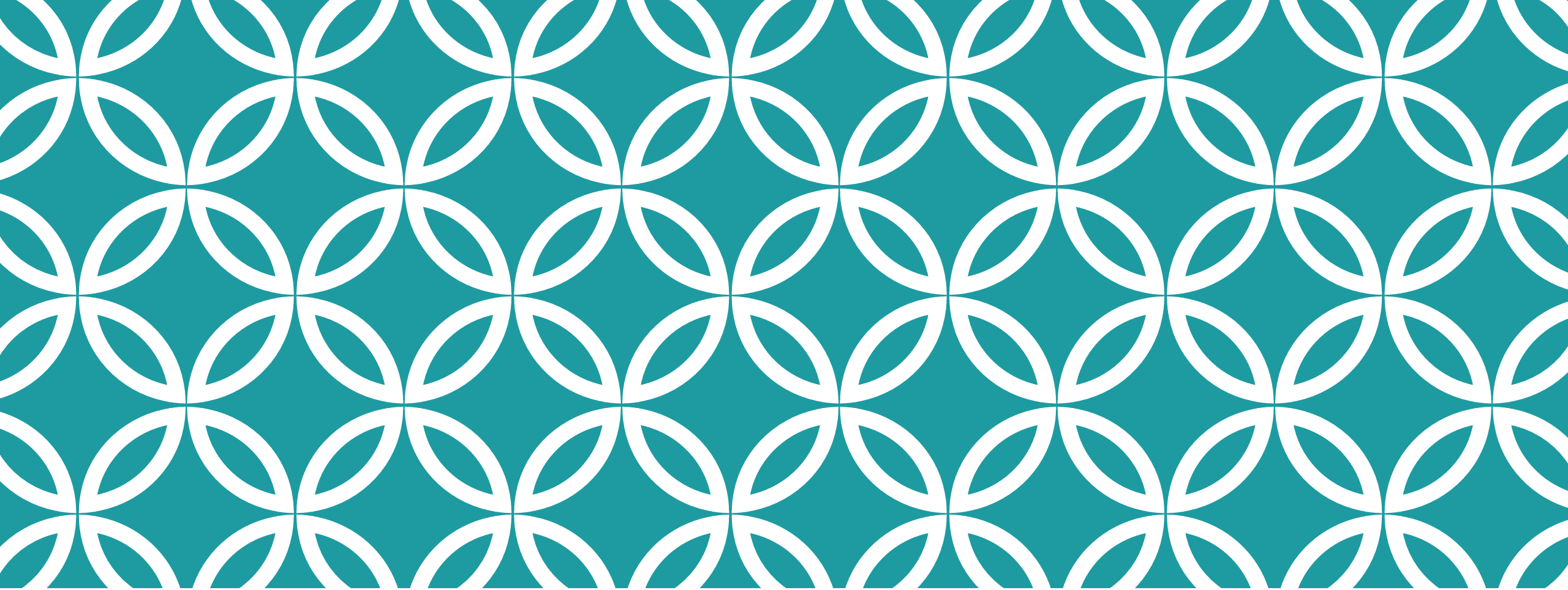
- Feature Scaling of Numeric Data
- Splitting data into Train and Test

Model Building

- Feature selection using RFE
- Optimal model designing using Logistic Regression
- Calculate
  - Accuracy
  - Sensitivity
  - Specificity
  - Precision
  - Recall
- Evaluate the model

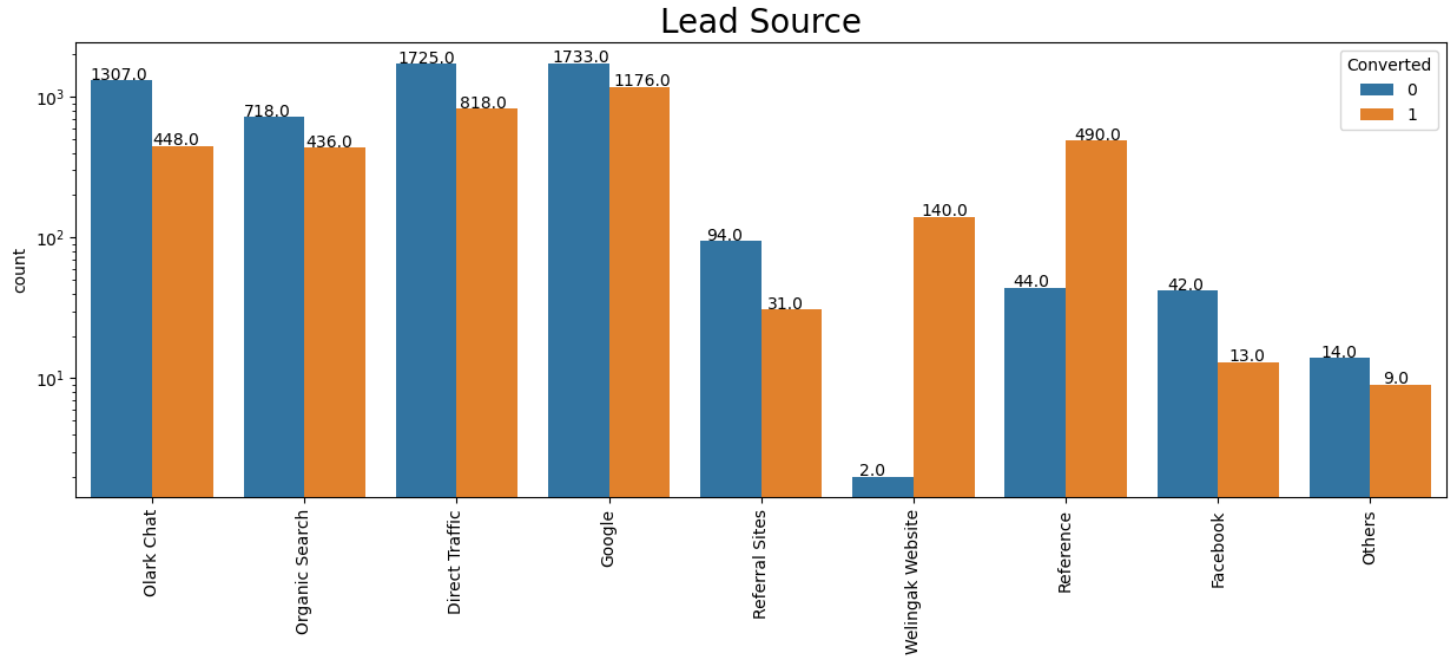
Result

- Determine the lead score and check if final prediction amounts to 80% conversion rate.
- Evaluate Final Prediction using Sensitivity and Specificity Metrics



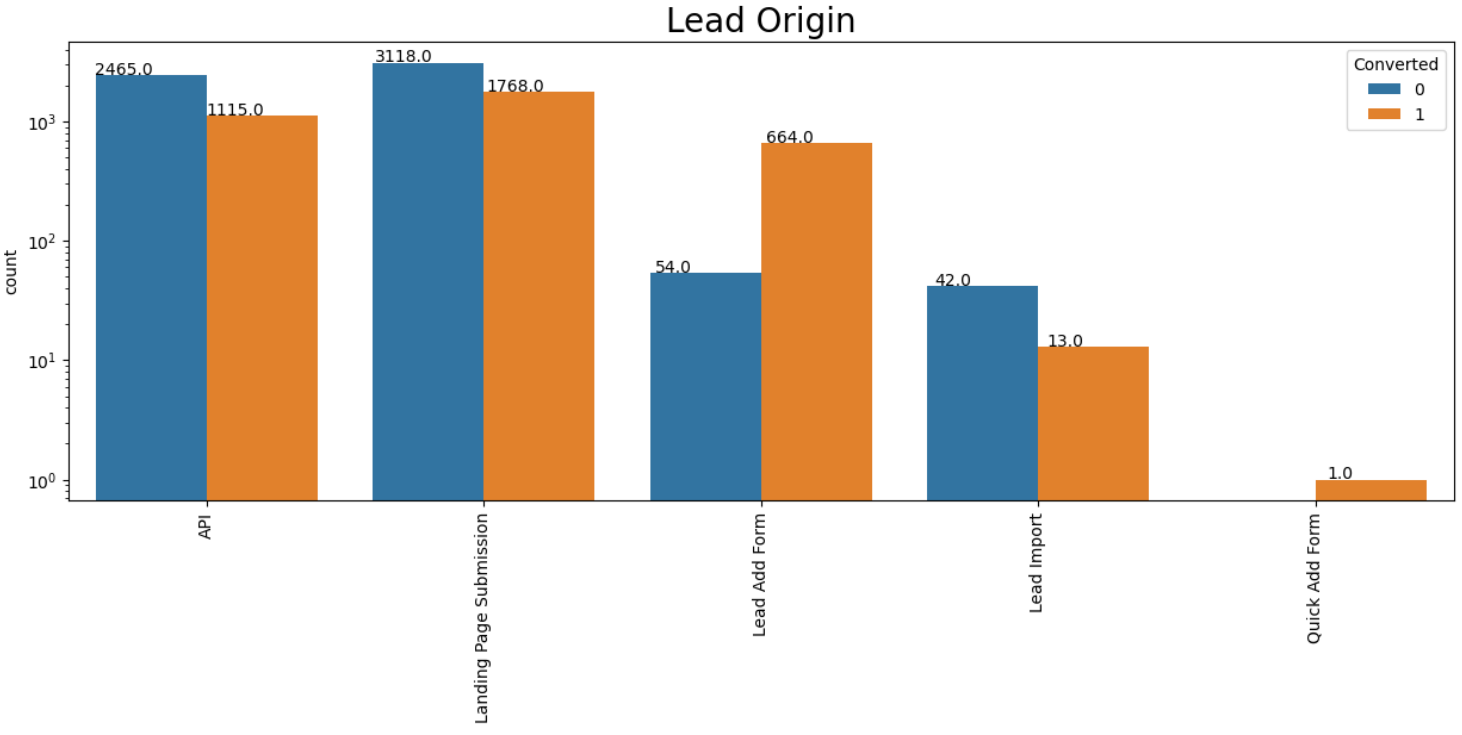
EXPLORATORY DATA ANALYSIS

Univariate Analysis



- **Top Generators:** Google and Direct traffic.

- **High Conversion Rates:** 'Reference' and 'Welingak Website'.

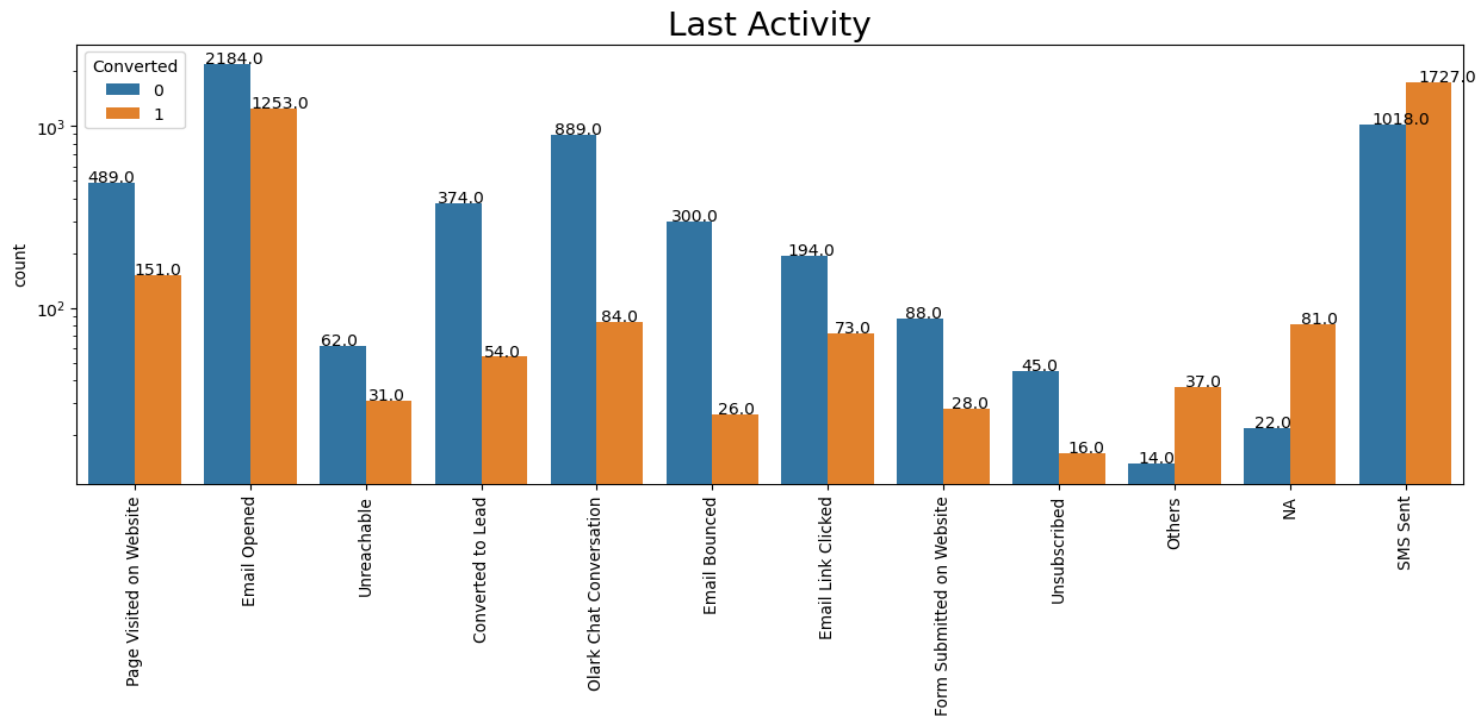


- **API Conversion Rate:** ~31%.

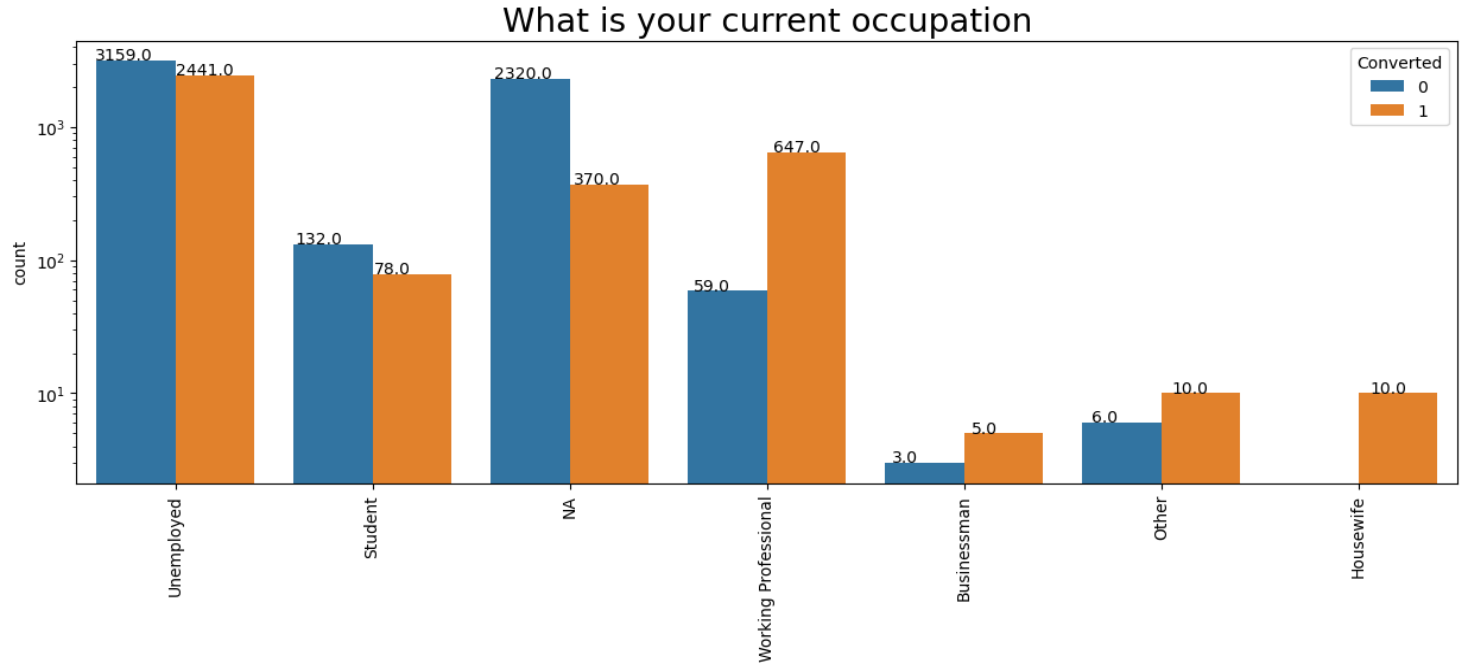
- **Landing Page Submission Conversion Rate:** ~36%.

- **Lead Add Form:** More successful conversions than unsuccessful.

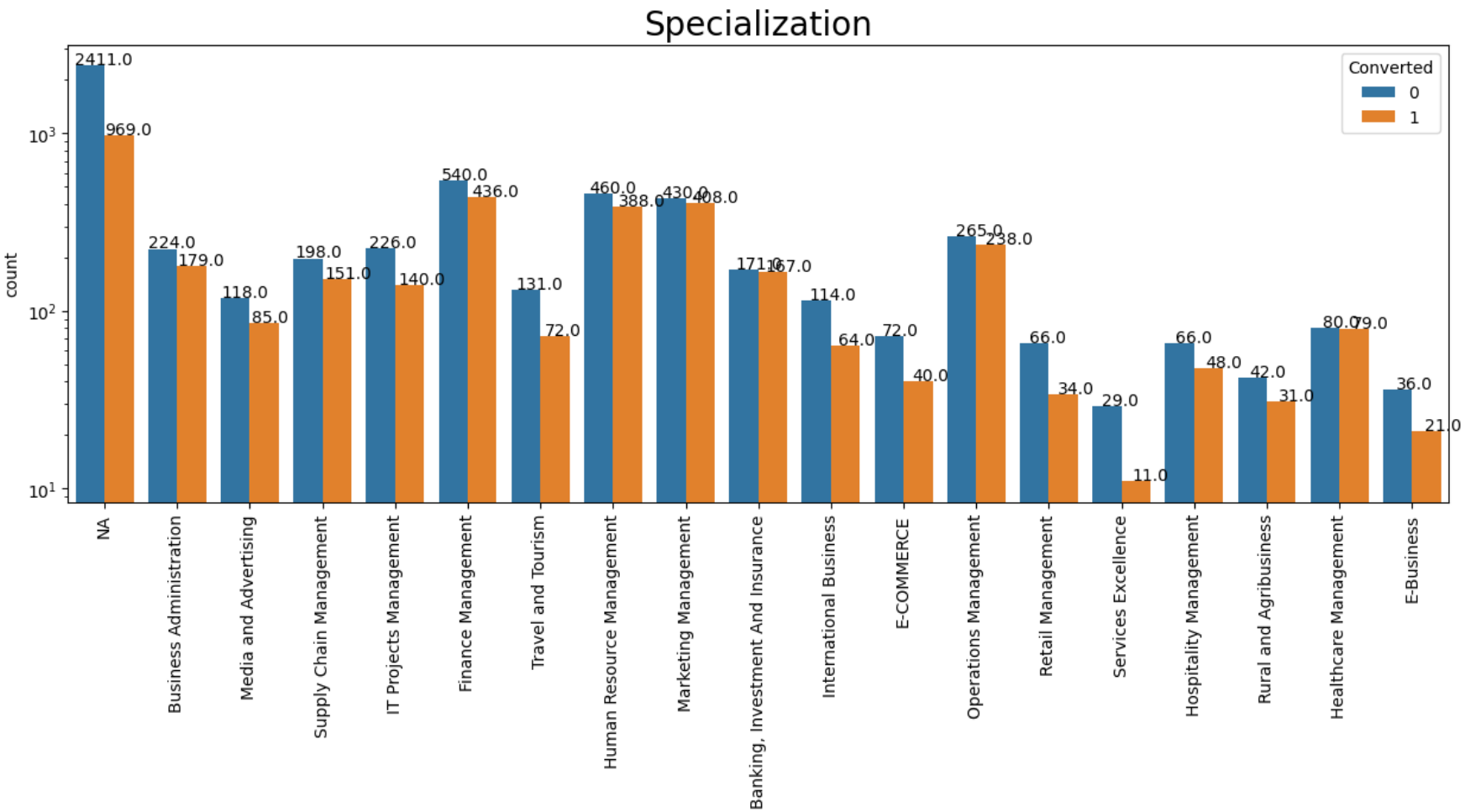
- **Lead Import:** Low count.



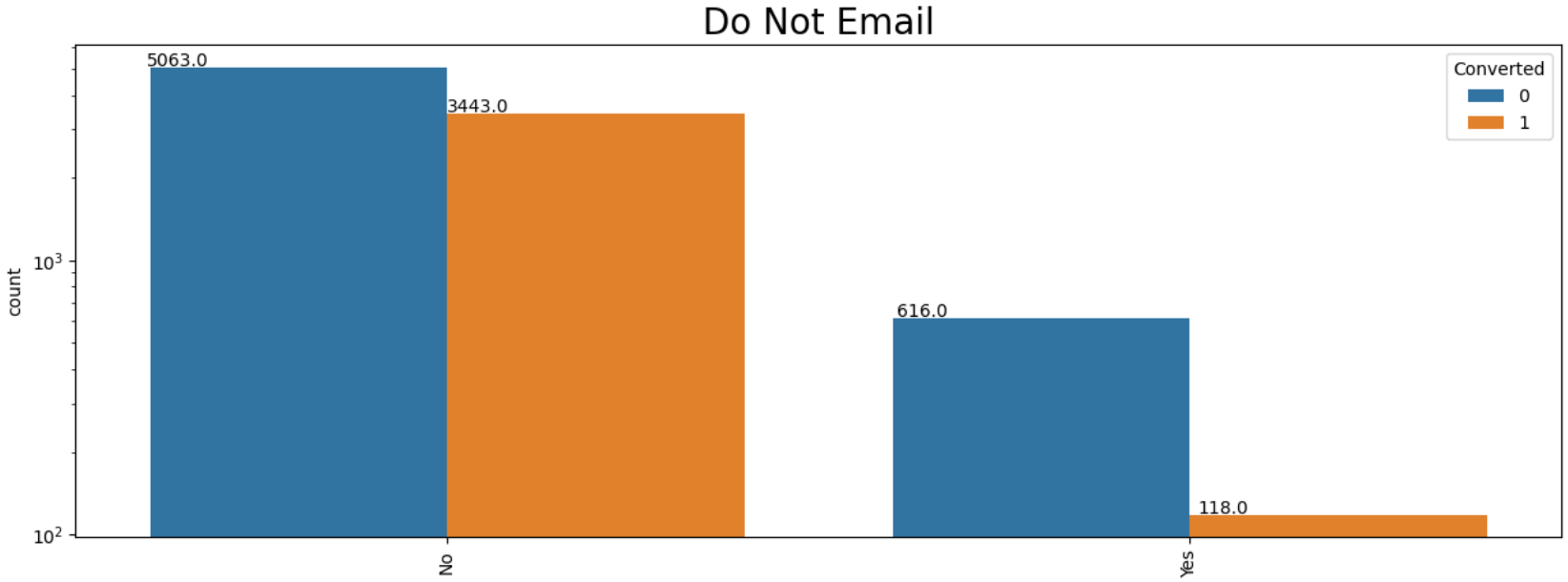
- **SMS Sent:** ~63% conversion rate.
- **Email Opened:** Most common last activity.

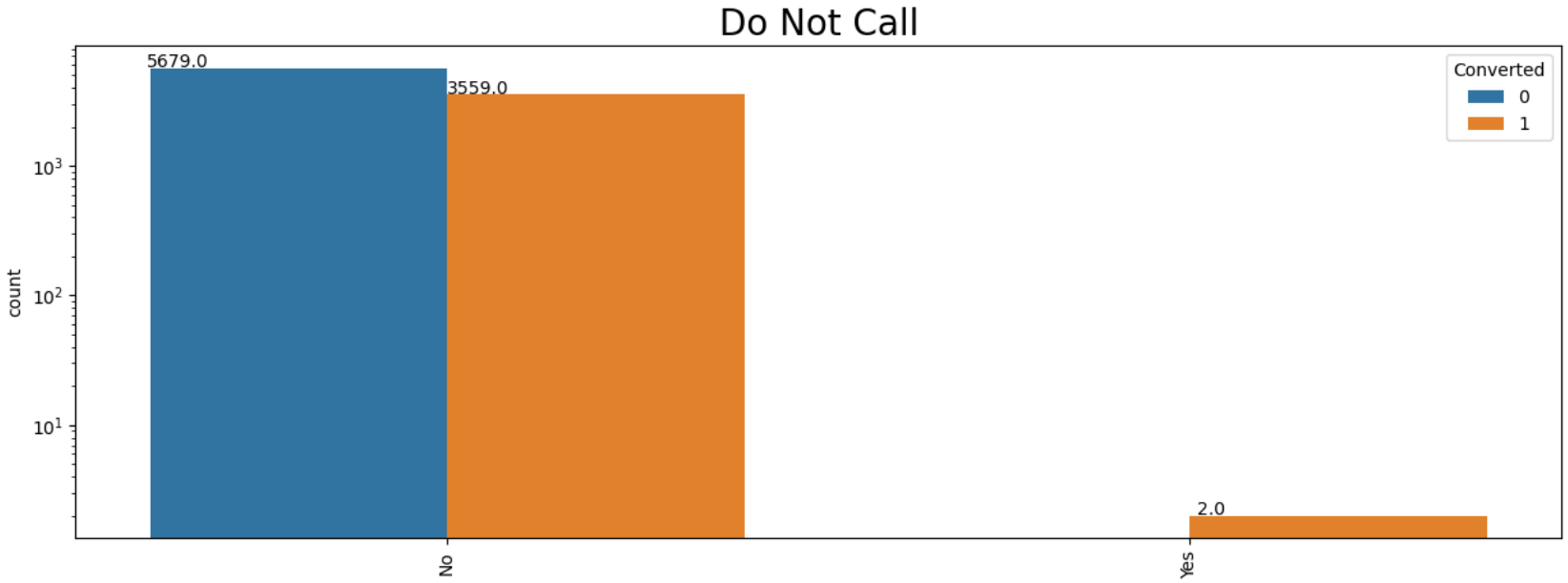


- **Unemployed Leads:** Generate more leads and have a ~45% conversion rate.
- **Working Professionals:** Higher conversion rate.



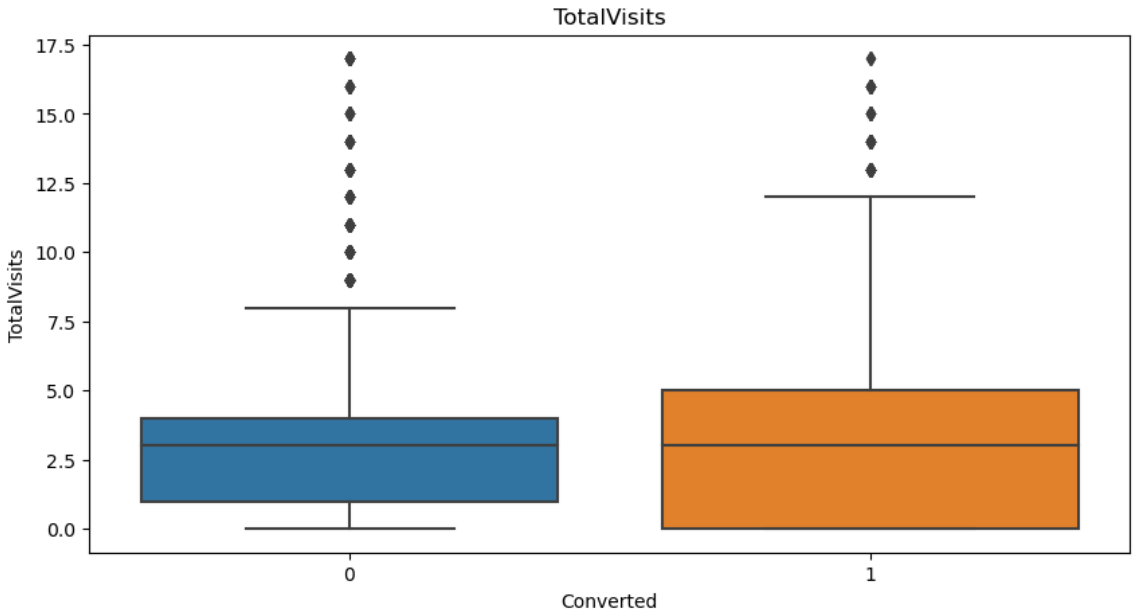
- **Management:** Generates the most leads.
- **NA Category:** Also generates a significant number of leads.

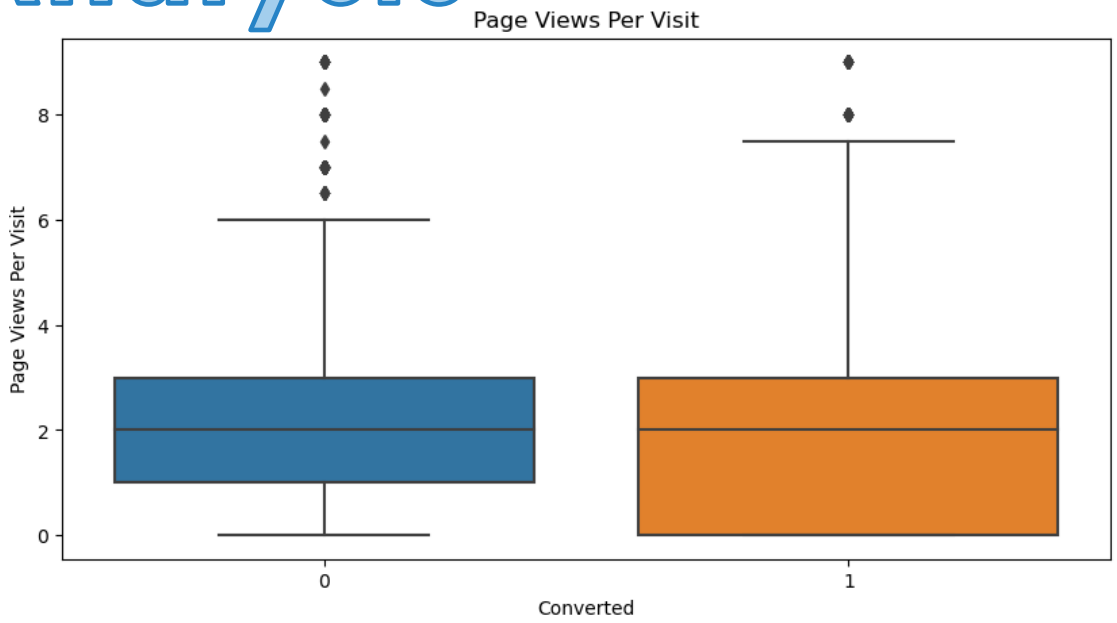


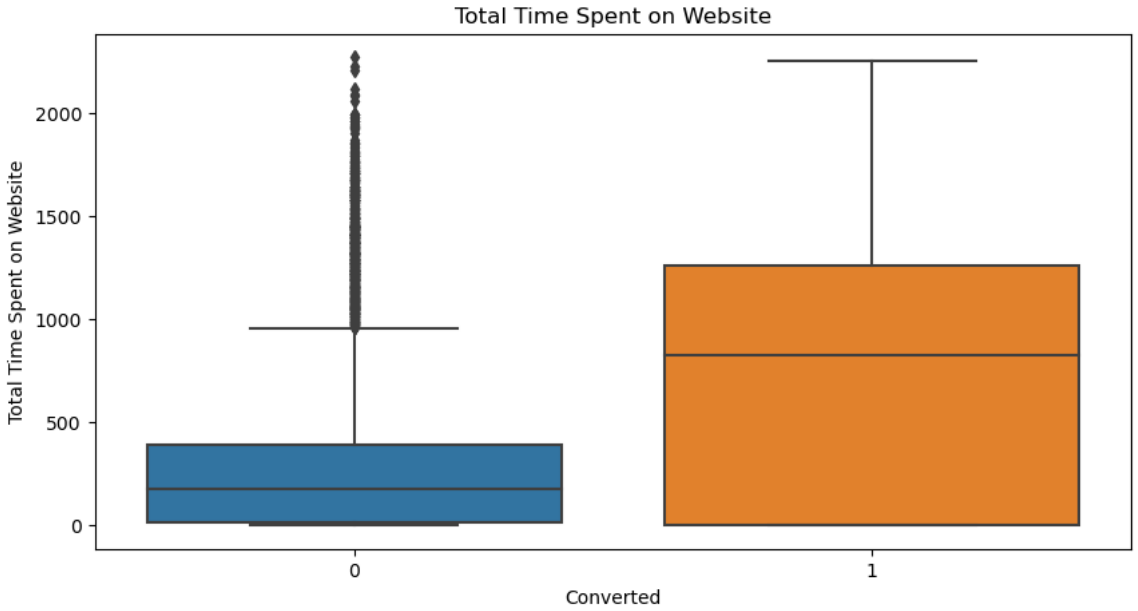


•**Maximum Sales –** Achieved through Over the Call Sales and Mail.

Bivariate Analysis







TotalVisits and Conversion: Median values for 'TotalVisits' are similar regardless of conversion status.

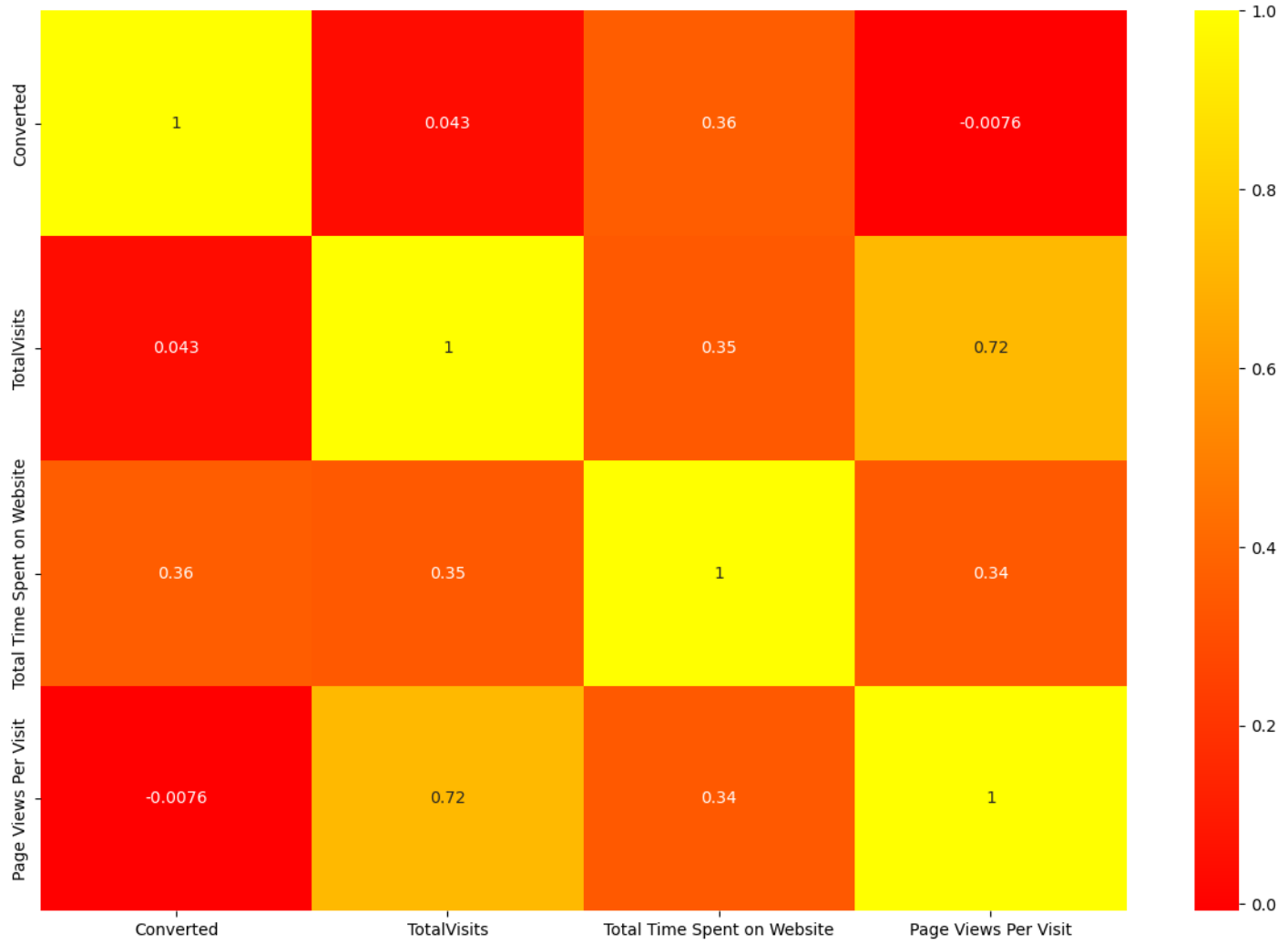
- Total Time Spent on Website: Leads spending more time on the website tend to convert more frequently.

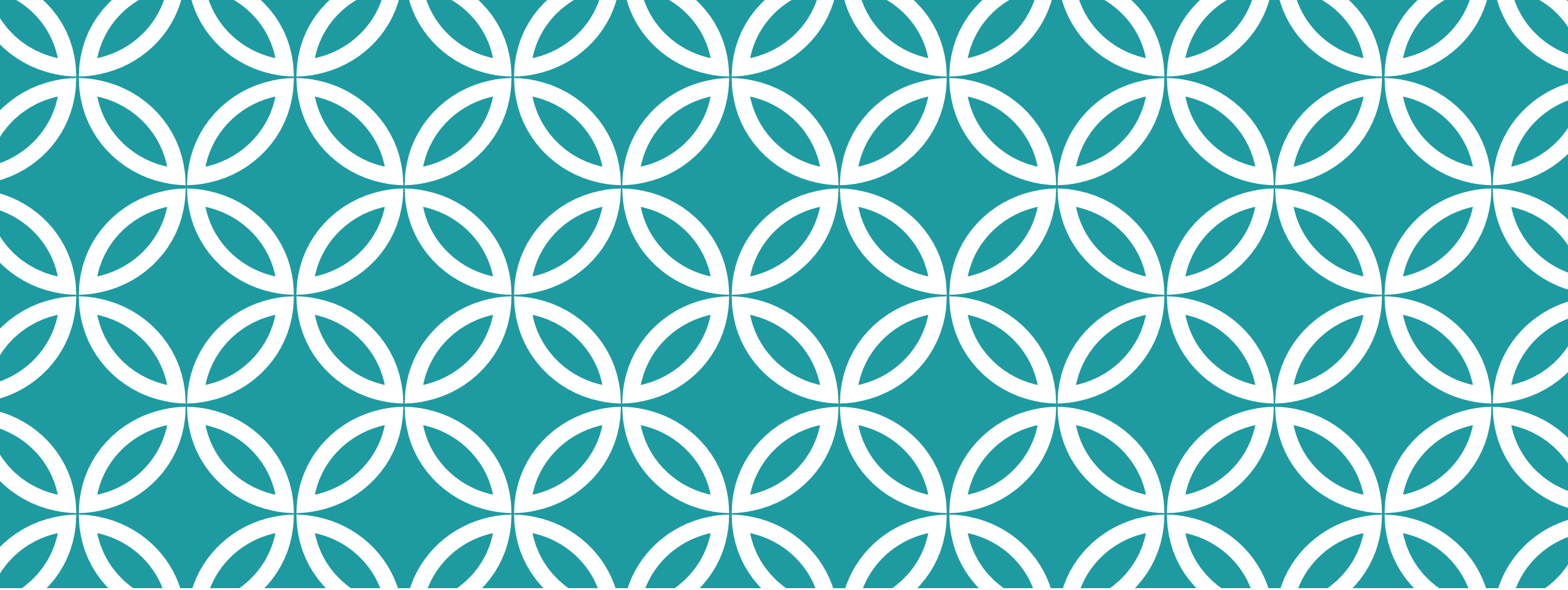
- Page Views per Visit: Similar median values for converted and non-converted leads suggest page views per visit have minimal impact on conversions.

Multivariate Analysis

•'TotalVisits' and 'Page Views per Visit': Highly correlated at 0.72, indicating a strong relationship.

•'Total Time Spent on Website': Moderately correlated at 0.36 with 'Converted', suggesting a potential influence on conversion.





MODEL BUILDING

1. Data Preparation Steps
    1. Converted binary variables (Yes/No) to 0 and 1 for model building.
    2. Created dummy variables for all category columns.
2. Train-test Split
    1. Split the data into train and test dataframe using 70-30% ratio respectively.
    2. At this stage we have to import train-test-split library from sklearn.
3. Model Building
    1. Build 1st Linear Regression model using all features.
    2. To build the best fit model, we used Recursive Feature Elimination (RFE) technique, to get the top 15 features to build our next model.
    3. For each model build we have to check for p-value, it should always be less than 0.05.
    4. To remove Multi-collinearity, calculated Variance Inflation Factor (VIF), to check if feature variables are nor correlated to each other.
    5. Drop the features which have high p-value and highly correlated on by one and recursively build the model to get the optimal model.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6363
Model:	GLM	**Df Residuals:**	6347
Model Family:	Binomial	**Df Model:**	15
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2533.8
Date:	Tue, 18 Jun 2024	**Deviance:**	5067.5
Time:	15:24:08	**Pearson chi2:**	6.69e+03
No. Iterations:	22	**Pseudo R-squ. (CS):**	0.4133
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9423	0.061	-15.481	0.000	-1.062	-0.823
Lead Origin_Lead Add Form	3.4171	0.198	17.275	0.000	3.029	3.805
Lead Source_Olark Chat	1.1977	0.104	11.509	0.000	0.994	1.402
Lead Source_Welingak Website	2.7390	1.026	2.669	0.008	0.728	4.750
Last Activity_Email Bounced	-1.5617	0.338	-4.627	0.000	-2.223	-0.900
Last Activity_Others	1.0767	0.594	1.812	0.070	-0.088	2.242
Last Activity_SMS Sent	1.2195	0.076	16.082	0.000	1.071	1.368
What is your current occupation_Housewife	23.9007	2.32e+04	0.001	0.999	-4.55e+04	4.56e+04
What is your current occupation_Working Professional	2.5493	0.190	13.418	0.000	2.177	2.922
What matters most to you in choosing a course_NA	-1.2458	0.089	-13.995	0.000	-1.420	-1.071
Asymmetrique Activity Index_03.Low	-1.8833	0.276	-6.814	0.000	-2.425	-1.342
Last Notable Activity_Had a Phone Conversation	22.9338	2.37e+04	0.001	0.999	-4.64e+04	4.64e+04
Last Notable Activity_Modified	-0.8408	0.080	-10.464	0.000	-0.998	-0.683
Last Notable Activity_Olark Chat Conversation	-1.3031	0.327	-3.987	0.000	-1.944	-0.663
Last Notable Activity_Unreachable	2.1128	0.621	3.403	0.001	0.896	3.330
Total Time Spent on Website	1.1056	0.041	27.021	0.000	1.025	1.186

	Features	VIF
0	Lead Origin_Lead Add Form	1.51
11	Last Notable Activity_Modified	1.48
1	Lead Source_Olark Chat	1.46
8	What matters most to you in choosing a course_NA	1.43
4	Last Activity_Others	1.34
10	Last Notable Activity_Had a Phone Conversation	1.32
2	Lead Source_Welingak Website	1.29
14	Total Time Spent on Website	1.23
5	Last Activity_SMS Sent	1.20
7	What is your current occupation_Working Profes...	1.17
3	Last Activity_Email Bounced	1.08
12	Last Notable Activity_Olark Chat Conversation	1.07
9	Asymmetrique Activity Index_03.Low	1.04
6	What is your current occupation_Housewife	1.00
13	Last Notable Activity_Unreachable	1.00

Model 1

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6363
Model:	GLM	**Df Residuals:**	6349
Model Family:	Binomial	**Df Model:**	13
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2543.8
Date:	Tue, 18 Jun 2024	**Deviance:**	5087.6
Time:	15:24:08	**Pearson chi2:**	6.80e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.4115
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9276	0.061	-15.304	0.000	-1.046	-0.809
Lead Origin_Lead Add Form	3.4243	0.197	17.338	0.000	3.037	3.811
Lead Source_Olark Chat	1.1876	0.104	11.433	0.000	0.984	1.391
Lead Source_Welingak Website	2.7207	1.026	2.651	0.008	0.710	4.732
Last Activity_Email Bounced	-1.5680	0.337	-4.646	0.000	-2.230	-0.907
Last Activity_Others	1.7073	0.520	3.284	0.001	0.688	2.726
Last Activity_SMS Sent	1.2096	0.076	15.971	0.000	1.061	1.358
What is your current occupation_Working Professional	2.5321	0.190	13.348	0.000	2.160	2.904
What matters most to you in choosing a course_NA	-1.2543	0.089	-14.099	0.000	-1.429	-1.080
Asymmetrique Activity Index_03.Low	-1.8945	0.276	-6.866	0.000	-2.435	-1.354
Last Notable Activity_Modified	-0.8455	0.080	-10.553	0.000	-1.002	-0.688
Last Notable Activity_Olark Chat Conversation	-1.3080	0.327	-4.003	0.000	-1.948	-0.668
Last Notable Activity_Unreachable	2.1037	0.621	3.388	0.001	0.887	3.321
Total Time Spent on Website	1.1034	0.041	27.025	0.000	1.023	1.183

	Features	VIF
0	Lead Origin_Lead Add Form	1.50
9	Last Notable Activity_Modified	1.48
1	Lead Source_Olark Chat	1.46
7	What matters most to you in choosing a course_NA	1.43
2	Lead Source_Welingak Website	1.29
12	Total Time Spent on Website	1.23
5	Last Activity_SMS Sent	1.20
6	What is your current occupation_Working Profes...	1.17
3	Last Activity_Email Bounced	1.08
10	Last Notable Activity_Olark Chat Conversation	1.07
8	Asymmetrique Activity Index_03.Low	1.04
4	Last Activity_Others	1.01
11	Last Notable Activity_Unreachable	1.00

Model 2

Generalized Linear Model Regression Results

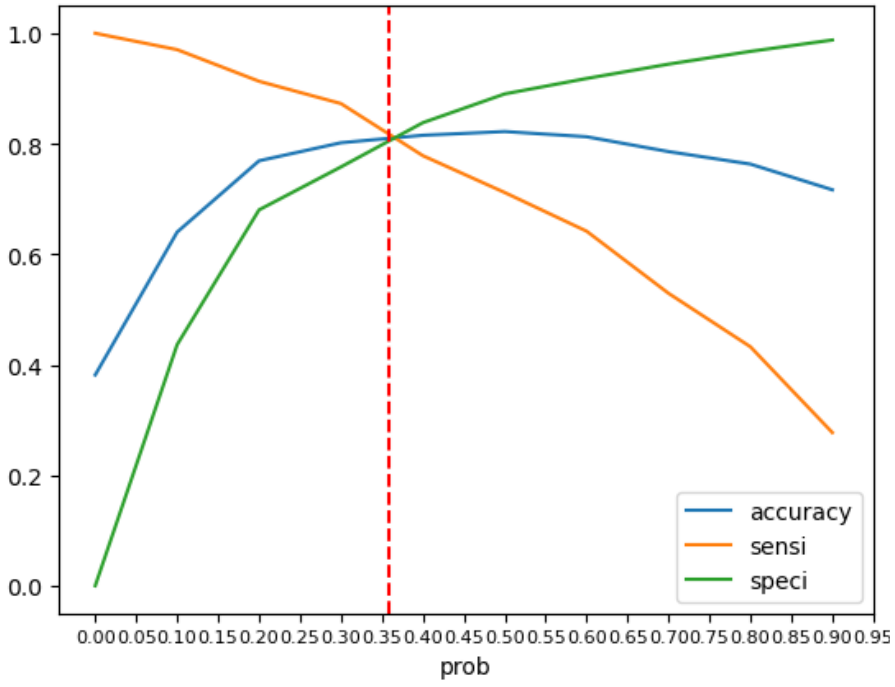
Dep. Variable:	Converted	**No. Observations:**	6363
Model:	GLM	**Df Residuals:**	6352
Model Family:	Binomial	**Df Model:**	10
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2564.0
Date:	Tue, 18 Jun 2024	**Deviance:**	5128.1
Time:	15:24:08	**Pearson chi2:**	6.81e+03
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.4077
Covariance Type:	nonrobust		

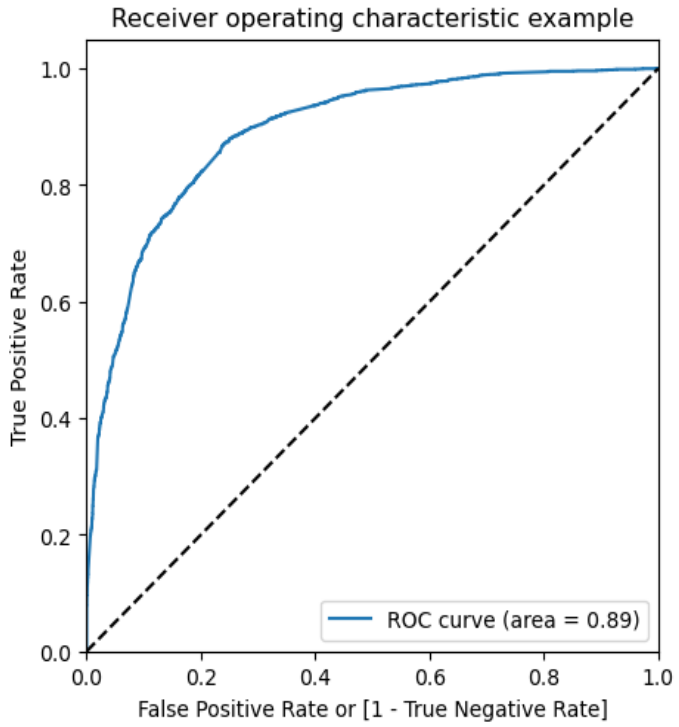
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8990	0.060	-14.964	0.000	-1.017	-0.781
Lead Origin_Lead Add Form	3.7538	0.190	19.724	0.000	3.381	4.127
Lead Source_Olark Chat	1.1721	0.104	11.321	0.000	0.969	1.375
Last Activity_Email Bounced	-1.5816	0.335	-4.720	0.000	-2.238	-0.925
Last Activity_SMS Sent	1.1854	0.075	15.738	0.000	1.038	1.333
What is your current occupation_Working Professional	2.5450	0.191	13.359	0.000	2.172	2.918
What matters most to you in choosing a course_NA	-1.2448	0.089	-14.054	0.000	-1.418	-1.071
Asymmetrique Activity Index_03.Low	-1.9011	0.277	-6.869	0.000	-2.444	-1.359
Last Notable Activity_Modified	-0.8545	0.080	-10.730	0.000	-1.011	-0.698
Last Notable Activity_Olark Chat Conversation	-1.3332	0.327	-4.076	0.000	-1.974	-0.692
Total Time Spent on Website	1.1024	0.041	27.074	0.000	1.023	1.182

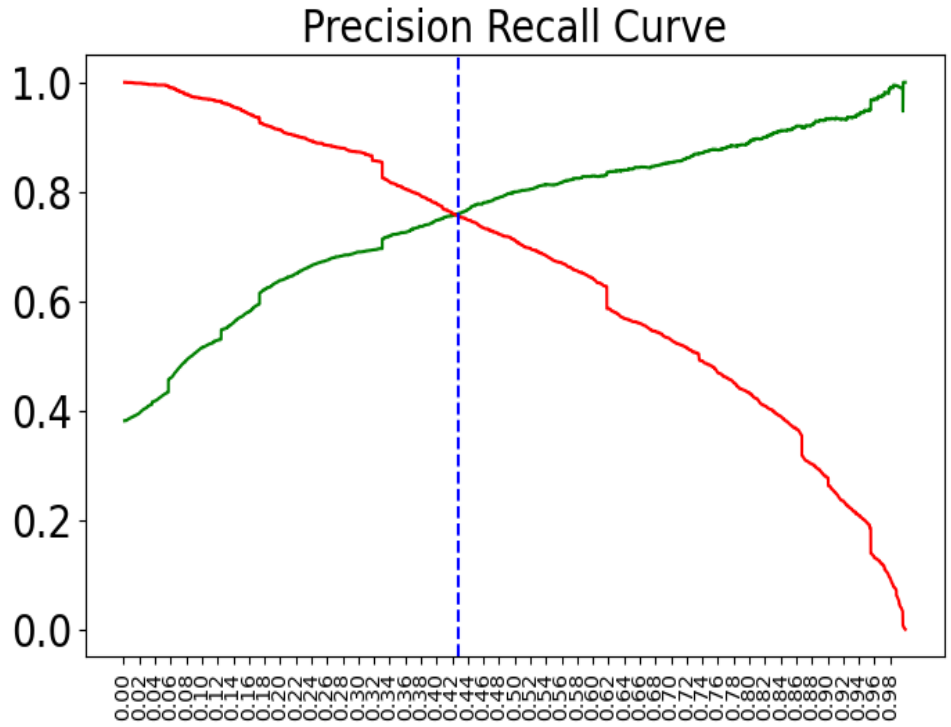
	Features	VIF
7	Last Notable Activity_Modified	1.47
1	Lead Source_Olark Chat	1.46
5	What matters most to you in choosing a course_NA	1.42
9	Total Time Spent on Website	1.22
0	Lead Origin_Lead Add Form	1.20
3	Last Activity_SMS Sent	1.20
4	What is your current occupation_Working Profes...	1.14
2	Last Activity_Email Bounced	1.08
8	Last Notable Activity_Olark Chat Conversation	1.07
6	Asymmetrique Activity Index_03.Low	1.04

Model 3

4. Model Evaluation –
    1. After getting optimal model, evaluate performance metrics score, Accuracy, Recall, Precision and Precision and Recall Score.
    2. ROC curve is then plotted, that shows the tradeoff between specificity and sensitivity (Both are inversely co-related)
        1. The closer the curve follows the left-hand border and top border of the ROC space, the more accurate the test.
        2. The closer the curve comes to the 45 Degree diagonal of the ROC curve, the less accurate the test.
    3. Calculate the cutoff point between accuracy, specificity, sensitivity. (Trade-off = 0.35)
    4. Plotting the Precision and Recall tradeoff as this will help us to identify the predicted converted versus the actual converted. (Trade-off = 0.427)







Model Evaluation – Train Data

Evaluation Metrics for the train Dataset:-
- Accuracy: 81.01%
- Sensitivity (Recall): 80.72%
- Specificity: 81.19%
- Precision: 72.59%
- Recall: 80.72%

Model Evaluation – Test Data

Evaluation Metrics for the test Dataset:-
- Accuracy: 79.72%
- Sensitivity (Recall): 78.63%
- Specificity: 80.42%
- Precision: 72.07%
- Recall: 78.63%

RECOMMENDATIONS

Conversion of the below is much easier and should be leveraged -

- **Lead Origin_Lead Add Form** - The leads produced from this source have been found to opt for the course most of the time.
- **What is your current occupation_Working Professional** - Working Professionals should be targeted for the sale of courses.
- **Last Activity_SMS Sent** - The leads converted mostly reach out via sms communication chain and are potential leads.
- **Lead Source_Olark Chat** - Leads sourced from Olark Chat have more conversion potential.
- **Total Time Spent on Website** - Higher the time spent on the website, results in higher conversion of the lead.