Lead Scoring Case Study

June 17, 2024

By
Harsh Agrawal, Aishwarya Girhare, Abhilash Siddaramareddy

Business Understanding

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, 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.

Problem Statement

- X education wants to know most promising leads.
- To identify the **hot leads** , we want to build a ML Model.
- The model should be flexible enough to accommodate changes in the data and ready for the future use.

Primary Goals:

- To build a **logistic regression model to assign a lead score** between 0 and 100 to each of the leads. This score determines the target potential leads for the company & Sales team.
- Build a flexible model in order to **handle the changes as per the requirements** in the future.

Steps

1. Data Sourcing & Loading

Data is already provided in the assignment.

Download and load the data into the dataframe.

2. Data Cleaning

Fixing rows and columns, like datatype, units, etc

Removing and Imputing the missing data in the columns

Handling Outliers

Binning

3. EDA

Univariate Analysis

- Categorical
- Numerical

Bivariate Analysis

- Correlation
- Numerical-Numerical
- Categorical-Numerical
- Categorical-Categorical

Contd..

Steps

4. Data Preparation

Preparations of Data for Model Building

Feature Scaling

Dummy Variables

Encoding of Data

5. Model Building

Use of Logistic Regression for Model development.

Building the ML Model using Statistics, RFE and VIF

6. Model Evaluation

Confusion Matrix

Overall Accuracy

Precision and Recall Tradeoff

7. Model Predictions

Identifying the top features

Final model line equation

Recommendations

Data Cleaning

Data Cleaning steps

1

- Total Rows: 9240, Total Columns: 37
- The dataset's shape remains unchanged before and after removing duplicates, indicating there are no duplicate values.

2

- Total **7 numeric columns** and **30 categorical columns.**
- None of the columns have inconsistent datatype. Hence, no conversion is required.

3

- Apart from actual **NULL** values, there are values that are marked as **Select** , which states that *Student had not selected the option for particular question/column.*
- These values are considered as missing values and have been replaced with **NULL** values.

Data Cleaning steps

4	• Columns with **30%+ NULL** Values: **10 columns** • **Dropped 8 columns** out of them as they are created by sales team after the follow up.
5	• **City (~40%NULL)** : Imputing this will make the data biased. So, dropping this as well. • Similarly, **Country** column also can be dropped.
6	• There are **12 columns** , in which only one value was majorly present for all data points. • Since, practically all of the values for these variables are **same**, it is best that we drop these columns as they won't help with our analysis.
7	• **Prospect ID** and **Lead Number** both do not have duplicate values. Means, these columns identify each data uniquely as unique Ids and will not make any significant impact on our model. So, it is better to drop these.

Data Analysis

Univariate Analysis

Imbalance

As per the problem statement **Converted** is our target variable.

The target variable indicates whether a lead has been successfully converted or not.

0: Not converted

1: Successfully converted

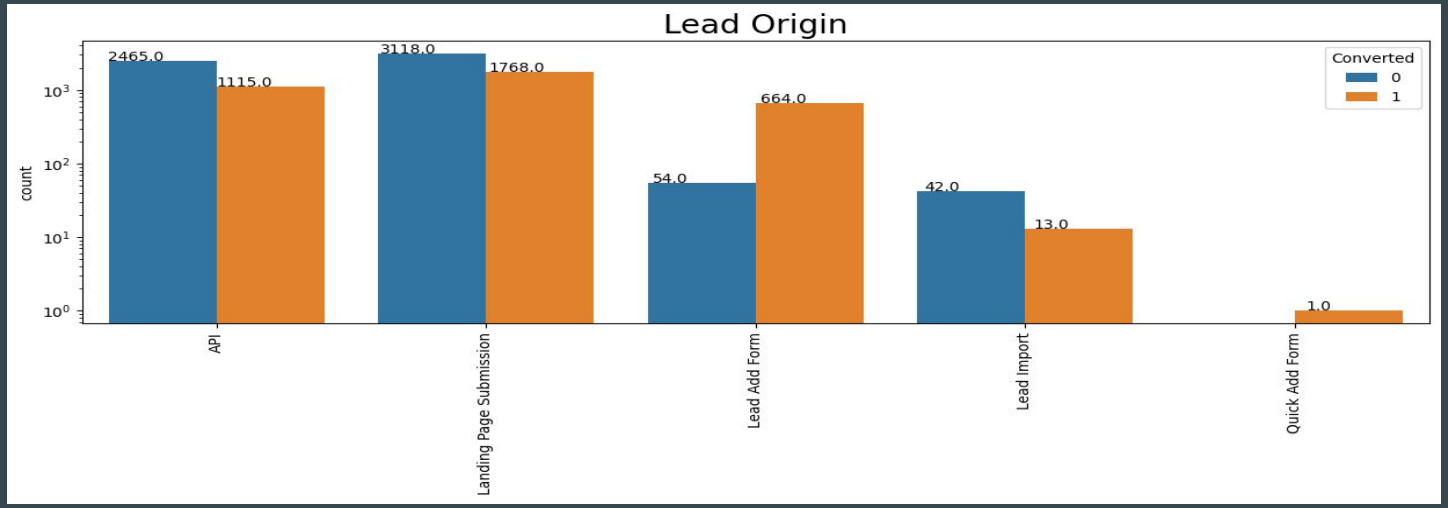
The imbalance ratio is **62.7%**

i.e. **count(Converted = 1) / count(Converted = 0)**

Univariate Analysis - Categorical - Lead Origin

Inference

- Conversion rate for API is ~ **31%** and for Landing Page Submission is **~36%** .
- For **Lead Add Form** number of successful conversions is more than unsuccessful ones.
- Count of Lead Import is lesser.

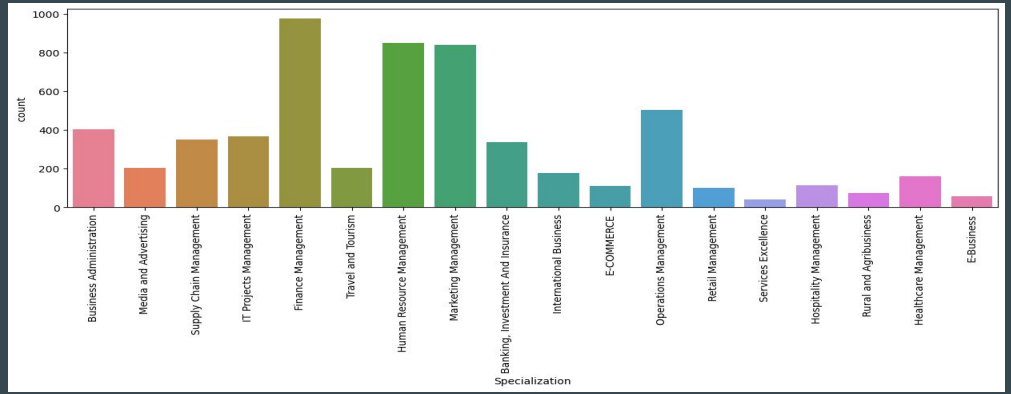


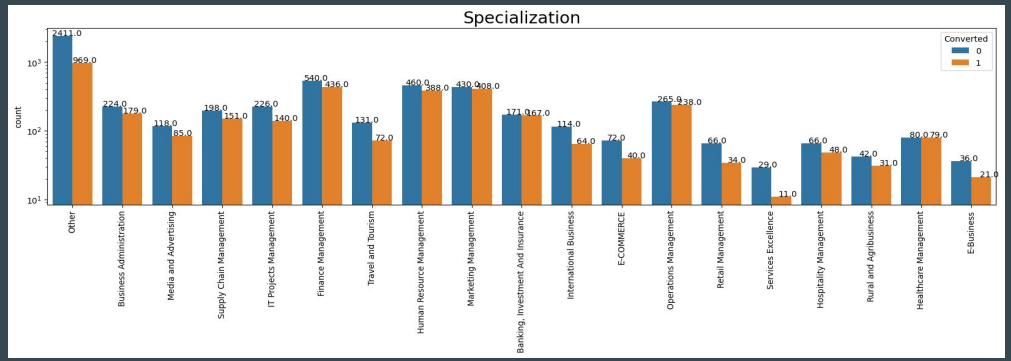
Data Imputing - Specialization

Inference

In these two graph,

- It would not be great to impute the values with most frequent values, as this will give us biased data.
- Hence, Labelling these rows as `**Other**`.
- **Management** under specialization has more number of leads.

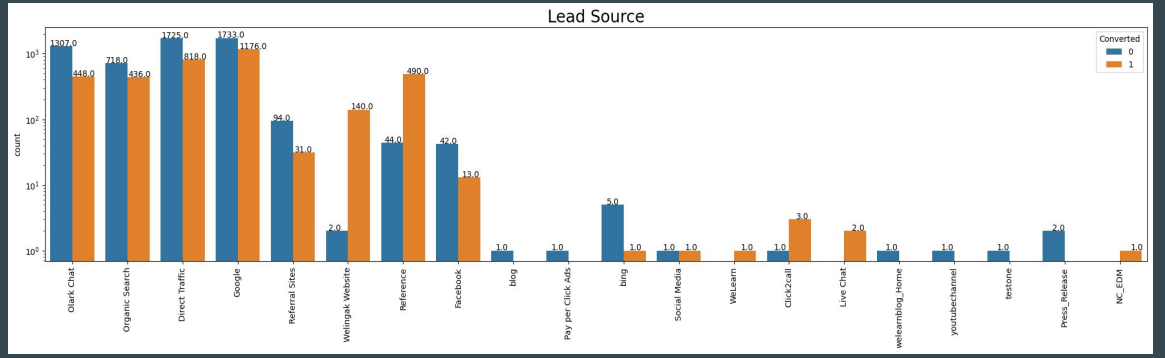


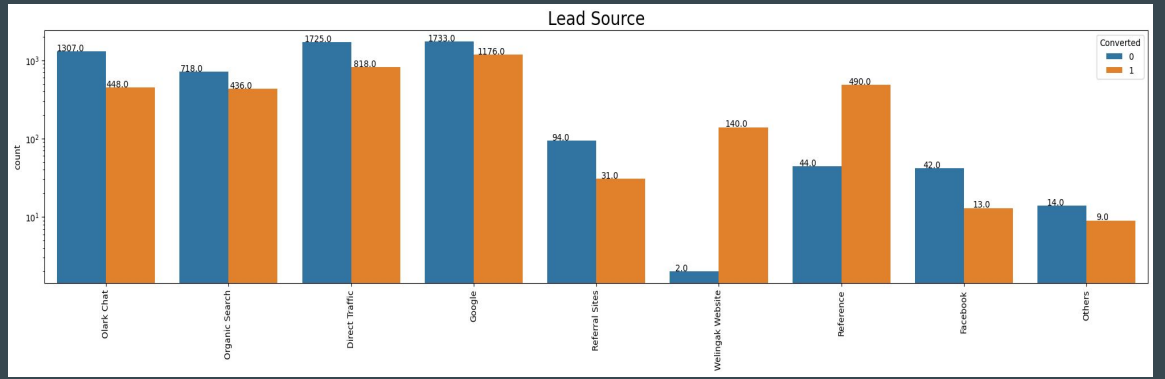


Univariate Analysis - Categorical - Lead Source

Inference

- The data is not significantly distributed among lower values. Hence, clubbing lower frequency values together under a common label **Others**.
- **Google** and **Direct traffic** generate maximum number of leads.
- **Conversion rate of Reference** and **Welingak Website** leads are high.

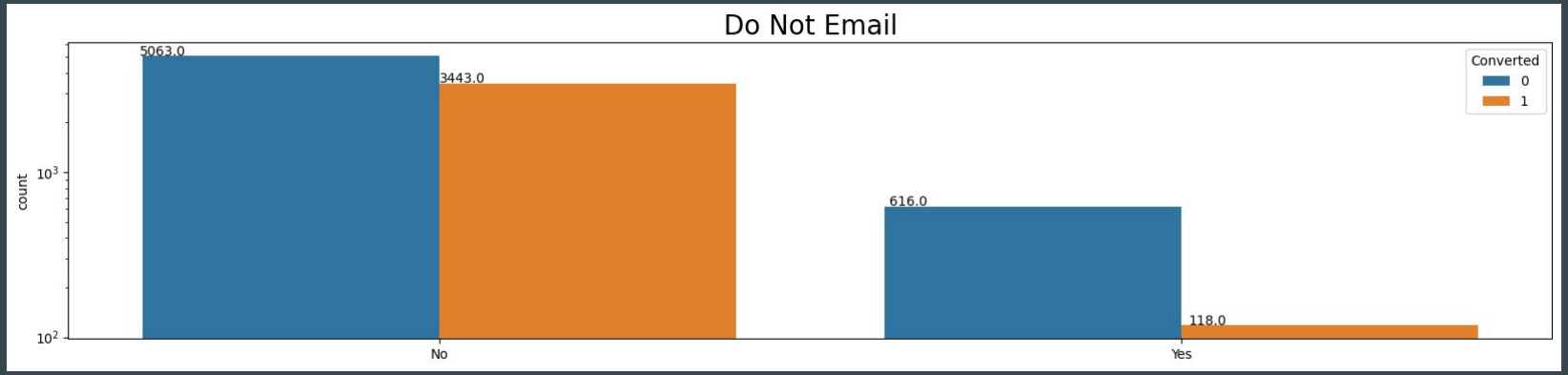




Univariate Analysis - Categorical - Do Not Email

Inference

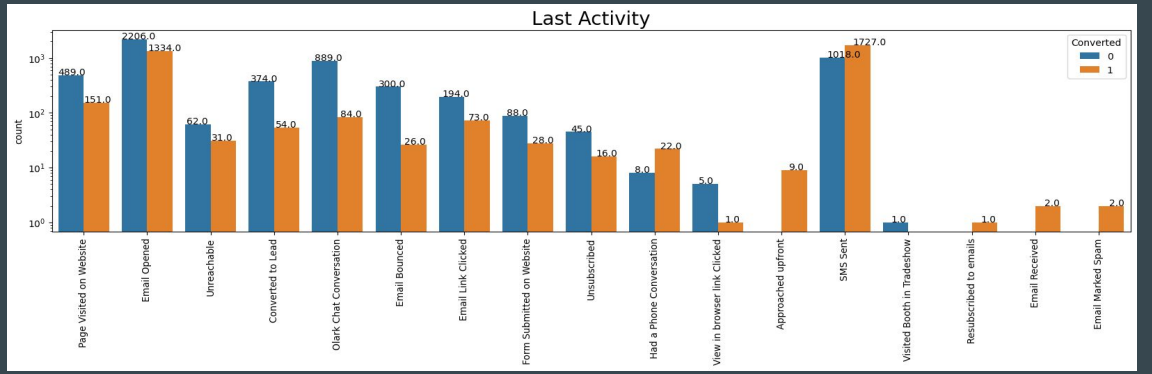
- People who opted for **mail** option are contributing to more leads.

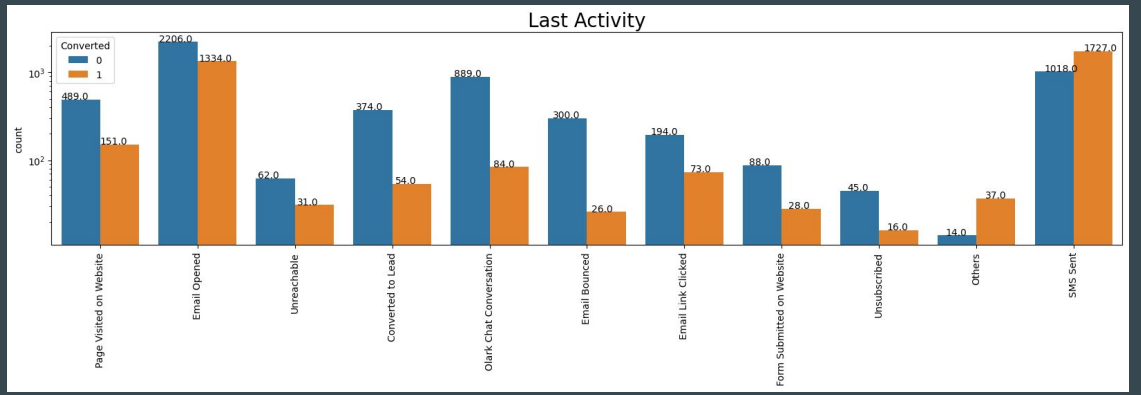


Univariate Analysis - Categorical - Last Activity

Inference

- Combining all low frequency values together under label **Others**.
- Conversion rate for last activity of **SMS Sent** is **~63%.**
- Highest last activity of leads is **Email Opened**.

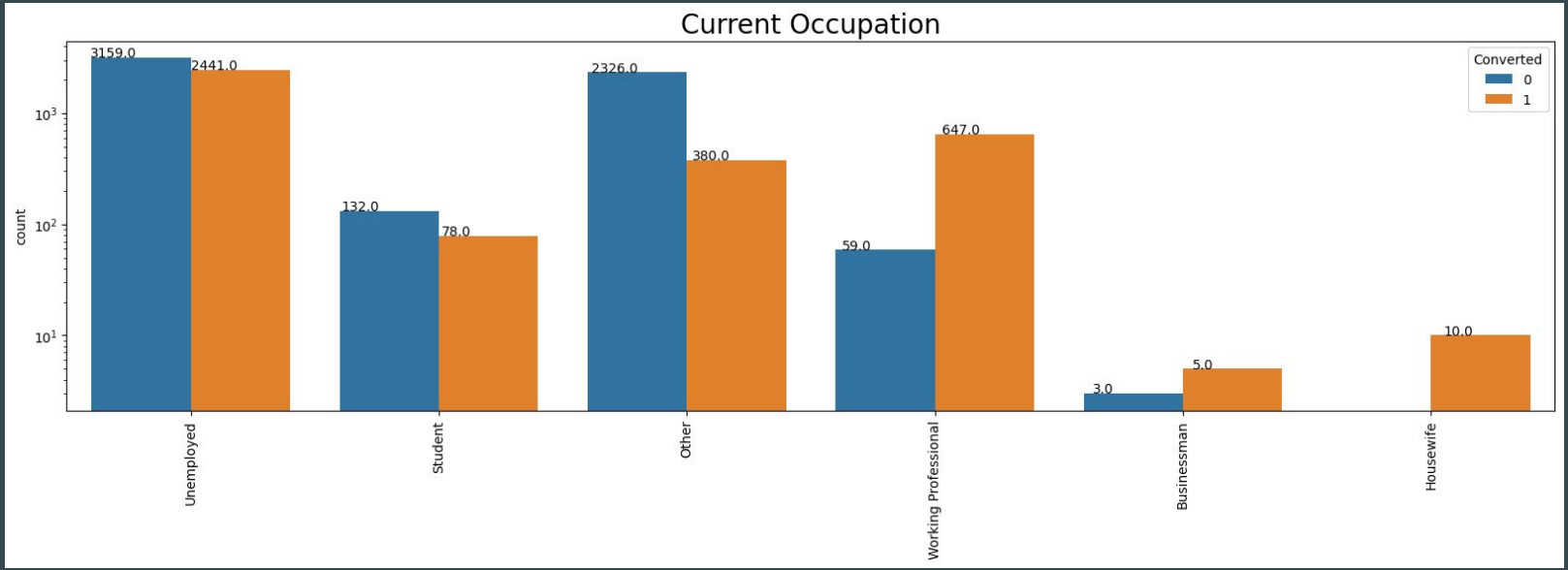




Univariate Analysis - Categorical - What is your current occupation

Inference

- **Unemployed** people are contributing to more number of leads and having **~45% conversion rate.**
- Conversion rate is higher for **Working Professionals.**

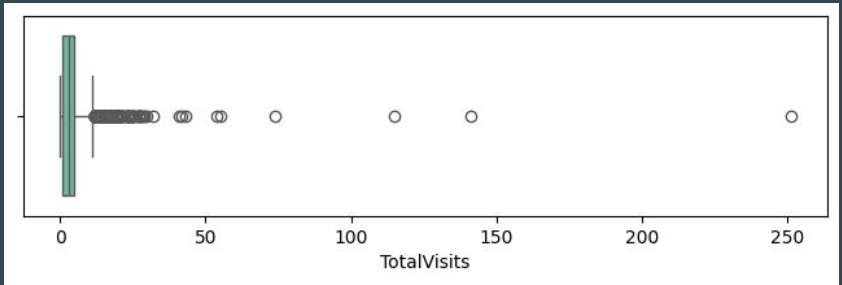


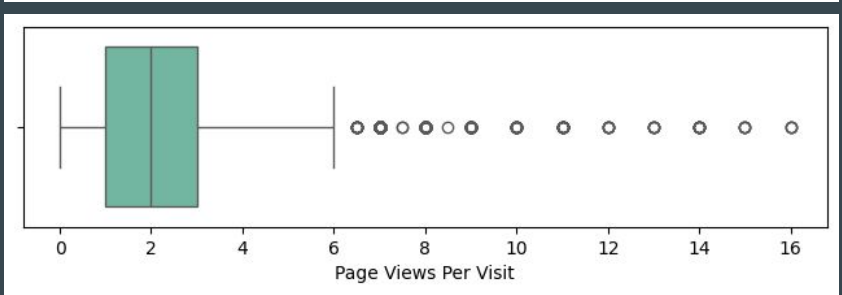
Univariate Analysis - Numerical - Total Visits, Page Views Per Visit

Inference

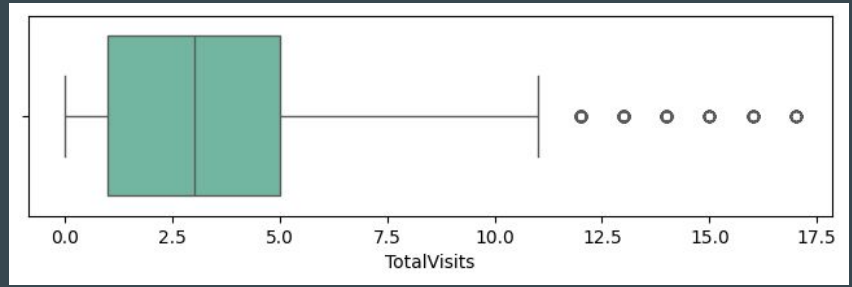
- We can see from below box plot that only upper range outliers are present in data, so need to treat outliers.
- Treating outliers by capping upper range to **0.99 percentile**

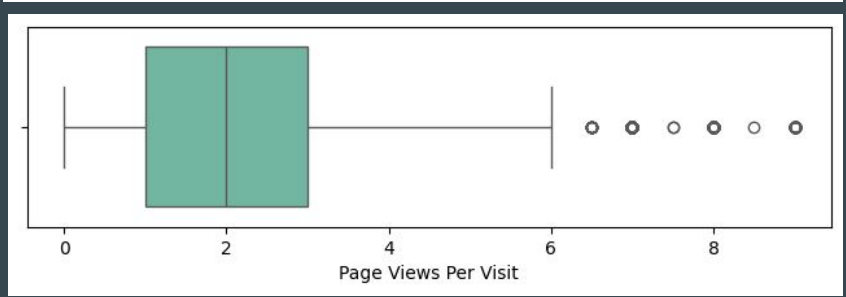
Before





After

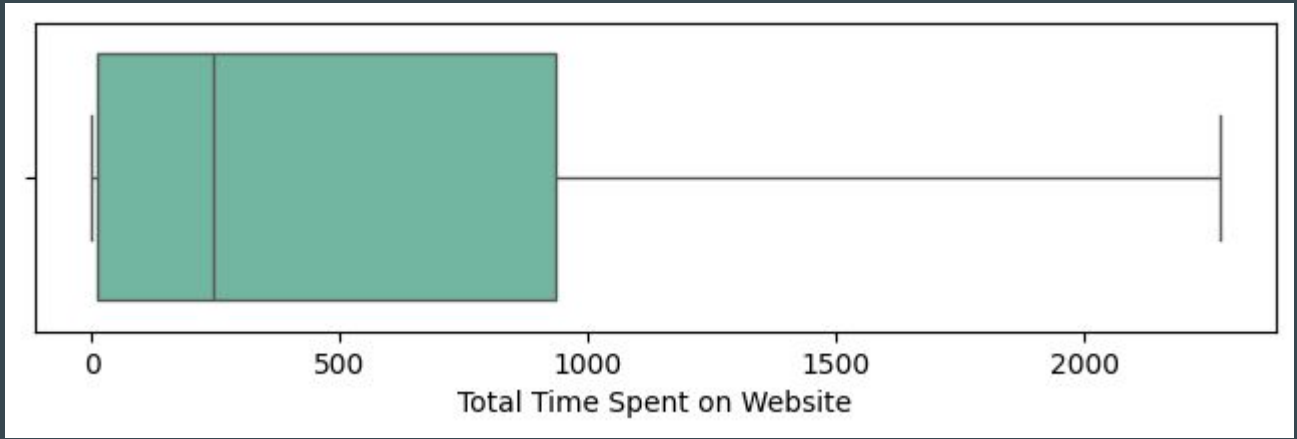




Univariate Analysis - Numerical - Total Time Spent on Website

Inference

- From below box plot we can see that there are no outliers in the data, so no outlier treatment is required.



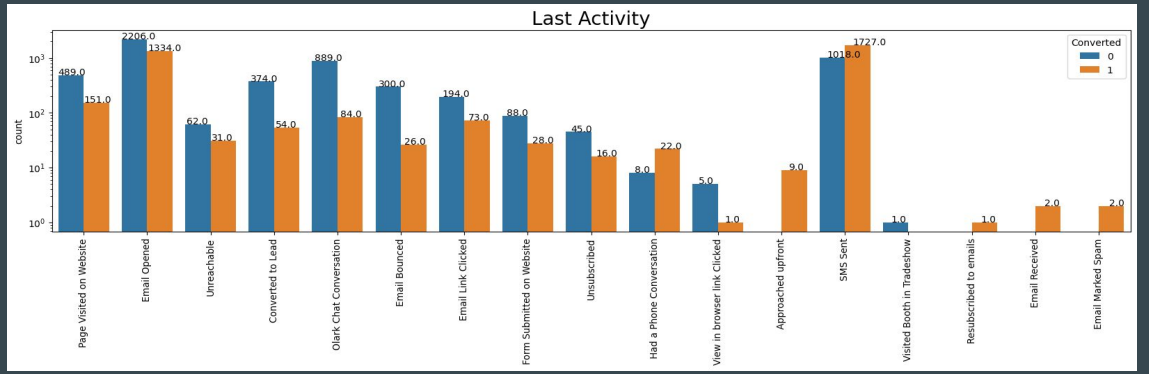
Data Analysis

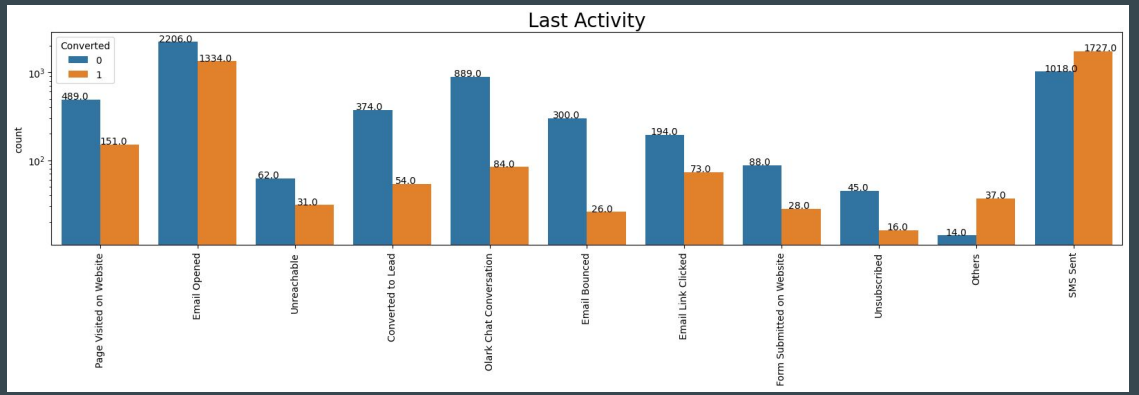
Bivariate Analysis

Univariate Analysis - Numerical - TotalVisits, Page Views Per Visit

Inference

- Combining all low frequency values together under label **Others**.
- Conversion rate for last activity of **SMS Sent** is **~63%.**
- Highest last activity of leads is **Email Opened**.

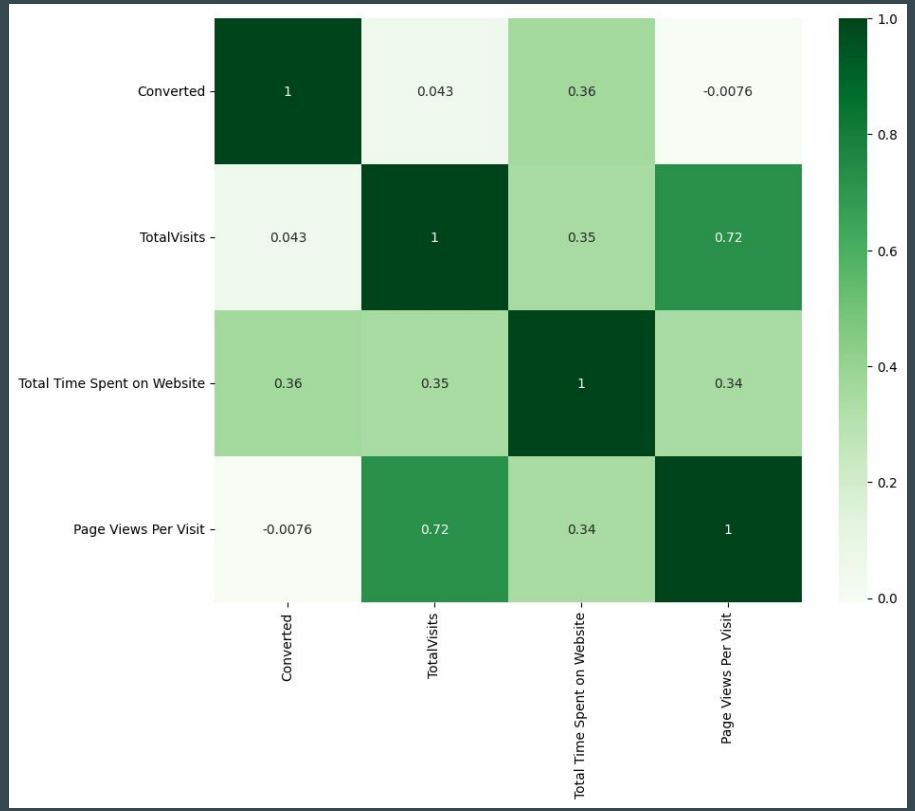




Bivariate Analysis - Correlation Matrix

Inference

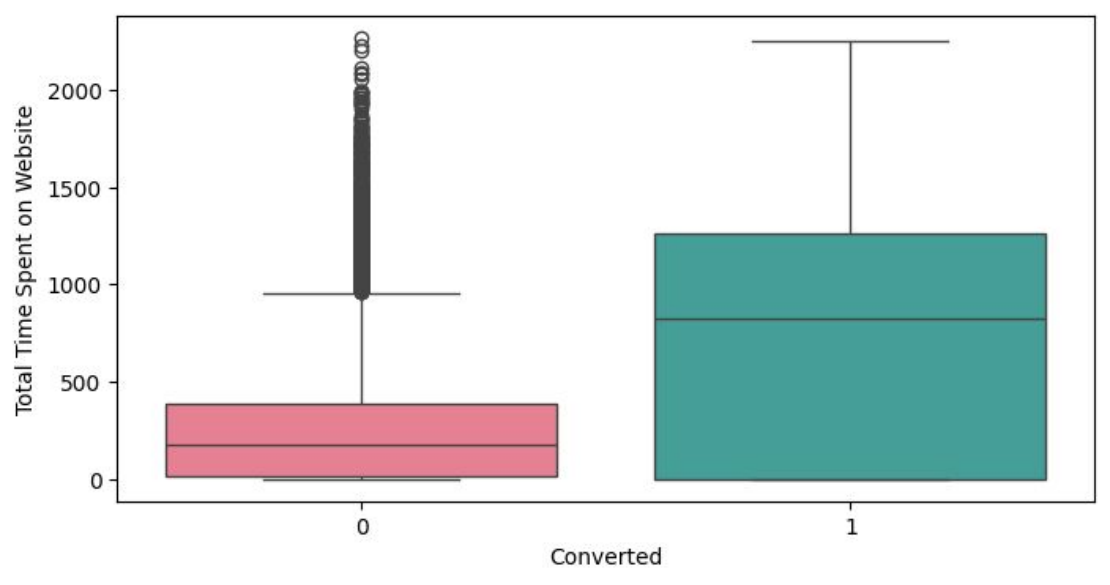
- **TotalVisits** and **Page Views Per Visit** are highly correlated with correlation of **0.72**.
- **Total Time Spent on Website** has correlation of **0.36** with target variable **Converted** .



Bivariate Analysis - Total Time Spent on Website vs Converted

Inference

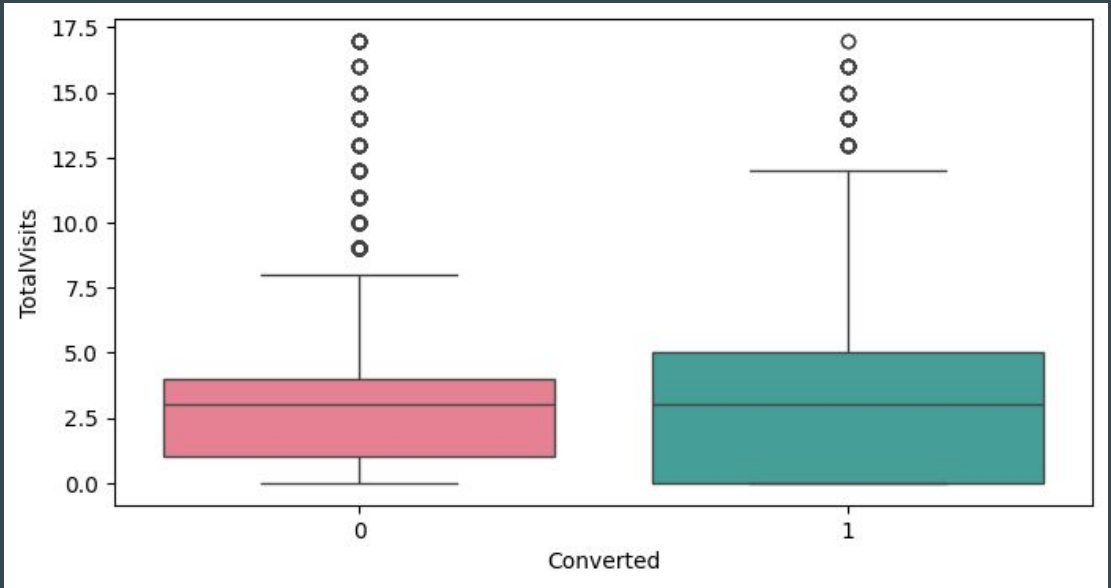
- Leads spending more time on website are more likely to opt for courses or get converted.



Bivariate Analysis - TotalVisits vs Converted

Inference

- From below plot, we can see that median for converted and unconverted is approximately similar.



Data Preparation

Data Preparation Steps

1

- Converted some binary variables (Yes/No) to 0/1

2

- For categorical variables with multiple levels, create dummy features
- Dummy Variables are created for object type variables

3

- Now, all the variables are numeric
- Total Rows available for Analysis: 9090
- Total Columns available for Analysis: 51

Data Preparation Steps

4

- Splitting the Data into Train and Test Sets
- The basic first step for regression is performing a train-test split, we have chosen 70:30 ratio to do the same.

5

- We are using **StandardScaler** for scaling.
- The heatmap (next page) clearly indicates the variables that are multicollinear and which have high collinearity with the target variable.

6

- We will use the heat map to build the logistic model, validating different correlated values. We will leverage **VIF** and **p-value** to identify which of the variables needs to be selected or eliminated.

Correlation Matrix - After creating Dummy variable

Inference

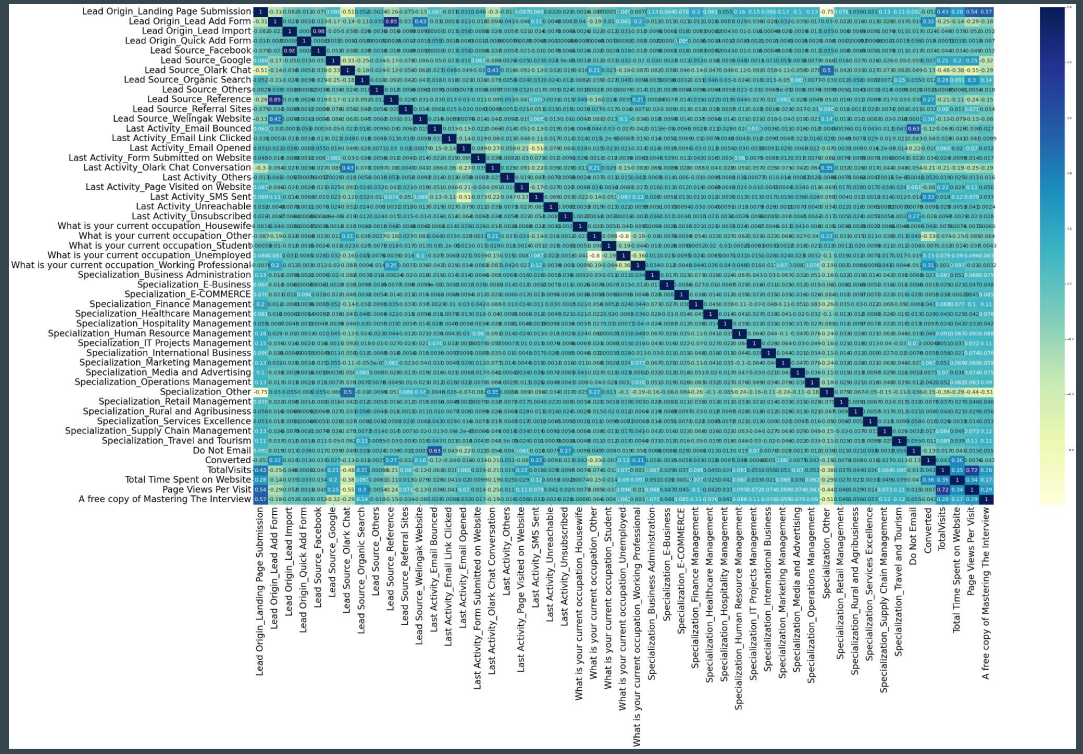
Lead Source_Facebook and **Lead Origin_Lead Import** have a high correlation of **0.98.**

Do Not Email and **Last Activity_Email Bounced** are **highly** correlated.

Lead Origin_Lead Add Form and **Lead Source_Reference** have a correlation of **0.85.**

TotalVisits and **Page Views Per Visit** have a correlation of **0.72.**

Lead Origin_Lead Add Form , **Lead Source_Welingak Website, Last Activity_SMS Sent** , and **What is your current Occupation_Working Professionals** show a positive correlation with the target variable **Converted** .



Model Building

Model Building Steps

1

- As you can see that there are a lot of variables present in the dataset which we cannot deal with. So the best way to approach this is to select a small set of features from this pool of variables using RFE.

2

- Use RFE for Feature Selection.
- Running RFE with 15 variables as output.

3

- Building Model by removing the variable whose **P-Value** is greater than **0.05** and **VIF** value is greater than **5.**

Model 1

Inference

Since the p-value for column **What is your current occupation_Housewife** is **0.999 (p>0.05)** , we can drop this column

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:**	-2567.5
Date:	Sat, 15 Jun 2024	**Deviance:**	5134.9
Time:	12:42:46	**Pearson chi2:**	6.92e+03
No. Iterations:	21	**Pseudo R-squ. (CS):**	0.4071
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3761	0.126	-2.988	0.003	-0.623	-0.129
Lead Origin_Landing Page Submission	-0.8685	0.129	-6.721	0.000	-1.122	-0.615
Lead Origin_Lead Add Form	2.9597	0.211	14.012	0.000	2.546	3.374
Lead Source_Olark Chat	1.1756	0.124	9.449	0.000	0.932	1.420
Lead Source_Welingak Website	3.1841	1.029	3.094	0.002	1.167	5.201
Last Activity_Olark Chat Conversation	-1.2072	0.167	-7.233	0.000	-1.534	-0.880
Last Activity_Others	1.2625	0.482	2.619	0.009	0.318	2.207
Last Activity_SMS Sent	1.4021	0.076	18.549	0.000	1.254	1.550
Last Activity_Unsubscribed	1.4655	0.449	3.263	0.001	0.585	2.346
What is your current occupation_Housewife	22.8356	1.39e+04	0.002	0.999	-2.73e+04	2.73e+04
What is your current occupation_Other	-1.1796	0.089	-13.276	0.000	-1.354	-1.005
What is your current occupation_Working Professional	2.3887	0.189	12.645	0.000	2.018	2.759
Specialization_Hospitality Management	-0.9458	0.336	-2.813	0.005	-1.605	-0.287
Specialization_Other	-0.8659	0.124	-6.987	0.000	-1.109	-0.623
Do Not Email	-1.5714	0.181	-8.706	0.000	-1.925	-1.218
Total Time Spent on Website	1.0715	0.040	26.577	0.000	0.992	1.151

	Features	VIF
12	Specialization_Other	2.19
2	Lead Source_Olark Chat	2.04
0	Lead Origin_Landing Page Submission	1.67
9	What is your current occupation_Other	1.63
1	Lead Origin_Lead Add Form	1.53
6	Last Activity_SMS Sent	1.52
4	Last Activity_Olark Chat Conversation	1.48
3	Lead Source_Welingak Website	1.32
14	Total Time Spent on Website	1.25
10	What is your current occupation_Working Profes...	1.20
13	Do Not Email	1.20
7	Last Activity_Unsubscribed	1.10
11	Specialization_Hospitality Management	1.02
5	Last Activity_Others	1.01
8	What is your current occupation_Housewife	1.01

Model 2

Inference

Dropping **Last Activity_Others** because of **p-value=0.01**

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6363
Model:	GLM	**Df Residuals:**	6348
Model Family:	Binomial	**Df Model:**	14
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2574.0
Date:	Sat, 15 Jun 2024	**Deviance:**	5148.0
Time:	12:51:11	**Pearson chi2:**	6.94e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.4058
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3662	0.126	-2.915	0.004	-0.613	-0.120
Lead Origin_Landing Page Submission	-0.8666	0.129	-6.716	0.000	-1.120	-0.614
Lead Origin_Lead Add Form	2.9724	0.211	14.089	0.000	2.559	3.386
Lead Source_Olark Chat	1.1745	0.124	9.443	0.000	0.931	1.418
Lead Source_Welingak Website	3.1699	1.029	3.081	0.002	1.153	5.187
Last Activity_Olark Chat Conversation	-1.2112	0.167	-7.256	0.000	-1.538	-0.884
Last Activity_Others	1.2546	0.482	2.601	0.009	0.309	2.200
Last Activity_SMS Sent	1.3941	0.076	18.462	0.000	1.246	1.542
Last Activity_Unsubscribed	1.4606	0.449	3.252	0.001	0.580	2.341
What is your current occupation_Other	-1.1839	0.089	-13.332	0.000	-1.358	-1.010
What is your current occupation_Working Professional	2.3804	0.189	12.603	0.000	2.010	2.751
Specialization_Hospitality Management	-0.9525	0.336	-2.834	0.005	-1.611	-0.294
Specialization_Other	-0.8715	0.124	-7.040	0.000	-1.114	-0.629
Do Not Email	-1.5757	0.180	-8.731	0.000	-1.929	-1.222
Total Time Spent on Website	1.0708	0.040	26.590	0.000	0.992	1.150

	Features	VIF
11	Specialization_Other	2.19
2	Lead Source_Olark Chat	2.04
0	Lead Origin_Landing Page Submission	1.66
8	What is your current occupation_Other	1.63
1	Lead Origin_Lead Add Form	1.53
6	Last Activity_SMS Sent	1.52
4	Last Activity_Olark Chat Conversation	1.48
3	Lead Source_Welingak Website	1.32
13	Total Time Spent on Website	1.25
9	What is your current occupation_Working Profes...	1.20
12	Do Not Email	1.20
7	Last Activity_Unsubscribed	1.10
10	Specialization_Hospitality Management	1.02
5	Last Activity_Others	1.01

Model 3

Inference

We observe that **the P-values of the variables are significant** and **the VIF values are below 3** . **No additional variables need to be dropped** , and we can proceed with making predictions using this model.

Generalized Linear Model Regression Results

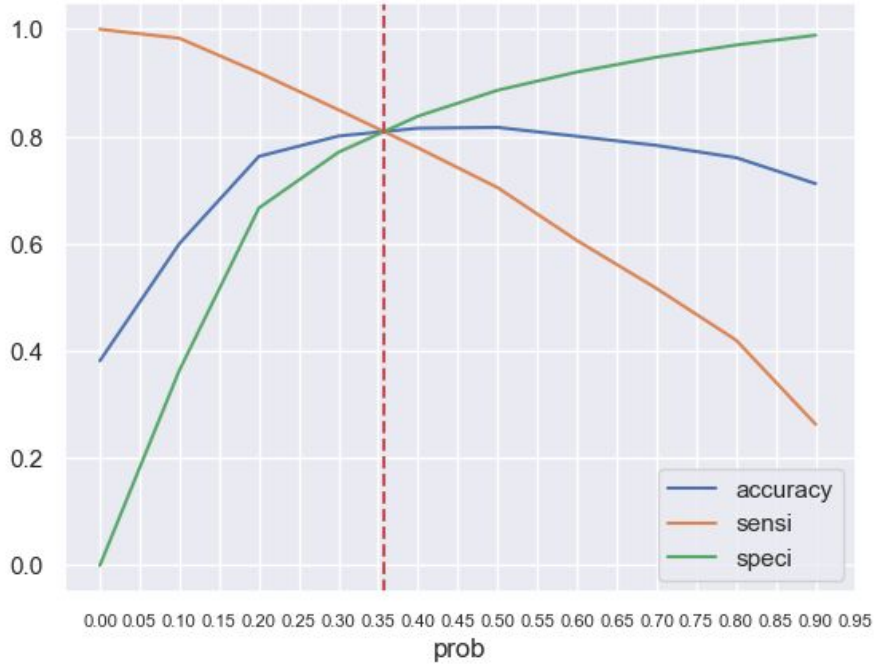
Dep. Variable:	Converted	**No. Observations:**	6363
Model:	GLM	**Df Residuals:**	6349
Model Family:	Binomial	**Df Model:**	13
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2577.6
Date:	Sat, 15 Jun 2024	**Deviance:**	5155.3
Time:	12:55:10	**Pearson chi2:**	6.80e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.4052
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3519	0.126	-2.803	0.005	-0.598	-0.106
Lead Origin_Landing Page Submission	-0.8696	0.129	-6.739	0.000	-1.123	-0.617
Lead Origin_Lead Add Form	2.9788	0.211	14.145	0.000	2.566	3.392
Lead Source_Olark Chat	1.1708	0.124	9.422	0.000	0.927	1.414
Lead Source_Welingak Website	3.1537	1.029	3.065	0.002	1.137	5.170
Last Activity_Olark Chat Conversation	-1.2222	0.167	-7.323	0.000	-1.549	-0.895
Last Activity_SMS Sent	1.3824	0.075	18.345	0.000	1.235	1.530
Last Activity_Unsubscribed	1.4457	0.449	3.219	0.001	0.565	2.326
What is your current occupation_Other	-1.1883	0.089	-13.389	0.000	-1.362	-1.014
What is your current occupation_Working Professional	2.3930	0.189	12.657	0.000	2.022	2.764
Specialization_Hospitality Management	-0.9632	0.336	-2.864	0.004	-1.622	-0.304
Specialization_Other	-0.8710	0.124	-7.038	0.000	-1.114	-0.628
Do Not Email	-1.5728	0.180	-8.738	0.000	-1.926	-1.220
Total Time Spent on Website	1.0724	0.040	26.650	0.000	0.993	1.151

	Features	VIF
10	Specialization_Other	2.18
2	Lead Source_Olark Chat	2.04
0	Lead Origin_Landing Page Submission	1.66
7	What is your current occupation_Other	1.62
1	Lead Origin_Lead Add Form	1.52
5	Last Activity_SMS Sent	1.51
4	Last Activity_Olark Chat Conversation	1.48
3	Lead Source_Welingak Website	1.31
12	Total Time Spent on Website	1.25
8	What is your current occupation_Working Profes...	1.20
11	Do Not Email	1.20
6	Last Activity_Unsubscribed	1.10
9	Specialization_Hospitality Management	1.02

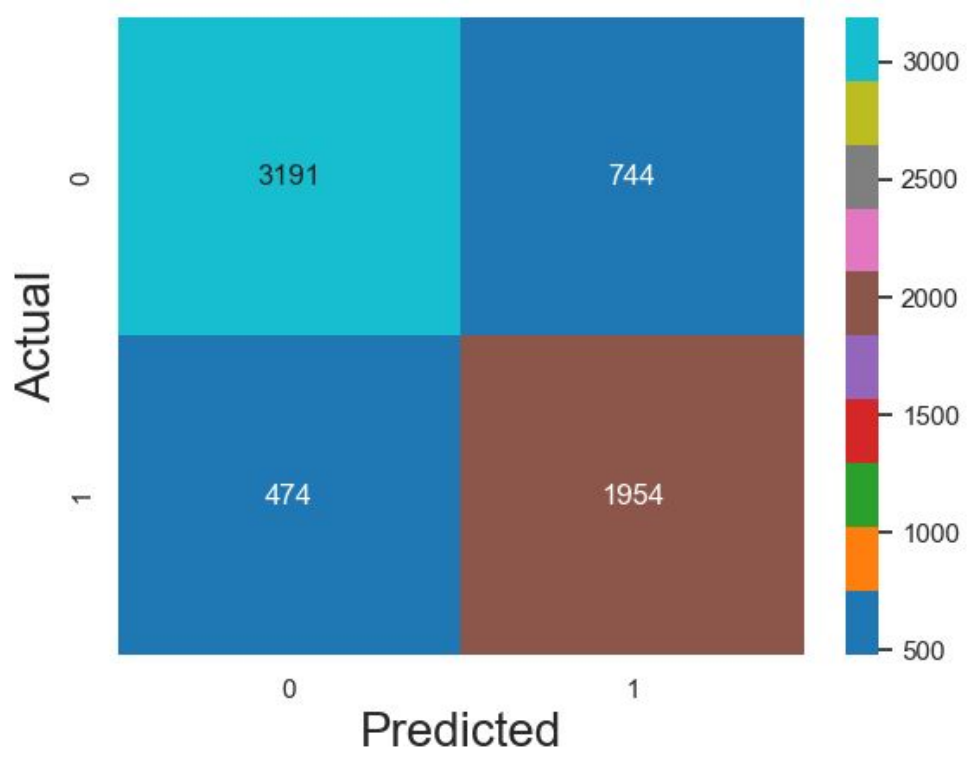
Optimal Cutoff Point

From the visualization below, it is evident that **0.357** is the optimal cutoff point.



Model Evaluation

Confusion Matrix - Train Dataset



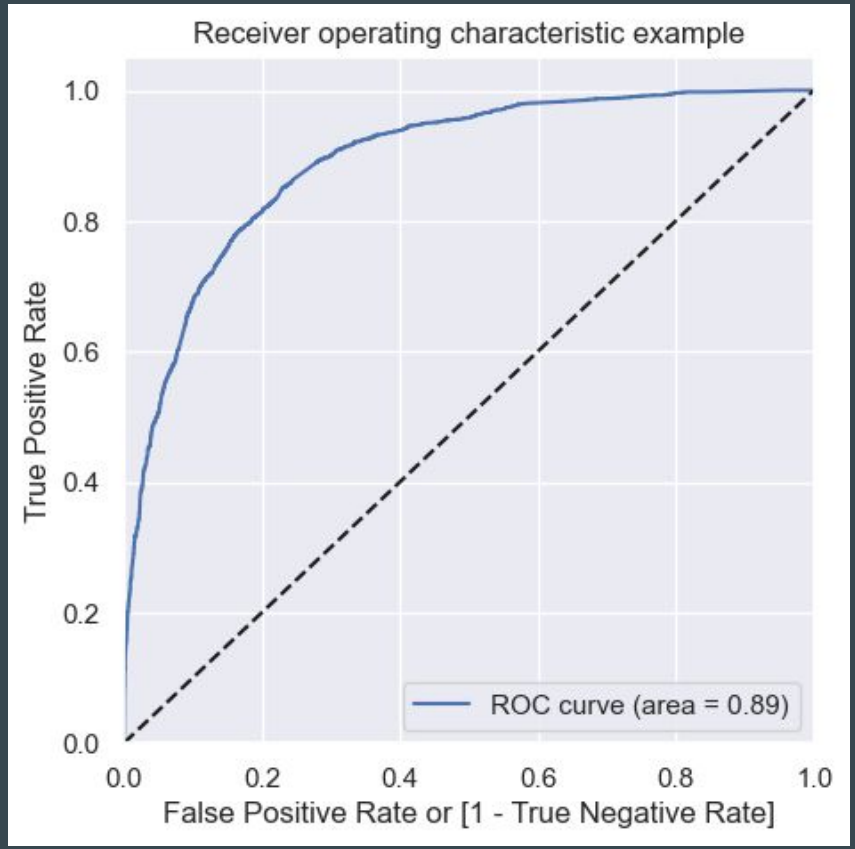
Observation

Accuracy:	**80.9%**
Sensitivity:	**80.5%**
Specificity:	**81.1%**
Precision:	**72.4%**
Recall:	**80.5%**

ROC Curve

Inference

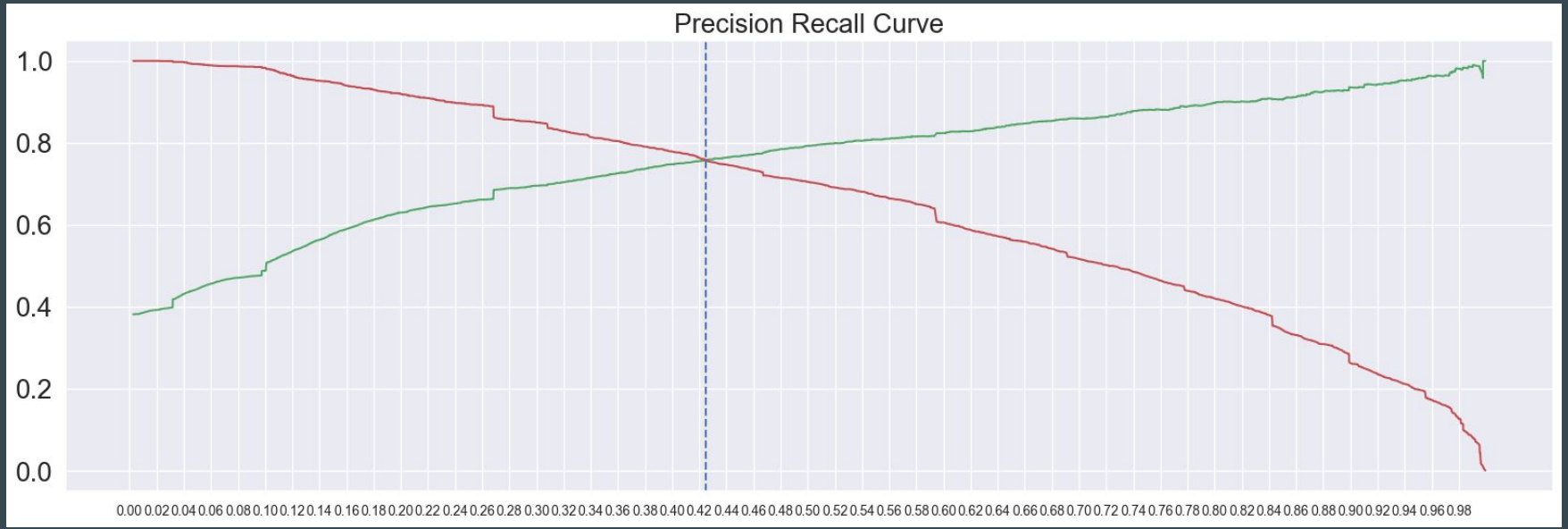
We have a strong ROC curve area of **0.89**, suggesting that our predictive model is effective. Because, **an ROC curve value close to 1 indicates a high level of accuracy.**



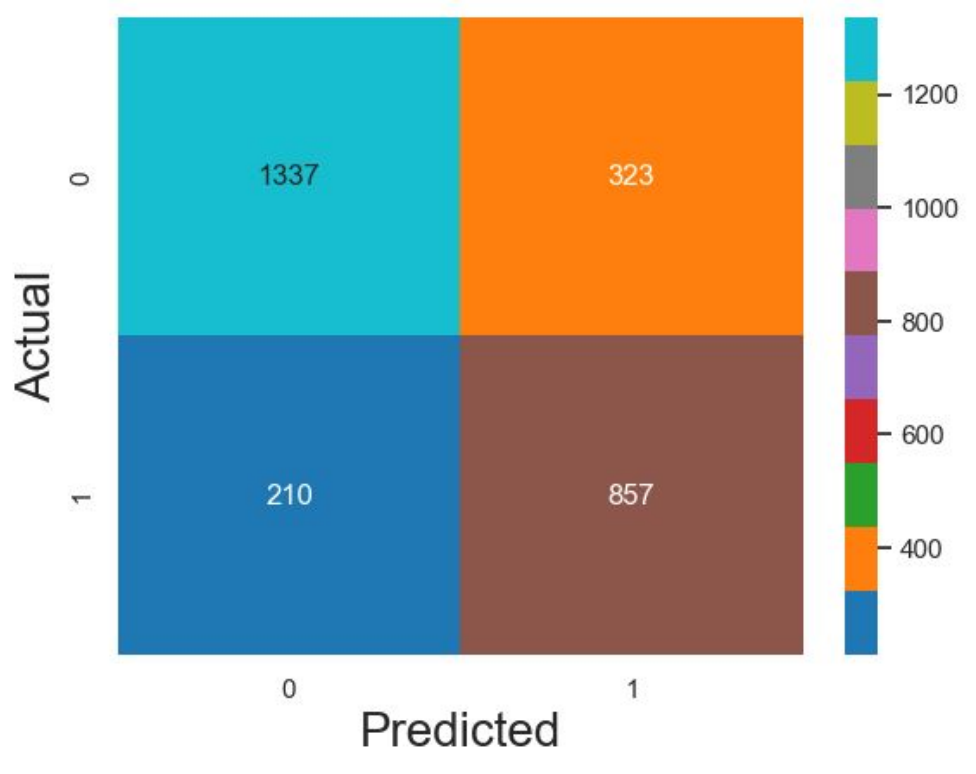
Precision and recall tradeoff

Inference

- From the **precision_recall_curve** below, we can see that the cutoff point is **0.427.**



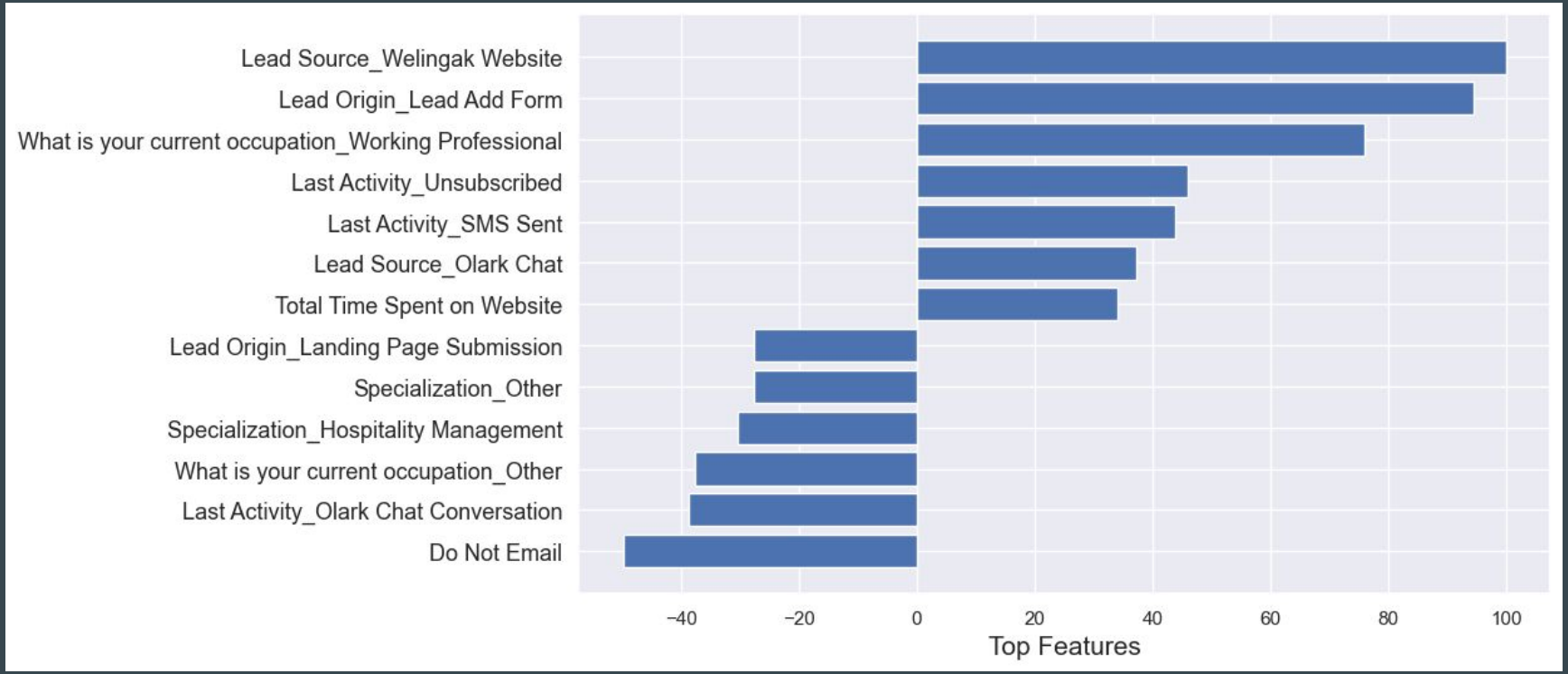
Confusion Matrix - Test Dataset



Observation

Accuracy:	**80.5%**
Sensitivity:	**80.3%**
Specificity:	**80.5%**
Precision:	**72.6%**
Recall:	**80.3%**

Top features based on the final model



Final model line equation

Converted = 0.261843 + 3.15 X Lead Source_Welingak Website + 2.98 X Lead Origin_Lead Add Form + 2.39 X What is your current occupation_Working Professional + 1.45 X Last Activity_Unsubscribed + 1.38 X Last Activity_SMS Sent + 1.17 X Lead Source_Olark Chat + 1.07 X Total Time Spent on Website - 0.87 X Lead Origin_Landing Page Submission - 0.87 X Specialization_Other - 0.96 X Specialization_Hospitality Management - 1.19 X What is your current occupation_Other - 1.22 X Last Activity_Olark Chat Conversation

Recommendations

To boost the potential lead conversion rate, X-Education should focus on the key features responsible for a higher conversion rate:

Leads from the Welingak Website show a higher conversion rate, so the company should prioritize this website to attract more potential leads.

Leads who engage through the **Lead Add Form have a higher conversion rate** , so focusing on this method can bring in more high-potential leads.

Leads who are working professionals have a higher conversion rate.
The company should target working professionals to increase the number of leads which has a higher probability of conversion.

Leads whose last activity involved **an SMS being sent are potential high-value leads** for the company.

Leads who **spend more time on the website** are more likely to convert, making them **potential high-value leads** .

Thank You!