Lead Scoring Case Study

Submitted by
Ratheesh Harshavardhanan
Uma Devi K
Deena Namreen

15-August-2023

Problem Statement

X Education company sells online courses to industry professionals.The company markets its courses on several websites and search engines like Google. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30% which is very poor.

Business Objective

X Education needs a model that will help them in identifying the leads that are most likely to convert into paying customers. The model should assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution Methodology

Step 1: Reading and Understanding the Data

Step 2: Data Cleaning

Step 3: Visualising the Data using EDA

Step 4: Create Dummy Variables

Step 5: Splitting the Data into Training and Testing Sets

Step 6: Feature Scaling

Step 7: Building a logistic regression model

Step 8: Feature Selection using RFE

Step 9: Plotting the ROC Curve and Evaluate the model by metrics like Accuracy, Specificity, Sensitivity, Precision and Recall

Step 10: Making Predictions on the test set using the Final Model

Reading, Understanding and Cleaning Data

1. Input file had 9240 rows and 37 columns.
2. Data types were Float, Integer and Objects.
3. There were no duplicate records in the data set.
4. 17 columns has NULL values in them.
5. Dropped columns which has > 30% of NULL values
6. There were multiple columns which had majority of values as ‘not selected.’ Hence converted them to null.
7. Dropped rows where columns have approx 1% of NULL values
8. The updated data frame had 9074 rows and 25 columns

Visualising Data using EDA

1. Plotted Box plot of numeric variables to identify outliers
2. Plotted Count Plot of categorical variables to observe the distribution of values
3. Dropped 15 columns which has highly biased value or which are deemed to be not having any predictive role.
4. The updated data frame had 9074 rows and 11 columns.
5. Dropped rows with high quantile values (>0.99)
6. There were not a significant drop. 96% of the records were retained. The updated data frame had 8863 rows and 11 columns.

Model Building

1. Created dummy variables for columns which are of object type

2. The updated data frame had 8863 rows and 65 columns.

3. Split the data frame into Train and Test data sets using 70:30 proportion.

4. Performed feature scaling on train dataset using standard scaler to convert categorical variables into numerical scale.

5. Build the a logistic regression model using all the variables.

6. Feature Selection using RFE. Selected top 20 features using Recursive Feature Elimination technique. Created a logistic regression model using these 20 and observed the P-Values and VIF score.

7. The model was run in multiple iterations and at the end of each iteration one feature with max P-Value was dropped. Finally after iteration 5 a final model with acceptable P-value and VIF values was determined. There were 16 features left in the final model.

8. Using the final model, created a new data frame with converted probability values. In first pass all probabilities > 0.5 were considered as converted(1) and remaining as not-converted (0).

Final Model's Stats

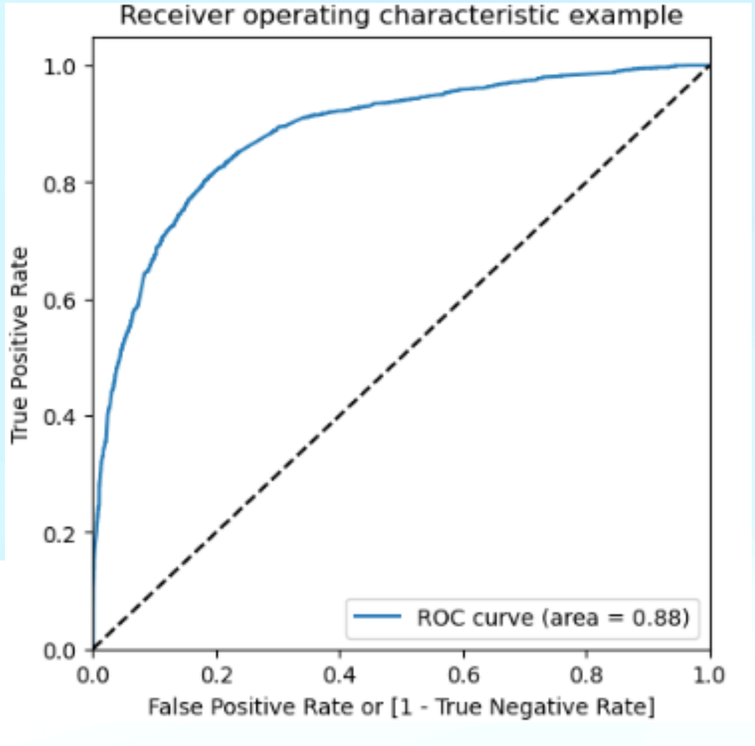
	coef	std err	z	P>|z|	[0.025	0.975]
const	2.8584	0.191	14.966	0.000	2.484	3.233
Total Time Spent on Website	1.1273	0.041	27.678	0.000	1.047	1.207
Lead Origin_Lead Add Form	3.9316	0.227	17.356	0.000	3.488	4.376
Lead Source_Olark Chat	1.2454	0.105	11.853	0.000	1.039	1.451
Lead Source_Welingak Website	2.6546	1.033	2.569	0.010	0.630	4.680
Do Not Email_Yes	-1.3379	0.198	-6.744	0.000	-1.727	-0.949
Last Activity_Converted to Lead	-0.9088	0.207	-4.395	0.000	-1.314	-0.504
Last Activity_Email Bounced	-1.1918	0.381	-3.130	0.002	-1.938	-0.446
Last Activity_Had a Phone Conversation	2.2325	0.943	2.367	0.018	0.384	4.081
Last Activity_Olark Chat Conversation	-1.3355	0.190	-7.011	0.000	-1.709	-0.962
What is your current occupation_Student	-2.3968	0.283	-8.480	0.000	-2.951	-1.843
What is your current occupation_Unemployed	-2.7432	0.183	-14.955	0.000	-3.103	-2.384
Last Notable Activity_Email Link Clicked	-1.9779	0.265	-7.469	0.000	-2.497	-1.459
Last Notable Activity_Email Opened	-1.4350	0.090	-15.996	0.000	-1.611	-1.259
Last Notable Activity_Modified	-1.6900	0.101	-16.669	0.000	-1.889	-1.491
Last Notable Activity_Olark Chat Conversation	-1.5719	0.372	-4.228	0.000	-2.301	-0.843
Last Notable Activity_Page Visited on Website	-1.7011	0.206	-8.272	0.000	-2.104	-1.298

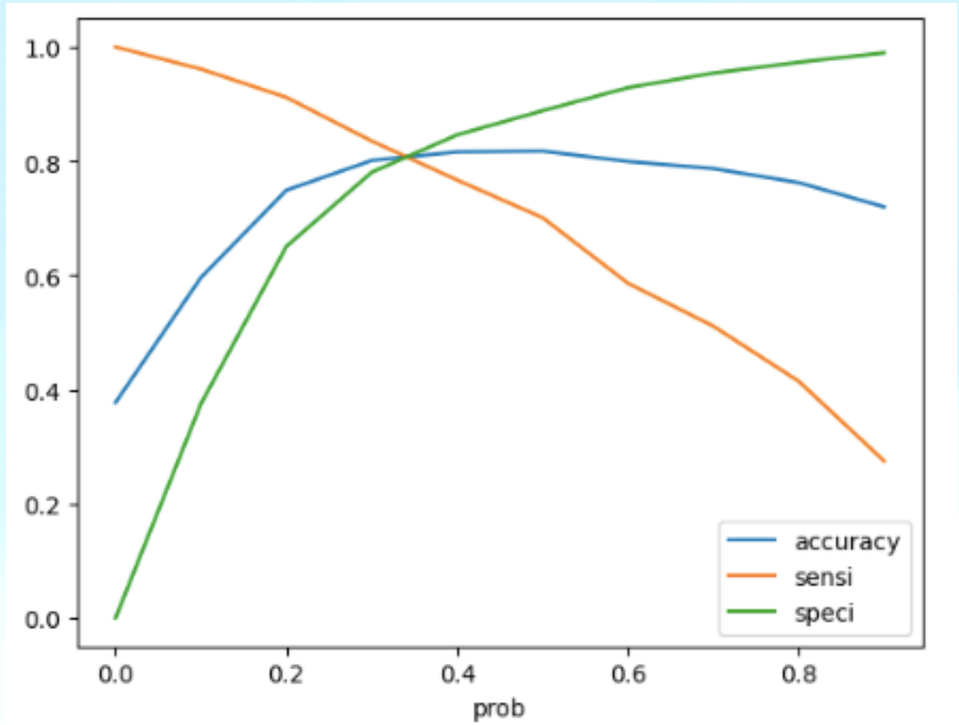
	Features	VIF
10	What is your current occupation_Unemployed	3.94
13	Last Notable Activity_Modified	2.98
12	Last Notable Activity_Email Opened	2.05
8	Last Activity_Olark Chat Conversation	2.03
4	Do Not Email_Yes	1.82
2	Lead Source_Olark Chat	1.78
6	Last Activity_Email Bounced	1.74
1	Lead Origin_Lead Add Form	1.42
14	Last Notable Activity_Olark Chat Conversation	1.40
3	Lead Source_Welingak Website	1.31
5	Last Activity_Converted to Lead	1.27
0	Total Time Spent on Website	1.25
15	Last Notable Activity_Page Visited on Website	1.12
9	What is your current occupation_Student	1.10
11	Last Notable Activity_Email Link Clicked	1.08
7	Last Activity_Had a Phone Conversation	1.00

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6204
Model:	GLM	Df Residuals:	6187
Model Family:	Binomial	Df Model:	16
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2559.2
Date:	Tue, 15 Aug 2023	Deviance:	5118.5
Time:	10:56:33	Pearson chi2:	6.53e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3937
Covariance Type:	nonrobust		

Plotting ROC Curve and determine Optimal CutOff Point





Plotted the ROC curve to observe the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity. We have observed area as 0.88 which is a good indication of the goodness of the model. The optimal Cutoff point was determined to be 0.35

Using the optimal cutoff point of 0.35, calculated the final Predicted value.

Accuracy 81.12%

Sensitivity 80.43%

Specificity 81.54%.

Precision 79.16%

Recall 70.1%

Making Predictions on the test set using the Final Model

1. Run the final model against the test data and determined the various metrics.

2. Identify the significant independent variables from the table listed below. The variables with positive sign should be focussed. They have higher chances to be potential customer.

3. The variables with negative values can be ignored. There is no need to focus on them.

Lead Origin_Lead Add Form	3.931557
const	2.858383
Lead Source_Welingak Website	2.654648
Last Activity_Had a Phone Conversation	2.232543
Lead Source_Olark Chat	1.245369
Total Time Spent on Website	1.127307
Last Activity_Converted to Lead	-0.908847
Last Activity_Email Bounced	-1.191781
Last Activity_Olark Chat Conversation	-1.335498
Do Not Email_Yes	-1.337927
Last Notable Activity_Email Opened	-1.434998
Last Notable Activity_Olark Chat Conversation	-1.571857
Last Notable Activity_Modified	-1.689989
Last Notable Activity_Page Visited on Website	-1.701051
Last Notable Activity_Email Link Clicked	-1.977854
What is your current occupation_Student	-2.396781
What is your current occupation_Unemployed	-2.743182

Model Evaluation Metrics: Train and Test

Train:

Accuracy: 81.12%

Sensitivity: 80.43%

Specificity: 81.54%

Precision: 79.16%

Recall: 70.1%

Test:

Accuracy: 80.59%

Sensitivity: 79.80%

Specificity: 81.07%

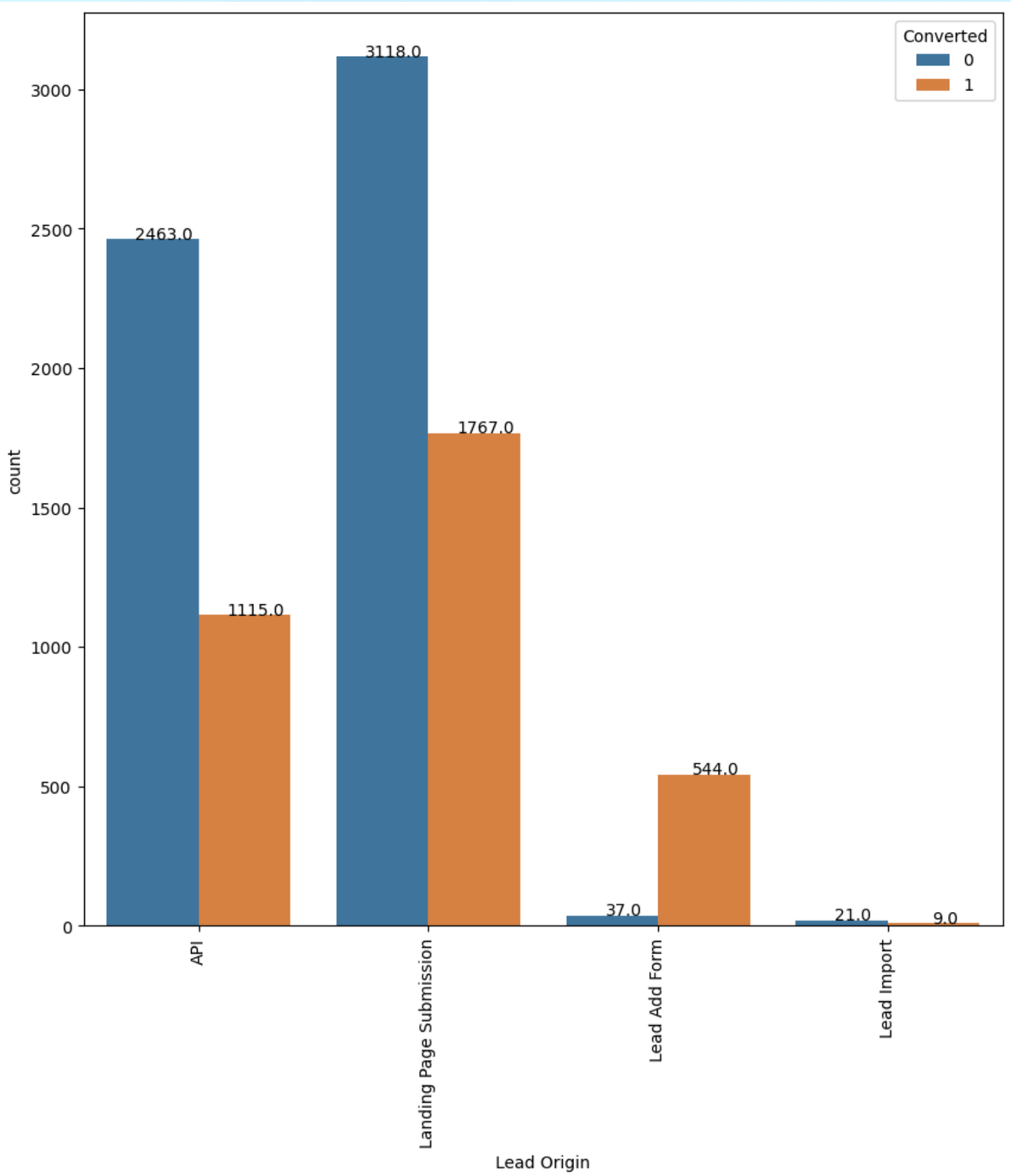
Precision: 72.09%

Recall: 79.80%

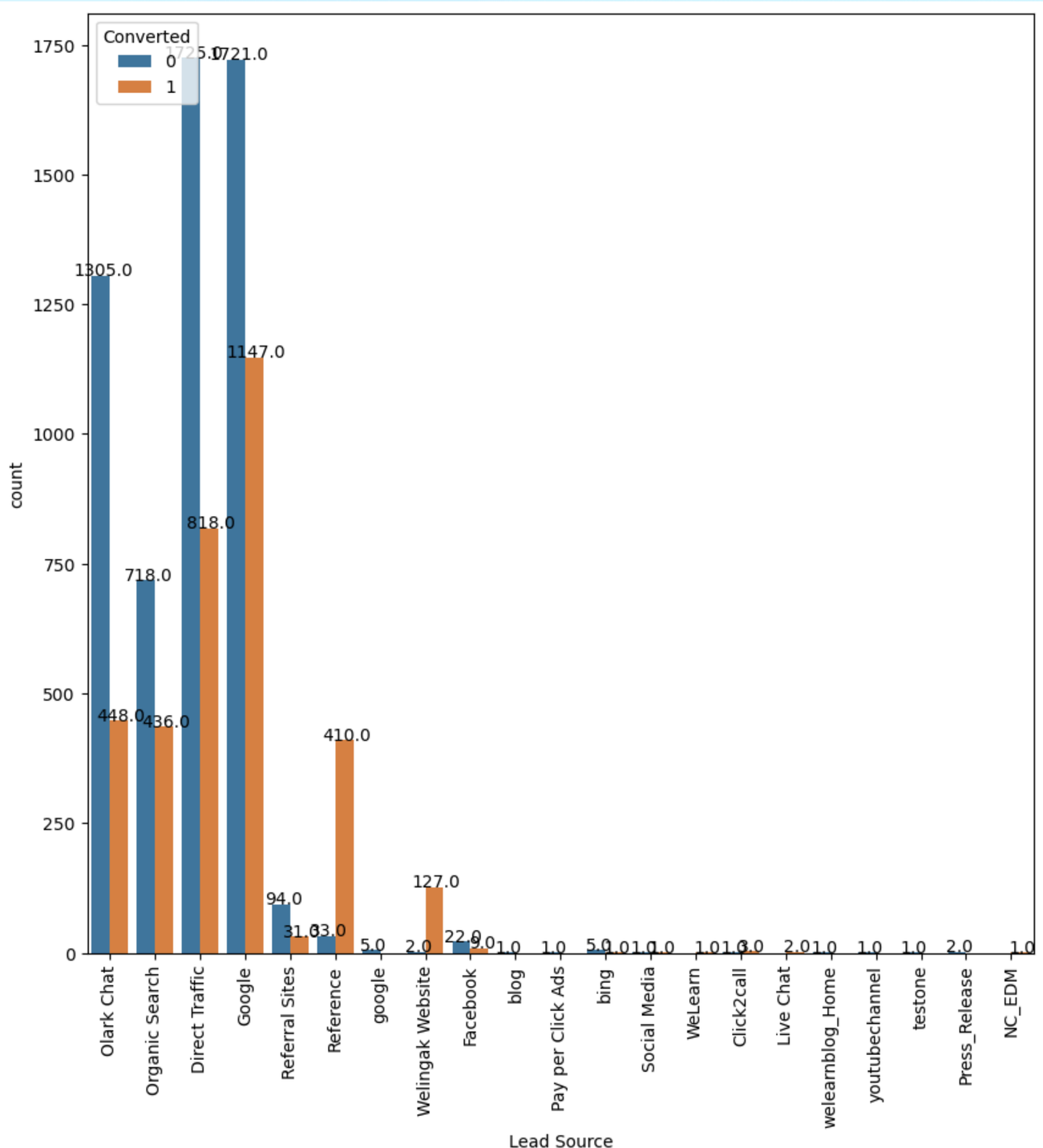
Inferences

Top three variables in the model which contribute most towards the probability of a lead getting converted are

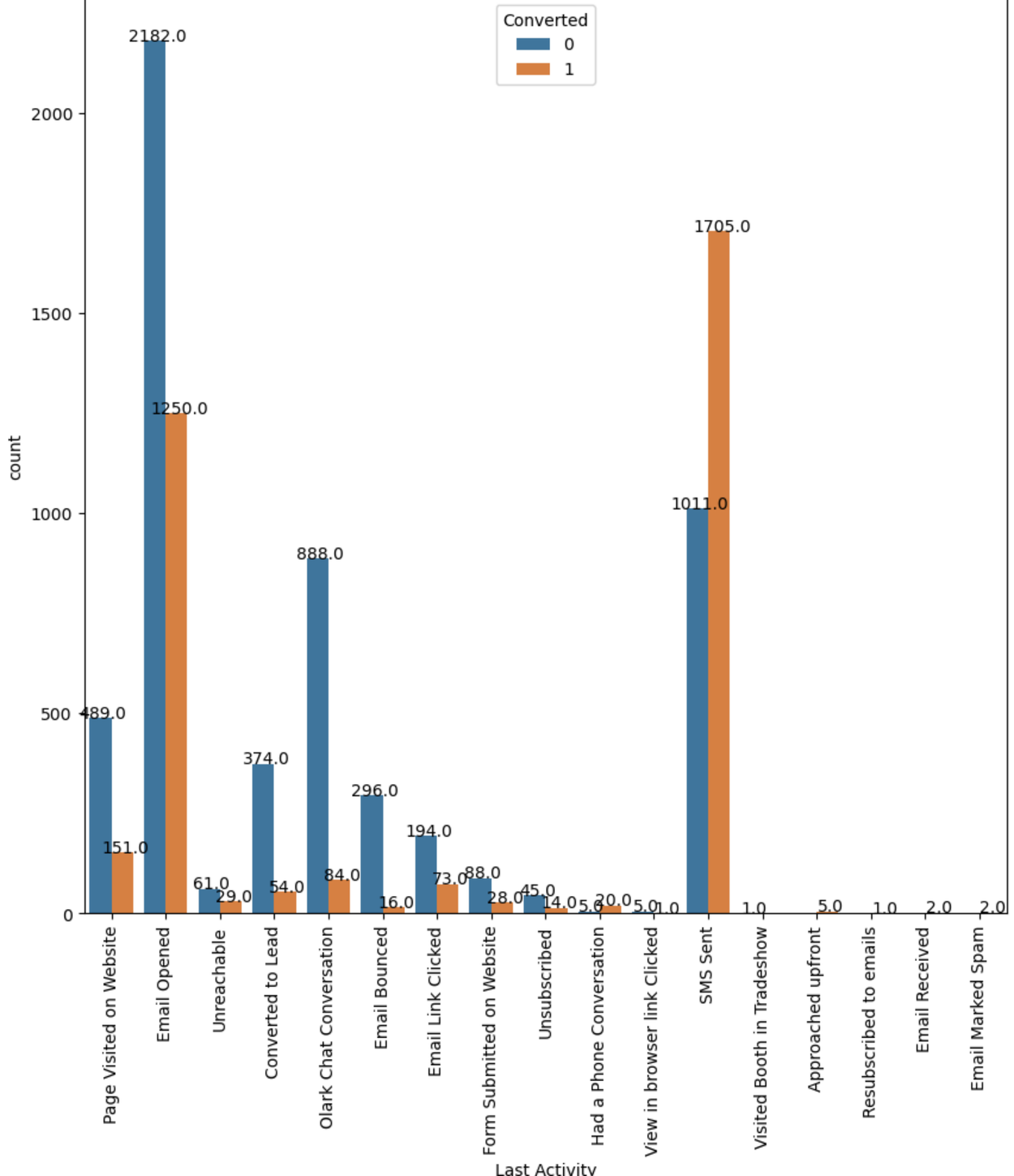
1. Lead Origin field
2. Lead Source field
3. Last Activity field



Customers with ‘Lead Origin’ value of Lead Add Form has the highest conversion rate.



Customers with Lead Source value of ‘Welingak Website’ has highest conversion rate



Customers with Last Activity value of ‘Had a Phone Conversation’ has highest conversion rate.