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. Imputed NULL values of Categorical and Numerical columns with Mean/Mode/Median appropriately
8. The updated data frame had 9240 rows and 26 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 9240 rows and 11 columns.
5. Dropped rows with high quantile values (>0.99)
6. There were not a significant drop. 98% of the records were retained. The updated data frame had 9029 rows and 11 columns.

Model Building

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

2. The updated data frame had 9029 rows and 66 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 6 a final model with acceptable P-value and VIF values was determined. There were 15 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

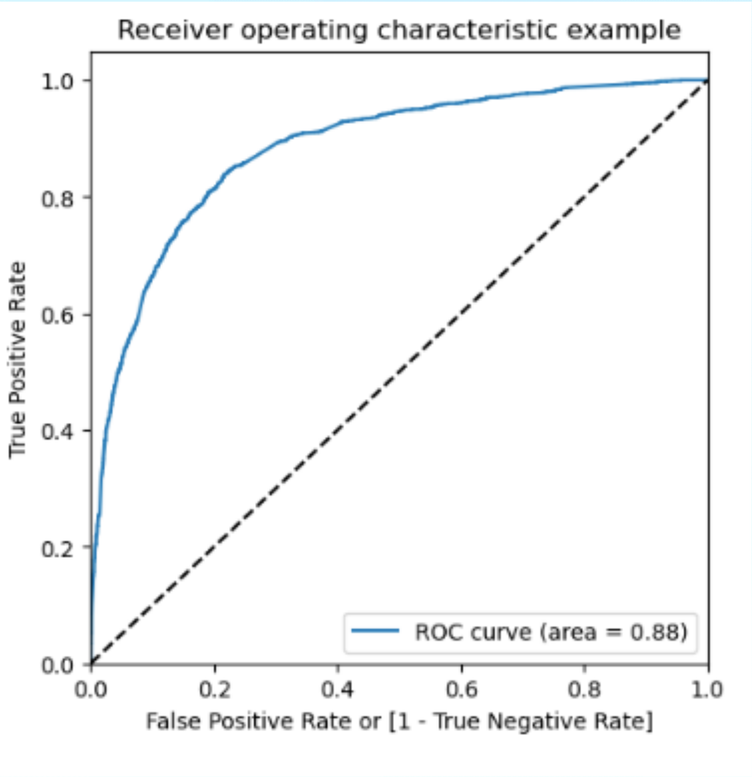
	coef	std err	z	P>|z|	[0.025	0.975]
const	1.2960	0.103	12.622	0.000	1.095	1.497
Total Time Spent on Website	1.0904	0.040	27.463	0.000	1.013	1.168
Lead Origin_Lead Add Form	3.2503	0.210	15.464	0.000	2.838	3.662
Lead Source_Direct Traffic	-1.2497	0.111	-11.248	0.000	-1.467	-1.032
Lead Source_Google	-0.9105	0.105	-8.657	0.000	-1.117	-0.704
Lead Source_Organic Search	-1.1112	0.133	-8.361	0.000	-1.372	-0.851
Do Not Email_Yes	-1.5987	0.170	-9.421	0.000	-1.931	-1.266
Last Activity_Converted to Lead	-0.8574	0.214	-4.014	0.000	-1.276	-0.439
Last Activity_Olark Chat Conversation	-1.2530	0.191	-6.564	0.000	-1.627	-0.879
Last Activity_Unavailable	-1.6835	0.425	-3.959	0.000	-2.517	-0.850
What is your current occupation_Working Professional	2.6041	0.182	14.285	0.000	2.247	2.961
Last Notable Activity_Email Link Clicked	-1.7752	0.258	-6.889	0.000	-2.280	-1.270
Last Notable Activity_Email Opened	-1.4528	0.088	-16.430	0.000	-1.626	-1.279
Last Notable Activity_Modified	-1.8527	0.101	-18.287	0.000	-2.051	-1.654
Last Notable Activity_Olark Chat Conversation	-1.5587	0.360	-4.330	0.000	-2.264	-0.853
Last Notable Activity_Page Visited on Website	-1.8121	0.208	-8.704	0.000	-2.220	-1.404

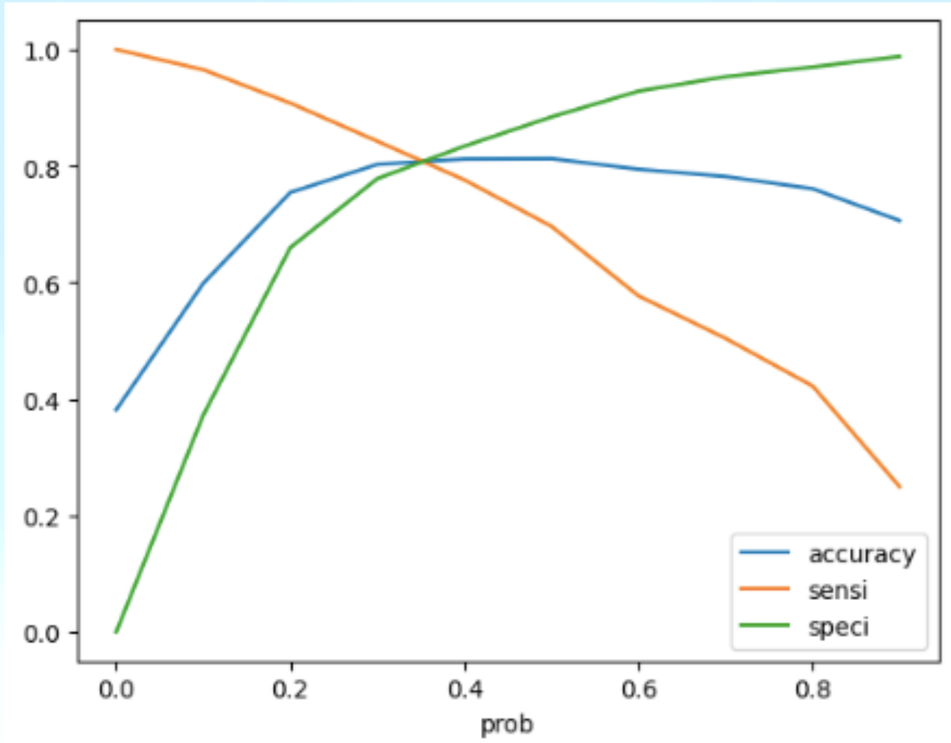
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6320
Model:	GLM	Df Residuals:	6304
Model Family:	Binomial	Df Model:	15
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2639.5
Date:	Tue, 08 Aug 2023	Deviance:	5279.0
Time:	14:35:47	Pearson chi2:	7.02e+03
No. Iterations:	6	Pseudo R-squ. (CS):	0.3900
Covariance Type:	nonrobust		

	Features	VIF
12	Last Notable Activity_Modified	2.65
7	Last Activity_Olark Chat Conversation	1.85
3	Lead Source_Google	1.78
2	Lead Source_Direct Traffic	1.76
11	Last Notable Activity_Email Opened	1.72
13	Last Notable Activity_Olark Chat Conversation	1.36
1	Lead Origin_Lead Add Form	1.30
4	Lead Source_Organic Search	1.30
6	Last Activity_Converted to Lead	1.26
0	Total Time Spent on Website	1.25
5	Do Not Email_Yes	1.19
8	Last Activity_Unavailable	1.19
9	What is your current occupation_Working Profes...	1.14
14	Last Notable Activity_Page Visited on Website	1.09
10	Last Notable Activity_Email Link Clicked	1.03

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 0.375

Using the optimal cutoff point of 0.375, calculated the final Predicted value.

Accuracy 80.7%

Sensitivity 79.3%

Specificity 81.6%.

Precision 78%

Recall 69%

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	3.250342
What is your current occupation_Working Professional	2.604062
const	1.296016
Total Time Spent on Website	1.090418
Last Activity_Converted to Lead	-0.857359
Lead Source_Google	-0.910481
Lead Source_Organic Search	-1.111224
Lead Source_Direct Traffic	-1.249724
Last Activity_Olark Chat Conversation	-1.253005
Last Notable Activity_Email Opened	-1.452772
Last Notable Activity_Olark Chat Conversation	-1.558720
Do Not Email_Yes	-1.598746
Last Activity_Unavailable	-1.683469
Last Notable Activity_Email Link Clicked	-1.775201
Last Notable Activity_Page Visited on Website	-1.812129
Last Notable Activity_Modified	-1.852651

Model Evaluation Metrics: Train and Test

Train:

Accuracy: 80.7%

Sensitivity: 79.3%

Specificity: 81.6%

Precision: 78%

Recall: 69%

Test:

Accuracy: 80.5%

Sensitivity: 79.08%

Specificity: 81.4%

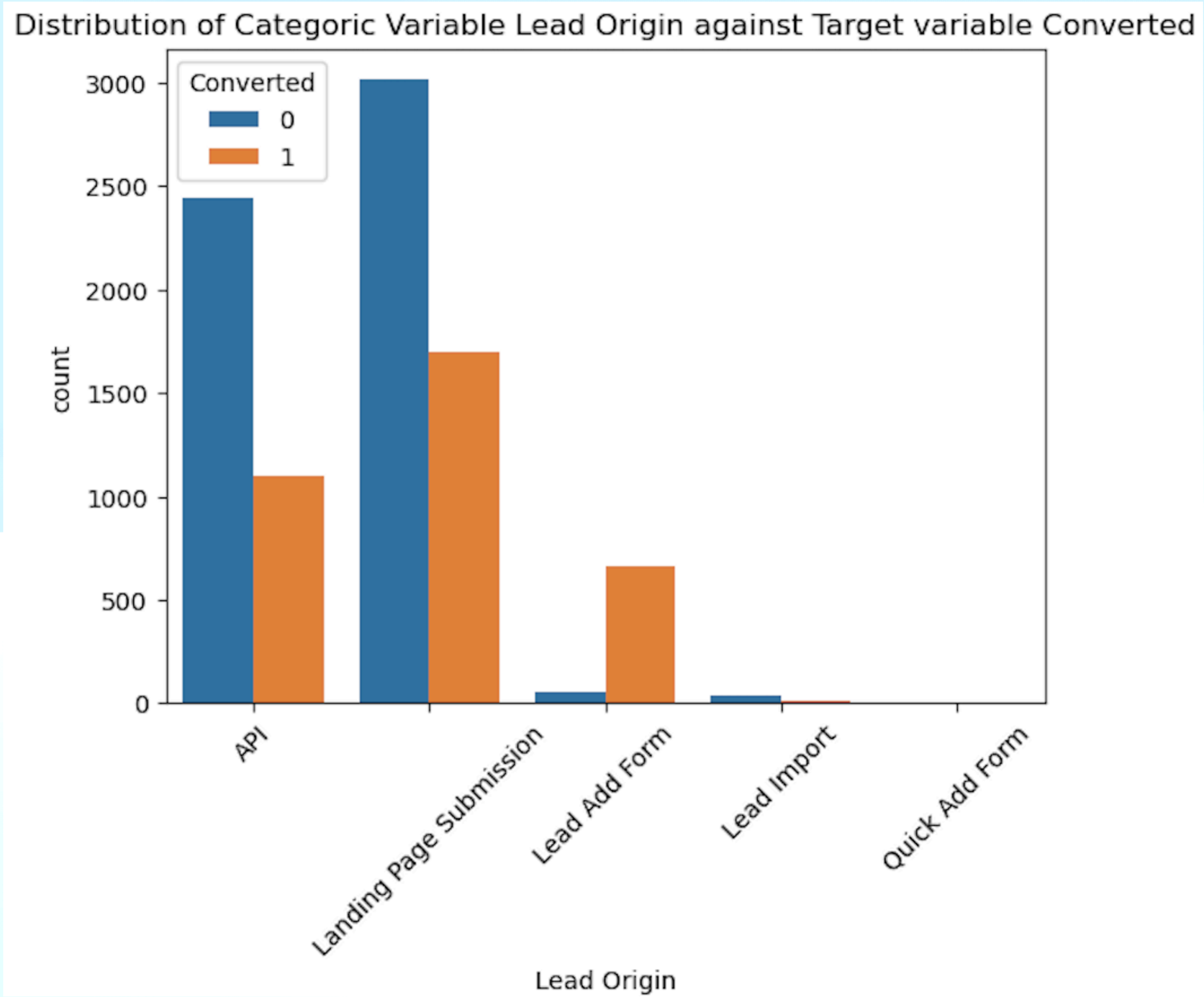
Precision: 73.4%

Recall: 79%

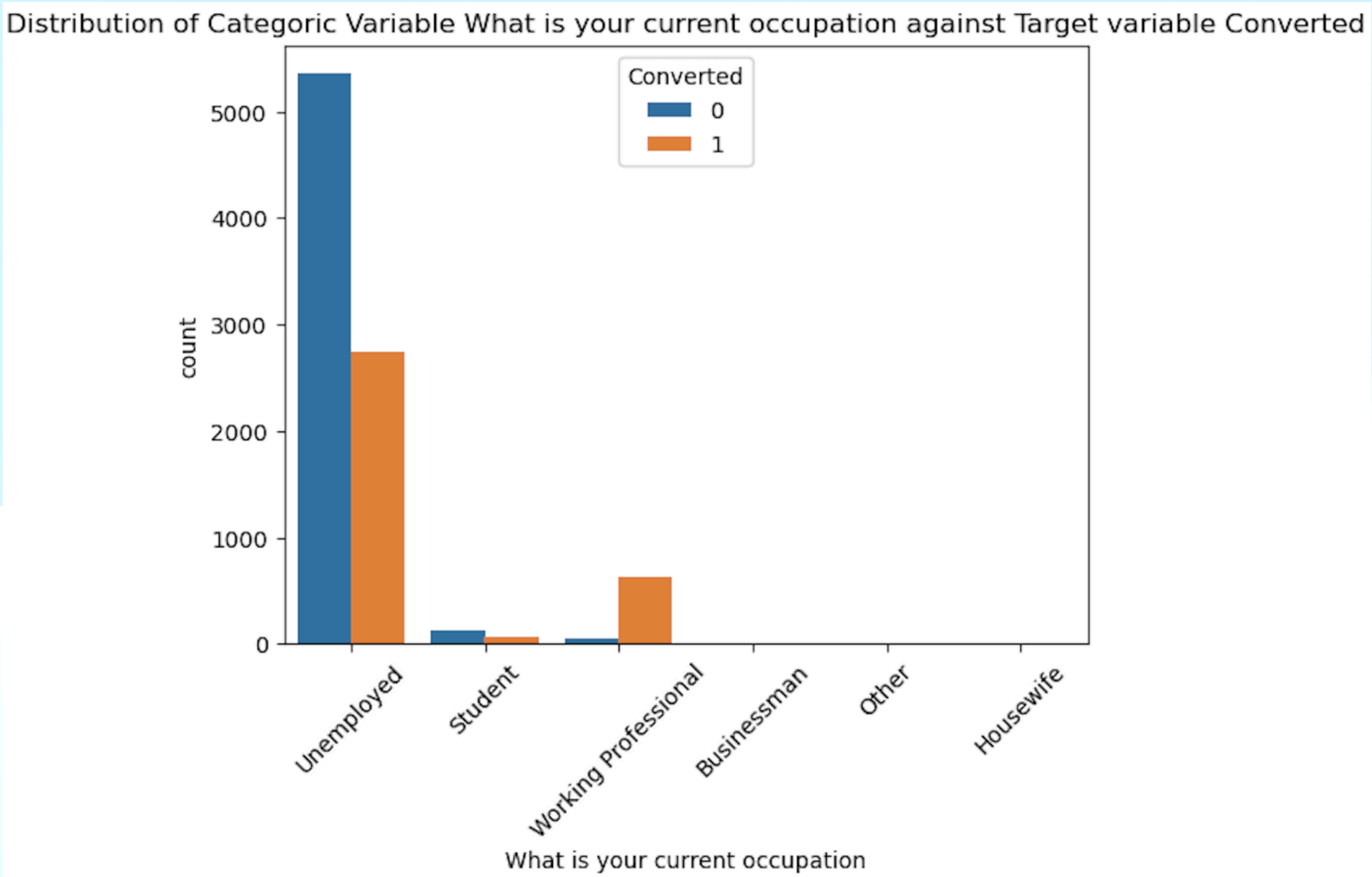
Inferences

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

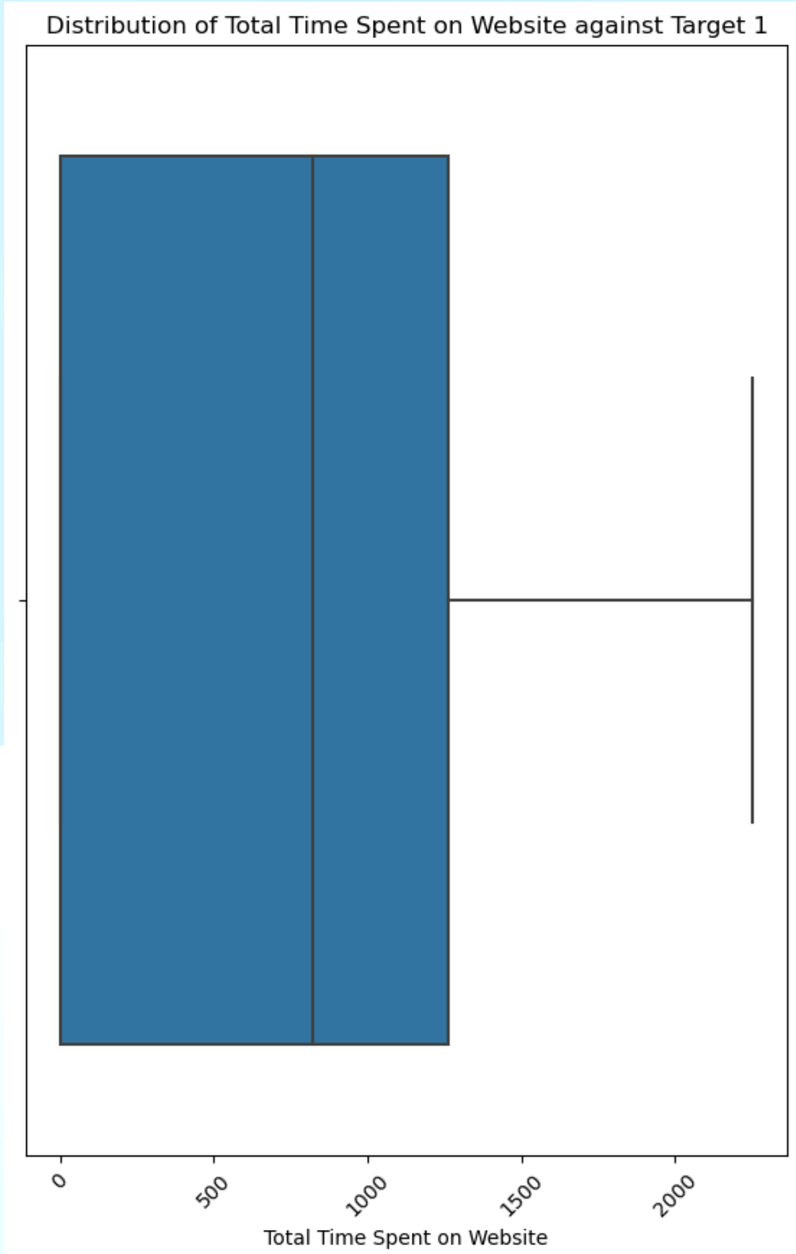
1. Lead Origin field
2. Current occupation field
3. Total Time Spent on Website field

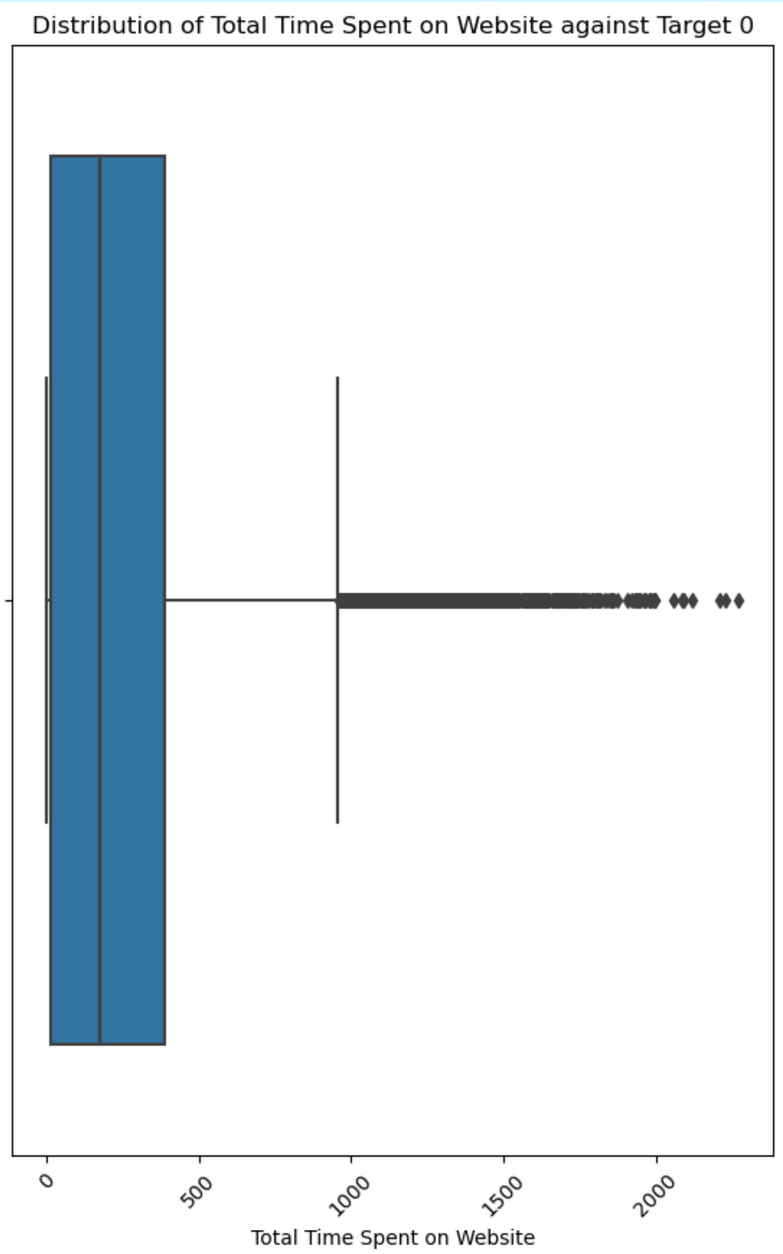


Customers with ‘Lead Origin’ value of Lead Add Form has the highest conversion rate.



Customers who are working professional has highest conversion rate.






```
Target 1
Mean 731.889847569744
Median 823.0
Mode 0
25% quantile:  0.0
50% quantile:  823.0
75% quantile:  1264.0
90% quantile:  1535.4
95% quantile:  1673.1999999999998
99% quantile:  1886.0
100th quantile:  0.0

Target 0
Mean 327.34636167146977
Median 176.0
Mode 0
25% quantile:  13.0
50% quantile:  176.0
75% quantile:  390.25
90% quantile:  1080.7000000000016
95% quantile:  1376.4499999999998
99% quantile:  1726.9399999999987
100th quantile:  0.0
```

Customers who spent more time on website has highest conversion rate