Lead Scoring Case Study

SUBMITTED BY: SHUBHAM SAURAV, SHAGUFTA KHAN, SAGAR KHADE

Problem statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses

Now, although X Education gets a lot of leads, its lead conversion rate is very poor.

Business Goal

X Education needs to identify the most promising leads—those who have the highest chance of becoming paying clients.

The business wants us to create a model in which every lead must be given a lead score, meaning that consumers who have higher lead scores are more likely to convert than those who have lower lead scores.

The intended lead conversion rate, as stated by the CEO, is approximately 80%.

Methodology

Understanding and reading the data

Inspecting the dataset

Exploratory data analysis (EDA) & dummy variable creation

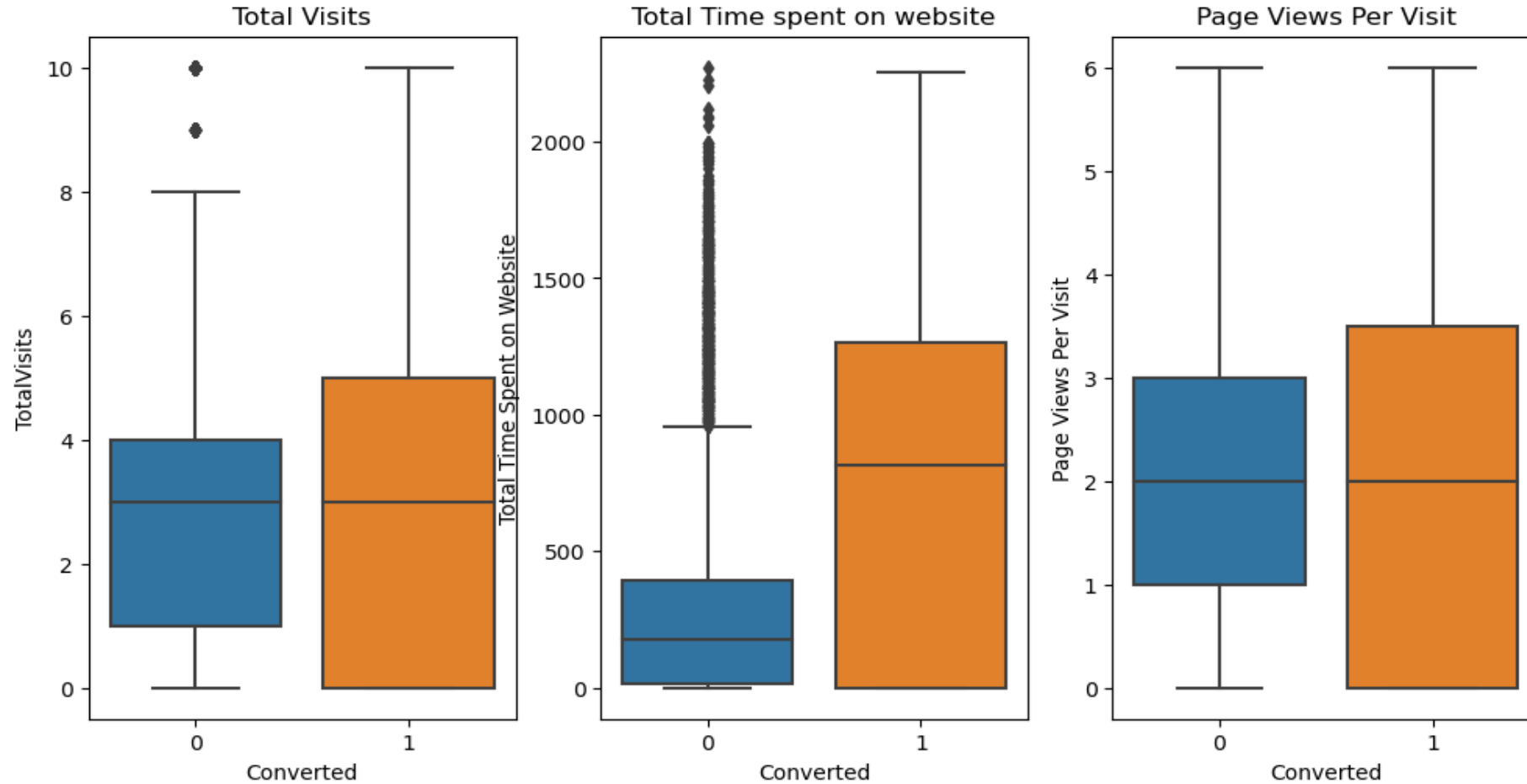Train test split & feature scaling

Model building

Finding optimal cut-off point & model evaluation

Plotting ROC curve

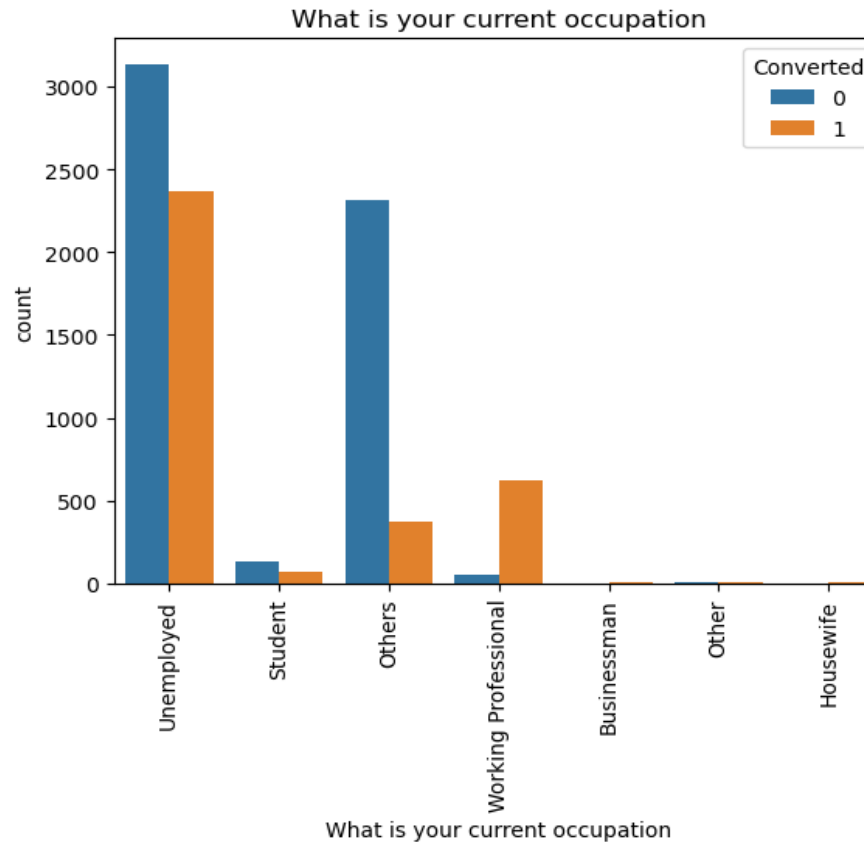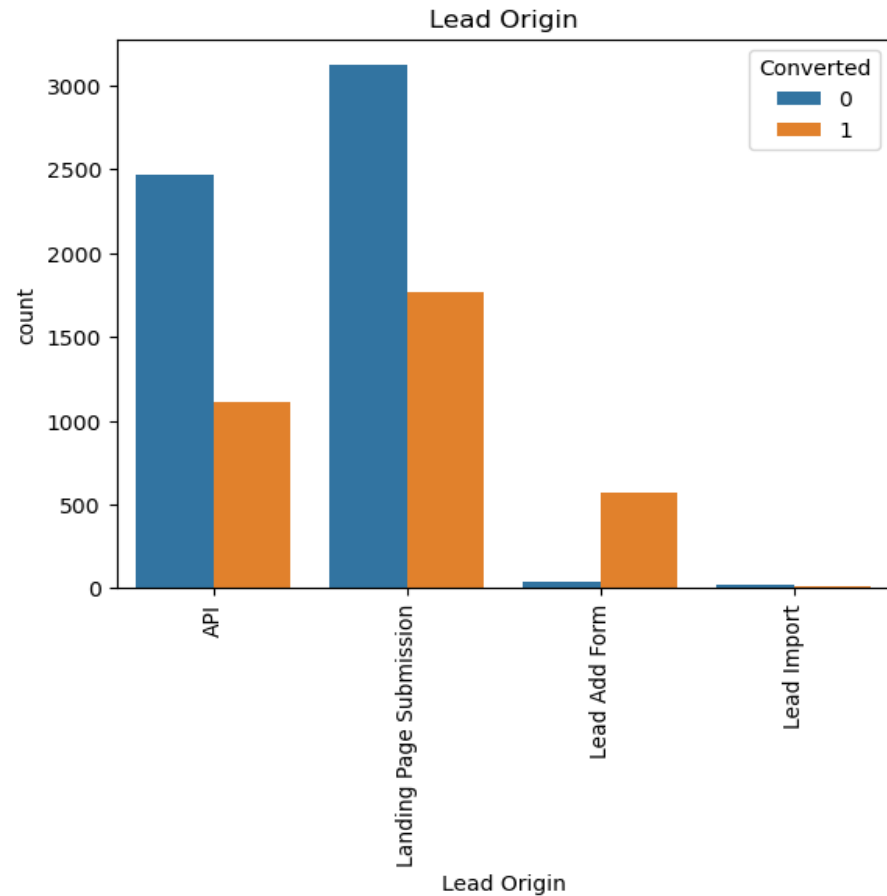Making predictions on the test set & test model evaluation

Determining top features & recommendation
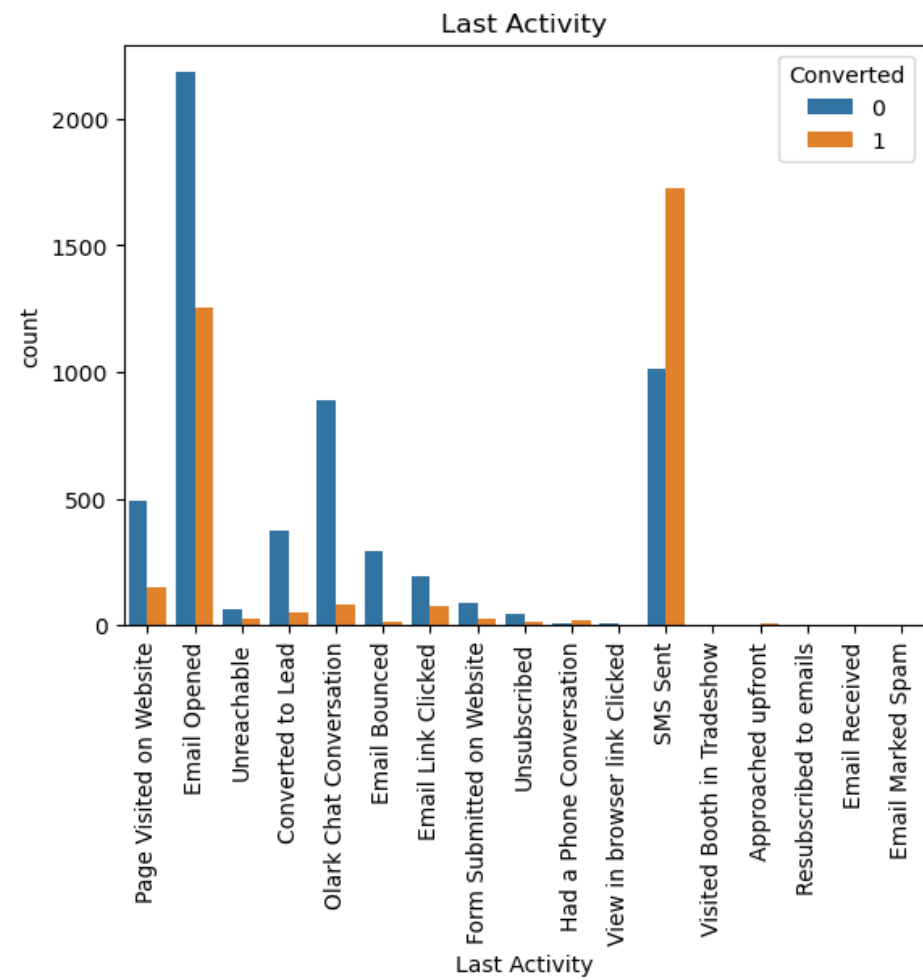
Exploratory Data Analysis

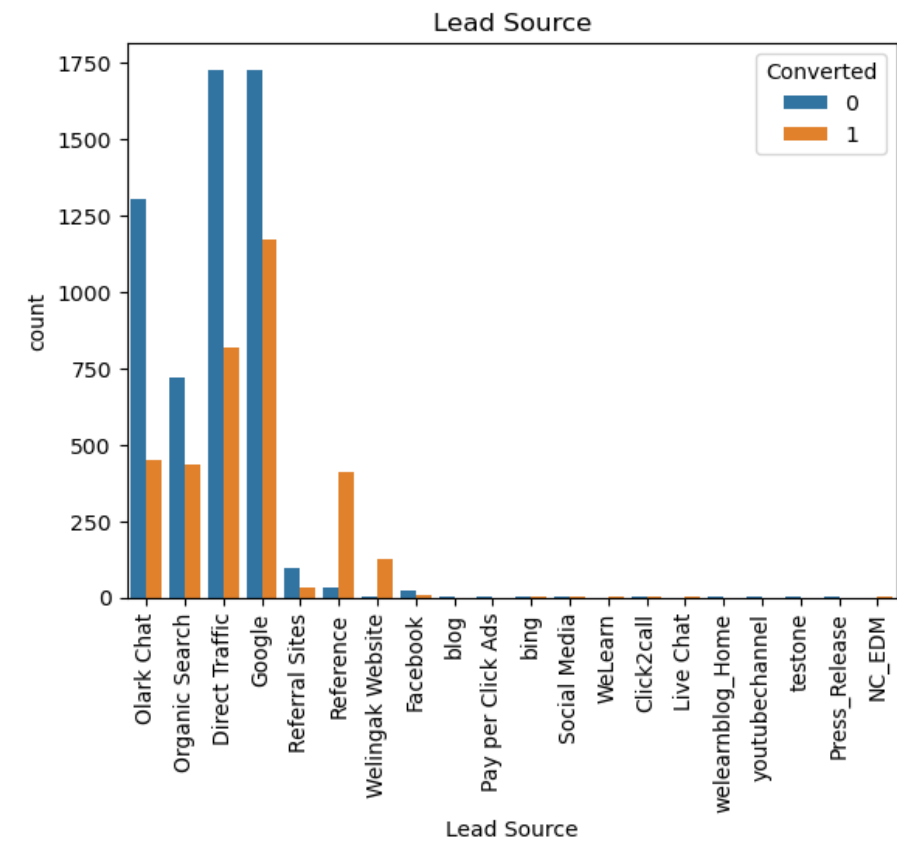Conversion rate is high for people who are visiting the page often or **spending more time** on the website

Exploratory Data Analysis

Converted leads surpass non-converted leads where lead origin is 'Ad form' and prospects who are 'Working Professionals'.

Exploratory Data Analysis

Converted leads surpass non-converted leads where lead source is 'Reference' and last activity is 'SMS sent'.

Final Model Results

```
==================================================================
Dep. Variable:              Converted    No. Observations:           6372
Model:                            GLM    Df Residuals:               6358
Model Family:                Binomial    Df Model:                     13
Link Function:                  Logit    Scale:                    1.0000
Method:                          IRLS    Log-Likelihood:           -2607.0
Date:                Tue, 21 Nov 2023    Deviance:                  5214.0
Time:                        17:27:29    Pearson chi2:            6.20e+03
No. Iterations:                     7    Pseudo R-squ. (CS):       0.3992
Covariance Type:            nonrobust
==================================================================
                                              coef    std err         z    P>|z|    [0.025    0.975]
------------------------------------------------------------------
const                                      -1.0988      0.055   -19.992    0.000    -1.206    -0.991
Total Time Spent on Website                 1.1153      0.040    27.651    0.000     1.036     1.194
Lead Origin_Lead Add Form                   3.6988      0.222    16.638    0.000     3.263     4.135
Lead Origin_Lead Import                     1.2577      0.445     2.824    0.005     0.385     2.131
Lead Source_Olark Chat                      1.2839      0.106    12.108    0.000     1.076     1.492
Lead Source_Welingak Website                1.8407      0.754     2.440    0.015     0.362     3.319
Last Activity_Converted to Lead            -1.3071      0.212    -6.153    0.000    -1.724    -0.891
Last Activity_Email Bounced                -1.9809      0.316    -6.277    0.000    -2.599    -1.362
Last Activity_Olark Chat Conversation      -1.3599      0.163    -8.362    0.000    -1.679    -1.041
What is your current occupation_Others     -1.1401      0.088   -12.971    0.000    -1.312    -0.968
What is your current occupation_Working Professional  2.5372  0.186  13.610  0.000  2.172  2.903
Last Notable Activity_Had a Phone Conversation  3.0168  1.171   2.576    0.010     0.721     5.312
Last Notable Activity_SMS Sent              1.3185      0.079    16.594    0.000     1.163     1.474
Last Notable Activity_Unreachable           1.9013      0.515     3.692    0.000     0.892     2.911
==================================================================
```

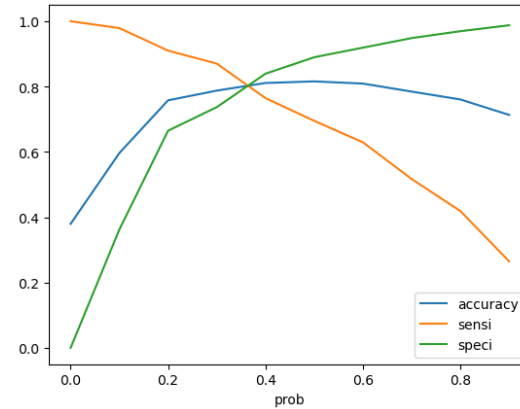| Features	VIF
Lead Source_Olark Chat	1.65
Lead Origin_Lead Add Form	1.50
Last Activity_Olark Chat Conversation	1.40
Lead Source_Welingak Website	1.32
What is your current occupation_Others	1.28
Total Time Spent on Website	1.27
Last Notable Activity_SMS Sent	1.19
What is your current occupation_Working Profes...	1.15
Last Activity_Email Bounced	1.03
Last Activity_Converted to Lead	1.02
Lead Origin_Lead Import	1.00
Last Notable Activity_Had a Phone Conversation	1.00
Last Notable Activity_Unreachable	1.00

All p-values were observed to be <0.05 (almost all are 0s) and VIF < 5
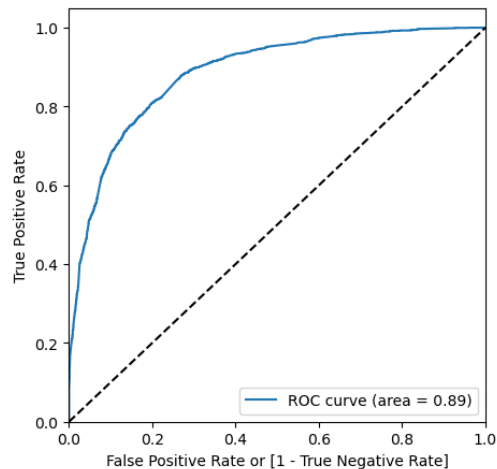
Training Model Evaluation

Confusion Matrix
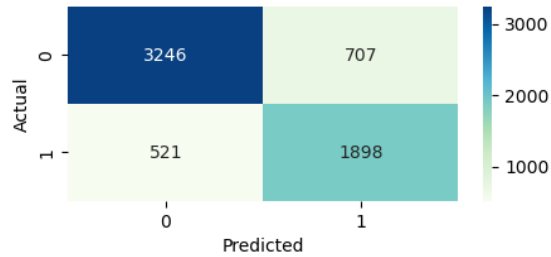
Accuracy Sensitivity and Specificity Plot

ROC Curve

Precision Recall curve

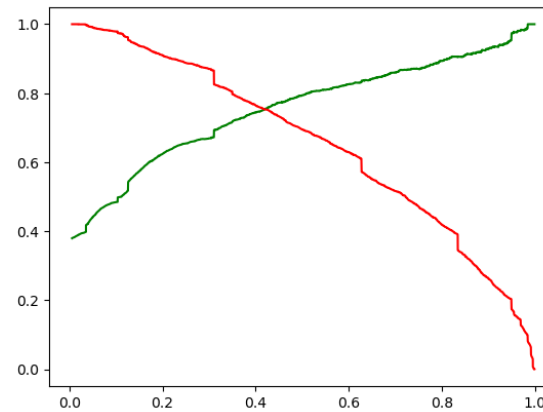**Training model metrics:**
- Accuracy: 81%
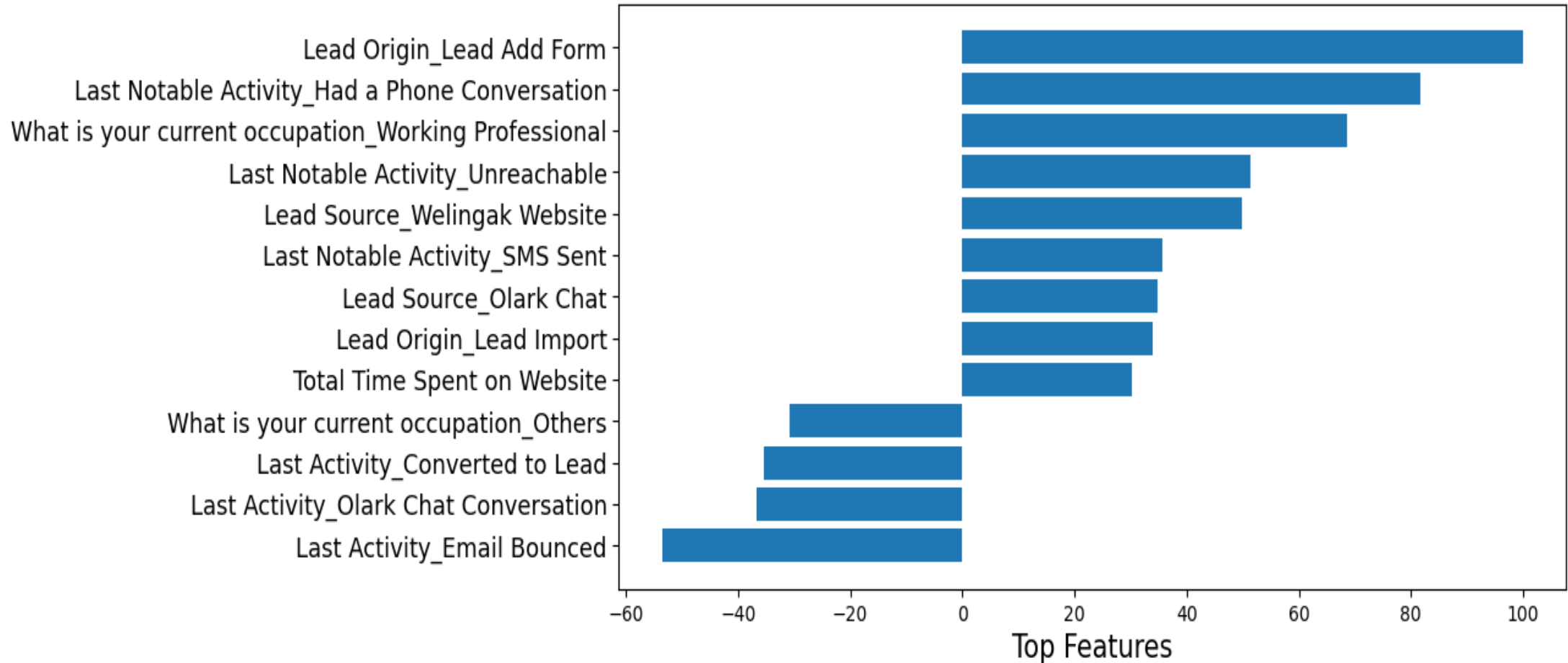- Sensitivity: 78%
- Specificity: 82%
- Precision: 73%
- Recall: 78%

Test model metrics:
- Accuracy: 81%
- Sensitivity: 79%
- Specificity: 82%
- Precision: 73%
- Recall: 79%

Top Features

Recommendations

X Education should focus on:
- Lead Origin: Ad form
- Profession: Working professionals
- Lead Source: Welingak website, Olark chat
- Last notable activity: Phone conversation, SMS sent
- Total time spent on website

- More budget/spend can be done on Welingak Website in terms of advertising, etc.
- Incentives/discounts for providing reference that convert to lead, encourage to provide more references.
- Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too

Keeping these in mind X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses