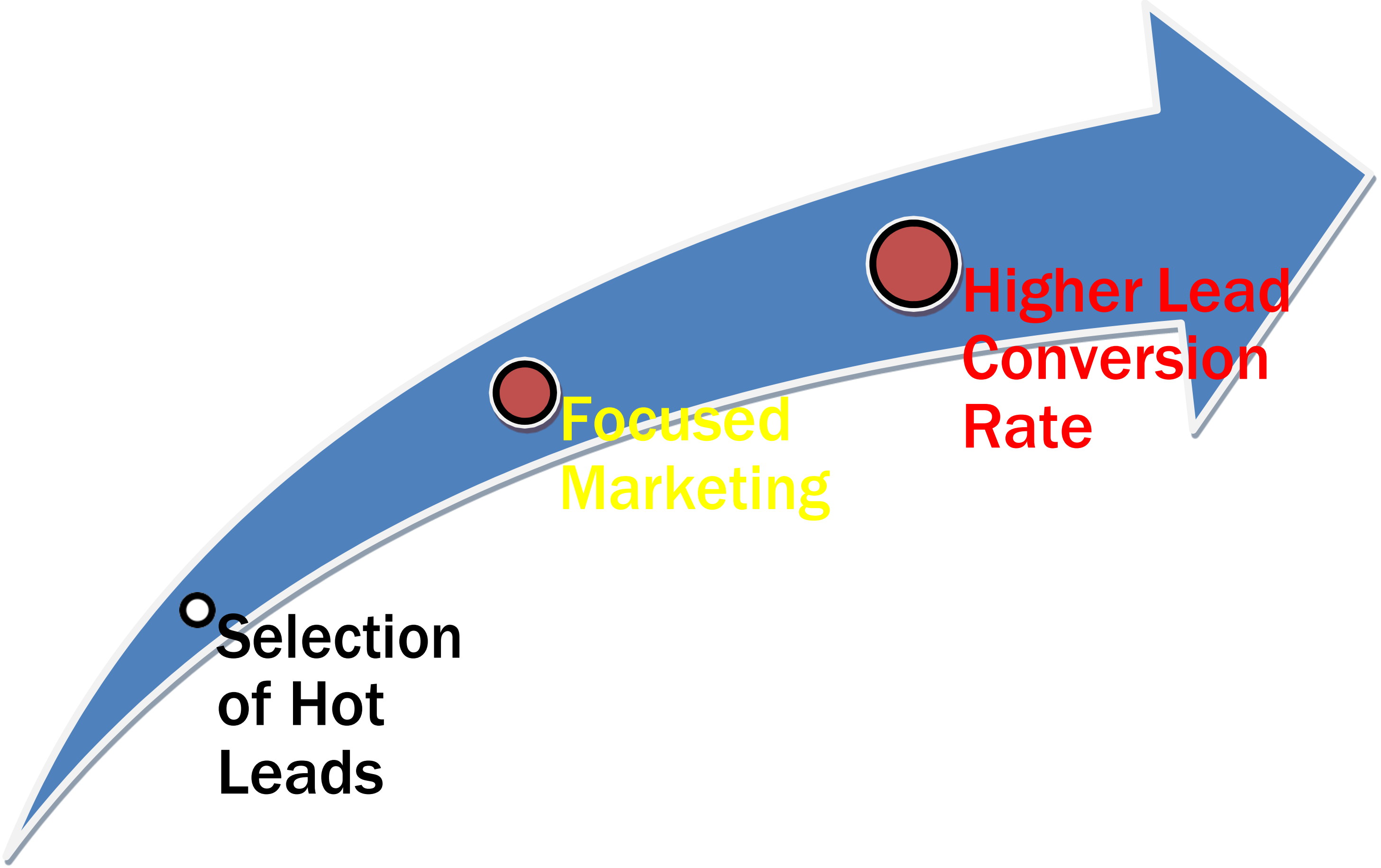
LEAD SCORING CASE STUDY

Problem Statement:

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a 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%.

Business Objective

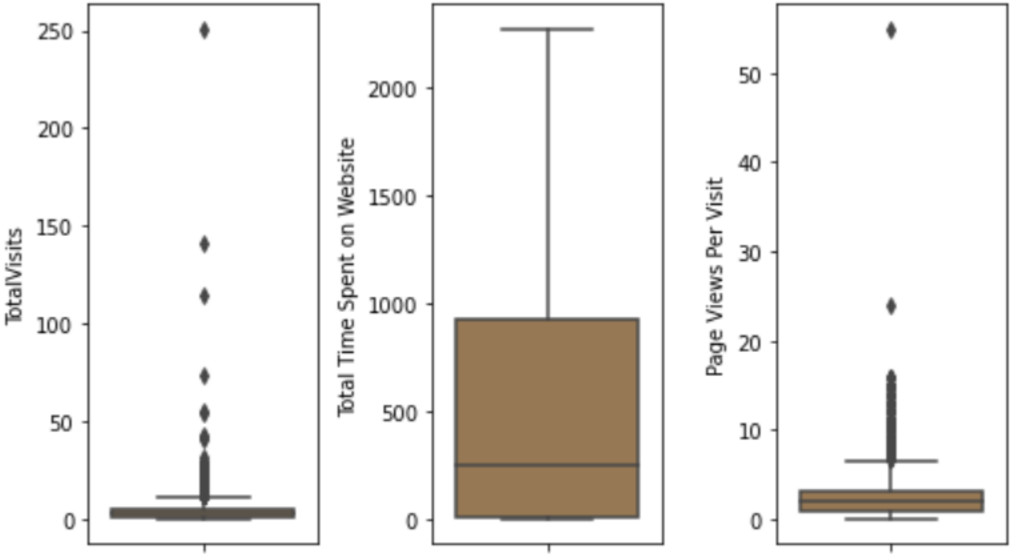
To help X Education select most promising leads ***(Hot Leads)***, i.e. the leads that are most likely to convert into paying customers.



DATA VISUALIZATION

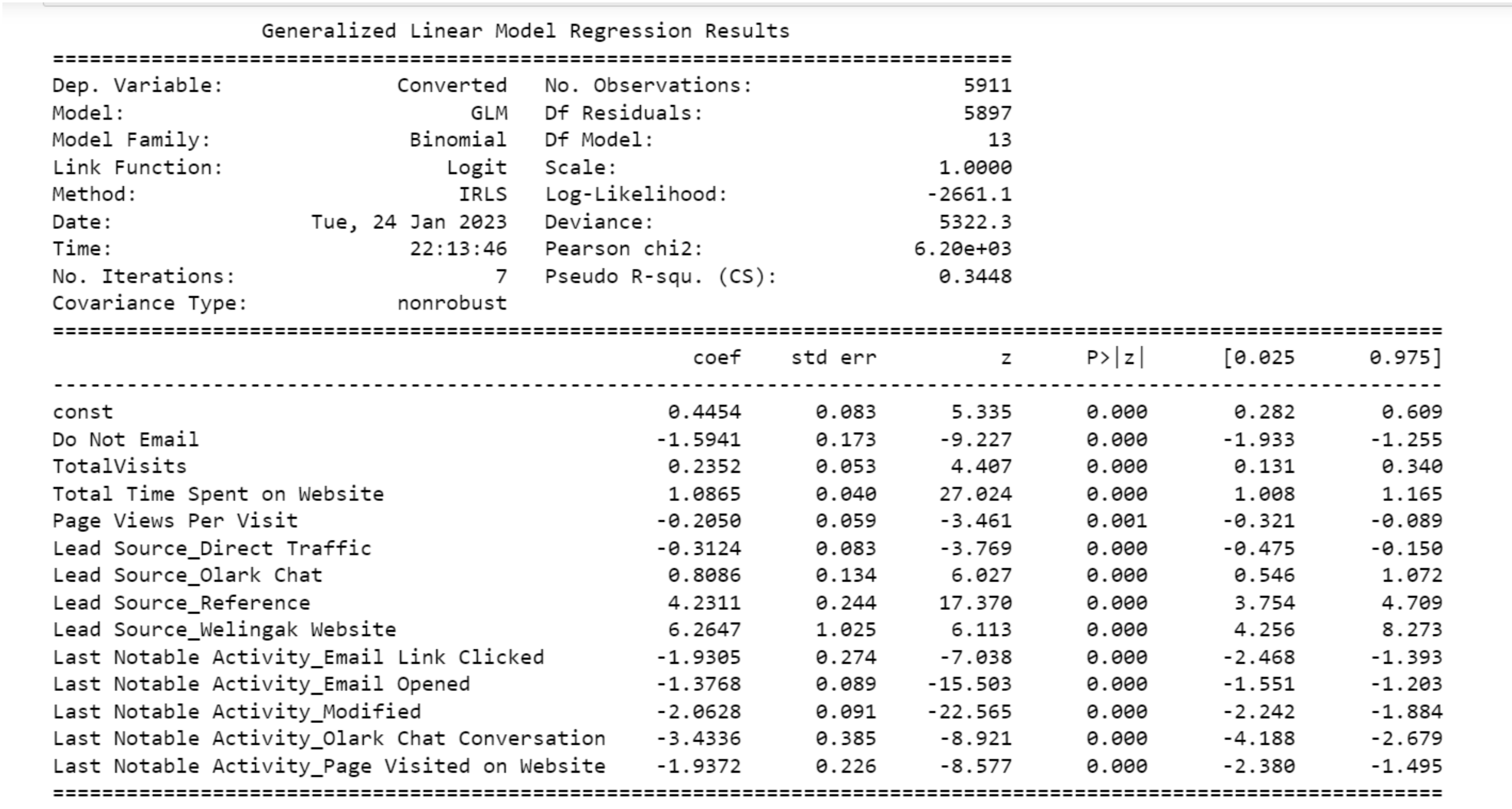
- To identify important features
- To get insights

Numerical Variables



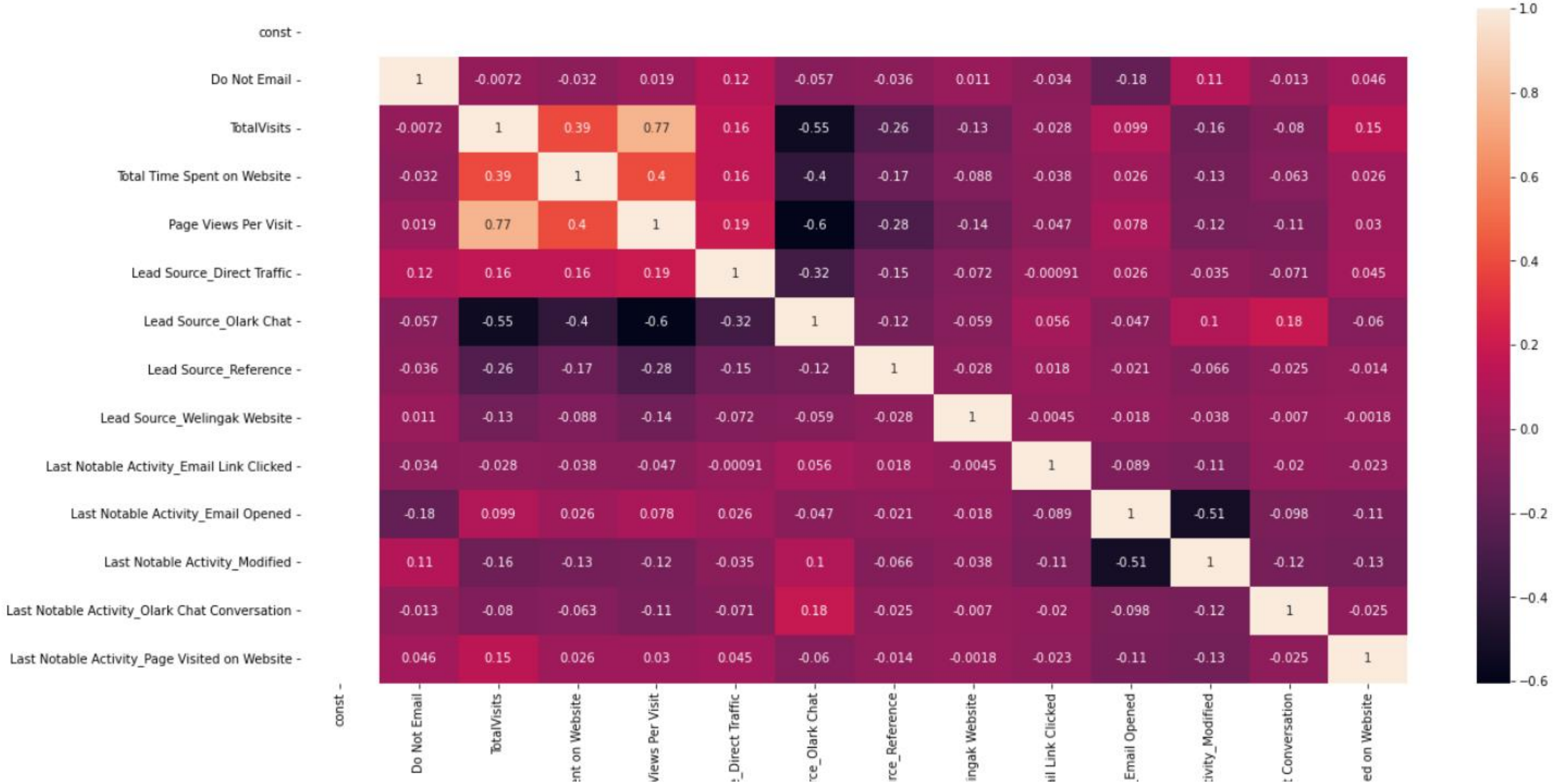
People spending more time on website are more likely to get converted.

MODEL EVALUATION



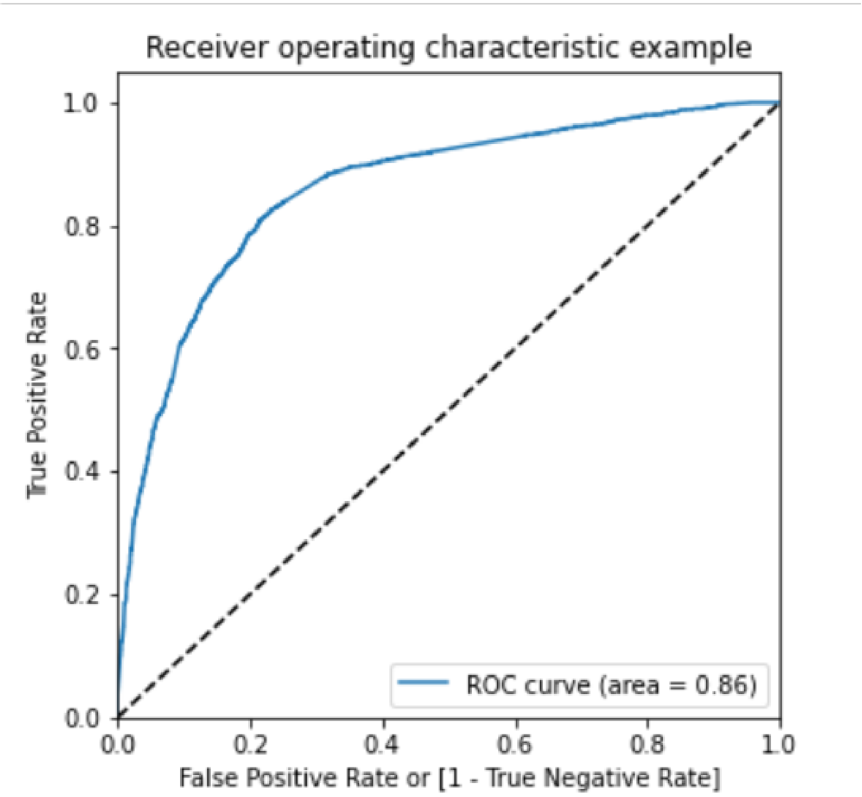
```
                 Generalized Linear Model Regression Results                  
==============================================================================
Dep. Variable:              Converted   No. Observations:                 5911
Model:                            GLM   Df Residuals:                     5897
Model Family:                Binomial   Df Model:                           13
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2661.1
Date:                Tue, 24 Jan 2023   Deviance:                       5322.3
Time:                        22:13:46   Pearson chi2:                 6.20e+03
No. Iterations:                     7   Pseudo R-squ. (CS):             0.3448
Covariance Type:            nonrobust                                         
=================================================================================================================
                                                    coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------------
const                                             0.4454      0.083      5.335      0.000       0.282       0.609
Do Not Email                                     -1.5941      0.173     -9.227      0.000      -1.933      -1.255
TotalVisits                                       0.2352      0.053      4.407      0.000       0.131       0.340
Total Time Spent on Website                       1.0865      0.040     27.024      0.000       1.008       1.165
Page Views Per Visit                             -0.2050      0.059     -3.461      0.001      -0.321      -0.089
Lead Source_Direct Traffic                       -0.3124      0.083     -3.769      0.000      -0.475      -0.150
Lead Source_Olark Chat                            0.8086      0.134      6.027      0.000       0.546       1.072
Lead Source_Reference                             4.2311      0.244     17.370      0.000       3.754       4.709
Lead Source_Welingak Website                      6.2647      1.025      6.113      0.000       4.256       8.273
Last Notable Activity_Email Link Clicked         -1.9305      0.274     -7.038      0.000      -2.468      -1.393
Last Notable Activity_Email Opened               -1.3768      0.089    -15.503      0.000      -1.551      -1.203
Last Notable Activity_Modified                   -2.0628      0.091    -22.565      0.000      -2.242      -1.884
Last Notable Activity_Olark Chat Conversation    -3.4336      0.385     -8.921      0.000      -4.188      -2.679
Last Notable Activity_Page Visited on Website    -1.9372      0.226     -8.577      0.000      -2.380      -1.495
=================================================================================================================
```

Final Model Summary: All p-values are zero



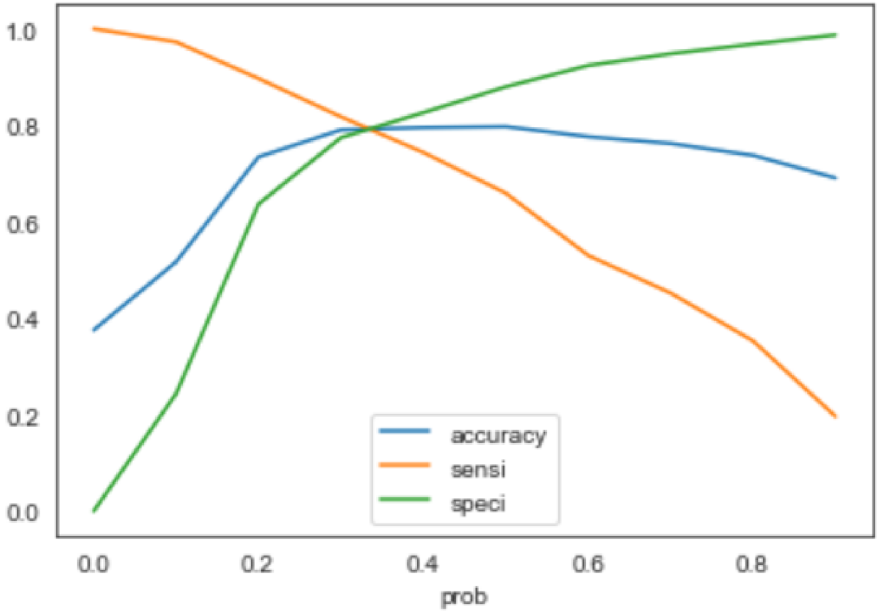
Correlations between features in the final model are **negligible.**

ROC curve



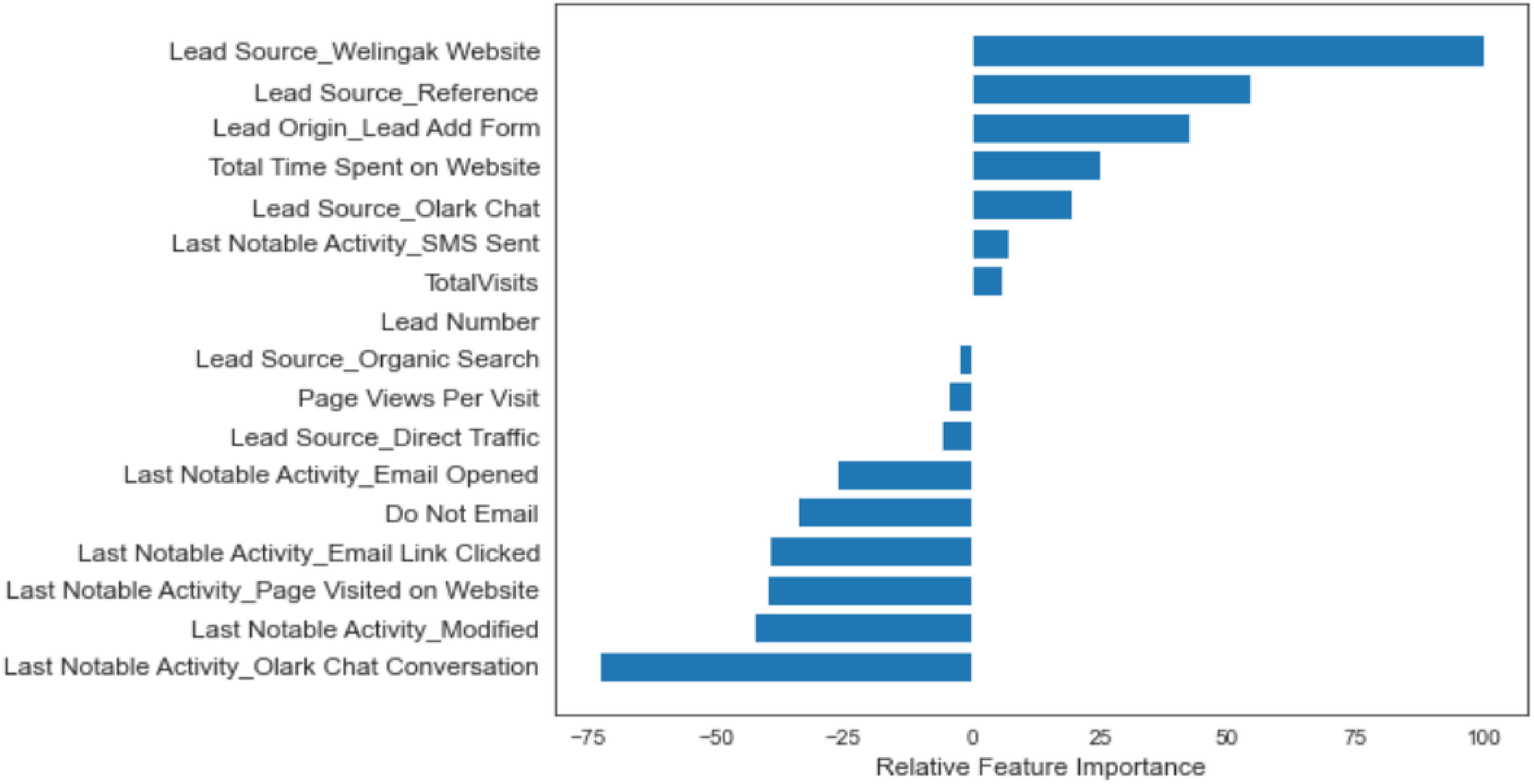
Area under curve = 0.86

Finding Optimal Threshold



Graph showing changes in Sensitivity, Specificity and Accuracy with changes in the probability threshold values
Optimal cutoff = 0.30

Relative Importance Of Features


Lead Source_Welingak Website
Lead Source_Reference
Lead Origin_Lead Add Form
Total Time Spent on Website
Lead Source_Olark Chat
Last Notable Activity_SMS Sent
TotalVisits
Lead Number
Lead Source_Organic Search
Page Views Per Visit
Lead Source_Direct Traffic
Last Notable Activity_Email Opened
Do Not Email
Last Notable Activity_Email Link Clicked
Last Notable Activity_Page Visited on Website
Last Notable Activity_Modified
Last Notable Activity_Olark Chat Conversation
Relative Feature Importance