X Education – Lead Scoring Case Study

Building a Logistic Regression Model to filter out the HOT Leads to focus more on them and thus enhancing the Conversion Ratio for X Education Company

Background

X Education Company

❑ An education company named X Education sells online courses to industry professionals

❑ Many interested professionals land on their website

❑ The company markets its courses on several websites like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos

Background

X Education Company

- ❑ When these people fill up a form providing their email address or phone number, they are classified to be a LEAD

- ❑ Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not

- ❑ The typical lead conversion rate at X education is around 30%

Problem Statement

- ❑ X Education gets a lot of leads but its lead conversion rate is very poor

- ❑ 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

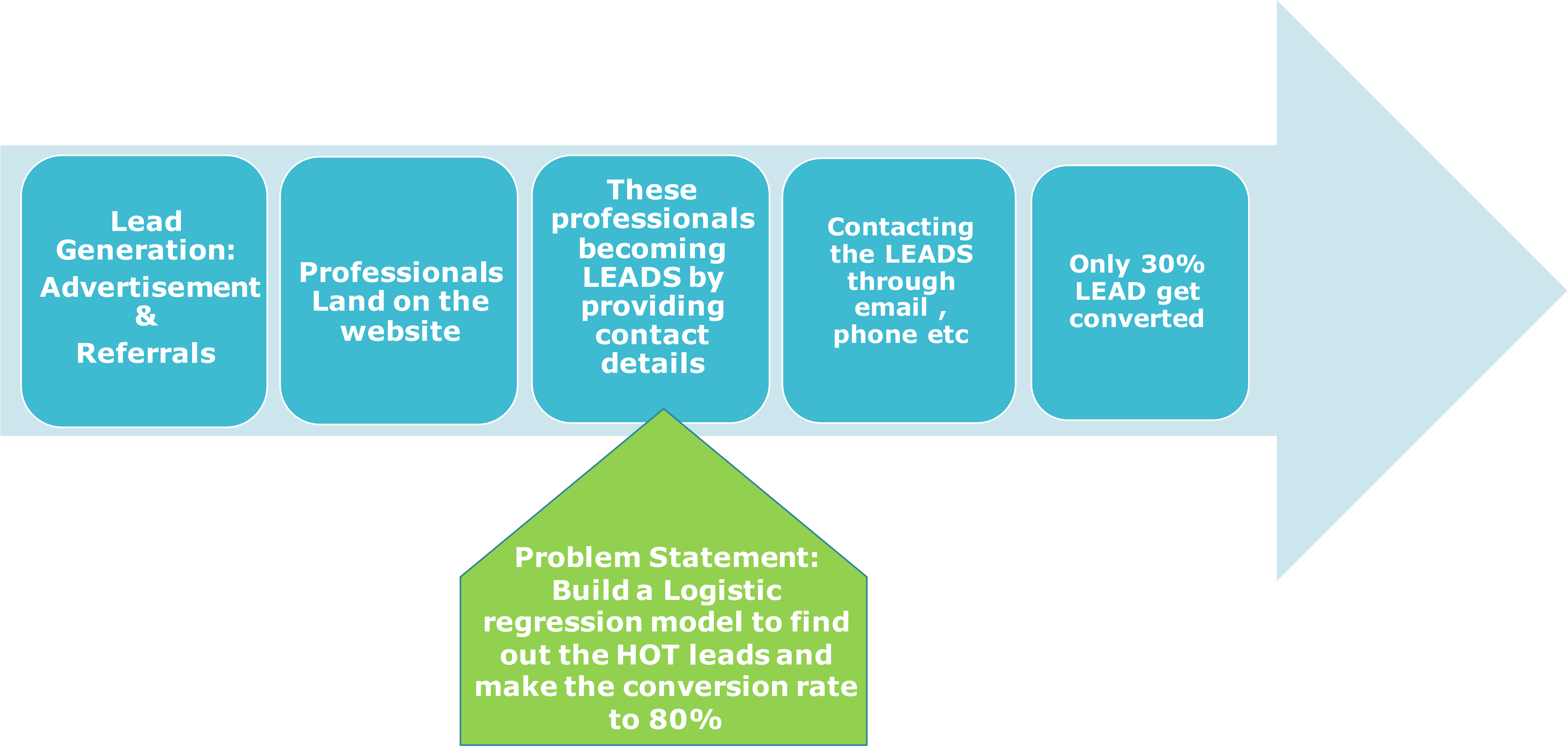
- We will help them to select the most promising leads, i.e. the leads that are most likely to convert into paying customers

- We are required to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance

- The CEO, in particular, has given a ballpark of the target lead conversion rate to be 80%.

How the things are working ?

Flow of LEAD Conversion



Lead Generation:
Advertisement
&
Referrals

Professionals Land on the website

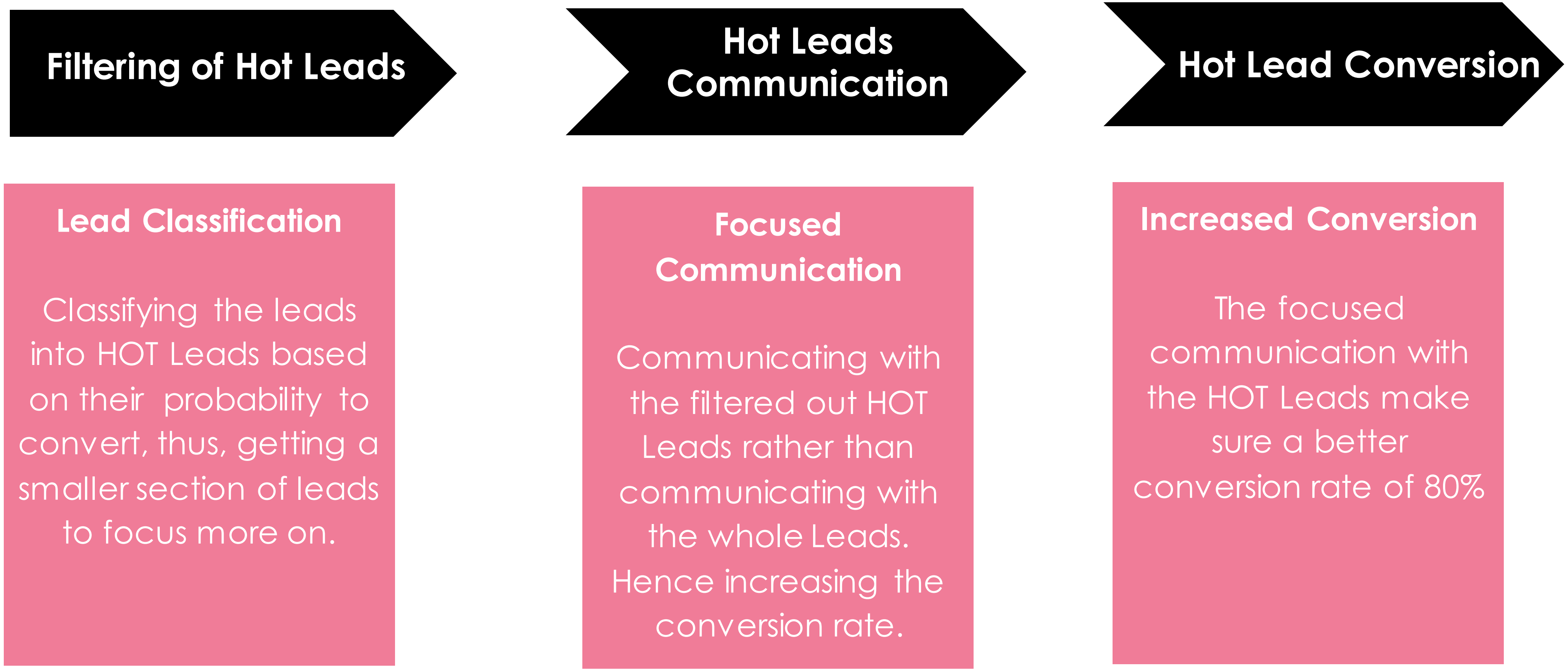
These professionals becoming LEADS by providing contact details

Contacting the LEADS through email , phone etc

Only 30% LEAD get converted

Problem Statement: Build a Logistic regression model to find out the HOT leads and make the conversion rate to 80%

Proposed Solution



Filtering of Hot Leads

Lead Classification

Classifying the leads into HOT Leads based on their probability to convert, thus, getting a smaller section of leads to focus more on.

Hot Leads Communication

Focused Communication

Communicating with the filtered out HOT Leads rather than communicating with the whole Leads. Hence increasing the conversion rate.

Hot Lead Conversion

Increased Conversion

The focused communication with the HOT Leads make sure a better conversion rate of 80%

Solution

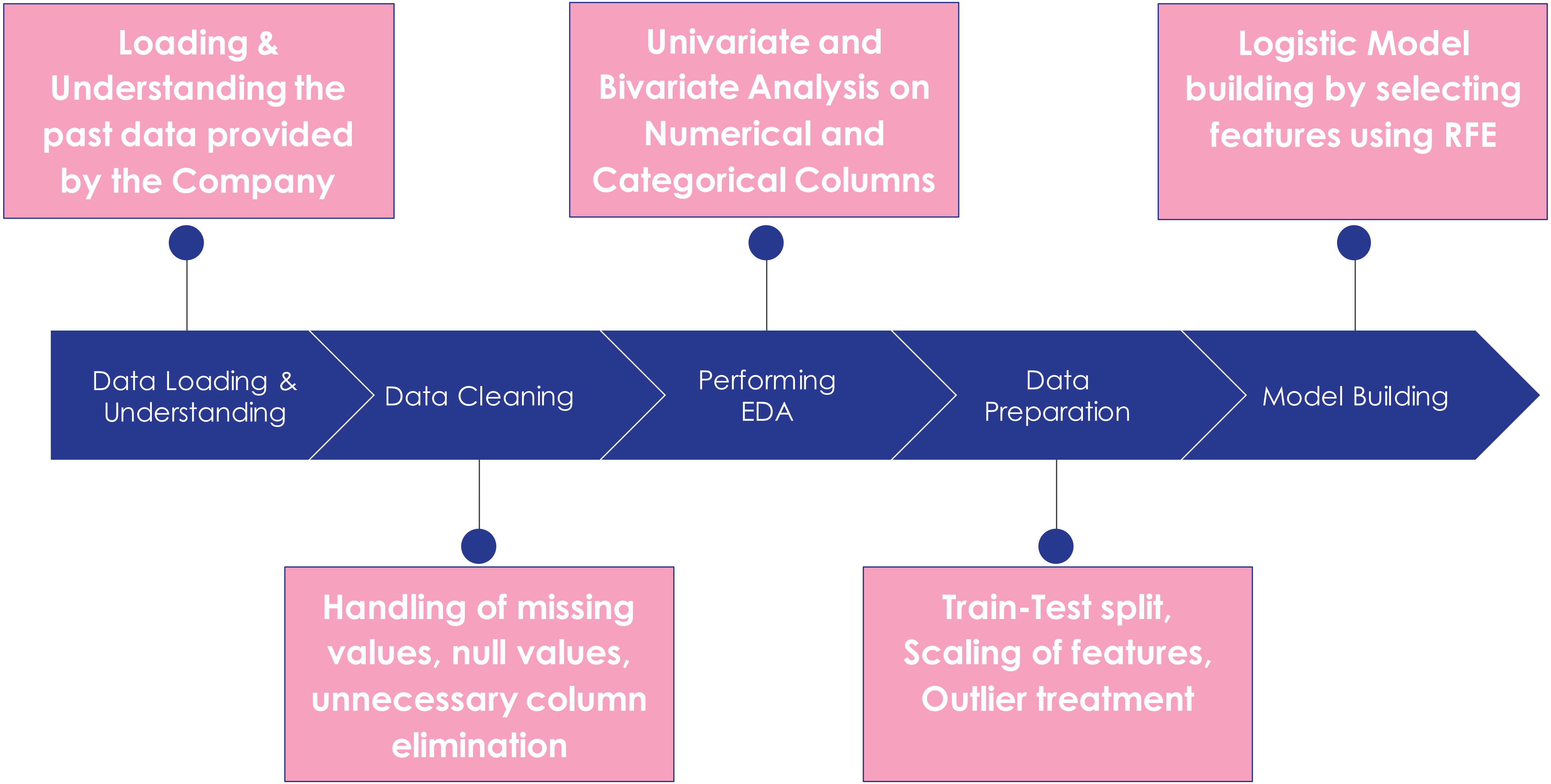
Selection of HOT Leads

- ❑ Filtering out the 'HOT Leads' by building a Logistic Regression Model

- ❑ In this business scenario we have to fil out the 80% of Actual HOT Leads correctly. Since the X Education company has a target of 80% conversion rate

- ❑ To make sure the Conversion rate of 80%, we have to build a model with high "Sensitivity"

FLOW OF IMPLEMENTATION



Data Loading & Understanding

Loading & Understanding the past data provided by the Company

Data Cleaning

Handling of missing values, null values, unnecessary column elimination

Performing EDA

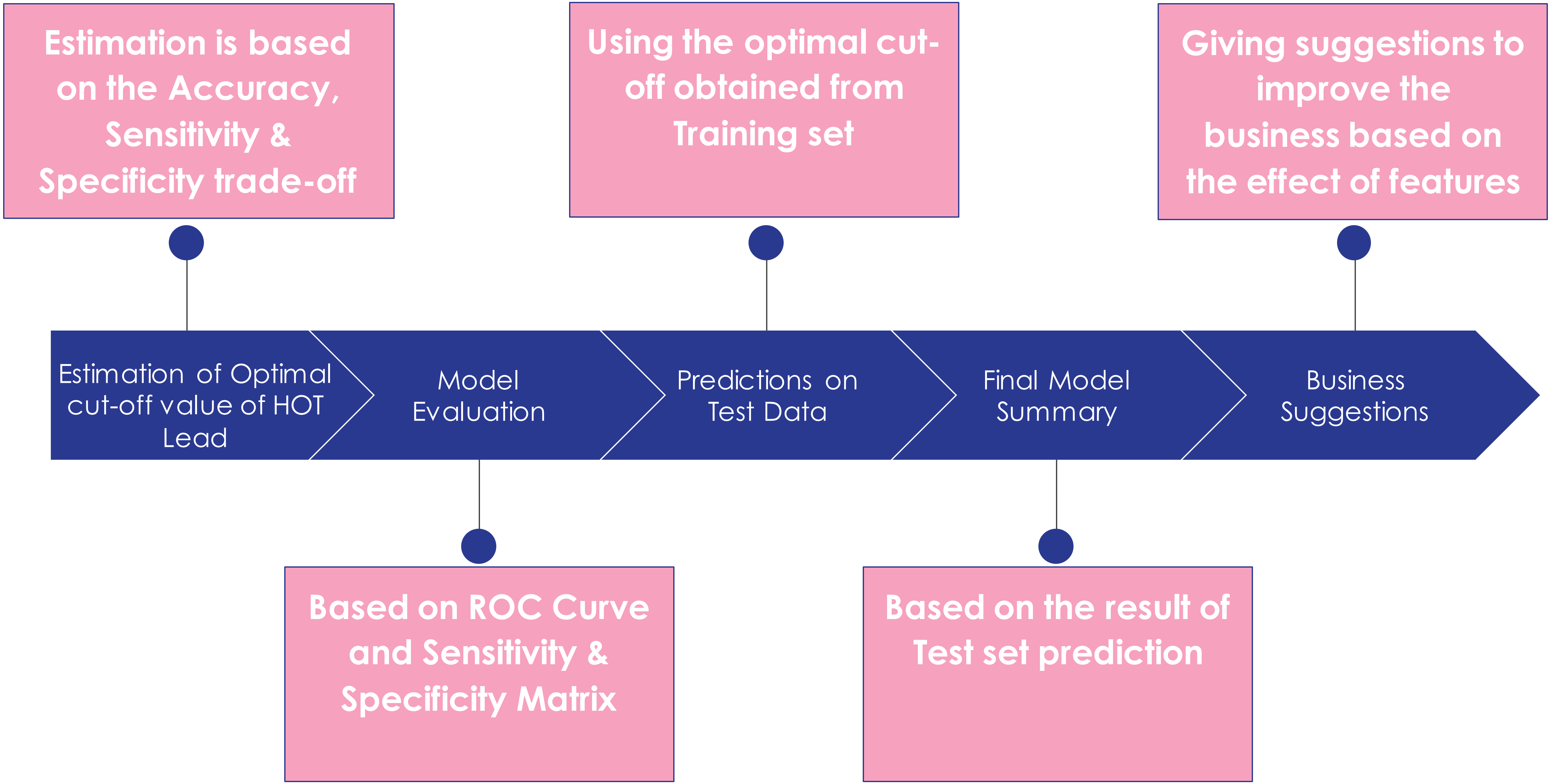
Univariate and Bivariate Analysis on Numerical and Categorical Columns

Data Preparation

Train-Test split, Scaling of features, Outlier treatment

Model Building

Logistic Model building by selecting features using RFE



Estimation of Optimal cut-off value of HOT Lead

Estimation is based on the Accuracy, Sensitivity & Specificity trade-off

Model Evaluation

Based on ROC Curve and Sensitivity & Specificity Matrix

Predictions on Test Data

Using the optimal cut-off obtained from Training set

Final Model Summary

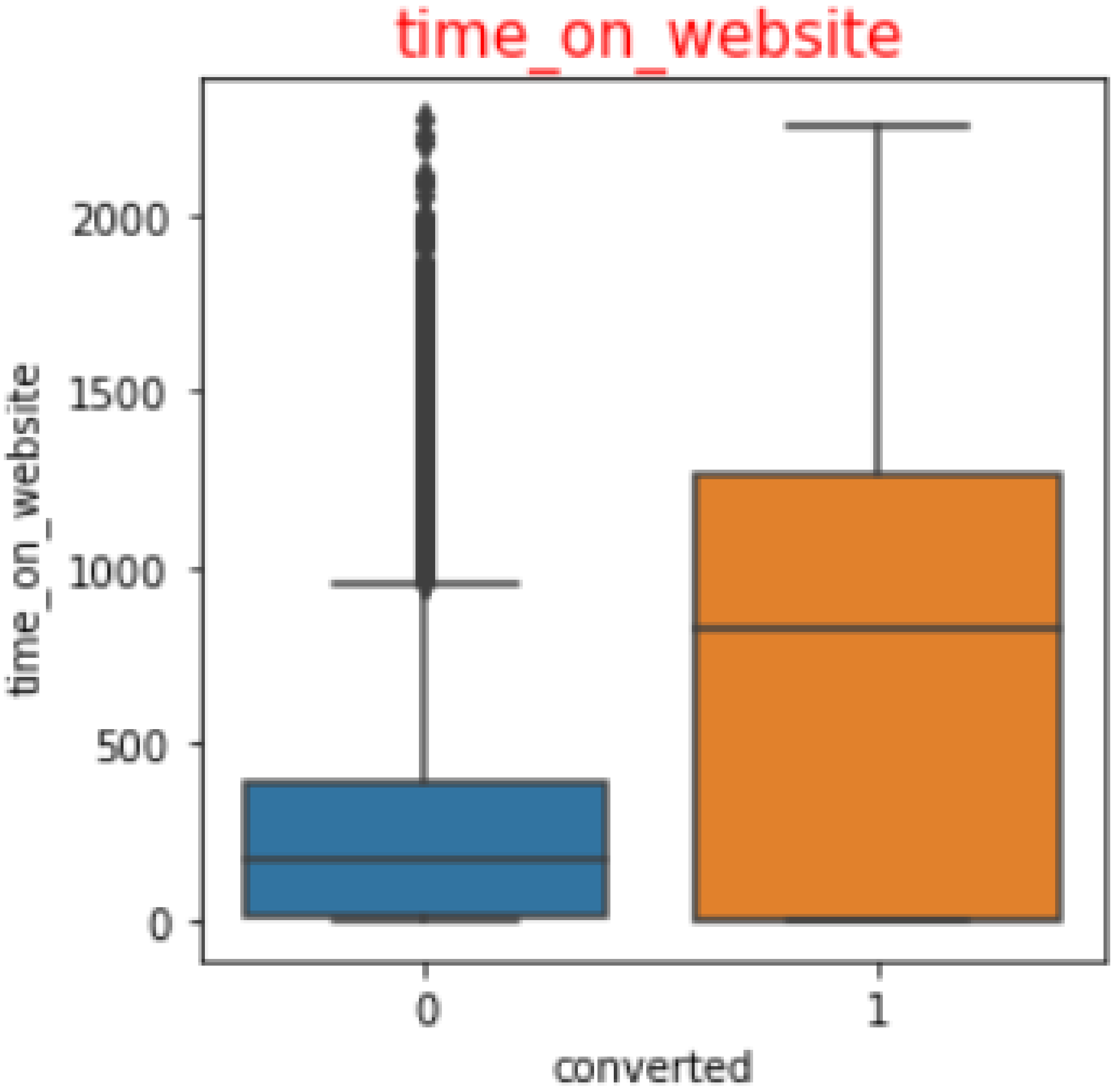
Based on the result of Test set prediction

Business Suggestions

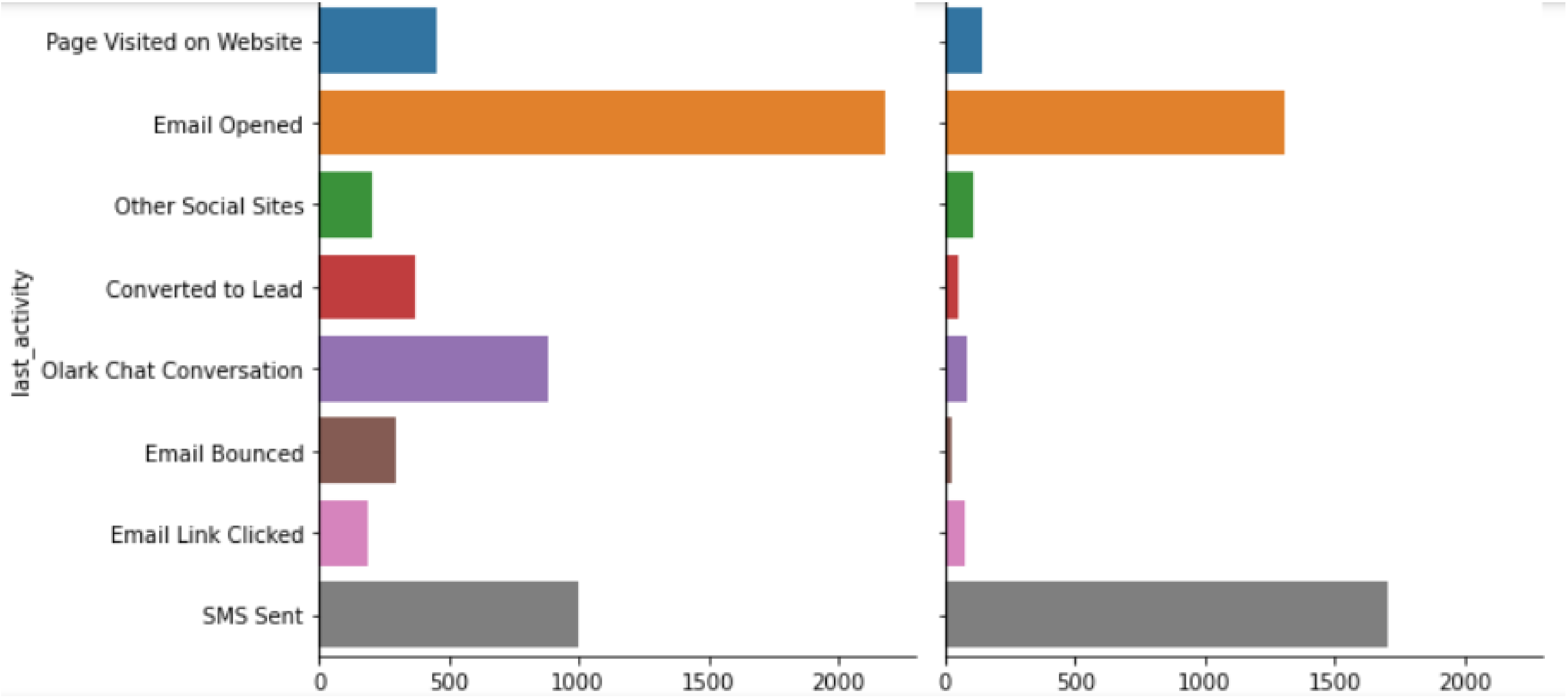
Giving suggestions to improve the business based on the effect of features

Insights From EDA

Total time spent on website & Target Variable



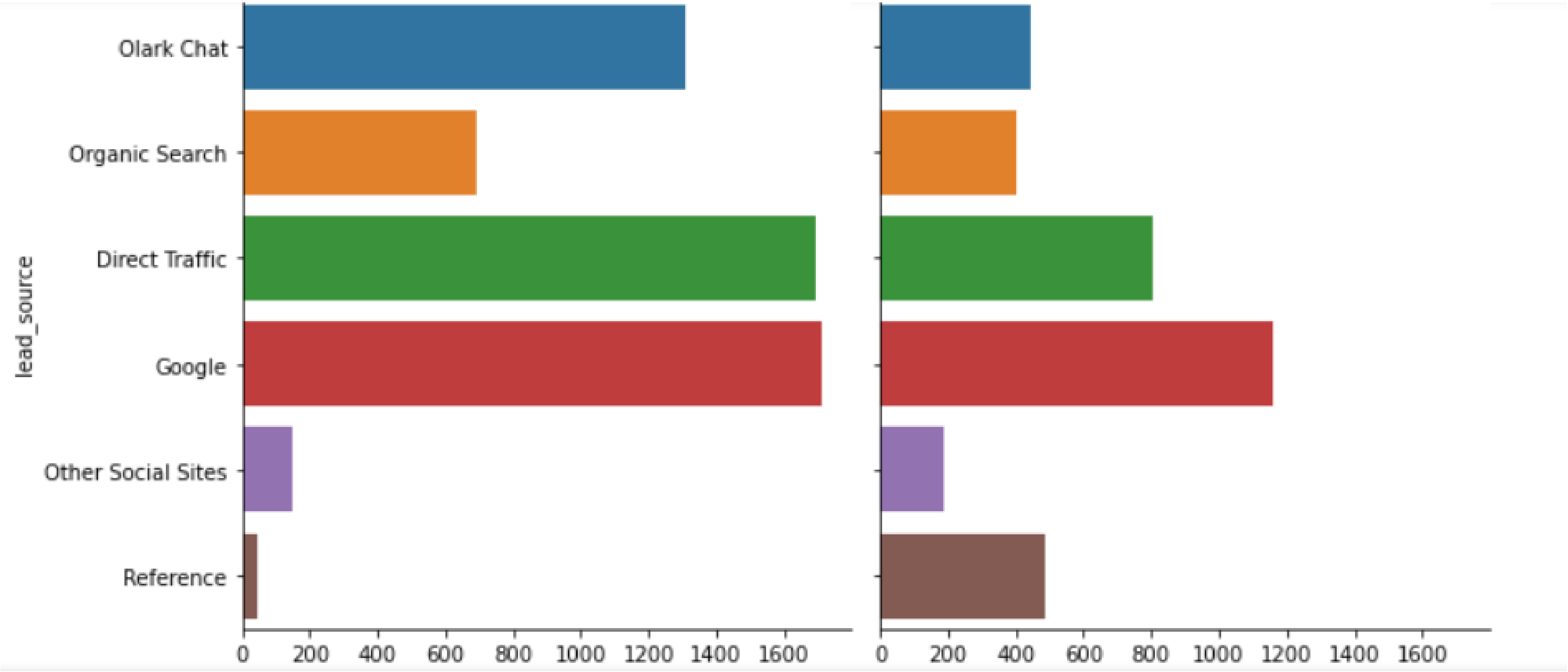
Last Activity & Target Variable



Convert = 0

Convert = 1

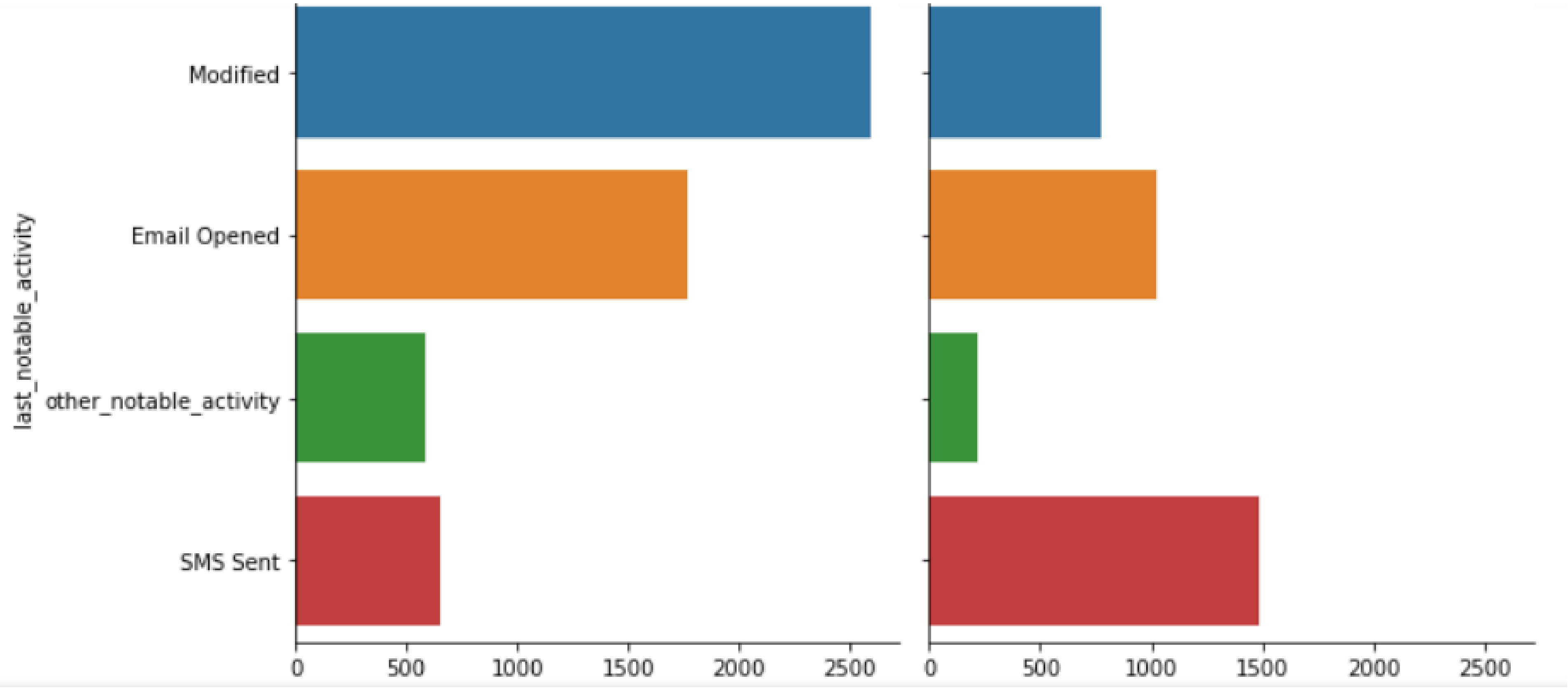
Lead Source & Target Variable



Convert = 0

Convert = 1

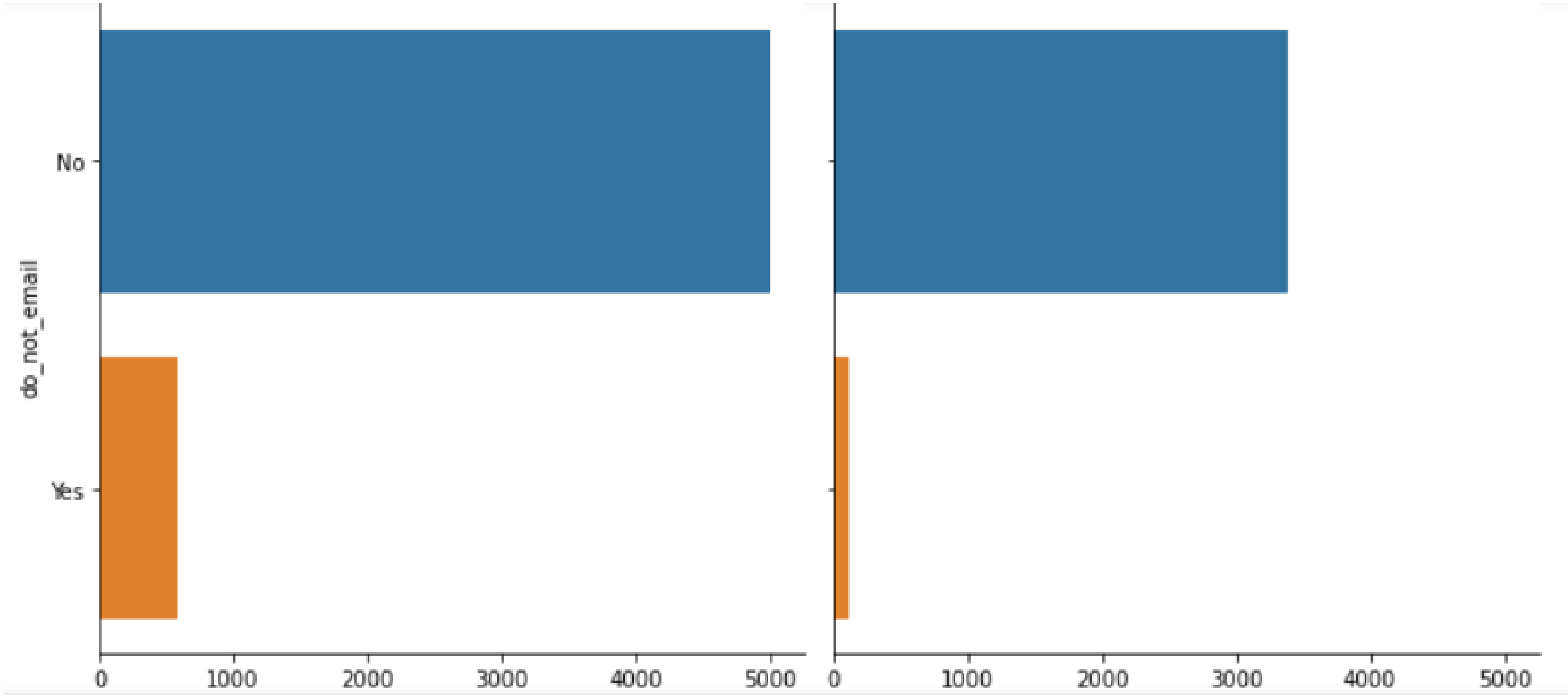
Last Notable Activity & Target Variable



Convert = 0

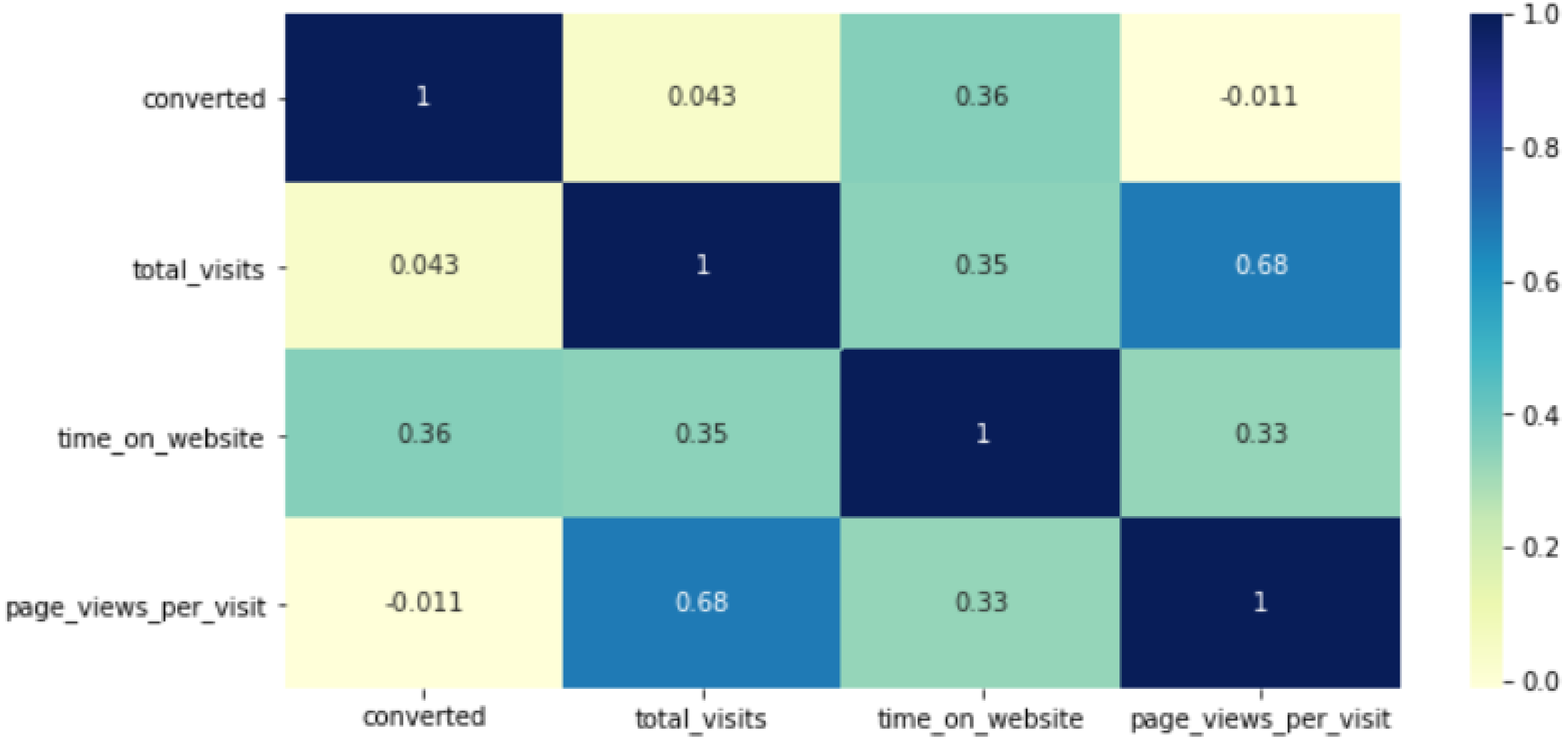
Convert = 1

Last Notable Activity & Target Variable



Convert = 0 **Convert = 1**

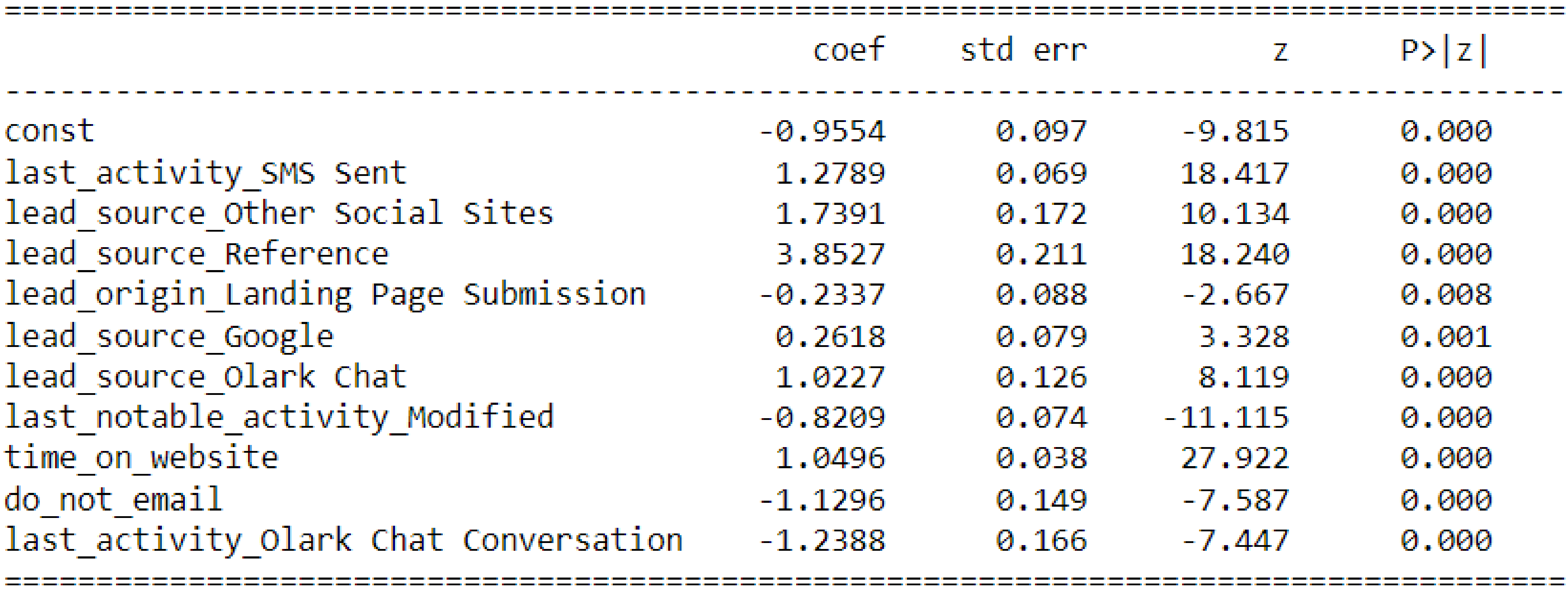
Heat Map – Correlation of all numeric columns



MODEL BUILDING

Significant Features of the Final Model

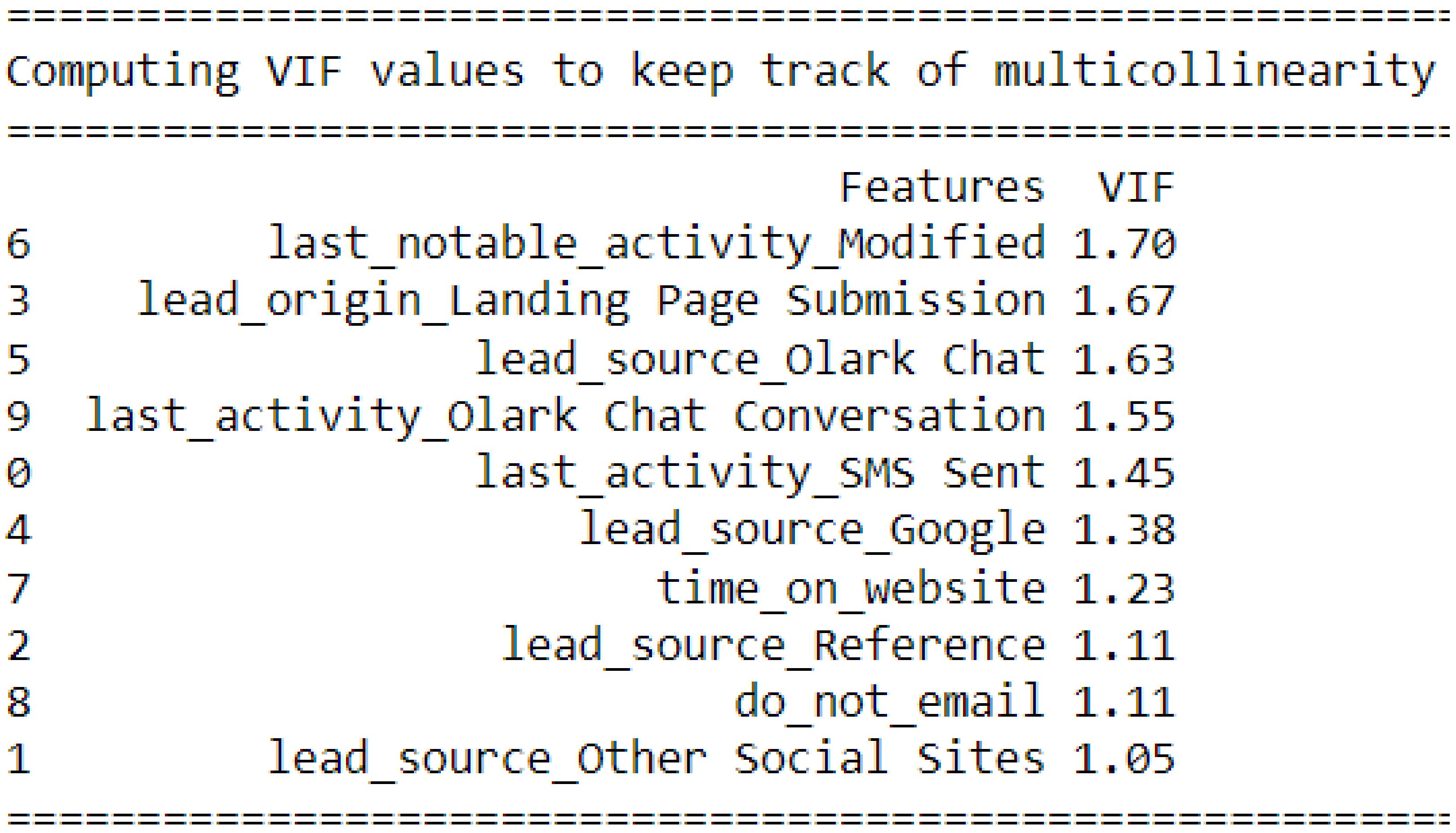
For all features the p – value is less than 0.05, which implies the features are significant



```
==========================================================================================
                                            coef    std err          z      P>|z|
------------------------------------------------------------------------------------------
const                                    -0.9554      0.097     -9.815      0.000
last_activity_SMS Sent                    1.2789      0.069     18.417      0.000
lead_source_Other Social Sites            1.7391      0.172     10.134      0.000
lead_source_Reference                     3.8527      0.211     18.240      0.000
lead_origin_Landing Page Submission      -0.2337      0.088     -2.667      0.008
lead_source_Google                        0.2618      0.079      3.328      0.001
lead_source_Olark Chat                    1.0227      0.126      8.119      0.000
last_notable_activity_Modified           -0.8209      0.074    -11.115      0.000
time_on_website                           1.0496      0.038     27.922      0.000
do_not_email                             -1.1296      0.149     -7.587      0.000
last_activity_Olark Chat Conversation    -1.2388      0.166     -7.447      0.000
==========================================================================================
```

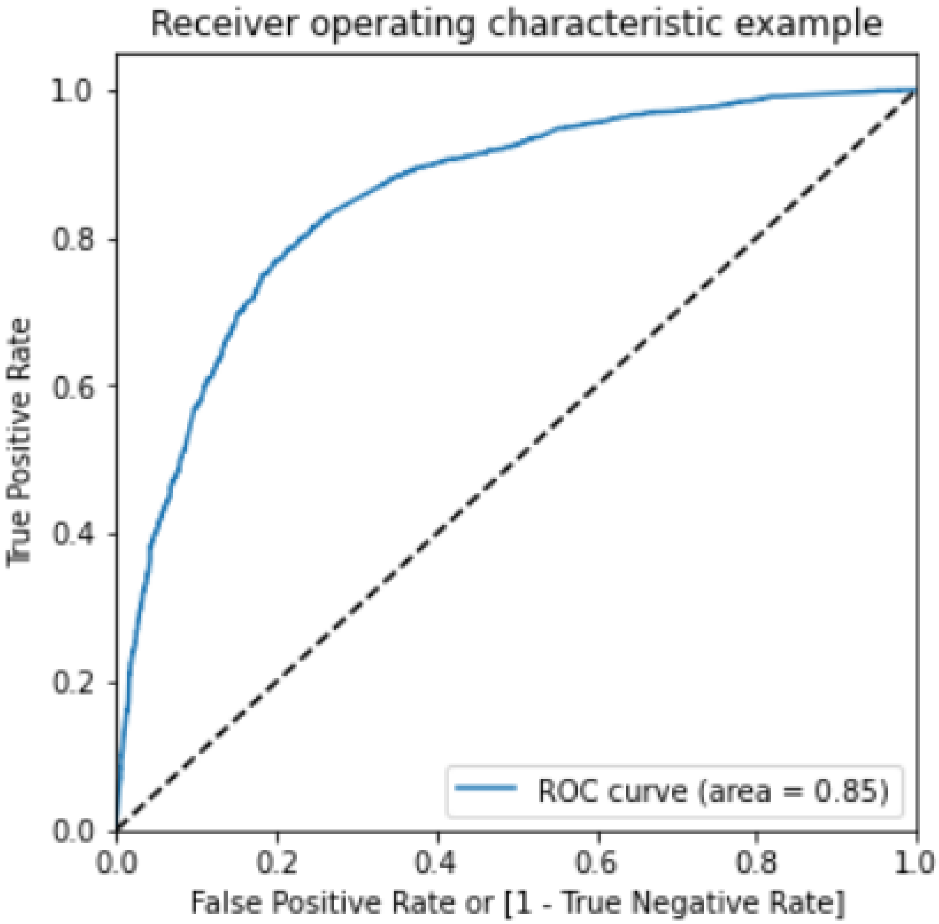
VIF of Final Model Features

For all features the VIF value is less than 2



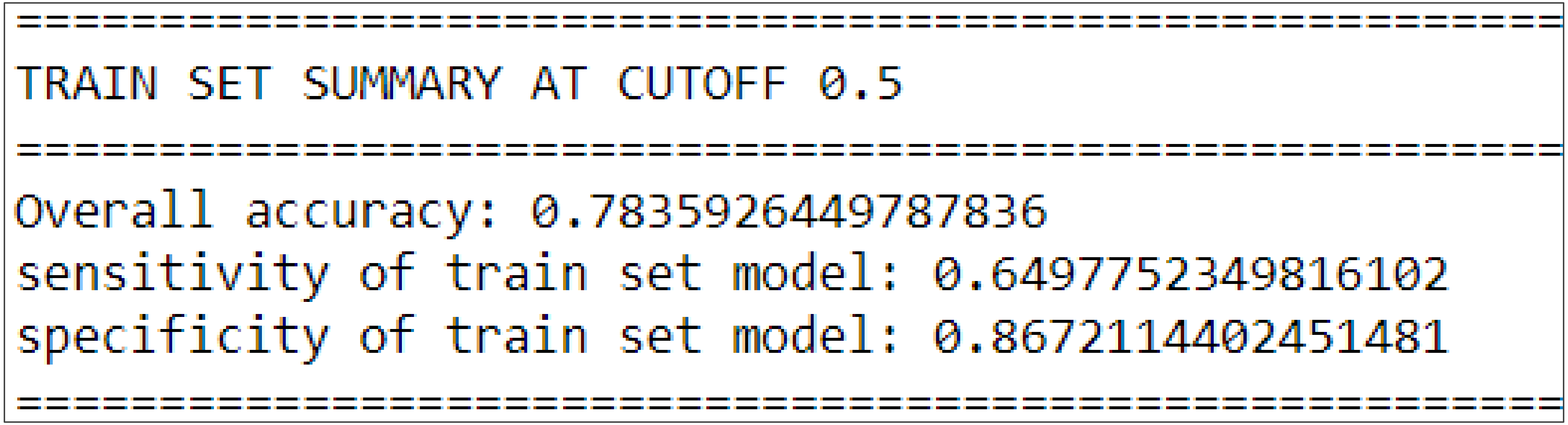
```
========================================================
Computing VIF values to keep track of multicollinearity
========================================================
                                Features  VIF
6         last_notable_activity_Modified 1.70
3    lead_origin_Landing Page Submission 1.67
5                 lead_source_Olark Chat 1.63
9  last_activity_Olark Chat Conversation 1.55
0                 last_activity_SMS Sent 1.45
4                     lead_source_Google 1.38
7                        time_on_website 1.23
2                  lead_source_Reference 1.11
8                           do_not_email 1.11
1         lead_source_Other Social Sites 1.05
========================================================
```

Receiver Operating Characteristic Curve of Final Model



Gini (Area under ROC Curve) - 0.85

Logistic Regression Final Model Parameters on Train set at the arbitrary Cut – off 0.5 :

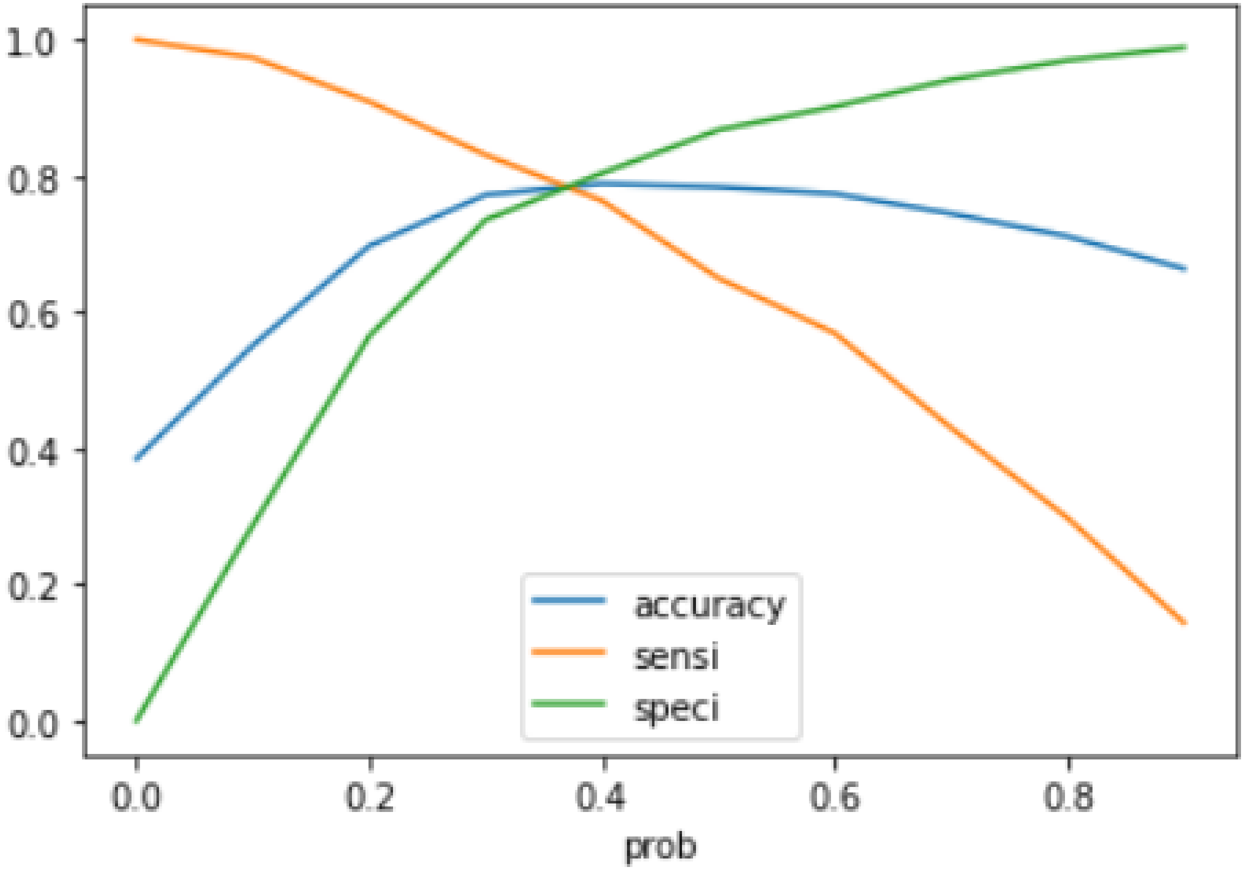


```
======================================================
TRAIN SET SUMMARY AT CUTOFF 0.5
======================================================
Overall accuracy: 0.7835926449787836
sensitivity of train set model: 0.6497752349816102
specificity of train set model: 0.8672114402451481
======================================================
```

Sensitivity of the model is Low at the arbitrary cut-off 0.5

Optimal cut-off of HOT Lead

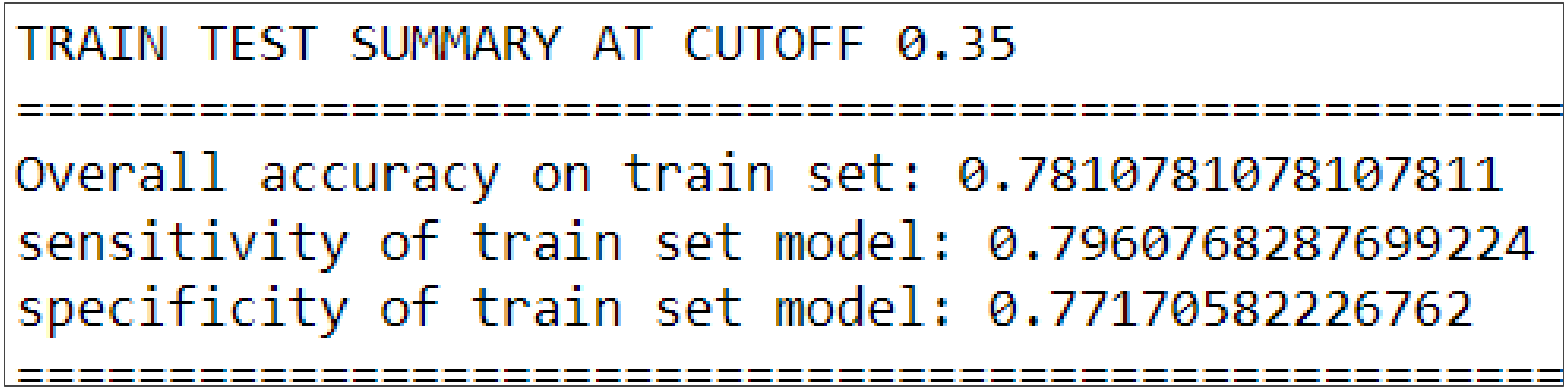
Accuracy, Sensitivity & Specificity Trade-off



Optimal cut-off of HOT Lead = 0.35

(Any Lead having Probability > 0.35 will be a HOT Lead)

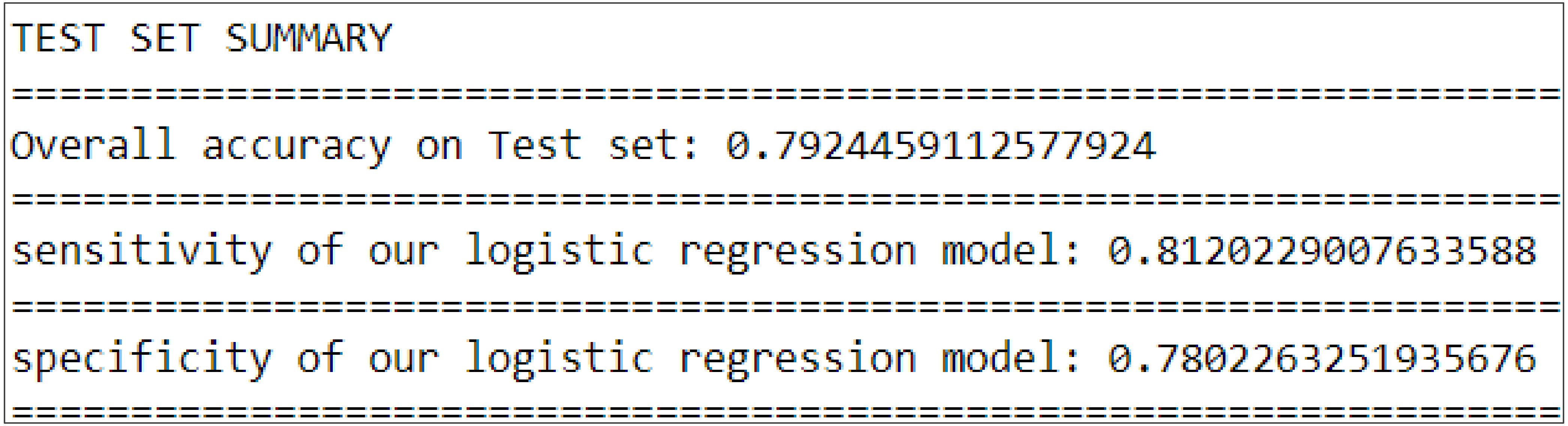
Logistic Regression Final Model Parameters on Train set at the Optimal Cut – off 0.35 :



```
TRAIN TEST SUMMARY AT CUTOFF 0.35
====================================================
Overall accuracy on train set: 0.7810781078107811
sensitivity of train set model: 0.7960768287699224
specificity of train set model: 0.7717058226762
====================================================
```

Sensitivity of the model is high almost 80% at the arbitrary cut-off 0.5

Logistic Regression Final Model Parameters on Test set



```
TEST SET SUMMARY
=================================================================
Overall accuracy on Test set: 0.7924459112577924
=================================================================
sensitivity of our logistic regression model: 0.8120229007633588
=================================================================
specificity of our logistic regression model: 0.7802263251935676
=================================================================
```

Sensitivity of the model on Test set is high 81.20%

Insights from the Final Model

Features Affecting The Lead Score

Time Spent on Website: Affecting positively

Last Activity SMS sent: Affecting positively

Lead Source-Reference: Affecting positively

Lead Source-Google: Affecting positively

Lead Source Other Social Media Sites: Affecting positively

Do not email: Affecting negatively

Olark Chat Conversation: Affecting negatively

Conclusions & Recommendations

HOT LEADS: The leads having probability greater than **0.35** are Hot Leads

Conversion rate increases with increase in the time spend on the website, therefore increase the user engagement in their wesite.

Try to give SMS notifications, since it improves the conversion rate

Use Email to Communicate with the Hot Leads

Improve the Olark conversation since it has a negative effect on Conversion Rate

Since reference has a positive effect on conversion provide better services to already converted leads to increase the reference

Improve the digital marketing to reach out to more people