Lead Scoring Case Study

X EDUCATION

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

Problem Statement

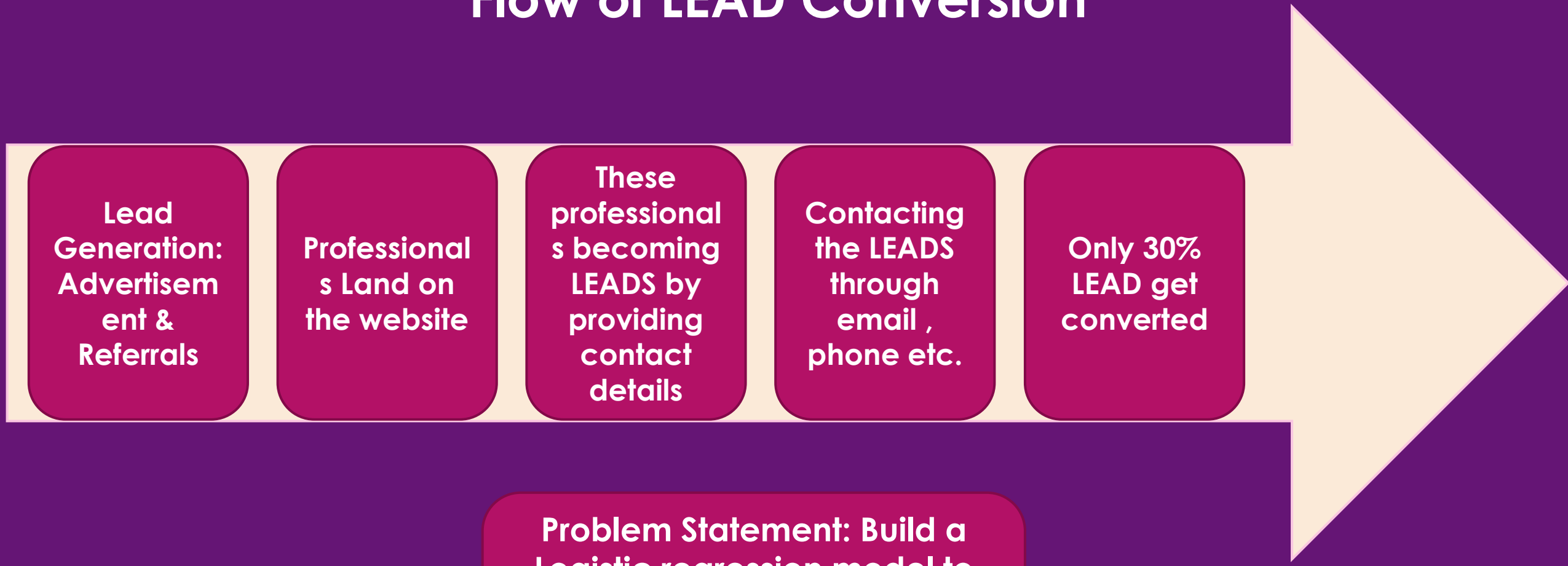
An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. 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%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.

Problem Statement

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 .There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion. 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%.

Flow of LEAD Conversion



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 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

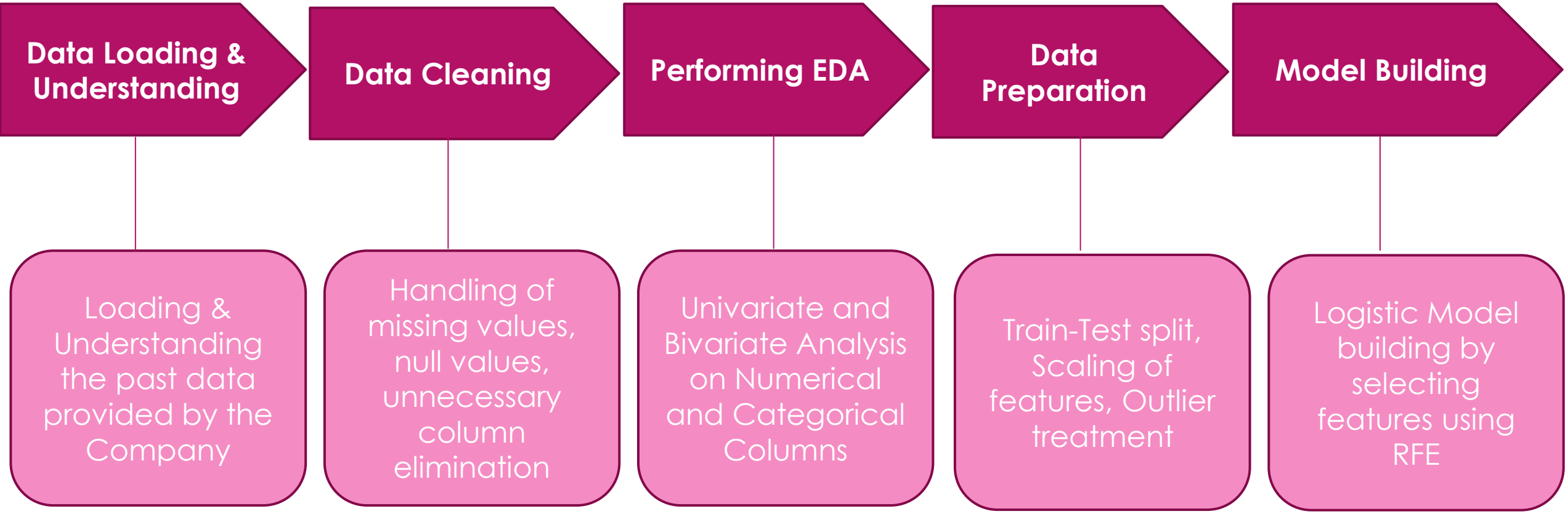
The focused communication with the HOT Leads make sure a better conversion rate of 80%

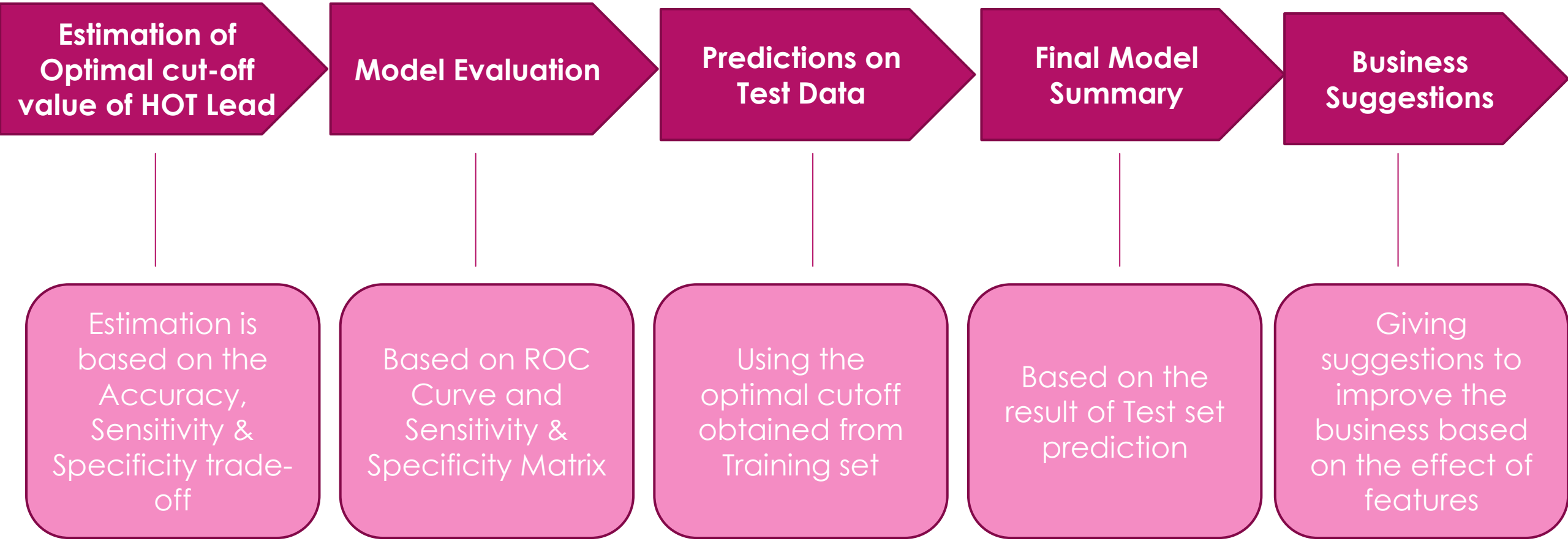
Solution

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”

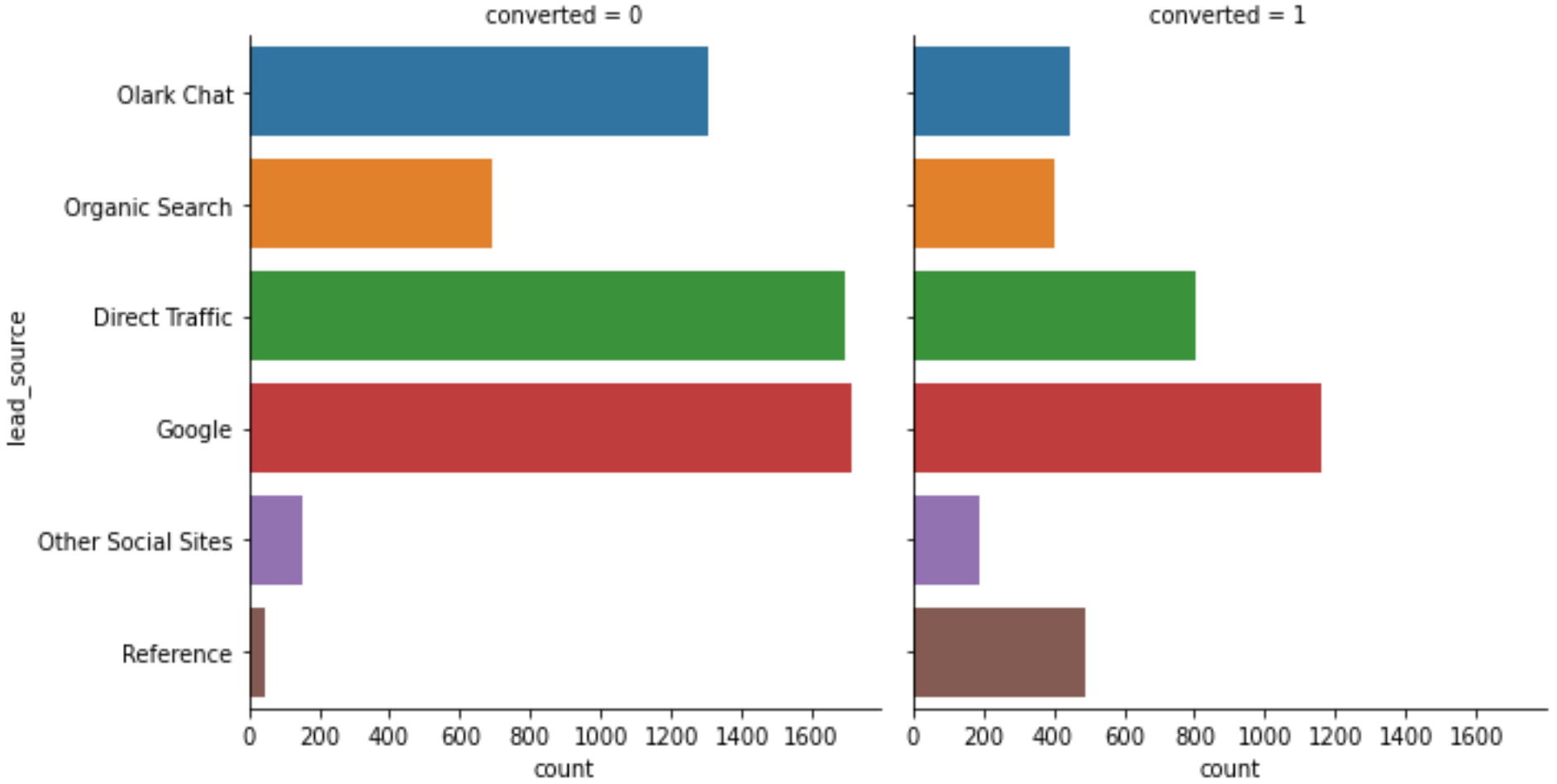
Flow of Implementation



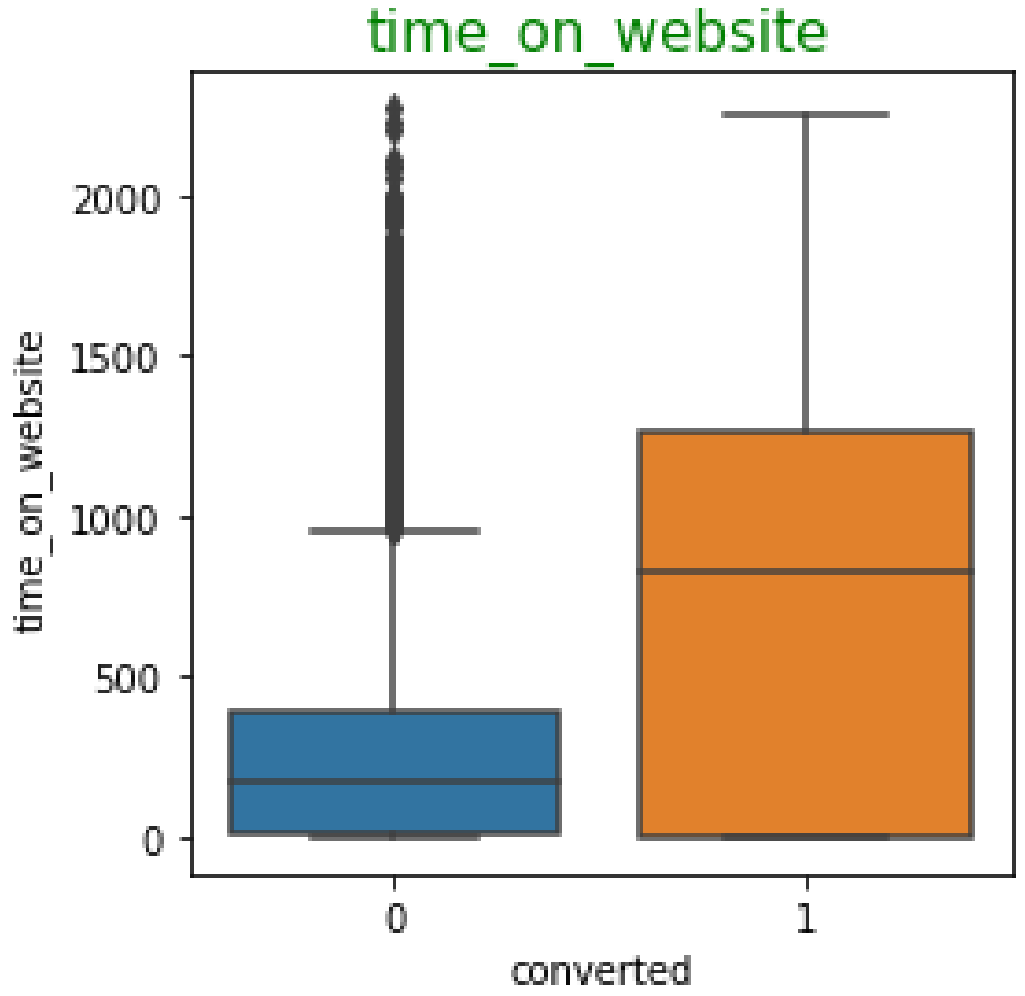

Estimation of Optimal cut-off value of HOT Lead
Model Evaluation
Predictions on Test Data
Final Model Summary
Business Suggestions
Estimation is based on the Accuracy, Sensitivity & Specificity trade-off
Based on ROC Curve and Sensitivity & Specificity Matrix
Using the optimal cutoff obtained from Training set
Based on the result of Test set prediction
Giving suggestions to improve the business based on the effect of features

Insights from EDA

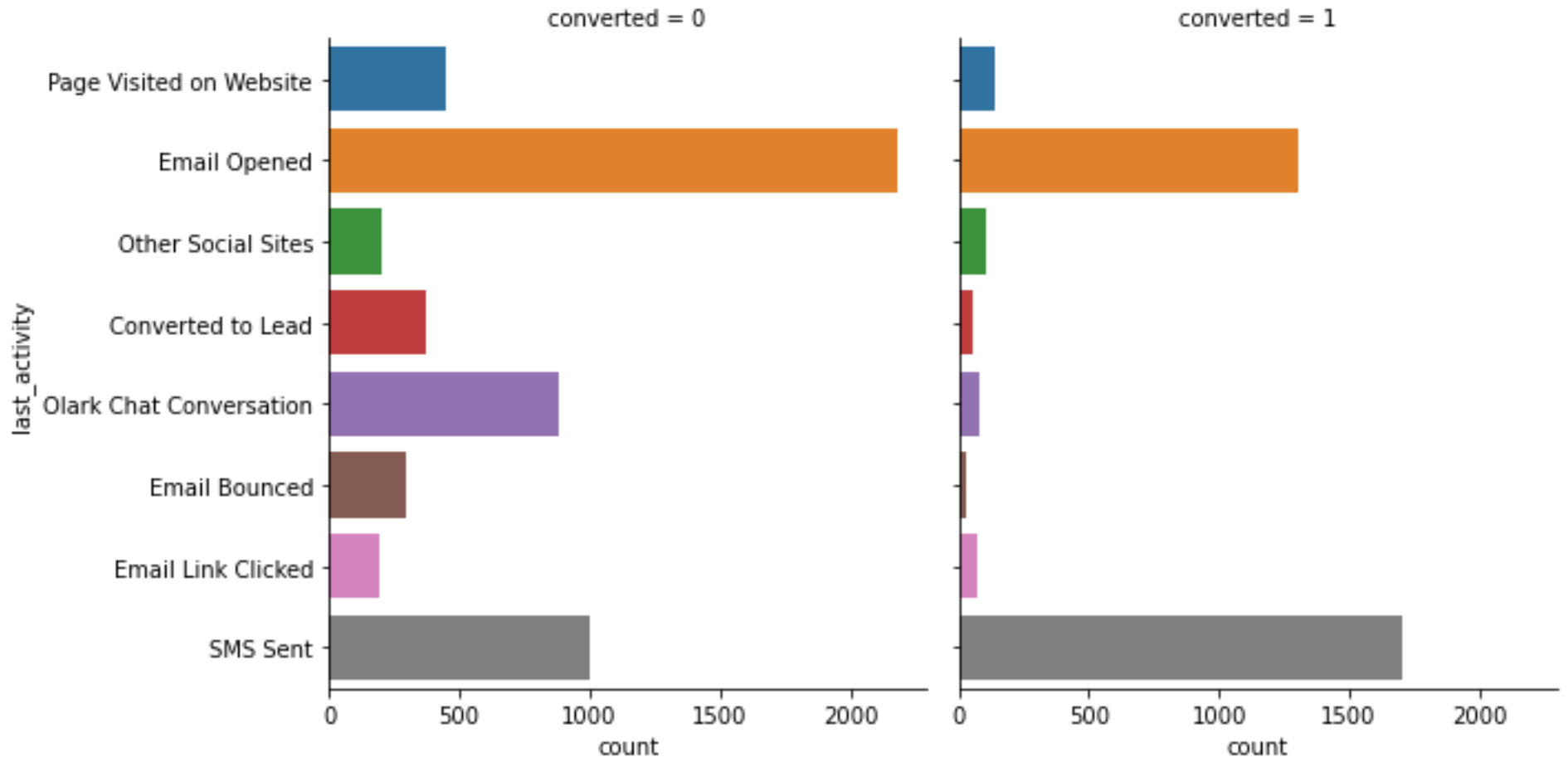
Lead source & Target variable



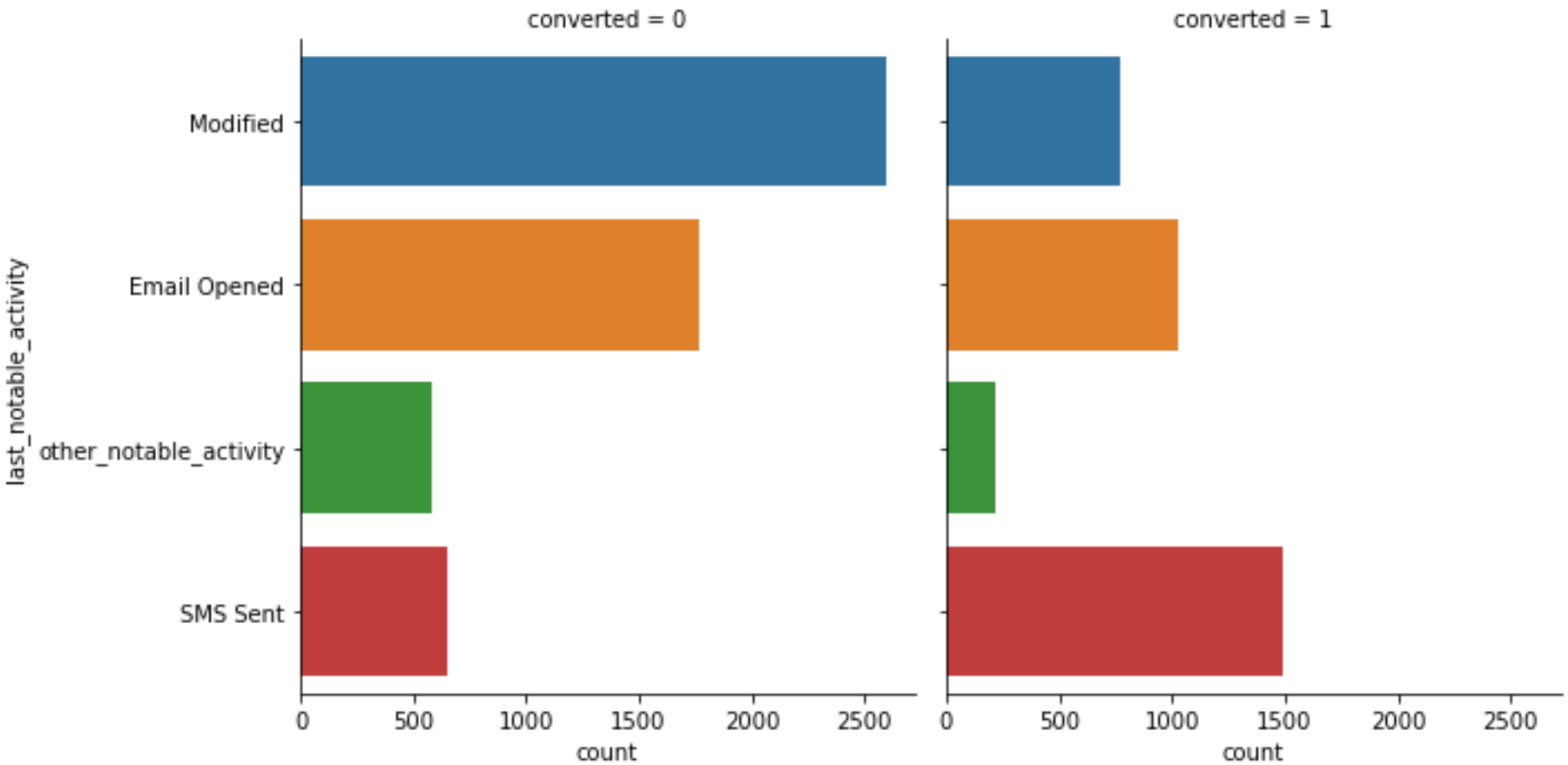
Total time spent on website & Target variable



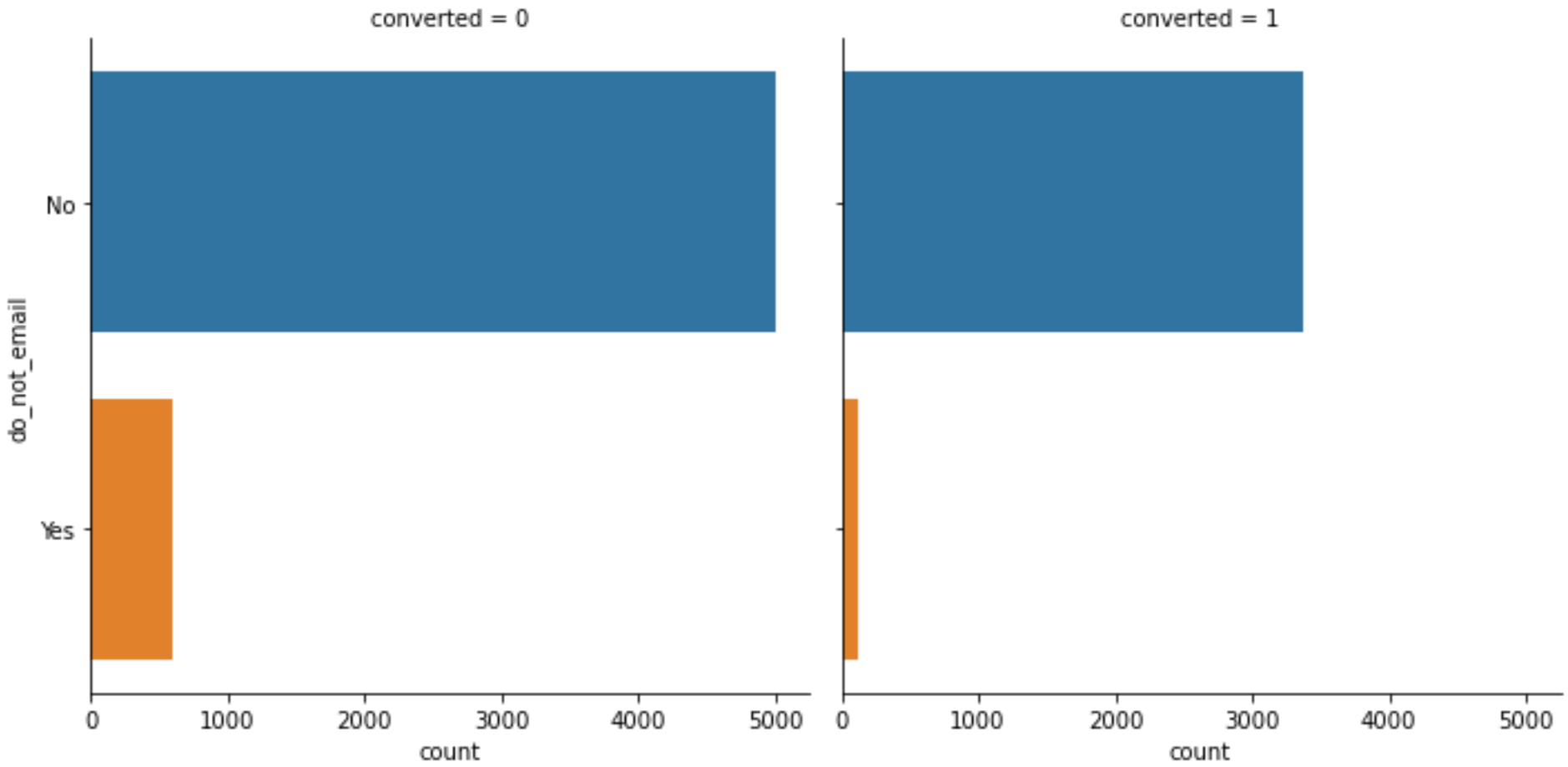
Last activity & Target variable



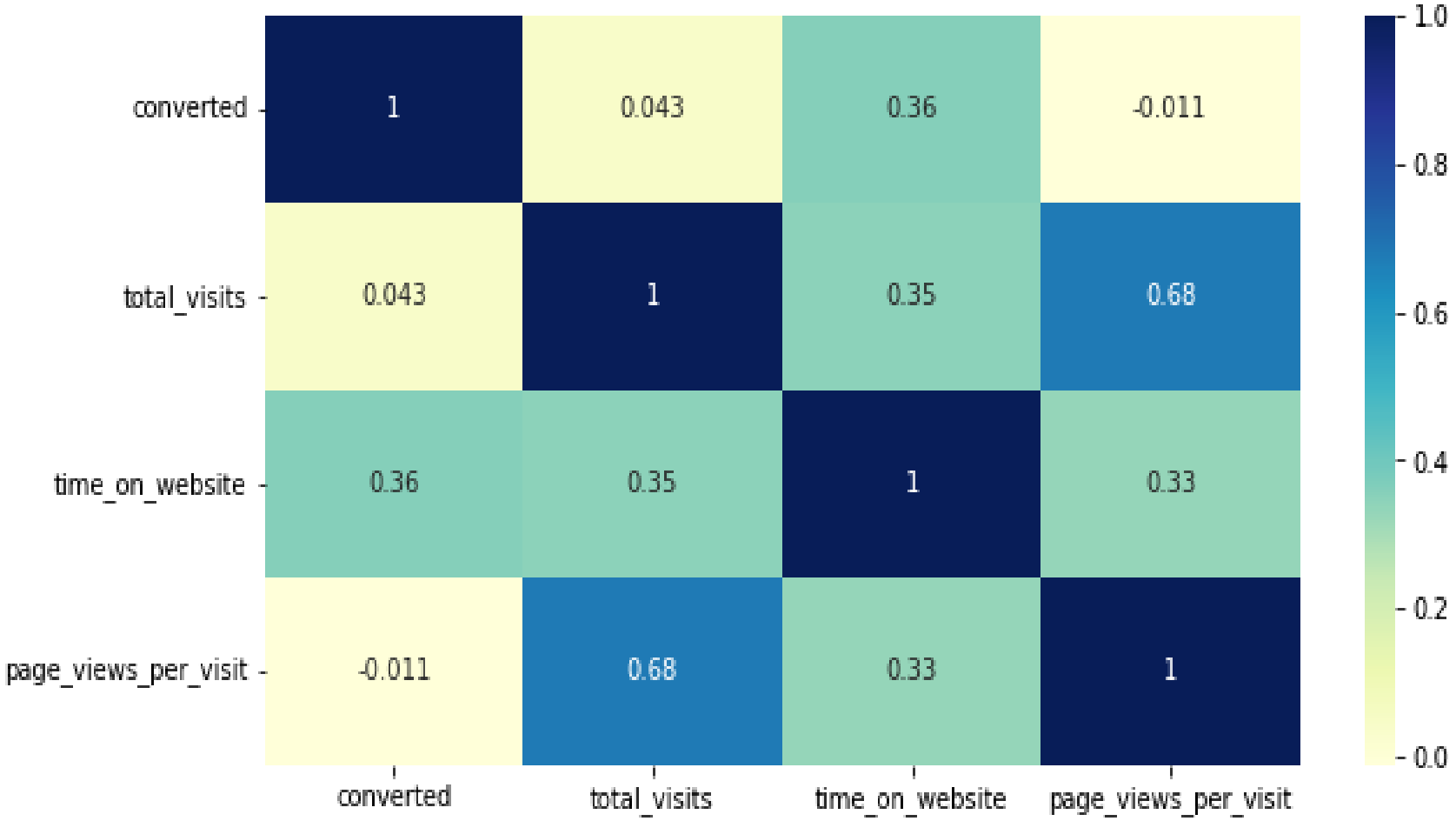
Last Notable Activity & Target Variable



Do not Emailed & Target Variable



Correlation of all Numeric columns
Heat Map



Model Building

Significant Features of the final model

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

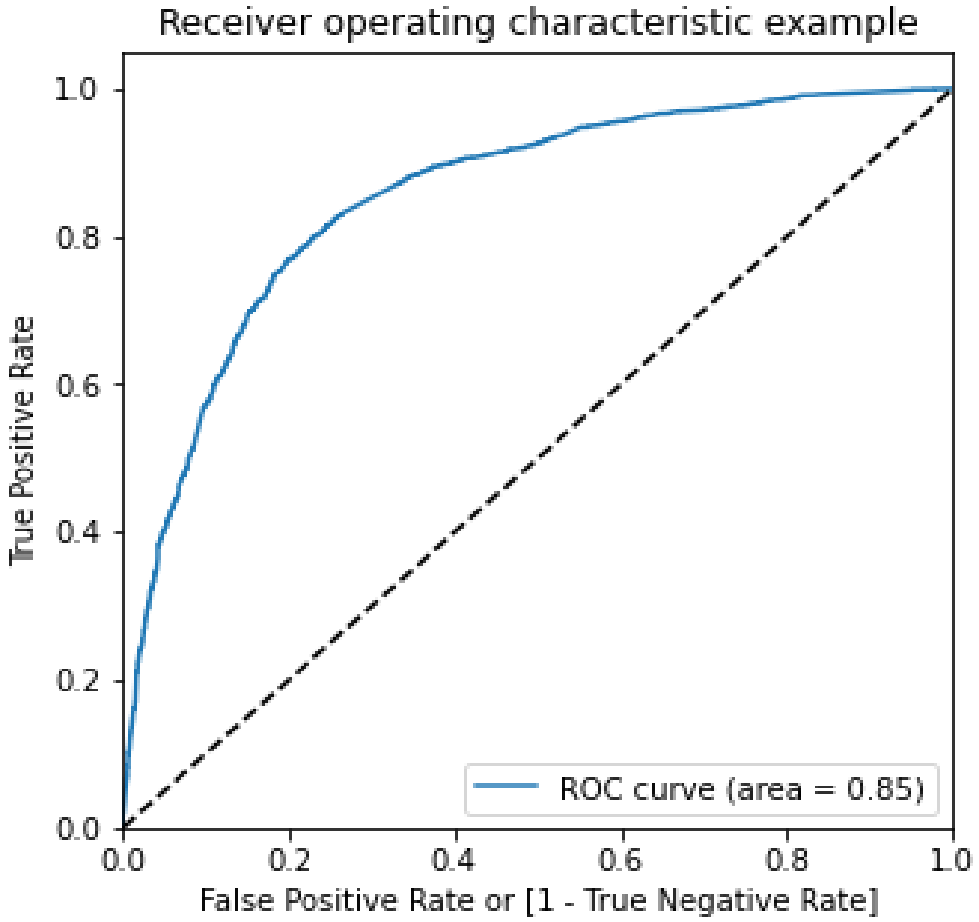
For all features the p – value is less than 0.05, which implies the features are significant

VIF of final model Features

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

For all features the VIF value is less than 2

ROC Curve of the Final model



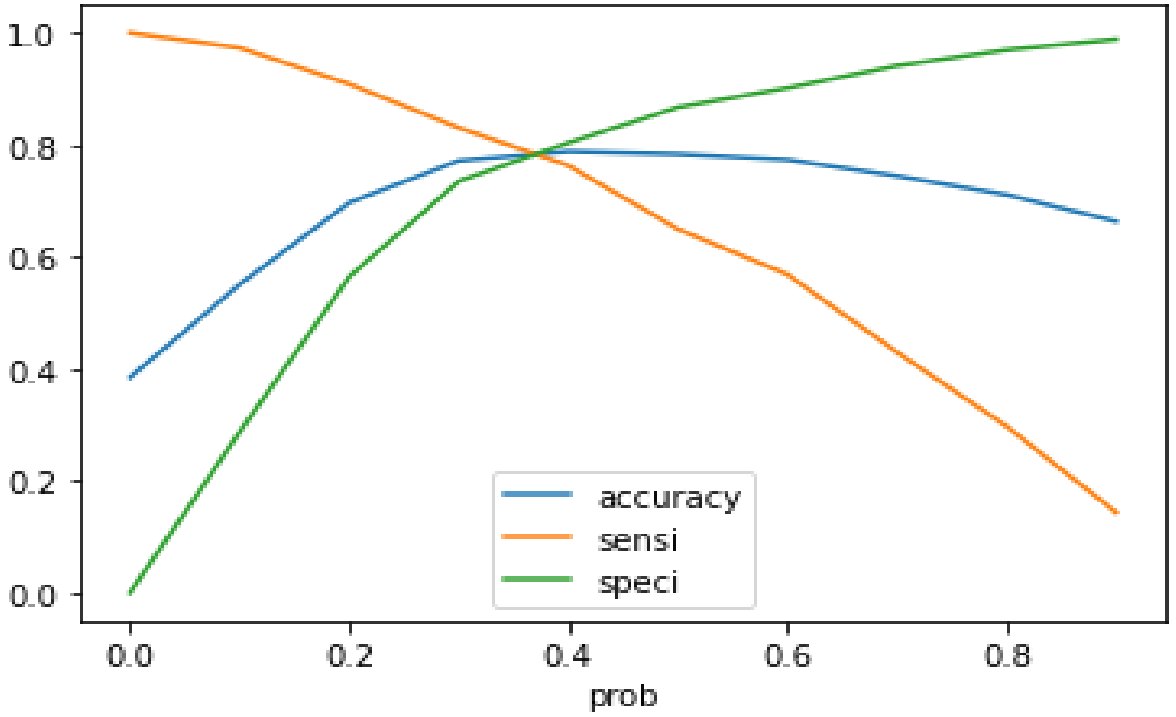
Gini (Area under ROC Curve) - 0.85

Logistic Regression Final Model Parameters on Train set at the Arbitrary cut

```
==================================================
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



Optimal cut-off of HOT Lead = 0.35

Logistic Regression Final Model Parameters on Train set at the Optimal cut

```
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: 0.77170582226762
=================================================
```

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

Affecting Positively

- time_on_website
- last_activity_SMS_Send
- lead_source_reference
- lead_source_google

Affecting Negatively

- last_notable_Activity_modified
- Olark_chat_conversation
- do_not_email

Conclusions

Recommendations

Hot Leads:

The most potential leads are the ones having Predicted Probability > 0.35.

- Use Email to Communicate with the Hot Leads
- Conversion rate increases with increase in the time spend on the website, therefore increase the user engagement in their website.
- Try to give SMS notifications, it improves the conversion rate
- Improve the digital marketing to reach out to more people