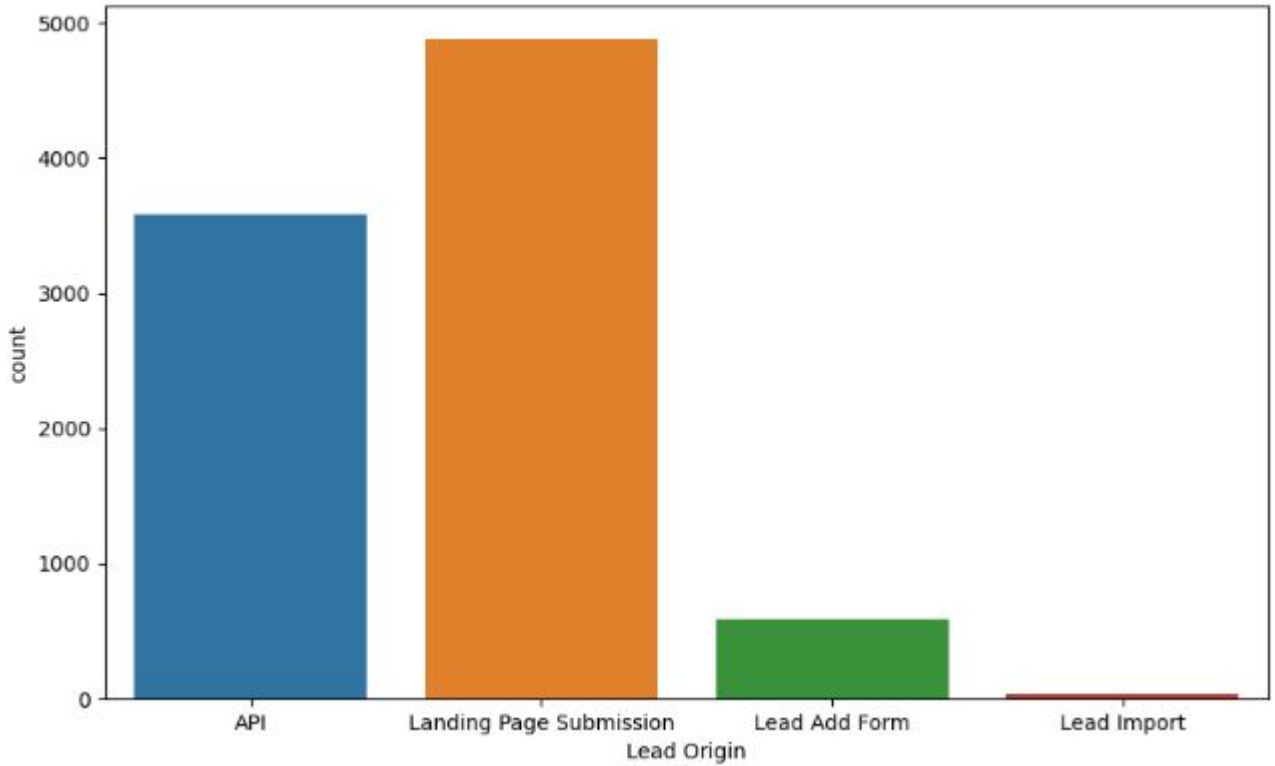
Lead Scoring Case Study

Done By:
Sivakami Chidambaram & Pallavi R

Data Visualization (Categorical Variables):



This is a graph of lead origin. Other Graphs are present in the Jupyter notebook.

Framing the Problem Statement

Problem Statement:

To increase conversion rate by 80% by only focusing on hot leads (Leads most likely to convert).

Objective:

To build a logistic regression model to predict lead score or conversion probability in order to find ‘Hot Leads’.

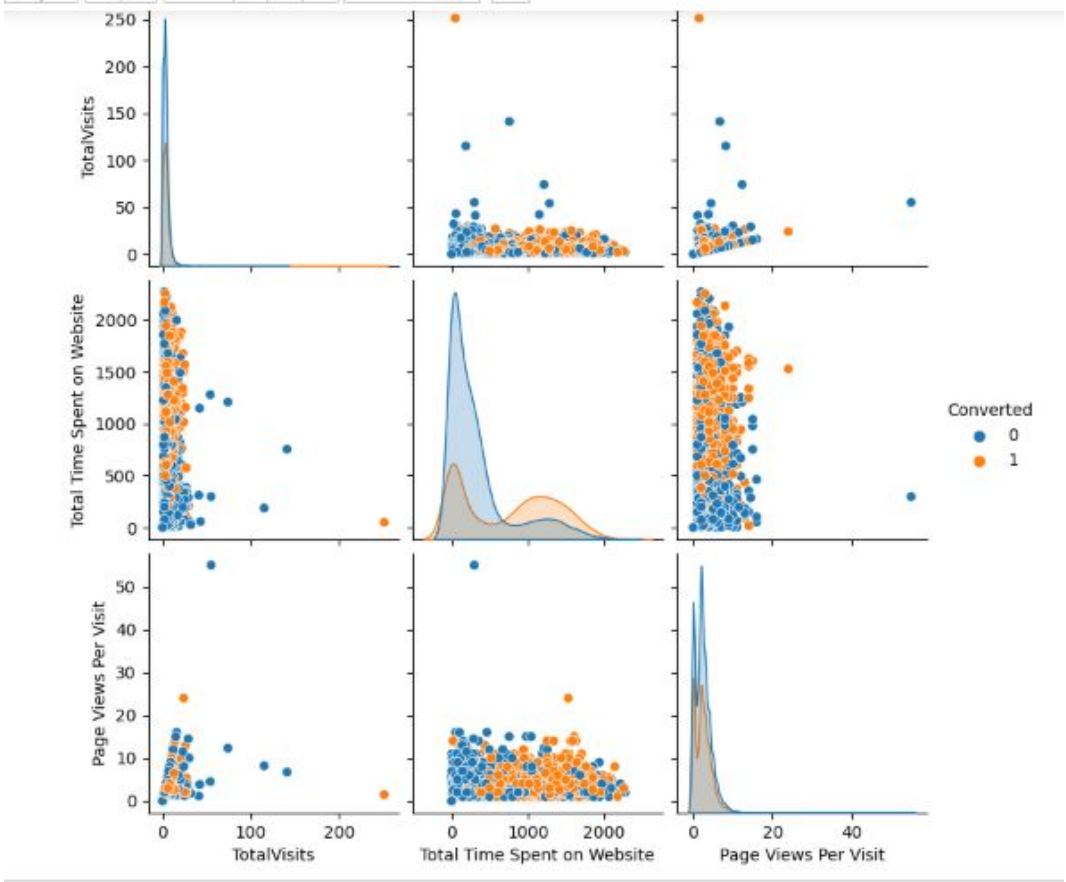
Procedure:

- Data Cleaning: also remove all ‘Select’ values because they are essentially null values
- Data Visualization: removing outliers and dropping columns with the same value in all cell
- Performing Variance Thresholding: Removes data with hardly any change in features
- Model Building : Logistic regression because target variable is a categorical variable
- Model Evaluation: Based on calculating the confusion matrix, accuracy, sensitivity, specificity on both test and train data.

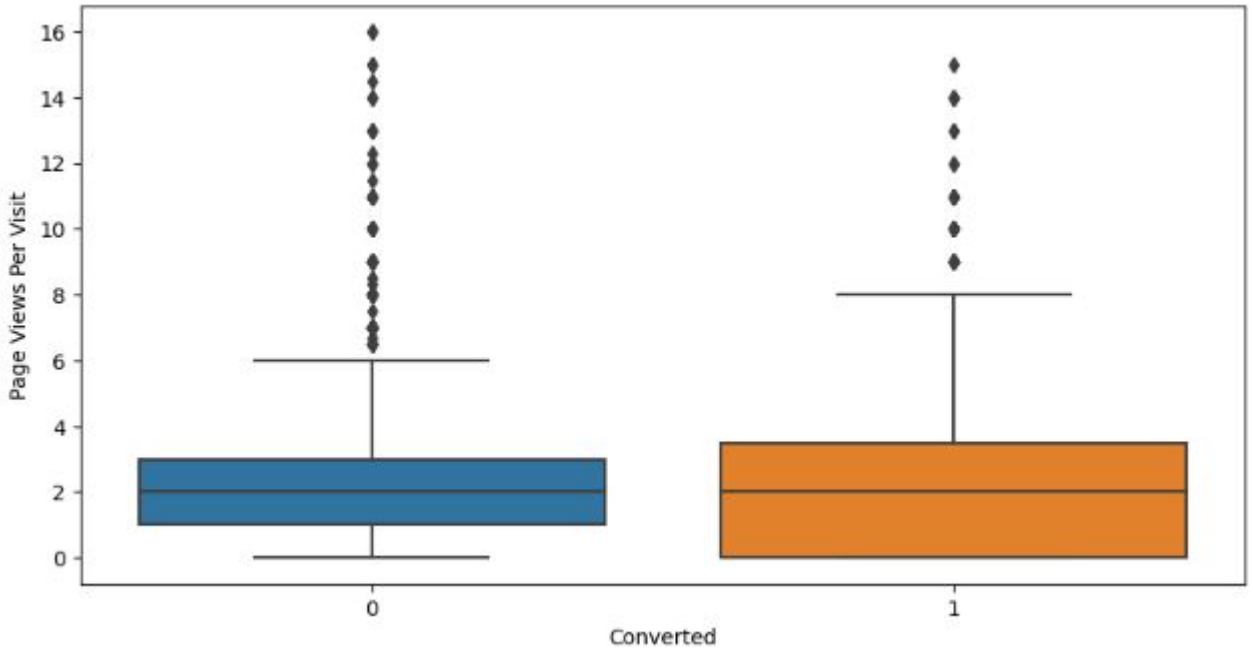
Data Understanding and Cleaning:

- The data has 9240 rows.
- The data has 37 columns out of which 7 are numeric and the rest are object type.
- Some of the columns have missing data point.
- 'Select' which is actually a null value because while filling online forms the default value is 'select' and if the customer does not fill a question then the value remains as 'select'. Thus ‘Select Value has to be converted to null value.
- 11 columns were dropped because they had a large percentage of missing values.
- After data cleaning the data frame had 9074

Data Visualization (Numerical Variables:Pair-Plot):



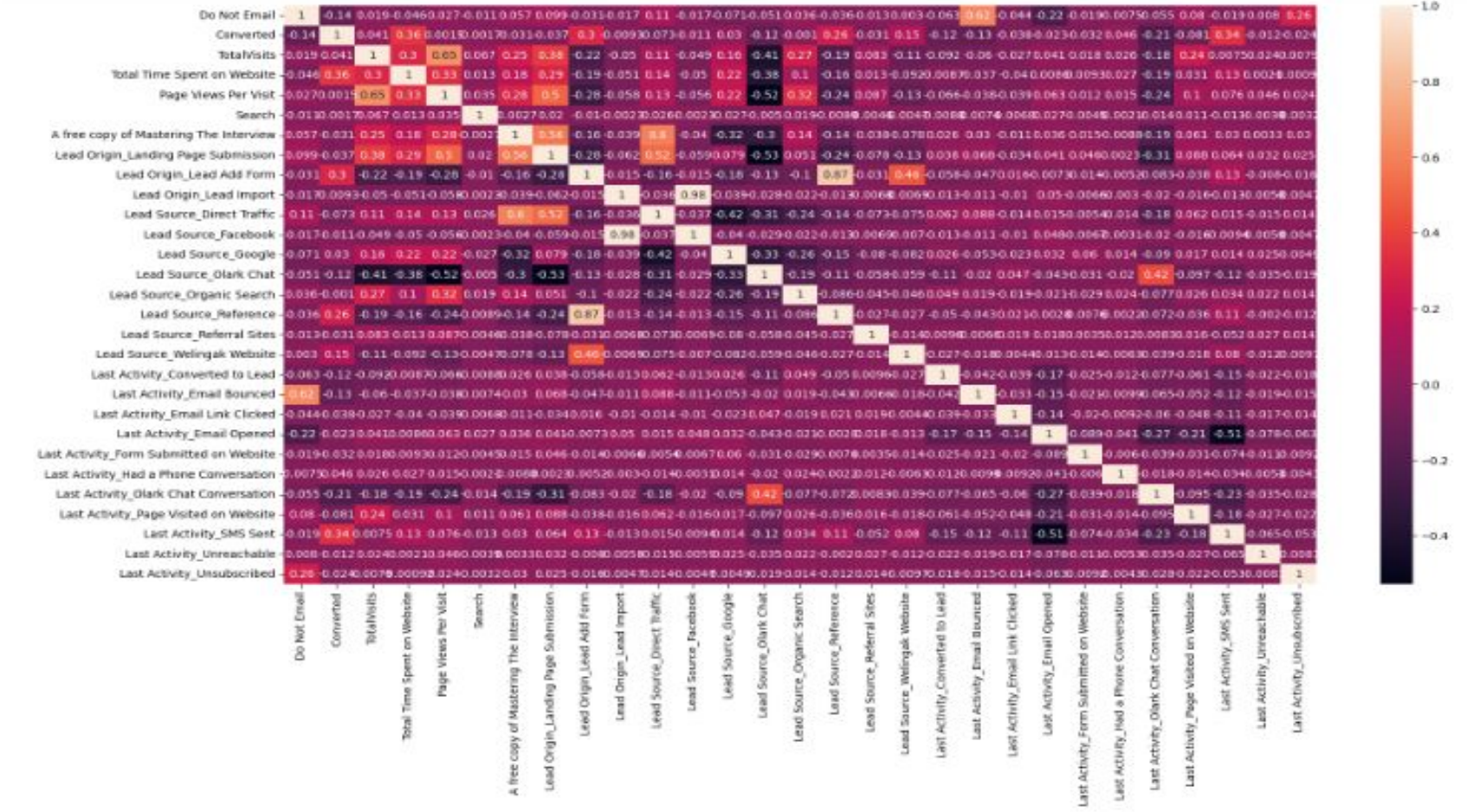
Data Visualization (Numerical Variables- Box plot):



Box plot between Page Views per visit and Converted after removing the outliers.

Graphs of other numeric variables are present in the Jupyter notebook.

Heat Map:


Do Not Email
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
Search
A free copy of Mastering The Interview
Lead Origin_Landing Page Submission
Lead Origin_Lead Add Form
Lead Origin_Lead Import
Lead Source_Direct Traffic
Lead Source_Facebook
Lead Source_Google
Lead Source_Olark Chat
Lead Source_Organic Search
Lead Source_Reference
Lead Source_Referral Sites
Lead Source_Welingak Website
Last Activity_Converted to Lead
Last Activity_Email Bounced
Last Activity_Email Link Clicked
Last Activity_Email Opened
Last Activity_Form Submitted on Website
Last Activity_Had a Phone Conversation
Last Activity_Olark Chat Conversation
Last Activity_Page Visited on Website
Last Activity_SMS Sent
Last Activity_Unreachable
Last Activity_Unsubscribed
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Model Building (RFE analysis):

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.4026	0.088	-27.343	0.000	-2.575	-2.230
Do Not Email	-1.7238	0.173	-9.957	0.000	-2.063	-1.384
TotalVisits	2.4993	0.578	4.323	0.000	1.366	3.632
Total Time Spent on Website	4.8234	0.163	29.532	0.000	4.503	5.144
Lead Origin_Lead Import	1.6495	0.441	3.743	0.000	0.786	2.513
Lead Source_Olark Chat	1.3672	0.109	12.590	0.000	1.154	1.580
Lead Source_Reference	4.4492	0.235	18.933	0.000	3.989	4.910
Lead Source_Welingak Website	5.8540	0.729	8.031	0.000	4.425	7.283
Last Activity_Converted to Lead	-0.9438	0.202	-4.676	0.000	-1.339	-0.548
Last Activity_Had a Phone Conversation	1.8440	0.594	3.106	0.002	0.680	3.008
Last Activity_Olark Chat Conversation	-1.5816	0.158	-9.995	0.000	-1.892	-1.271
Last Activity_Page Visited on Website	-0.6165	0.143	-4.312	0.000	-0.897	-0.336
Last Activity_SMS Sent	1.2346	0.074	16.788	0.000	1.091	1.379
Last Activity_Unsubscribed	1.3070	0.448	2.916	0.004	0.429	2.185

This is the Final logistic regression model.It has 13 variables, Lead Source_Welingak Website has the largest coefficient hence it is the most influential factor, followed by Total Time Spent on Website and Lead Source_Reference.

Model Prediction:

```
y_train_pred_final['Lead Score']= y_train_pred_final.Converted_Prob.map(lambda x: round(x*100,2))
y_train_pred_final['predicted'] = y_train_pred_final.Converted_Prob.map(lambda x: 1 if x > 0.8 else 0)

# Let's see the head
y_train_pred_final.head()
```

	Converted	Converted_Prob	Lead Score	predicted
5061	0	0.111273	11.13	0
5055	0	0.059566	5.96	0
8167	0	0.131229	13.12	0
1197	0	0.891166	89.12	1
1374	0	0.604525	60.45	0

Since the problem statement states that the conversion rates should be increased to 80% we are taking deciding probability as 0.8, only of the converted probability is greater than 0.8 will the model predict the data as a lead.

Train Data Results and Explanation: Accuracy, sensitivity/Recall, Specificity:

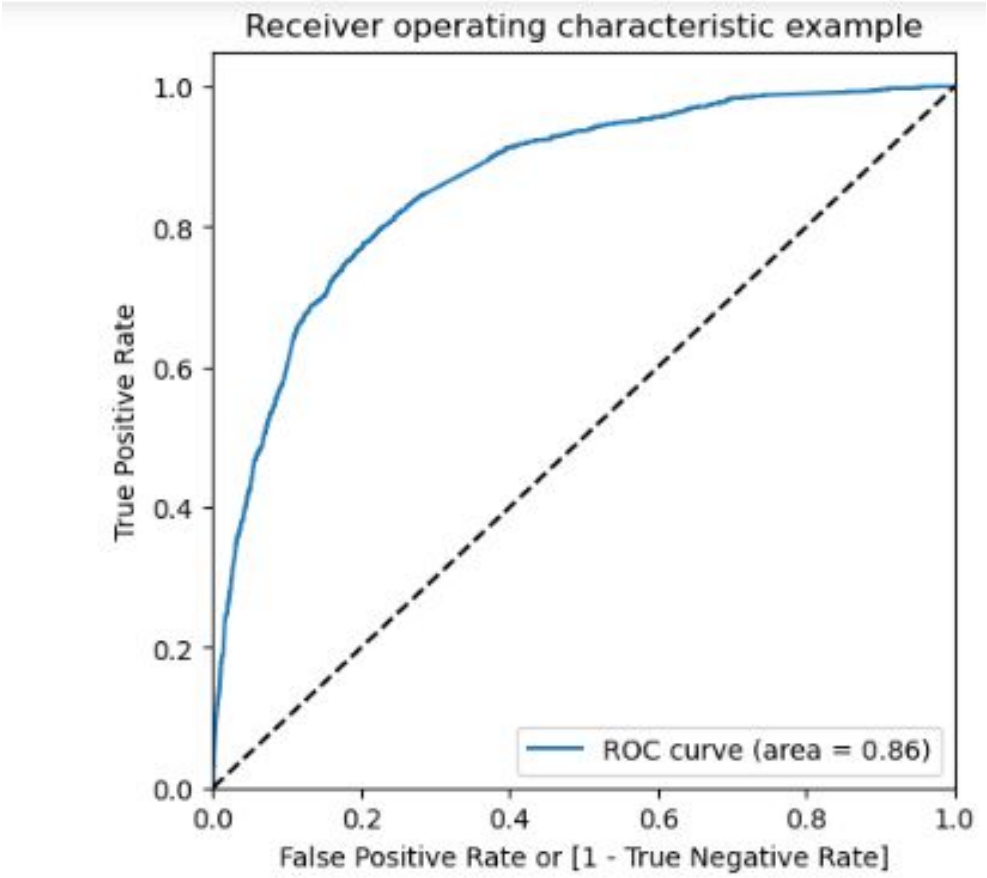
- Accuracy: 73.31
- sensitivity/Recall:34.76
- Specificity:96.88

```
# Confusion matrix
confusion = metrics
print(confusion)
```

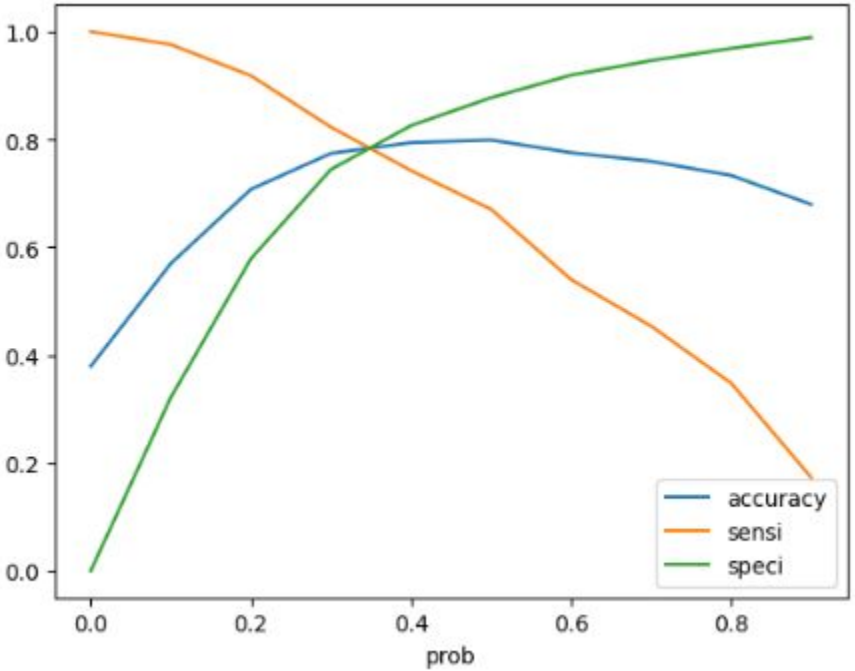
```
[[3817  123]
 [1571  837]]
```

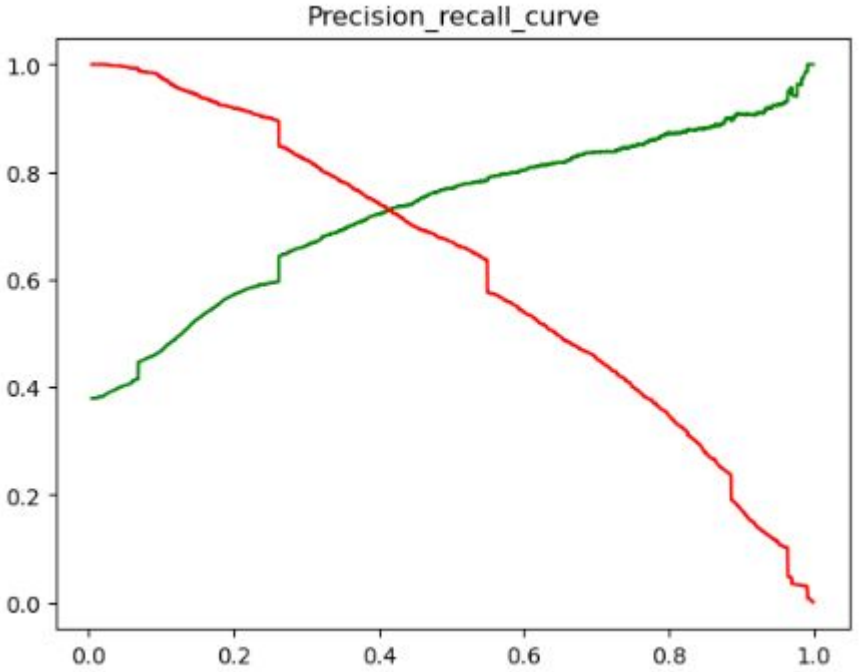
For our model sensitivity is low and specificity is high because we are only looking for hot leads with very high probability of conversion.

ROC Curve For Train Data:



Additional Graphs:





Test Data results:

- Accuracy: 72.51
- sensitivity/Recall:33.85
- Specificity:95.88

```
confusion2 = metrics.confusion_matr
confusion2
```

```
array([[1626,   70],
       [ 678,  347]], dtype=int64)
```

As we can see Accuracy, sensitivity and specificity of the test data is very similar to the train data. To increase sensitivity (since specificity is around 95% and we only need 80%, we could reduce the probability cut off from 80%.

Analysing Possible Solutions:

- The biggest influencing factor in the list of 13 factors we have considered in our model is ‘Lead Source_Welingak Website’. Which means that people who reach the company through this website have a higher chance of converting, so the company could invest in this avenue by increasing adds to this website.
- The next most important contributing factor is ‘Total Time Spent on Website’ which is understandable more time a customer spends on the website , more the customer is interested in the program. So the sales team could target customers who spend more time on the website.
- The third most import factor is ‘Lead Source_Reference’. If a person is referred to the program by someone they know and trust they seem to join the program

Analysing Possible Solutions:

- Factors like ‘Do Not Email’,’ Last Activity_Olark Chat Conversation’, have negative coefficients thus these factors contribute to the customer not getting converted to lead. By doing root cause analysis we could find possible cause and eliminate the same.
- Other factors like ’Page Views Per Visit’ which was not even a part of the model could also logically be used to identify a lead.
- Some of the dropped features (Because of large amount of missing values)like Lead profile, Lead quality, What matters most to you in choosing a course are logically good factors in identifying a lead, finding the missing values and building the model again with those factors could lead to a better model.

THANK YOU