\

LEADING SCORE CASE STUDY REPORT

Linear Regression Model

STUDY GROUP

1) Veda Pranathi Peddisetti
2) Raman Sharma
3) Pranshu Sharma

PRESENTATION AGENDA

1) Problem Statement

2) Objectives

3) Approach

4) Data Insights

5) Factors Responsible in Driving Leads

6) Terminologies Required

7) Model Metrics

8) Conclusion

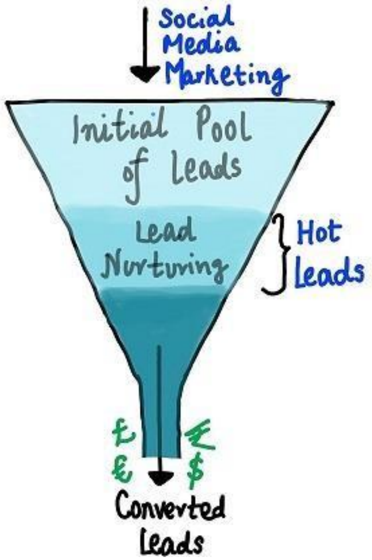
Problem Statement

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’. 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.

Fig. Lead Conversion Process



Objectives

- To help the company **in selecting the** most potential leads, also known as **'Hot Leads'** whose lead **conversion rate is around 80%.**
- **To build a model wherein a lead score is assigned** to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- Help the sales team to divert their focus on potential leads & avoid them from making useless phone calls.

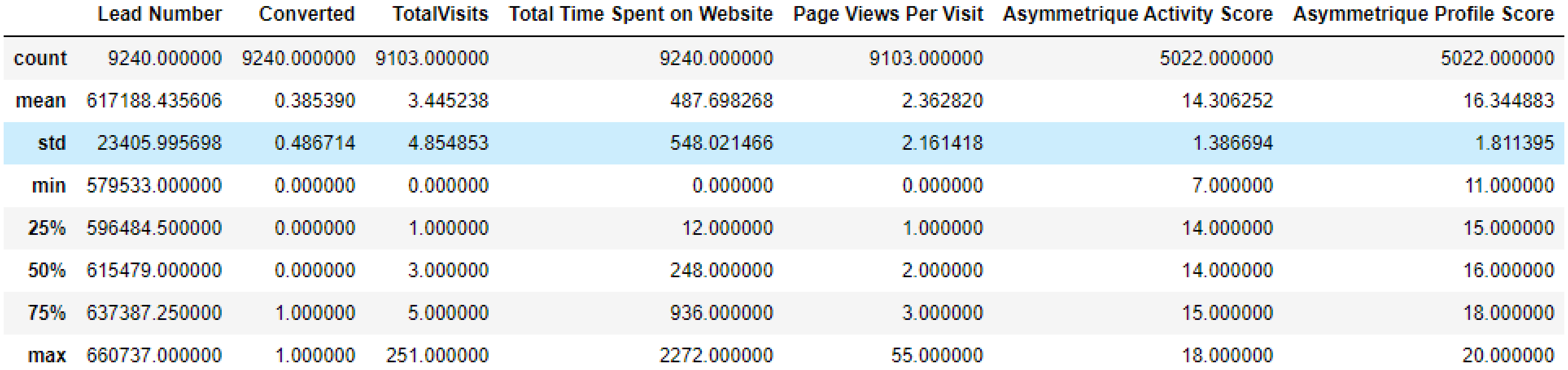
Approach

- Analysing Patterns:

  ➢ Using Exploratory Data Analysis, we have analyzed the patterns present in the Dataset which will provide us intuition that the which features will help in driving the lead conversion.

- Driving Factors:

  ➢ Looking at the below data we get an intuition that how the variables are distributed.



	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000	5022.000000	5022.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820	14.306252	16.344883
std	23405.995698	0.486714	4.854853	548.021466	2.161418	1.386694	1.811395
min	579533.000000	0.000000	0.000000	0.000000	0.000000	7.000000	11.000000
25%	596484.500000	0.000000	1.000000	12.000000	1.000000	14.000000	15.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000	14.000000	16.000000
75%	637387.250000	1.000000	5.000000	936.000000	3.000000	15.000000	18.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000	18.000000	20.000000

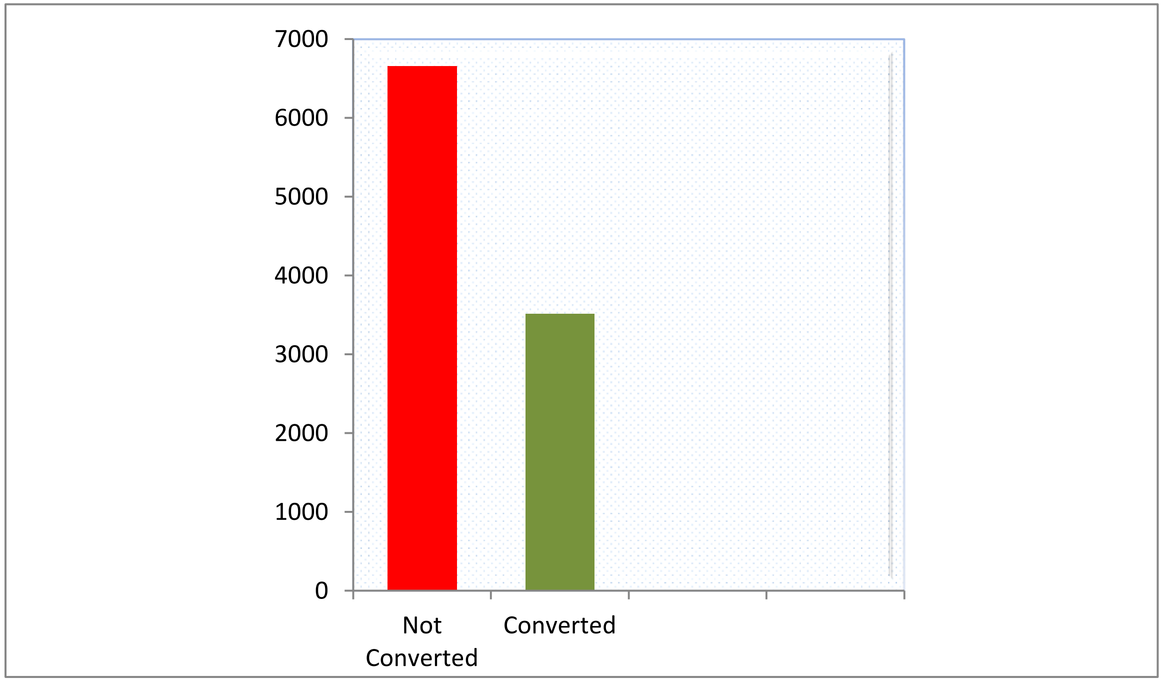
- Correlations:
  - Identifying correlations amongst variables to identify the variability in data and identify most important features that can help in driving the conversion of leads.
- Recommendations:
  - Focus on features that can expedite the conversion of leads.

Data Insights

1. We have total 9240 entries of unique customers and we needs to identify out of these which have the highest probability of getting converted.

   Decision Criteria:

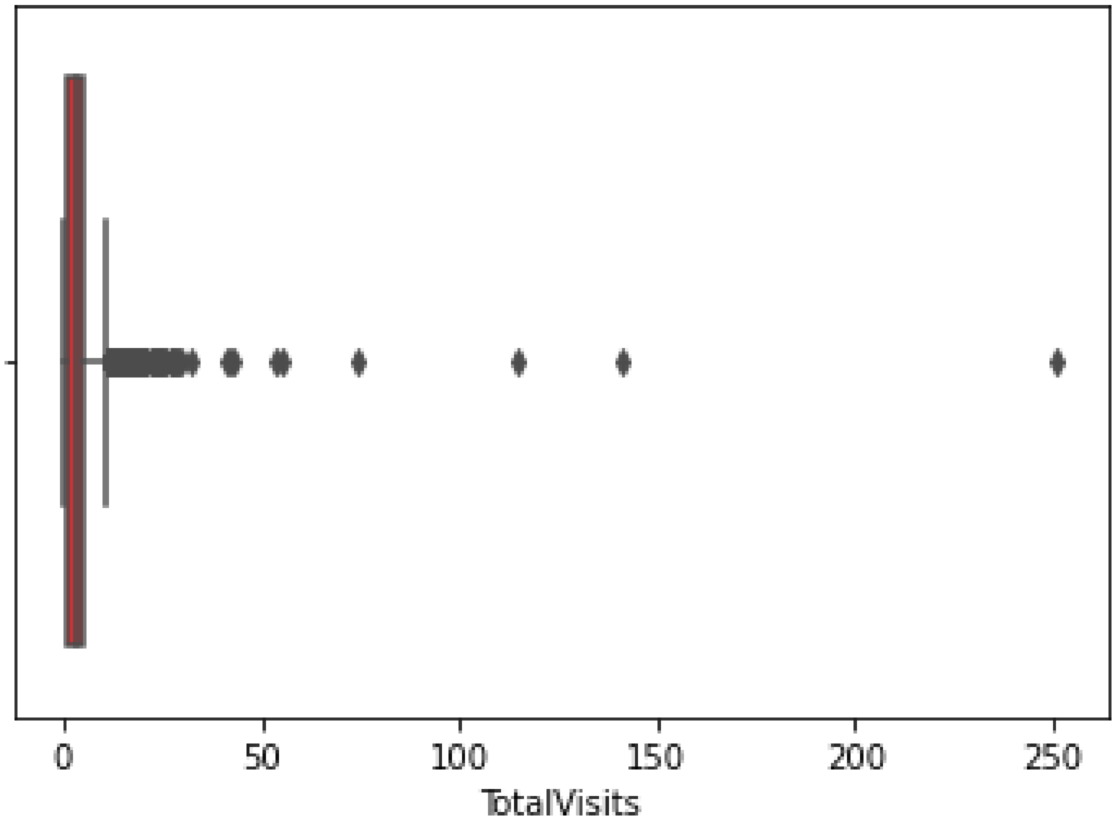
   - Potential Leads can be bifurcated on the basis of Leads Score (which is probability of getting converted).
   - Out of 9240 entries we see that around 38% of leads are converted and 72% of leads are not converted.

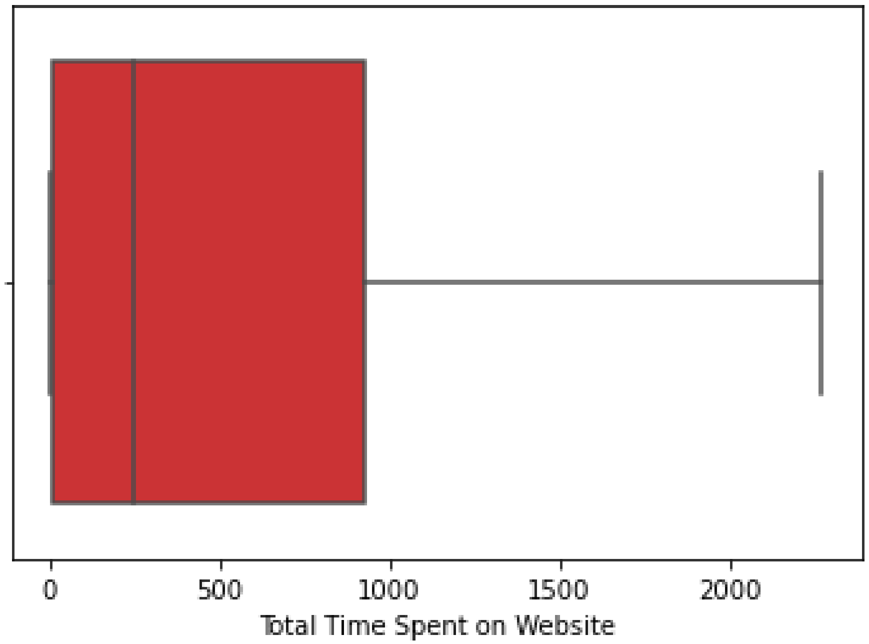


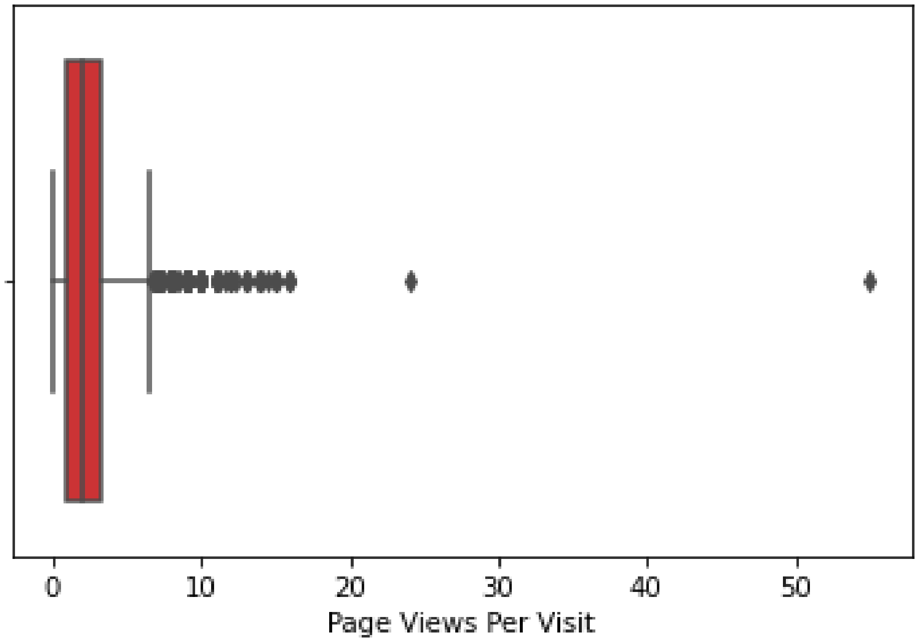
Task:

Identify solution so that the lead conversion rate could be increased.

Let us see the **spread of numerical columns**.



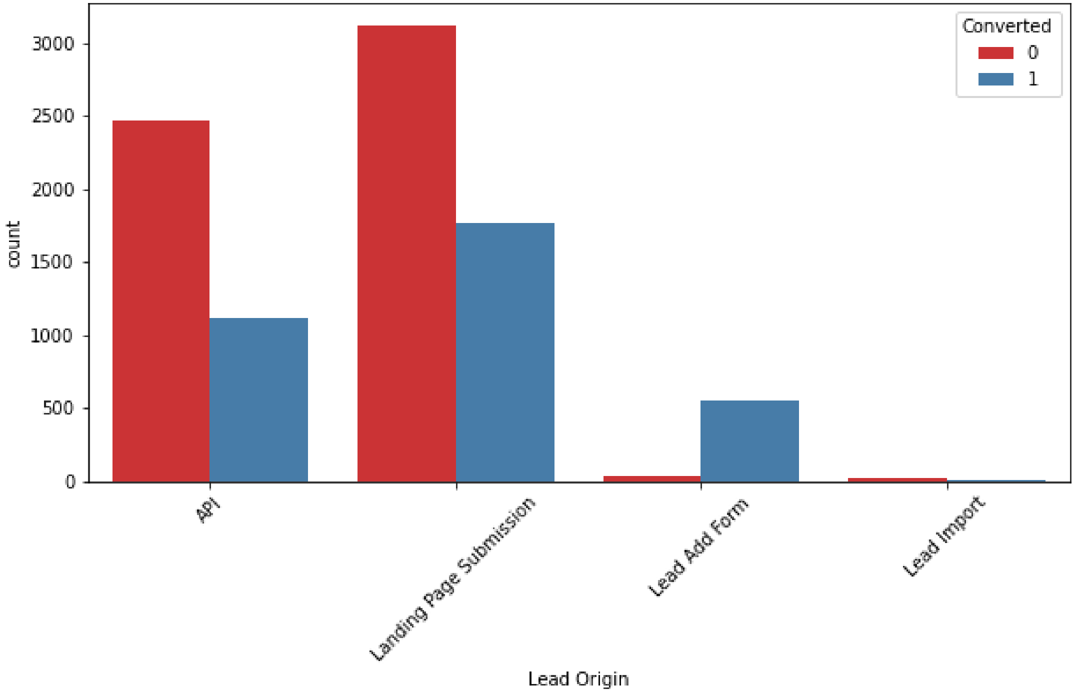




Observations:

We observe that our data is skewed.

Let us see the **spread of Categorical Columns w.r.t Converted Columns**.



Lead Origin

1.) Customers who were identified as Leads from Landing Page submission, constitute most of the leads.

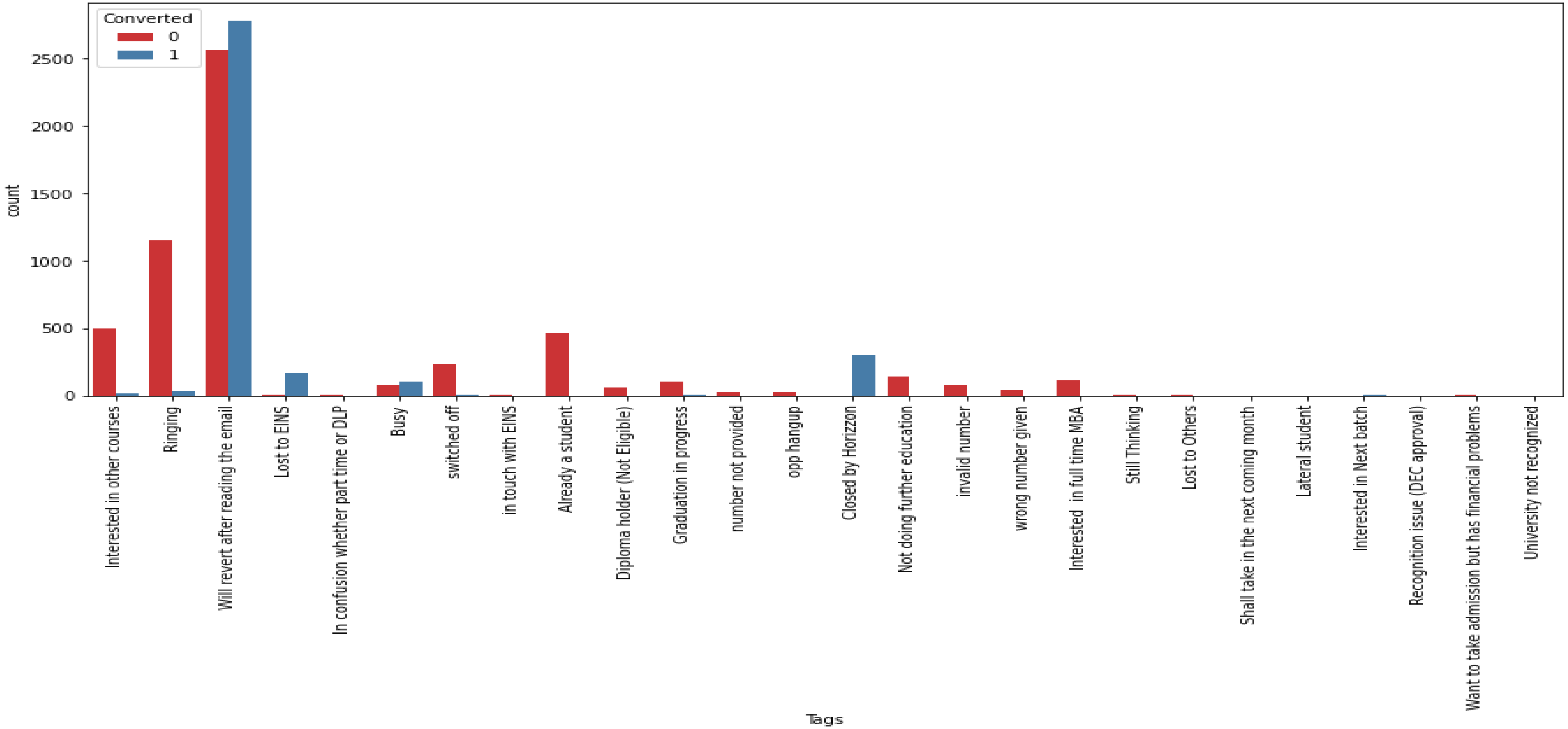
2.) Customers originating from Lead Add Form have high probability of conversion. These Customers are very few.

3.) Lead origin-API & Lead Import have the least conversion rate. Customers from Lead Import are very few.

To improve overall lead conversion rate, we need to focus more on improving lead conversion rate of Customers originating from API and Landing Page Submission and generate more leads from Lead Add Form.

Tags

More focus shall be given on the leads as will revert after reading the email & others as these are potential leads and have higher rate of conversion.

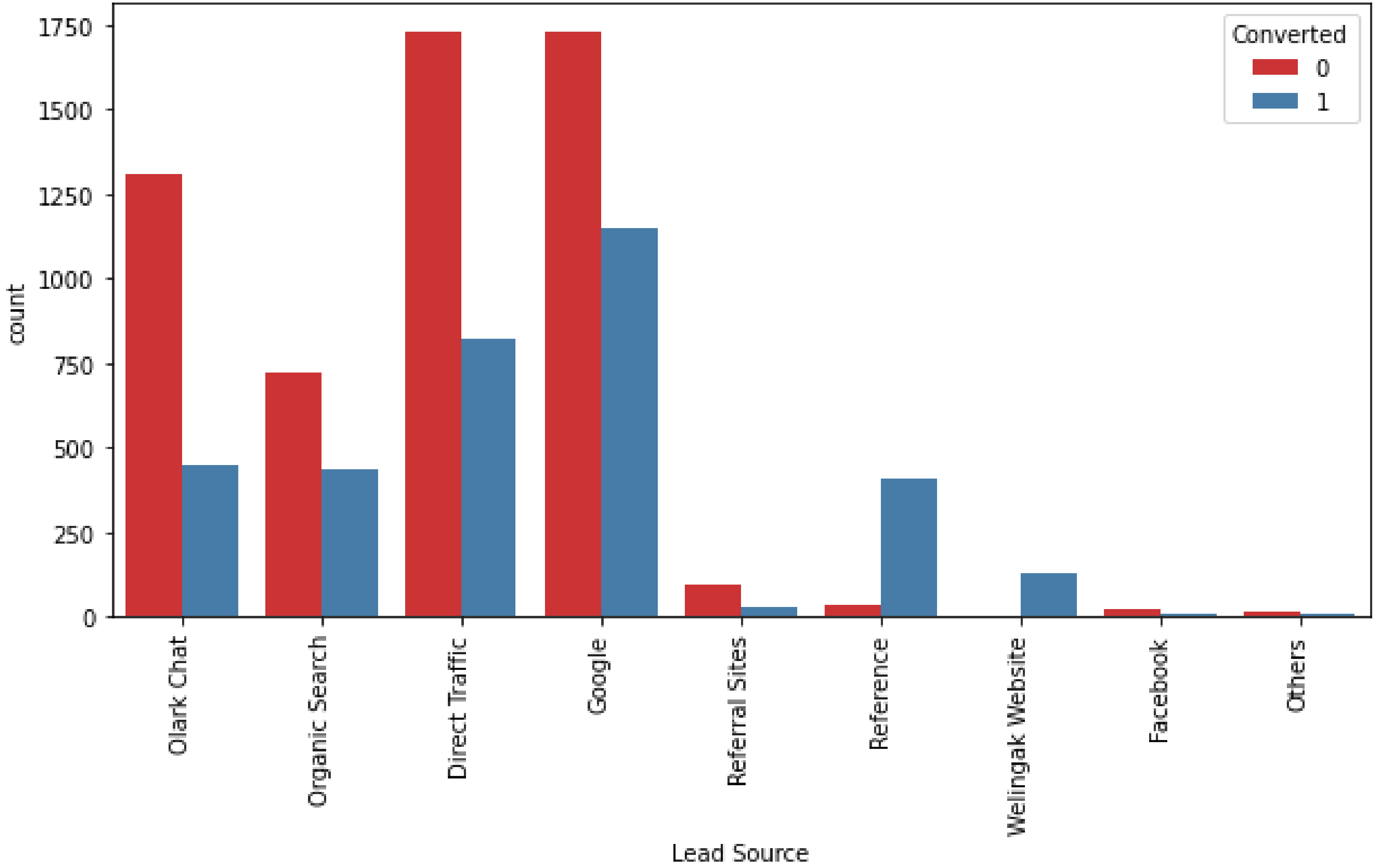


Lead Source

1.) Majority source of the lead is Google & Direct Traffic.

2.) Lead source from Google has highest probability of conversion.

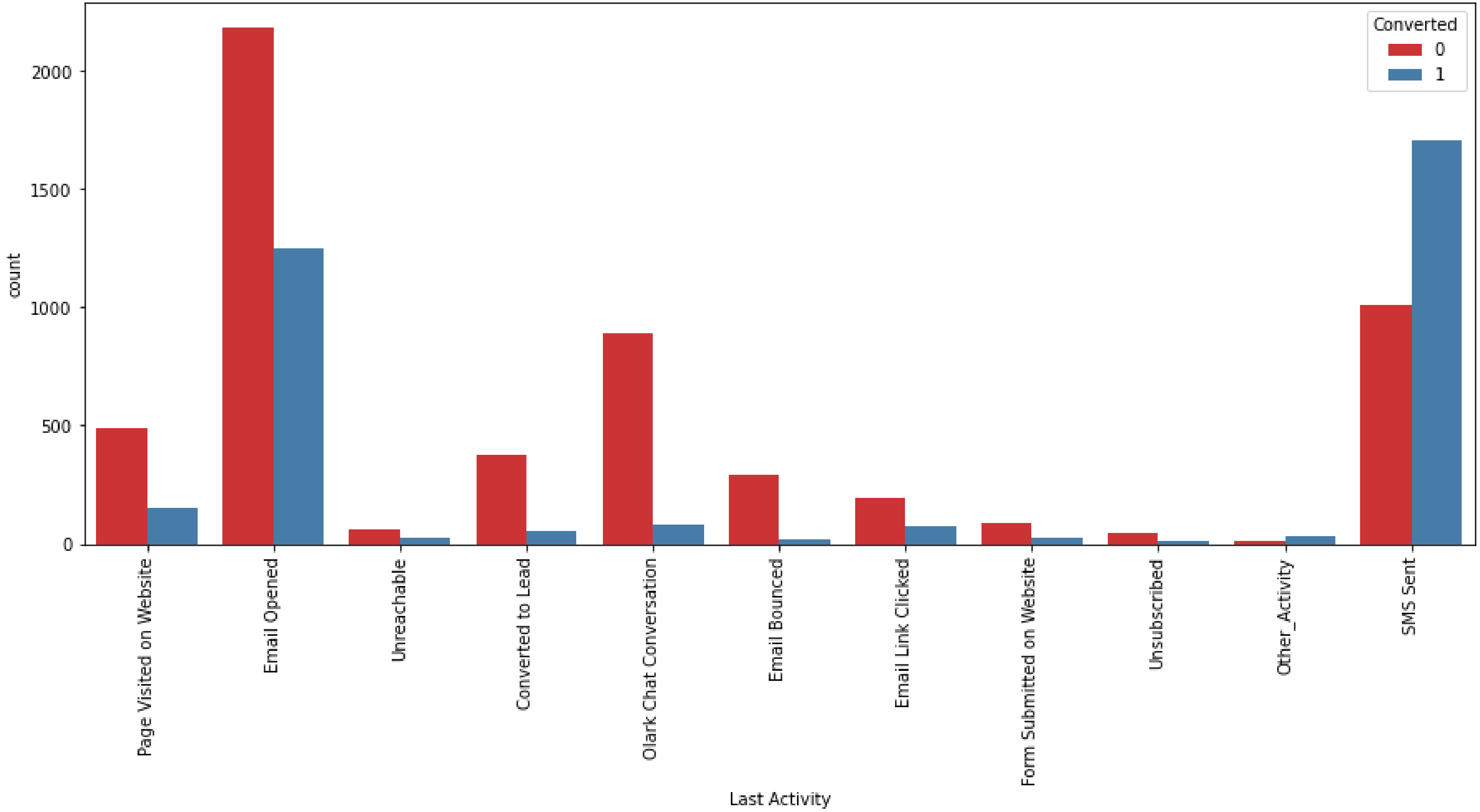
3.) Leads with source Reference has maximum probability of conversion.



Last Activity

1.) Customers whose last activity was SMS Sent have higher conversion rate which is around 63%.

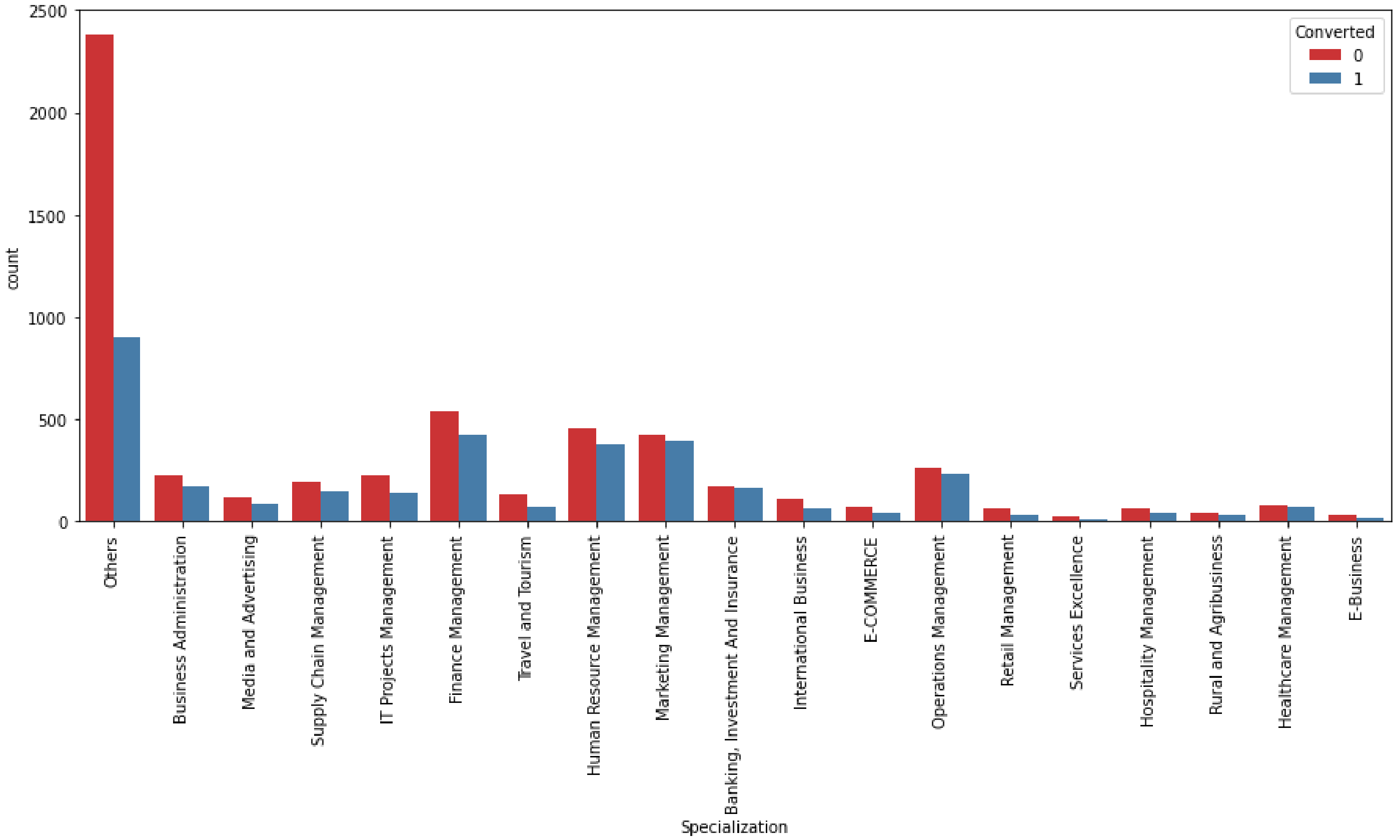
2.) Customers who last activity was Email Opened constitute majority of the customers. They have around 36% of conversion rate.



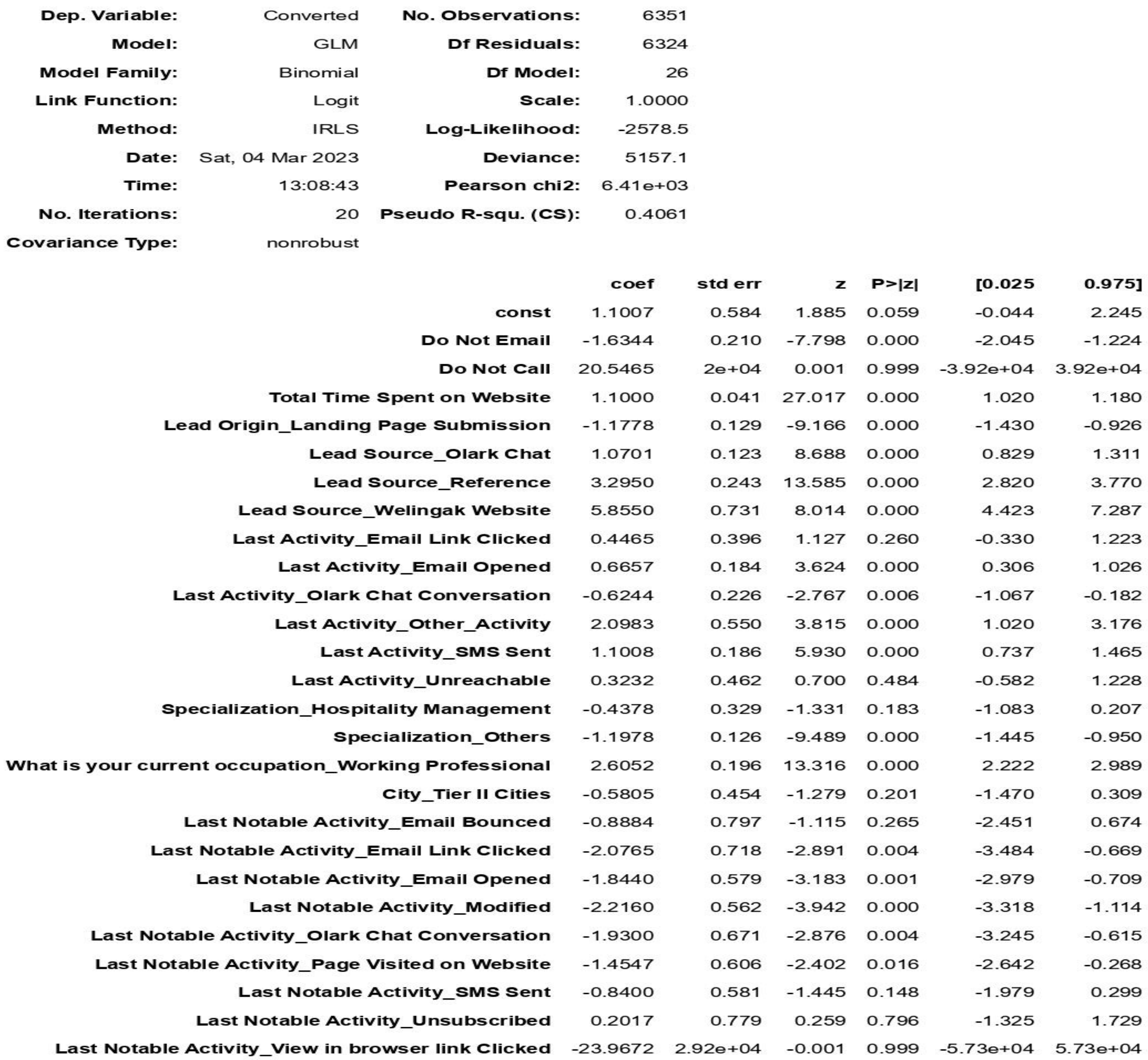
To improve overall lead conversion rate, we need to focus more on improving lead conversion rate of Customers whose last activity was Email Opened and generate more leads from the ones whose last activity was SMS Sent.

Specialization

1) Maximum amount of leads have Specialization as Management & Others.
2) Agri-Business leads have least probability of conversion.



Factors Responsible in Driving Leads - Generalized Linear Model Regression Results



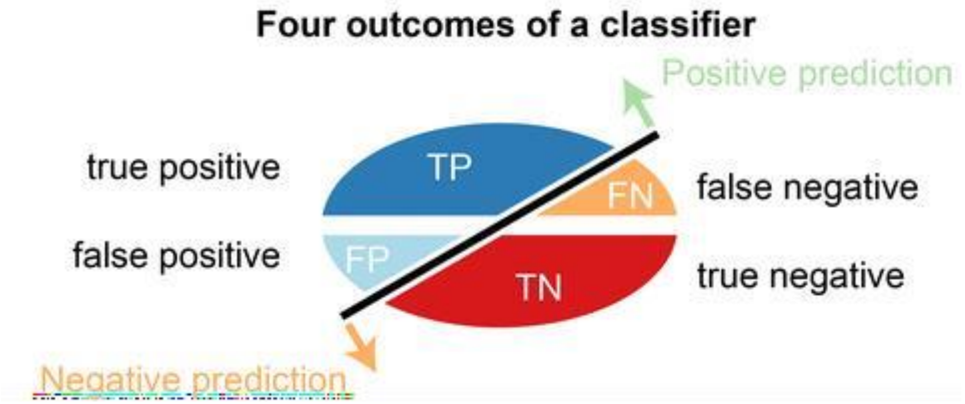
```
    Dep. Variable:        Converted    No. Observations:      6351
           Model:               GLM        Df Residuals:      6324
    Model Family:          Binomial            Df Model:        26
   Link Function:             Logit               Scale:    1.0000
          Method:              IRLS      Log-Likelihood:   -2578.5
            Date:  Sat, 04 Mar 2023            Deviance:    5157.1
            Time:          13:08:43        Pearson chi2:  6.41e+03
  No. Iterations:                20  Pseudo R-squ. (CS):    0.4061
 Covariance Type:         nonrobust

                                                       coef   std err        z  P>|z|     [0.025    0.975]
                                          const      1.1007     0.584    1.885  0.059     -0.044     2.245
                                   Do Not Email     -1.6344     0.210   -7.798  0.000     -2.045    -1.224
                                    Do Not Call     20.5465     2e+04    0.001  0.999  -3.92e+04  3.92e+04
                    Total Time Spent on Website      1.1000     0.041   27.017  0.000      1.020     1.180
             Lead Origin_Landing Page Submission    -1.1778     0.129   -9.166  0.000     -1.430    -0.926
                         Lead Source_Olark Chat      1.0701     0.123    8.688  0.000      0.829     1.311
                          Lead Source_Reference      3.2950     0.243   13.585  0.000      2.820     3.770
                    Lead Source_Welingak Website      5.8550     0.731    8.014  0.000      4.423     7.287
                Last Activity_Email Link Clicked      0.4465     0.396    1.127  0.260     -0.330     1.223
                     Last Activity_Email Opened      0.6657     0.184    3.624  0.000      0.306     1.026
            Last Activity_Olark Chat Conversation    -0.6244     0.226   -2.767  0.006     -1.067    -0.182
                   Last Activity_Other_Activity      2.0983     0.550    3.815  0.000      1.020     3.176
                         Last Activity_SMS Sent      1.1008     0.186    5.930  0.000      0.737     1.465
                      Last Activity_Unreachable      0.3232     0.462    0.700  0.484     -0.582     1.228
         Specialization_Hospitality Management    -0.4378     0.329   -1.331  0.183     -1.083     0.207
                          Specialization_Others    -1.1978     0.126   -9.489  0.000     -1.445    -0.950
 What is your current occupation_Working Professional  2.6052   0.196   13.316  0.000      2.222     2.989
                              City_Tier II Cities    -0.5805     0.454   -1.279  0.201     -1.470     0.309
              Last Notable Activity_Email Bounced    -0.8884     0.797   -1.115  0.265     -2.451     0.674
         Last Notable Activity_Email Link Clicked    -2.0765     0.718   -2.891  0.004     -3.484    -0.669
               Last Notable Activity_Email Opened    -1.8440     0.579   -3.183  0.001     -2.979    -0.709
                  Last Notable Activity_Modified    -2.2160     0.562   -3.942  0.000     -3.318    -1.114
     Last Notable Activity_Olark Chat Conversation    -1.9300     0.671   -2.876  0.004     -3.245    -0.615
      Last Notable Activity_Page Visited on Website    -1.4547     0.606   -2.402  0.016     -2.642    -0.268
                  Last Notable Activity_SMS Sent    -0.8400     0.581   -1.445  0.148     -1.979     0.299
              Last Notable Activity_Unsubscribed      0.2017     0.779    0.259  0.796     -1.325     1.729
   Last Notable Activity_View in browser link Clicked  -23.9672  2.92e+04  -0.001  0.999  -5.73e+04  5.73e+04
```

Terminologies Required

Before proceeding ahead, we need to understand few terminologies

- **Conversion of categorical columns to numerical.** This step is done as our algorithm runs only on numerical data.
- **Feature Scaling.** This is done to bring our data into same scale.
- **Data Splitting:** We have split the data into 80:20 and named it as train data and test data. We run model on train data and validate our model on test data.
- **Confusion Matrix:**

	Predicted No	Predicted Yes
Actual No	True Negative	False Negative
Actual Yes	False Positive	True Positive



Where,

True positive (TP): correct positive prediction False
positive (FP): incorrect positive prediction True
negative (TN): correct negative prediction False
negative (FN): incorrect negative prediction

Above Metrics is known as **Confusion Matrix**, using above metrics we derived following things:

1. **Accuracy** = (True Negative + True Positive)/Total

   This metrics provides the accuracy of the model, where total is TP + FN + FP +FN

2. **Sensitivity** = True Positive / (True Positive + False Positive)

   Sensitivity (SN) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true positive rate (TPR). The best **sensitivity** is 1.0, whereas the worst is 0.0.

3. **Specificity =** True Negative/ (True Negative + False Negative)

   Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR). The best **specificity** is 1.0, whereas the worst is 0.0.

4. **Precision =** True Positive/ (True Positives +False Positives)

Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives.

5. **Recall** = True Positives/(True Positives +False Negatives)

The precise definition of **recall** is the number of true positives divided by the number of true positives plus the number of false negatives. True positives are data point classified as positive by the model that actually are positive (meaning they are correct), and false negatives are data points the model identifies as negative that actually are positive (incorrect).

Model Metrics

Running model on features selected we get following metrics:

1. **Train Data:**
   - **Confusion Matrix**

	Not Converted Leads	Converted Leads
Not Converted Leads	3634	379
Converted Leads	273	2182

   - Accuracy: 81.0%
   - Sensitivity: 81.7%
   - Specificity: 80.6%
   - Precision: 79.0%
   - Recall: 71.0%

2. **Test Data:**

- **Confusion Matrix**

	Not Converted Leads	Converted Leads
Not Converted Leads	1570	117
Converted Leads	164	921

- Accuracy: 80.4%
- Sensitivity: 80.4%
- Specificity: 80.5%
- Precision: 79.0%
- Recall:71.0%

The model seems to predict the conversion rate very well. We should be able to help the education company select the most promising Leads or the Hot Leads.

Conclusion

Focus:

Company should focus on following features to increase the leads

- **Tags_Closed by Horizon:** Leads that have been assigned Tags as ‘closed by horizon’ have the highest probability of conversion.
- **Tags_Lost:** Leads that have been tagged as ‘Lost ‘also contribute to the conversion to a considerable extent.
- **Tags_Will revert after reading the email:** Leads that have been tagged as ‘will revert after reading the mail’ also have significant correlation with the conversion.

Expansion:

Company should also focus on Lead Score (which are the probabilities obtained via algorithm) which are greater than 80% to expedite the conversion

Thank You!