Lead Scoring Case Study

Presented By

- Dheivameena K
- Rahul R
- Shantam Garg

Problem Statement

X Education needs help in selecting 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%.

Approach

We have build this model using Logistic regression along with RFE, to get top features and based on that we have provided recommendations to the company.

Steps Followed

1. EDA
2. Data Pre Processing
3. Train & Test Data Split
4. Model Building
5. Metrics Score & Analysis

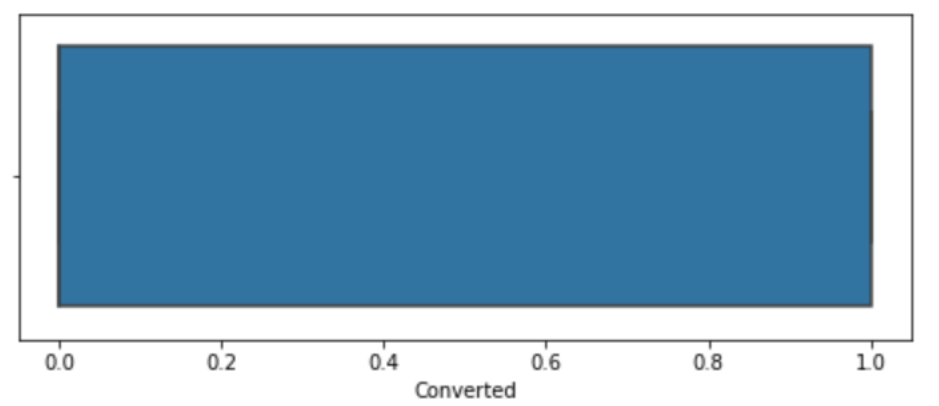
EDA

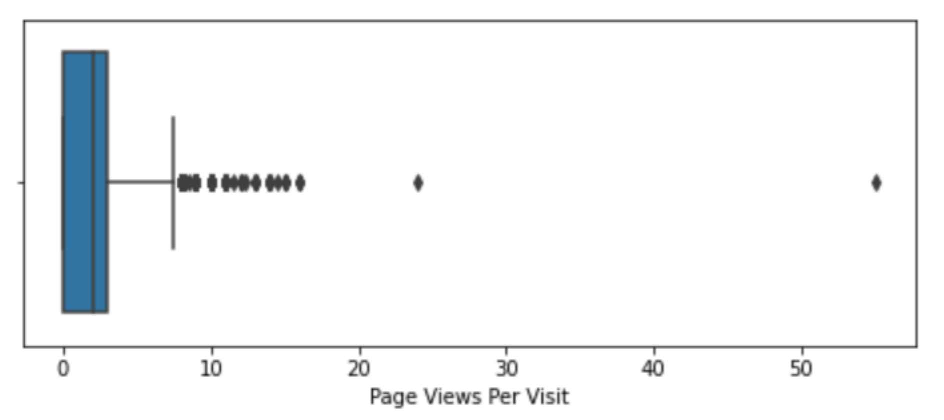
1. *Understating the Data Set:*

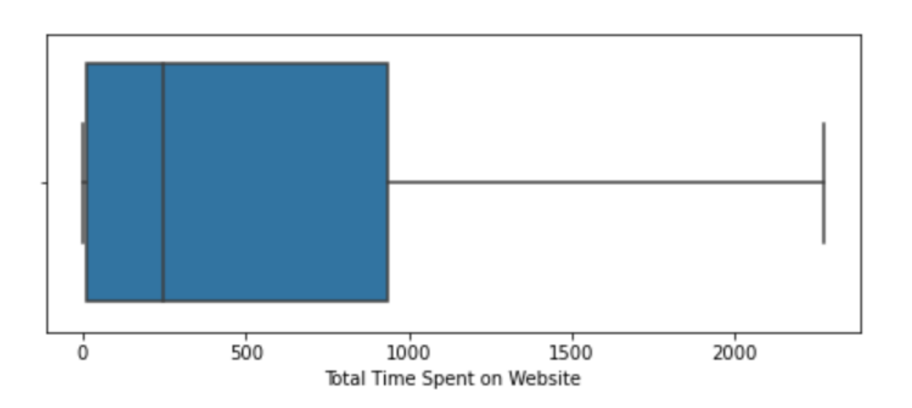
- In The given data set there was total of **9247 records** with **37 attributes.**
- Data contains high number of missing values which we have handled by capped the null values to 40%, anything above 40% was dropped.
- The country column is dropped As most of the records belongs to 'India' this variable is not significant and will not help much in classification, it is better to drop this column.
- After observing the columns we found the biasing in certain columns(i.e., one class is relatively higher than other). We need to drop these columns because they lack variation.
- Finally we have cut down to **9247 records** and **16 attributes.**

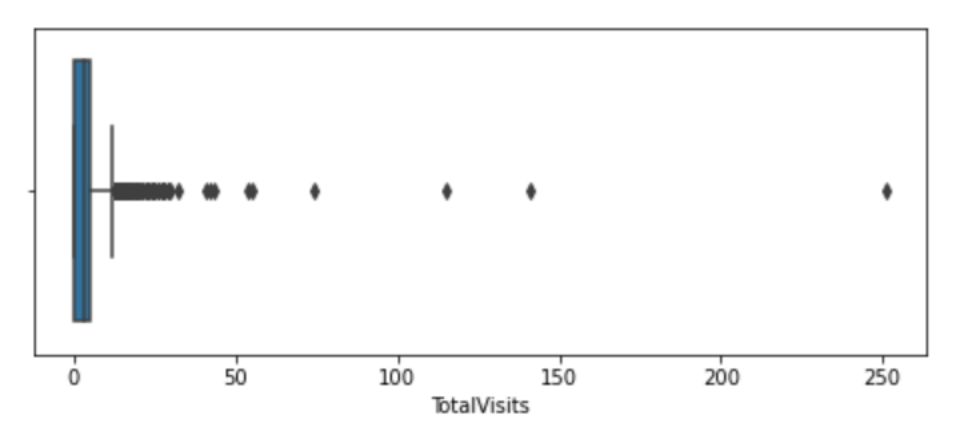
EDA

1. *<u>Outlier Check:</u>* We did some univariate analysis and then outlier treatment these were some potential outliers we did capping of 99%





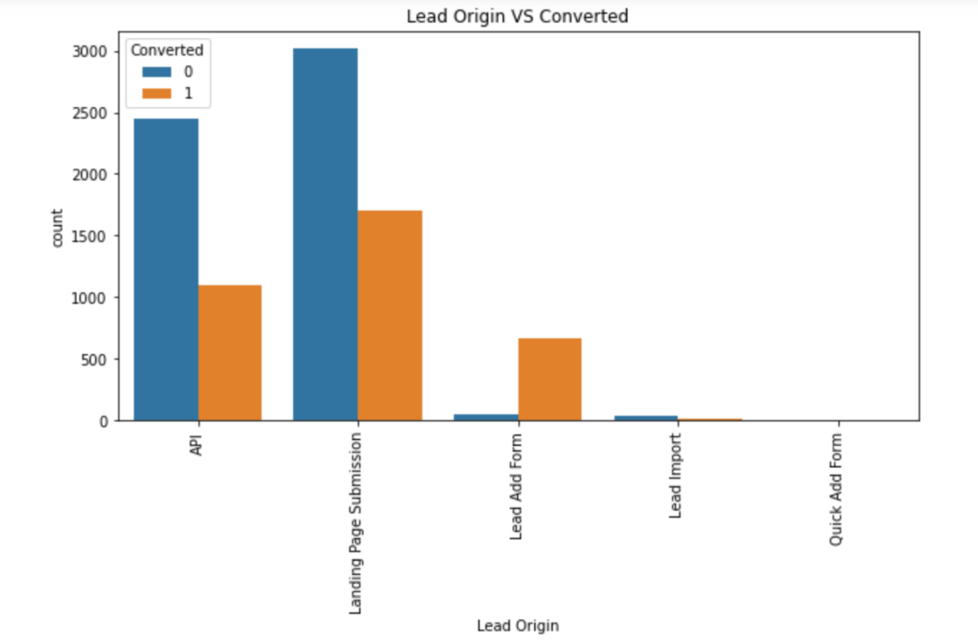


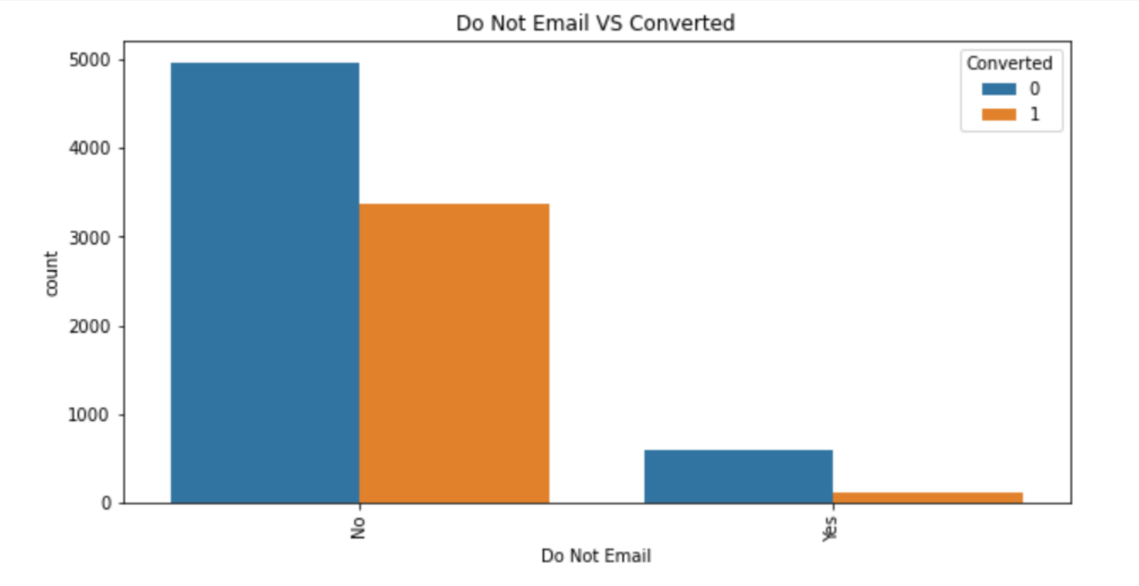


EDA

2. *Visualizing The Data:* We did some bivariate analysis and these are the inferences

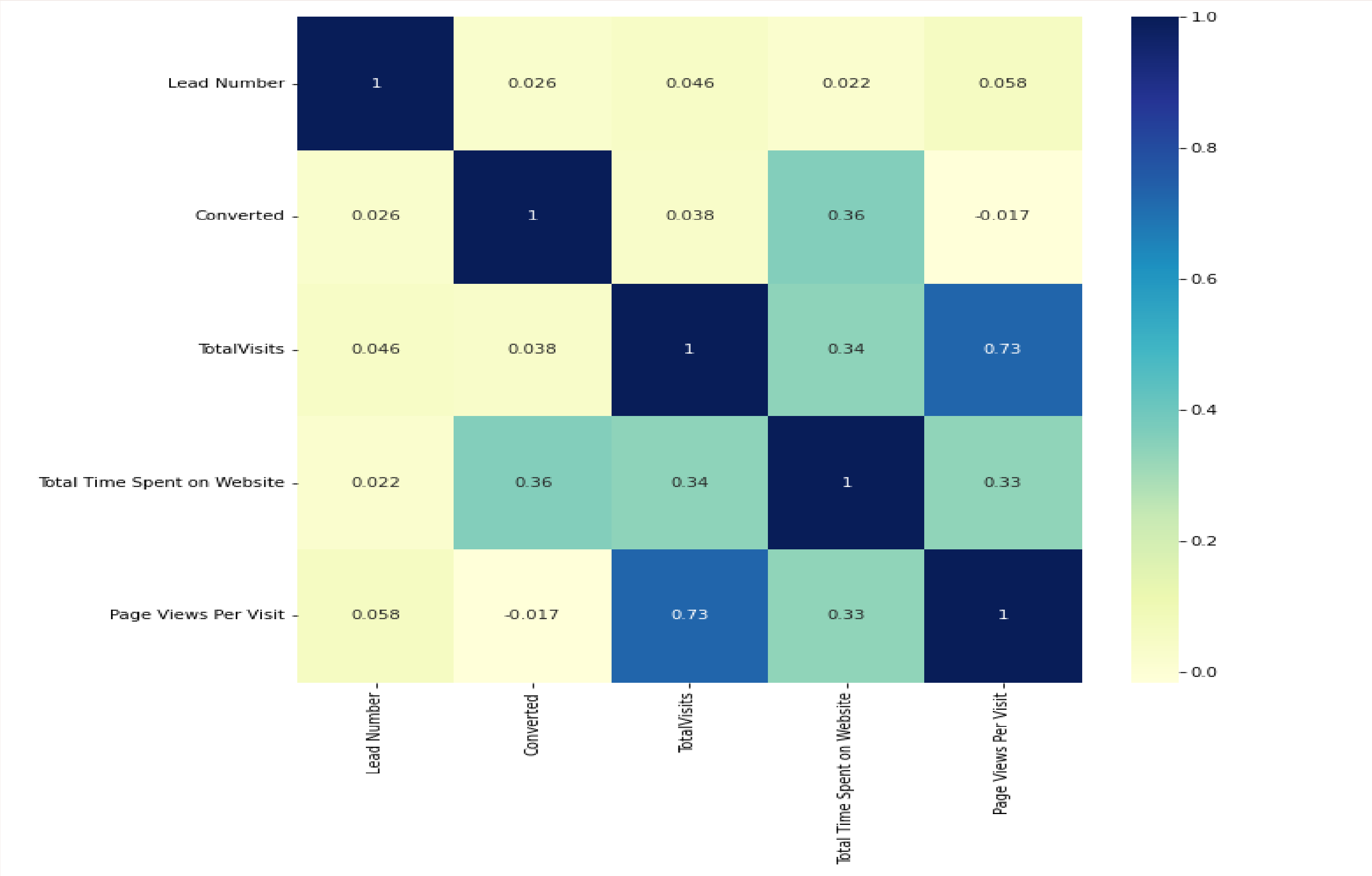
- Comparative to other lead origins categories 'Lead Add Form' category has the highest conversion ratio.
- 'Reference' category in 'Lead Source' column is doing good followed by 'Google' & 'Direct Traffic' in conversions.
- Chance of conversion increases when the customer do not decline for Email.
- 'SMS Sent' category in 'Last Activity' column has highest conversion ratio followed by 'Email Opened'.
- 'Working Professional' seems to convert more as compared to 'Unemployed' Ones.





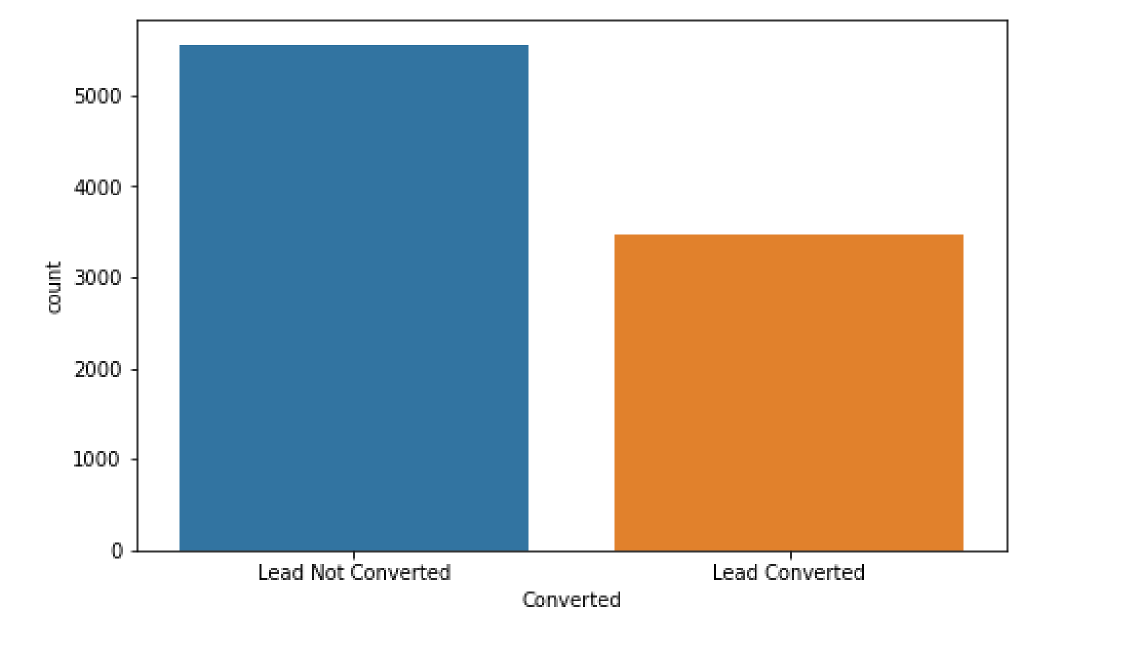
EDA

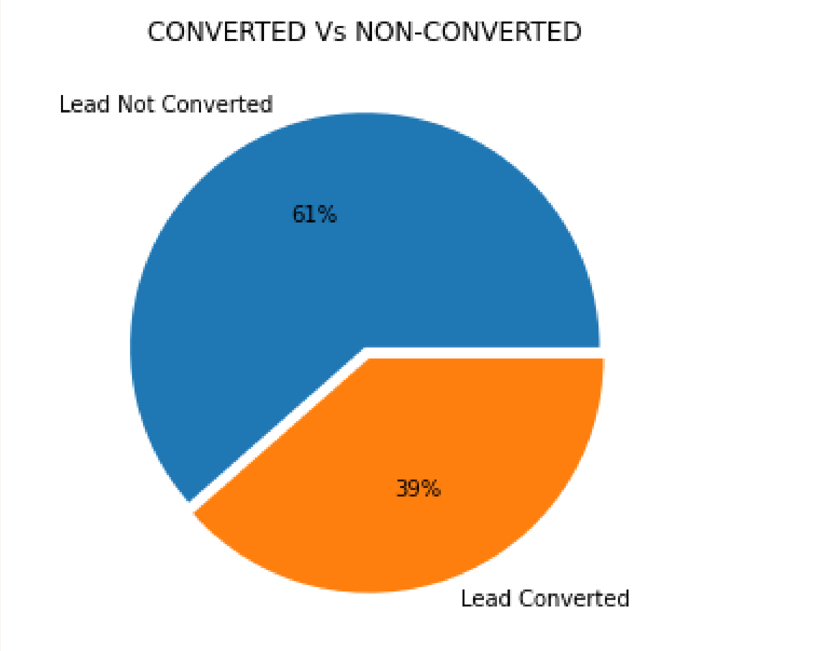
- Below is the correlation matrix, total visits' have high correlation with 'leads number'



EDA

- There doesn't seem to be data imbalance in the 'Target Variable'.





Building Model

- In The 1st Model Built, We observed 'Lead Origin_Quick Add Form' has High p-value and hence it is Insignificant.

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6320
Model:                            GLM   Df Residuals:                     6299
Model Family:                Binomial   Df Model:                           20
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1464.5
Date:                Mon, 23 Jan 2023   Deviance:                       2928.9
Time:                        02:34:37   Pearson chi2:                 7.97e+03
No. Iterations:                    19   Pseudo R-squ. (CS):             0.5826
Covariance Type:            nonrobust
=================================================================================================
                                           coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                                    3.6585      0.734      4.987      0.000       2.221       5.096
Do Not Email                            -1.0095      0.254     -3.971      0.000      -1.508      -0.511
Total Time Spent on Website              1.0797      0.055     19.458      0.000       0.971       1.188
Lead Origin_Landing Page Submission     -0.4537      0.125     -3.627      0.000      -0.699      -0.209
Lead Origin_Lead Add Form                1.8881      0.332      5.683      0.000       1.237       2.539
Lead Origin_Quick Add Form              21.4276   1.77e+04      0.001      0.999   -3.47e+04    3.48e+04
Lead Source_Olark Chat                   1.0588      0.158      6.683      0.000       0.748       1.369
Lead Source_Welingak Website             3.6616      0.792      4.623      0.000       2.109       5.214
Last Activity_Email Bounced             -1.1005      0.536     -2.053      0.040      -2.151      -0.050
Last Activity_Email Opened               0.7025      0.152      4.631      0.000       0.405       1.000
Last Activity_Olark Chat Conversation   -0.9013      0.242     -3.727      0.000      -1.375      -0.427
Last Activity_SMS Sent                   1.1401      0.210      5.439      0.000       0.729       1.551
Specialization_Travel and Tourism       -0.5952      0.361     -1.648      0.099      -1.303       0.112
current_occup_Working Professional       1.3978      0.360      3.878      0.000       0.691       2.104
Tags_Interested in other courses        -8.3210      0.821    -10.138      0.000      -9.930      -6.712
Tags_Others                             -6.1314      0.725     -8.454      0.000      -7.553      -4.710
Tags_Ringing                            -9.0175      0.754    -11.954      0.000     -10.496      -7.539
Tags_Tags_Not_Specified                 -5.8460      0.724     -8.080      0.000      -7.264      -4.428
Tags_Will revert after reading the email -1.4798     0.738     -2.004      0.045      -2.927      -0.032
Last Notable Activity_Other_Notable_activity 1.3212  0.453      2.919      0.004       0.434       2.208
Last Notable Activity_SMS Sent           1.6147      0.189      8.541      0.000       1.244       1.985
=================================================================================================
```

Building Model

- In The 2nd Model Built, We observed ‘Specialization Travel and Tourism’ has High p-value and hence it is Insignificant.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6320
Model:	GLM	Df Residuals:	6300
Model Family:	Binomial	Df Model:	19
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1465.6
Date:	Mon, 23 Jan 2023	Deviance:	2931.2
Time:	02:34:37	Pearson chi2:	7.98e+03
No. Iterations:	8	Pseudo R-squ. (CS):	0.5824
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	3.6644	0.734	4.995	0.000	2.227	5.102
Do Not Email	-0.9997	0.254	-3.938	0.000	-1.497	-0.502
Total Time Spent on Website	1.0845	0.055	19.569	0.000	0.976	1.193
Lead Origin_Landing Page Submission	-0.4617	0.125	-3.695	0.000	-0.707	-0.217
Lead Origin_Lead Add Form	1.8848	0.332	5.673	0.000	1.234	2.536
Lead Source_Olark Chat	1.0576	0.158	6.678	0.000	0.747	1.368
Lead Source_Welingak Website	3.6558	0.791	4.619	0.000	2.105	5.207
Last Activity_Email Bounced	-0.9526	0.512	-1.860	0.063	-1.956	0.051
Last Activity_Email Opened	0.6995	0.152	4.615	0.000	0.402	0.997
Last Activity_Olark Chat Conversation	-0.9052	0.242	-3.745	0.000	-1.379	-0.431
Last Activity_SMS Sent	1.1368	0.210	5.424	0.000	0.726	1.548
Specialization_Travel and Tourism	-0.6005	0.361	-1.663	0.096	-1.308	0.107
current_occup_Working Professional	1.3980	0.360	3.881	0.000	0.692	2.104
Tags_Interested in other courses	-8.3264	0.821	-10.143	0.000	-9.935	-6.717
Tags_Others	-6.1320	0.725	-8.454	0.000	-7.554	-4.710
Tags_Ringing	-9.0183	0.754	-11.955	0.000	-10.497	-7.540
Tags_Tags_Not_Specified	-5.8444	0.724	-8.078	0.000	-7.262	-4.426
Tags_Will revert after reading the email	-1.4809	0.738	-2.005	0.045	-2.928	-0.034
Last Notable Activity_Other_Notable_activity	1.2623	0.449	2.812	0.005	0.382	2.142
Last Notable Activity_SMS Sent	1.6154	0.189	8.539	0.000	1.245	1.986

Building Model

- In The 3rd Model Built, We observed 'Last Activity_Email Bounced' has High p-value and hence it is Insignificant.

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6320
Model:                            GLM   Df Residuals:                     6301
Model Family:                Binomial   Df Model:                           18
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1467.1
Date:                Mon, 23 Jan 2023   Deviance:                       2934.1
Time:                        02:34:38   Pearson chi2:                 7.97e+03
No. Iterations:                     8   Pseudo R-squ. (CS):             0.5822
Covariance Type:            nonrobust
==================================================================================================================
                                           coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------
const                                    3.6398      0.733      4.964      0.000       2.203       5.077
Do Not Email                            -1.0042      0.253     -3.974      0.000      -1.499      -0.509
Total Time Spent on Website              1.0855      0.055     19.592      0.000       0.977       1.194
Lead Origin_Landing Page Submission     -0.4793      0.125     -3.846      0.000      -0.724      -0.235
Lead Origin_Lead Add Form                1.8915      0.332      5.703      0.000       1.241       2.542
Lead Source_Olark Chat                   1.0597      0.158      6.693      0.000       0.749       1.370
Lead Source_Welingak Website             3.6544      0.791      4.619      0.000       2.104       5.205
Last Activity_Email Bounced             -0.9791      0.514     -1.905      0.057      -1.987       0.028
Last Activity_Email Opened               0.7075      0.152      4.668      0.000       0.410       1.005
Last Activity_Olark Chat Conversation   -0.8958      0.242     -3.709      0.000      -1.369      -0.422
Last Activity_SMS Sent                   1.1342      0.209      5.419      0.000       0.724       1.545
current_occup_Working Professional       1.3981      0.359      3.900      0.000       0.695       2.101
Tags_Interested in other courses        -8.3134      0.821    -10.132      0.000      -9.922      -6.705
Tags_Others                             -6.1141      0.725     -8.432      0.000      -7.535      -4.693
Tags_Ringing                            -8.9916      0.754    -11.926      0.000     -10.469      -7.514
Tags_Tags_Not_Specified                 -5.8278      0.723     -8.057      0.000      -7.246      -4.410
Tags_Will revert after reading the email -1.4791     0.738     -2.003      0.045      -2.926      -0.032
Last Notable Activity_Other_Notable_activity 1.2919  0.448      2.882      0.004       0.413       2.171
Last Notable Activity_SMS Sent           1.6208      0.189      8.579      0.000       1.251       1.991
==================================================================================================================
```

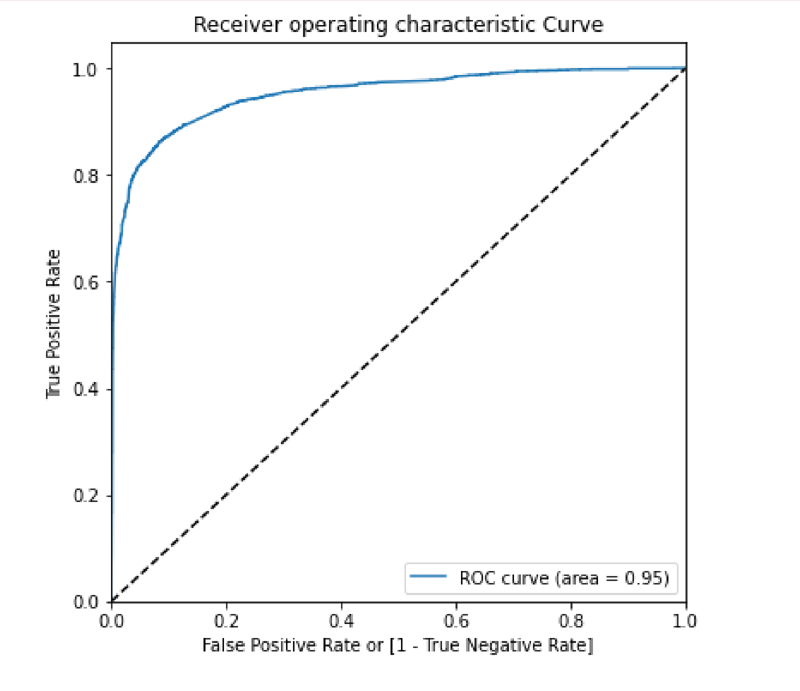
Building Model

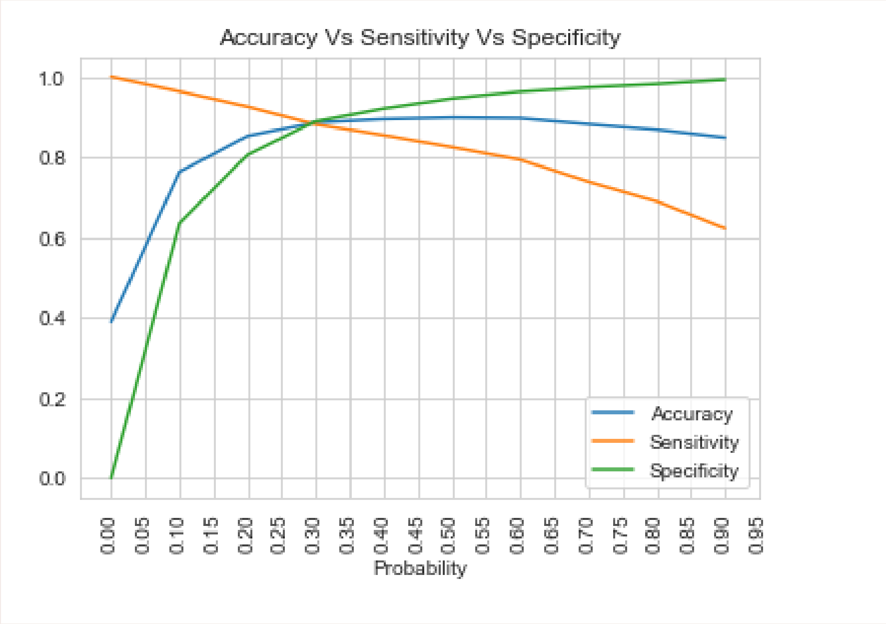
After multiple model build in The 7th Model Built, We observed the p-value are less than 0.05 and VIF values are less than 5. Therefore it seems that all the variables are significant and have low multicollinearity. So we can go ahead and make predictions using this model.

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6320
Model:                            GLM   Df Residuals:                     6305
Model Family:                Binomial   Df Model:                           14
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1648.4
Date:                Mon, 23 Jan 2023   Deviance:                       3296.8
Time:                        02:34:39   Pearson chi2:                 1.39e+04
No. Iterations:                     7   Pseudo R-squ. (CS):             0.5576
Covariance Type:            nonrobust
==============================================================================================================
                                              coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------------
const                                      -1.5087      0.127    -11.845      0.000      -1.758      -1.259
Do Not Email                               -1.4110      0.218     -6.486      0.000      -1.837      -0.985
Total Time Spent on Website                 1.1006      0.052     21.000      0.000       0.998       1.203
Lead Origin_Landing Page Submission        -0.2453      0.117     -2.100      0.036      -0.474      -0.016
Lead Origin_Lead Add Form                   3.4598      0.258     13.435      0.000       2.955       3.965
Lead Source_Olark Chat                      0.9673      0.150      6.439      0.000       0.673       1.262
Lead Source_Welingak Website                2.3933      0.762      3.140      0.002       0.899       3.887
Last Activity_Email Opened                  0.3174      0.113      2.808      0.005       0.096       0.539
Last Activity_Olark Chat Conversation      -1.1898      0.205     -5.801      0.000      -1.592      -0.788
current_occup_Working Professional          2.1326      0.297      7.187      0.000       1.551       2.714
Tags_Interested in other courses           -3.0622      0.404     -7.588      0.000      -3.853      -2.271
Tags_Others                                -0.7350      0.111     -6.593      0.000      -0.954      -0.517
Tags_Ringing                               -3.6568      0.238    -15.377      0.000      -4.123      -3.191
Tags_Will revert after reading the email    3.8268      0.186     20.576      0.000       3.462       4.191
Last Notable Activity_SMS Sent              2.1583      0.134     16.138      0.000       1.896       2.420
==============================================================================================================
```

Metrics Check and Analysis

We did some analysis using roc curve, After observing 'Accuracy Vs Sensitivity Vs Specificity' 0.3 Probability seems to be optimal cutoff.





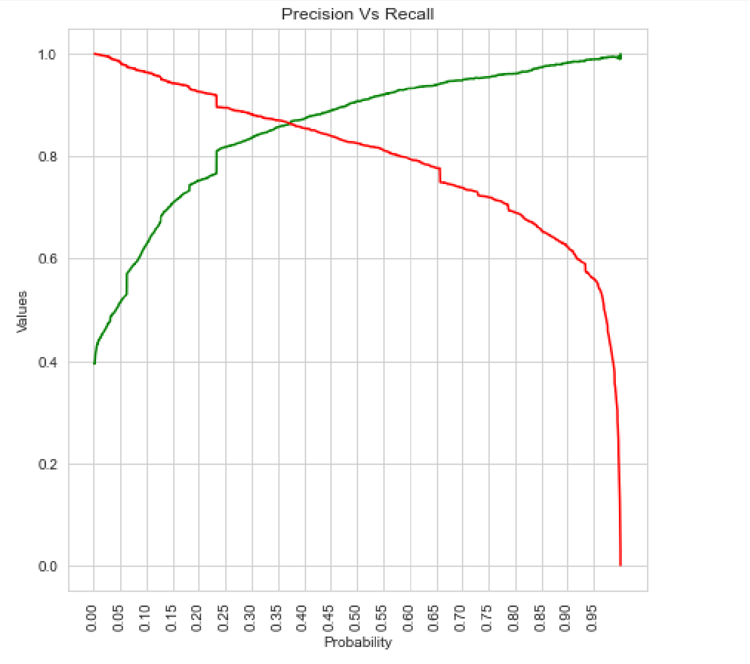
	Probability	Accuracy	Sensitivity	Specificity
0.0	0.0	0.389557	1.000000	0.000000
0.1	0.1	0.762816	0.963851	0.634526
0.2	0.2	0.852373	0.925670	0.805599
0.3	0.3	0.886392	0.881803	0.889321
0.4	0.4	0.894937	0.854184	0.920943
0.5	0.5	0.898734	0.824939	0.945827
0.6	0.6	0.897152	0.793664	0.963193
0.7	0.7	0.882278	0.738424	0.974080
0.8	0.8	0.868196	0.689683	0.982115
0.9	0.9	0.848576	0.622665	0.992742

Metrics Check and Analysis

After observing 'Precision Vs Recall' and dataframe 0.37 seems to be the optimal cutoff.

	precision	recall	threshold
0	0.394425	1.000000	0.001845
1	0.394328	0.999594	0.001864
2	0.394391	0.999594	0.001872
3	0.394454	0.999594	0.001893
4	0.394517	0.999594	0.001898

After Observing both 0.3 and 0.37 Cutoffs:
- 0.37 Gives a bit higher accuracy score.
- All the metrics seems good.
- Also false positive rate is lower for 0.37 cutoff which will help in reduction of falsely predictions.



Metrics Check and Analysis

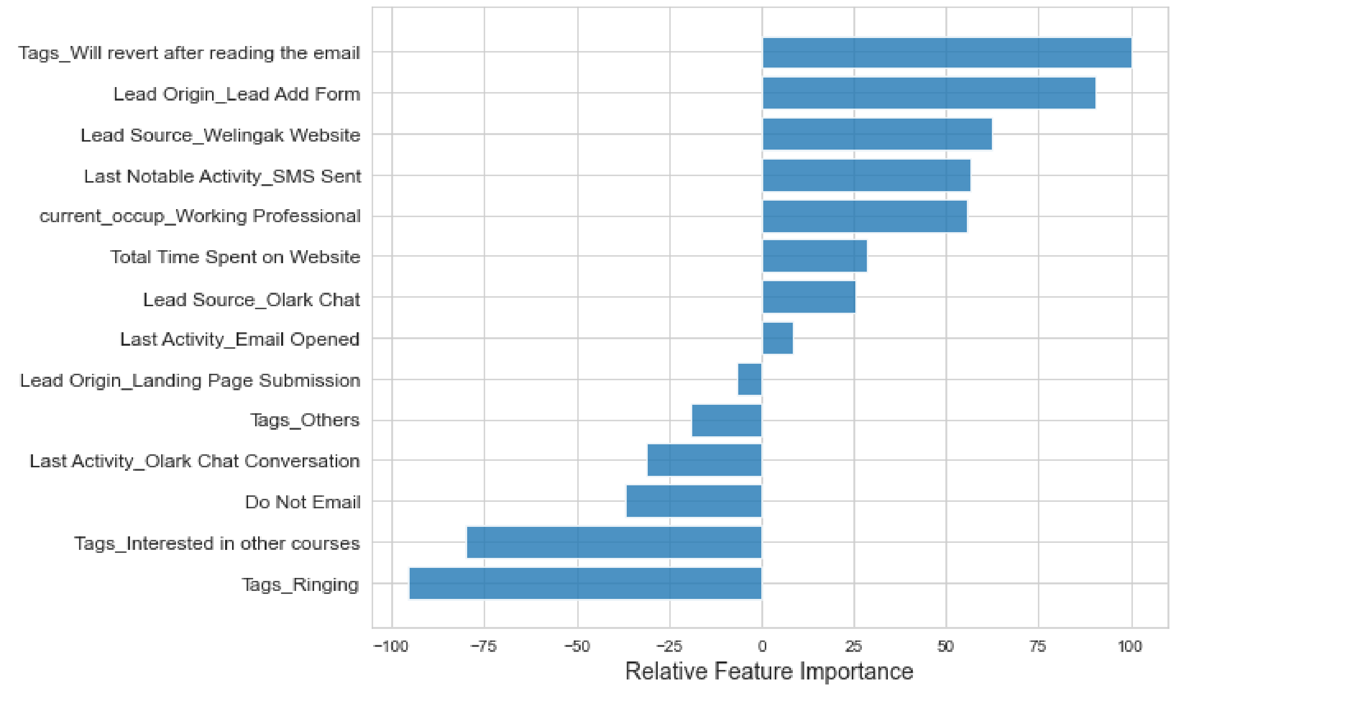
Training Data Set		Vs	Test Data Set	
Sensitivity is	: 86.43	|	Sensitivity is	: 86.90
Specificity is	: 91.16	|	Specificity is	: 89.73
True Positive Rate is	: 86.43	|	True Positive Rate is	: 86.90
False Positive Rate is	: 08.84	|	False Positive Rate is	: 10.27
Precision is	: 86.19	|	Precision is	: 83.52
Recall is	: 86.43	|	Recall is	: 86.90
Accuracy score is	: 89.32	|	Accuracy score is	: 88.67

- The difference b/w train and test data's performance metrics is under 3%. This means that the final model did not overfit training data and is performing well.¶
- High Sensitivity will make sure that all possible leads who are likely to convert are correctly predicted, where as high Specificity will ensure that the leads that are on the brink of the probability of getting converted or not are not selected.
- Based on the business requirement, we can increase or decrease the probability threshold value which in turn will decrease or increase the Sensitivity and increase or decrease the Specificity of the model as required.

Metrics Check and Analysis

The Top 3 Factors which can help in generating more successful leads:

- Tags Will revert after reading the email
- Lead Origin Lead Add Form
- Lead Source_Welingak Website



Recommendations

It Was found that the Variables that mattered the most in the potential buyers are:

- The total time spent on website
- When the last activity was on: SMS sent , olark chat and email opened
- When the Lead origin is lead add form
- Whether the customer is working professional
- When the lead source was: Direct Traffic or Welingak Website

Keeping the Above Factors in Mind The X Education can flourish as they have a very high chance to get almost all the potential buyers to buy their courses.

Top 3 Factors which can help in generating more successful leads:
Tags_Will revert after reading the email,
Lead Origin_Lead Add Form,
Lead Source_Welingak Website.

Thank you