X-EDUCATION LEAD IMPROVEMENTS

SANDEEP SUNKARA

CHETAN PRAKASH

THE PROBLEM

What is the problem?	• Target Lead conversion to be around 80%
Who has this problem?	• Online Course provider X-Education
Why should this problem be solved?	• It would provide focus on potential leads and thus increase conversion
How will I know this problem has been solved?	• Review of the lead conversion rate basis the model definition

BACKGROUND INFORMATION

- X Education sells online courses to industry professionals
- The company markets its courses on several websites and search engines like Google. Company also get leads through referrals
- A lead is one where email and mobile number is entered on their website
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.
- X-Education has realized their lead conversion is poor and they are not focusing on the potential leads

DATA CLEANING & IMPUTING

Step #1

- Removing the Asymmetrique null rows as these are values defined a predefined rational outside the scope of the model.
- Removing Tags, Lead Quality, Lead Profile as these are subjective values.
- Removing City & country as this is not relevant for a company focused on delivering content online
- Removing Last Activity data that are null as these are subjective

Step #2

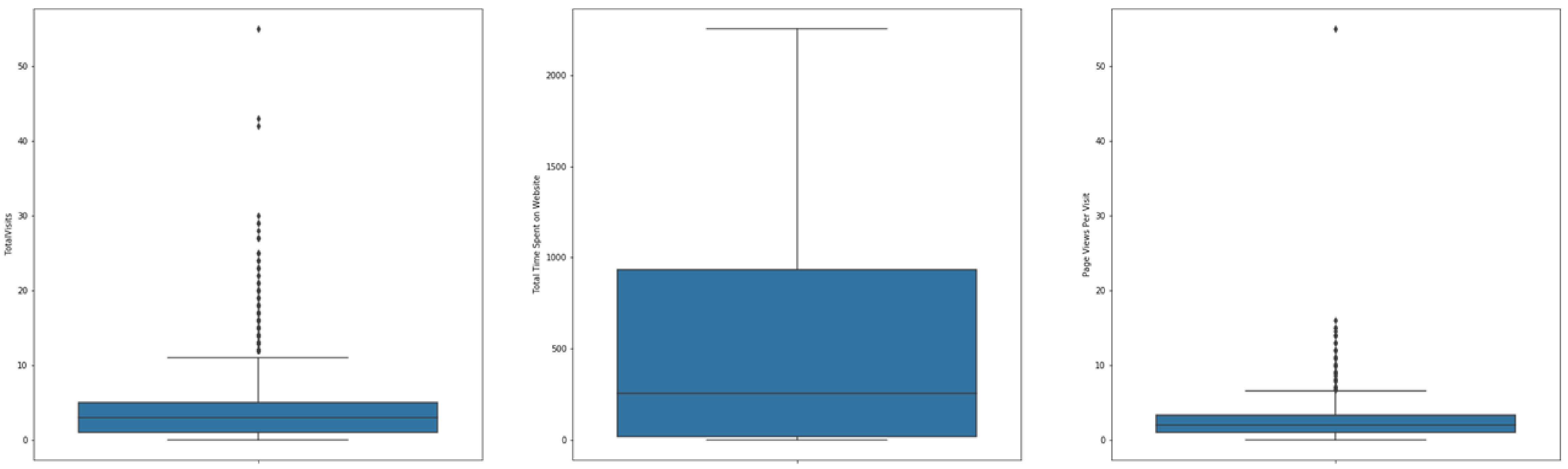
- Using existing distribution logic for Course expectation & Current Occupation field by using mode values
- Imputing Select options as null for Specialization & How they heard about the company fields. Post analysis as these fields had high null values removing these columns

Step #3

- Imputing Total Visits & Page Views per visit using the mode values of the time spent on website for similar data range
- Imputing missing values for Lead Source basis Lead origin and their mode values for the data

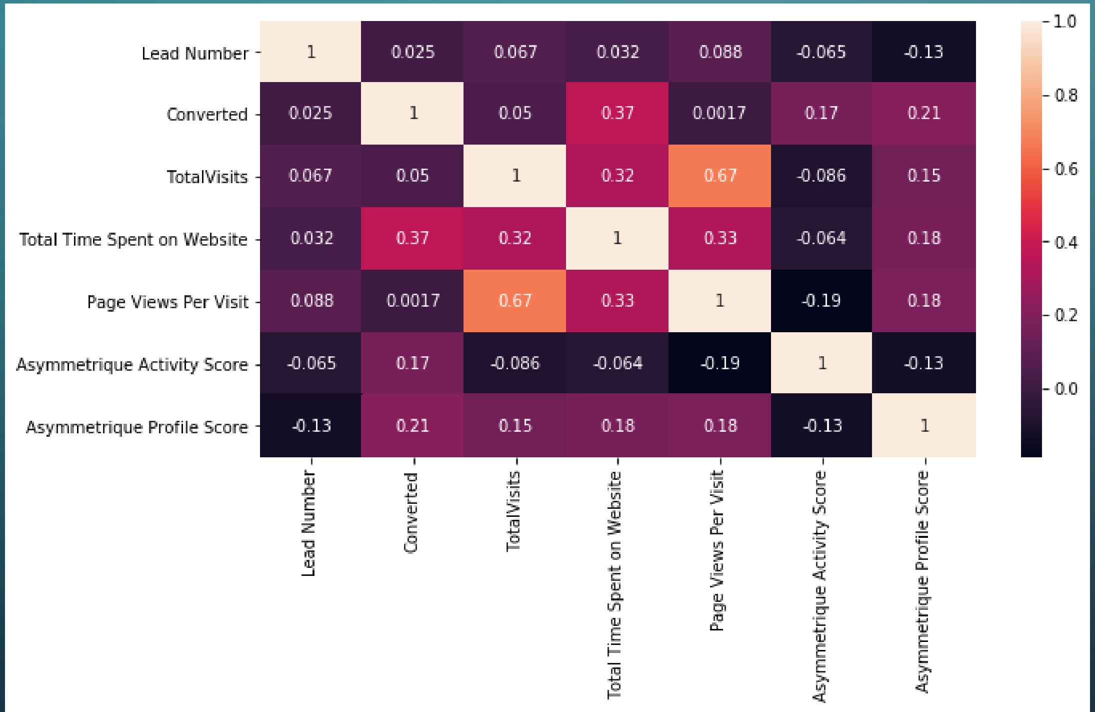
OUTLIER ANALYSIS

- Removing the outlier for Total Visits and Page Views per visit.



CORRELATION

- Quick view of the correlation shows certain negative correlations among numeric fields



	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
Lead Number	1	0.025	0.067	0.032	0.088	-0.065	-0.13
Converted	0.025	1	0.05	0.37	0.0017	0.17	0.21
TotalVisits	0.067	0.05	1	0.32	0.67	-0.086	0.15
Total Time Spent on Website	0.032	0.37	0.32	1	0.33	-0.064	0.18
Page Views Per Visit	0.088	0.0017	0.67	0.33	1	-0.19	0.18
Asymmetrique Activity Score	-0.065	0.17	-0.086	-0.064	-0.19	1	-0.13
Asymmetrique Profile Score	-0.13	0.21	0.15	0.18	0.18	-0.13	1

CATEGORICAL VARIABLE TREATING

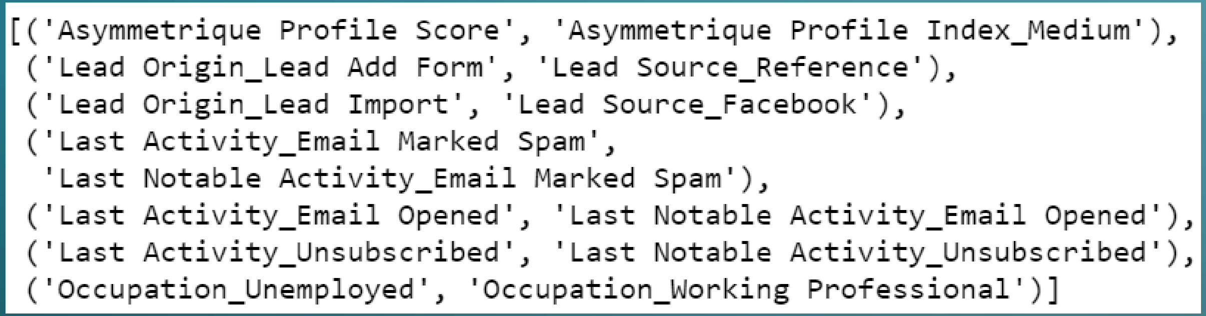
Step #1

- Treating Binary fields like Yes & No with 1 & 0.
- Removing Binary fields that have no variance in their information like Magazine, X Education Forums, Newspaper, Receive More Updates About Our Courses, Update me on Supply Chain Content, Get updates on DM Content & I agree to pay the amount through cheque

Step #2

- Replacing other categorical variables with dummy variables where n-categorical variables are replaced by n-1 ordinal columns

SCALING OF DATA & CORRELATIONS

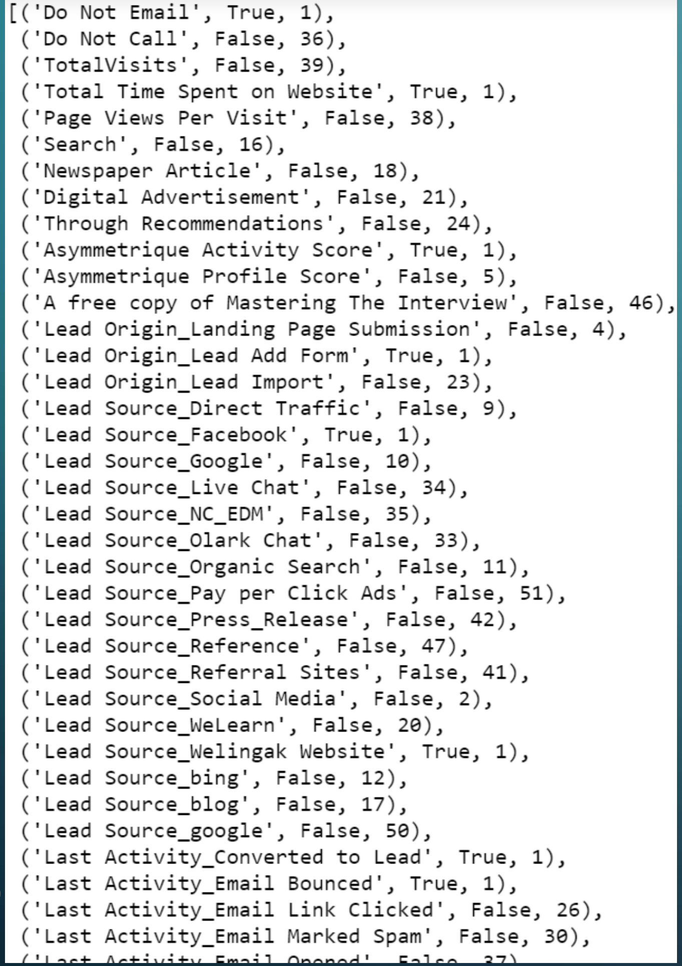


```
[('Asymmetrique Profile Score', 'Asymmetrique Profile Index_Medium'),
 ('Lead Origin_Lead Add Form', 'Lead Source_Reference'),
 ('Lead Origin_Lead Import', 'Lead Source_Facebook'),
 ('Last Activity_Email Marked Spam',
  'Last Notable Activity_Email Marked Spam'),
 ('Last Activity_Email Opened', 'Last Notable Activity_Email Opened'),
 ('Last Activity_Unsubscribed', 'Last Notable Activity_Unsubscribed'),
 ('Occupation_Unemployed', 'Occupation_Working Professional')]
```

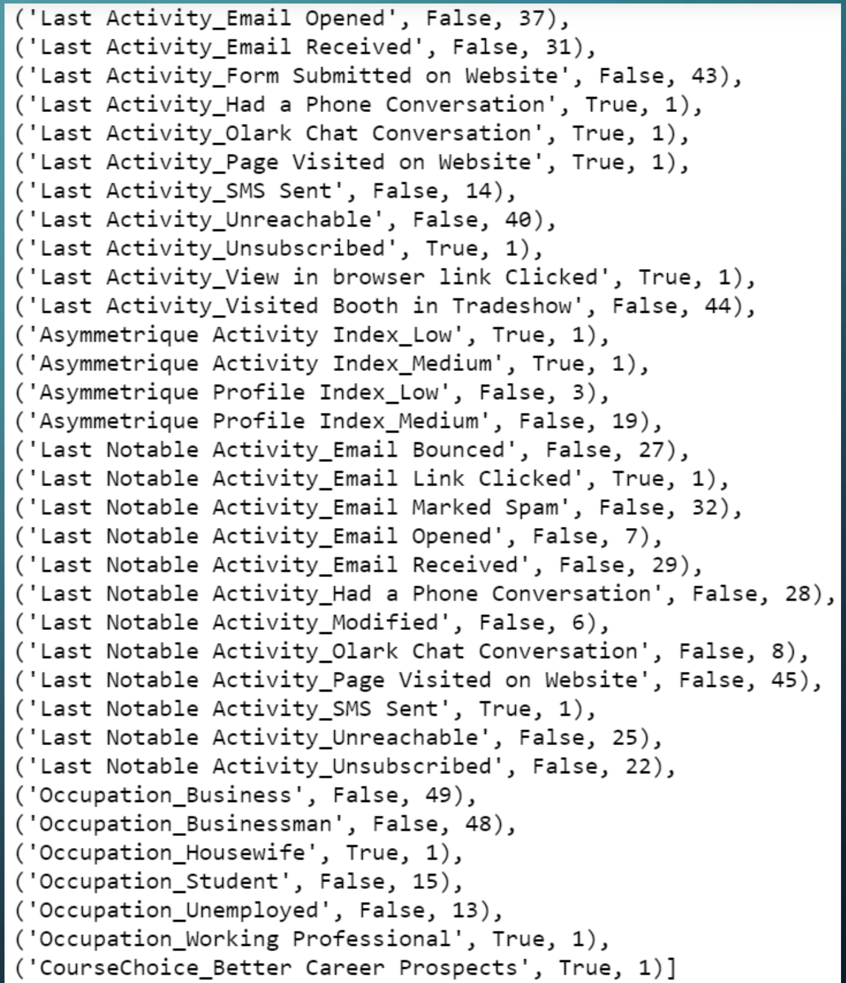
- Scaling the numeric datatypes to bring them to same base.
- Checking the correlation of variables post scaling and converting categorical variables to dummy variables

FEATURE SELECTION

- Using RFE for feature selection of 20 variables to provide options for the company to work on multiple parameters to convert the leads to potential leads. Below are the list of variables selected



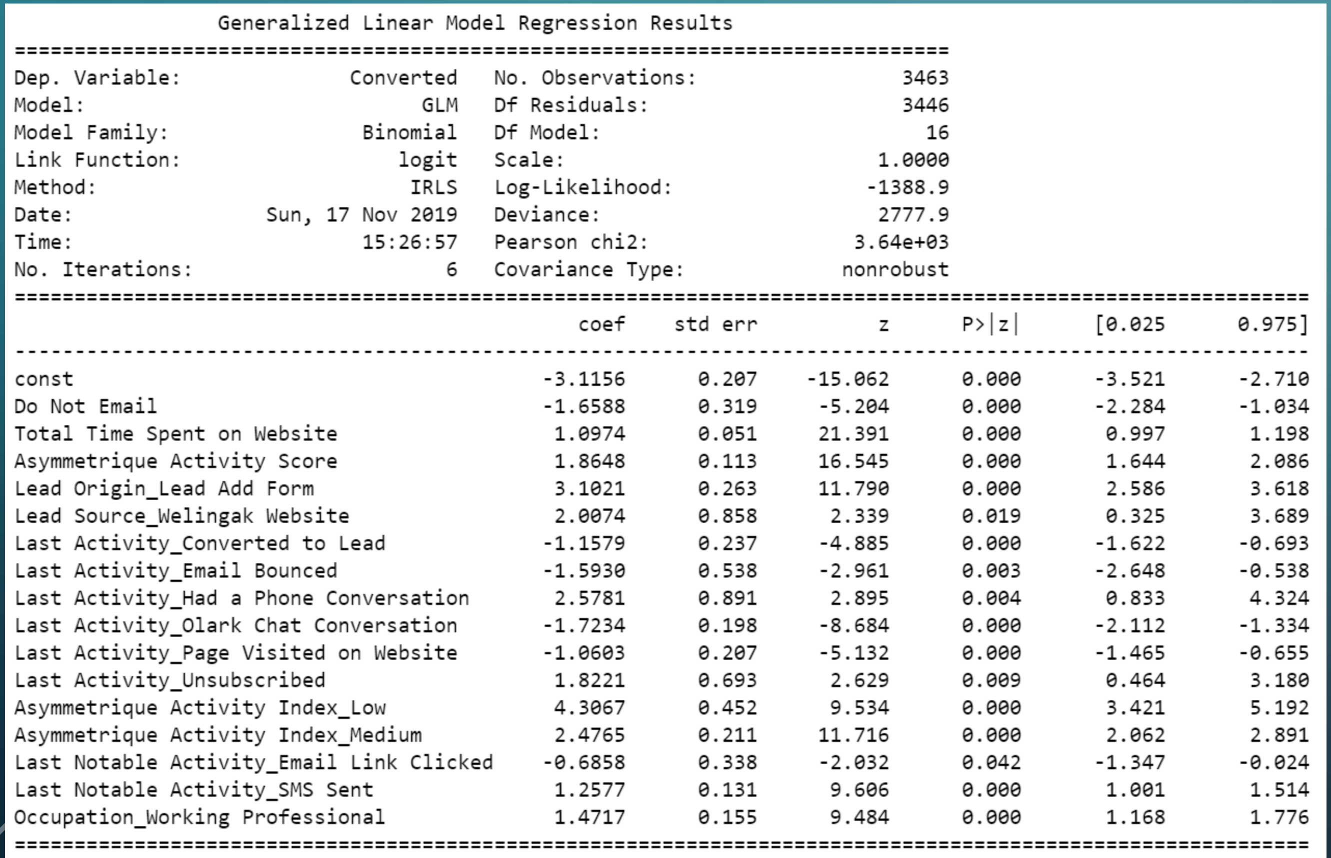
```
[('Do Not Email', True, 1),
 ('Do Not Call', False, 36),
 ('TotalVisits', False, 39),
 ('Total Time Spent on Website', True, 1),
 ('Page Views Per Visit', False, 38),
 ('Search', False, 16),
 ('Newspaper Article', False, 18),
 ('Digital Advertisement', False, 21),
 ('Through Recommendations', False, 24),
 ('Asymmetrique Activity Score', True, 1),
 ('Asymmetrique Profile Score', False, 5),
 ('A free copy of Mastering The Interview', False, 46),
 ('Lead Origin_Landing Page Submission', False, 4),
 ('Lead Origin_Lead Add Form', True, 1),
 ('Lead Origin_Lead Import', False, 23),
 ('Lead Source_Direct Traffic', False, 9),
 ('Lead Source_Facebook', True, 1),
 ('Lead Source_Google', False, 10),
 ('Lead Source_Live Chat', False, 34),
 ('Lead Source_NC_EDM', False, 35),
 ('Lead Source_Olark Chat', False, 33),
 ('Lead Source_Organic Search', False, 11),
 ('Lead Source_Pay per Click Ads', False, 51),
 ('Lead Source_Press_Release', False, 42),
 ('Lead Source_Reference', False, 47),
 ('Lead Source_Referral Sites', False, 41),
 ('Lead Source_Social Media', False, 2),
 ('Lead Source_WeLearn', False, 20),
 ('Lead Source_Welingak Website', True, 1),
 ('Lead Source_bing', False, 12),
 ('Lead Source_blog', False, 17),
 ('Lead Source_google', False, 50),
 ('Last Activity_Converted to Lead', True, 1),
 ('Last Activity_Email Bounced', True, 1),
 ('Last Activity_Email Link Clicked', False, 26),
 ('Last Activity_Email Marked Spam', False, 30),
 ('Last Activity_Email Opened', False, 37),
```



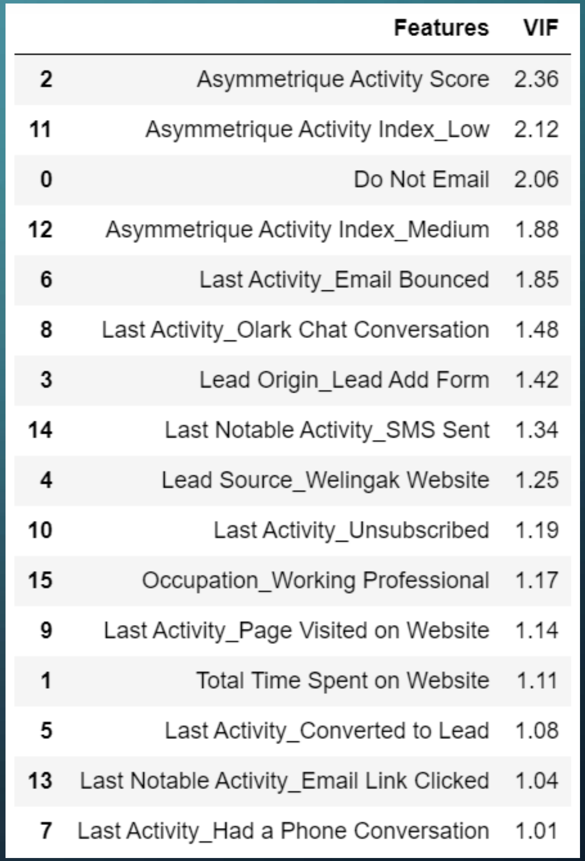
```
('Last Activity_Email Opened', False, 37),
('Last Activity_Email Received', False, 31),
('Last Activity_Form Submitted on Website', False, 43),
('Last Activity_Had a Phone Conversation', True, 1),
('Last Activity_Olark Chat Conversation', True, 1),
('Last Activity_Page Visited on Website', True, 1),
('Last Activity_SMS Sent', False, 14),
('Last Activity_Unreachable', False, 40),
('Last Activity_Unsubscribed', True, 1),
('Last Activity_View in browser link Clicked', True, 1),
('Last Activity_Visited Booth in Tradeshow', False, 44),
('Asymmetrique Activity Index_Low', True, 1),
('Asymmetrique Activity Index_Medium', True, 1),
('Asymmetrique Profile Index_Low', False, 3),
('Asymmetrique Profile Index_Medium', False, 19),
('Last Notable Activity_Email Bounced', False, 27),
('Last Notable Activity_Email Link Clicked', True, 1),
('Last Notable Activity_Email Marked Spam', False, 32),
('Last Notable Activity_Email Opened', False, 7),
('Last Notable Activity_Email Received', False, 29),
('Last Notable Activity_Had a Phone Conversation', False, 28),
('Last Notable Activity_Modified', False, 6),
('Last Notable Activity_Olark Chat Conversation', False, 8),
('Last Notable Activity_Page Visited on Website', False, 45),
('Last Notable Activity_SMS Sent', True, 1),
('Last Notable Activity_Unreachable', False, 25),
('Last Notable Activity_Unsubscribed', False, 22),
('Occupation_Business', False, 49),
('Occupation_Businessman', False, 48),
('Occupation_Housewife', True, 1),
('Occupation_Student', False, 15),
('Occupation_Unemployed', False, 13),
('Occupation_Working Professional', True, 1),
('CourseChoice_Better Career Prospects', True, 1)]
```

FINAL MODEL

- Below is the final model with 16 variables



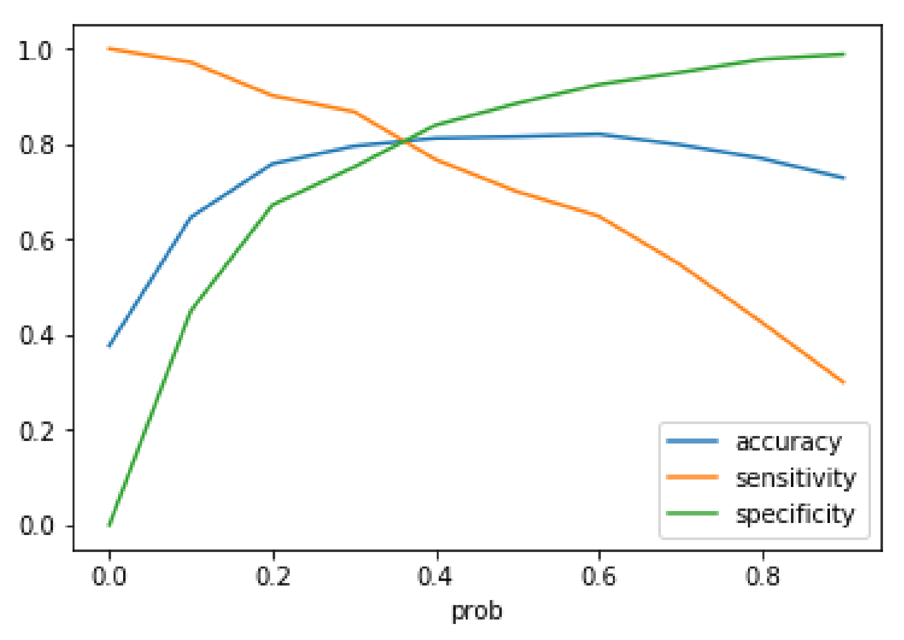
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 3463
Model:                            GLM   Df Residuals:                     3446
Model Family:                Binomial   Df Model:                           16
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1388.9
Date:                Sun, 17 Nov 2019   Deviance:                       2777.9
Time:                        15:26:57   Pearson chi2:                 3.64e+03
No. Iterations:                     6   Covariance Type:             nonrobust
============================================================================================================
                                              coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------
const                                      -3.1156      0.207    -15.062      0.000      -3.521      -2.710
Do Not Email                               -1.6588      0.319     -5.204      0.000      -2.284      -1.034
Total Time Spent on Website                 1.0974      0.051     21.391      0.000       0.997       1.198
Asymmetrique Activity Score                 1.8648      0.113     16.545      0.000       1.644       2.086
Lead Origin_Lead Add Form                   3.1021      0.263     11.790      0.000       2.586       3.618
Lead Source_Welingak Website                2.0074      0.858      2.339      0.019       0.325       3.689
Last Activity_Converted to Lead            -1.1579      0.237     -4.885      0.000      -1.622      -0.693
Last Activity_Email Bounced                -1.5930      0.538     -2.961      0.003      -2.648      -0.538
Last Activity_Had a Phone Conversation      2.5781      0.891      2.895      0.004       0.833       4.324
Last Activity_Olark Chat Conversation      -1.7234      0.198     -8.684      0.000      -2.112      -1.334
Last Activity_Page Visited on Website      -1.0603      0.207     -5.132      0.000      -1.465      -0.655
Last Activity_Unsubscribed                  1.8221      0.693      2.629      0.009       0.464       3.180
Asymmetrique Activity Index_Low             4.3067      0.452      9.534      0.000       3.421       5.192
Asymmetrique Activity Index_Medium          2.4765      0.211     11.716      0.000       2.062       2.891
Last Notable Activity_Email Link Clicked   -0.6858      0.338     -2.032      0.042      -1.347      -0.024
Last Notable Activity_SMS Sent              1.2577      0.131      9.606      0.000       1.001       1.514
Occupation_Working Professional             1.4717      0.155      9.484      0.000       1.168       1.776
============================================================================================================
```

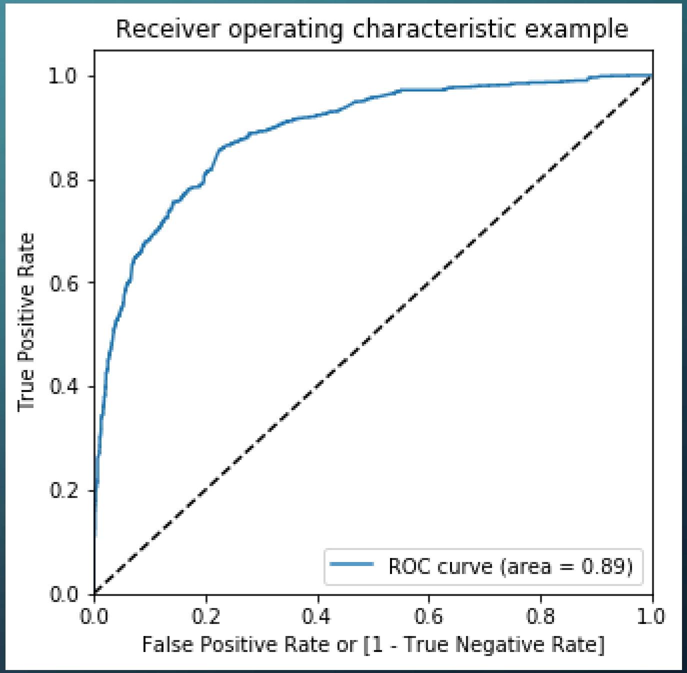


	Features	VIF
2	Asymmetrique Activity Score	2.36
11	Asymmetrique Activity Index_Low	2.12
0	Do Not Email	2.06
12	Asymmetrique Activity Index_Medium	1.88
6	Last Activity_Email Bounced	1.85
8	Last Activity_Olark Chat Conversation	1.48
3	Lead Origin_Lead Add Form	1.42
14	Last Notable Activity_SMS Sent	1.34
4	Lead Source_Welingak Website	1.25
10	Last Activity_Unsubscribed	1.19
15	Occupation_Working Professional	1.17
9	Last Activity_Page Visited on Website	1.14
1	Total Time Spent on Website	1.11
5	Last Activity_Converted to Lead	1.08
13	Last Notable Activity_Email Link Clicked	1.04
7	Last Activity_Had a Phone Conversation	1.01

OPTIMUM CUT-OFF & ROC

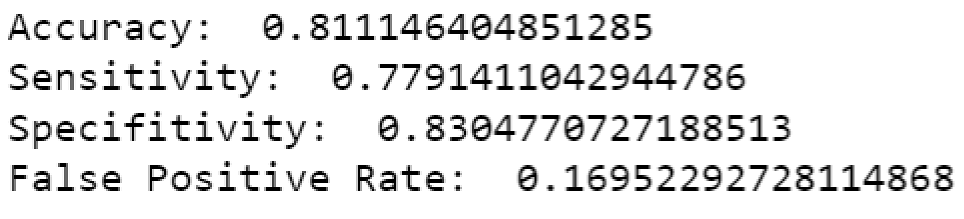
- An optimum Cut-off of 0.38 is observed considering a balance between True Positive Rate & True Negative Rate.





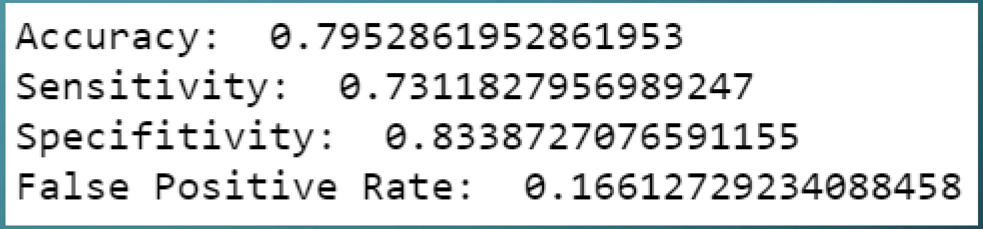
FINAL RESULT

- Accuracy, Specificity, Sensitivity & False Positive Rates for Train & Test Data Sets.
- Also the final dataset has a Lead Score ranging from 0 – 100 which will be used by



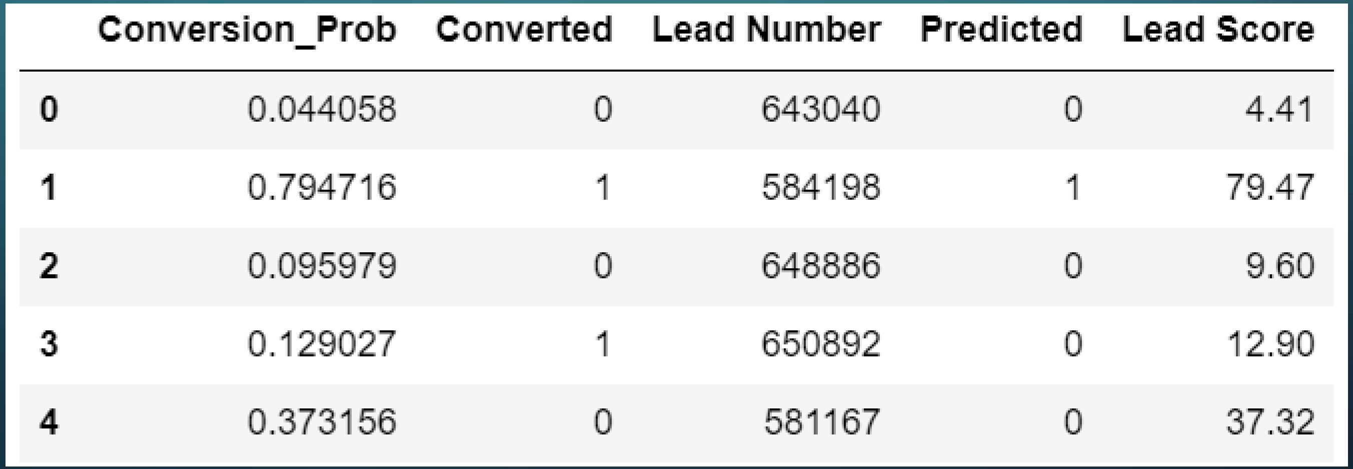
```
Accuracy:   0.811146404851285
Sensitivity:   0.779141104294786
Specifitivity:   0.8304770727188513
False Positive Rate:   0.16952292728114868
```

Train Dataset



```
Accuracy:   0.7952861952861953
Sensitivity:   0.7311827956989247
Specifitivity:   0.8338727076591155
False Positive Rate:   0.1661272923408458
```

Test Dataset



	Conversion_Prob	Converted	Lead Number	Predicted	Lead Score
0	0.044058	0	643040	0	4.41
1	0.794716	1	584198	1	79.47
2	0.095979	0	648886	0	9.60
3	0.129027	1	650892	0	12.90
4	0.373156	0	581167	0	37.32