Presentation
Lead Scoring Case Study

Problem Statement

X Education 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 .

Since the sales team will now be focusing more on communicating with the potential leads

rather than making calls to everyone.

Data Cleaning: Step 1

Columns with Value "Select“ filled by np.nan values

1.Specialization,
2.How did you hear about X Education,
3.Lead Profile,
4.City

Data Cleaning: Step 2

Dropping below Columns with more than 45 % null values

1.How did you hear about X Education Dropping since around 78% data missing
2.Lead Quality Dropping since 51% data missing and is someone's judgement not based on facts collected
3.Lead Profile Dropping since 74% data missing and is someone's judgement not based on facts collected
4.Asymmetrique Activity Index Dropping since 46% data missing and is someone's judgement not based on facts collected
5.Asymmetrique Profile Index Dropping since 46% data missing and is someone's judgement not based on facts collected
6.Asymmetrique Activity Score Dropping since 46% data missing and is someone's judgement not based on facts collected
7.Asymmetrique Profile Score Dropping since 46% data missing and is someone's judgement not based on facts
8.Tags Dropping since not available at the time of prediction but after sales is closed

Data Cleaning: Step 3

Imputing Values for the columns having less than 45 % missing values

Imputing Unknown for missing value in Country Column
Imputing Other for missing value in Specialization Column
Imputing Other for missing value in Column What is your current occupation
Imputing Unknown for missing value in Column What matters most to you in choosing a course
Imputing Other City for missing value in City Column

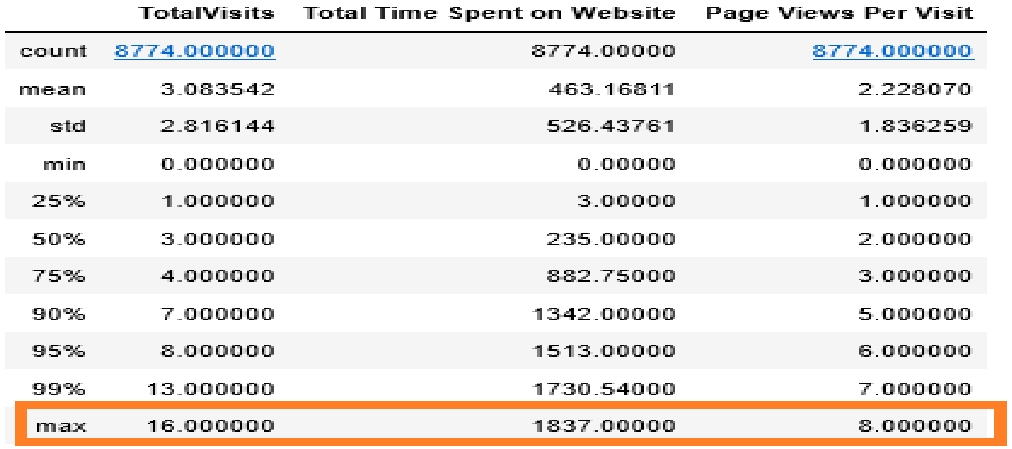
Data Cleaning: Step 4

Dropping below Categorical Columns with single Level or if Particular Level is repetitive for majority of leads

#Do Not Call
# Magazine
#Newspaper Article
#Digital Advertisement
#Through Recommendations
#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
#Newspaper
#Search

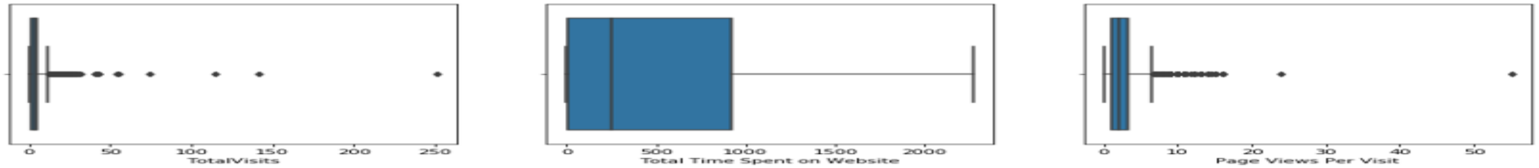
Data Cleaning: Step 5

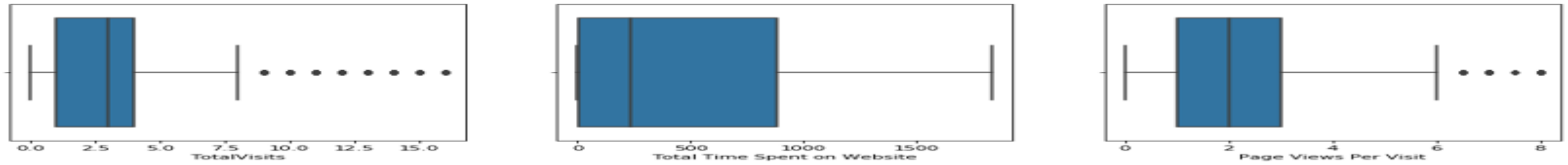
Removing 1% outliers outside 99% percentile



	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	8774.000000	8774.00000	8774.000000
mean	3.083542	463.16811	2.228070
std	2.816144	526.43761	1.836259
min	0.000000	0.00000	0.000000
25%	1.000000	3.00000	1.000000
50%	3.000000	235.00000	2.000000
75%	4.000000	882.75000	3.000000
90%	7.000000	1342.00000	5.000000
95%	8.000000	1513.00000	6.000000
99%	13.000000	1730.54000	7.000000
max	16.000000	1837.00000	8.000000

Before Dropping 1% Outliers after 99 Percentile





After Dropping 1% Outliers after 99 Percentile

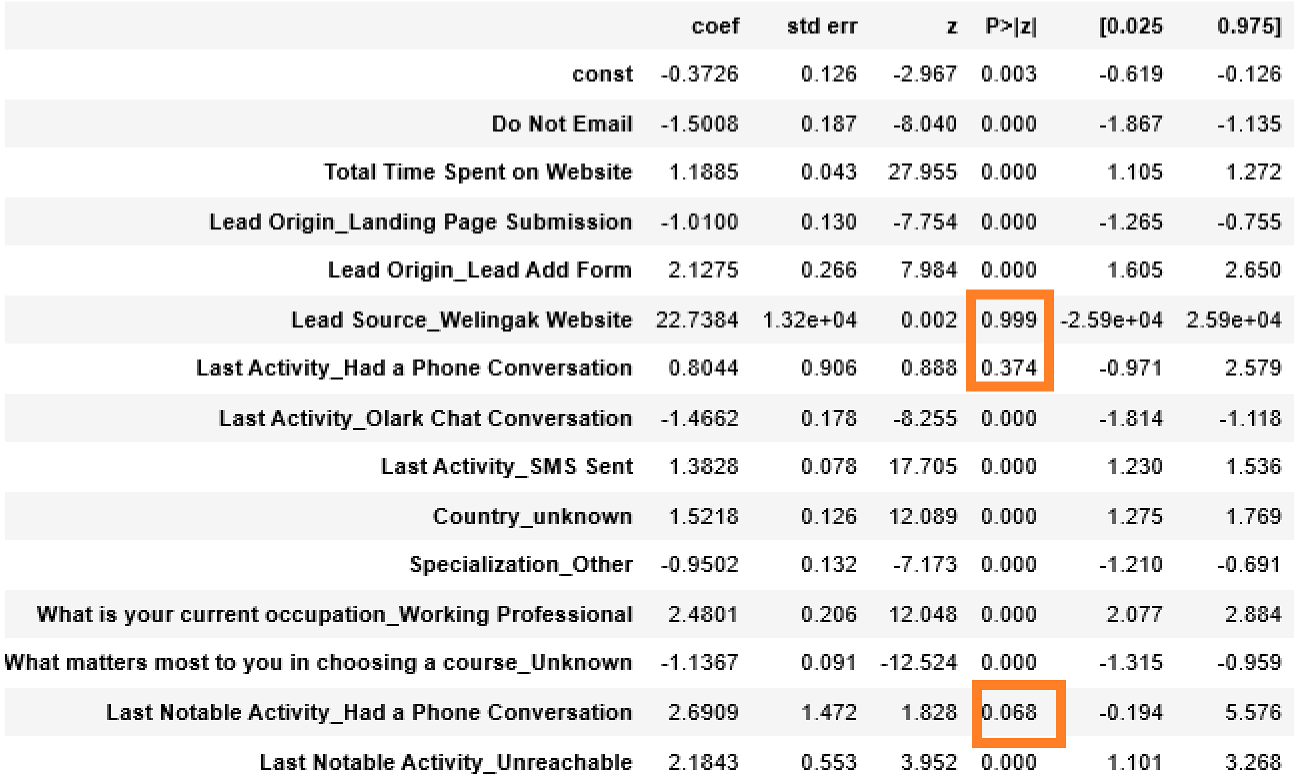
Modelling :Step 1

Creating Dummy Variables and using RFE to select 15 features:

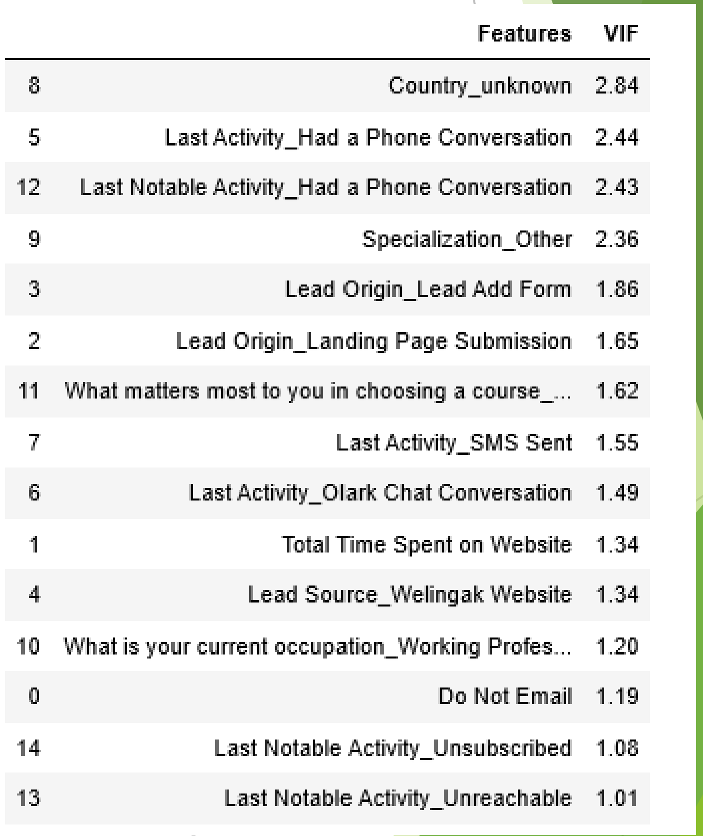
- Do Not Email,
- Total Time Spent on Website
- Lead Origin_Landing Page Submission
- Lead Origin_Lead Add Form,
- Lead Source_Welingak Website,
- Last Activity_Had a Phone Conversation,
- Last Activity_Olark Chat Conversation
- Last Activity_SMS Sent
- Country_unknown
- Specialization_Other
- What is your current occupation_Working Professional
- What matters most to you in choosing a course_Unknown
- Last Notable Activity_Had a Phone Conversation
- Last Notable Activity_Unreachable
- Last Notable Activity_Unsubscribed

Modelling: Step 2

Running model iteratively looking for P value > 0.05 significance and VIF more than 5 to drop correlated columns making coefficients swing or unstable

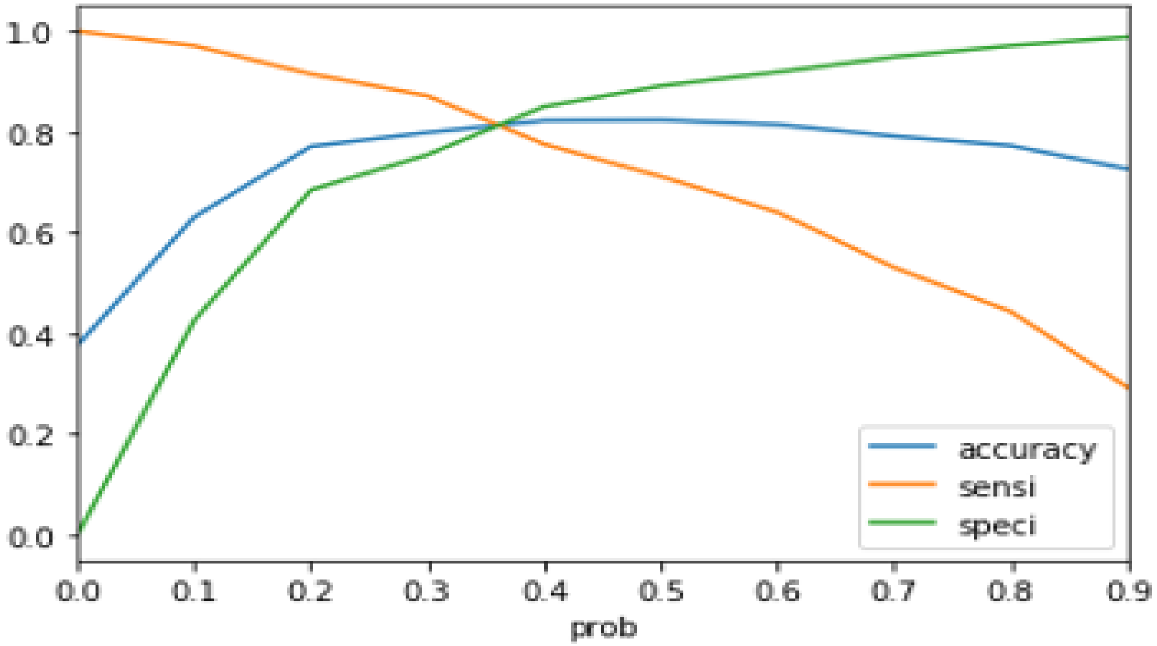


	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3726	0.126	-2.967	0.003	-0.619	-0.126
Do Not Email	-1.5008	0.187	-8.040	0.000	-1.867	-1.135
Total Time Spent on Website	1.1885	0.043	27.955	0.000	1.105	1.272
Lead Origin_Landing Page Submission	-1.0100	0.130	-7.754	0.000	-1.265	-0.755
Lead Origin_Lead Add Form	2.1275	0.266	7.984	0.000	1.605	2.650
Lead Source_Welingak Website	22.7384	1.32e+04	0.002	0.999	-2.59e+04	2.59e+04
Last Activity_Had a Phone Conversation	0.8044	0.906	0.888	0.374	-0.971	2.579
Last Activity_Olark Chat Conversation	-1.4662	0.178	-8.255	0.000	-1.814	-1.118
Last Activity_SMS Sent	1.3828	0.078	17.705	0.000	1.230	1.536
Country_unknown	1.5218	0.126	12.089	0.000	1.275	1.769
Specialization_Other	-0.9502	0.132	-7.173	0.000	-1.210	-0.691
What is your current occupation_Working Professional	2.4801	0.206	12.048	0.000	2.077	2.884
What matters most to you in choosing a course_Unknown	-1.1367	0.091	-12.524	0.000	-1.315	-0.959
Last Notable Activity_Had a Phone Conversation	2.6909	1.472	1.828	0.068	-0.194	5.576
Last Notable Activity_Unreachable	2.1843	0.553	3.952	0.000	1.101	3.268

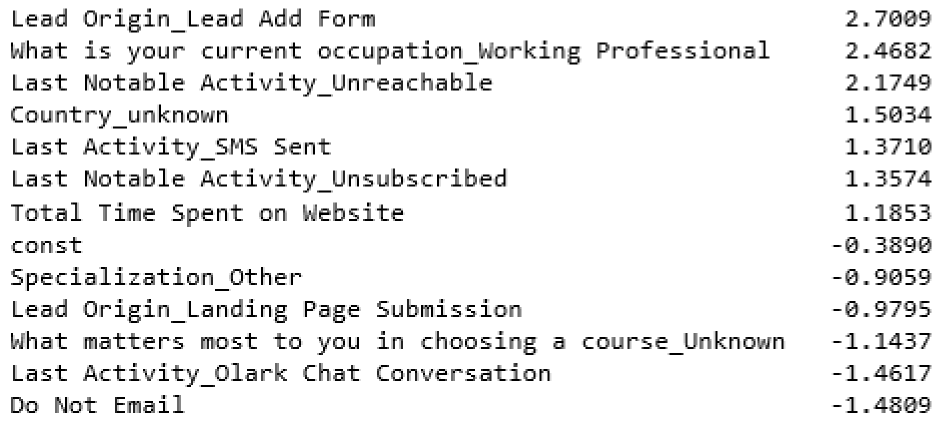


	Features	VIF
8	Country_unknown	2.84
5	Last Activity_Had a Phone Conversation	2.44
12	Last Notable Activity_Had a Phone Conversation	2.43
9	Specialization_Other	2.36
3	Lead Origin_Lead Add Form	1.86
2	Lead Origin_Landing Page Submission	1.65
11	What matters most to you in choosing a course_...	1.62
7	Last Activity_SMS Sent	1.55
6	Last Activity_Olark Chat Conversation	1.49
1	Total Time Spent on Website	1.34
4	Lead Source_Welingak Website	1.34
10	What is your current occupation_Working Profes...	1.20
0	Do Not Email	1.19
14	Last Notable Activity_Unsubscribed	1.08
13	Last Notable Activity_Unreachable	1.01

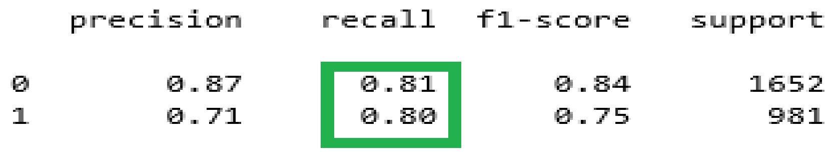
Finding threshold Probability for 80% Recall of both class 1 and 0



Final Model With 80% Recall for 1 and 0 class



```
Lead Origin_Lead Add Form                                  2.7009
What is your current occupation_Working Professional       2.4682
Last Notable Activity_Unreachable                          2.1749
Country_unknown                                            1.5034
Last Activity_SMS Sent                                     1.3710
Last Notable Activity_Unsubscribed                         1.3574
Total Time Spent on Website                                1.1853
const                                                     -0.3890
Specialization_Other                                      -0.9059
Lead Origin_Landing Page Submission                       -0.9795
What matters most to you in choosing a course_Unknown     -1.1437
Last Activity_Olark Chat Conversation                     -1.4617
Do Not Email                                              -1.4809
```



```
          precision    recall  f1-score   support

   0          0.87      0.81      0.84      1652
   1          0.71      0.80      0.75       981
```