Lead Score Assignment for X Education

- **Business Problem:** X Education sells online courses to industry professionals. Everyday many professionals who are interested in the courses land on their website and browse for courses. These users are classified as potential leads to take the course. Currently the conversion rate is very poor. The company wishes to identify the most potential leads, also known as 'Hot Leads'. With the Hot Leads, the Sales team will focus more on communicating with the them rather than making calls to everyone.

- **Dataset Description:** The Dataset provided for this problem has 9240 rows and 37 columns. Among all these columns only 5 are numerical and rest are categorical. Reading the description of all the columns we understood that all the columns are not important. So we started some data visualization to understand the pattern of the features of the data and based on that as well as applying few intuition we removed few columns.

- **Data Preparation:** Following columns have been removed as those were ubique identification numbers for the instances:

  1. Prospect ID
  2. Lead Number.

- We created few categories for the missing values for all the columns based on the percentage of missing values present in each column, Thus we got below findings:

  1. Number of columns having less than or equal to 20% -> 4
  2. Number of columns having null value from 20 to 40% -> 6
  3. Number of columns having null value from 40 to 50% -> 4
  4. Number of columns having null value from 50 to 60% -> 1
  5. Number of columns having null value from 60 to 70% -> 6
  6. Number of columns having null value from 70 to 80% -> 2
  7. Number of columns having more than 80% null values-> 0

- We dropped off all those columns which have more than 40% null values.

- There are few columns which have a value called 'Select'. So, we assumed here that whenever filling any online application form, users didn't like to reveal some information for some non-mandatory fields. As a result, 'Select' is appearing in those fields. We changed those values as null values.

- **Null Imputation:** We can use different methods for null imputation.

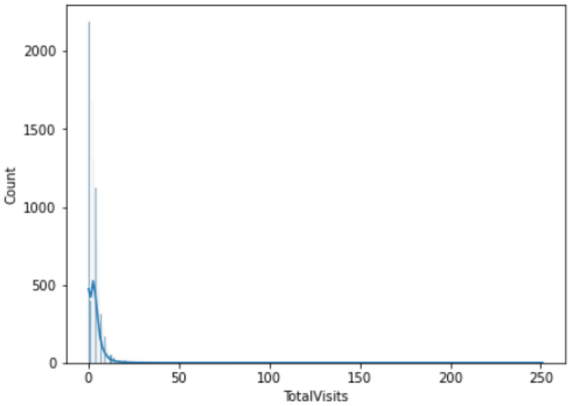
  1. We can distribute all the missing values from an attribute equally among all the categories of that attribute.
  2. We can impute the null values using mode or median based on the nature of the column.
  3. We can group the null values and rename those and treat as a separate category all together. This way had been followed for **'Specialization'** attribute. Similar approach has been used for **'Current_Occupation', 'city'** and '**Tags'** attribute.
  4. For '**country'** column we used mode for imputing nulls as it was a categorical column, but due to being highly skewed in nature we omitted this column.
  5. For '**TotalVisits'** column, which is numerical in nature, we used median for imputation.

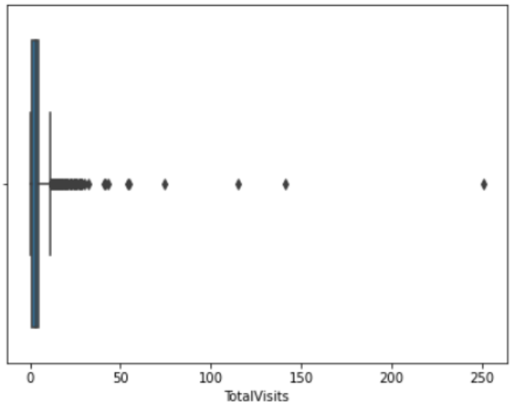
6. For the column ‘**Lead Source’** having only 36 null instances, the rows having null instances were dropped from the data. Similar approach had been taken for “Last Activity” column.

- **Data Cleaning:** For Data cleaning, we first applied our intuition and based on that few columns were dropped as those didn’t seem to be very significant or important, Also, we checked the skewness and also the presence of only unique category in some columns came into the decision of rejecting few columns. Such examples include - **‘Prospect ID’, ‘Lead Number’ ,‘Country’, ’Motivation’, ’ Through Recommendations’,’ Newspaper Article’, ‘Newspaper’,’ Digital Advertisement’,’ X Education Forums’,’ Search’,’ Get_update’,’ Get updates on DM Content’,’ Last Notable Activity’,’ Update me on Supply Chain Content’,’ Payment_Mode’,’ Do Not Call’,’ Magazine’** . Data Visualization helped a lot in taking this decision.

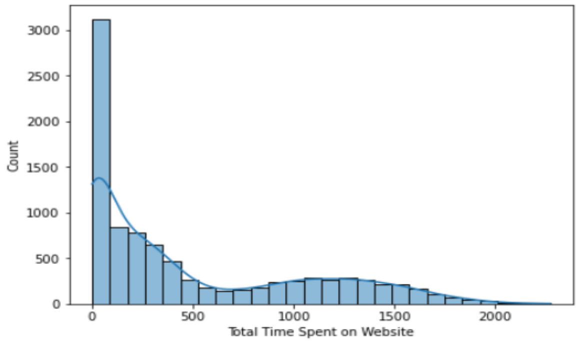
- **Data Visualization:** For **univariate** and **bi-variate** analysis we performed data visualization and exploratory data analysis(EDA) .Also identification of outliers were identified when we did distribution plot for the numerical columns.

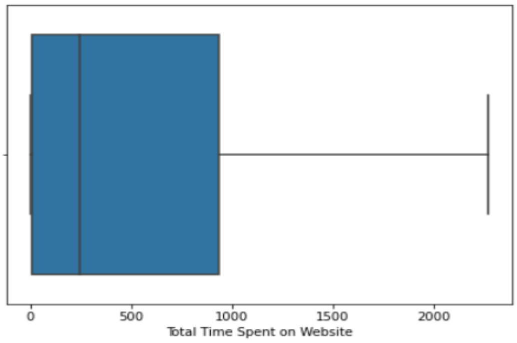
1. **“TotalVisits**” column has skewed distribution and presence of outliers were noticed in the attribute. The column is a numerical column.



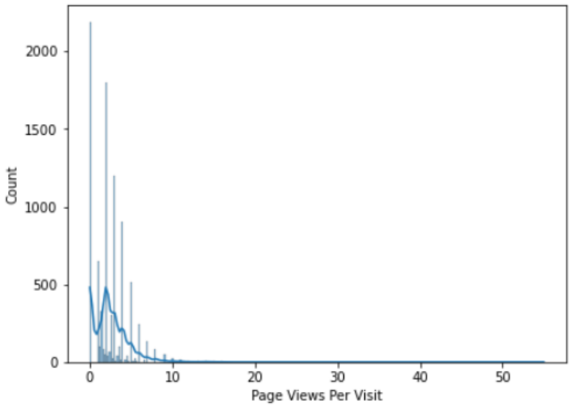


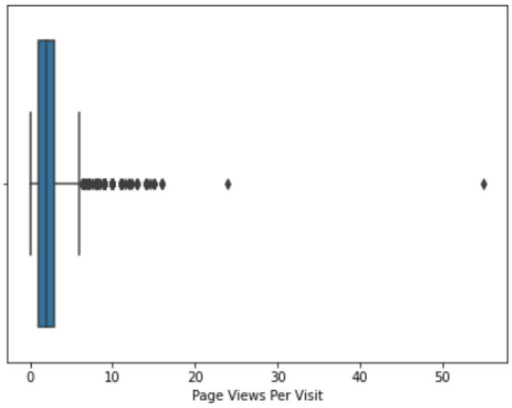
2. **“Total Time Spent on Website”** column also has skewed distribution but there was no outlier in the feature.





3. **“Page Views Per Visit’** has a skewed distribution and presence of outliers were noticed there.

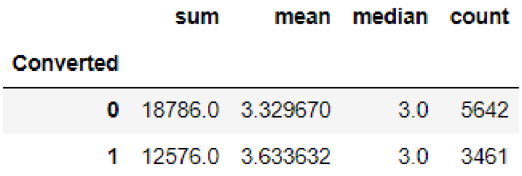




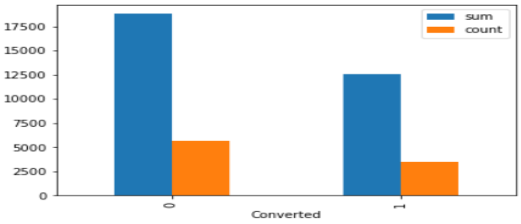
4. **“Converted”** is the target column where we found that 61.46% is 0 i.e, ‘not-converted’ and 38.53% is 1 i.e, ‘converted’. Slight Data imbalance is found but it’s not too high. If there would have been high data imbalance then we would have used some balancing technique as: ‘Upsampling’, ‘Downsampling’ or ‘SMOTE’.

5. Performed several groupby operation as a part of **bi-variate** analysis to understand how much the target variable’s count is for each category of any feature. This was done for categorical and numerical variables.

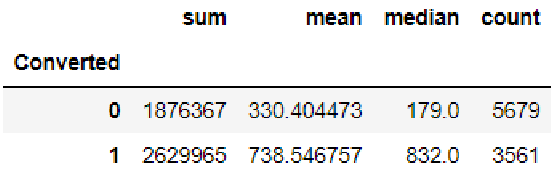
6. **“Converted” vs “TotalVisits”:**



	sum	mean	median	count
Converted				
0	18786.0	3.329670	3.0	5642
1	12576.0	3.633632	3.0	3461

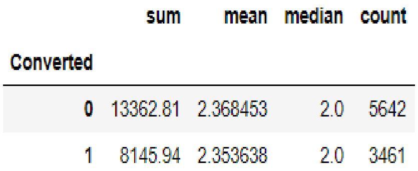


7. **“Converted” vs “Total_Time_Spent”:**

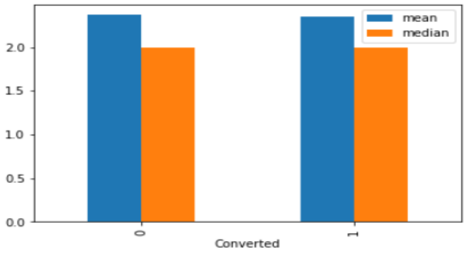


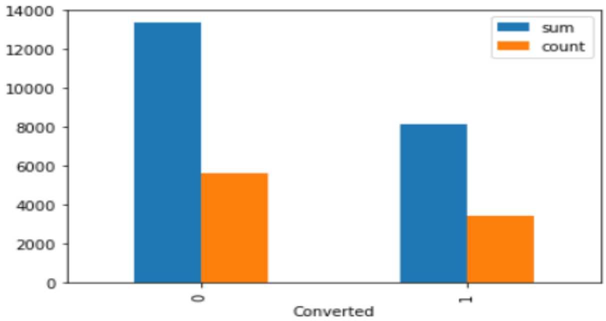
	sum	mean	median	count
Converted				
0	1876367	330.404473	179.0	5679
1	2629965	738.546757	832.0	3561

8. **“Converted” vs “Page Views Per Visit”:**



	sum	mean	median	count
Converted				
0	13362.81	2.368453	2.0	5642
1	8145.94	2.353638	2.0	3461

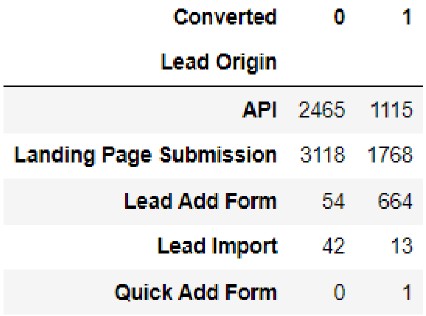




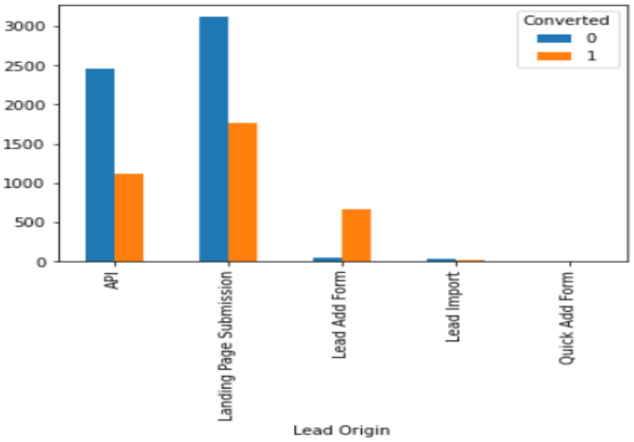
Inference:

- The Converted leads have less count for total visit then non-converted one.
- The average time spent visited on website by converted leads are higher than that of non-converted.
- The total number of pages viewed in every visit by the non-converted leads is higher than that of converted leads.

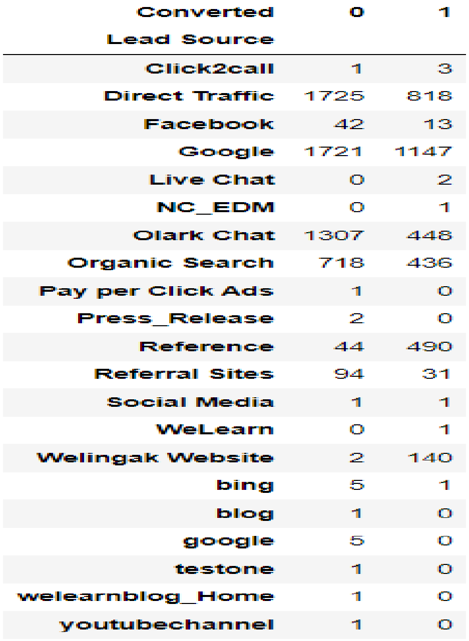
9. “Lead Origin” vs “Converted”:



Converted	0	1
Lead Origin		
API	2465	1115
Landing Page Submission	3118	1768
Lead Add Form	54	664
Lead Import	42	13
Quick Add Form	0	1

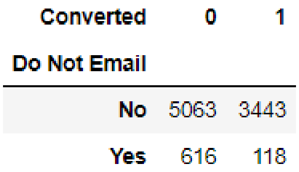


10. “Lead Source” vs “Converted”:



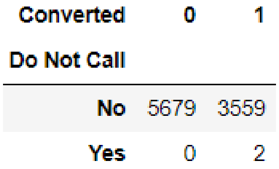
Converted	0	1
Lead Source		
Click2call	1	3
Direct Traffic	1725	818
Facebook	42	13
Google	1721	1147
Live Chat	0	2
NC_EDM	0	1
Olark Chat	1307	448
Organic Search	718	436
Pay per Click Ads	1	0
Press_Release	2	0
Reference	44	490
Referral Sites	94	31
Social Media	1	1
WeLearn	0	1
Welingak Website	2	140
bing	5	1
blog	1	0
google	5	0
testone	1	0
welearnblog_Home	1	0
youtubechannel	1	0

11. “Do Not Email” vs “Converted”:



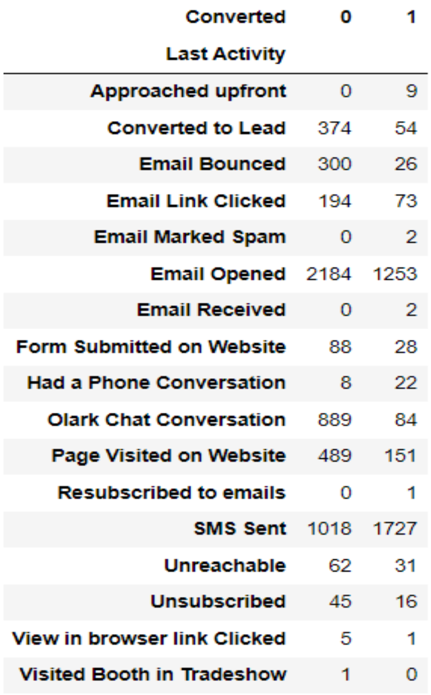
Converted	0	1
Do Not Email		
No	5063	3443
Yes	616	118

12. “Do Not Call” vs “Converted”:

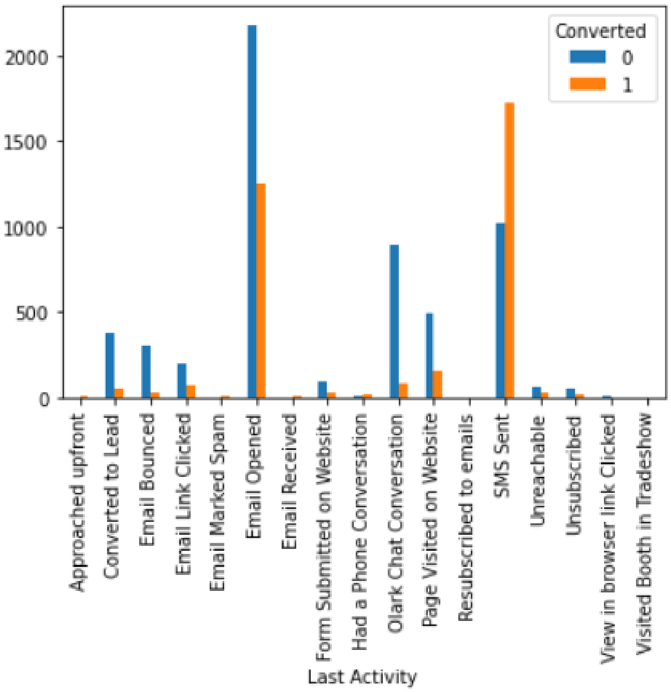


Converted	0	1
Do Not Call		
No	5679	3559
Yes	0	2

13. “Last Activity” vs “Converted”:



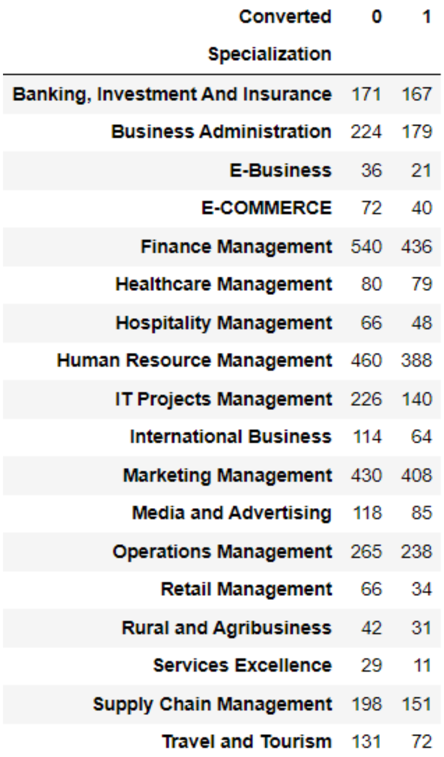
Converted	0	1
Last Activity		
Approached upfront	0	9
Converted to Lead	374	54
Email Bounced	300	26
Email Link Clicked	194	73
Email Marked Spam	0	2
Email Opened	2184	1253
Email Received	0	2
Form Submitted on Website	88	28
Had a Phone Conversation	8	22
Olark Chat Conversation	889	84
Page Visited on Website	489	151
Resubscribed to emails	0	1
SMS Sent	1018	1727
Unreachable	62	31
Unsubscribed	45	16
View in browser link Clicked	5	1
Visited Booth in Tradeshow	1	0



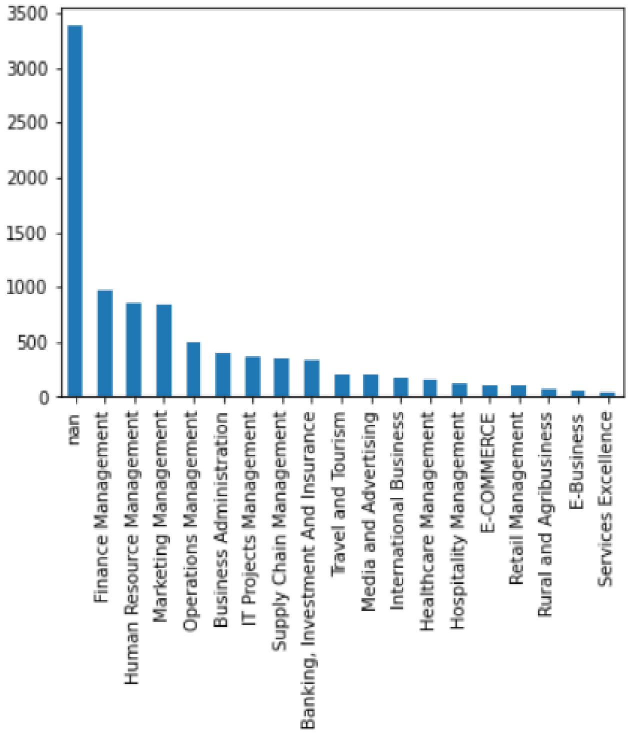
Inference:

- The leads who have opened email has high percentage of non-conversion
- Leads with SMS Activity has high percentage of conversion among all.
- ‘Email Opened’ also has high percentage of conversion. Overall count of “Email Opened” and “SMS Sent” are pretty high.

14. “Specialization” vs “Converted”:

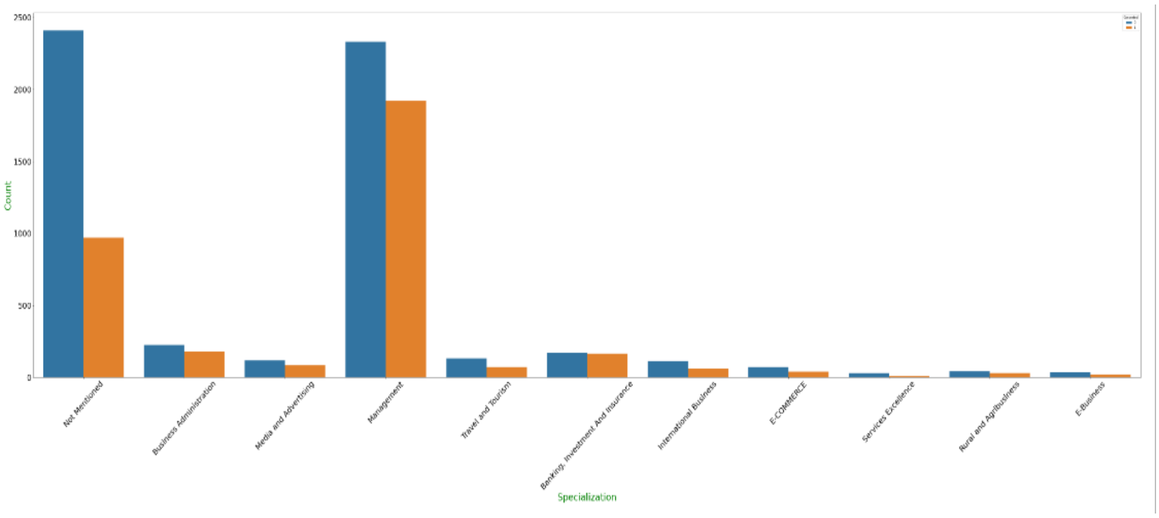


Converted	0	1
Specialization		
Banking, Investment And Insurance	171	167
Business Administration	224	179
E-Business	36	21
E-COMMERCE	72	40
Finance Management	540	436
Healthcare Management	80	79
Hospitality Management	66	48
Human Resource Management	460	388
IT Projects Management	226	140
International Business	114	64
Marketing Management	430	408
Media and Advertising	118	85
Operations Management	265	238
Retail Management	66	34
Rural and Agribusiness	42	31
Services Excellence	29	11
Supply Chain Management	198	151
Travel and Tourism	131	72

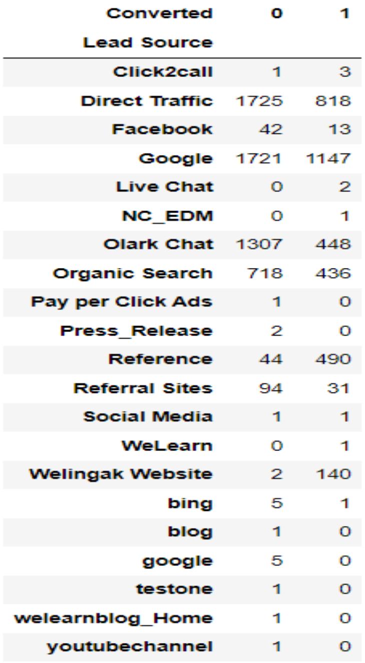


In “Specialization”, we can see many Management specialization which we have categorized into same group “Management”.

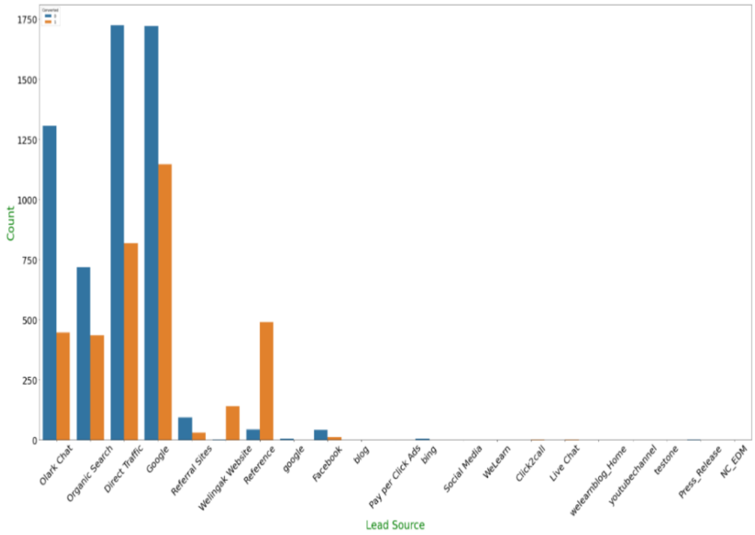
We can see that, in ‘Specialization’ we have many null values. All those null values have been renamed into “Not Mentioned” category.



15. “Lead Source” vs “Converted”:



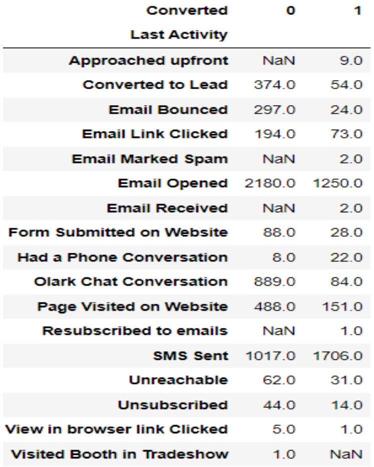
Converted	0	1
Lead Source		
Click2call	1	3
Direct Traffic	1725	818
Facebook	42	13
Google	1721	1147
Live Chat	0	2
NC_EDM	0	1
Olark Chat	1307	448
Organic Search	718	436
Pay per Click Ads	1	0
Press_Release	2	0
Reference	44	490
Referral Sites	94	31
Social Media	1	1
WeLearn	0	1
Welingak Website	2	140
bing	5	1
blog	1	0
google	5	0
testone	1	0
welearnblog_Home	1	0
youtubechannel	1	0



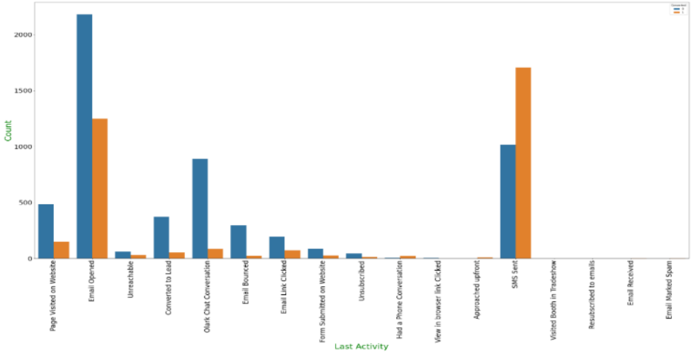
Inference:

- We can see that maximum leads have come from 'Olark Chat', 'Organic Search', 'Direct Traffic' and 'Google'. Very few from 'Referral Sites','Welingak Website', 'Reference' and 'Facebook'. Rest are very negligible.
- we can notice here 'Google' and 'google' - these same things are present but with different spellings. So we can convert these into same.
- Reference and Welingak Website - these two sources have higher conversion chance
- We saw that there are two categories ‘google’ and ‘Google’ both types of categories. So we made these same.
- We saw that there are 4 categories ‘blog’, ‘testone’, ‘welearnblog_Home’, ‘youtubechannel’ which have total 4 number of instances overall, So dropped off these instances.

16. “Last Activity” vs “Converted”:

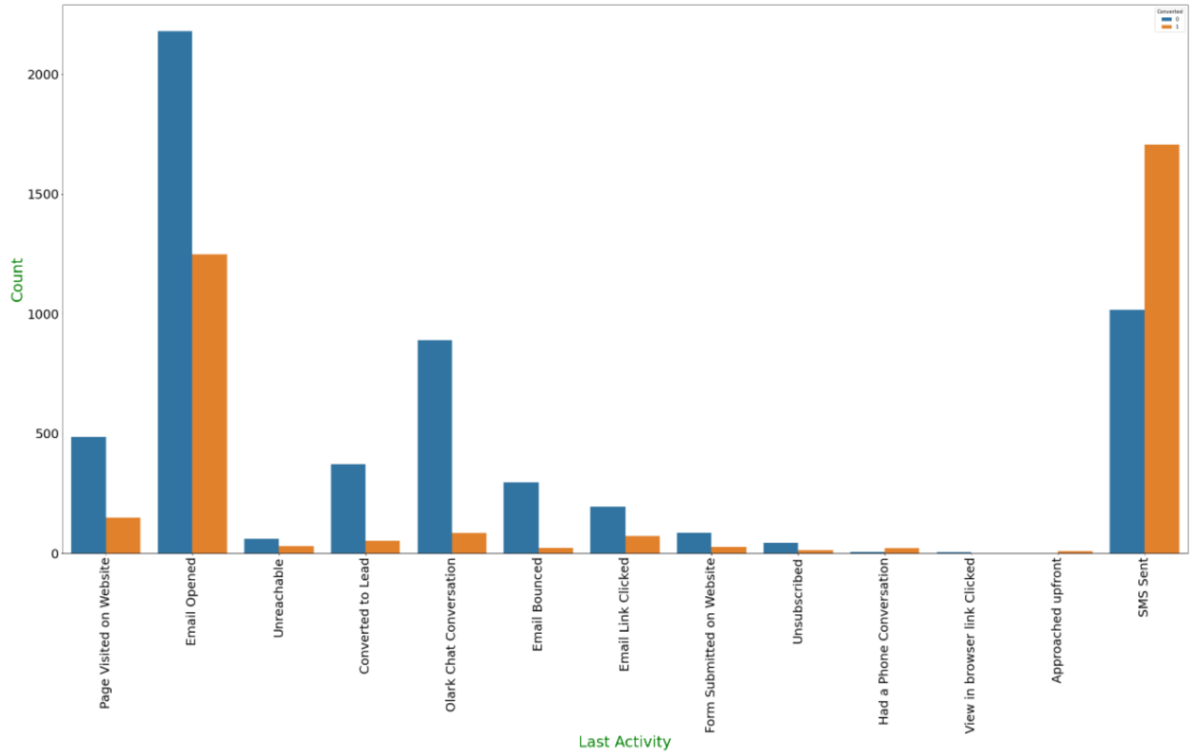


Converted	0	1
Last Activity		
Approached upfront	NaN	9.0
Converted to Lead	374.0	54.0
Email Bounced	297.0	24.0
Email Link Clicked	194.0	73.0
Email Marked Spam	NaN	2.0
Email Opened	2180.0	1250.0
Email Received	NaN	2.0
Form Submitted on Website	88.0	28.0
Had a Phone Conversation	8.0	22.0
Olark Chat Conversation	889.0	84.0
Page Visited on Website	488.0	151.0
Resubscribed to emails	NaN	1.0
SMS Sent	1017.0	1706.0
Unreachable	62.0	31.0
Unsubscribed	44.0	14.0
View in browser link Clicked	5.0	1.0
Visited Booth in Tradeshow	1.0	NaN

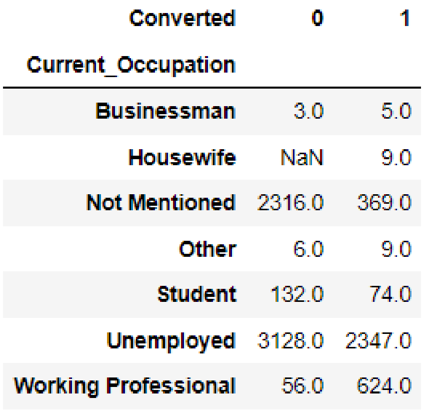


We saw that there are 6 categories of Last Activity have very less count. So deleted those instances.

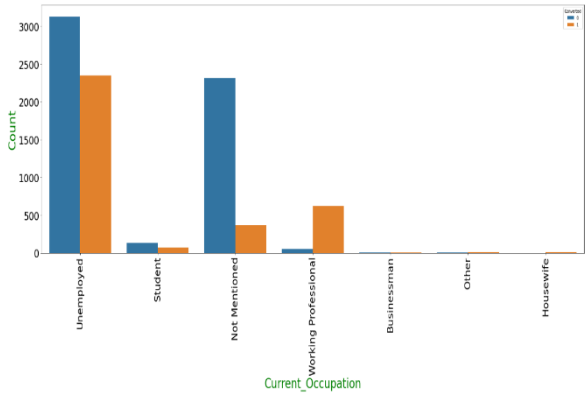
After deleting those, the plot looked like below:



17. **“Current_Occupation” vs “Convereted”:** All null values were categorized under a single name “Unempolyed”.



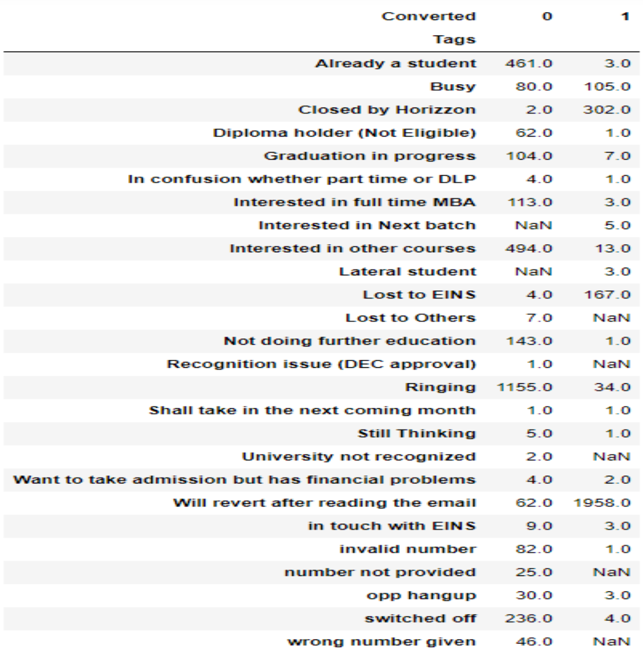
Converted	0	1
Current_Occupation		
Businessman	3.0	5.0
Housewife	NaN	9.0
Not Mentioned	2316.0	369.0
Other	6.0	9.0
Student	132.0	74.0
Unemployed	3128.0	2347.0
Working Professional	56.0	624.0



Also, some other categories like: “Businessman”, “Other” and “Housewife” were grouped together and made a new category “Other” which was used further.

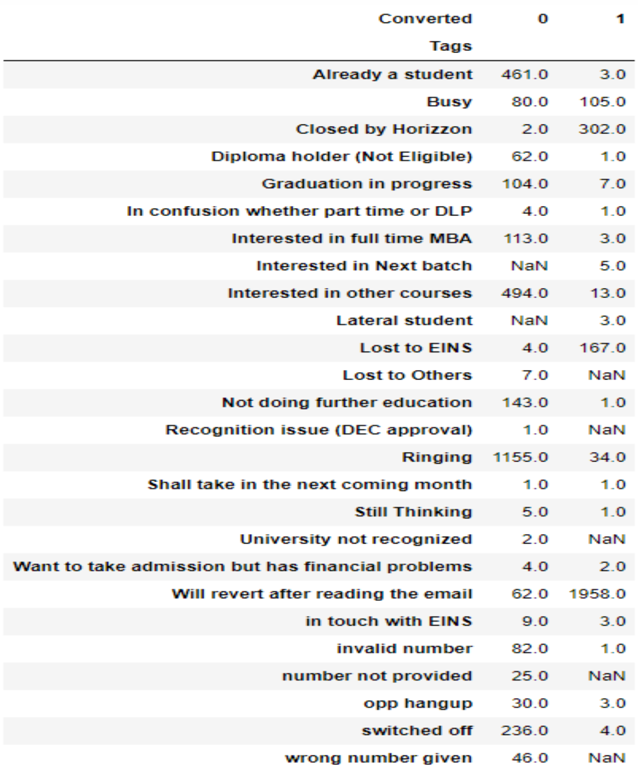
18. **“Tags” vs “Converted”:**

The unique value counts of this attribute were as follows:



Converted	0	1
Tags		
Already a student	461.0	3.0
Busy	80.0	105.0
Closed by Horizzon	2.0	302.0
Diploma holder (Not Eligible)	62.0	1.0
Graduation in progress	104.0	7.0
In confusion whether part time or DLP	4.0	1.0
Interested in full time MBA	113.0	3.0
Interested in Next batch	NaN	5.0
Interested in other courses	494.0	13.0
Lateral student	NaN	3.0
Lost to EINS	4.0	167.0
Lost to Others	7.0	NaN
Not doing further education	143.0	1.0
Recognition issue (DEC approval)	1.0	NaN
Ringing	1155.0	34.0
Shall take in the next coming month	1.0	1.0
Still Thinking	5.0	1.0
University not recognized	2.0	NaN
Want to take admission but has financial problems	4.0	2.0
Will revert after reading the email	62.0	1958.0
in touch with EINS	9.0	3.0
invalid number	82.0	1.0
number not provided	25.0	NaN
opp hangup	30.0	3.0
switched off	236.0	4.0
wrong number given	46.0	NaN

The category wise count for this attribute was as follows:

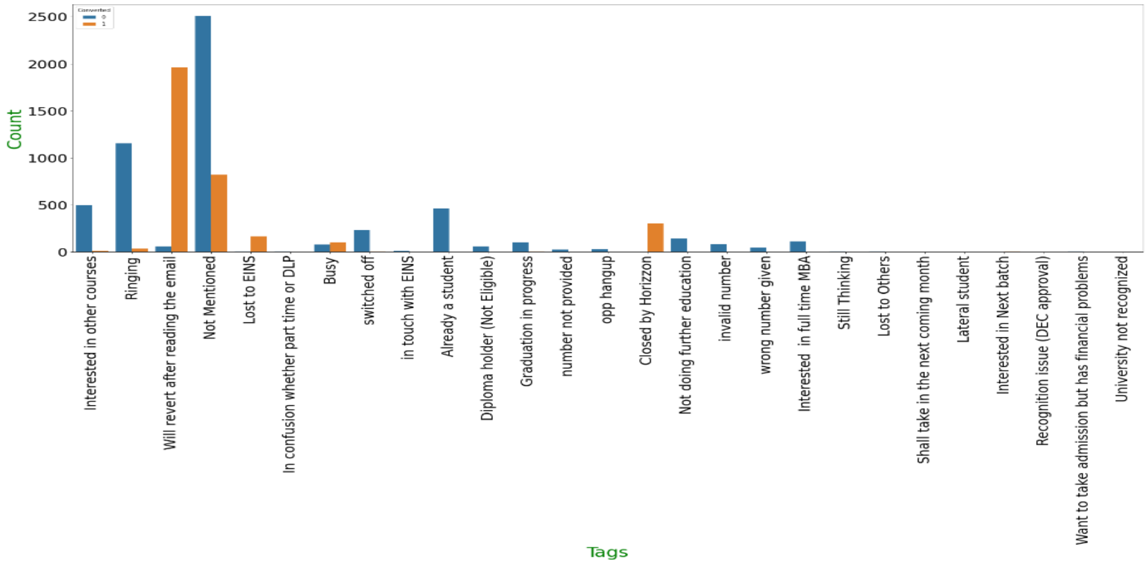


Converted	0	1
Tags		
Already a student	461.0	3.0
Busy	80.0	105.0
Closed by Horizzon	2.0	302.0
Diploma holder (Not Eligible)	62.0	1.0
Graduation in progress	104.0	7.0
In confusion whether part time or DLP	4.0	1.0
Interested in full time MBA	113.0	3.0
Interested in Next batch	NaN	5.0
Interested in other courses	494.0	13.0
Lateral student	NaN	3.0
Lost to EINS	4.0	167.0
Lost to Others	7.0	NaN
Not doing further education	143.0	1.0
Recognition issue (DEC approval)	1.0	NaN
Ringing	1155.0	34.0
Shall take in the next coming month	1.0	1.0
Still Thinking	5.0	1.0
University not recognized	2.0	NaN
Want to take admission but has financial problems	4.0	2.0
Will revert after reading the email	62.0	1958.0
in touch with EINS	9.0	3.0
invalid number	82.0	1.0
number not provided	25.0	NaN
opp hangup	30.0	3.0
switched off	236.0	4.0
wrong number given	46.0	NaN

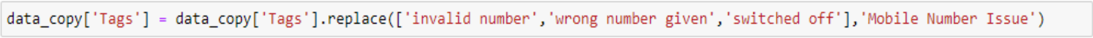
After the initial study of the attribute the inferences were as below:

- We can say that for many leads , the conversion didn't happen and they are just not interested to take the call , as a result the 'Ringing' category count is so high approximately 1155.
- for 236 non- converted leads it has been found that their phone is switched off.
- 1960 leads have converted who have reverted after reading the mail.
- 496 leads are interested in other courses.
- 462 leads are already student - as a result the didn't turn up.

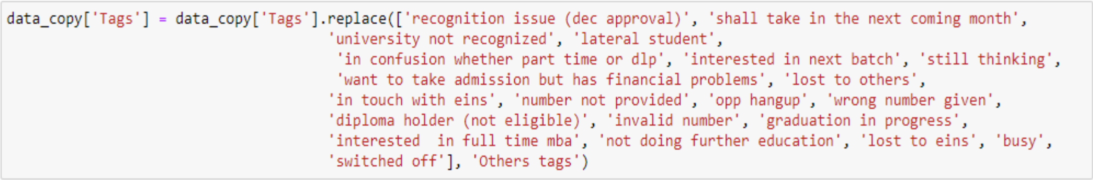
The groupby operation of this attribute for target variable gave us a picture like below:



After this we performed a grouping operation again for this attribute which is as below:



```
data_copy['Tags'] = data_copy['Tags'].replace(['invalid number','wrong number given','switched off'],'Mobile Number Issue')
```



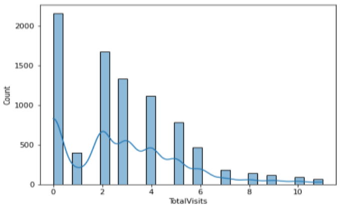
```
data_copy['Tags'] = data_copy['Tags'].replace(['recognition issue (dec approval)', 'shall take in the next coming month',
                                       'university not recognized', 'lateral student',
                                        'in confusion whether part time or dlp', 'interested in next batch', 'still thinking',
                                        'want to take admission but has financial problems', 'lost to others',
                                       'in touch with eins', 'number not provided', 'opp hangup', 'wrong number given',
                                       'diploma holder (not eligible)', 'invalid number', 'graduation in progress',
                                       'interested  in full time mba', 'not doing further education', 'lost to eins', 'busy',
                                       'switched off'], 'Others tags')
```

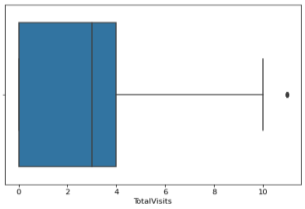
- **Outlier Detection:**
  We tried to detect the outliers of the numerical variables. The numerical variables for which we identified outliers and imputed are:
  a) **"TotalVisits"** - It has total 266 outliers which actually contributed to almost 3% of the full data.
  b) **"Page Views Per Visit"** - It has total 293 outliers which gives 3.32% of the whole data.
  c) **"Total_Time_Spent"** – It has total 2 outliers which is not even contributing to 1% of whole data.

  As the total number of outliers are present in very less percentage, so we dropped those rows which actually have these outliers. Actually, we consider a threshold value, usually in maximum cases this threshold is 5%. If the percentage of outliers are below this threshold, then we can delete the rows, which are actually contributing to the existence of these outliers and if the percentage of outliers is higher than the threshold value then we can impute those by capping or can adopt some other transformation process.

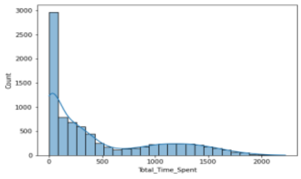
  Below are the visualizations of the attributes after outlier imputation:

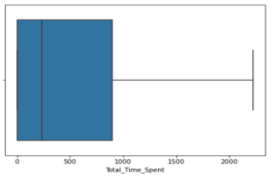
```
<AxesSubplot:xlabel='TotalVisits'>
```



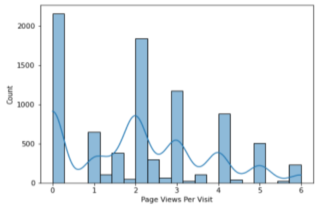


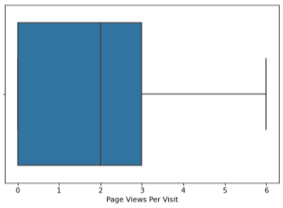
```
<AxesSubplot:xlabel='Total_Time_Spent'>
```



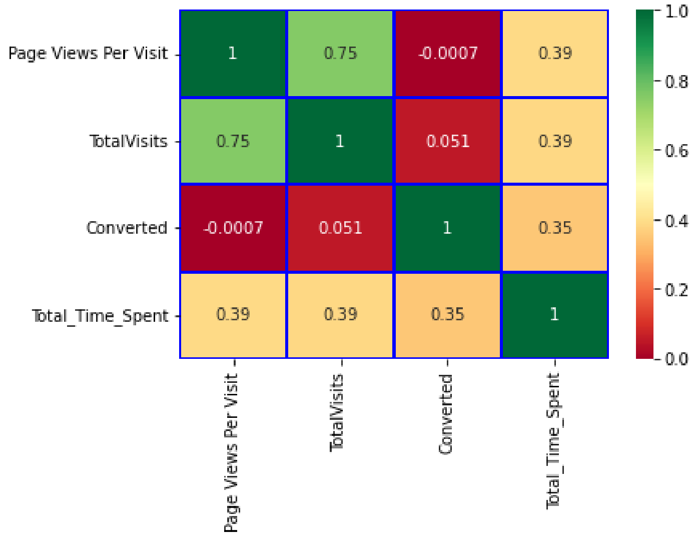


```
<AxesSubplot:xlabel='Page Views Per Visit'>
```

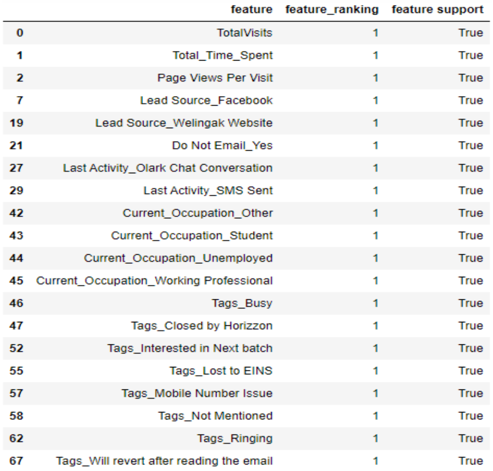




- **Checking Correlation:**
  In the dataset we checked the correlation among all the numerical columns. From there we found that total independent variables “TotalVisits” and “Page Views Per Visit” are multicollinear and that coefficient is 0.75. The target variable “Converted” has correlation coefficient 0.35 with “Total_Time_Spent” variable. The figure is given below:

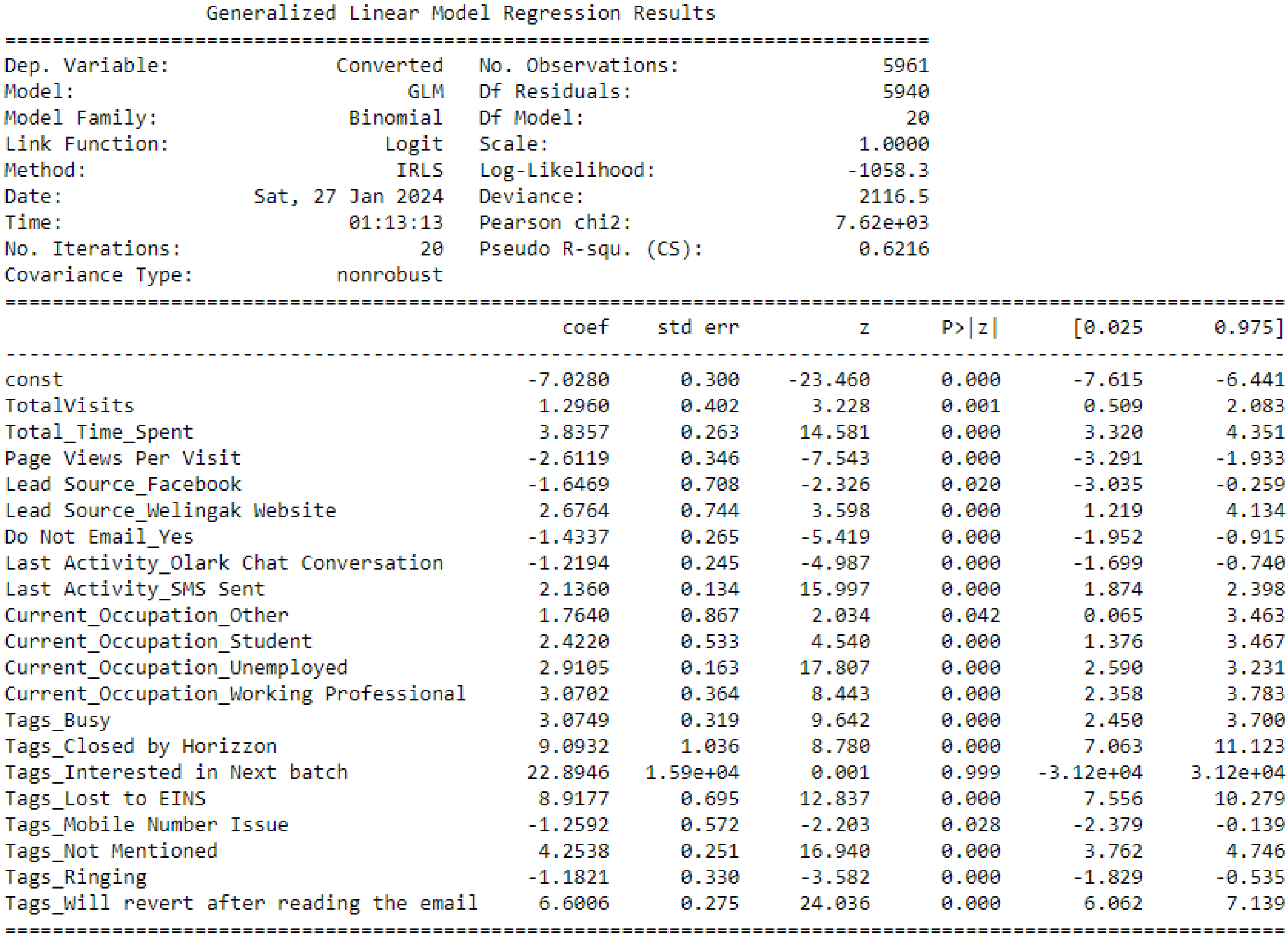


- **Dummy Variable Creation:**
  We created dummy variables for all the categorical variables. After that the size of the whole dataset became 8517 x 71 i.e, 8517 rows and 71 columns.
- **Train-Test Split:**
  The whole dataset was divided into two parts – 70% was Train set and 30% became Test Set.
  From their we dropped the target variable and in that way we made y_train and y_test . X_train and X_test were also made.
- **Scaling of Numerical Data:**
  We scaled the numerical columns of Train set (X_train) and we applied min max scaler as it will bring all the variables between 0 and 1. So in that way all the variables including categorical variables became between 0 and 1.
- **RFE (Recursive Feature Elimination):**
  Before starting modelling we checked that total number of variables till now is 70 which is too high. So, we ran RFE to understand which features are actually important and the total number of required variables were set to 20. Though Specialization column seemed to be important but for RFE , the ‘Specialization’ related dummy variables were eliminated in this process. So , total 20 features, which we found to be most important are as follows:



	feature	feature_ranking	feature support
0	TotalVisits	1	True
1	Total_Time_Spent	1	True
2	Page Views Per Visit	1	True
7	Lead Source_Facebook	1	True
19	Lead Source_Welingak Website	1	True
21	Do Not Email_Yes	1	True
27	Last Activity_Olark Chat Conversation	1	True
29	Last Activity_SMS Sent	1	True
42	Current_Occupation_Other	1	True
43	Current_Occupation_Student	1	True
44	Current_Occupation_Unemployed	1	True
45	Current_Occupation_Working Professional	1	True
46	Tags_Busy	1	True
47	Tags_Closed by Horizzon	1	True
52	Tags_Interested in Next batch	1	True
55	Tags_Lost to EINS	1	True
57	Tags_Mobile Number Issue	1	True
58	Tags_Not Mentioned	1	True
62	Tags_Ringing	1	True
67	Tags_Will revert after reading the email	1	True

- **Modeling:**
  As this is a classification problem, we used Logistic Regression. We made total 4 models and in each step we reduced few variables as the p-value of few variables seemed to be high i.e, greater than 0.05.
  - Model 1:



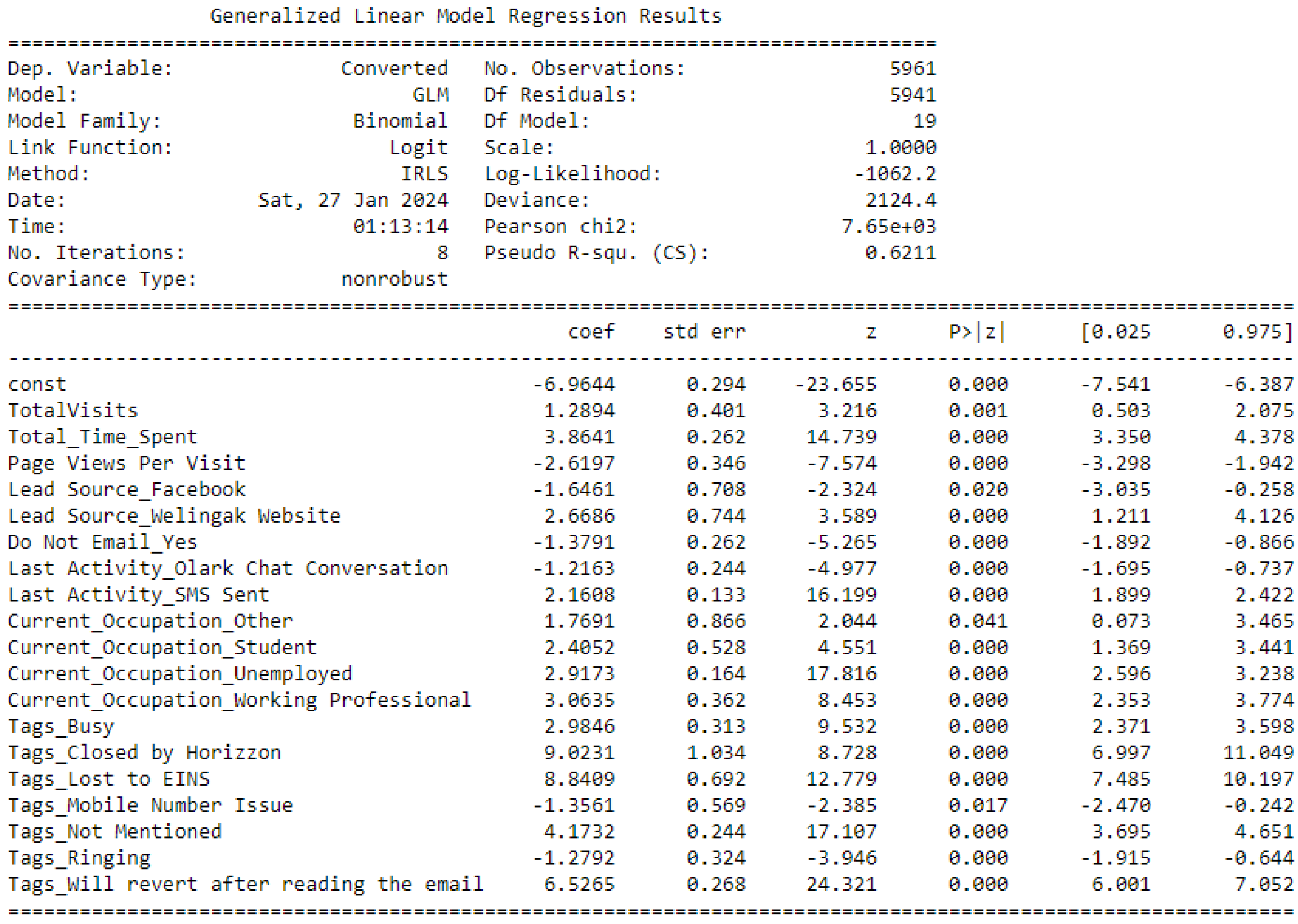
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 5961
Model:                            GLM   Df Residuals:                     5940
Model Family:                Binomial   Df Model:                           20
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1058.3
Date:                Sat, 27 Jan 2024   Deviance:                       2116.5
Time:                        01:13:13   Pearson chi2:                 7.62e+03
No. Iterations:                    20   Pseudo R-squ. (CS):             0.6216
Covariance Type:            nonrobust
===========================================================================================================
                                              coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------
const                                      -7.0280      0.300    -23.460      0.000      -7.615      -6.441
TotalVisits                                 1.2960      0.402      3.228      0.001       0.509       2.083
Total_Time_Spent                            3.8357      0.263     14.581      0.000       3.320       4.351
Page Views Per Visit                       -2.6119      0.346     -7.543      0.000      -3.291      -1.933
Lead Source_Facebook                       -1.6469      0.708     -2.326      0.020      -3.035      -0.259
Lead Source_Welingak Website                2.6764      0.744      3.598      0.000       1.219       4.134
Do Not Email_Yes                           -1.4337      0.265     -5.419      0.000      -1.952      -0.915
Last Activity_Olark Chat Conversation      -1.2194      0.245     -4.987      0.000      -1.699      -0.740
Last Activity_SMS Sent                      2.1360      0.134     15.997      0.000       1.874       2.398
Current_Occupation_Other                    1.7640      0.867      2.034      0.042       0.065       3.463
Current_Occupation_Student                  2.4220      0.533      4.540      0.000       1.376       3.467
Current_Occupation_Unemployed               2.9105      0.163     17.807      0.000       2.590       3.231
Current_Occupation_Working Professional     3.0702      0.364      8.443      0.000       2.358       3.783
Tags_Busy                                   3.0749      0.319      9.642      0.000       2.450       3.700
Tags_Closed by Horizzon                     9.0932      1.036      8.780      0.000       7.063      11.123
Tags_Interested in Next batch              22.8946   1.59e+04      0.001      0.999   -3.12e+04    3.12e+04
Tags_Lost to EINS                           8.9177      0.695     12.837      0.000       7.556      10.279
Tags_Mobile Number Issue                   -1.2592      0.572     -2.203      0.028      -2.379      -0.139
Tags_Not Mentioned                          4.2538      0.251     16.940      0.000       3.762       4.746
Tags_Ringing                               -1.1821      0.330     -3.582      0.000      -1.829      -0.535
Tags_Will revert after reading the email    6.6006      0.275     24.036      0.000       6.062       7.139
===========================================================================================================
```

- We can see that few variables have high p-value(>0.05) So, all these seem to be insignificant variable for the model.

Interpretation: From the model's result we understood that only one variable has higher p-value (>0.05). The variable is "Tags_Interested in Next batch".

Then we dropped that variable and rebuilt the model .The result showed that all the variables are significant as all the variables' p-values are less than or equal to 0.05. The image is below:

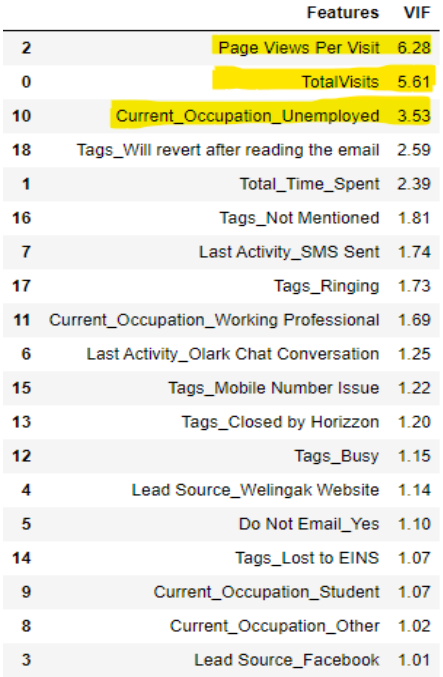
- Model 2:



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 5961
Model:                            GLM   Df Residuals:                     5941
Model Family:                Binomial   Df Model:                           19
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1062.2
Date:                Sat, 27 Jan 2024   Deviance:                       2124.4
Time:                        01:13:14   Pearson chi2:                 7.65e+03
No. Iterations:                     8   Pseudo R-squ. (CS):             0.6211
Covariance Type:            nonrobust
===========================================================================================================
                                              coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------
const                                      -6.9644      0.294    -23.655      0.000      -7.541      -6.387
TotalVisits                                 1.2894      0.401      3.216      0.001       0.503       2.075
Total_Time_Spent                            3.8641      0.262     14.739      0.000       3.350       4.378
Page Views Per Visit                       -2.6197      0.346     -7.574      0.000      -3.298      -1.942
Lead Source_Facebook                       -1.6461      0.708     -2.324      0.020      -3.035      -0.258
Lead Source_Welingak Website                2.6686      0.744      3.589      0.000       1.211       4.126
Do Not Email_Yes                           -1.3791      0.262     -5.265      0.000      -1.892      -0.866
Last Activity_Olark Chat Conversation      -1.2163      0.244     -4.977      0.000      -1.695      -0.737
Last Activity_SMS Sent                      2.1608      0.133     16.199      0.000       1.899       2.422
Current_Occupation_Other                    1.7691      0.866      2.044      0.041       0.073       3.465
Current_Occupation_Student                  2.4052      0.528      4.551      0.000       1.369       3.441
Current_Occupation_Unemployed               2.9173      0.164     17.816      0.000       2.596       3.238
Current_Occupation_Working Professional     3.0635      0.362      8.453      0.000       2.353       3.774
Tags_Busy                                   2.9846      0.313      9.532      0.000       2.371       3.598
Tags_Closed by Horizzon                     9.0231      1.034      8.728      0.000       6.997      11.049
Tags_Lost to EINS                           8.8409      0.692     12.779      0.000       7.485      10.197
Tags_Mobile Number Issue                   -1.3561      0.569     -2.385      0.017      -2.470      -0.242
Tags_Not Mentioned                          4.1732      0.244     17.107      0.000       3.695       4.651
Tags_Ringing                               -1.2792      0.324     -3.946      0.000      -1.915      -0.644
Tags_Will revert after reading the email    6.5265      0.268     24.321      0.000       6.001       7.052
===========================================================================================================
```

- We can see that all the variables have p-value lesser than 0.05

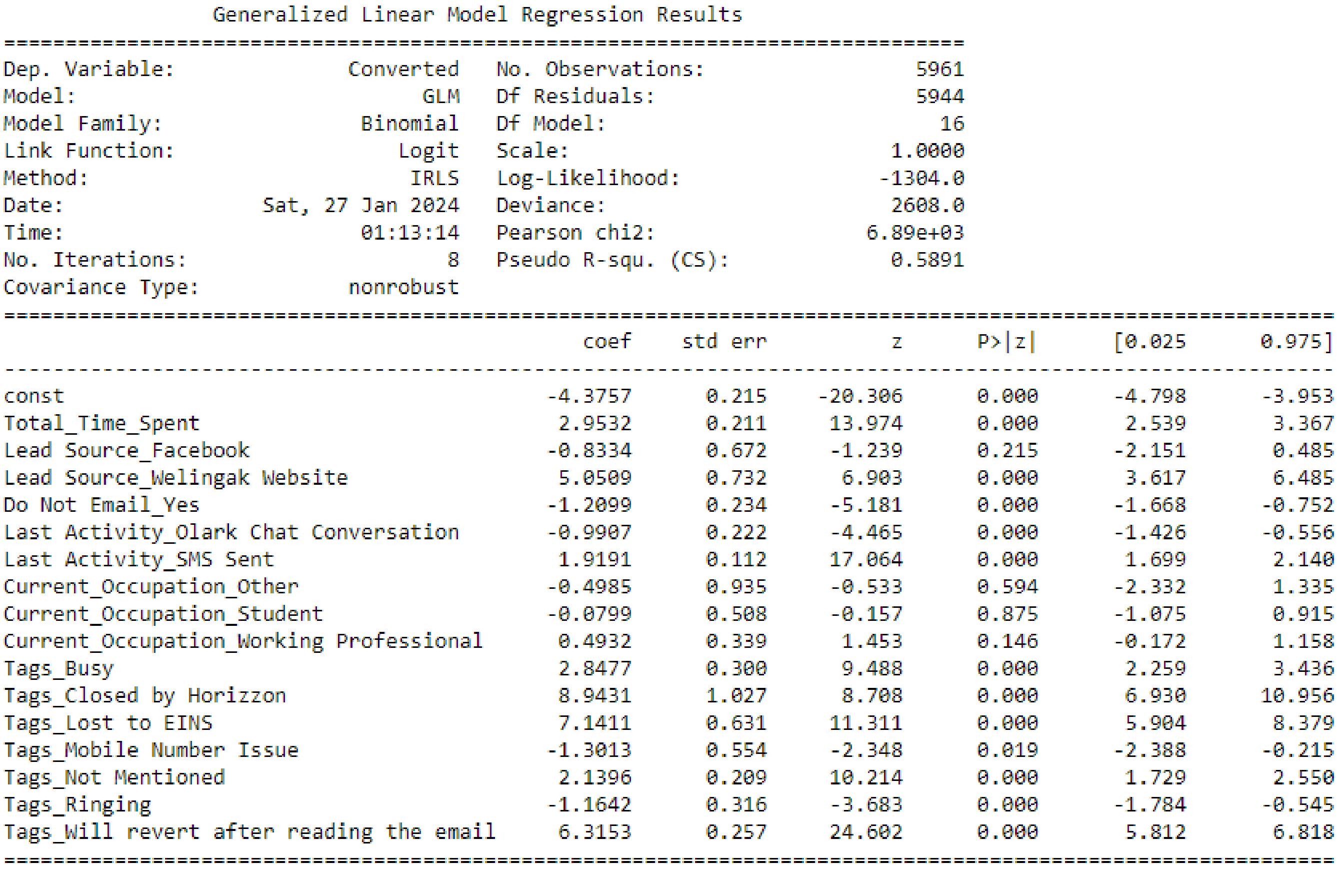
We ran **VIF(Variance Inflation Factor)** for all the independent variables and we kept the threshold as 3. So whichever were found more than or equal to 3 , were dropped off from the variable list. The list of variables and VIF are is as below:



	Features	VIF
2	Page Views Per Visit	6.28
0	TotalVisits	5.61
10	Current_Occupation_Unemployed	3.53
18	Tags_Will revert after reading the email	2.59
1	Total_Time_Spent	2.39
16	Tags_Not Mentioned	1.81
7	Last Activity_SMS Sent	1.74
17	Tags_Ringing	1.73
11	Current_Occupation_Working Professional	1.69
6	Last Activity_Olark Chat Conversation	1.25
15	Tags_Mobile Number Issue	1.22
13	Tags_Closed by Horizzon	1.20
12	Tags_Busy	1.15
4	Lead Source_Welingak Website	1.14
5	Do Not Email_Yes	1.10
14	Tags_Lost to EINS	1.07
9	Current_Occupation_Student	1.07
8	Current_Occupation_Other	1.02
3	Lead Source_Facebook	1.01

After this total number of variables became 16. Again, we ran Logistic regression model. The result was found as below:

- Model 3:

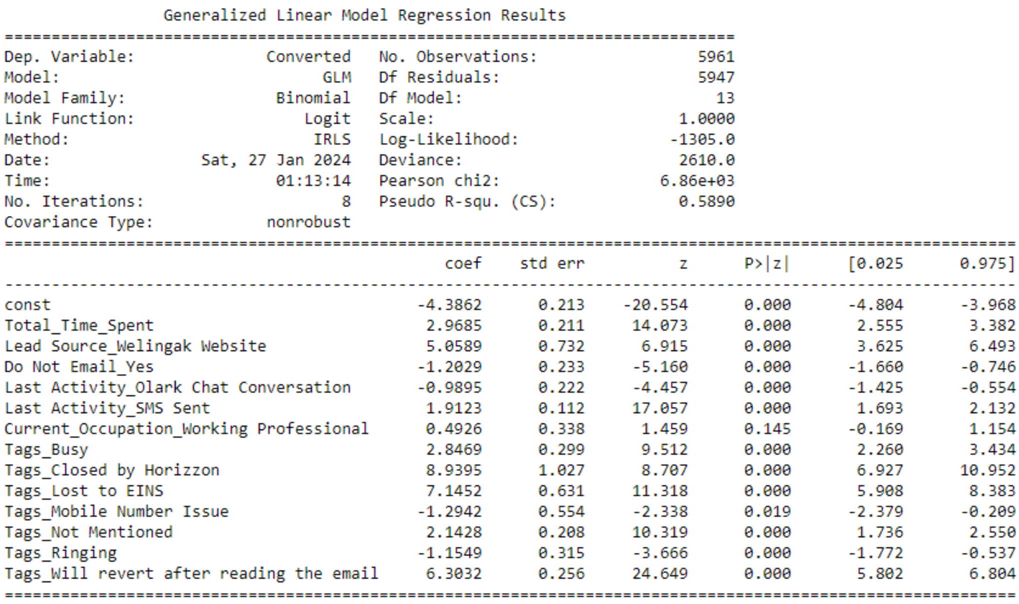


```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 5961
Model:                            GLM   Df Residuals:                     5944
Model Family:                Binomial   Df Model:                           16
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1304.0
Date:                Sat, 27 Jan 2024   Deviance:                       2608.0
Time:                        01:13:14   Pearson chi2:                 6.89e+03
No. Iterations:                     8   Pseudo R-squ. (CS):             0.5891
Covariance Type:            nonrobust
============================================================================================================
                                              coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------
const                                      -4.3757      0.215    -20.306      0.000      -4.798      -3.953
Total_Time_Spent                            2.9532      0.211     13.974      0.000       2.539       3.367
Lead Source_Facebook                       -0.8334      0.672     -1.239      0.215      -2.151       0.485
Lead Source_Welingak Website                5.0509      0.732      6.903      0.000       3.617       6.485
Do Not Email_Yes                           -1.2099      0.234     -5.181      0.000      -1.668      -0.752
Last Activity_Olark Chat Conversation      -0.9907      0.222     -4.465      0.000      -1.426      -0.556
Last Activity_SMS Sent                      1.9191      0.112     17.064      0.000       1.699       2.140
Current_Occupation_Other                   -0.4985      0.935     -0.533      0.594      -2.332       1.335
Current_Occupation_Student                 -0.0799      0.508     -0.157      0.875      -1.075       0.915
Current_Occupation_Working Professional     0.4932      0.339      1.453      0.146      -0.172       1.158
Tags_Busy                                   2.8477      0.300      9.488      0.000       2.259       3.436
Tags_Closed by Horizzon                     8.9431      1.027      8.708      0.000       6.930      10.956
Tags_Lost to EINS                           7.1411      0.631     11.311      0.000       5.904       8.379
Tags_Mobile Number Issue                   -1.3013      0.554     -2.348      0.019      -2.388      -0.215
Tags_Not Mentioned                          2.1396      0.209     10.214      0.000       1.729       2.550
Tags_Ringing                               -1.1642      0.316     -3.683      0.000      -1.784      -0.545
Tags_Will revert after reading the email    6.3153      0.257     24.602      0.000       5.812       6.818
============================================================================================================
```

- The 4 variales with p-value > 0.05 are **"Lead Source_Facebook", "Current_Occupation_Other", "Current_Occupation_Student"** and **"Current_Occupation_Working_Proessional"**.
- By intuition, **Current_Occupation_Working_Professional** seemed to be an important variable because according to the business problem, X_Education tries to sell its courses to the Industry Professionals. So, this variable was not dropped from the variable list. Now, total number of independent variables are 13.

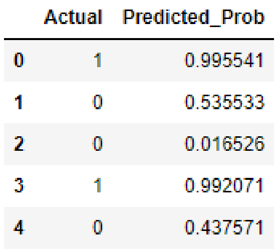
- Model 4:

With these 13 variables we ran a model and that model was considered as the final model. The screenshot is as below:



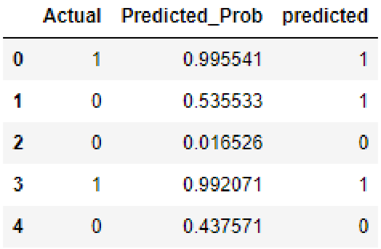
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 5961
Model:                            GLM   Df Residuals:                     5947
Model Family:                Binomial   Df Model:                           13
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1305.0
Date:                Sat, 27 Jan 2024   Deviance:                       2610.0
Time:                        01:13:14   Pearson chi2:                 6.86e+03
No. Iterations:                     8   Pseudo R-squ. (CS):             0.5890
Covariance Type:            nonrobust
============================================================================================================
                                              coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------
const                                      -4.3862      0.213    -20.554      0.000      -4.804      -3.968
Total_Time_Spent                            2.9685      0.211     14.073      0.000       2.555       3.382
Lead Source_Welingak Website                5.0589      0.732      6.915      0.000       3.625       6.493
Do Not Email_Yes                           -1.2029      0.233     -5.160      0.000      -1.660      -0.746
Last Activity_Olark Chat Conversation      -0.9895      0.222     -4.457      0.000      -1.425      -0.554
Last Activity_SMS Sent                      1.9123      0.112     17.057      0.000       1.693       2.132
Current_Occupation_Working Professional     0.4926      0.338      1.459      0.145      -0.169       1.154
Tags_Busy                                   2.8469      0.299      9.512      0.000       2.260       3.434
Tags_Closed by Horizzon                     8.9395      1.027      8.707      0.000       6.927      10.952
Tags_Lost to EINS                           7.1452      0.631     11.318      0.000       5.908       8.383
Tags_Mobile Number Issue                   -1.2942      0.554     -2.338      0.019      -2.379      -0.209
Tags_Not Mentioned                          2.1428      0.208     10.319      0.000       1.736       2.550
Tags_Ringing                               -1.1549      0.315     -3.666      0.000      -1.772      -0.537
Tags_Will revert after reading the email    6.3032      0.256     24.649      0.000       5.802       6.804
============================================================================================================
```

- **Prediction of Train Set:**
  The Train set was predicted as below:



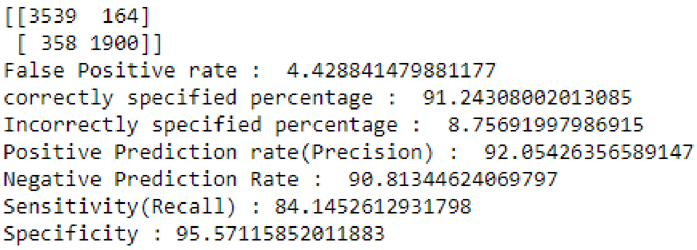
	Actual	Predicted_Prob
0	1	0.995541
1	0	0.535533
2	0	0.016526
3	1	0.992071
4	0	0.437571

The probabilities were rounded off based on a threshold 0.5. Whichever has higher probability greater than 0.5 they were coded as1 and whichever was founded as lower than 0.5 they were coded as 0.



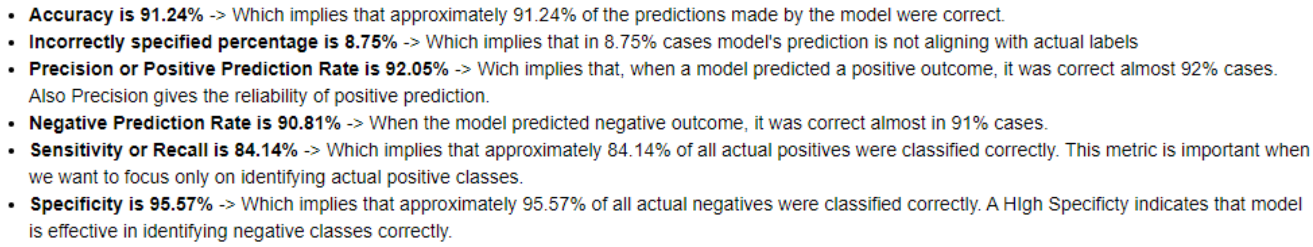
	Actual	Predicted_Prob	predicted
0	1	0.995541	1
1	0	0.535533	1
2	0	0.016526	0
3	1	0.992071	1
4	0	0.437571	0

- **Performance Metrics and Confusion Matrix :**



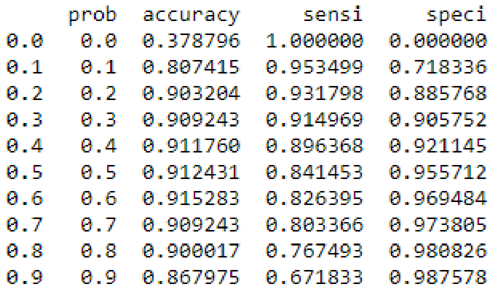
```
[[3539  164]
 [ 358 1900]]
False Positive rate :  4.428841479881177
correctly specified percentage :  91.24308002013085
Incorrectly specified percentage :  8.75691997986915
Positive Prediction rate(Precision) :  92.05426356589147
Negative Prediction Rate :  90.81344624069797
Sensitivity(Recall) : 84.1452612931798
Specificity : 95.57115852011883
```

  - **Inference:**

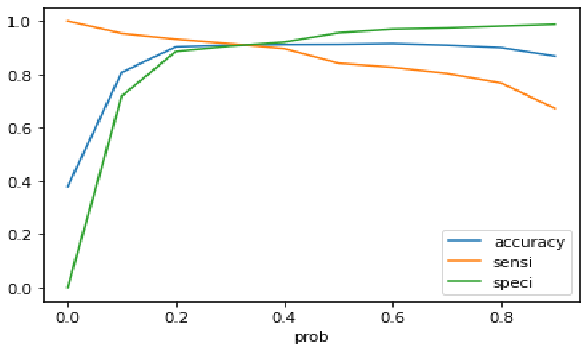


- **Accuracy is 91.24%** -> Which implies that approximately 91.24% of the predictions made by the model were correct.
- **Incorrectly specified percentage is 8.75%** -> Which implies that in 8.75% cases model's prediction is not aligning with actual labels
- **Precision or Positive Prediction Rate is 92.05%** -> Wich implies that, when a model predicted a positive outcome, it was correct almost 92% cases. Also Precision gives the reliability of positive prediction.
- **Negative Prediction Rate is 90.81%** -> When the model predicted negative outcome, it was correct almost in 91% cases.
- **Sensitivity or Recall is 84.14%** -> Which implies that approximately 84.14% of all actual positives were classified correctly. This metric is important when we want to focus only on identifying actual positive classes.
- **Specificity is 95.57%** -> Which implies that approximately 95.57% of all actual negatives were classified correctly. A HIgh Specificty indicates that model is effective in identifying negative classes correctly.

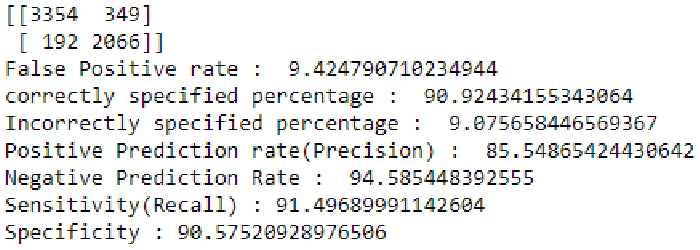
    - The whole interpretation shows that the model is pretty good and it's quite capable in classifying the leads into two classes (1 and 0).
    - Post that we tried to identify the optimum threshold value. Basically, for several values between 0 to 1 we tried to check the round off the predicted probability values. Thereafter for those values we derived accuracy, sensitivity and specificity. These graphs were plotted. The intersection point was taken as our optimum point for further analysis.



```
     prob  accuracy     sensi     speci
0.0   0.0  0.378796  1.000000  0.000000
0.1   0.1  0.807415  0.953499  0.718336
0.2   0.2  0.903204  0.931798  0.885768
0.3   0.3  0.909243  0.914969  0.905752
0.4   0.4  0.911760  0.896368  0.921145
0.5   0.5  0.912431  0.841453  0.955712
0.6   0.6  0.915283  0.826395  0.969484
0.7   0.7  0.909243  0.803366  0.973805
0.8   0.8  0.900017  0.767493  0.980826
0.9   0.9  0.867975  0.671833  0.987578
```



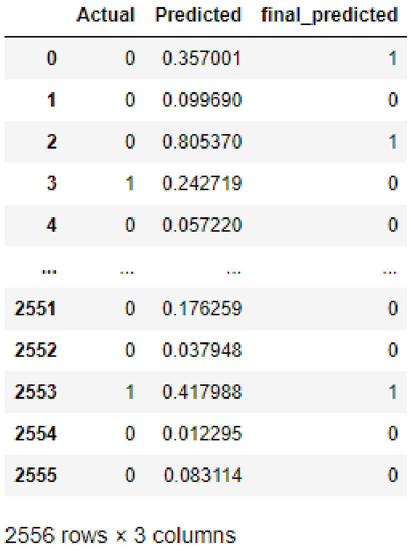
- From the graph it was clearly understood that 0.3 is our threshold point and based on that our confusion matrix and other metrics along with accuracy came as below:



```
[[3354  349]
 [ 192 2066]]
False Positive rate :  9.424790710234944
correctly specified percentage :  90.92434155343064
Incorrectly specified percentage :  9.075658446569367
Positive Prediction rate(Precision) :  85.54865424430642
Negative Prediction Rate :  94.585448392555
Sensitivity(Recall) : 91.49689991142604
Specificity : 90.57520928976506
```

Which can be interpreted as the model can provide accuracy of approximately 91% and also Sensitivity is found to be 91.5% and specificity is 91%.

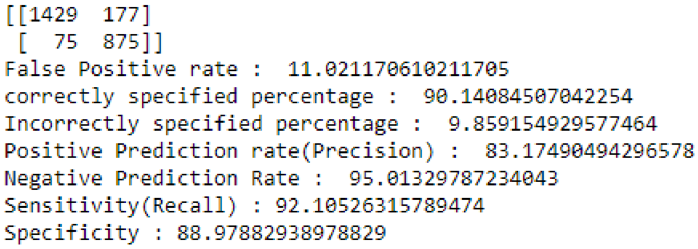
- With this model we predicted Test set where the result is found as below:



	Actual	Predicted	final_predicted
0	0	0.357001	1
1	0	0.099690	0
2	0	0.805370	1
3	1	0.242719	0
4	0	0.057220	0
...	...	...	...
2551	0	0.176259	0
2552	0	0.037948	0
2553	1	0.417988	1
2554	0	0.012295	0
2555	0	0.083114	0

2556 rows × 3 columns

- The Predicted column has been rounded off based on the probability value 0.3. Whichever probability is greater than 0.3 was written as 1 else written as 0.
- **The Performance Metric on Test Set:**
  The confusion matrix and other performance metrics along with accuracy are shown in the below image:



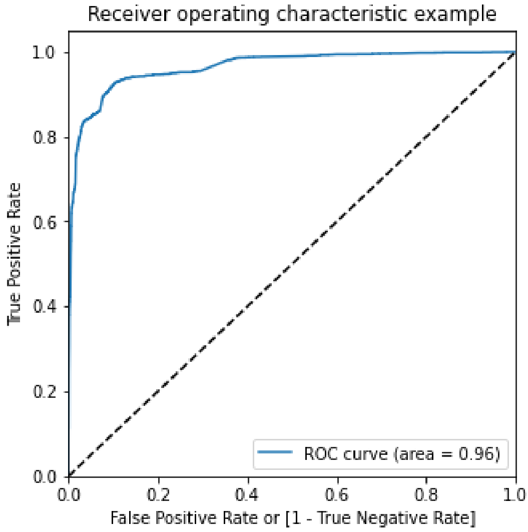
```
[[1429  177]
 [  75  875]]
False Positive rate :  11.021170610211705
correctly specified percentage :  90.14084507042254
Incorrectly specified percentage :  9.859154929577464
Positive Prediction rate(Precision) :  83.17490494296578
Negative Prediction Rate :  95.01329787234043
Sensitivity(Recall) : 92.10526315789474
Specificity : 88.97882938978829
```

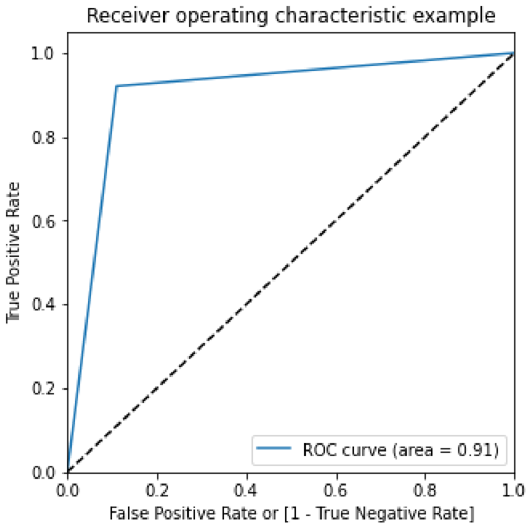
Which can be interpreted as:

- for the test data or unknown data the model can provide 90.14% accuracy i.e, approximately 90.14% predictions made by the model are correct.

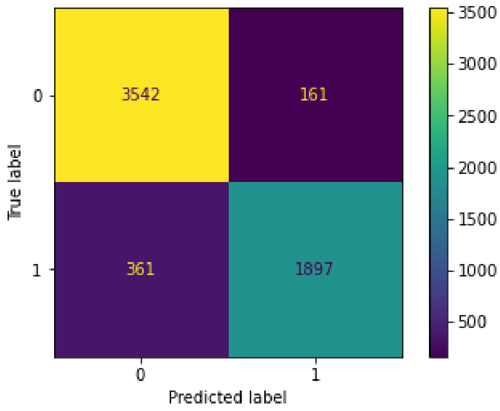
- 9.85% predictions by the model are erroneous.
- Precision is 83.17% i.e, when the model predicts something positive, it's found to be correct almost for 83.17%.
- Negative Prediction rate is 95.1% which says that out of all predicted negative by the model, approximately 95.1% is correct.
- Sensitivity or Recall is 92.10% which implies that the model can correctly identify approximately 92.10% Actual Positives out of all actual positives.
- Specificity is 88.97% which implies that the model can correctly identify approximately 88.97% Actual Negatives out of all actual negatives.

- **The AUC-ROC Curve:**
  For Train and Test set the AUC-ROC score was found to be 0.96 and 0.91 respectively. The graphs are shown below:

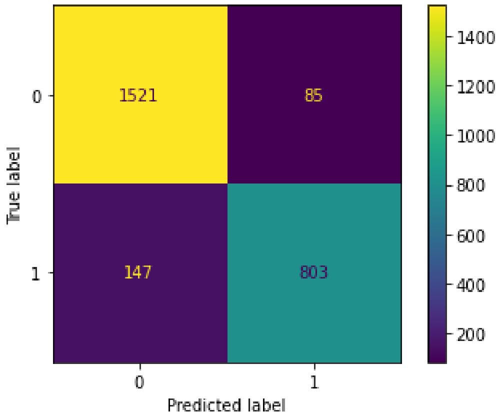




- **The Colourful Confusion Matrix for both Train and Test Set:**



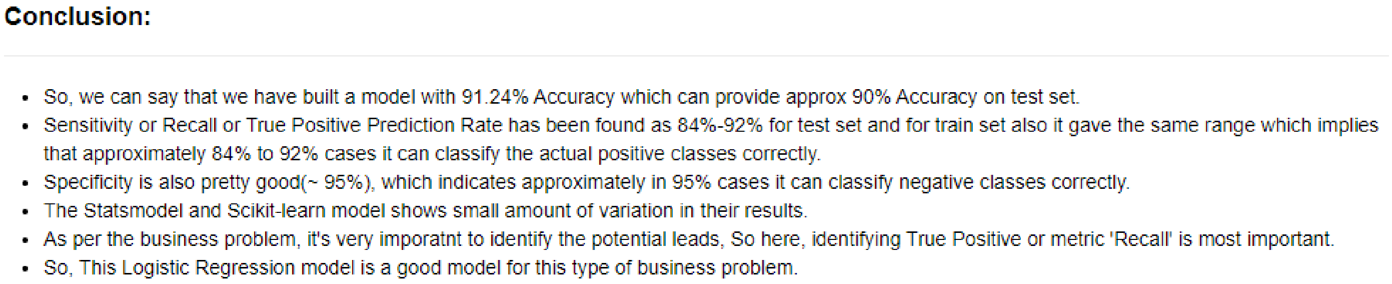
Train Set Confusion Matrix



Test Set Confusion Matrix

As X_education tries to identify the potential leads , so for that purpose Sensitivity can be one of the most important metric. Our model can provide a good trade off between Sensitivity and Specificity wit a descent Accuracy score. Though many modifications can still be done on the final model.

- **The Conclusion:**



Conclusion:

- So, we can say that we have built a model with 91.24% Accuracy which can provide approx 90% Accuracy on test set.
- Sensitivity or Recall or True Positive Prediction Rate has been found as 84%-92% for test set and for train set also it gave the same range which implies that approximately 84% to 92% cases it can classify the actual positive classes correctly.
- Specificity is also pretty good(~ 95%), which indicates approximately in 95% cases it can classify negative classes correctly.
- The Statsmodel and Scikit-learn model shows small amount of variation in their results.
- As per the business problem, it's very imporatnt to identify the potential leads, So here, identifying True Positive or metric 'Recall' is most important.
- So, This Logistic Regression model is a good model for this type of business problem.