Lead Score Case Study

By:

Ranip Hore

Sanchari Gautam

Amith Pradhan

Pratik Nath

Abstract:

- **Agenda**:

X Education wants to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

- **Terminologies used**:
  - <u>Leads</u> - The company markets its courses on several websites and search engines like Google. Once visitors browse the courses or fill up a form for the course or watch some videos, they fill up a form providing their email address or phone number. These people are classified to be a lead.
  - <u>Conversion Rate</u> – The process of conversion of leads to paid customers is called conversion and the rate of conversion of leads is known as conversion rate.
  - <u>Hot Leads</u> - The company wishes to identify the most potential leads that is the leads who are mostly likely to convert into paid customers, known as 'Hot Leads'.

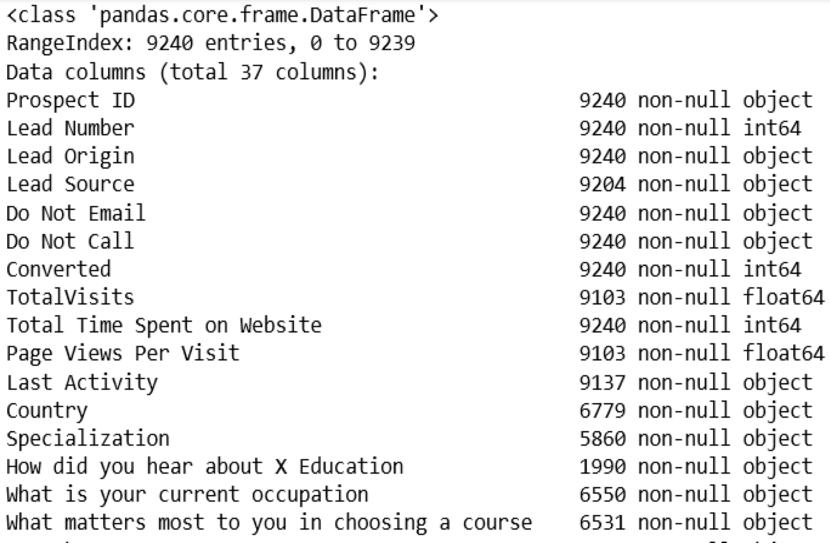
- **Method Applied**:
  - After data cleaning and outlier treatment, we have applied scaling on all numerical variables and created dummy variables for all categorical variables.
  - On application of logistic regression model, the important features are chosen using RFE and then the predicted variables are chosen taking the stipulated cut off for the convert probability, so that the precision value can reach 80% with minimum recall value.

Data Dictionary:

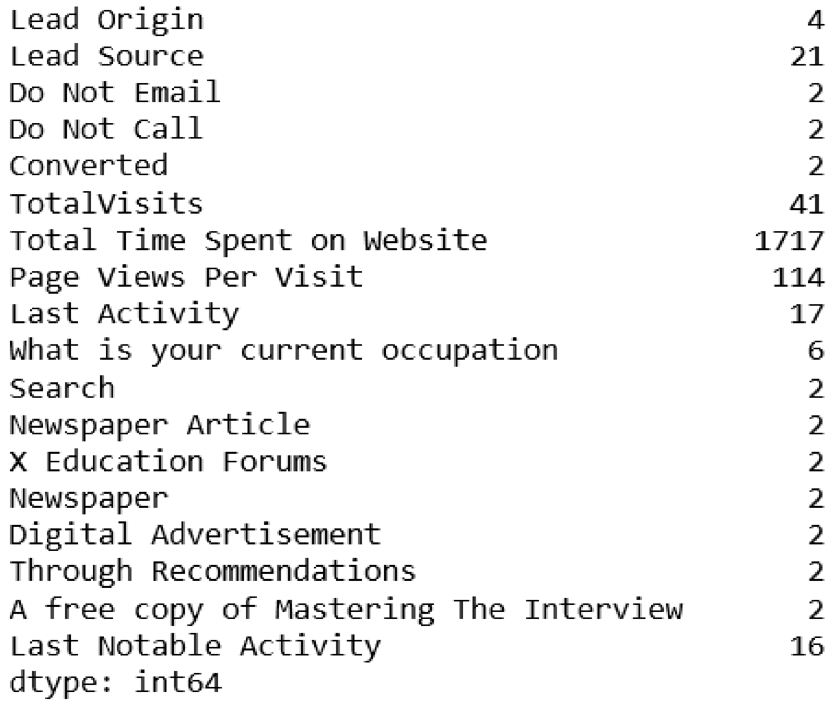
Variables	Description
Prospect ID	A unique ID with which the customer is identified.
Lead Number	A lead number assigned to each lead procured.
Lead Origin	The origin identifier with which the customer was identified to be a lead. Includes API, Landing Page Submission, etc.
Lead Source	The source of the lead. Includes Google, Organic Search, Olark Chat, etc.
Do Not Email	An indicator variable selected by the customer wherein they select whether of not they want to be emailed about the course or not.
Do Not Call	An indicator variable selected by the customer wherein they select whether of not they want to be called about the course or not.
Converted	The target variable. Indicates whether a lead has been successfully converted or not.
TotalVisits	The total number of visits made by the customer on the website.
Total Time Spent on Website	The total time spent by the customer on the website.
Page Views Per Visit	Average number of pages on the website viewed during the visits.
Last Activity	Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.
Country	The country of the customer.
Specialization	The industry domain in which the customer worked before. Includes the level 'Select Specialization' which means the customer had not selected this option while filling the form.
How did you hear about X Education	The source from which the customer heard about X Education.
What is your current occupation	Indicates whether the customer is a student, umemployed or employed.
What matters most to you in choosing this course	An option selected by the customer indicating what is their main motto behind doing this course.
Search	Indicating whether the customer had seen the ad in any of the listed items.
Magazine	
Newspaper Article	
X Education Forums	
Newspaper	
Digital Advertisement	
Through Recommendations	Indicates whether the customer came in through recommendations.
Receive More Updates About Our Courses	Indicates whether the customer chose to receive more updates about the courses.
Tags	Tags assigned to customers indicating the current status of the lead.
Lead Quality	Indicates the quality of lead based on the data and intuition the the employee who has been assigned to the lead.
Update me on Supply Chain Content	Indicates whether the customer wants updates on the Supply Chain Content.
Get updates on DM Content	Indicates whether the customer wants updates on the DM Content.
Lead Profile	A lead level assigned to each customer based on their profile.
City	The city of the customer.
Asymmetrique Activity Index	An index and score assigned to each customer based on their activity and their profile
Asymmetrique Profile Index	
Asymmetrique Activity Score	
Asymmetrique Profile Score	
I agree to pay the amount through cheque	Indicates whether the customer has agreed to pay the amount through cheque or not.
a free copy of Mastering The Interview	Indicates whether the customer wants a free copy of 'Mastering the Interview' or not.
Last Notable Activity	The last notable activity performed by the student.

Data Understanding:

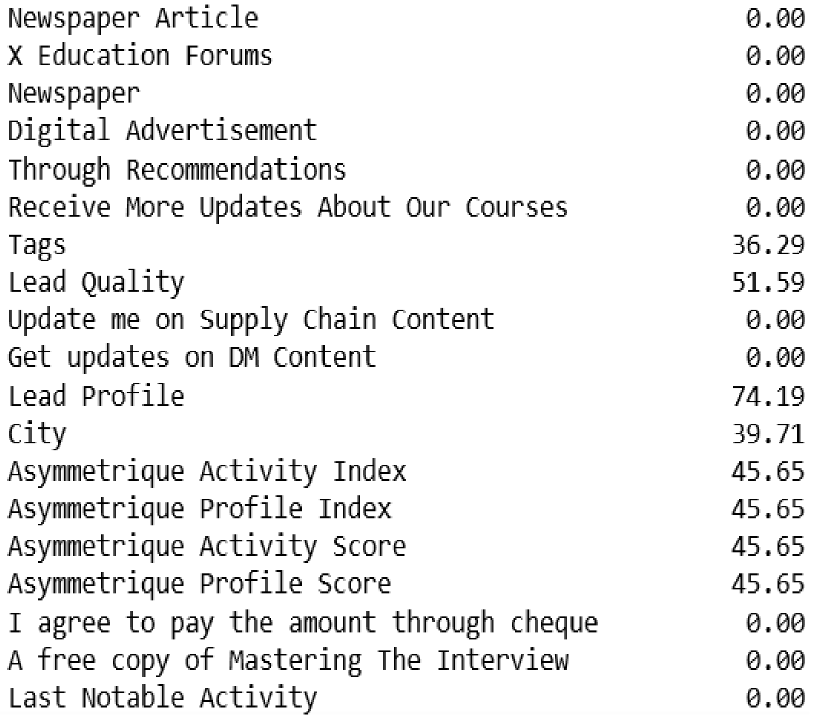
- The dataset has 37 columns and 9240 rows in total.
- There are several columns which have more than 30% null value data. It is better to drop those columns.
- After outlier treatment, dropping all the duplicate values and removing all null values, the final shape comes out to (8513,11).



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
Prospect ID                                      9240 non-null object
Lead Number                                      9240 non-null int64
Lead Origin                                      9240 non-null object
Lead Source                                      9204 non-null object
Do Not Email                                     9240 non-null object
Do Not Call                                      9240 non-null object
Converted                                        9240 non-null int64
TotalVisits                                      9103 non-null float64
Total Time Spent on Website                      9240 non-null int64
Page Views Per Visit                             9103 non-null float64
Last Activity                                    9137 non-null object
Country                                          6779 non-null object
Specialization                                   5860 non-null object
How did you hear about X Education               1990 non-null object
What is your current occupation                  6550 non-null object
What matters most to you in choosing a course    6531 non-null object
```



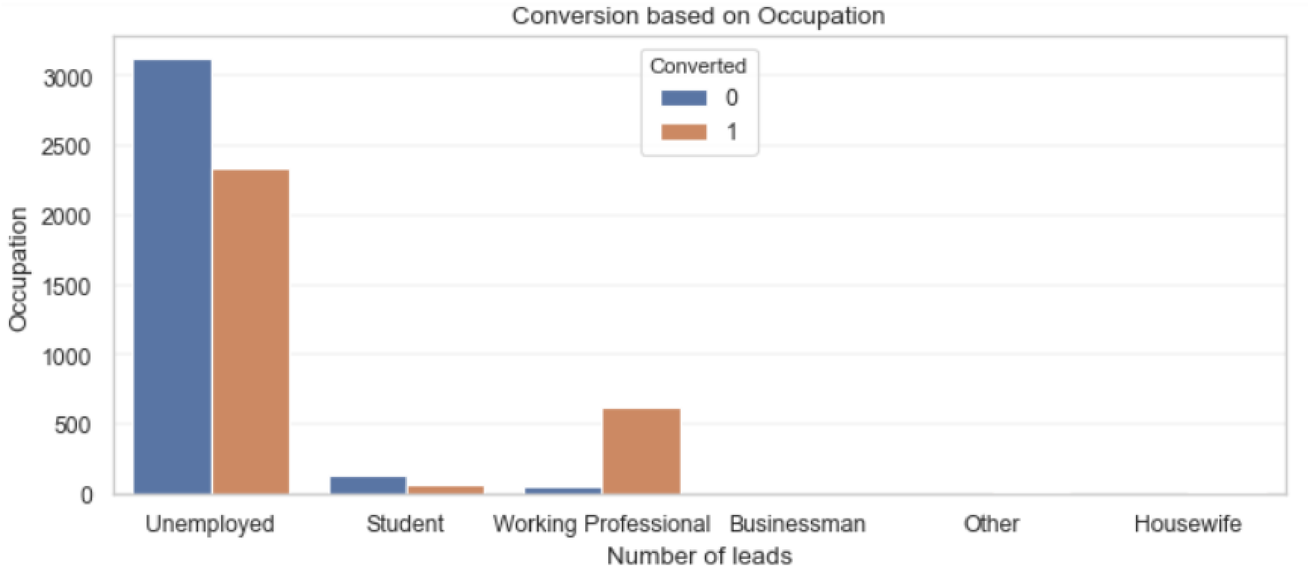
```
Lead Origin                                  4
Lead Source                                 21
Do Not Email                                 2
Do Not Call                                  2
Converted                                    2
TotalVisits                                 41
Total Time Spent on Website               1717
Page Views Per Visit                       114
Last Activity                               17
What is your current occupation              6
Search                                       2
Newspaper Article                            2
X Education Forums                           2
Newspaper                                    2
Digital Advertisement                        2
Through Recommendations                      2
A free copy of Mastering The Interview       2
Last Notable Activity                       16
dtype: int64
```



```
Newspaper Article                                 0.00
X Education Forums                                0.00
Newspaper                                         0.00
Digital Advertisement                             0.00
Through Recommendations                           0.00
Receive More Updates About Our Courses            0.00
Tags                                             36.29
Lead Quality                                     51.59
Update me on Supply Chain Content                 0.00
Get updates on DM Content                         0.00
Lead Profile                                     74.19
City                                             39.71
Asymmetrique Activity Index                      45.65
Asymmetrique Profile Index                       45.65
Asymmetrique Activity Score                      45.65
Asymmetrique Profile Score                       45.65
I agree to pay the amount through cheque          0.00
A free copy of Mastering The Interview            0.00
Last Notable Activity                             0.00
```

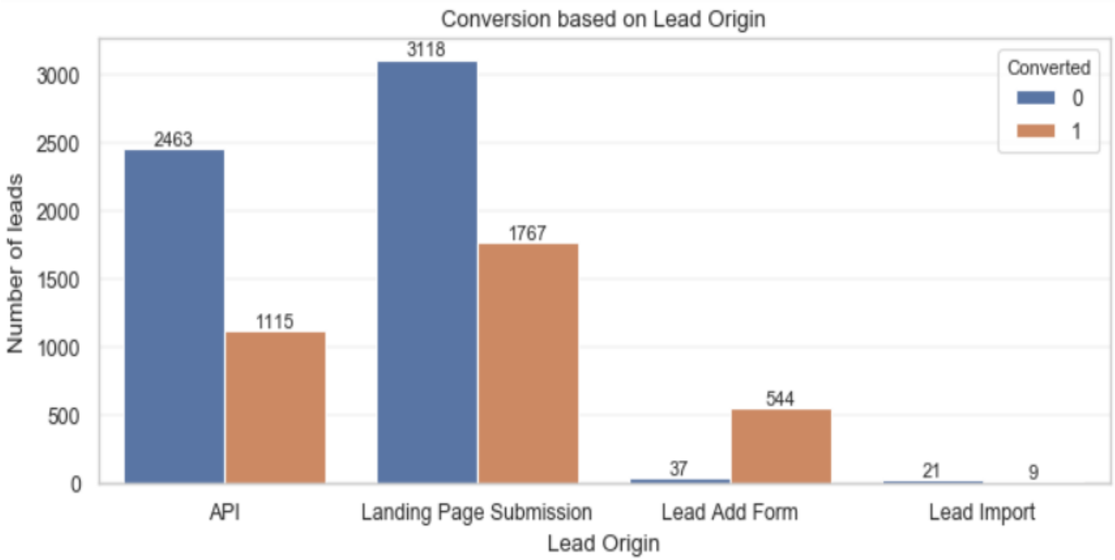
EDA on the Dataset:

- Conversion based on Occupation:



Comments: The conversion rate is the highest for the unemployed people.

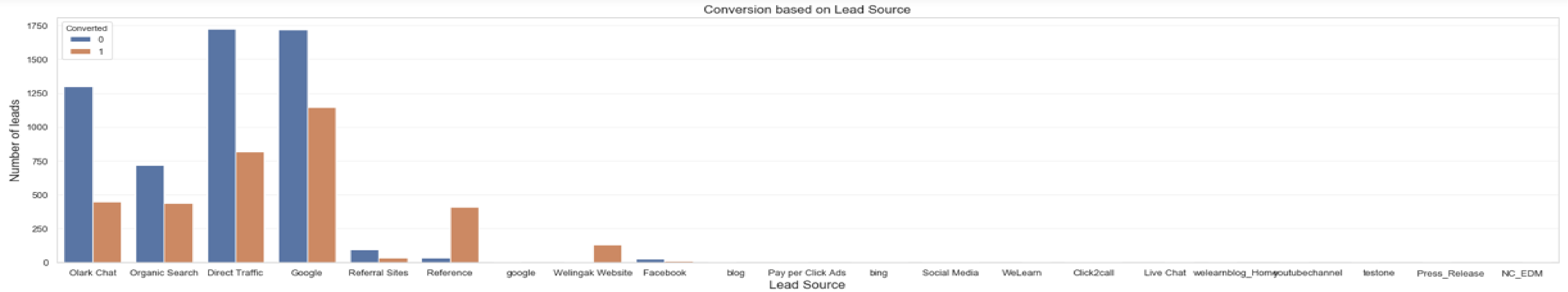
- Conversion based on Lead Origin:



Comments: The conversion rate is the highest for the leads signing in the landing page.

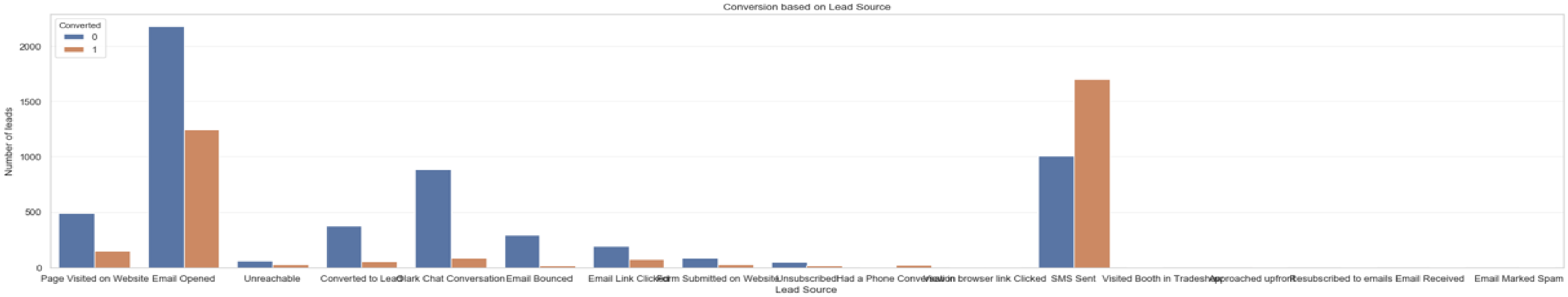
EDA on the Dataset:

- Conversion based on Lead Source:



Comments: The conversion rate is the highest for the source as Google.

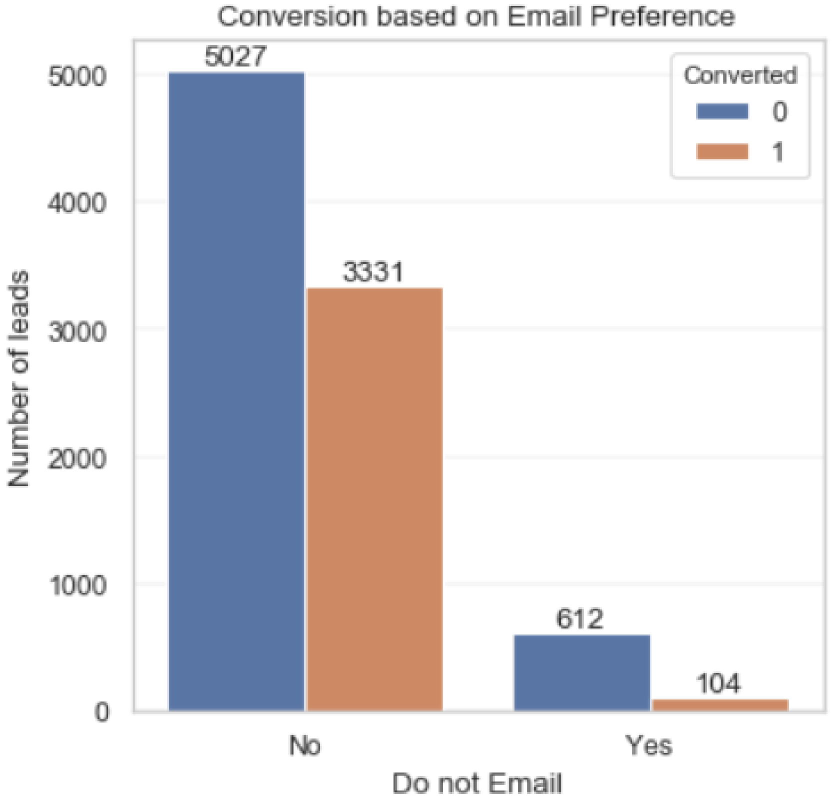
- Conversion based on Last Activity:



Comments: The conversion rate is the highest for the SMS sent communication.

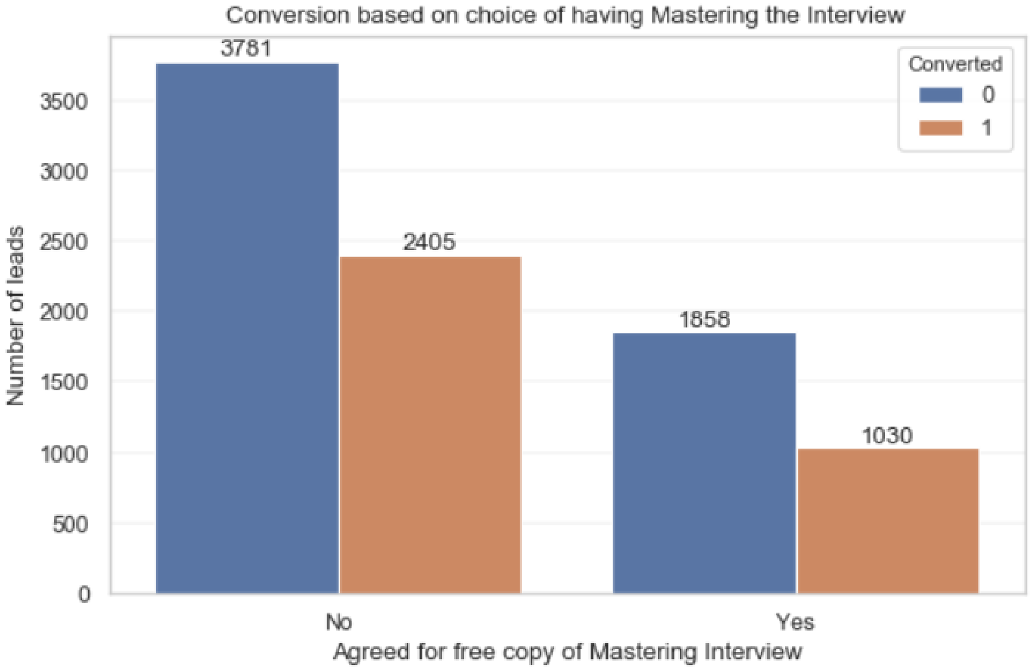
EDA on the Dataset:

- Conversion based on Email Preference:



Comments: The conversion rate is the high amongst those who have chosen Email as the way of contact.

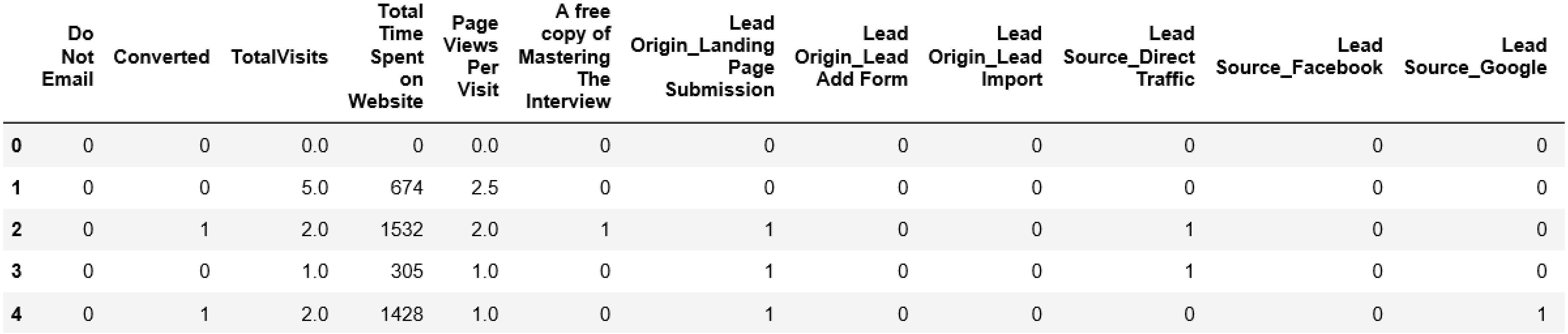
- Conversion based on Interview Preparation:



Comments: The conversion rate is high for the leads if they subscribe for the copy of interview questions.

Data Preparation:

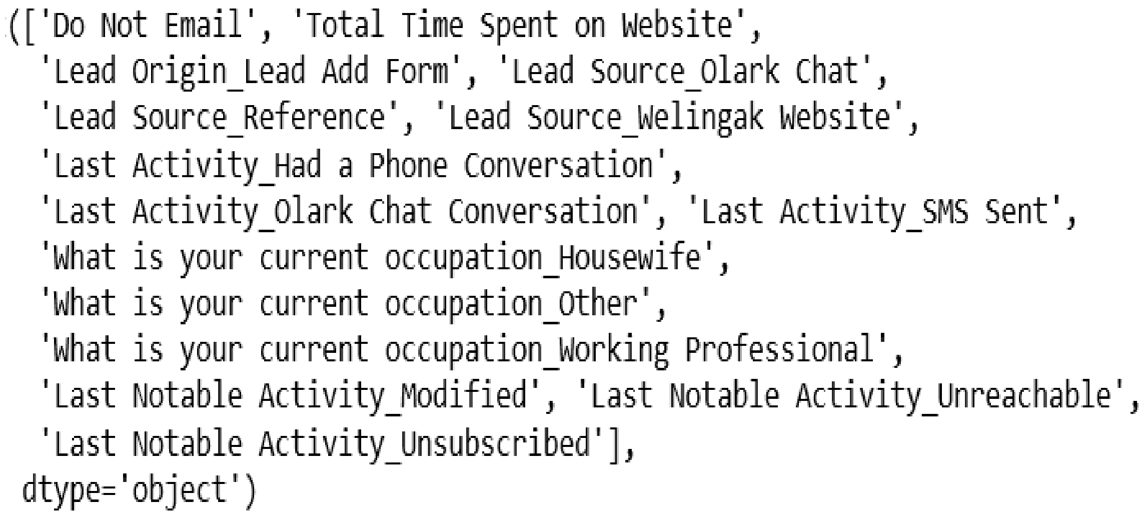
- Binary mapping is done for the columns: 'Do Not Email‘ and 'A free copy of Mastering The Interview‘.
- Dummy variables are created for the rest of the categorical variables: 'Lead Origin', 'Lead Source', 'Last Activity', 'What is your current occupation‘ and 'Last Notable Activity‘.
- The final shape of the dataset comes to (8513,65).



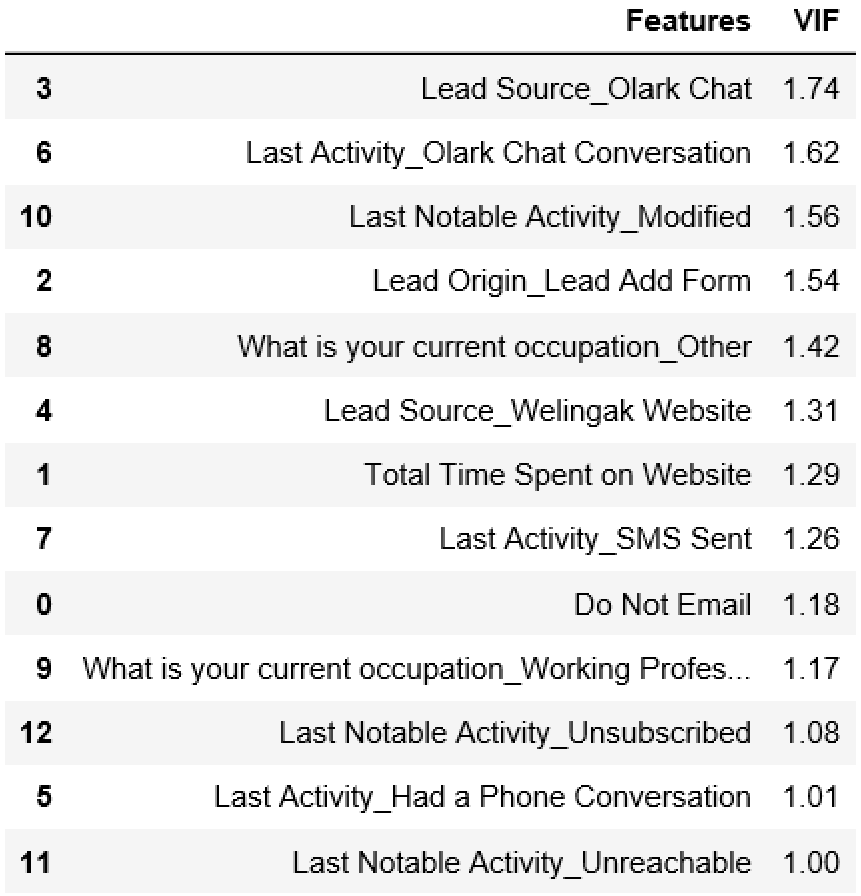
	Do Not Email	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	A free copy of Mastering The Interview	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Source_Direct Traffic	Lead Source_Facebook	Lead Source_Google
0	0	0	0.0	0	0.0	0	0	0	0	0	0	0
1	0	0	5.0	674	2.5	0	0	0	0	0	0	0
2	0	1	2.0	1532	2.0	1	1	0	0	1	0	0
3	0	0	1.0	305	1.0	0	1	0	0	1	0	0
4	0	1	2.0	1428	1.0	0	1	0	0	0	0	1

RFE Selection

The RFE selection selected some of the columns (attached in the screenshot) from which the columns having higher p value and vif value are removed. The final set is also attached below. On that dataset, we have applied conversion probability cut off based on which the final model is chosen.

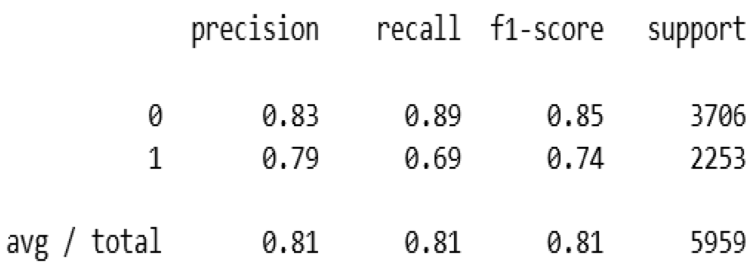


```
([' Do Not Email', 'Total Time Spent on Website',
  'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat',
  'Lead Source_Reference', 'Lead Source_Welingak Website',
  'Last Activity_Had a Phone Conversation',
  'Last Activity_Olark Chat Conversation', 'Last Activity_SMS Sent',
  'What is your current occupation_Housewife',
  'What is your current occupation_Other',
  'What is your current occupation_Working Professional',
  'Last Notable Activity_Modified', 'Last Notable Activity_Unreachable',
  'Last Notable Activity_Unsubscribed'],
 dtype='object')
```

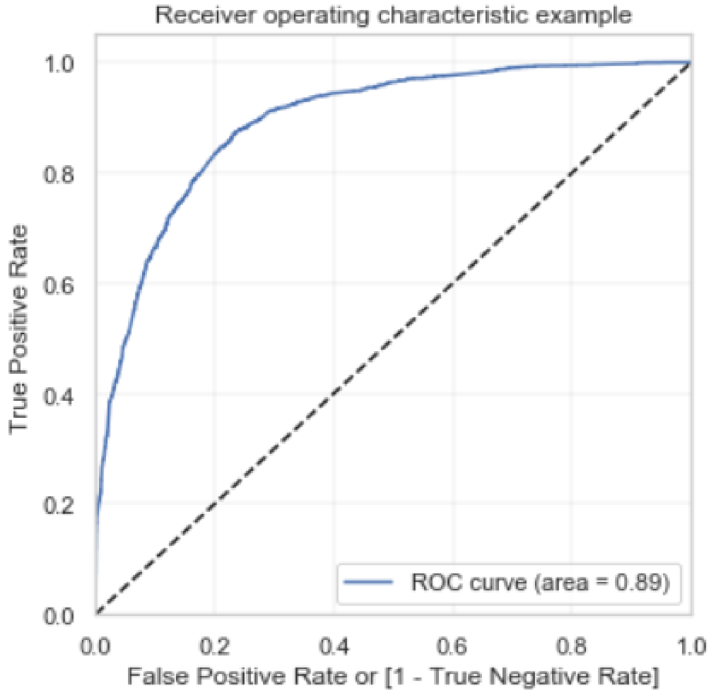


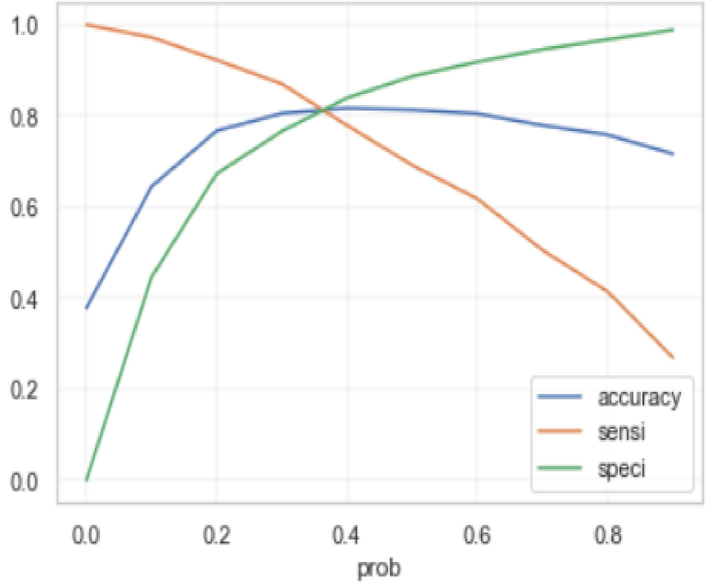
	Features	VIF
3	Lead Source_Olark Chat	1.74
6	Last Activity_Olark Chat Conversation	1.62
10	Last Notable Activity_Modified	1.56
2	Lead Origin_Lead Add Form	1.54
8	What is your current occupation_Other	1.42
4	Lead Source_Welingak Website	1.31
1	Total Time Spent on Website	1.29
7	Last Activity_SMS Sent	1.26
0	Do Not Email	1.18
9	What is your current occupation_Working Profes...	1.17
12	Last Notable Activity_Unsubscribed	1.08
5	Last Activity_Had a Phone Conversation	1.01
11	Last Notable Activity_Unreachable	1.00

Prediction based on Cut Off – 0.5



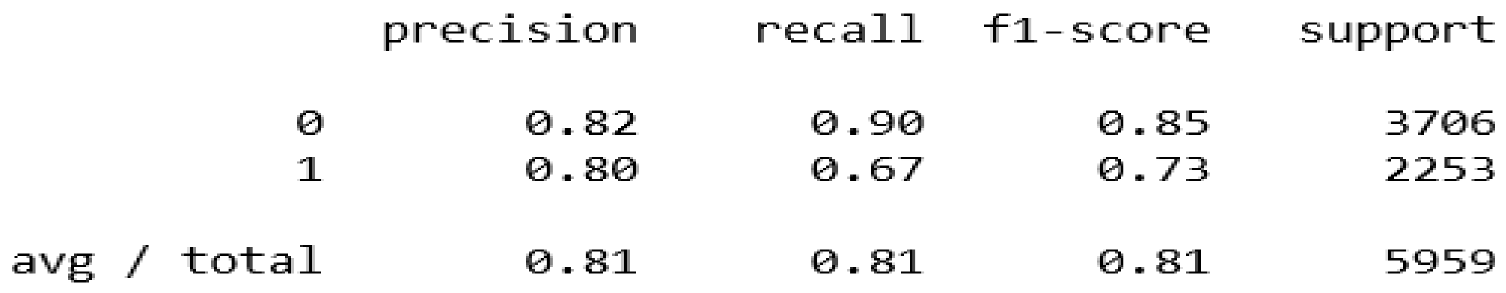
```
             precision    recall  f1-score   support

          0       0.83      0.89      0.85      3706
          1       0.79      0.69      0.74      2253

avg / total       0.81      0.81      0.81      5959
```



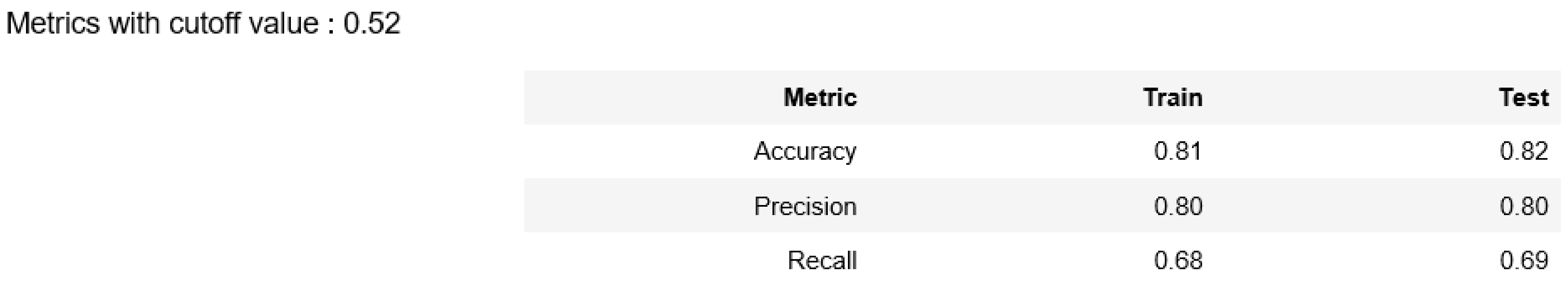


With cut off 0.5, we have reached precision 0.79 and recall 0.69. Our aim is to reach .80 according to CEO of the company.

Prediction based on Cut Off – 0.52



```
             precision    recall  f1-score   support

          0       0.82      0.90      0.85      3706
          1       0.80      0.67      0.73      2253

avg / total       0.81      0.81      0.81      5959
```



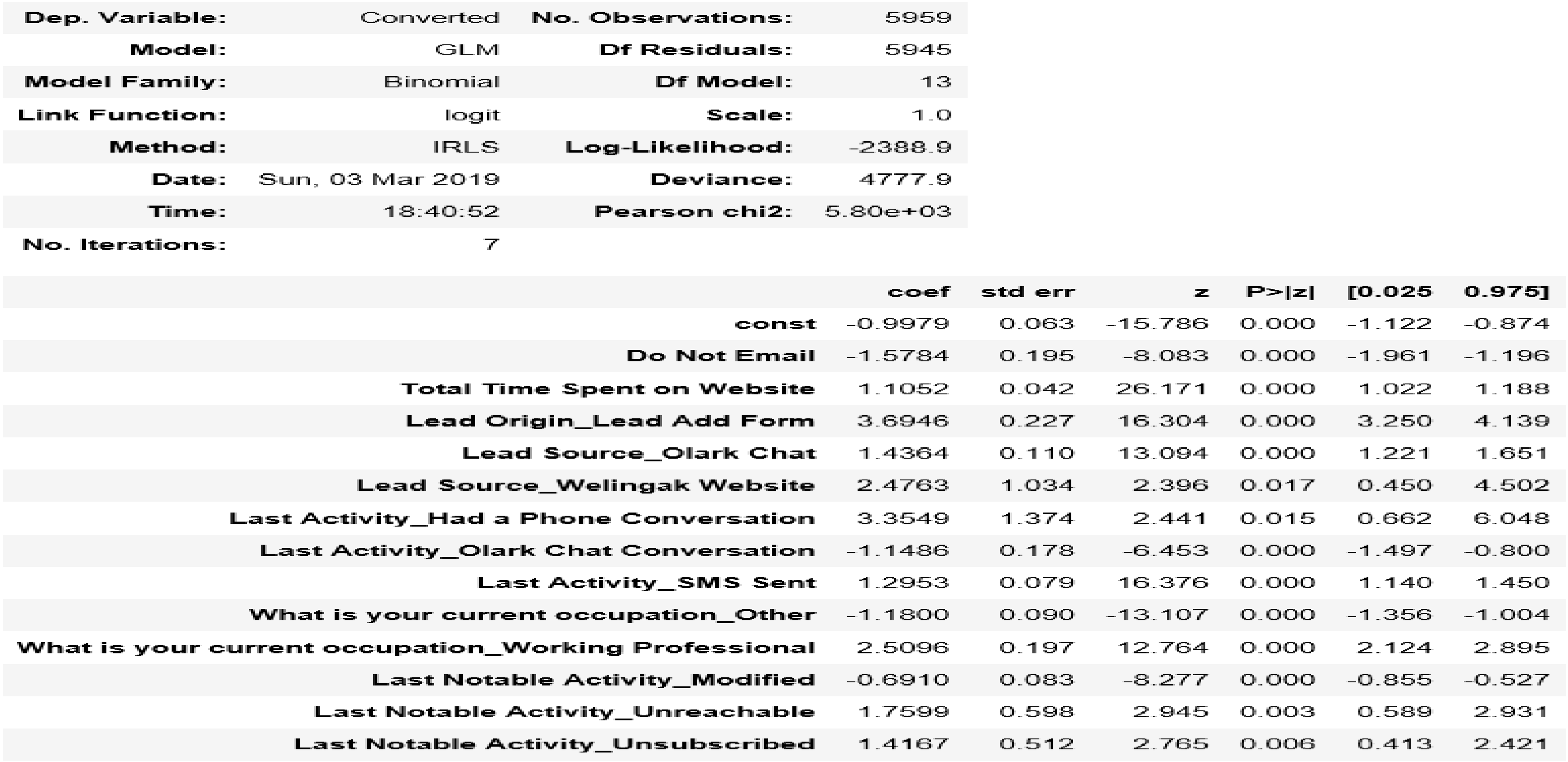
Metrics with cutoff value : 0.52

Metric	Train	Test
Accuracy	0.81	0.82
Precision	0.80	0.80
Recall	0.68	0.69

With cut off 0.52, we have reached precision 0.80 and recall 0.68. Hence it is better to settle for 0.52 as cut off.

Business Problems:

- Which are the top three variables in your model which contribute most towards the probability of a lead getting converted?



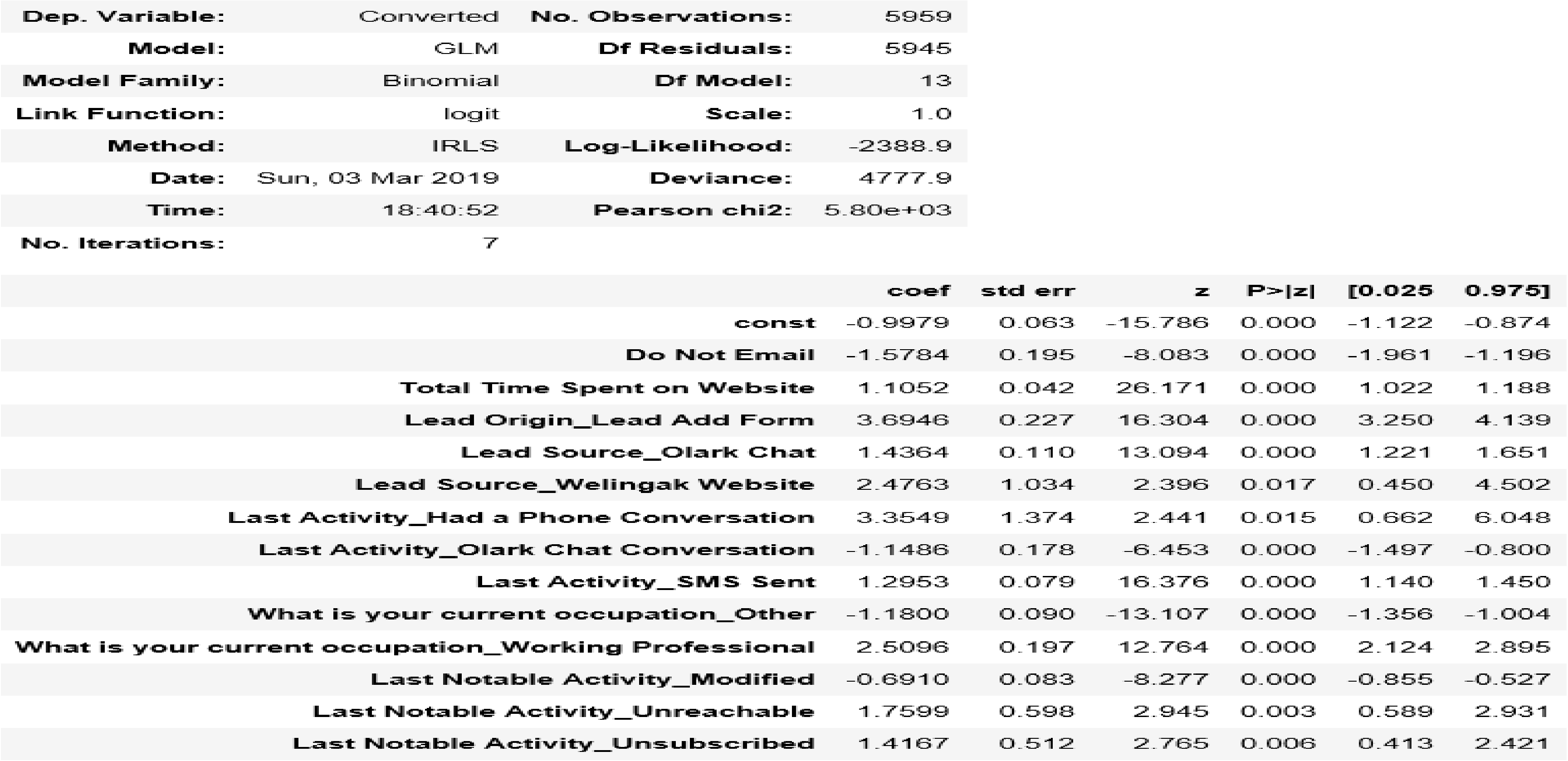
Dep. Variable:	Converted	No. Observations:	5959
Model:	GLM	Df Residuals:	5945
Model Family:	Binomial	Df Model:	13
Link Function:	logit	Scale:	1.0
Method:	IRLS	Log-Likelihood:	-2388.9
Date:	Sun, 03 Mar 2019	Deviance:	4777.9
Time:	18:40:52	Pearson chi2:	5.80e+03
No. Iterations:	7		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9979	0.063	-15.786	0.000	-1.122	-0.874
Do Not Email	-1.5784	0.195	-8.083	0.000	-1.961	-1.196
Total Time Spent on Website	1.1052	0.042	26.171	0.000	1.022	1.188
Lead Origin_Lead Add Form	3.6946	0.227	16.304	0.000	3.250	4.139
Lead Source_Olark Chat	1.4364	0.110	13.094	0.000	1.221	1.651
Lead Source_Welingak Website	2.4763	1.034	2.396	0.017	0.450	4.502
Last Activity_Had a Phone Conversation	3.3549	1.374	2.441	0.015	0.662	6.048
Last Activity_Olark Chat Conversation	-1.1486	0.178	-6.453	0.000	-1.497	-0.800
Last Activity_SMS Sent	1.2953	0.079	16.376	0.000	1.140	1.450
What is your current occupation_Other	-1.1800	0.090	-13.107	0.000	-1.356	-1.004
What is your current occupation_Working Professional	2.5096	0.197	12.764	0.000	2.124	2.895
Last Notable Activity_Modified	-0.6910	0.083	-8.277	0.000	-0.855	-0.527
Last Notable Activity_Unreachable	1.7599	0.598	2.945	0.003	0.589	2.931
Last Notable Activity_Unsubscribed	1.4167	0.512	2.765	0.006	0.413	2.421

After application of the logistic regression model on the dataset, the analysis report (as shown above) depicts that the top three variables which contribute the most towards the probability of a lead getting converted are: Lead Origin , Last Activity and Current Occupation of the lead because, these three variables have the highest coefficient.

Business Problems:

- What are the top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion?



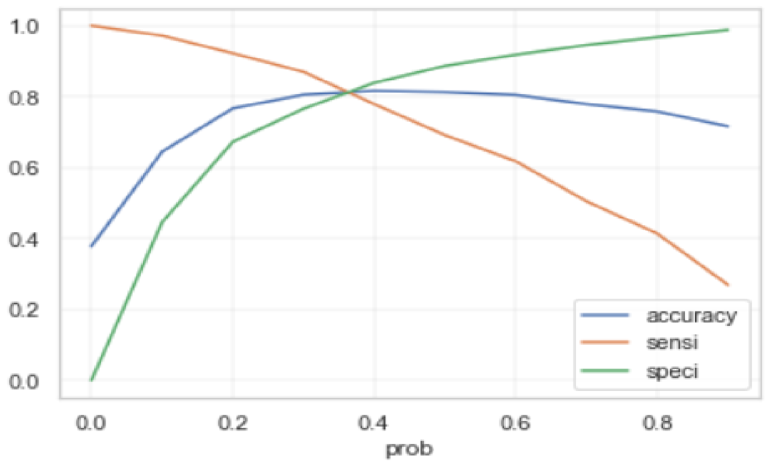
```
    Dep. Variable:          Converted  No. Observations:      5959
            Model:                GLM      Df Residuals:      5945
     Model Family:           Binomial          Df Model:        13
    Link Function:              logit             Scale:       1.0
           Method:               IRLS    Log-Likelihood:   -2388.9
             Date:   Sun, 03 Mar 2019          Deviance:    4777.9
             Time:           18:40:52     Pearson chi2:   5.80e+03
  No. Iterations:                  7

                                                   coef  std err        z  P>|z|  [0.025  0.975]
                                       const    -0.9979    0.063  -15.786  0.000  -1.122  -0.874
                                Do Not Email    -1.5784    0.195   -8.083  0.000  -1.961  -1.196
                 Total Time Spent on Website     1.1052    0.042   26.171  0.000   1.022   1.188
                  Lead Origin_Lead Add Form      3.6946    0.227   16.304  0.000   3.250   4.139
                     Lead Source_Olark Chat      1.4364    0.110   13.094  0.000   1.221   1.651
                Lead Source_Welingak Website     2.4763    1.034    2.396  0.017   0.450   4.502
        Last Activity_Had a Phone Conversation   3.3549    1.374    2.441  0.015   0.662   6.048
       Last Activity_Olark Chat Conversation    -1.1486    0.178   -6.453  0.000  -1.497  -0.800
                     Last Activity_SMS Sent      1.2953    0.079   16.376  0.000   1.140   1.450
          What is your current occupation_Other -1.1800  0.090  -13.107  0.000  -1.356  -1.004
What is your current occupation_Working Professional 2.5096 0.197 12.764 0.000  2.124   2.895
             Last Notable Activity_Modified     -0.6910    0.083   -8.277  0.000  -0.855  -0.527
          Last Notable Activity_Unreachable      1.7599    0.598    2.945  0.003   0.589   2.931
         Last Notable Activity_Unsubscribed      1.4167    0.512    2.765  0.006   0.413   2.421
```

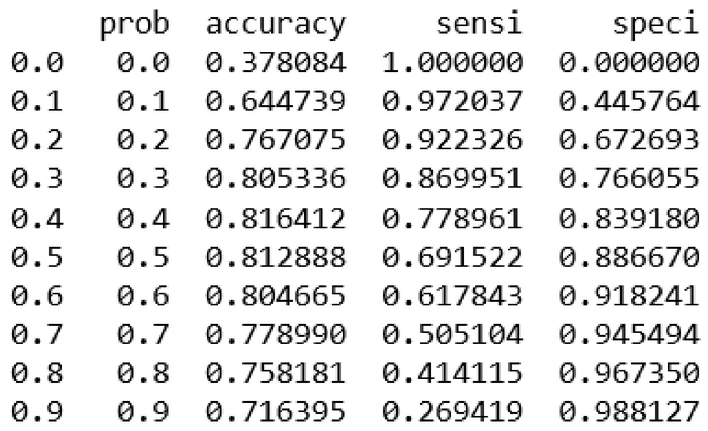
The dummy variables which impact the most on the conversion rate are ‘Lead Origin_Lead Add Form’ , ‘Last Activity_Had a Phone Conversation’ and ‘What is your current occupation_Working Professional’, which are the categorical variables having the highest coefficient.

Business Problems:

- X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.

Since the company wants to make phone calls as much as possible, so in this case even if we identify some leads which are not going to convert as hot lead that won't make any difference since the company is trying to reach maximum leads as possible. So the company need to increase the cutoff for the model they have built for predicting the hot leads. In this case the model needs to have less false negative count and higher true positive count. So ideally the company should focus on the sensitivity of the model and would prefer a higher sensitivity. The sensitivity will increase with increasing cutoff value. We had the below plot for sensitivity, accuracy and specificity earlier:





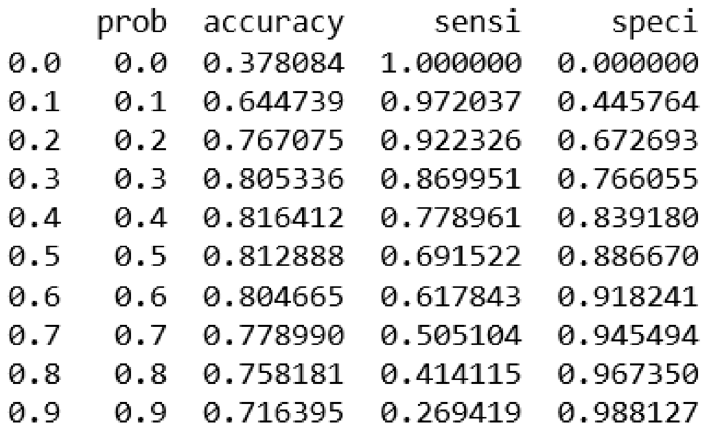
```
     prob  accuracy     sensi     speci
0.0   0.0  0.378084  1.000000  0.000000
0.1   0.1  0.644739  0.972037  0.445764
0.2   0.2  0.767075  0.922326  0.672693
0.3   0.3  0.805336  0.869951  0.766055
0.4   0.4  0.816412  0.778961  0.839180
0.5   0.5  0.812888  0.691522  0.886670
0.6   0.6  0.804665  0.617843  0.918241
0.7   0.7  0.778990  0.505104  0.945494
0.8   0.8  0.758181  0.414115  0.967350
0.9   0.9  0.716395  0.269419  0.988127
```

From the above table and plot we can see that for cutoff value around 0.3 we are getting sensitivity of around 87% without compromising accuracy (80%). So the company can predict based on this and this will give them a lot of leads whom they can now contact for trying to convert them to paying customers.

Business Problems:

- Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company’s aim is to not make phone calls unless it’s extremely necessary, i.e. they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.

In this case the company doesn't want to make useless phone calls , so the model which is used to predict leads needs to have low FPR(false positivity rate). The cutoff values should be high so that false positives are avoided. Since FPR = 1 - Specificity, the Specificity of the model needs to be high in this scenario. The specificity of the model will increase with increasing in cutoff value.



```
     prob  accuracy     sensi     speci
0.0   0.0  0.378084  1.000000  0.000000
0.1   0.1  0.644739  0.972037  0.445764
0.2   0.2  0.767075  0.922326  0.672693
0.3   0.3  0.805336  0.869951  0.766055
0.4   0.4  0.816412  0.778961  0.839180
0.5   0.5  0.812888  0.691522  0.886670
0.6   0.6  0.804665  0.617843  0.918241
0.7   0.7  0.778990  0.505104  0.945494
0.8   0.8  0.758181  0.414115  0.967350
0.9   0.9  0.716395  0.269419  0.988127
```

For cutoff value 0.6, the specificity is around 0.91 and accuracy is 0.8. So based on this cutoff the company can pursue the identified hot leads(converted value 1) and it will ensure that the company is making as less phone calls as possible.

Analysis and Recommendations:

- With 80% conversion rate, the model has been built with the conversion probability as 0.52 and taking Lead Origin , Last Activity and Current Occupation of the lead as the most important factors behind the conversion.
- If X Education focusses most on these factors, they will be able to increase their hot leads count as also shown in the exploratory data analysis.
- With this model, the company will be able to meet their 80% conversion rate successfully.
- By marketing more in the most trending Lead Origin (ie Google), or communicating more in the SMS mode amongst Unemployed people will help their company grow more.
- The approach should be made more to the unemployed people with exciting deals so that the course gets it worth.
- The advertisement should be made in a tempting way in Google more so that people feel more likely to click on it and register on the landing page.
- The exciting deals can be communicated via SMS, so that they can go through it in their leisure time. By calls, mostly people feel reluctant to pay any attention and can go out of opportunity.

Thank You!