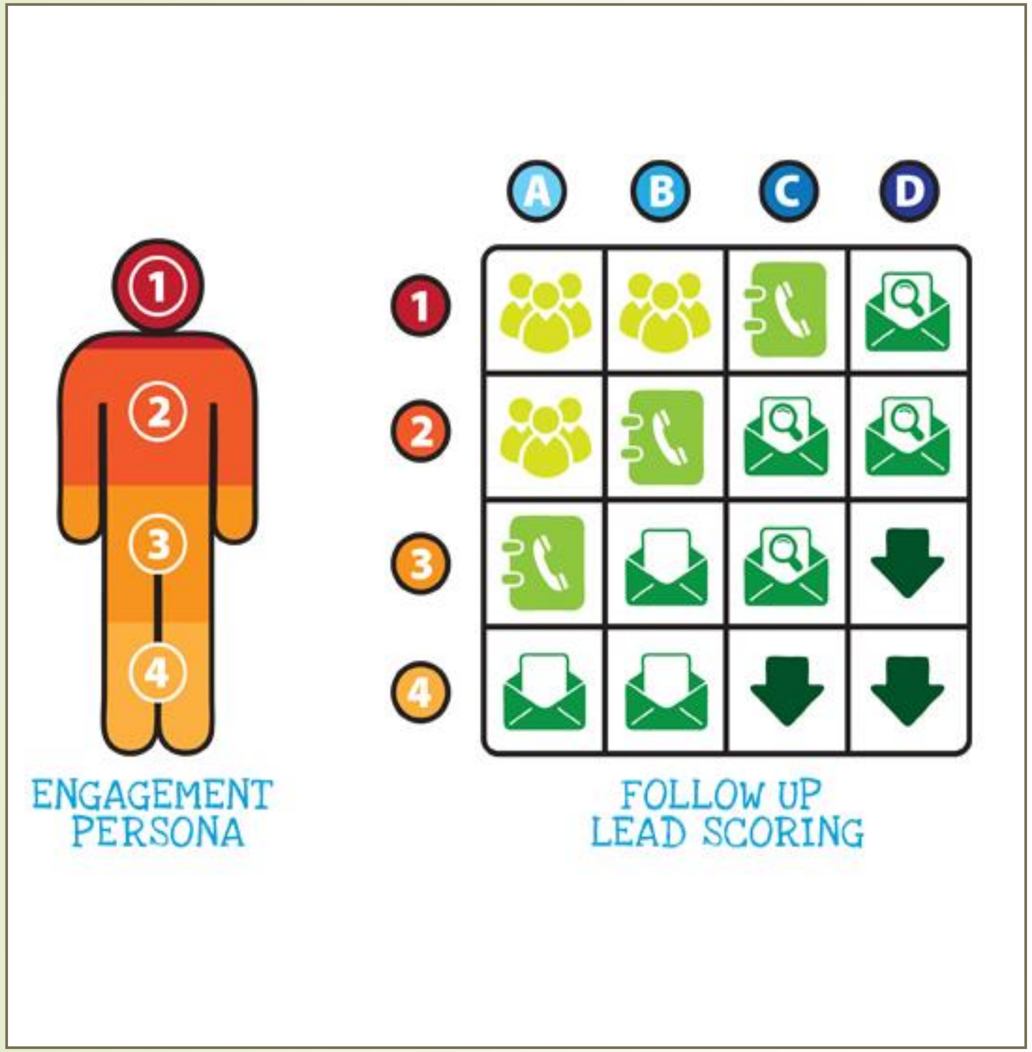


X Education Lead Score Analysis

(Group Case Study)

By

Akansha Khandelwal

And

Neelanjan Basu

13/10/2021

Agenda

Topic	Section
Business Objective	3
Dataset Parameters	4
Univariate Analysis	5 – 21
Bivariate Analysis	22 – 25
ML Model Build and Evaluation	26 - 29
Observations and Recommendations	30

Business Objective / Problem Statement

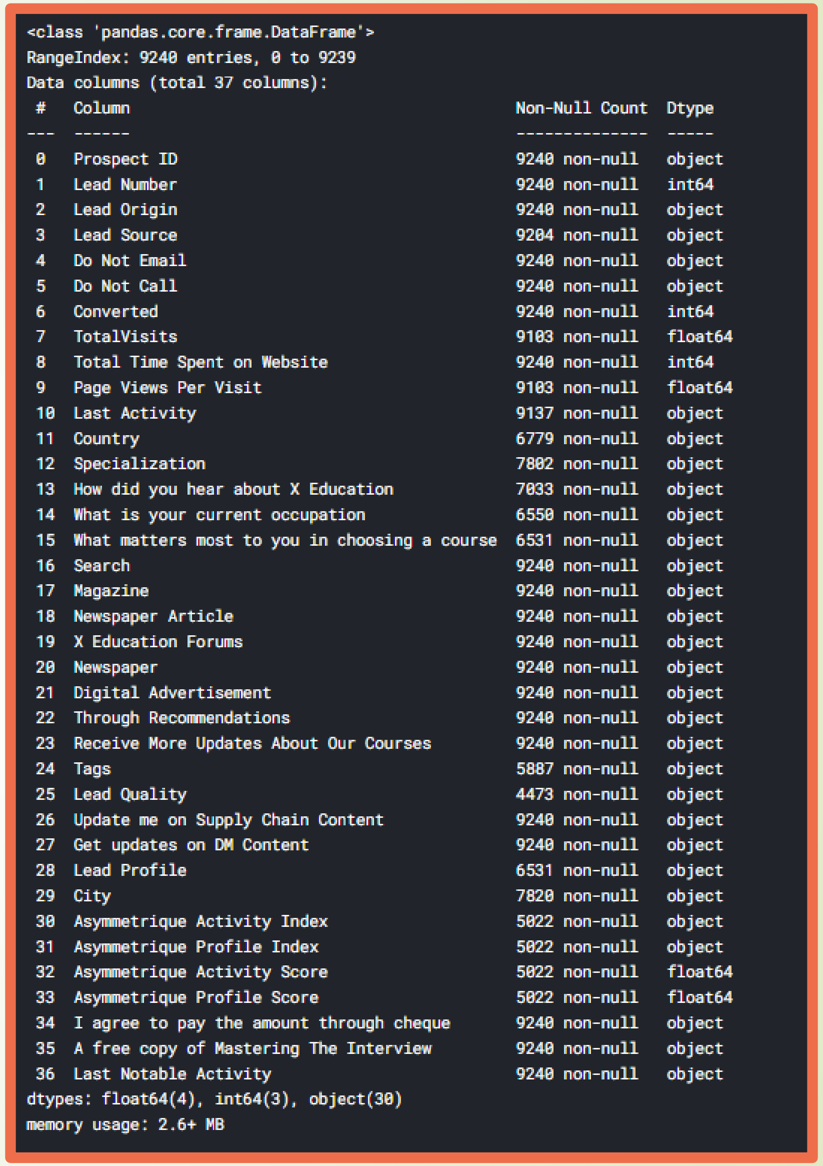
- X Education sells online courses to industry professionals based on the leads they receive via different channels the company has used to market. If a user fills up a form, the company treats that as a lead and tries to convert the Lead into a customer via various channels.
- With the current strategy, the company associates get into touch with the Leads and convert only 38-39% of the Lead population, which is low.
- Therefore, as part of the assignment, the company has requested an ML model that can successfully determine the possible customers with > 80% accuracy and possibly save the company from contacting every Lead via different channels.
- In summary, the goal of the Case Study is to:
  - ✓ Understand the relationship between the variables the company have captured via different channels
  - ✓ Build an ML model to assign a Lead Score between 0 – 100. One hundred would mean that the Lead is hot.
  - ✓ The ML model should identify users with 80% accuracy who is willing to take up any course.
  - ✓ Identify the top variables which contribute to the model

Dataset and its Parameters

As part of the case study two datasets were provided as follows

- ***'Leads.csv'*** contains all the information collection for a users at the time they filled up the application form application.

- ***'Leads Data Dictionary.csv'*** *contains the definition of the columns in the data set*



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count  Dtype
---  ------                                         --------------  -----
 0   Prospect ID                                    9240 non-null   object
 1   Lead Number                                    9240 non-null   int64
 2   Lead Origin                                    9240 non-null   object
 3   Lead Source                                    9204 non-null   object
 4   Do Not Email                                   9240 non-null   object
 5   Do Not Call                                    9240 non-null   object
 6   Converted                                      9240 non-null   int64
 7   TotalVisits                                    9103 non-null   float64
 8   Total Time Spent on Website                    9240 non-null   int64
 9   Page Views Per Visit                           9103 non-null   float64
 10  Last Activity                                  9137 non-null   object
 11  Country                                        6779 non-null   object
 12  Specialization                                 7802 non-null   object
 13  How did you hear about X Education             7033 non-null   object
 14  What is your current occupation                6550 non-null   object
 15  What matters most to you in choosing a course  6531 non-null   object
 16  Search                                         9240 non-null   object
 17  Magazine                                       9240 non-null   object
 18  Newspaper Article                              9240 non-null   object
 19  X Education Forums                             9240 non-null   object
 20  Newspaper                                      9240 non-null   object
 21  Digital Advertisement                          9240 non-null   object
 22  Through Recommendations                        9240 non-null   object
 23  Receive More Updates About Our Courses         9240 non-null   object
 24  Tags                                           5887 non-null   object
 25  Lead Quality                                   4473 non-null   object
 26  Update me on Supply Chain Content              9240 non-null   object
 27  Get updates on DM Content                      9240 non-null   object
 28  Lead Profile                                   6531 non-null   object
 29  City                                           7820 non-null   object
 30  Asymmetrique Activity Index                    5022 non-null   object
 31  Asymmetrique Profile Index                     5022 non-null   object
 32  Asymmetrique Activity Score                    5022 non-null   float64
 33  Asymmetrique Profile Score                     5022 non-null   float64
 34  I agree to pay the amount through cheque       9240 non-null   object
 35  A free copy of Mastering The Interview         9240 non-null   object
 36  Last Notable Activity                          9240 non-null   object
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB
```

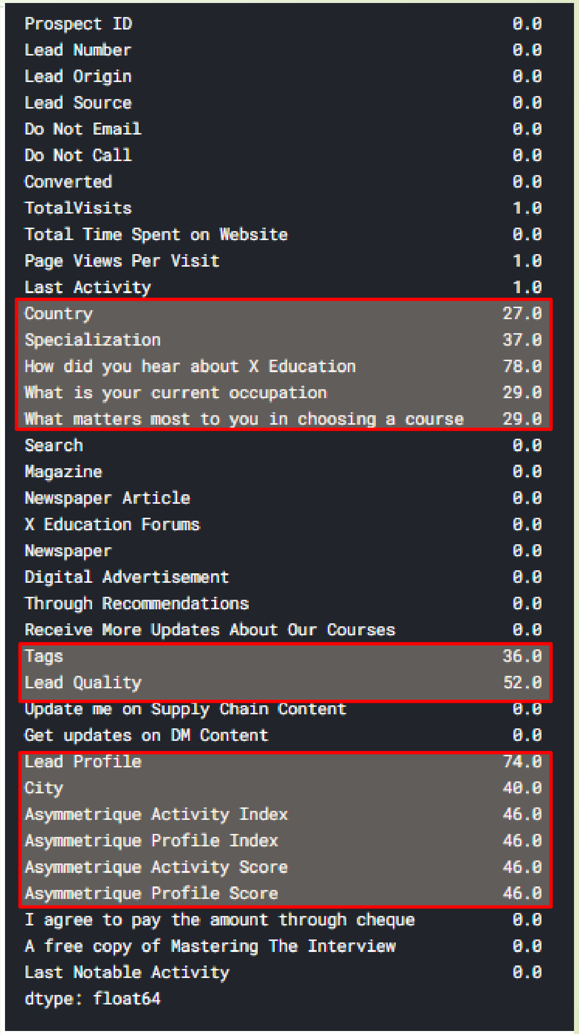
Univariate Analysis – Null Value Treatment

'Leads.csv' contains around **9240 rows x 37 columns** , which we would use for EDA and then feed the relevant columns to the model.

It's observed that the many columns contain 'Select' as not fields are mandatory. Therefore, as part we replaced the 'Select' with Null

Columns with very high % NULL values and there is no way to derive the values from other columns was dropped.

- 'How did you hear about X Education' - 78% Null Values
- 'Lead Profile' - 74% Null values
- 'Lead Quality' - 52% Null Values
- 'Asymmetrique Activity Index' - 46% Null Values
- 'Asymmetrique Profile Index' - 46% Null Values
- 'Asymmetrique Activity Score' - 46% Null Values
- 'Asymmetrique Profile Score' - 46% Null Values



```
Prospect ID                                       0.0
Lead Number                                       0.0
Lead Origin                                       0.0
Lead Source                                       0.0
Do Not Email                                      0.0
Do Not Call                                       0.0
Converted                                         0.0
TotalVisits                                       1.0
Total Time Spent on Website                       0.0
Page Views Per Visit                              1.0
Last Activity                                     1.0
Country                                          27.0
Specialization                                   37.0
How did you hear about X Education               78.0
What is your current occupation                  29.0
What matters most to you in choosing a course    29.0
Search                                            0.0
Magazine                                          0.0
Newspaper Article                                 0.0
X Education Forums                                0.0
Newspaper                                         0.0
Digital Advertisement                             0.0
Through Recommendations                           0.0
Receive More Updates About Our Courses            0.0
Tags                                             36.0
Lead Quality                                     52.0
Update me on Supply Chain Content                 0.0
Get updates on DM Content                         0.0
Lead Profile                                     74.0
City                                             40.0
Asymmetrique Activity Index                      46.0
Asymmetrique Profile Index                       46.0
Asymmetrique Activity Score                      46.0
Asymmetrique Profile Score                       46.0
I agree to pay the amount through cheque          0.0
A free copy of Mastering The Interview            0.0
Last Notable Activity                             0.0
dtype: float64
```

Univariate Analysis – Null Value Treatment

For the columns having **NULL values < 45%,** it was not dropped straight away as few columns may be imputed/derived for the same column or from other column.

- 'City' - 40% Null Values

  There is a 'Other Cities' Category in the City column therefore all City Null values was replaced with 'Other City'

- 'Specialization' -37% Null Values

  All Specialization Null Values was replaced with 'Not Sure'

- 'Tags' - 36% Null Values

  'Tags' is a columns which is filled by the company employee after they spoke with the Lead. But as per the case study we should be able to determine whom the company employee should call. Hence the ML model should not depend on 'Tags'. Hence dropping 'Tags'

  Same is the case for 'Lead Quality

- 'What is your current occupation' -29% Null Values

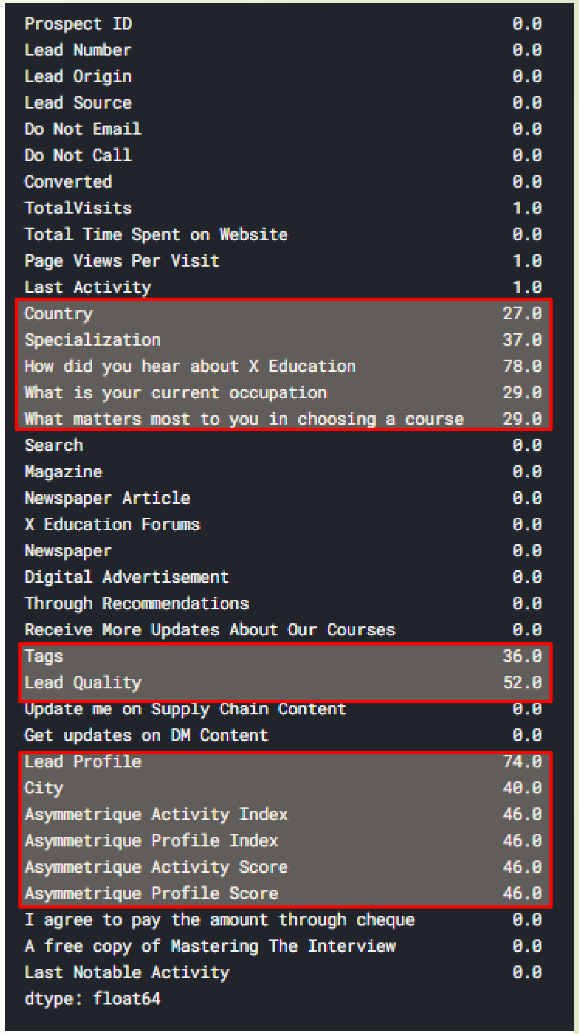
  For the current occupation we have replaced null values with 'Other'

- 'What matters most to you in choosing a course' - 29% Null Values

  There is no much Variance in the column therefore will drop the column

- 'Country' - 27% Null Values

  Country value can be derived from the City values. Also, if the city value is 'Other Cities' we have replaced with 'Other Country'



```
Prospect ID                                       0.0
Lead Number                                       0.0
Lead Origin                                       0.0
Lead Source                                       0.0
Do Not Email                                      0.0
Do Not Call                                       0.0
Converted                                         0.0
TotalVisits                                       1.0
Total Time Spent on Website                       0.0
Page Views Per Visit                              1.0
Last Activity                                     1.0
Country                                          27.0
Specialization                                   37.0
How did you hear about X Education               78.0
What is your current occupation                  29.0
What matters most to you in choosing a course    29.0
Search                                            0.0
Magazine                                          0.0
Newspaper Article                                 0.0
X Education Forums                                0.0
Newspaper                                         0.0
Digital Advertisement                             0.0
Through Recommendations                           0.0
Receive More Updates About Our Courses            0.0
Tags                                             36.0
Lead Quality                                     52.0
Update me on Supply Chain Content                 0.0
Get updates on DM Content                         0.0
Lead Profile                                     74.0
City                                             40.0
Asymmetrique Activity Index                      46.0
Asymmetrique Profile Index                       46.0
Asymmetrique Activity Score                      46.0
Asymmetrique Profile Score                       46.0
I agree to pay the amount through cheque          0.0
A free copy of Mastering The Interview            0.0
Last Notable Activity                             0.0
dtype: float64
```

Univariate Analysis – Null Value Treatment

For the columns having **NULL values <= 1%,** it was not dropped straight away as few columns may be imputed/derived for the same column or from other column.

- 'Total Visits' - 1% Null Values

  Since the number null values is less therefore, imputed the vales with median of the column

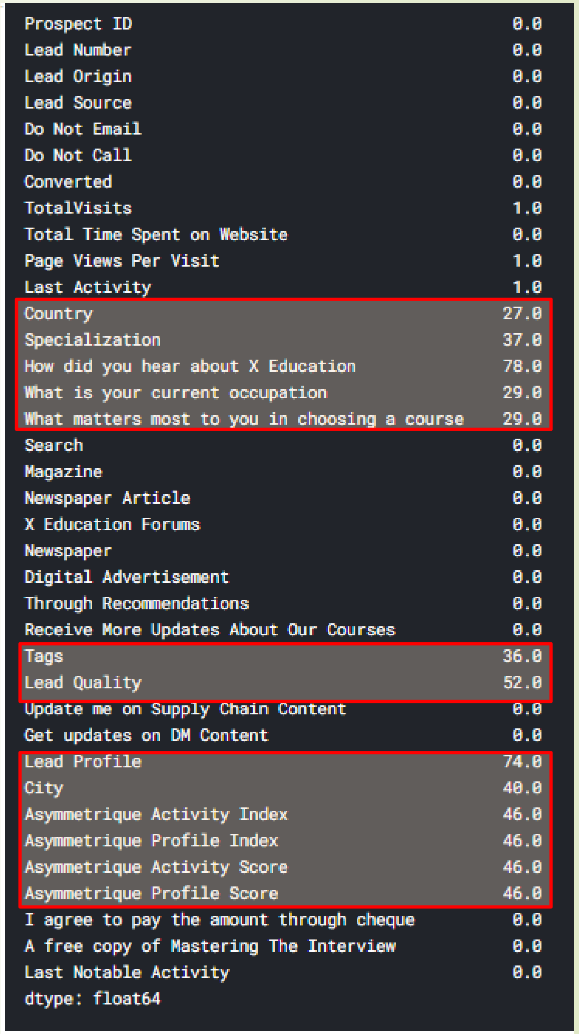
- 'Page Views Per Visit' - 1% Null Values

  Since the number null values is less therefore, imputed the vales with median of the column

- Last Activity - 1% Null Values

  Since the number null values is less therefore, we will drop the values

Post all the null value treatment there is no Null Value in the dataset now.



```
Prospect ID                                       0.0
Lead Number                                       0.0
Lead Origin                                       0.0
Lead Source                                       0.0
Do Not Email                                      0.0
Do Not Call                                       0.0
Converted                                         0.0
TotalVisits                                       1.0
Total Time Spent on Website                       0.0
Page Views Per Visit                              1.0
Last Activity                                     1.0
Country                                          27.0
Specialization                                   37.0
How did you hear about X Education               78.0
What is your current occupation                  29.0
What matters most to you in choosing a course    29.0
Search                                            0.0
Magazine                                          0.0
Newspaper Article                                 0.0
X Education Forums                                0.0
Newspaper                                         0.0
Digital Advertisement                             0.0
Through Recommendations                           0.0
Receive More Updates About Our Courses            0.0
Tags                                             36.0
Lead Quality                                     52.0
Update me on Supply Chain Content                 0.0
Get updates on DM Content                         0.0
Lead Profile                                     74.0
City                                             40.0
Asymmetrique Activity Index                      46.0
Asymmetrique Profile Index                       46.0
Asymmetrique Activity Score                      46.0
Asymmetrique Profile Score                       46.0
I agree to pay the amount through cheque          0.0
A free copy of Mastering The Interview            0.0
Last Notable Activity                             0.0
dtype: float64
```

For our benefit during the EDA, we have categorized the individual dataset columns as below.

Dataset: '***Leads.csv***'

- **Lead Unique Identifiers**
  - *Prospect ID*
  - *Lead Number*

- **Categorical Features**
  - *Lead Origin*
  - *Lead Source*
  - *Do Not Email*
  - *Do Not Call*
  - *Country*
  - *Specialization*
  - *What is your current occupation*
  - *Search / Magazine/' Newspaper Article' / 'X Education Forums' / 'Newspaper' / 'Digital Advertisement / Through Recommendation*
  - *'Receive More Updates About Our Courses'*
  - *Update me on Supply Chain Content*
  - *Get updates on DM Content*
  - *City*
  - *I agree to pay the amount through cheque*
  - *A free copy of Mastering The Interview*
  - *Last Notable Activity*

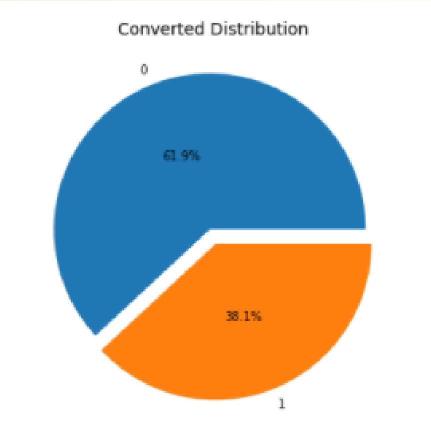
- **Numerical Features**
  - *Converted [Target Variable]*
  - *Total Visits*
  - *Total Time Spent on Website*
  - *Page Views Per Visit*

Univariate Analysis - Lead Unique Identifiers

- The columns under ‘**Lead Unique Identifiers**’ identifies the user’s uniquely therefore its of no use to us while building the model therefore have dropped the columns

- **Lead Unique Identifiers column:**
  - *Prospect ID*
  - *Lead Number*

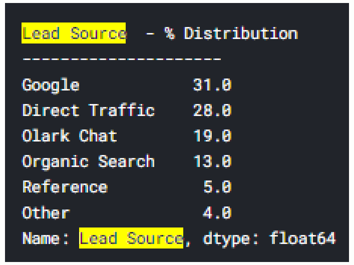
- Currently the it seems only 38% of the Leads we can convert to customers. **Target Variable distribution** is as below



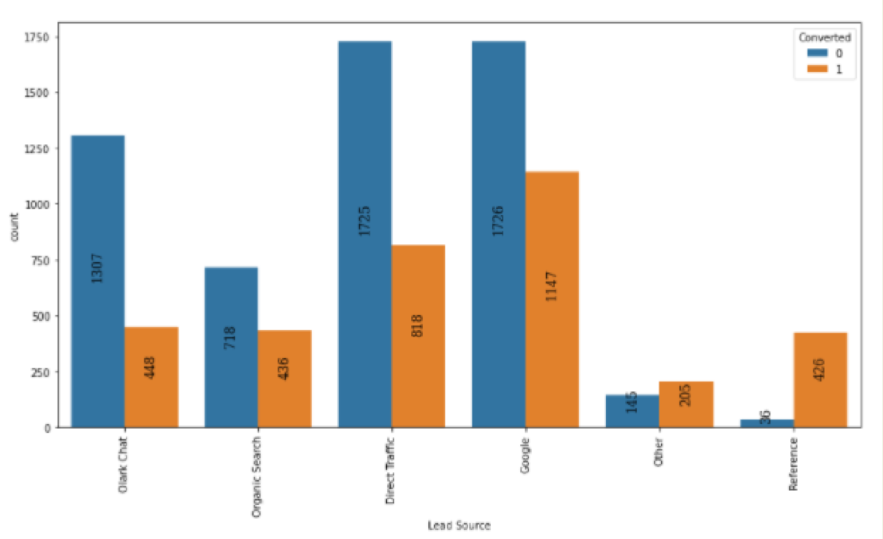
Univariate Analysis - Categorical Features

Lead Source :

- ✓ Contains 0 % null values
- ✓ To minimize the number of category, any Lead Source category below <= 1% is converted to **'Others'**

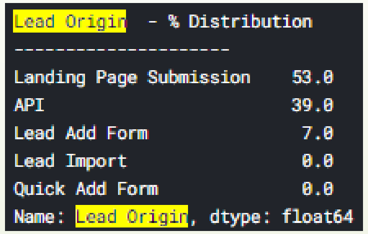


```
Lead Source  - % Distribution
---------------------
Google           31.0
Direct Traffic   28.0
Olark Chat       19.0
Organic Search   13.0
Reference         5.0
Other             4.0
Name: Lead Source, dtype: float64
```

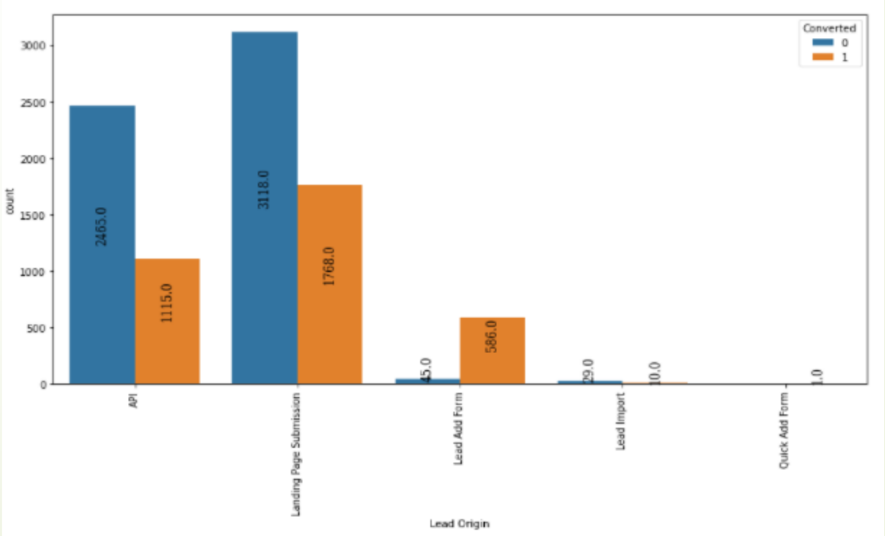


Lead Origin :

- ✓ Contains 0 % null values
- ✓ Users who landed on the landing Page seems to have converted more than any other Lead Origin



```
Lead Origin  - % Distribution
---------------------
Landing Page Submission    53.0
API                        39.0
Lead Add Form               7.0
Lead Import                 0.0
Quick Add Form              0.0
Name: Lead Origin, dtype: float64
```



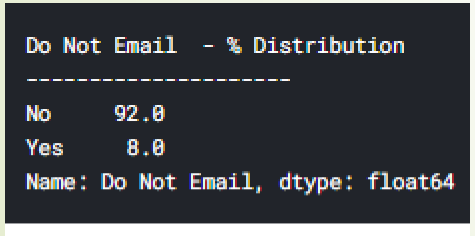
Univariate Analysis - Categorical Features

Do Not Email :

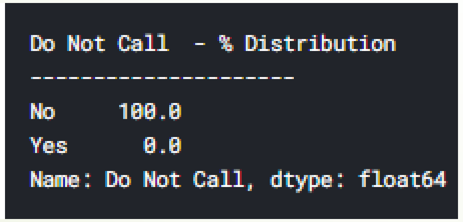
- ✓ Contains 0 % null values
- ✓ Majority users are happy to receive emails

Do Not Call :

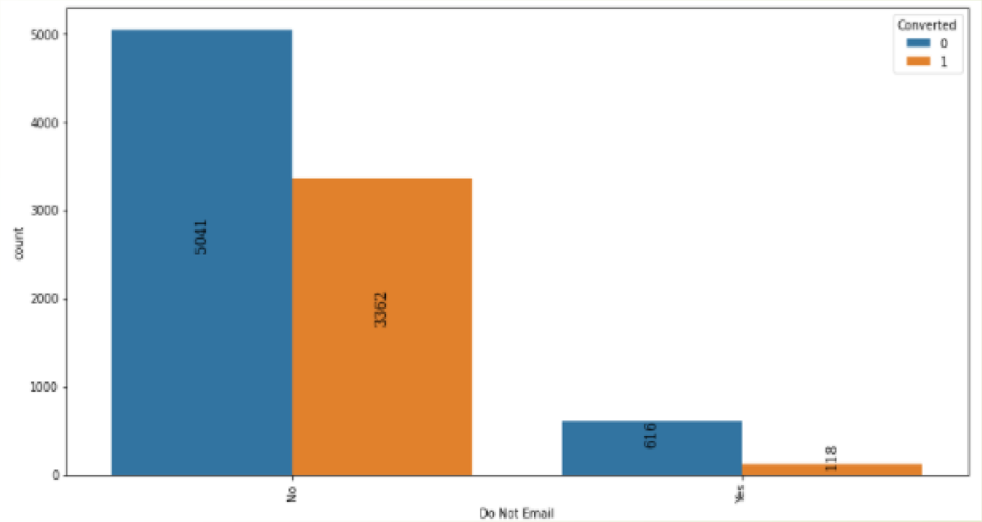
- ✓ Contains 0 % null values
- ✓ Majority users are happy to receive calls
- ✓ There is no variance in the column therefore will drop the column

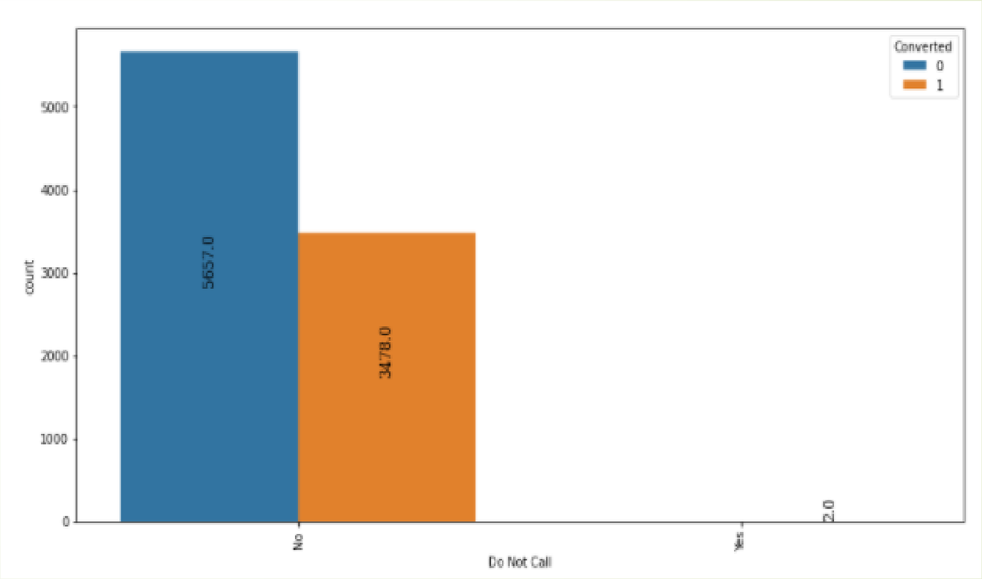


```
Do Not Email  - % Distribution
---------------------
No     92.0
Yes     8.0
Name: Do Not Email, dtype: float64
```



```
Do Not Call  - % Distribution
---------------------
No     100.0
Yes      0.0
Name: Do Not Call, dtype: float64
```

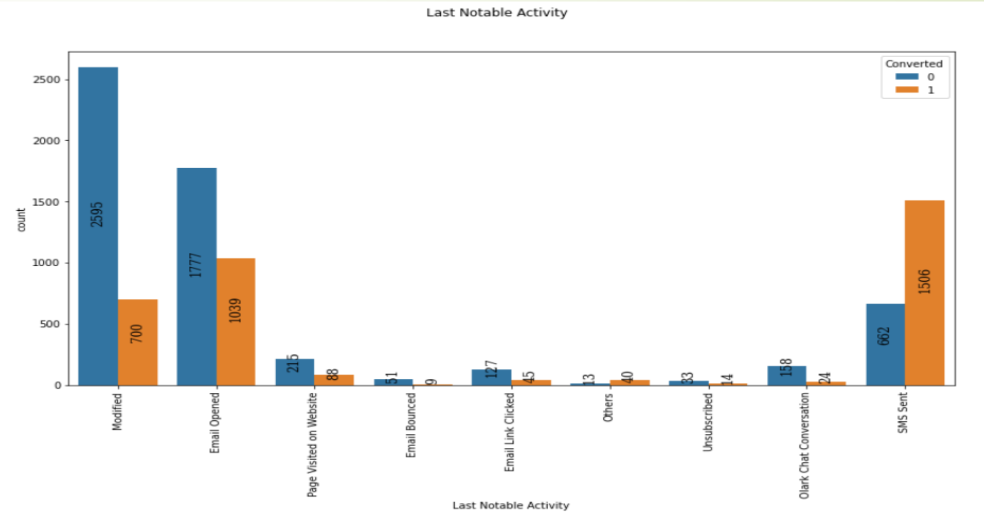




Univariate Analysis - Categorical Features

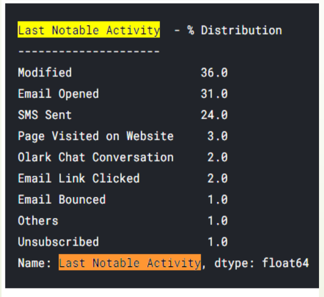
Last Activity :

- ✓ Contains 1 % null values. Therefore, dropped the null values

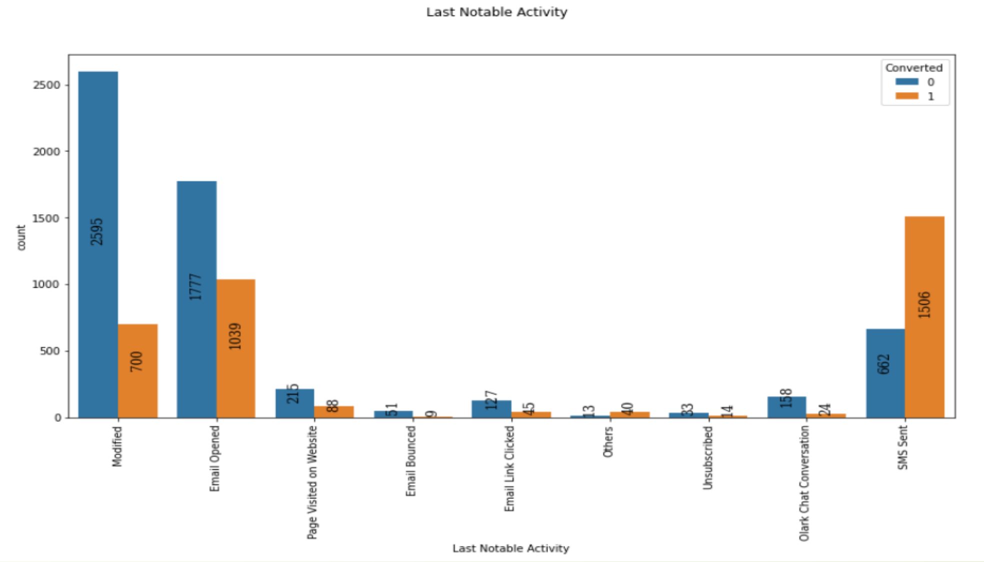


Last Notable Activity:

- ✓ Contains 0% null values .
- ✓ To minimize the number, of categories have clubbed categories <=2% to 'Others'



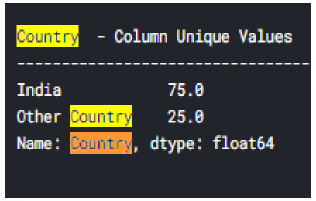
```
Last Notable Activity  - % Distribution
--------------------
Modified                   36.0
Email Opened               31.0
SMS Sent                   24.0
Page Visited on Website     3.0
Olark Chat Conversation     2.0
Email Link Clicked          2.0
Email Bounced               1.0
Others                      1.0
Unsubscribed                1.0
Name: Last Notable Activity, dtype: float64
```



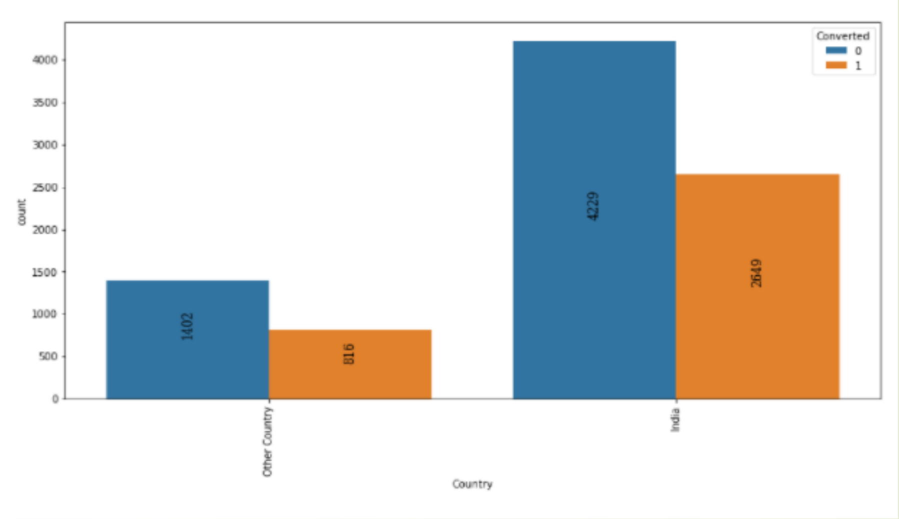
Univariate Analysis - Categorical Features

Country:

- ✓ Contains 27% null values.
- ✓ There is a lot of different countries reported. Therefore, whatever country we were able to derive from City we imputed else have changed imputed with 'Other Country'

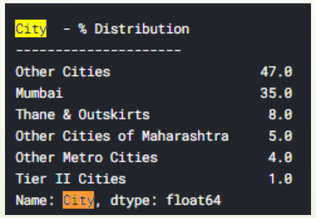


```
Country  - Column Unique Values
--------------------------------
India            75.0
Other Country    25.0
Name: Country, dtype: float64
```

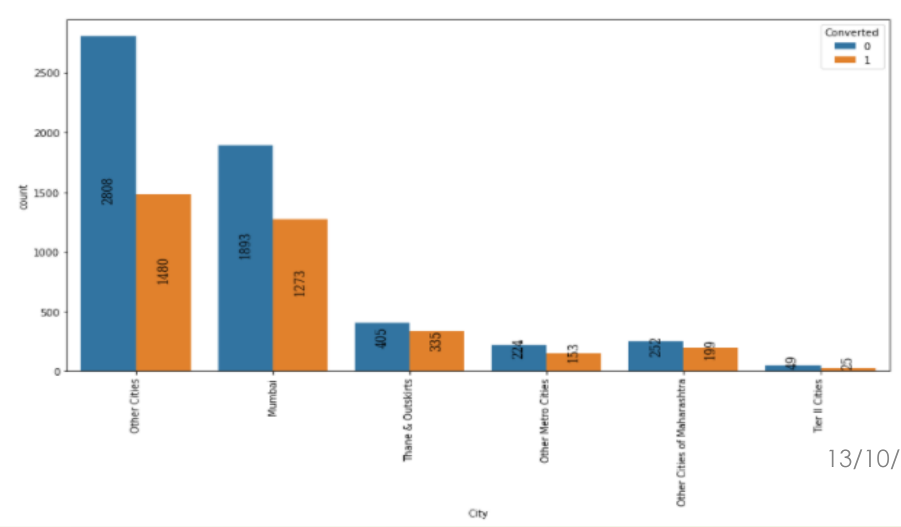


City:

- ✓ Contains 40% null values.
- ✓ Since there is no way to impute the City from Country, we have replaced all Nulls with 'Other Cities'



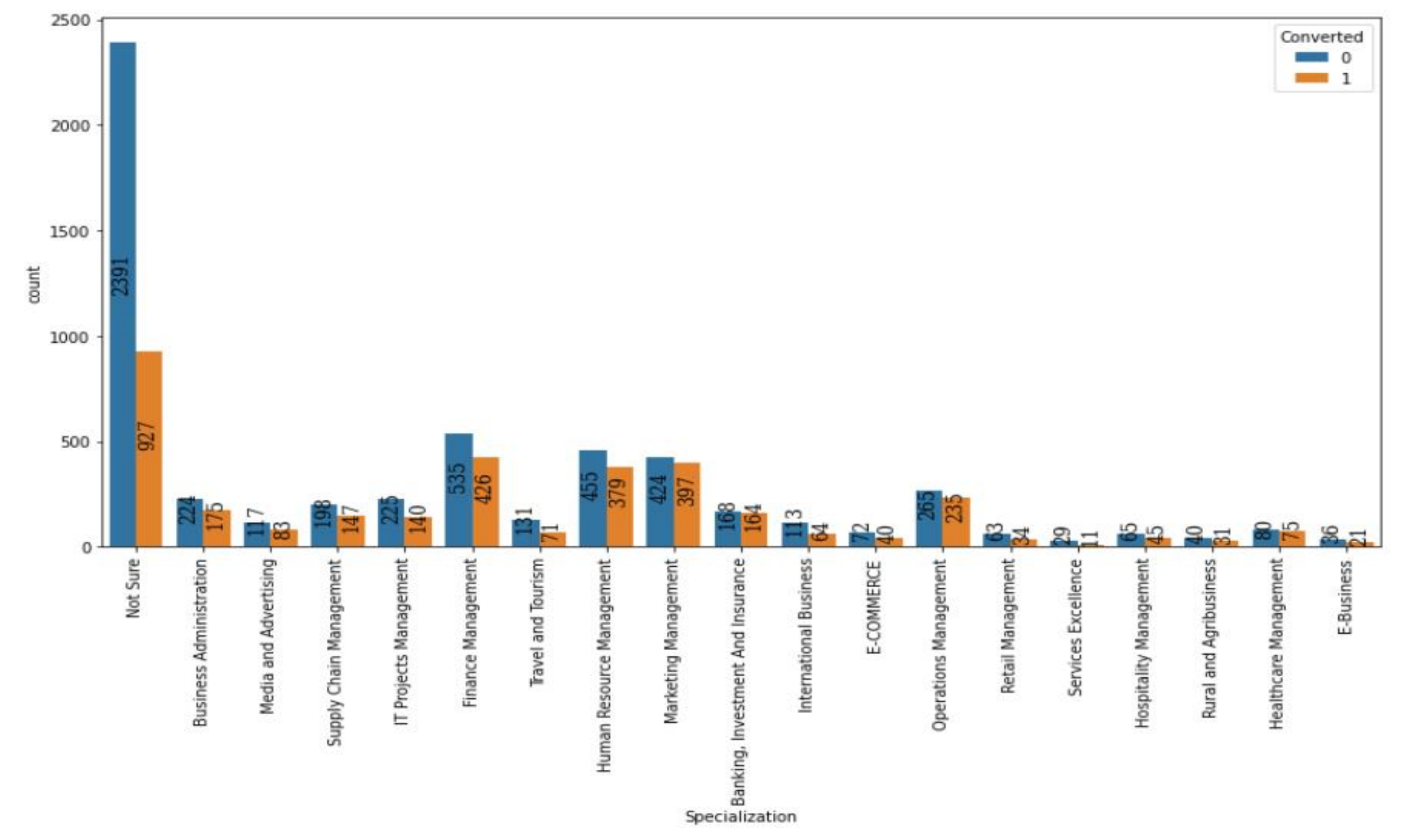
```
City  - % Distribution
---------------------
Other Cities                   47.0
Mumbai                         35.0
Thane & Outskirts               8.0
Other Cities of Maharashtra     5.0
Other Metro Cities              4.0
Tier II Cities                  1.0
Name: City, dtype: float64
```

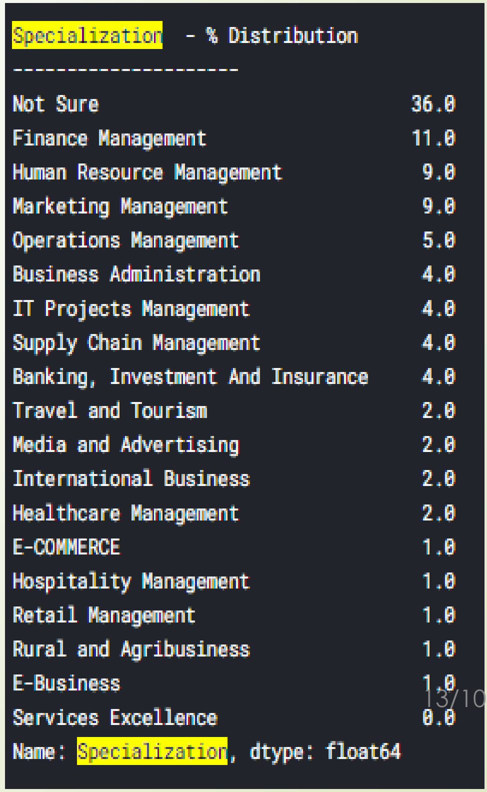


Univariate Analysis - Categorical Features

Specialization:

- ✓ Contains 37% null values.
- ✓ Imputed all null values of the column with ‘Not Sure’



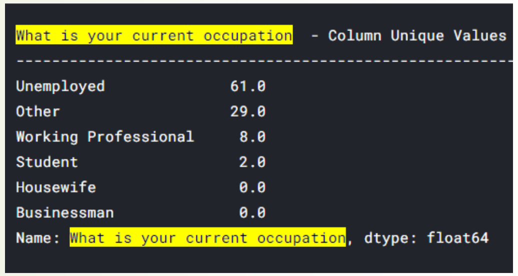


```
Specialization  - % Distribution
--------------------
Not Sure                             36.0
Finance Management                   11.0
Human Resource Management             9.0
Marketing Management                  9.0
Operations Management                 5.0
Business Administration               4.0
IT Projects Management                4.0
Supply Chain Management               4.0
Banking, Investment And Insurance     4.0
Travel and Tourism                    2.0
Media and Advertising                 2.0
International Business                2.0
Healthcare Management                 2.0
E-COMMERCE                            1.0
Hospitality Management                1.0
Retail Management                     1.0
Rural and Agribusiness                1.0
E-Business                            1.0
Services Excellence                   0.0
Name: Specialization, dtype: float64
```

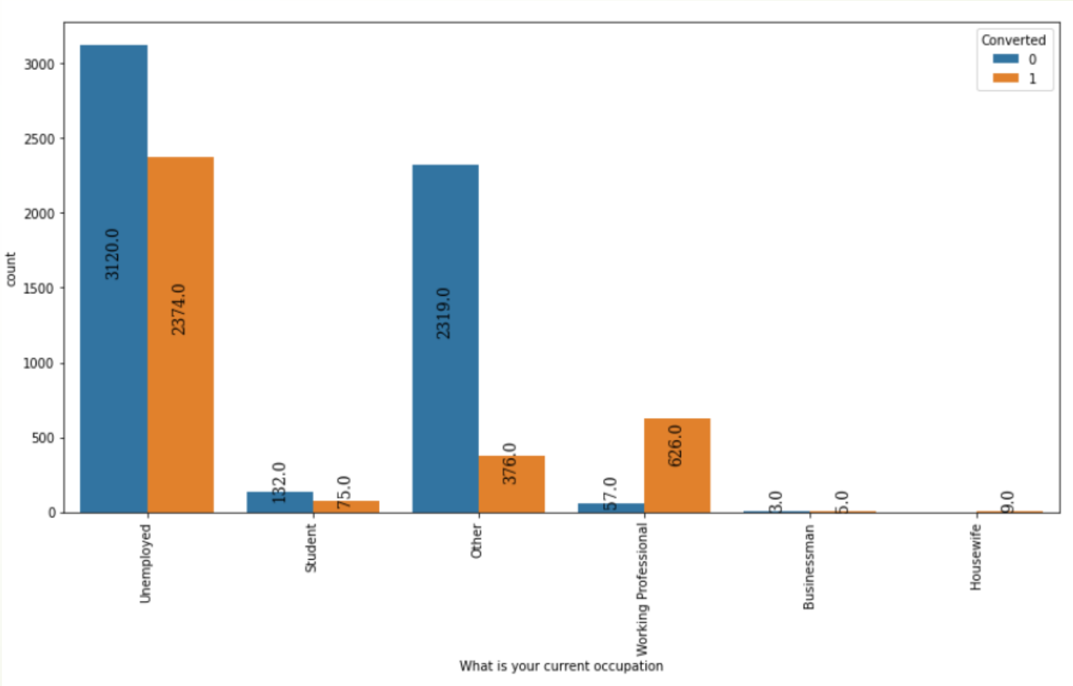
Univariate Analysis - Categorical Features

What is your current occupation:

- ✓ Contains 29% null values.
- ✓ Imputed all null values of the column with ‘Other’



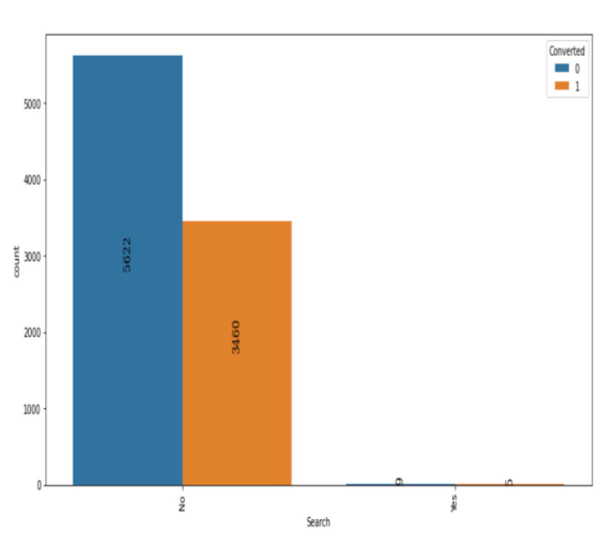
```
What is your current occupation   - Column Unique Values
-----------------------------------------------------------
Unemployed              61.0
Other                   29.0
Working Professional     8.0
Student                  2.0
Housewife                0.0
Businessman              0.0
Name: What is your current occupation, dtype: float64
```

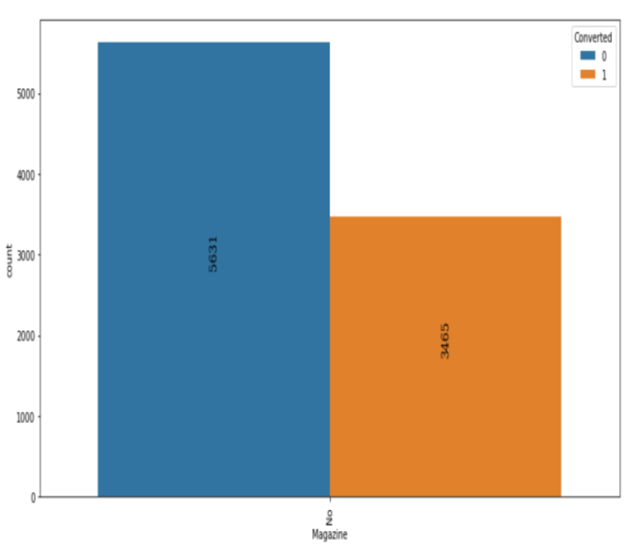


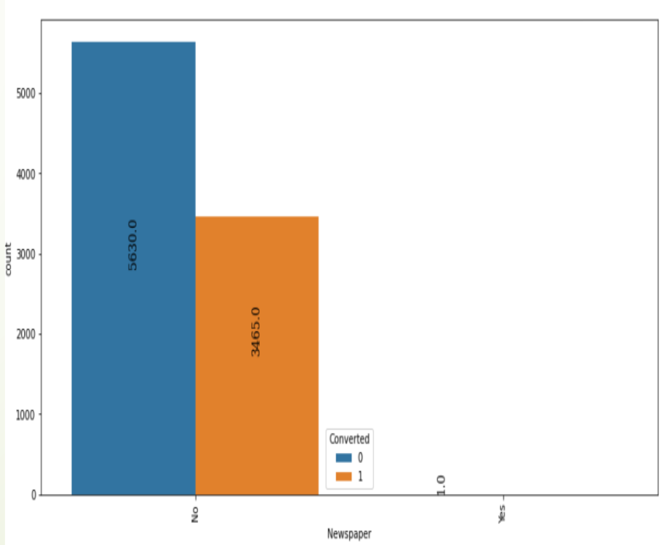
Univariate Analysis - Categorical Features

Search / Magazine/' Newspaper Article' / 'X Education Forums' / 'Newspaper' / 'Digital Advertisement / Through Recommendation / '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':

- ✓ Contains 0% null values.
- ✓ The columns above contains 0% variance (class imbalance) hence it will be of no use for the ML model. Will drop the columns

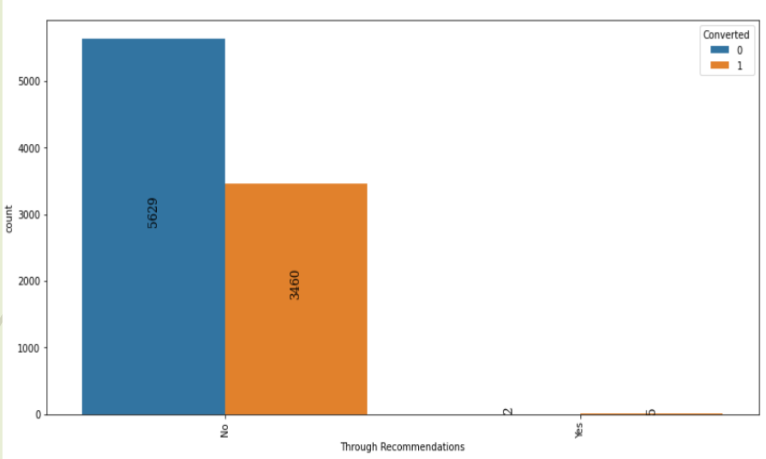


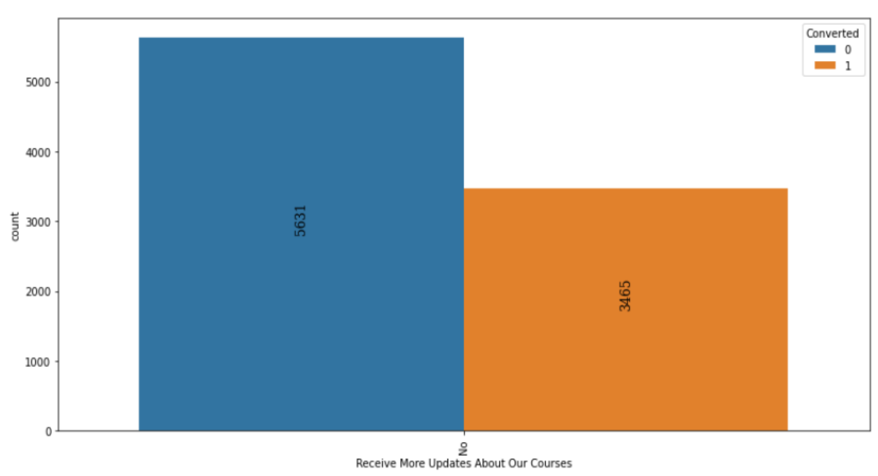


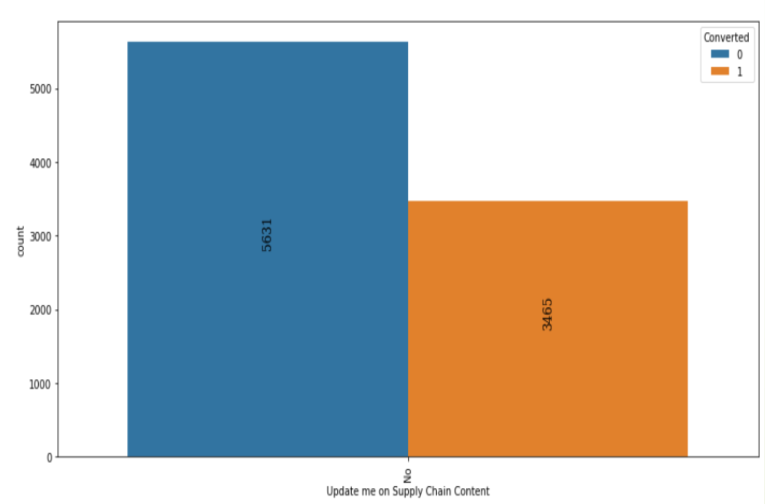


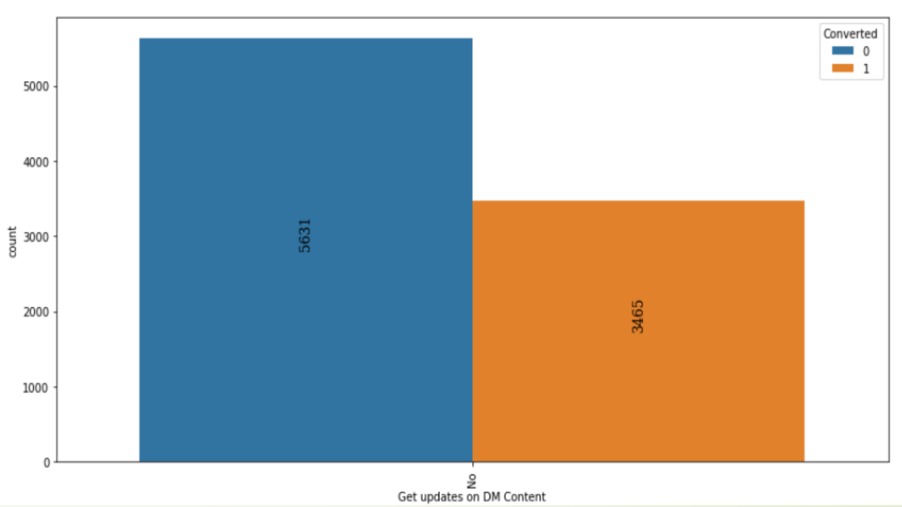
Continued…..

Univariate Analysis - Categorical Features



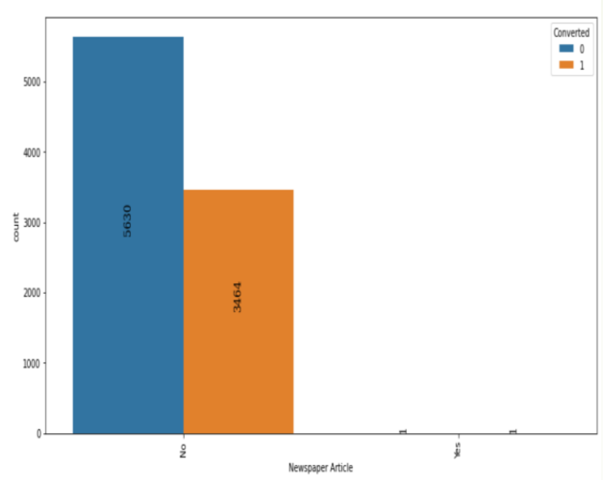


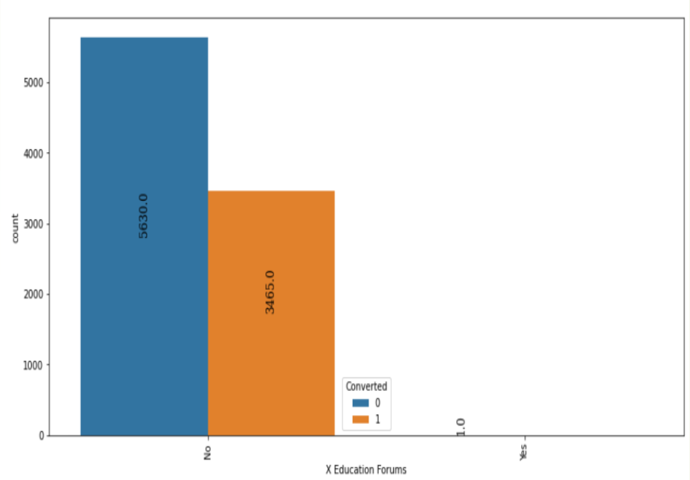


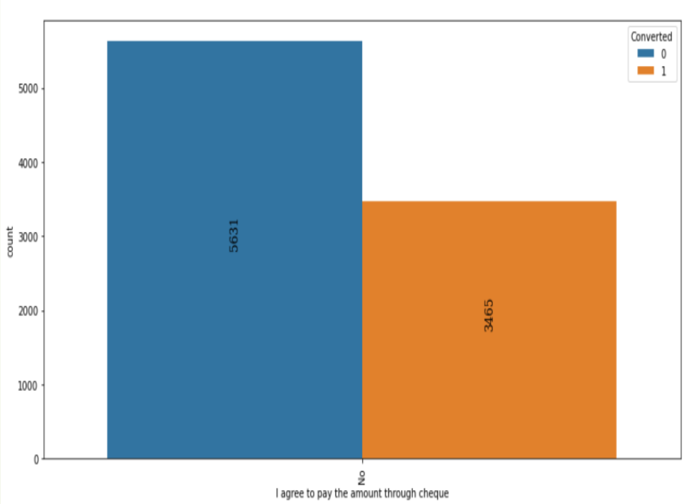


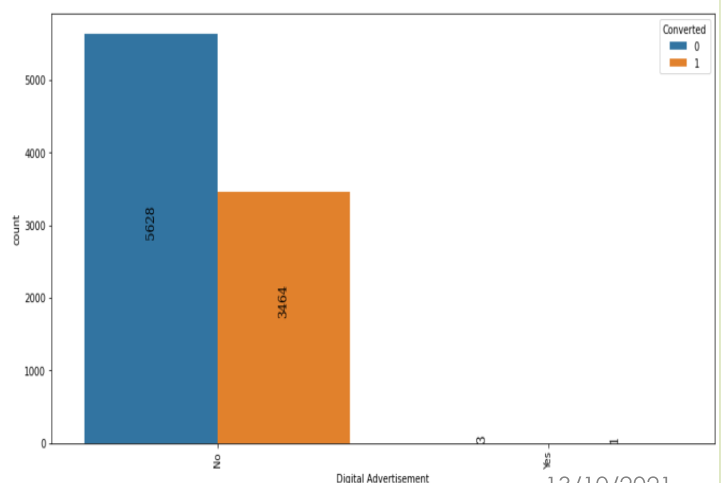
Continued…..

Univariate Analysis - Categorical Features





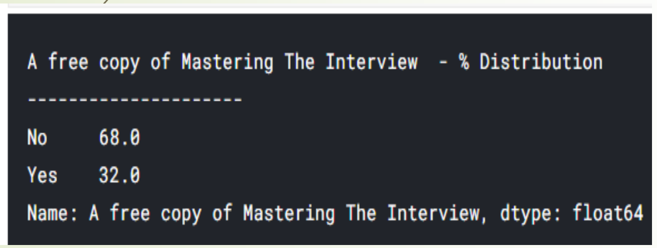




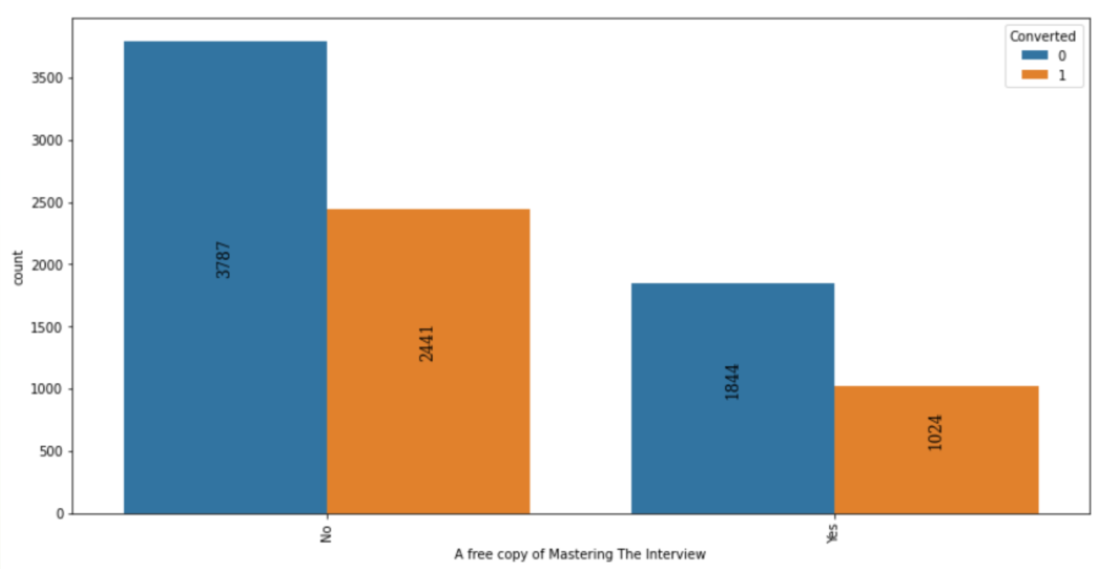
Univariate Analysis - Categorical Features

A free copy of Mastering The Interview:

- ✓ Contains 0% null values.
- ✓ It seems people who have not requested for a “Free copy of Mastering the Interview” have also signed up.



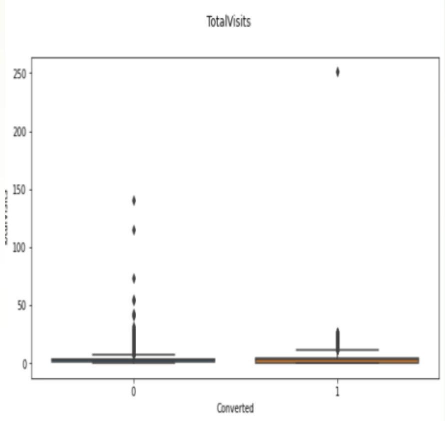
```
A free copy of Mastering The Interview  - % Distribution
--------------------
No     68.0
Yes    32.0
Name: A free copy of Mastering The Interview, dtype: float64
```

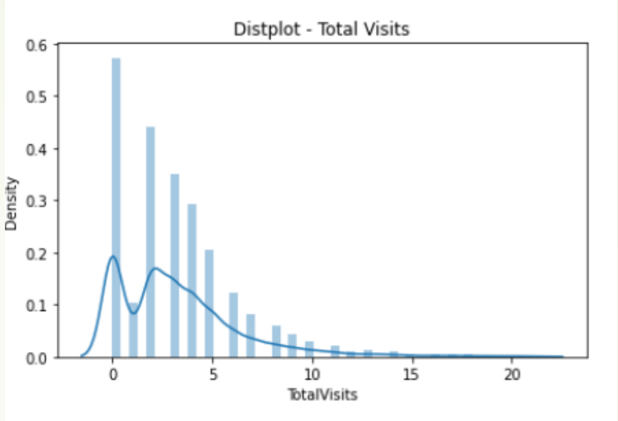


Univariate Analysis – Numerical Features

Total Visits:

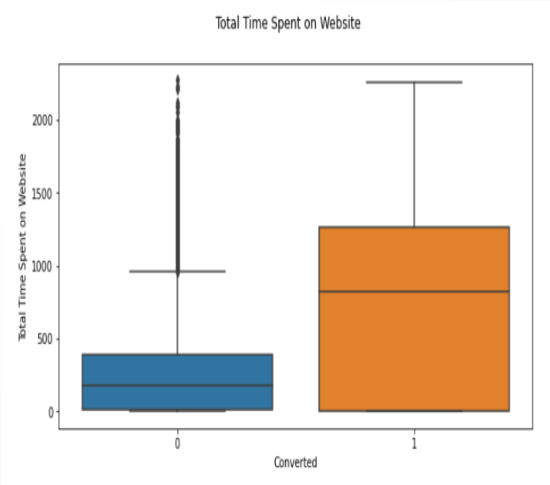
- ✓ Contains 1% null values.
- ✓ Since the number of null values is less that 1% will impute it with the median value
- ✓ Removed the rows where Total Visit > 99.5%

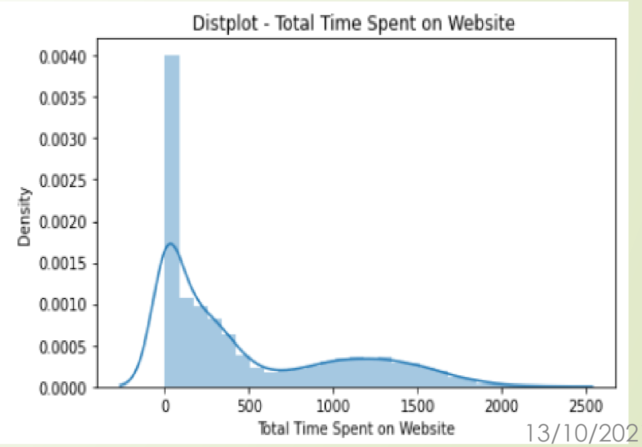

TotalVisits


Distplot - Total Visits

Total Time Spent on Website:

- ✓ Contains 0% null values.
- ✓ It seems people who have spent more time on the website is more likely to sign up for a course

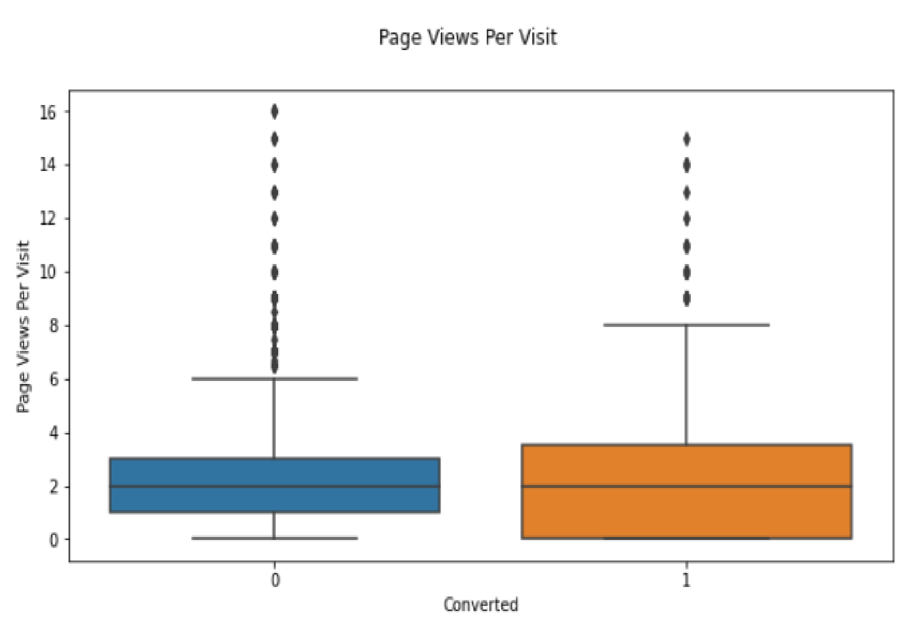

Total Time Spent on Website


Distplot - Total Time Spent on Website

Univariate Analysis – Numerical Features

Page Views Per Visit:

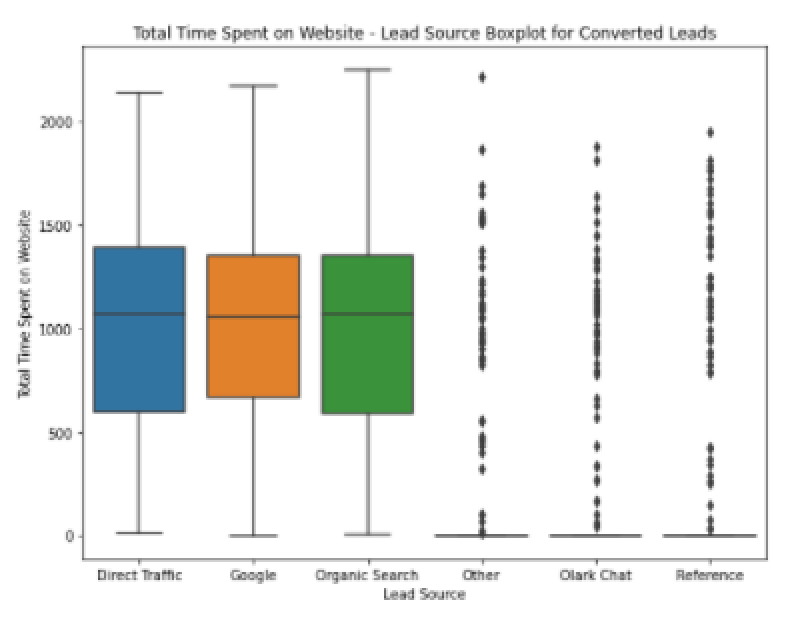
- ✓ Contains 1% null values.
- ✓ Since the number of null values is less that 1% will impute it with the median value.
- ✓ Cannot say the variable is having outliers

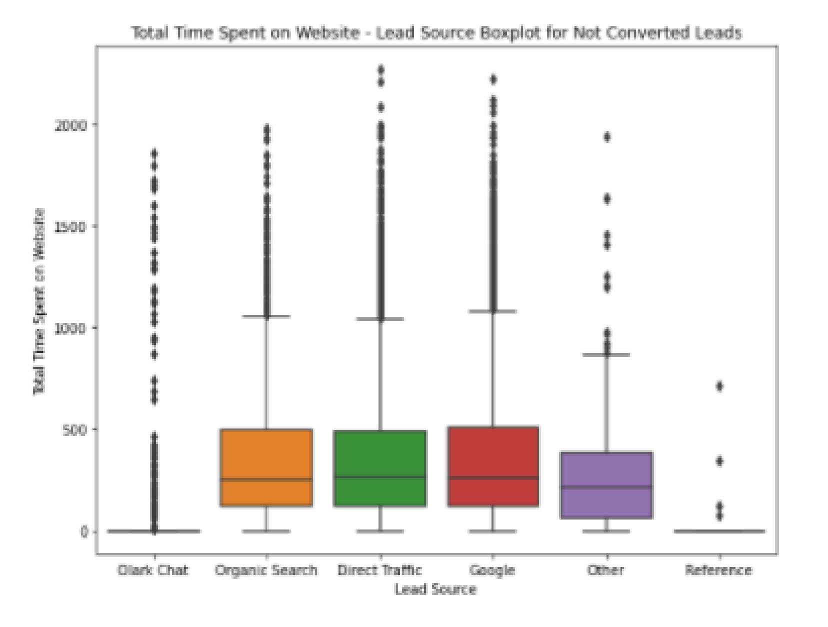


Bivariate Analysis

Total Time Spent on Website Vs Lead Source:

- ✓ For Converted Leads - Total Time Spent on Website via direct traffic / Google / Organic Search seems to have a similar range.
- ✓ For Non-Converted Leads - Total Time Spent on Website via direct traffic / Google / Organic Search seems to have a similar range.
- ✓ The 'Total Time Spent' median for Converted Leads via channels direct traffic / Google / Organic Search seems to be almost double compared to a Non-Converted Leads user



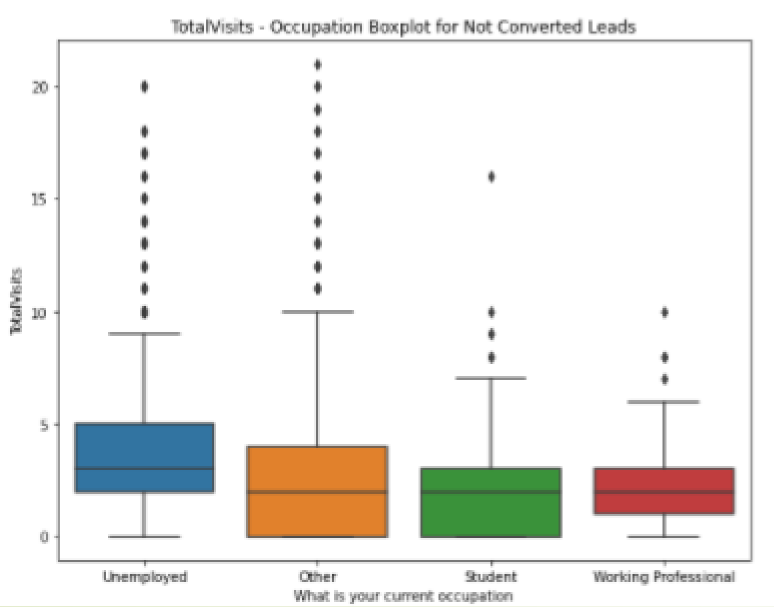


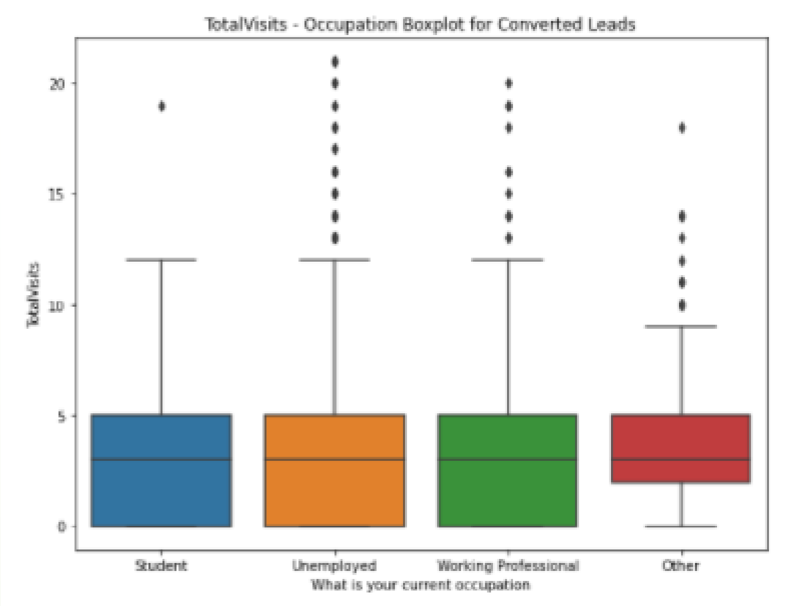
- ✓ Majority traffic seems to come from Google/Organic Search/Direct Traffic

Bivariate Analysis

Total Visits Vs Occupation:

- ✓ There is no significant difference in the Total Visits per profession
- ✓ For Non-Converted Leads – Students/Working Professional/Professional/Other visit the website less(Median) compared to a Converted Lead

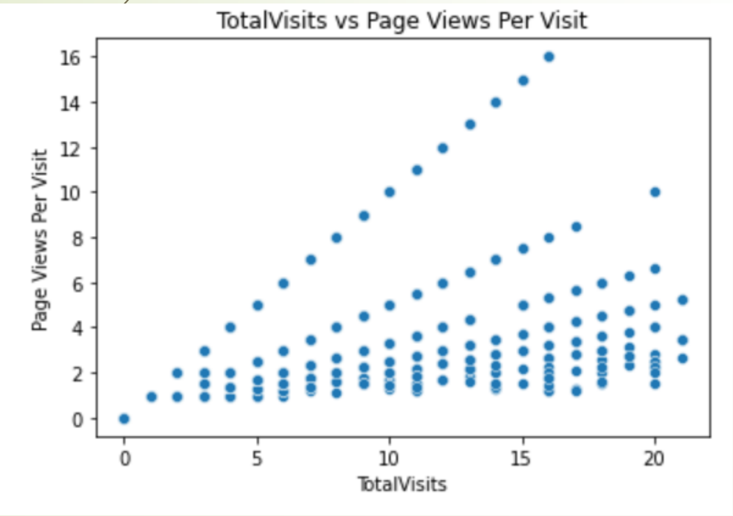


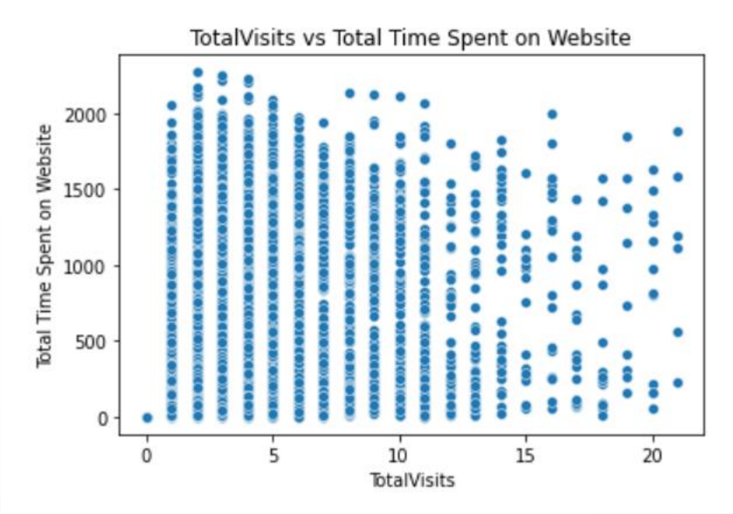


Bivariate Analysis

Total Visits Vs Page Views Vs Total Time Spent:

- ✓ There seems to be a linear relation between the Total Visits Vs Page Views per visit
- ✓ Also, it seems with an increase in the number of visits there is a drop in the Total Time Spent on the Website. Or It may be that that lesser users visit the website multiple times.



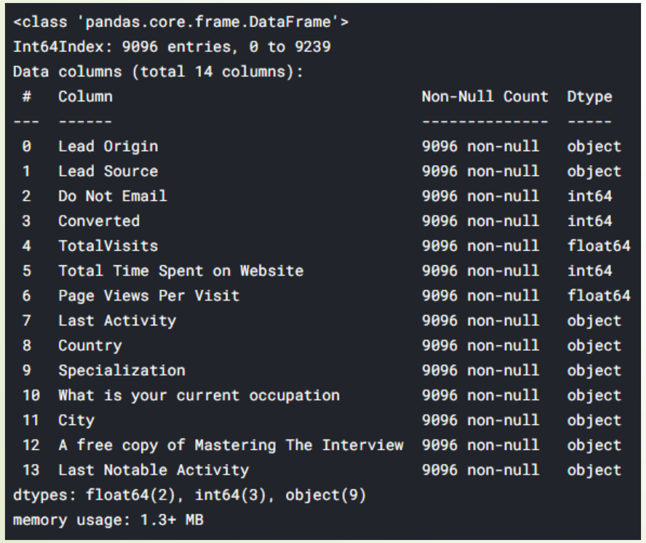


Bivariate Analysis

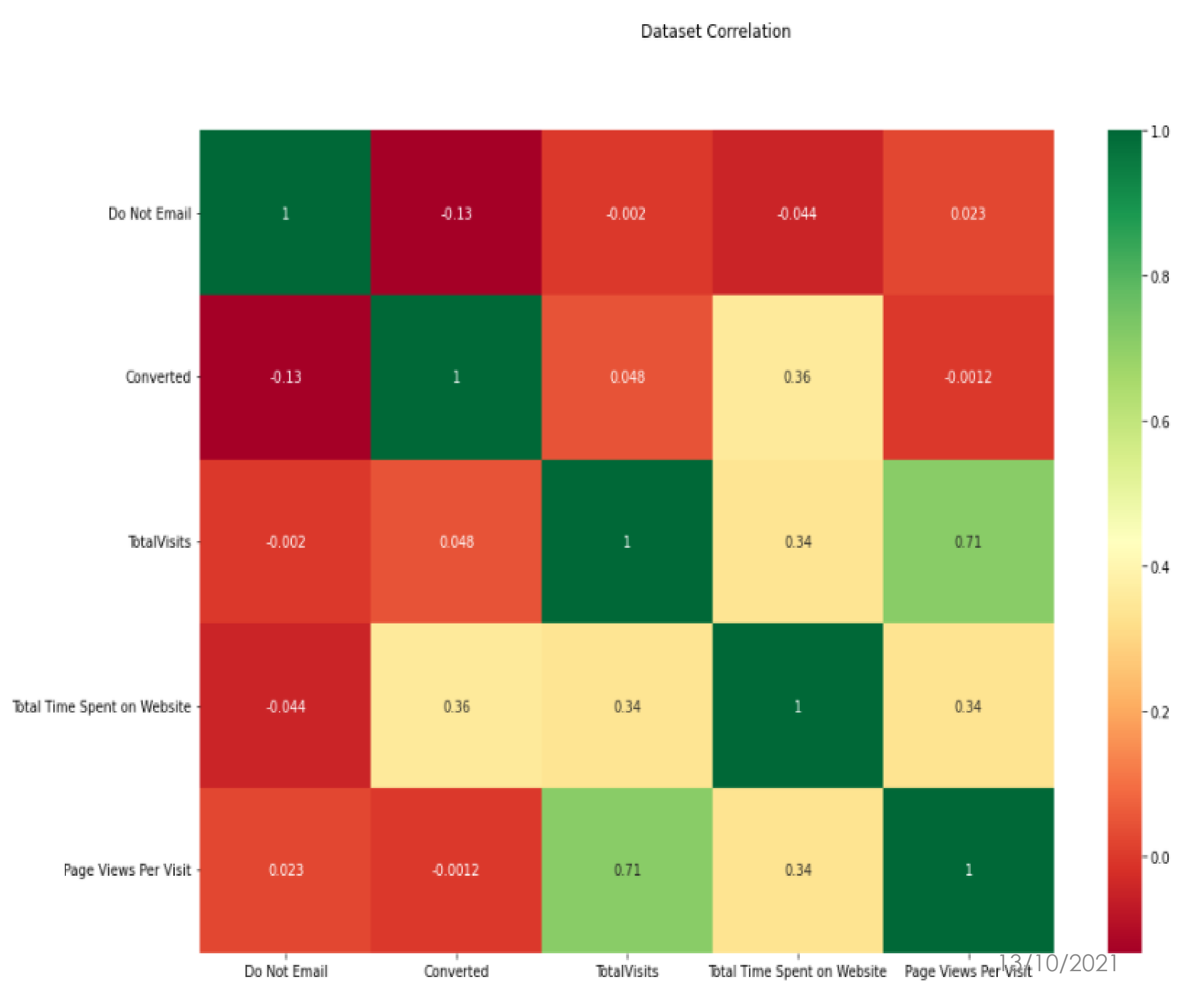
Dataset Correlation

(for the left-over columns after cleaning) :

- ✓ There is some collinearity between Total Visits and Total Time Spent on Website
- ✓ Other than that, there is not much collinearity present in the dataset therefore we are good for building the model with the existing dataset



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9096 entries, 0 to 9239
Data columns (total 14 columns):
 #   Column                                   Non-Null Count  Dtype
---  ------                                   --------------  -----
 0   Lead Origin                              9096 non-null   object
 1   Lead Source                              9096 non-null   object
 2   Do Not Email                             9096 non-null   int64
 3   Converted                                9096 non-null   int64
 4   TotalVisits                              9096 non-null   float64
 5   Total Time Spent on Website              9096 non-null   int64
 6   Page Views Per Visit                     9096 non-null   float64
 7   Last Activity                            9096 non-null   object
 8   Country                                  9096 non-null   object
 9   Specialization                           9096 non-null   object
 10  What is your current occupation          9096 non-null   object
 11  City                                     9096 non-null   object
 12  A free copy of Mastering The Interview   9096 non-null   object
 13  Last Notable Activity                    9096 non-null   object
dtypes: float64(2), int64(3), object(9)
memory usage: 1.3+ MB
```



Machine Learning Model – Logistic Regression

Create Dummy / Scale the variables :

- ✓ The dataset now contains 14 columns(9 categorical and 6 numerical columns).
- ✓ The categorial columns were be converted to dummy variable.
- ✓ After 70:30 split of the data into Train and Test. Train & Test dataset contains 6367 and 2729 rows.
- ✓ Used MinMaxScaler to scale the Variables. "**Converted**" is the [**Target Variable]** variable here.
- ✓ Used Recursive Feature Elimination (RFE) with 15 columns to find out the most significant columns.

Machine Learning Model – Logistic Regression

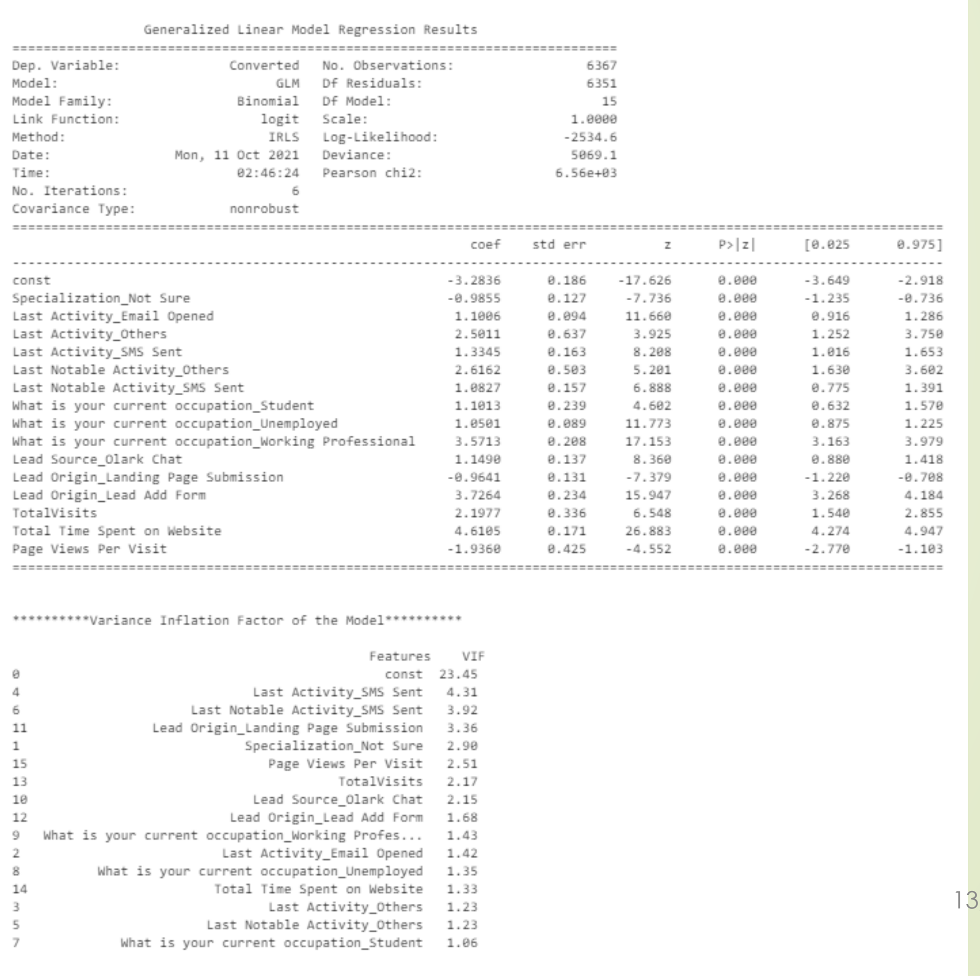
Model Evaluation (Model 2 - Train data):

- ✓ Model Parameters – (With Prob – 0.36)
  - ✓ Accuracy Score - 81.12
  - ✓ Sensitivity - 81.0
  - ✓ Specificity - 81.0
  - ✓ False Positive - 19.0
  - ✓ Positive Predictive Value - 73.0
  - ✓ Negative Predictive Value - 87.0
  - ✓ F1 Score - 81.27

Model Evaluation (Model 2 - Test data):

- ✓ Model Parameters (With Prob – 0.36)
  - ✓ Accuracy Score - 80.54
  - ✓ Sensitivity - 80.0
  - ✓ Specificity - 81.0
  - ✓ False Positive - 19.0
  - ✓ Positive Predictive Value - 71.0
  - ✓ Negative Predictive Value - 87.0
  - ✓ F1 Score - 80.75

*As per the plot (placed in the next slide) between **Accuracy / Sensitivity and Specificity it seems cutoff of 0.36** is the ideal cut-off for the model*



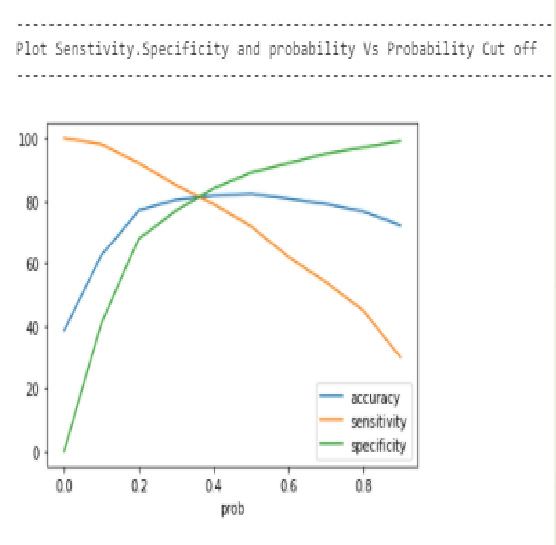
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6367
Model:                            GLM   Df Residuals:                     6351
Model Family:                Binomial   Df Model:                           15
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2534.6
Date:                Mon, 11 Oct 2021   Deviance:                       5069.1
Time:                        02:46:24   Pearson chi2:                 6.56e+03
No. Iterations:                     6
Covariance Type:            nonrobust
========================================================================================================================
                                                           coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------------
const                                                   -3.2836      0.186    -17.626      0.000      -3.649      -2.918
Specialization_Not Sure                                 -0.9855      0.127     -7.736      0.000      -1.235      -0.736
Last Activity_Email Opened                               1.1006      0.094     11.660      0.000       0.916       1.286
Last Activity_Others                                     2.5011      0.637      3.925      0.000       1.252       3.750
Last Activity_SMS Sent                                   1.3345      0.163      8.208      0.000       1.016       1.653
Last Notable Activity_Others                             2.6162      0.503      5.201      0.000       1.630       3.602
Last Notable Activity_SMS Sent                           1.0827      0.157      6.888      0.000       0.775       1.391
What is your current occupation_Student                  1.1013      0.239      4.602      0.000       0.632       1.570
What is your current occupation_Unemployed               1.0501      0.089     11.773      0.000       0.875       1.225
What is your current occupation_Working Professional     3.5713      0.208     17.153      0.000       3.163       3.979
Lead Source_Olark Chat                                   1.1490      0.137      8.360      0.000       0.880       1.418
Lead Origin_Landing Page Submission                     -0.9641      0.131     -7.379      0.000      -1.220      -0.708
Lead Origin_Lead Add Form                                3.7264      0.234     15.947      0.000       3.268       4.184
TotalVisits                                              2.1977      0.336      6.548      0.000       1.540       2.855
Total Time Spent on Website                              4.6105      0.171     26.883      0.000       4.274       4.947
Page Views Per Visit                                    -1.9360      0.425     -4.552      0.000      -2.770      -1.103
========================================================================================================================


**********Variance Inflation Factor of the Model**********

                                               Features    VIF
0                                                 const  23.45
4                                Last Activity_SMS Sent   4.31
6                        Last Notable Activity_SMS Sent   3.92
11                  Lead Origin_Landing Page Submission   3.36
1                               Specialization_Not Sure   2.90
15                                 Page Views Per Visit   2.51
13                                          TotalVisits   2.17
10                               Lead Source_Olark Chat   2.15
12                            Lead Origin_Lead Add Form   1.68
9   What is your current occupation_Working Profes...   1.43
2                            Last Activity_Email Opened   1.42
8            What is your current occupation_Unemployed   1.35
14                          Total Time Spent on Website   1.33
3                                  Last Activity_Others   1.23
5                          Last Notable Activity_Others   1.23
7               What is your current occupation_Student   1.06
```

Machine Learning Model – Logistic Regression

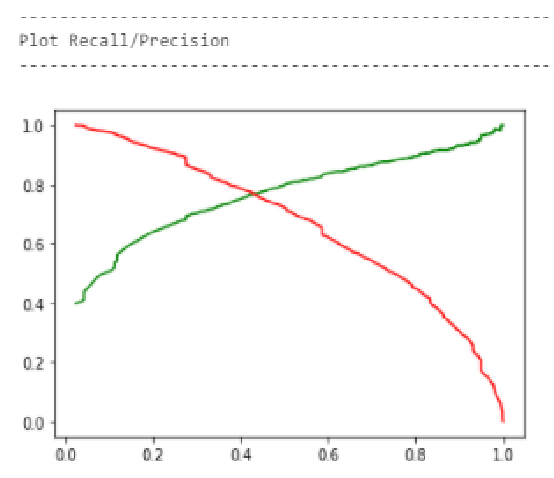
- As per the plot (placed in the next slide) between **Accuracy / Sensitivity and Specificity it seems cutoff of 0.36** is the ideal cut-off for the model

Probability	Accuracy	Sensitivity	Specificity
0	38.57	100	0
0.1	62.6	98	41
0.2	77.08	92	68
0.3	80.41	85	77
0.4	81.69	79	84
0.5	82.25	72	89
0.6	80.74	62	92
0.7	79.08	54	95
0.8	76.69	45	97
0.9	72.31	30	99

- Please note that the Lead score is the Probability % returned by the Logistic Regression model.



```
--------------------------------------------------------------------------------
Plot Senstivity.Specificity and probability Vs Probability Cut off
--------------------------------------------------------------------------------
```

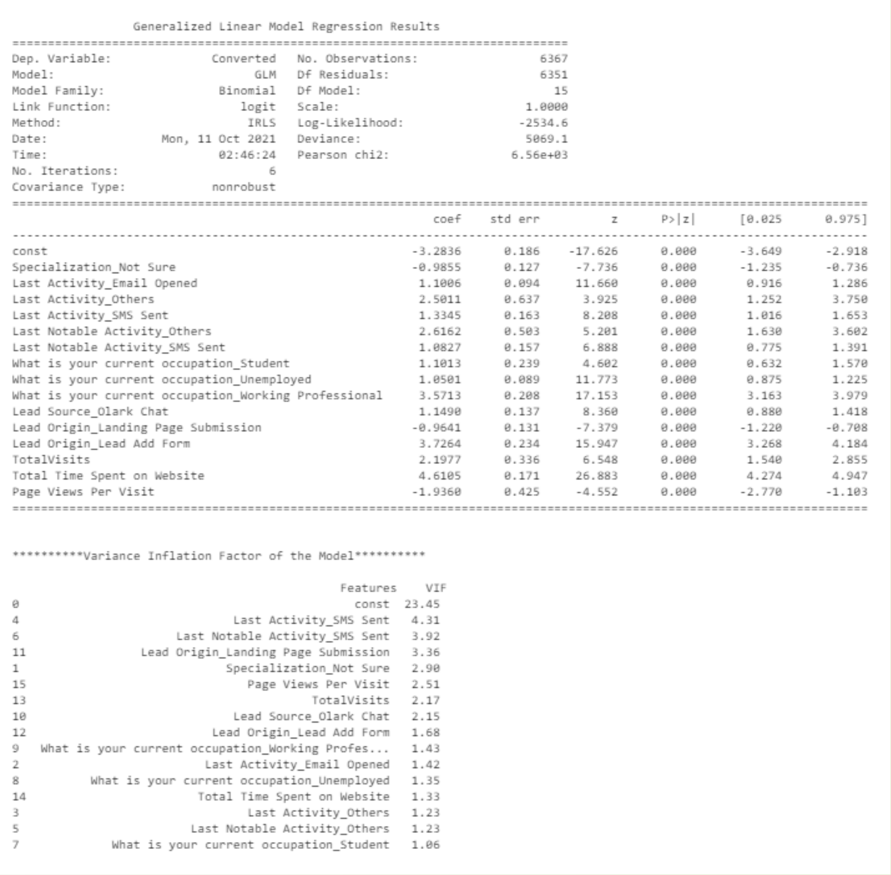


```
--------------------------------------------------------------------------------
Plot Recall/Precision
--------------------------------------------------------------------------------
```

Machine Learning Model – Logistic Regression

Top Features:

Serial Number	Features	Feature Coefficient
1	Total Time Spent on Website	4.610469
2	Lead Origin_Lead Add Form	3.726397
3	What is your current occupation_Working Professional	3.571323
4	const	3.283554
5	Last Notable Activity_Others	2.616187
6	Last Activity_Others	2.501128
7	TotalVisits	2.197668
8	Page Views Per Visit	1.936048
9	Last Activity_SMS Sent	1.334504
10	Lead Source_Olark Chat	1.148977
11	What is your current occupation_Student	1.101285
12	Last Activity_Email Opened	1.100598
13	Last Notable Activity_SMS Sent	1.082652
14	What is your current occupation_Unemployed	1.050106
15	Specialization_Not Sure	0.985521
16	Lead Origin_Landing Page Submission	0.964084



```
                 Generalized Linear Model Regression Results                  
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6367
Model:                            GLM   Df Residuals:                     6351
Model Family:                Binomial   Df Model:                           15
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2534.6
Date:                Mon, 11 Oct 2021   Deviance:                       5069.1
Time:                        02:46:24   Pearson chi2:                 6.56e+03
No. Iterations:                     6                                         
Covariance Type:            nonrobust                                         
=======================================================================================================================
                                                            coef    std err          z      P>|z|     [0.025     0.975]
-----------------------------------------------------------------------------------------------------------------------
const                                                    -3.2836      0.186    -17.626      0.000     -3.649     -2.918
Specialization_Not Sure                                  -0.9855      0.127     -7.736      0.000     -1.235     -0.736
Last Activity_Email Opened                                1.1006      0.094     11.660      0.000      0.916      1.286
Last Activity_Others                                      2.5011      0.637      3.925      0.000      1.252      3.750
Last Activity_SMS Sent                                    1.3345      0.163      8.208      0.000      1.016      1.653
Last Notable Activity_Others                              2.6162      0.503      5.201      0.000      1.630      3.602
Last Notable Activity_SMS Sent                            1.0827      0.157      6.888      0.000      0.775      1.391
What is your current occupation_Student                   1.1013      0.239      4.602      0.000      0.632      1.570
What is your current occupation_Unemployed                1.0501      0.089     11.773      0.000      0.875      1.225
What is your current occupation_Working Professional      3.5713      0.208     17.153      0.000      3.163      3.979
Lead Source_Olark Chat                                    1.1490      0.137      8.360      0.000      0.880      1.418
Lead Origin_Landing Page Submission                      -0.9641      0.131     -7.379      0.000     -1.220     -0.708
Lead Origin_Lead Add Form                                 3.7264      0.234     15.947      0.000      3.268      4.184
TotalVisits                                               2.1977      0.336      6.548      0.000      1.540      2.855
Total Time Spent on Website                               4.6105      0.171     26.883      0.000      4.274      4.947
Page Views Per Visit                                     -1.9360      0.425     -4.552      0.000     -2.770     -1.103
=======================================================================================================================


**********Variance Inflation Factor of the Model**********

                                             Features    VIF
0                                               const  23.45
4                              Last Activity_SMS Sent   4.31
6                      Last Notable Activity_SMS Sent   3.92
11                Lead Origin_Landing Page Submission   3.36
1                             Specialization_Not Sure   2.90
15                               Page Views Per Visit   2.51
13                                        TotalVisits   2.17
10                             Lead Source_Olark Chat   2.15
12                          Lead Origin_Lead Add Form   1.68
9   What is your current occupation_Working Profes...   1.43
2                          Last Activity_Email Opened   1.42
8          What is your current occupation_Unemployed   1.35
14                        Total Time Spent on Website   1.33
3                                Last Activity_Others   1.23
5                        Last Notable Activity_Others   1.23
7             What is your current occupation_Student   1.06
```

Conclusion

Summary:

Instead of contacting all the Leads, the Sales team of the X Education company should use the Lead Scoring as a fundamental methodology to determine which leads have the higher potential to transform into a buyer.

The model presented (in slide 27) could identify 80-81% of the leads, possible buyers correctly compared to 38% before.

This means if the Sales team were spending X days to contact 100 Leads, they were able to get only 38 buyers.

But now, with the help of this model Sales team will spend X days to contact 100 Leads identified by the model, out of which 80-81 buyers should signup, which is a straight 110% jump.

Therefore, the Sales team of the company

- Will likely to spend time on the leads more likely to convert into customers and lower marketing costs. Also means Targets can be met well before agreed time.
- Revenue to increase quite a few folds with the same Sale FTE due to higher conversion rates
- The sales team can get 20% of the non-converted leads and recommend it to another group to nurture them a little bit before they can be converted to a possible buyer.