LEAD SCORING CASE STUDY

Submitted by –

Ankita Gupta

& Satish Chowdary

(Batch - DSC43)

Index

1. Problem Statement
2. Approach
3. EDA and Univariate/ Bivariate Analysis
4. Overall Approach
5. Conclusion
6. Final Recommendation

Problem Statement

X Education Company is an education company which sells Online Courses to professionals. Marketing of the company is done via several websites and search engines. When people land on the website and fill up a form for enquiry, they are classified as lead. Sales Team approach to these leads for sale of their product (conversion).

Current Lead Conversion Rate of the company is 30% and company intends to increase the Conversion rate to 80% by identifying "Hot Leads" which has comparatively high probability for conversion.

Now there is 1 dataset provided, which contains details of around 9000+ leads on 37 features.

We have following Goals as a Data Analyst -

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

Approach

Following is the step-by-step approach which is used to analyze the dataset and draw the inferences based on the findings. There are multiple ways to approach a problem and the approach which is used in this assignment is one of the ways to solve the problem. Same Steps are also followed in the Python file attached.

1. Import Libraries and Data source
2. EDA – Exploratory Data Analysis
    a. Handling NULL Values
        a. Converted "Select" values to Null wherever applicable
        b. Removed Columns where null value is greater than 45%
        c. Removed Rows where null value is less than 2%
        d. Imputed values based on data available, wherever necessary
    b. Column Selection
        a. Removed Columns where single value available across all data points
        b. Removed Columns where there was high data imbalance >85%
        c. Removed Redundant Columns from the dataset

Approach contd.

3. Univariate / Bivariate Analysis
    a. Standardizing categories with multiple values
    b. Combining multiple categories as single category which are contributing very miniscule in a feature
    c. Handling outliers in the dataset

4. Data Preparation for Modelling
    a. Conversion of Categorical Binary variable to 1/0
    b. Dummy Variable Creation for categorical columns
    c. Train and Test Split
    d. Feature Scaling
        a. Train Dataset as fit_transform
        b. Test Dataset as transform
    e. Correlation in Features

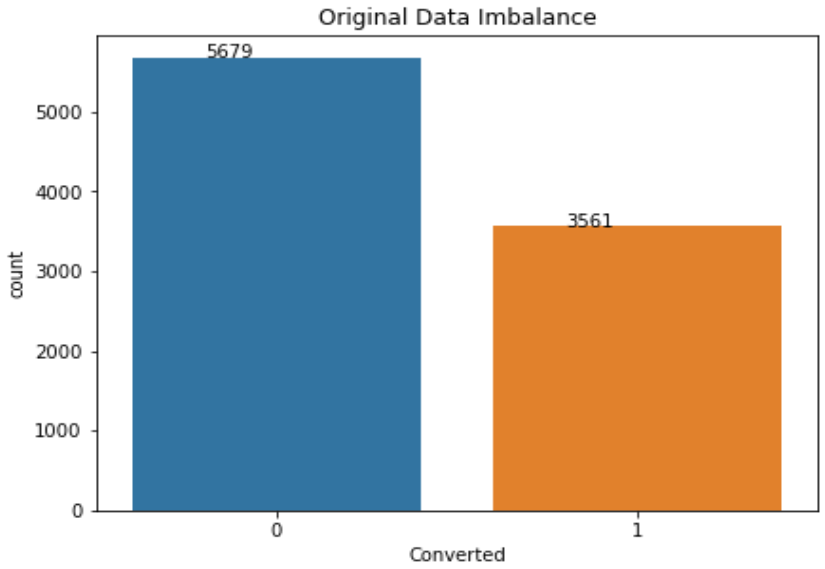
5. Model Building – Logistic Regression
    a. Model Building based on p-value and VIF value
    b. Parameter tracking – Confusion Matrix, Accuracy, Recall, Precision, Specificity, False Positive Rate, Negative Predictive Value, etc.

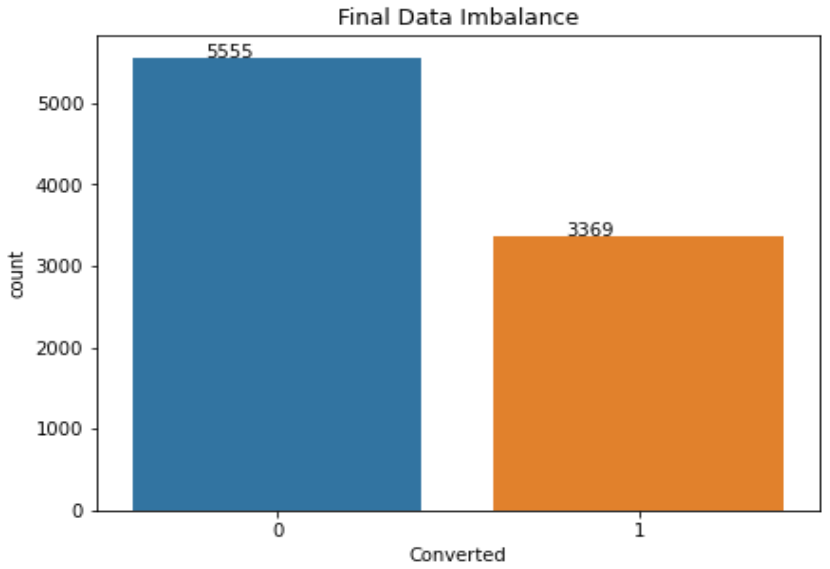
Approach contd.

6. ROC Curve
7. Optimal Cutoff Point
8. Precision Recall Curve
9. Predictions on Test Set
10. Final Train v/s Test Set Comparison on Parameters

EDA – Data Imbalance and Shape

Data Imbalance Ratio





Data Set Shape

```
Original Shape of Dataset:
--------------------
(9240, 37)

Final Shape of Dataset:
--------------------
(8924, 13)

Final Dataset ROWS retained post null value handling and univariate/bivariate analysis is :
-------------------------------------------------------------------------------------------

96.58
```

0 implies Converted
1 implies Not Converted

EDA – Null Value Handling and Column Selection

	Column Name	Missing Value
13	How did you hear about X Education	78.463203
28	Lead Profile	74.188312
25	Lead Quality	51.590909
33	Asymmetrique Profile Score	45.649351
32	Asymmetrique Activity Score	45.649351
30	Asymmetrique Activity Index	45.649351
31	Asymmetrique Profile Index	45.649351
29	City	39.707792
12	Specialization	36.580087

Columns where Null Val is >45% are removed

22	Receive More Updates About Our Courses	1
24	Update me on Supply Chain Content	1
25	Get updates on DM Content	1
16	Magazine	1
27	I agree to pay the amount through cheque	1

Columns with same value across all data points are removed

```
Search :
No     0.998485
Yes    0.001515
Name: Search, dtype: float64
------------------------------

Newspaper Article :
No     0.999784
Yes    0.000216
Name: Newspaper Article, dtype: float64
------------------------------

X Education Forums :
No     0.999892
Yes    0.000108
Name: X Education Forums, dtype: float64
------------------------------

Newspaper :
No     0.999892
Yes    0.000108
Name: Newspaper, dtype: float64
------------------------------

Digital Advertisement :
No     0.999567
Yes    0.000433
Name: Digital Advertisement, dtype: float64
------------------------------

Through Recommendations :
No     0.999242
Yes    0.000758
Name: Through Recommendations, dtype: float64
------------------------------
```

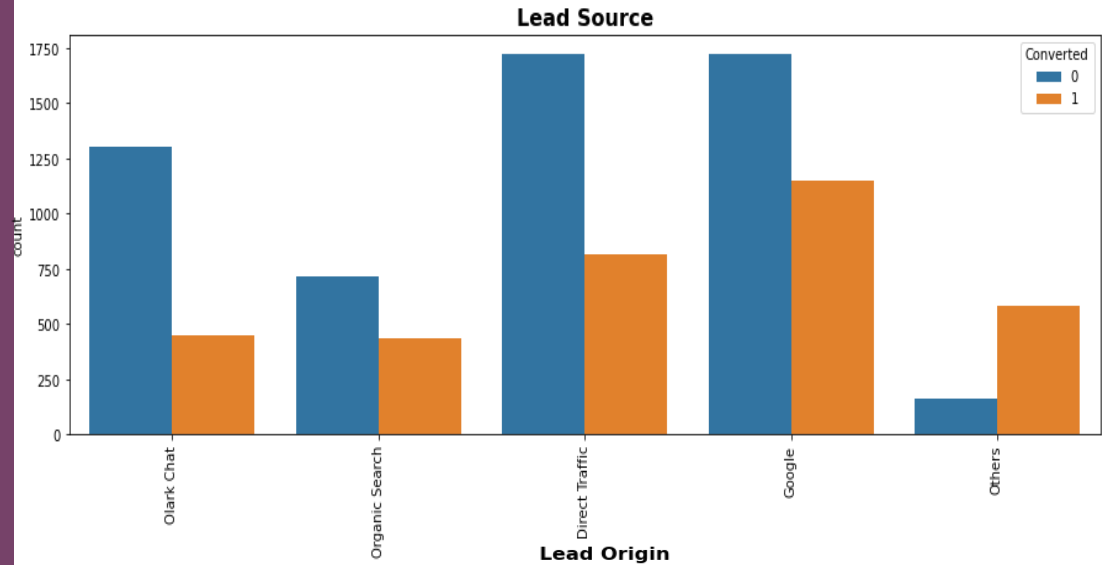
Columns with high Data Imbalance are also removed

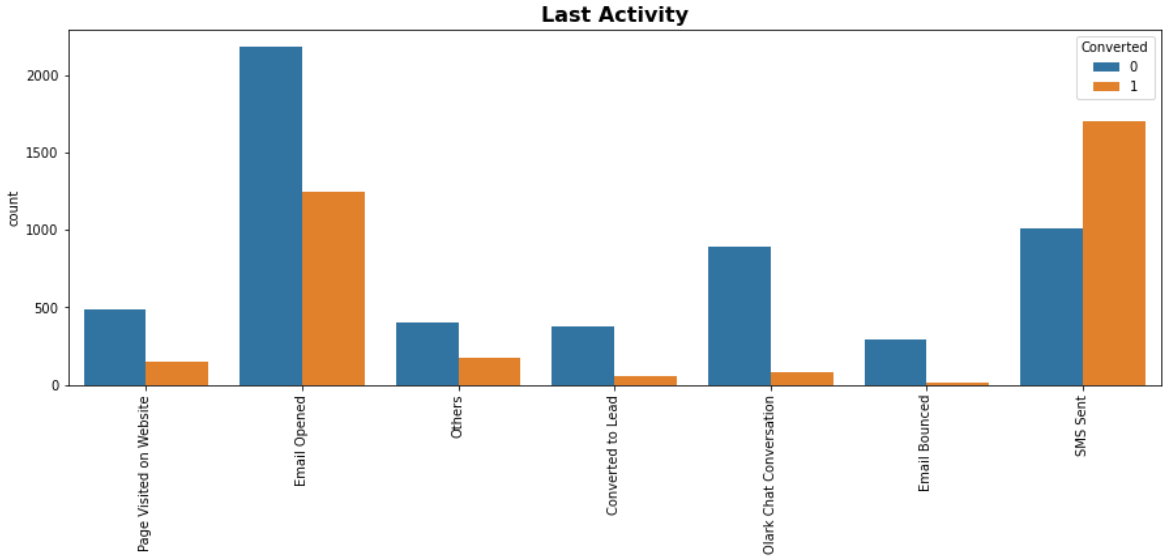
Columns with intermediate missing Values are imputed based on the data understanding

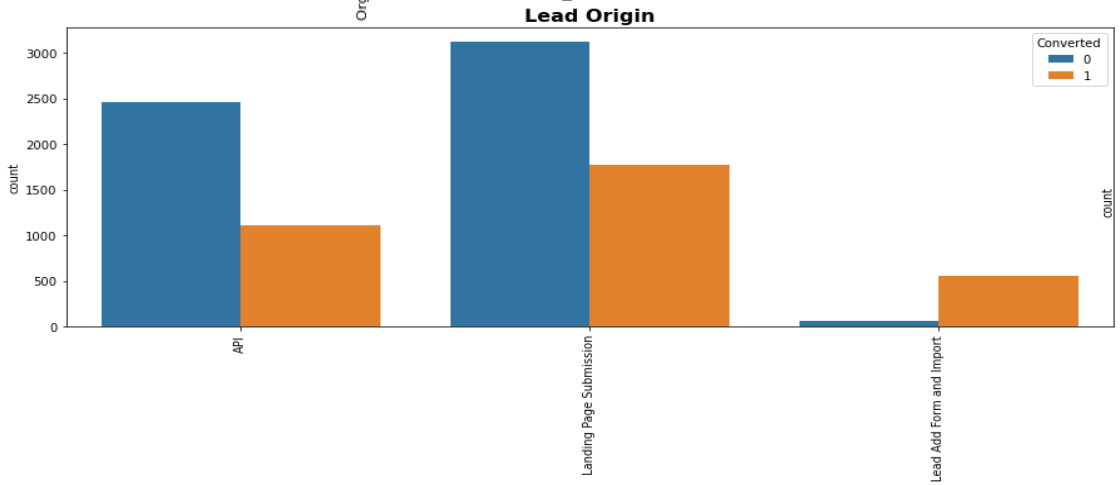
	Column Name	Missing Value
10	City	39.707792
8	Specialization	36.580087
9	Tags	36.287879
4	TotalVisits	1.482684
6	Page Views Per Visit	1.482684
7	Last Activity	1.114719
2	Lead Source	0.389610

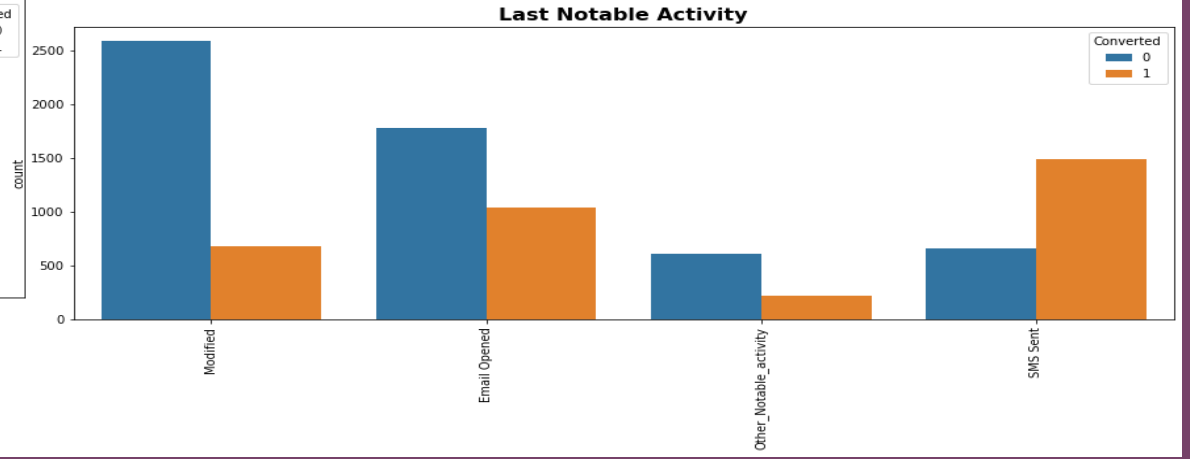
Columns where missing Values are <2%, those rows are removed

EDA – Univariate/ Bivariate Analysis and Final Column Selection

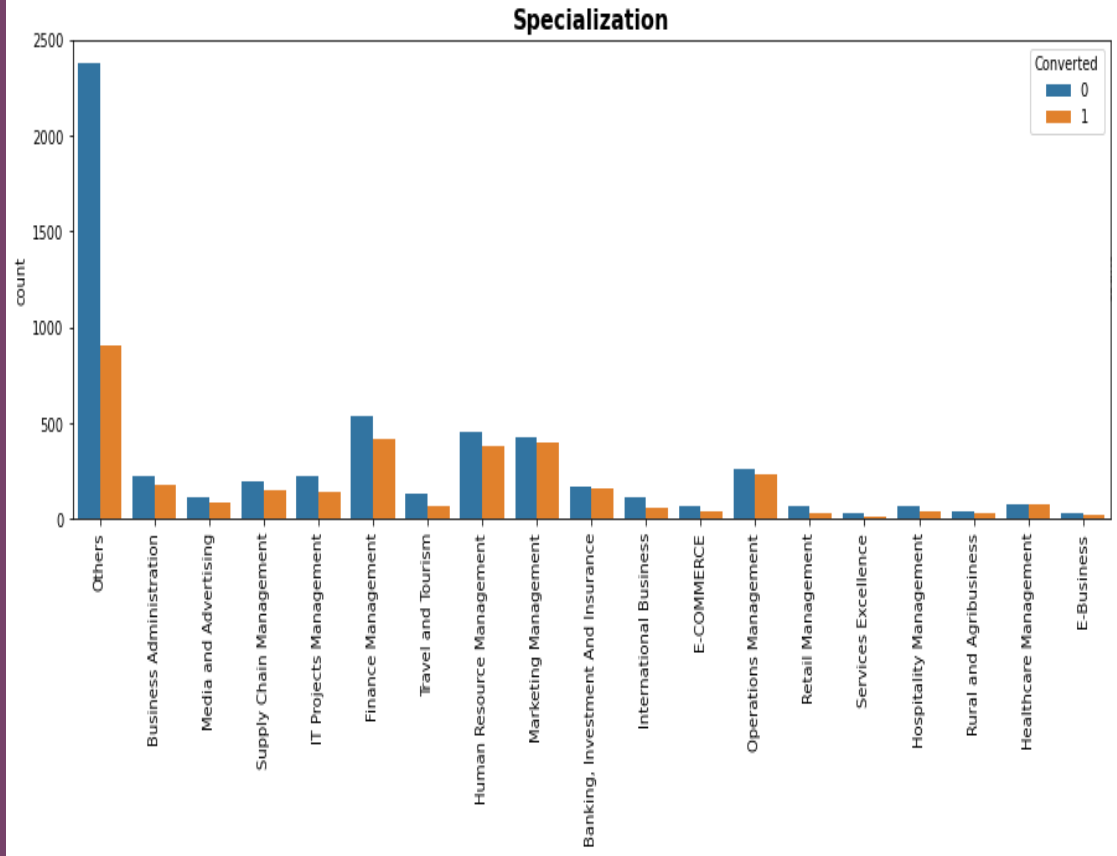


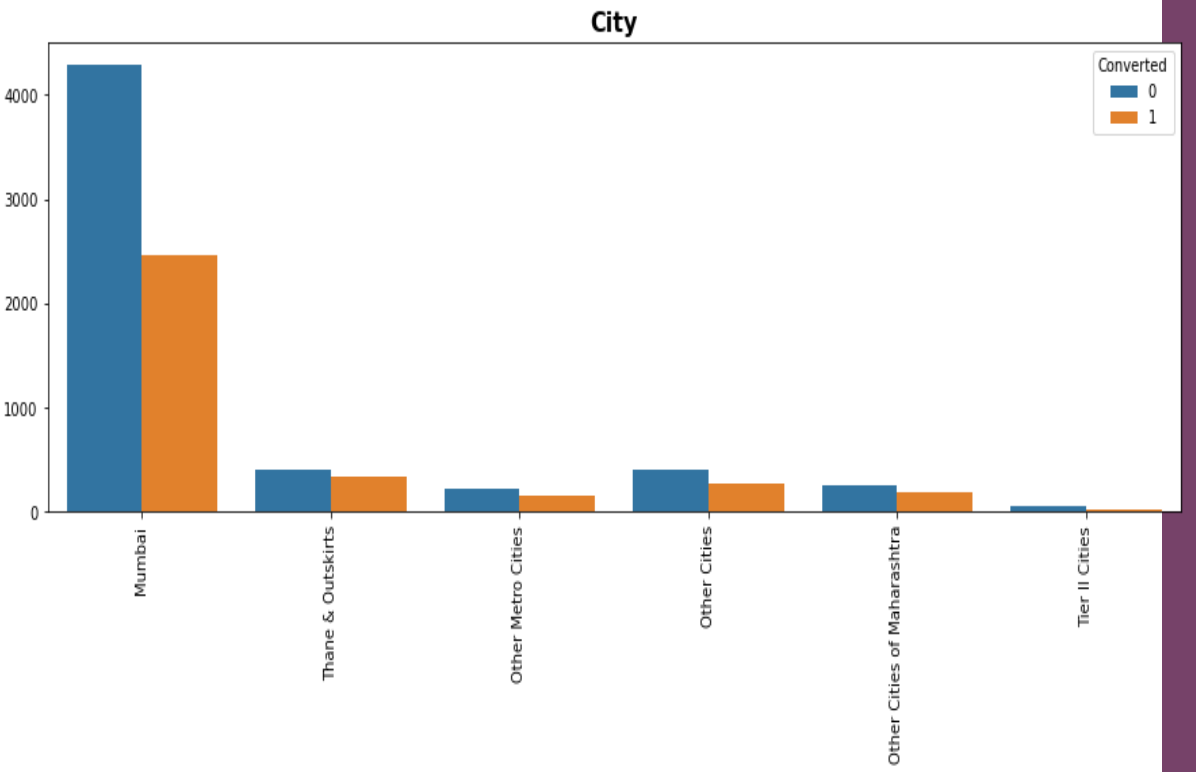




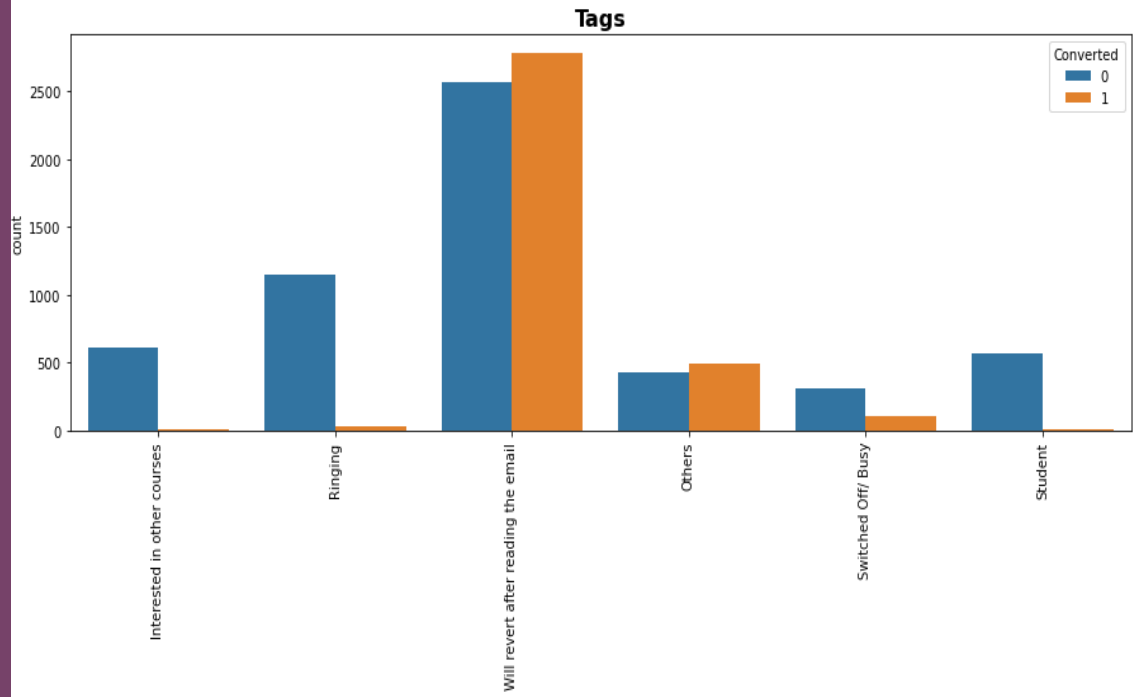


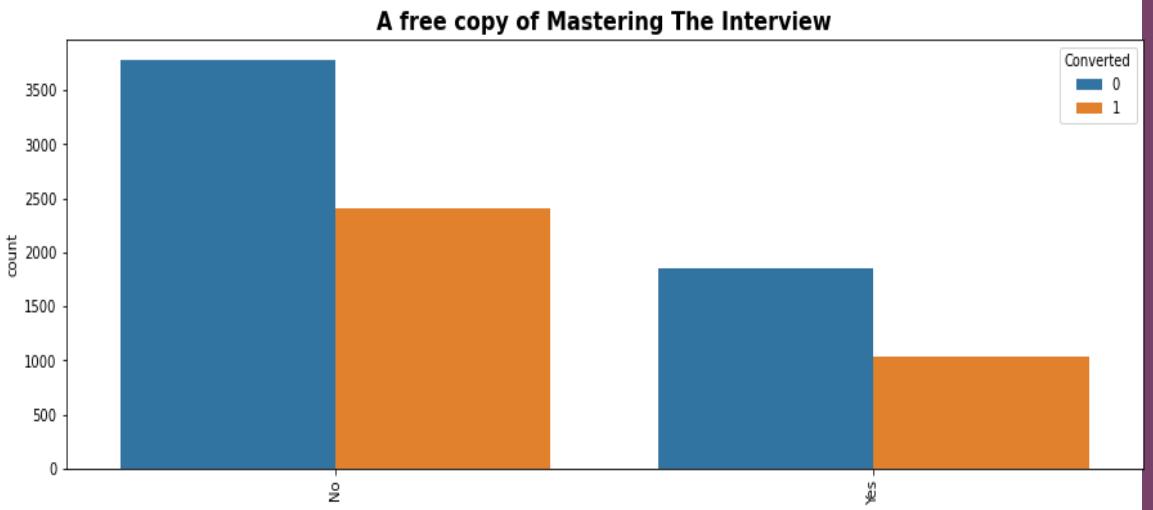
EDA – Univariate/ Bivariate Analysis and Final Column Selection

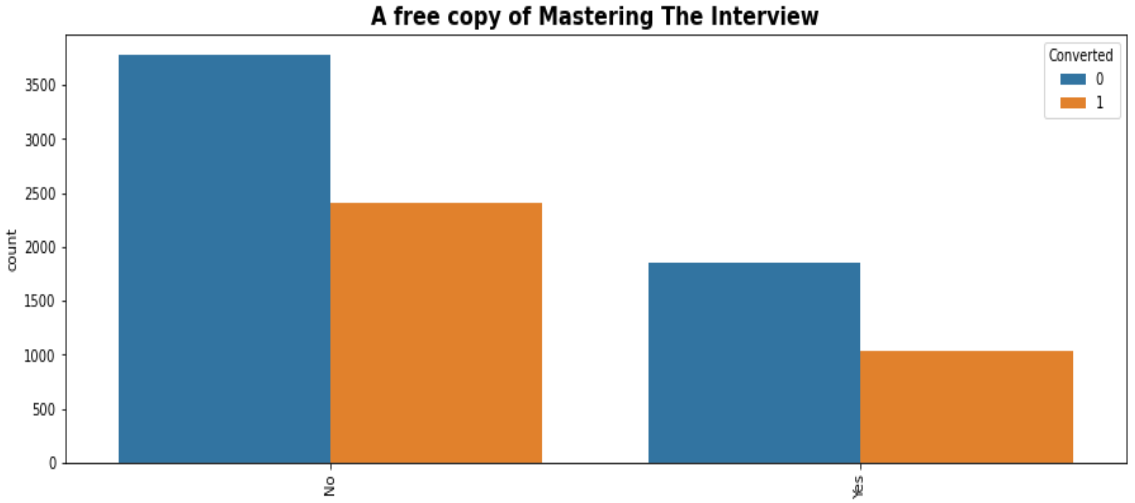




EDA – Univariate/ Bivariate Analysis and Final Column Selection



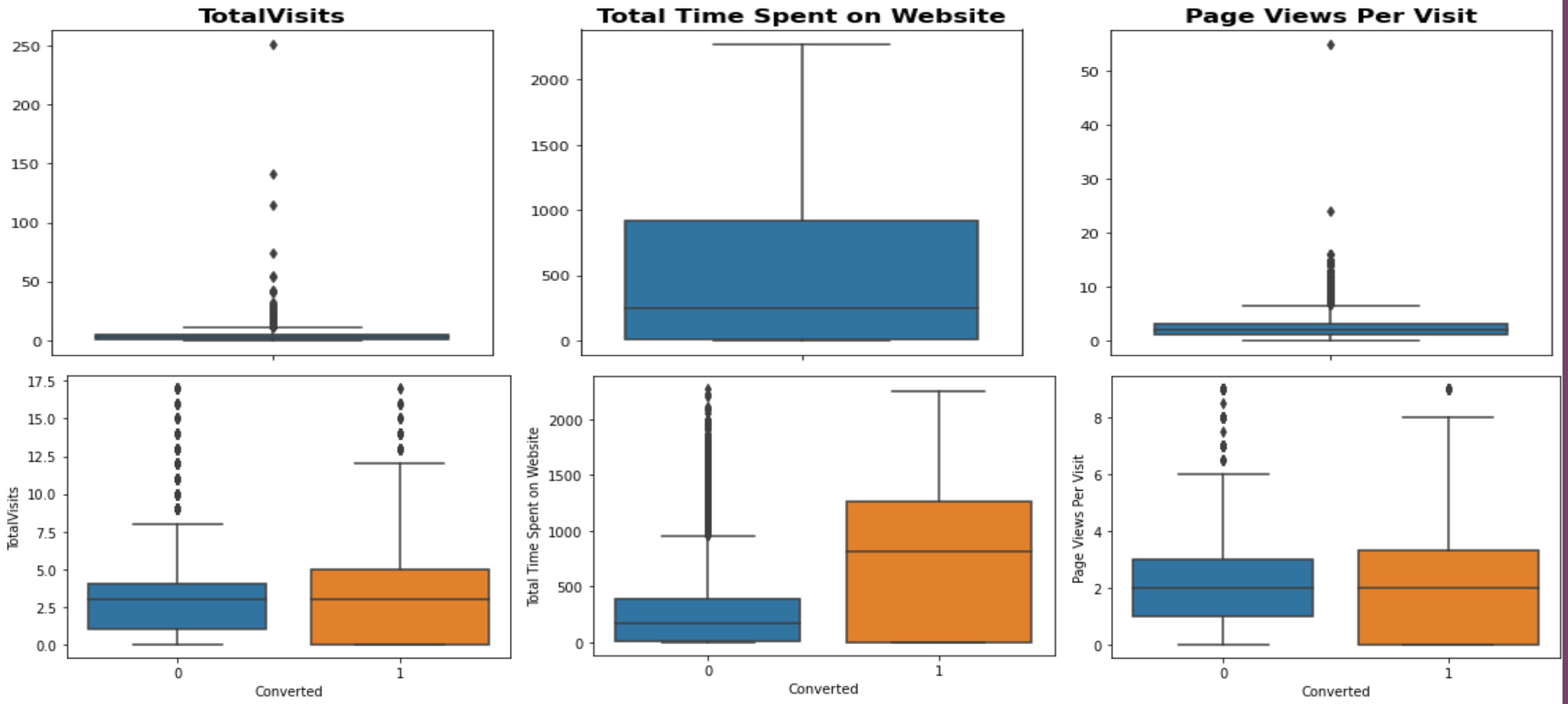




EDA – Univariate/ Bivariate Analysis and Final Column Selection

I

II



I implies Analysis of Numeric Variable

I I implies Analysis of Numeric Variable post outlier treatment, in respect to Converted/Non-Converted Flag

Data Preparation Steps for Modelling

1. Columns with Binary Values Such as Yes/ No were converted to 1 and 0, like "*A free copy of Mastering The Interview*"
2. Dummy Values were created for all remaining 7 columns and then Original Columns were dropped

```
Final DataFrame Shape for Modelling:
------------------------

(8924, 49)
```

3. Train (70%) and Test (30%) Dataset split on random_state 100
4. For Numeric Columns, Feature Scaling was done using Standard Scaler. On train dataset, Feature Scaling was done using fit_transform and test dataseet, Feature Scaling was done using transform.

```
Train and Test Dataset shape:
X_train shape: (6246, 47)
y_train shape: (6246,)
X_test shape: (2678, 47)
y_test shape: (2678,)
```

```
Data columns (total 13 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Lead Number                             8924 non-null   int64
 1   Lead Origin                             8924 non-null   object
 2   Lead Source                             8924 non-null   object
 3   Converted                               8924 non-null   int64
 4   TotalVisits                             8924 non-null   float64
 5   Total Time Spent on Website             8924 non-null   int64
 6   Page Views Per Visit                    8924 non-null   float64
 7   Last Activity                           8924 non-null   object
 8   Specialization                          8924 non-null   object
 9   Tags                                    8924 non-null   object
 10  City                                    8924 non-null   object
 11  A free copy of Mastering The Interview  8924 non-null   object
 12  Last Notable Activity                   8924 non-null   object
dtypes: float64(2), int64(3), object(8)
memory usage: 1.2+ MB
```

Post EDA Final Column List which will be used in Modelling

Model Building – Logistic Regression

1. Model Building is done using RFE with initial 20 variables
2. Model was built by removing features one by one where p-values are less than 0.010 and VIF Value is less than 2
3. Predictions on Test Data was performed and compared with the actual Data.

```
Confusuion Matrix:
[[3504  367]
 [ 514 1861]]

Accuracy:    85.89
Recall:    78.36
Precision:    83.53
Specificity:    90.52
False Positive Rate:    9.48
Negative Predictive Value:    87.21
```

	Features	VIF
2	Source_Olark Chat	1.55
5	Activity_Olark Chat Conversation	1.38
0	Total Time Spent on Website	1.30
14	Last Notable Activity_SMS Sent	1.28
1	Origin_Lead Add Form and Import	1.15
11	Tags_Ringing	1.11
7	Specialization_Human Resource Management	1.09
10	Tags_Interested in other courses	1.09
8	Specialization_Marketing Management	1.08
12	Tags_Student	1.08
3	Activity_Converted to Lead	1.06
6	Activity_Page Visited on Website	1.05
13	Tags_Switched Off/ Busy	1.05
9	Specialization_Operations Management	1.04
4	Activity_Email Bounced	1.02

Dep. Variable:	Converted	**No. Observations:**	6246
Model:	GLM	**Df Residuals:**	6230
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2122.0
Date:	Sun, 16 Oct 2022	**Deviance:**	4244.0
Time:	14:35:50	**Pearson chi2:**	7.75e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.6353	0.065	-9.790	0.000	-0.762	-0.508
Total Time Spent on Website	1.1742	0.047	24.842	0.000	1.082	1.267
Origin_Lead Add Form and Import	3.8024	0.216	17.634	0.000	3.380	4.225
Source_Olark Chat	0.8978	0.113	7.956	0.000	0.677	1.119
Activity_Converted to Lead	-1.0816	0.231	-4.688	0.000	-1.534	-0.629
Activity_Email Bounced	-2.6165	0.369	-7.088	0.000	-3.340	-1.893
Activity_Olark Chat Conversation	-1.6546	0.168	-9.850	0.000	-1.984	-1.325
Activity_Page Visited on Website	-0.7044	0.154	-4.566	0.000	-1.007	-0.402
Specialization_Human Resource Management	0.4587	0.139	3.305	0.001	0.187	0.731
Specialization_Marketing Management	0.6000	0.135	4.437	0.000	0.335	0.865
Specialization_Operations Management	0.4300	0.165	2.600	0.009	0.106	0.754
Tags_Interested in other courses	-3.4509	0.334	-10.325	0.000	-4.106	-2.796
Tags_Ringing	-4.8719	0.252	-19.333	0.000	-5.366	-4.378
Tags_Student	-3.4698	0.380	-9.134	0.000	-4.214	-2.725
Tags_Switched Off/ Busy	-1.8708	0.178	-10.501	0.000	-2.220	-1.522
Last Notable Activity_SMS Sent	1.9521	0.103	18.974	0.000	1.750	2.154

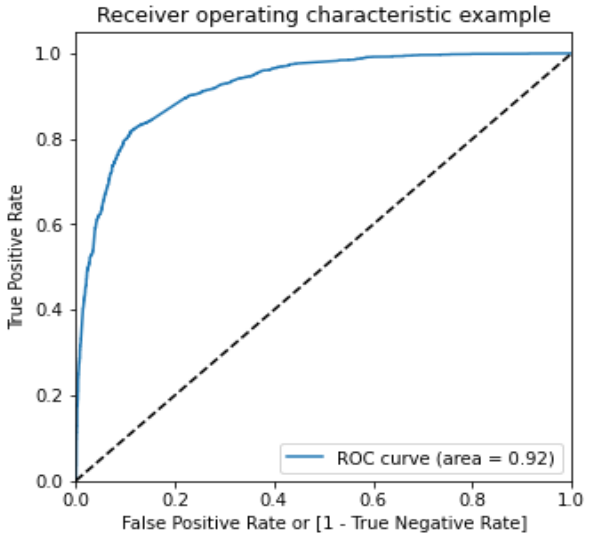
ROC Curve, Optimum Cutoff Point and Precision Recall Curve

Based on our Model, we have got below characteristics:

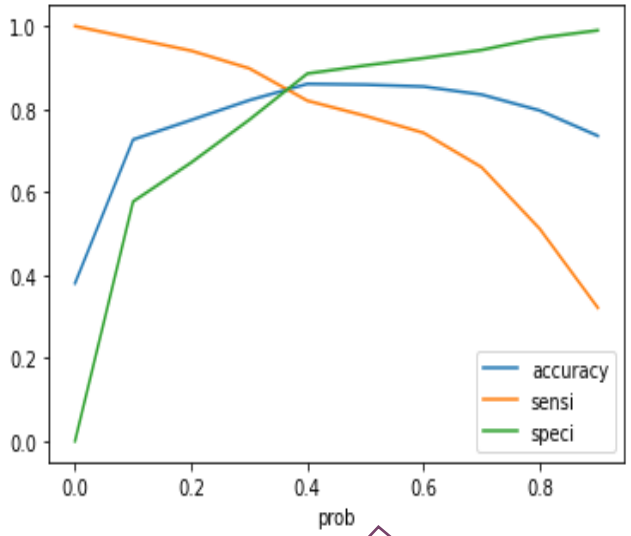
1. ROC Curve Value as 0.92
2. Optimum Cutoff Point as 0.35

Based on this Cutoff Point of 0.35 we have got the following metrics on Train set:

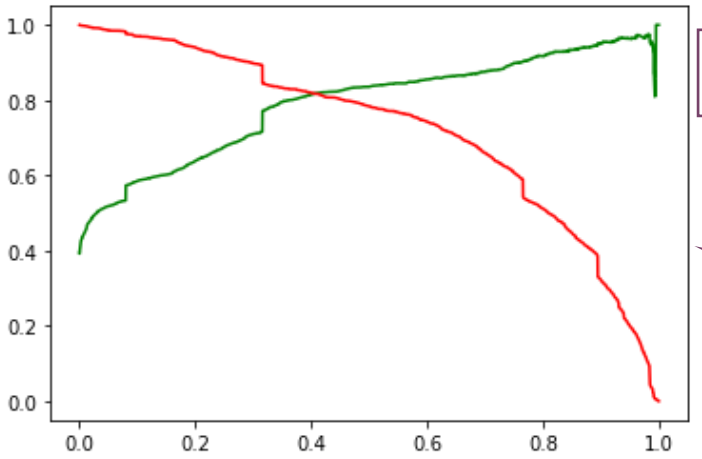
```
Confusion Matrix:

array([[3362,  509],
       [ 396, 1979]], dtype=int64)


Accuracy:   85.51
Recall:   83.33
Precision:   79.54
Specificity:   86.85
False Positive Rate:   13.15
Negative Predictive Value:   89.46
```



ROC Curve



Optimum Cut Off Point



Precision Recall Curve

Final Recommendation

Based in the above analysis we can conclude on the following points –

1. Lead Source -
    a. Olark Chat
2. Last Activity -
    a. Olark Chat Conversation
    b. Converted to Lead
    c. Page visited on Website
    d. Email Bounced
3. Total Time Spent on Website
4. Last Notable Activity -
    a. SMS Sent
5. Lead Origin -
    a. Lead Add Form and Import
6. Tags -
    a. Ringing
    b. Intrested in other courses
    c. Student
    d. Switched off/ Busy
7. Specialization -
    a. HR Management
    b. Marketing Management
    c. Ops Management


THANK YOU