Lead Scoring Assignment Case Study

Gaurav B R [gauravhsn8@gmail.com]

Mirza Burhan [mirzaburhan23@gmail.com]

Sai Vaibhav Naidu [svnhpt.vaibhav@gmail.com]

Introduction

- An education company named X Education sells online courses to industry professionals
- The company markets its courses on several websites and search engines like Google
- Upon landing on the website, people browse the courses or fill up a form for the course or watch some videos
- Once the form is filled providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals
- After the leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not
- The typical lead conversion rate at X education is around 30%

Problem Statement

- Although X Education gets a lot of leads, its lead conversion rate is very poor
- For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted
- The company wants to identify the most potential leads (hot leads)
- If they successfully identify this set of leads, the lead conversion rate should go up as focus would be on these leads

Lead Conversion Process

This can be represented using the following funnel:

- There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom
- in order to get a higher lead conversion, we need to nurture the potential leads well
    - By educating them about the product
    - Constantly communicating

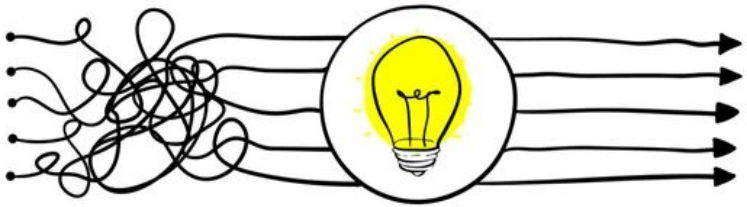
Goals and Objective

- To help the company select the most promising leads (leads that are most likely to convert into paying customers)
- 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
- There are some more problems presented by the company which our model should be able to adjust to if the company's requirement changes in the future

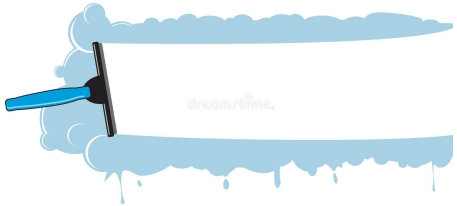
Steps Involved

We use Python in Google Colab to execute our work. Below mentioned are the step-wise procedure into analyzing and predicting the results

1. Importing necessary libraries
2. Importing the provided dataset
3. Data Understanding & Cleaning
4. Exploratory Data Analysis (Variables Inspection)
5. Data Preparation
6. Model Building (Logistic Regression)
7. Model Evaluation (Logistic Regression Metrics)
8. Model Testing
9. Model Inference
10. Recommendation based on our results



Data Understanding and Cleaning



- After we import the dataset, we have to go through the entire data present and make some key observations
- Checking the overall dimensions of the set
- Checking the column formats and correcting if any irregularities found.
- Checking for any NULL values present in the data
- Deal with NULL values by either imputing those rows or replacing with either mean or median values

```
# Imputing the null values for numerical columns (Excluding Binary columns having 0 and 1)
num_cols = ['Lead Number', 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit']
leads[num_cols].isna().mean()*100
```

```
Lead Number                     0.000000
TotalVisits                     1.482684
Total Time Spent on Website     0.000000
Page Views Per Visit            1.482684
dtype: float64
```

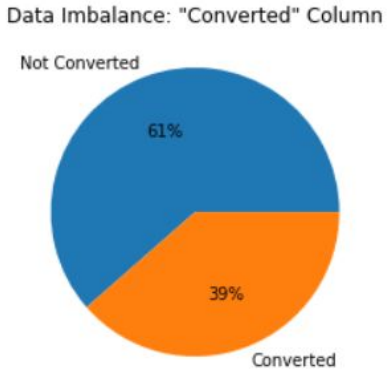
Contd...

- Dealing with the NULL values of categorical column
- After all the NULL values are dealt with, we drop few columns which are of no use in the analysis
- In our case we have dropped the following columns as they either add no value or has only 1 unique value to them
    - ‘Lead Number’
    - ‘Prospect ID’
    - ‘Magazine’
    - ‘Receive More Updates About Our Courses’
    - ‘Update me on Supply Chain Content’
    - ‘Get updates on DM Content’
    - ‘agree to pay the amount through cheque

Exploratory Data Analysis

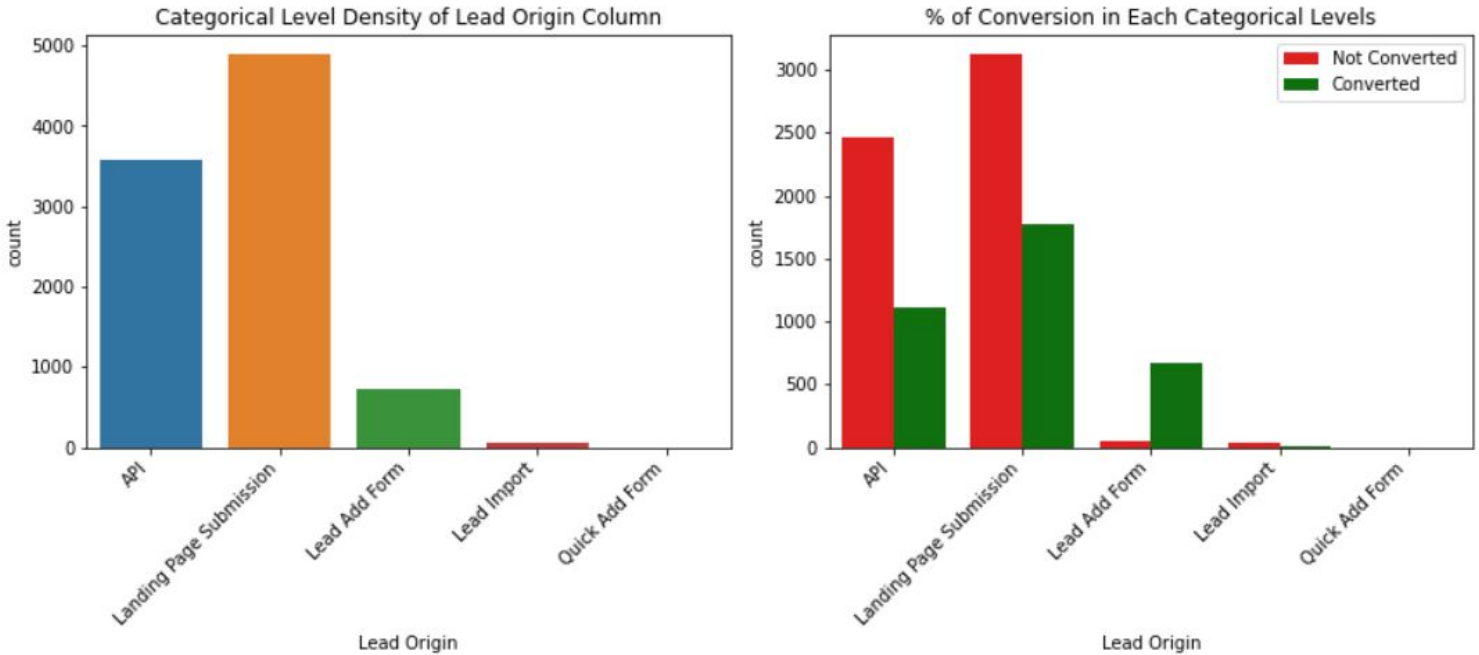
- Plotting graphs to evaluate the required variables
- Conducting Univariate and Multivariate analysis on both numerical and categorical data
- Considering the target variable 'Converted' in both categorical and numerical column for visualization for better understanding

- Performed Data imbalance check on ‘Converted’ column
- The ratio was (Not Converted to Converted) is 1.6:1



Univariate and Multivariate Categorical Analysis

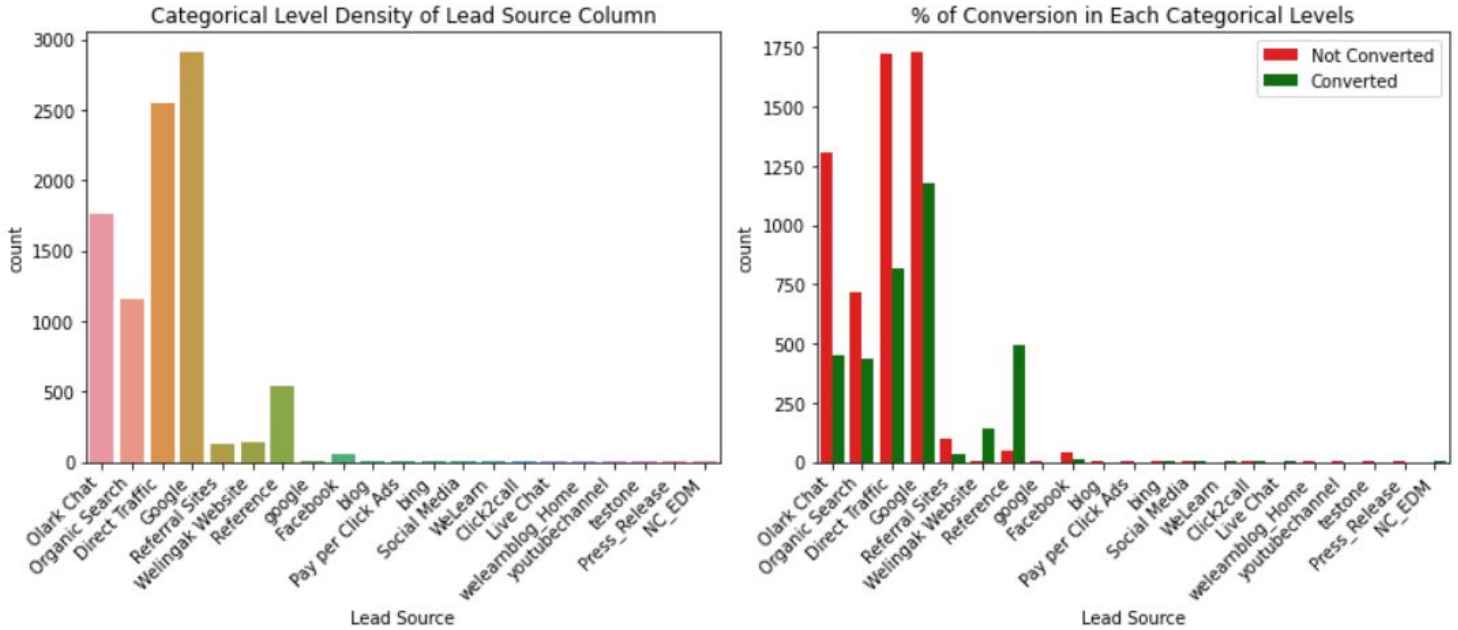
‘Lead Origin’ column



- More Conversions are happening for Landing Page Submissions and Lead Add Form

Univariate and Multivariate Categorical Analysis

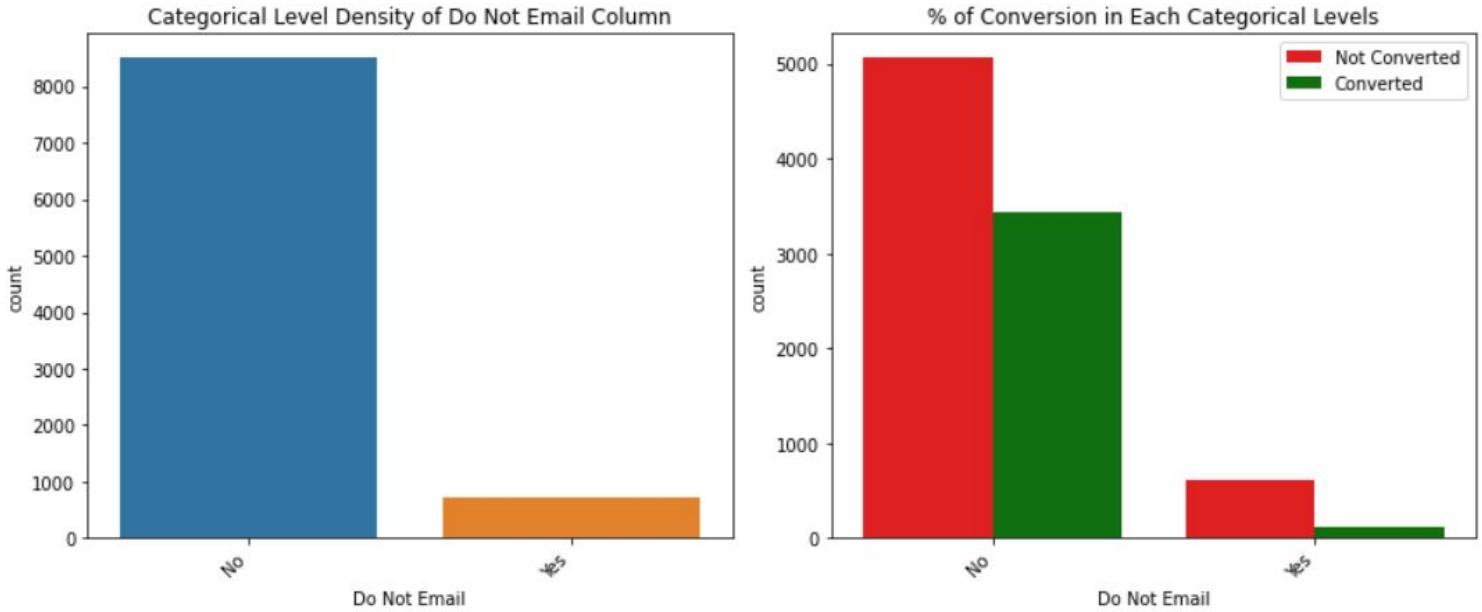
‘Lead Source’ column



- More Conversions can be seen from the leads that came from sites like Google, Organic Search, Direct Traffic and Referrals

Univariate and Multivariate Categorical Analysis

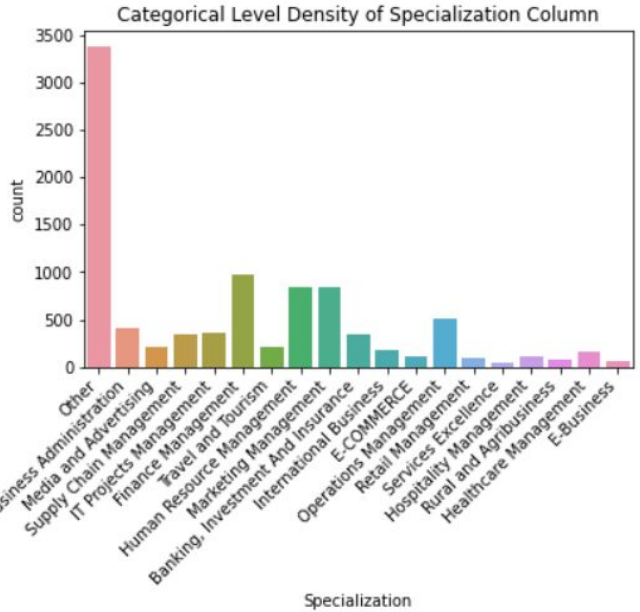
‘Do Not Email’ column

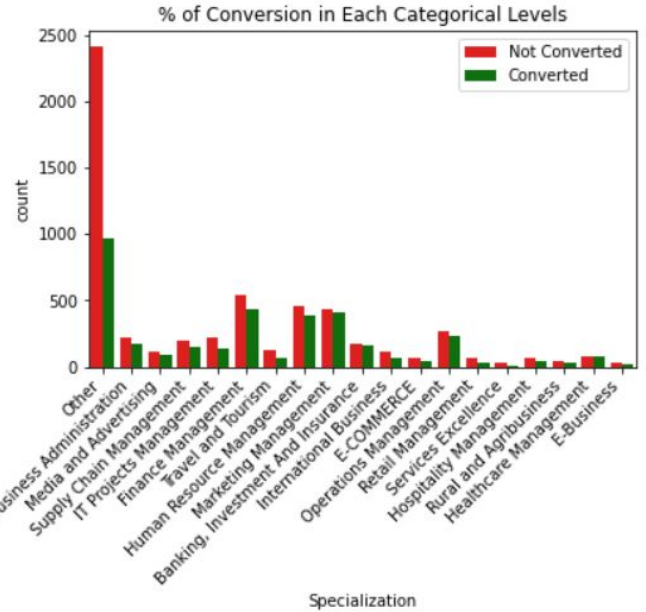


- Not much insight was drawn here
- Also doesn't make much sense as customers who have asked for not to email them have converted more

Univariate and Multivariate Categorical Analysis

‘Specialization’ column

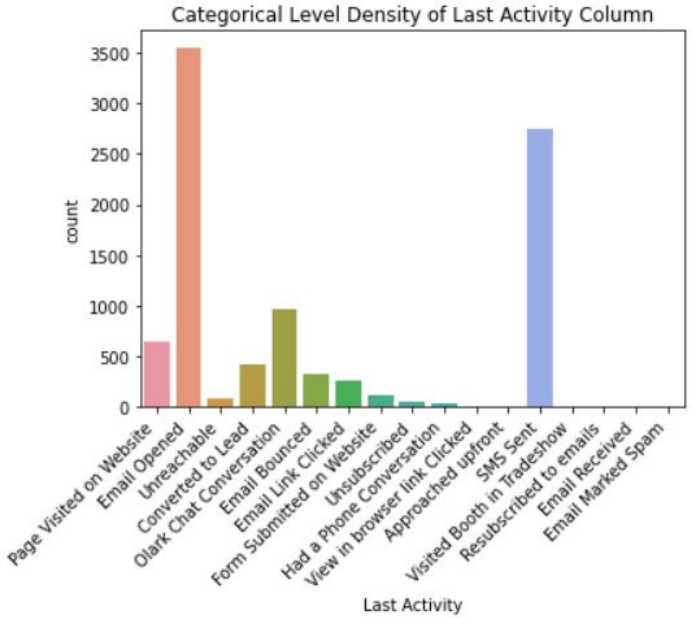


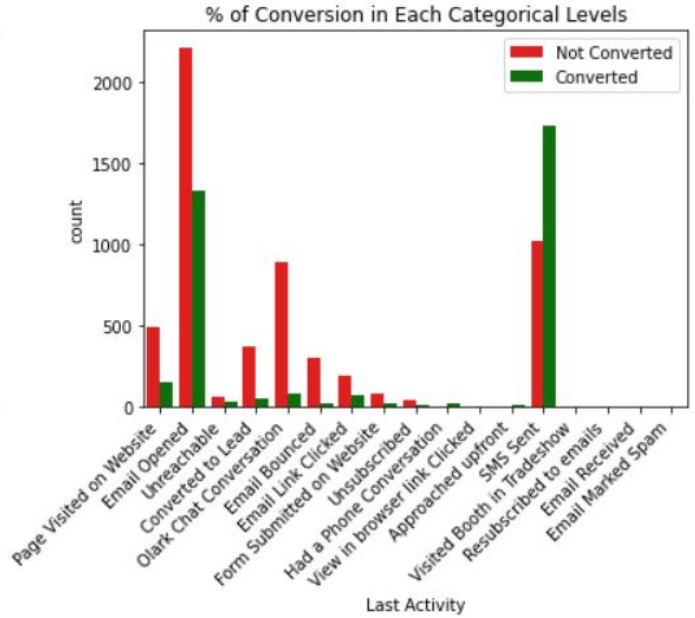


- Mostly customers who worked in the Finance, HR, Marketing, Operations and Banking sector tend to convert more

Univariate and Multivariate Categorical Analysis

‘Last Activity’ column

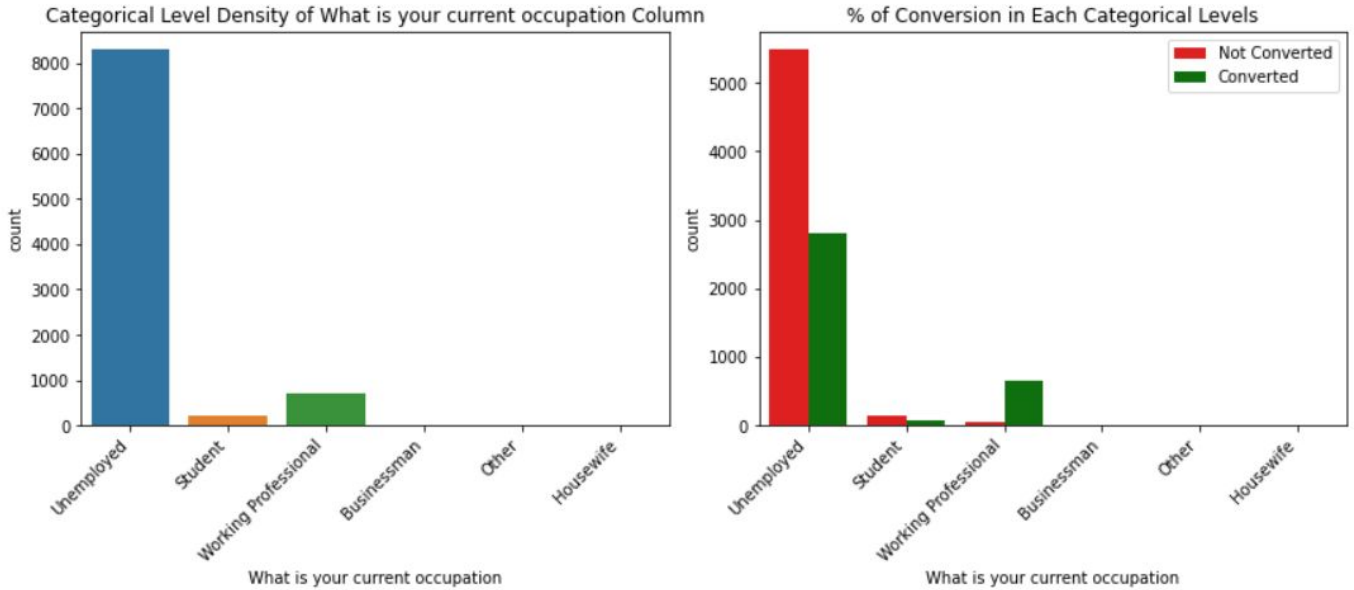




- High conversions were seen through SMS and Email marketing leads

Univariate and Multivariate Categorical Analysis

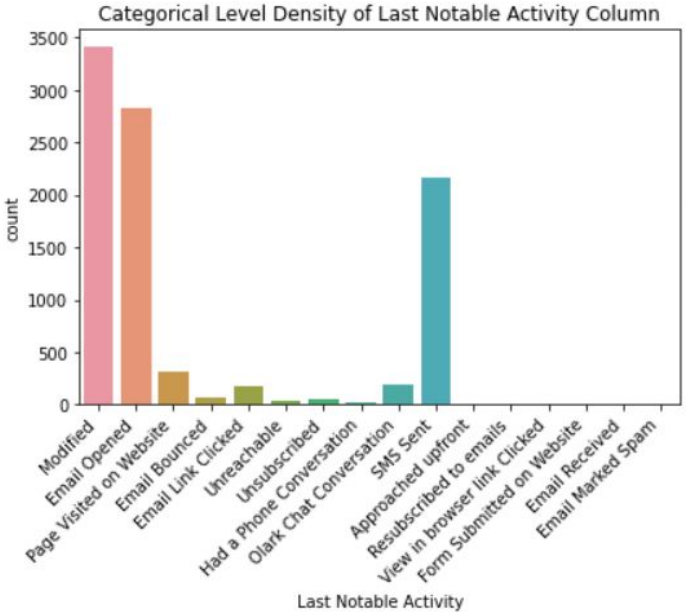
‘What is your current occupation’ column

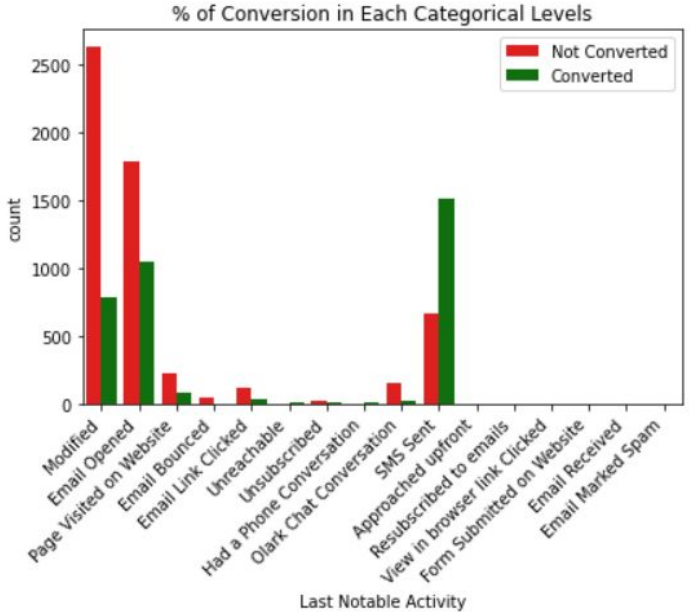


- Data Imbalance can be seen here. Considering only the unemployed segment, ~35% of them tend to convert
- Working professions also have a good conversion rate

Univariate and Multivariate Categorical Analysis

‘Last Notable Activity’ column

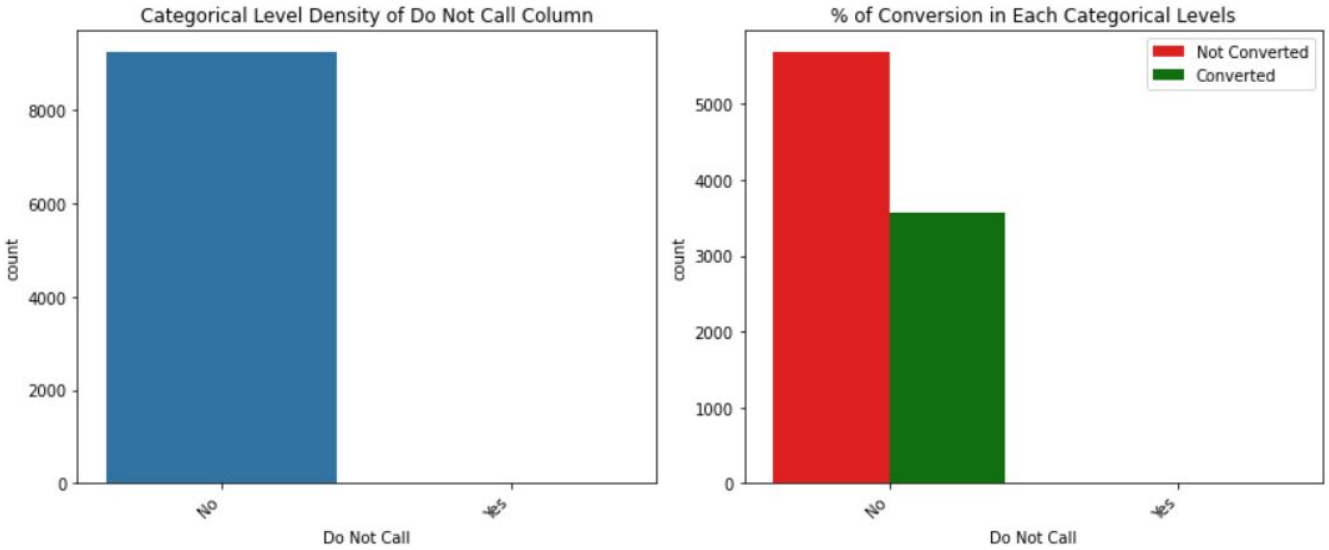




- Insights are same as that of the column Last Activity. High collinearity can be can be seen between these two columns

Univariate and Multivariate Categorical Analysis

‘Do Not Call’ column



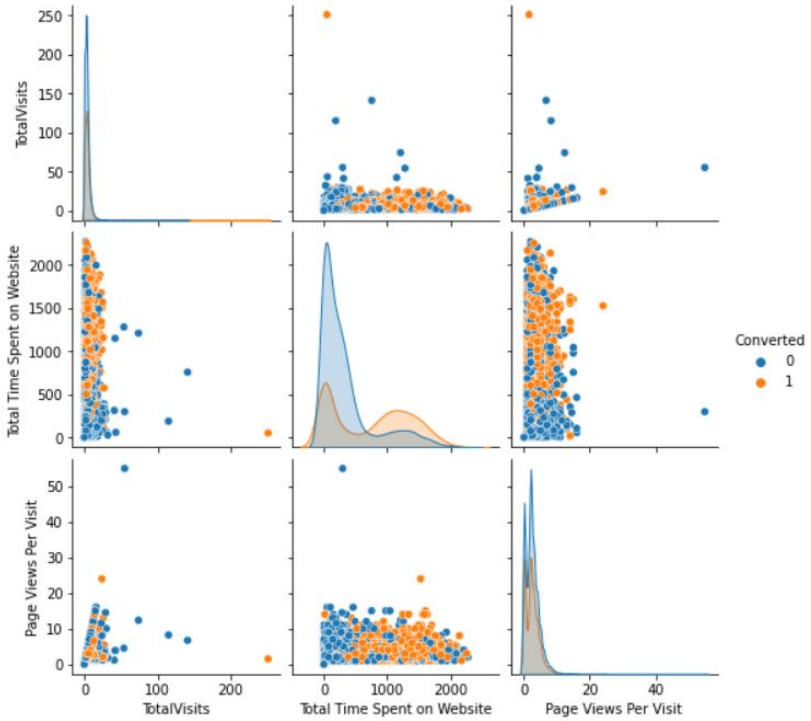
- No case drawn, same as ‘Do Not Email’ column

Lastly Drop all the columns that have very high data imbalance and won’t add much value to the model and may overfit

Univariate and Multivariate Numerical Analysis

'Converted', 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit'

- Leads tend to get converted who are likely to spend more time on website
- Lesser TotalVisits and Equal distribution of Total Time Spent on website
- Page Views Per Visit is <20 no matter how long the user spend time on website



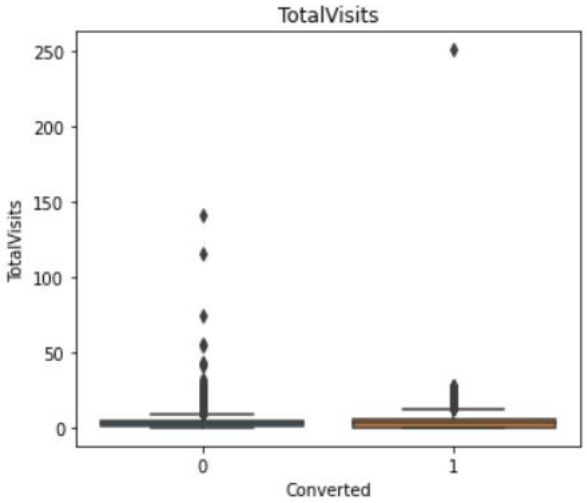
Univariate and Multivariate Numerical Analysis

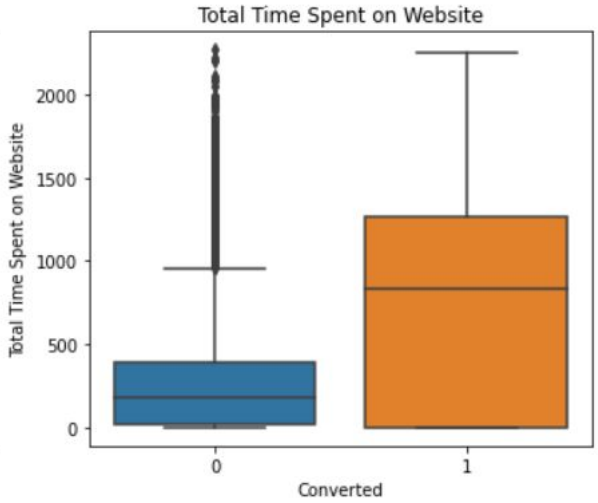
Correlation Check

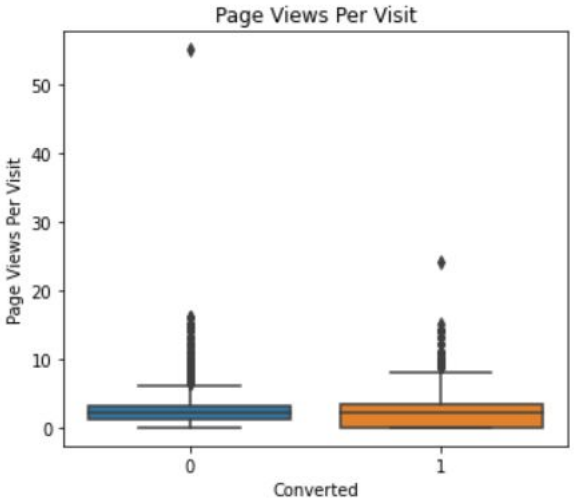
Some correlation can be seen between TotalVisits and Total Time Spent on Website

Univariate and Multivariate Numerical Analysis

Boxplot Analysis





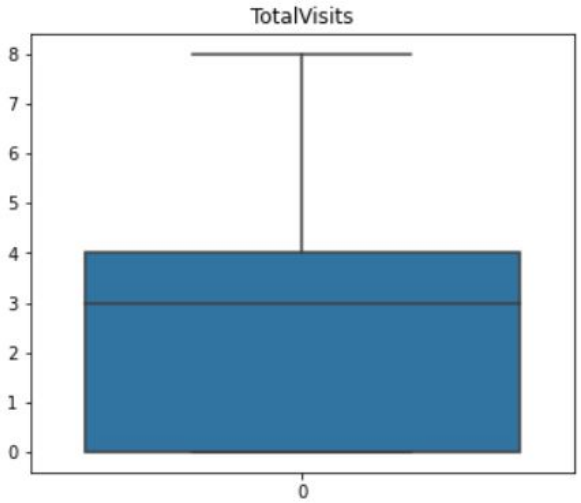


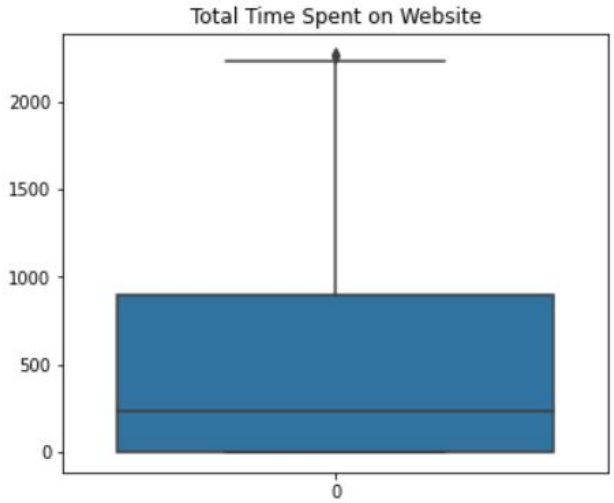
- High conversion rate can be seen on the leads which spend more time on the website before filling the lead form
- Also there are outliers that needs to be treated

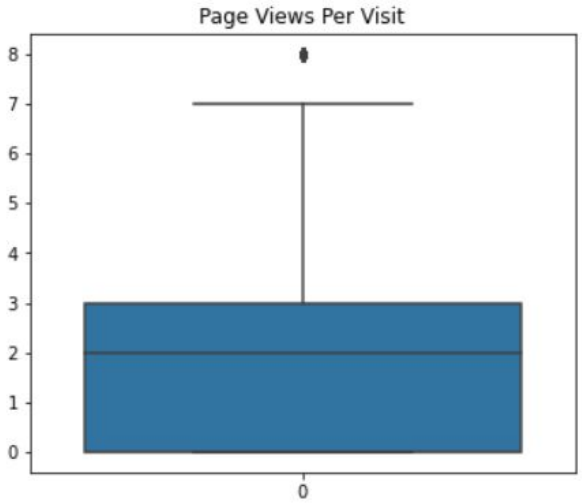
Univariate and Multivariate Numerical Analysis

Outliners Treatment

We will trim the columns upto 93 percentile and check







- Most of the outliners are handled. But we cannot reduce it completely as we may lose the data

Data Preparation

- Here we prepare the overall data for the model building
- We first collect unique values and map binary categorical columns to 0 and 1
- Then we get the dummy variables
- Heatmap cannot be performed as the number of features are more
- We then split the data in Train set(70%) and Test set(30%)
- Lastly we rescale all numerical columns using MinMax scaling

	count	mean	std	min	25%	50%	75%	max
Do Not Email	2583.0	0.075494	0.264237	0.0	0.0	0.000000	0.000000	1.000000
Converted	2583.0	0.379404	0.485333	0.0	0.0	0.000000	1.000000	1.000000
TotalVisits	2583.0	0.335075	0.272553	0.0	0.0	0.375000	0.500000	1.000000
Total Time Spent on Website	2583.0	0.202646	0.236545	0.0	0.0	0.098151	0.380282	0.979754
Page Views Per Visit	2583.0	0.270807	0.229403	0.0	0.0	0.250000	0.375000	1.000000
...	...	...	...	...	...	...	...	...
What is your current occupation_Housewife	2583.0	0.001549	0.039329	0.0	0.0	0.000000	0.000000	1.000000
What is your current occupation_Other	2583.0	0.001549	0.039329	0.0	0.0	0.000000	0.000000	1.000000
What is your current occupation_Student	2583.0	0.021680	0.145665	0.0	0.0	0.000000	0.000000	1.000000
What is your current occupation_Unemployed	2583.0	0.894696	0.307004	0.0	1.0	1.000000	1.000000	1.000000
What is your current occupation_Working Professional	2583.0	0.080139	0.271561	0.0	0.0	0.000000	0.000000	1.000000

Train set
—------->

Test Set
<—-----

	count	mean	std	min	25%	50%	75%	max
Do Not Email	6026.0	0.080319	0.271808	0.0	0.0	0.000000	0.000000	1.0
Converted	6026.0	0.379356	0.485267	0.0	0.0	0.000000	1.000000	1.0
TotalVisits	6026.0	0.338243	0.273724	0.0	0.0	0.375000	0.500000	1.0
Total Time Spent on Website	6026.0	0.209070	0.240201	0.0	0.0	0.104533	0.400418	1.0
Page Views Per Visit	6026.0	0.269999	0.225857	0.0	0.0	0.250000	0.375000	1.0
...	...	...	...	...	...	...	...	...
What is your current occupation_Housewife	6026.0	0.000996	0.031541	0.0	0.0	0.000000	0.000000	1.0
What is your current occupation_Other	6026.0	0.001494	0.038621	0.0	0.0	0.000000	0.000000	1.0
What is your current occupation_Student	6026.0	0.023731	0.152221	0.0	0.0	0.000000	0.000000	1.0
What is your current occupation_Unemployed	6026.0	0.899436	0.300776	0.0	1.0	1.000000	1.000000	1.0
What is your current occupation_Working Professional	6026.0	0.073515	0.261001	0.0	0.0	0.000000	0.000000	1.0

RFE method of selection

We first assign x-train, y_train and x_set, y_set

Then, using RFE method, we select top 16 out of 74 variables

Final data is now prepared

```
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   Do Not Email                                    6026 non-null   int64
 1   What is your current occupation_Housewife       6026 non-null   uint8
 2   Specialization_Other                            6026 non-null   uint8
 3   Last Activity_Unsubscribed                      6026 non-null   uint8
 4   Last Activity_SMS Sent                          6026 non-null   uint8
 5   Last Activity_Olark Chat Conversation           6026 non-null   uint8
 6   Last Activity_Had a Phone Conversation          6026 non-null   uint8
 7   Last Activity_Form Submitted on Website         6026 non-null   uint8
 8   Last Activity_Converted to Lead                 6026 non-null   uint8
 9   Lead Source_Welingak Website                    6026 non-null   uint8
 10  Lead Source_Olark Chat                          6026 non-null   uint8
 11  What is your current occupation_Working Professional  6026 non-null   uint8
 12  Lead Origin_Lead Add Form                       6026 non-null   uint8
 13  Lead Origin_Landing Page Submission             6026 non-null   uint8
 14  Total Time Spent on Website                     6026 non-null   float64
 15  Page Views Per Visit                            6026 non-null   float64
dtypes: float64(2), int64(1), uint8(13)
memory usage: 264.8 KB
```

Model Building - Logistic Regression

We use VIF method for the Multicollinearity check

Model 1

- 'What is your current occupation_Housewife' has very high p-value (0.99)
- Hence it can be dropped
- As all the columns have VIF <5, we shall focus on p-value now

	Features	VIF
15	Page Views Per Visit	3.856075
13	Lead Origin_Landing Page Submission	3.671051
2	Specialization_Other	2.672579
10	Lead Source_Olark Chat	2.182981
14	Total Time Spent on Website	2.062256
4	Last Activity_SMS Sent	1.582400
5	Last Activity_Olark Chat Conversation	1.492540
12	Lead Origin_Lead Add Form	1.454328
9	Lead Source_Welingak Website	1.286554
11	What is your current occupation_Working Profes...	1.208725
0	Do Not Email	1.197562
8	Last Activity_Converted to Lead	1.123609
3	Last Activity_Unsubscribed	1.082888
7	Last Activity_Form Submitted on Website	1.024828
6	Last Activity_Had a Phone Conversation	1.008573
1	What is your current occupation_Housewife	1.005756

Variance Inflation Factor

Model 2

Multicollinearity check

- There not much change in the VIF scores.
- 'Last Activity_Had a Phone Conversation' has a very high p-value.
- Hence needs to be dropped.

	Features	VIF
14	Page Views Per Visit	3.855443
12	Lead Origin_Landing Page Submission	3.671045
1	Specialization_Other	2.671950
9	Lead Source_Olark Chat	2.182457
13	Total Time Spent on Website	2.061749
3	Last Activity_SMS Sent	1.581576
4	Last Activity_Olark Chat Conversation	1.492523
11	Lead Origin_Lead Add Form	1.447024
8	Lead Source_Welingak Website	1.285658
10	What is your current occupation_Working Profes...	1.208030
0	Do Not Email	1.197508
7	Last Activity_Converted to Lead	1.123600
2	Last Activity_Unsubscribed	1.082884
6	Last Activity_Form Submitted on Website	1.024793
5	Last Activity_Had a Phone Conversation	1.008568

Variance Inflation Factor

Model 3

Multicollinearity check

- Again there not much change in the VIF scores
- Hence, let's observe p-values
- 'Page Views Per Visit' has a very high p-value.
- Hence needs to be dropped in order to make the model stable

	Features	VIF
13	Page Views Per Visit	3.853616
11	Lead Origin_Landing Page Submission	3.671003
1	Specialization_Other	2.670549
8	Lead Source_Olark Chat	2.182399
12	Total Time Spent on Website	2.061168
3	Last Activity_SMS Sent	1.578234
4	Last Activity_Olark Chat Conversation	1.492112
10	Lead Origin_Lead Add Form	1.446755
7	Lead Source_Welingak Website	1.285603
9	What is your current occupation_Working Profes...	1.204511
0	Do Not Email	1.197356
6	Last Activity_Converted to Lead	1.123274
2	Last Activity_Unsubscribed	1.082868
5	Last Activity_Form Submitted on Website	1.024721

Variance Inflation Factor

Model 4

Multicollinearity check

- Now the highest VIF value is <3 and p-values are less
- Let's drop the feature 'Last Activity_Form Submitted on Website'

	Features	VIF
1	Specialization_Other	2.316272
11	Lead Origin_Landing Page Submission	2.054427
8	Lead Source_Olark Chat	2.028502
12	Total Time Spent on Website	1.916104
3	Last Activity_SMS Sent	1.567910
4	Last Activity_Olark Chat Conversation	1.490277
10	Lead Origin_Lead Add Form	1.440454
7	Lead Source_Welingak Website	1.281038
9	What is your current occupation_Working Profes...	1.200051
0	Do Not Email	1.197072
6	Last Activity_Converted to Lead	1.111269
2	Last Activity_Unsubscribed	1.081637
5	Last Activity_Form Submitted on Website	1.024322

Variance Inflation Factor

Model 5

Multicollinearity check

- Specialization_Other was the category created by us while handling the null values
- So it is nothing but null values imputed to 'Other' category
- Here it also has high VIF and can be dropped as it is nothing but an imputed column

	Features	VIF
1	Specialization_Other	2.316221
10	Lead Origin_Landing Page Submission	2.031798
7	Lead Source_Olark Chat	2.024180
11	Total Time Spent on Website	1.913781
3	Last Activity_SMS Sent	1.557966
4	Last Activity_Olark Chat Conversation	1.488799
9	Lead Origin_Lead Add Form	1.438700
6	Lead Source_Welingak Website	1.281008
8	What is your current occupation_Working Profes...	1.200027
0	Do Not Email	1.196572
5	Last Activity_Converted to Lead	1.110043
2	Last Activity_Unsubscribed	1.081553

Variance Inflation Factor

Model 6

Multicollinearity check

- Now the VIF's are very less. Let's focus on p-values
- We drop 'Last Activity_Unsubscribed' as it has high p-value

	Features	VIF
9	Lead Origin_Landing Page Submission	1.909354
10	Total Time Spent on Website	1.771170
2	Last Activity_SMS Sent	1.504316
6	Lead Source_Olark Chat	1.444362
8	Lead Origin_Lead Add Form	1.414369
3	Last Activity_Olark Chat Conversation	1.400155
5	Lead Source_Welingak Website	1.256735
0	Do Not Email	1.180179
7	What is your current occupation_Working Profes...	1.178671
1	Last Activity_Unsubscribed	1.081306
4	Last Activity_Converted to Lead	1.057101

Variance Inflation Factor

Model 7

Multicollinearity check

- We drop 'Lead Origin_Landing Page Submission' as it has high p-value

	Features	VIF
8	Lead Origin_Landing Page Submission	1.909290
9	Total Time Spent on Website	1.769953
1	Last Activity_SMS Sent	1.500516
5	Lead Source_Olark Chat	1.444306
7	Lead Origin_Lead Add Form	1.414189
2	Last Activity_Olark Chat Conversation	1.399765
4	Lead Source_Welingak Website	1.256682
6	What is your current occupation_Working Profes...	1.178654
0	Do Not Email	1.101718
3	Last Activity_Converted to Lead	1.057044

Variance Inflation Factor

Model 8

Multicollinearity check

- We drop 'Lead Source_Welingak Website' as it has high p-value

	Features	VIF
5	Lead Source_Olark Chat	1.430037
2	Last Activity_Olark Chat Conversation	1.397671
1	Last Activity_SMS Sent	1.395993
7	Lead Origin_Lead Add Form	1.387508
8	Total Time Spent on Website	1.345348
4	Lead Source_Welingak Website	1.256609
6	What is your current occupation_Working Profes...	1.175706
0	Do Not Email	1.038399
3	Last Activity_Converted to Lead	1.029026

Variance Inflation Factor

Model 9

Multicollinearity check

- The final model is ready

	Features	VIF
4	Lead Source_Olark Chat	1.430008
2	Last Activity_Olark Chat Conversation	1.397635
1	Last Activity_SMS Sent	1.393613
7	Total Time Spent on Website	1.345348
5	What is your current occupation_Working Profes...	1.153631
6	Lead Origin_Lead Add Form	1.123152
0	Do Not Email	1.038398
3	Last Activity_Converted to Lead	1.029003

Let's do the evaluation part

Model Evaluation

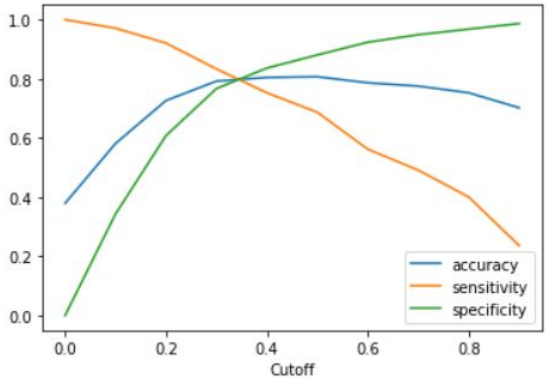
- We first check the model performance before the using optimal cutoff (default at 0.5)

```
Model Performance
```

```
Accuracy Score: 0.81
Sensitivity/Recall: 0.69
Specificity: 0.88
Precision: 0.78
F1 Score: 0.7298930729893073
```

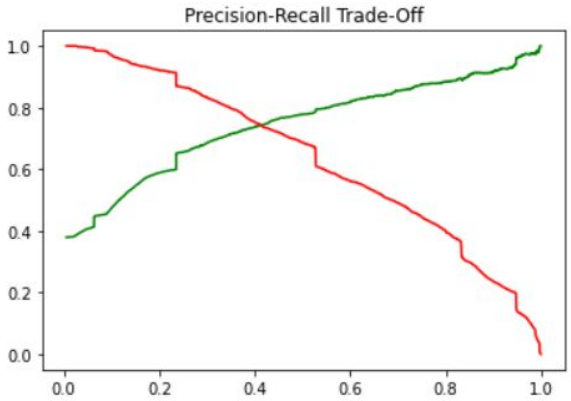
From the plot, 0.35 seem to be the optimal cutoff value

Model performance can be further improved by choosing the optimal cutoff point



Contd...

- Again we do the precision recall tradeoff to get the optimal cutoff
- From this plot, the cutoff seems to be 0.4
- Let's check the model performance for both of these cutoffs



Model performance of 0.4

```
Model Performance
-----------------------------------------------------
Accuracy Score: 0.8
Sensitivity/Recall: 0.75
Specificity: 0.84
Precision: 0.74
F1 Score: 0.7451829400303096
```

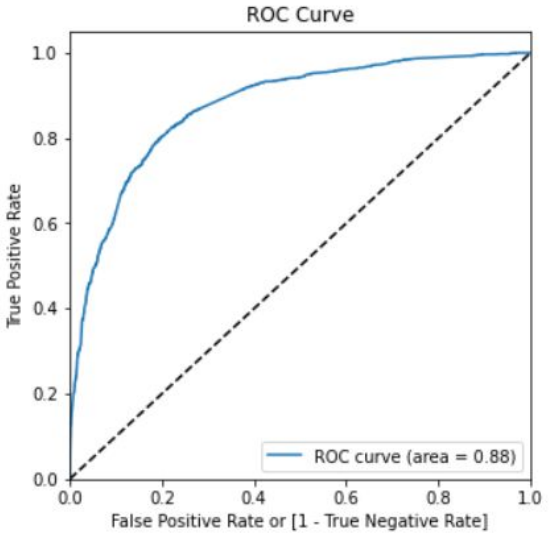
Model performance of 0.35

```
Model Performance
-----------------------------------------------------
Accuracy Score: 0.8
Sensitivity/Recall: 0.8
Specificity: 0.81
Precision: 0.71
F1 Score: 0.7541322314049588
```

As Sensitivity and F1 score is high for cutoff point '0.35', lets choose this as the optimal cutoff value

Contd...

- We then then check the ROC curve for the final model
- Area under the ROC curve is 0.88 which is good
- We then assign the lead scores and the predicted values to the final train data



Train data has an:-

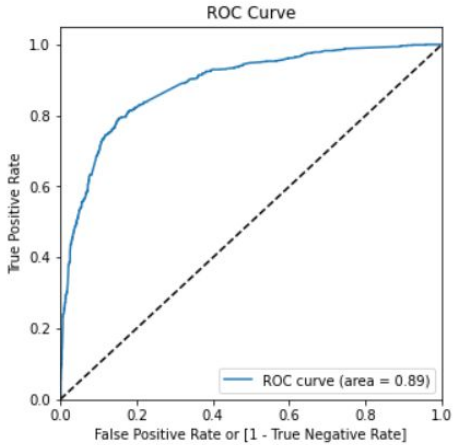
- accuracy score of 0.8
- Sensitivity: 0.8
- Specificity: 0.81
- Precision: 0.71
- F1 Score: 0.75

Model Testing

- We go with those columns which were finalized in column 9
- We then add constants
- Fit the logistic regression model

```
Model Performance

-----------------------------------------------------------
Accuracy Score: 0.82
Sensitivity/Recall: 0.8
Specificity: 0.84
Precision: 0.75
F1 Score: 0.7711319822046465
```



- The model performance is good on test data
- Area under the ROC curve is 0.89.
- Hence, our model is good is with high accuracy score, sensitivity and ROC curve area

Model Inference

Features Used in the Final Model:- 'Do Not Email', 'Last Activity_SMS Sent', 'Last Activity_Olark Chat Conversation', 'Last Activity_Converted to Lead', 'Lead Source_Olark Chat', 'What is your current occupation_Working Professional', 'Lead Origin_Lead Add Form', 'Total Time Spent on Website'

Optimal Cutoff: 0.35

Trained model has an:-

- Accuracy Score: 80%
- Sensitivity: 80%
- Specificity: 81%
- Precision: 71%
- F1 Score: 75%
- ROC Curve Area: 88%

Tested model has an:-

- Accuracy Score: 82%
- Sensitivity: 80%
- Specificity: 84%
- Precision: 75%
- F1 Score: 77%
- ROC Curve Area: 89%

Top 3 predictors which are impacting the lead conversion rate are:-

1. **Total Time Spent on Website:-** How long the customer spend time on our website before he/she fills up the lead form
2. **Lead Origin_Lead Add Form:-** Leads coming from the source 'Lead Add Form' tend to have high conversion rate.
3. **What is your current occupation_Working Professional:-** Working professionals have high impact in the lead conversion rate.

Summary (Recommendations)

Sales team need to keep the following things in mind:-

- More Conversions are happening for Landing Page Submissions and Lead Add Form.
- More Conversions can be seen from the leads that came from sites like Google, Organic Search, Direct Traffic and Referrals.
- High conversions were seen through SMS and Email marketing leads.
- Mostly customers who worked in the Finance, HR, Marketing, Operations and Banking sector tend to convert more.
- Customers who have chosen the option of "Better Career Prospects" for the career outcome tend to convert more.
- Leads tend to get converted more who are likely to spend more time on website. If a customer spends more time on the website, he/she is researching a lot and is interested in the course.
- Try to reduce the bounce rate of the website as the customer engagement time increases, the chances of him/her getting converted will also be high.
- High importance needs to be given for the leads that came from Lead Add Form. Also this form can be used across key areas in order to generate qualifying leads.
- Sales team needs to focus on working professionals as they dont have any financial restrictions and also they tend to convert more by enrolling to the courses.
- Lead Score can be taken into consideration for giving the importance to each leads. Lead score having >0.35 tend to convert more. Model **accuracy score is 80%** and hence can be relied.

THANK YOU

Link to our project GitHub repo: Click here