Lead Scoring Case Study

Logistic Regression

Performed by:
- Mohammed Hussain Chitapulla
- Sweta Singh

Problem Statement

The purpose of this case study is to provide a comprehensive research and build a logistic regression model which can help X education select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The case study aims to assign lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. 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.

Analysis Approach

Data Cleaning

- Data Inspection
- Dealing Missing values
- Unique value checks
- Outlier Analysis
- Check on % rows left after cleaning

EDA

- Univariate Analysis
- Bivariate Analysis
- Correlation Matrix

Data Prep

- Creating dummies for categorical vars
- Train-Test split
- Scaling numerical vars
- Collinearity through heatmap

Data Modelling

- Variable selection through RFE
- Model Creation
- ROC Curve
- Predictions
- Metric on train data
- Confusion Matrix
- Metric on test data

Data Cleaning

Missing Values

A significant number of columns contain missing values which needs to be handled

Strategy used for imputation of Missing Values

Categorical columns

1) Checked for actual null value count column wise
2) Deleted columns with missing values **more than 45%**
3) Checked row wise null values and handled
4) Checked for unique values in categorical columns
5) Deleted category columns with less than 2 unique categories

```
Magazine ['No'] 1

Receive More Updates About Our Courses ['No'] 1

Update me on Supply Chain Content ['No'] 1

Get updates on DM Content ['No'] 1

I agree to pay the amount through cheque ['No'] 1
```

6) Checked for null values in other forms : Observed **SELECT** value in columns and dealt with it as a null value.
Column list : *'Specialization', 'How did you hear about X Education', 'Lead Profile', 'City'*

7) **Less than 13% missing values**
Imputed value with mode for categorical columns : *'Lead Source' , 'Last Activity'*

8) **Between than 13% to 40% missing values**
Created new category as "Not Available"

Merging multiple categories into single

The columns which consists of large number of categories with multiple categories having very less value counts could lead to unnecessary increase in number of dummy variables.
Therefore merged them into single category named "Others"

1) What is your current occupation

Unemployed	85.496183
Working Professional	10.778626
Student	3.206107
Other	0.244275
Housewife	0.152672
Businessman	0.122137

'Other', 'Housewife' , 'Businessman' merged to *'Others'*

Unemployed	60.608686
Not available	29.102134
Working Professional	7.646485
Others	2.642695

2) *What is your current occupation*

India	95.766337
United States	1.017849
United Arab Emirates	0.781826
Singapore	0.354035
Saudi Arabia	0.309780
United Kingdom	0.221272
Australia	0.191769
Qatar	0.147514
Hong Kong	0.103260
Bahrain	0.103260
France	0.088509
Oman	0.088509
unknown	0.073757
Canada	0.059006
South Africa	0.059006
Germany	0.059006
Kuwait	0.059006
Nigeria	0.059006
Sweden	0.044254
China	0.029503
Asia/Pacific Region	0.029503
Philippines	0.029503

- 95% of total leads are from "***India"*** *therefore created category* ***"Outside India"*** *merging all other countries*

India	70.248023
Not available	26.654392
Outside India	3.097585

Missing value treatment *Numerical Variables*

1) ***Page Views Per Visit*** *and* ***TotalVisits***: For both the numerical columns less than 13% missing values found.

.

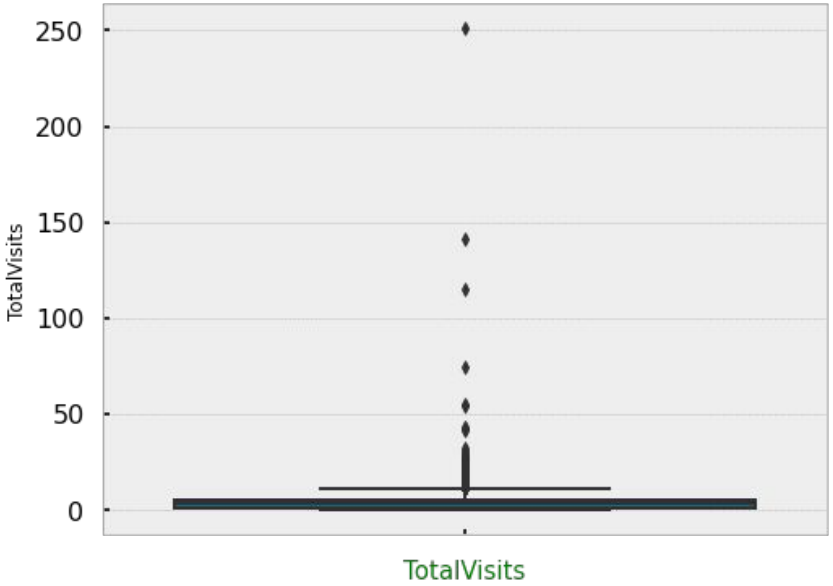
```
**Page Views Per Visit**
Percentage of missing values 1.48
count    9103.000000
mean        2.362820
std         2.161418
min         0.000000
50%         2.000000
90%         5.000000
95%         6.000000
99%         9.000000
max        55.000000
```

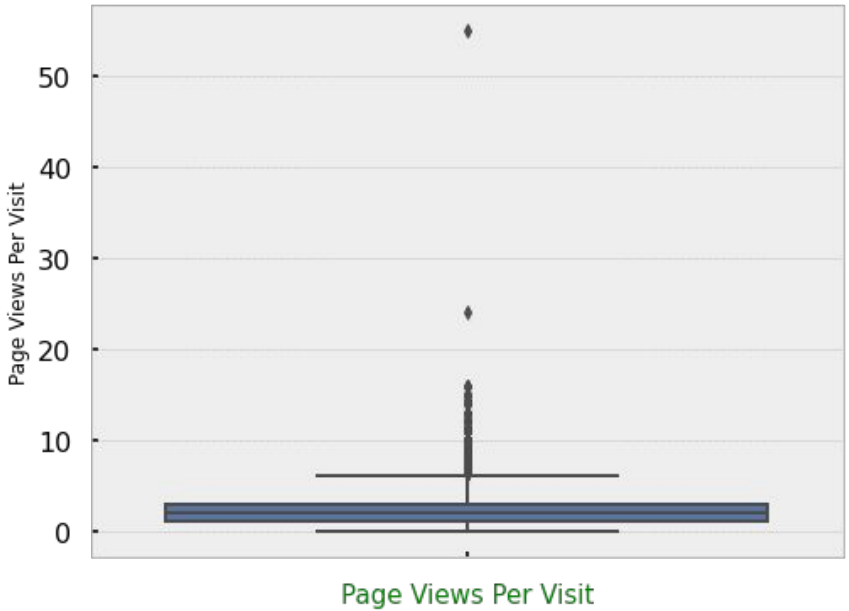
```
**TotalVisits**
Percentage Missing value 1.48
count    9103.000000
mean        3.445238
std         4.854853
min         0.000000
50%         3.000000
90%         7.000000
95%        10.000000
99%        17.000000
max       251.000000
```

****Median*** *method is used to input the missing values due to presence of outliers in the data.*

Outlier Analysis

Outlier analysis is performed on numerical columns using the box plots : "***Page Views Per Visit***", '***TotalVisits***'



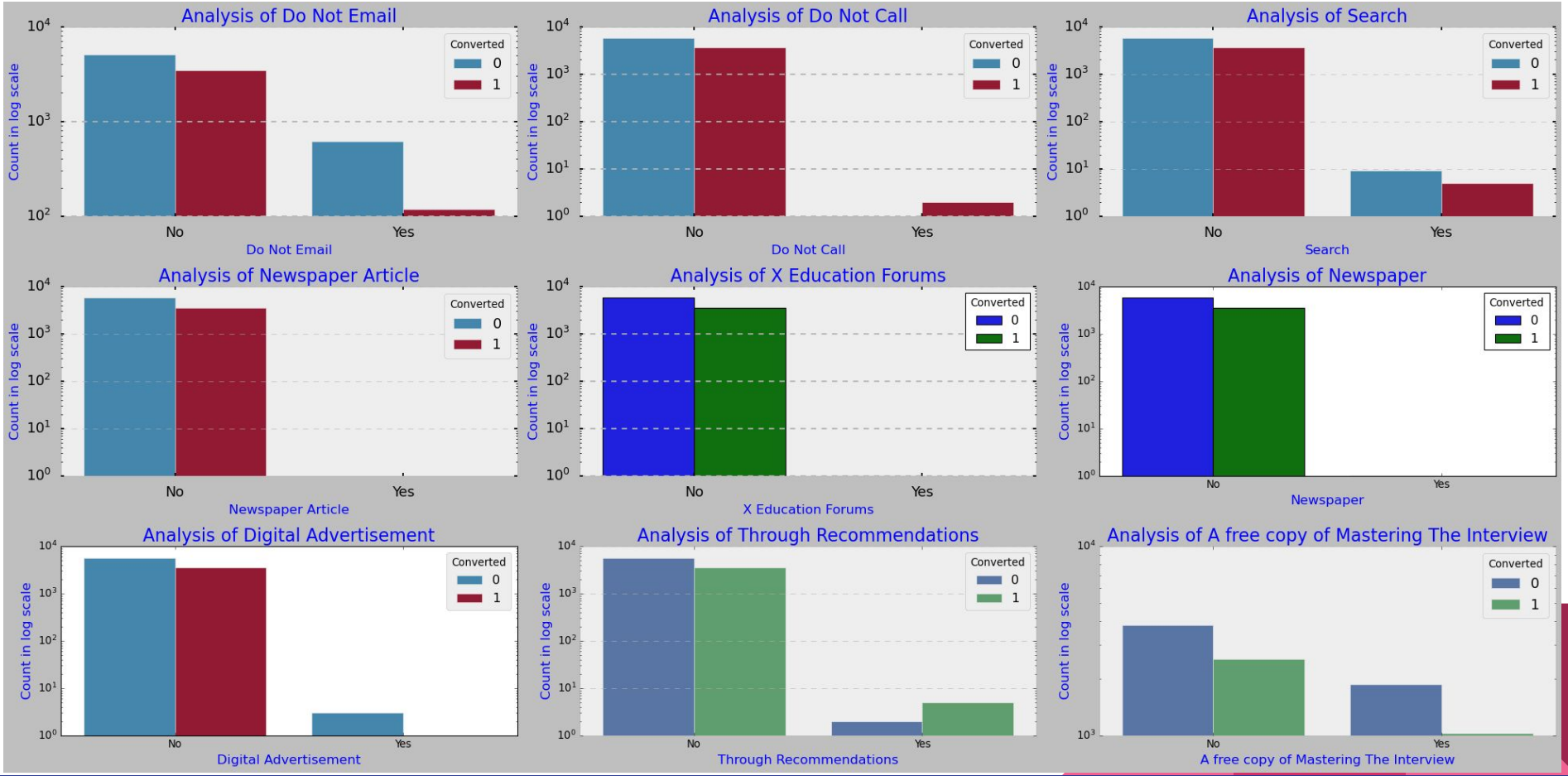


1. There are Extreme outliers in the column after 50
2. Outliers are continuous from 17 to 30

1. After removing outlier we can observe the median is at 2
2. There are no extreme outliers in the data

Exploratory Data Analysis

Binary Categorical Variables VS Target Variable



Inferences from previous plot

1. Most of the converted leads prefer to receive Email about the course

2. Both converted and non-converted leads do not prefer to recieve Calls

3. Both converted and non-converted leads have not found Seen the Ad through Search

4. Both leads have not seen the add through Newspaper Article

5. Both leads have not seen the Ad through X education forums

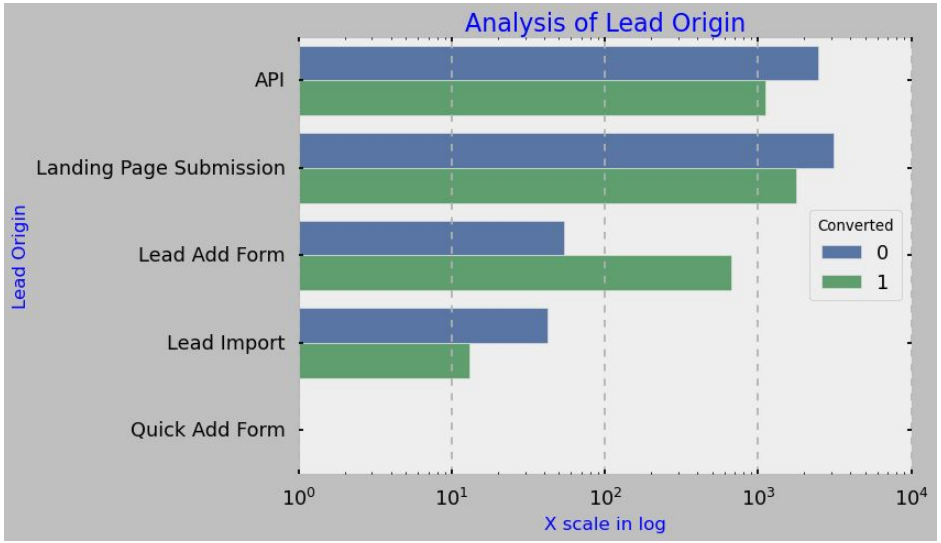
6. Both leads have not seen the Ad through Newspaper

7. Both leads have not seen the Ad through Digital Advertisement, very few non converted leads have seen the Ad through Digital Advertisement

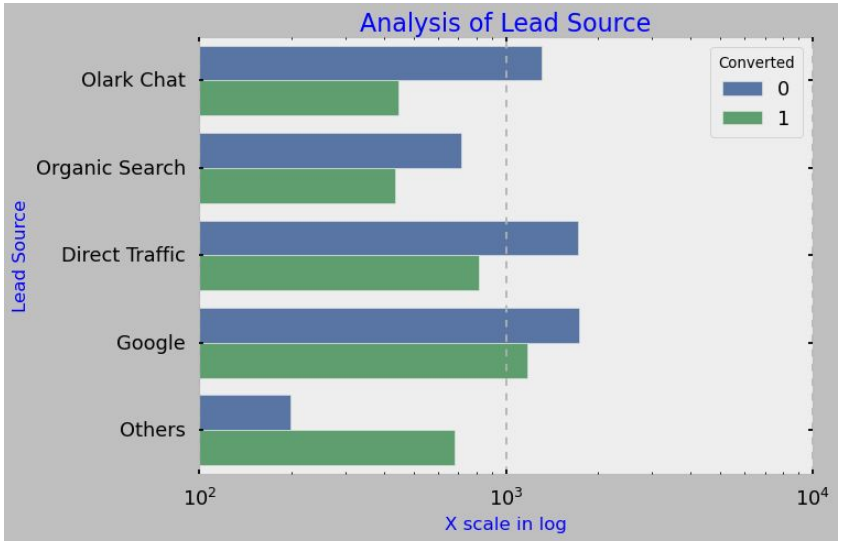
8. We can Observe most of the Converted Leads are from Non recommendation, we can also observe the conversion of leads through recommendation is higher than non converted leads

9. Most of the converted leads do not prefer copy of mastering the interview, whereas Non converted leads prefer a copy

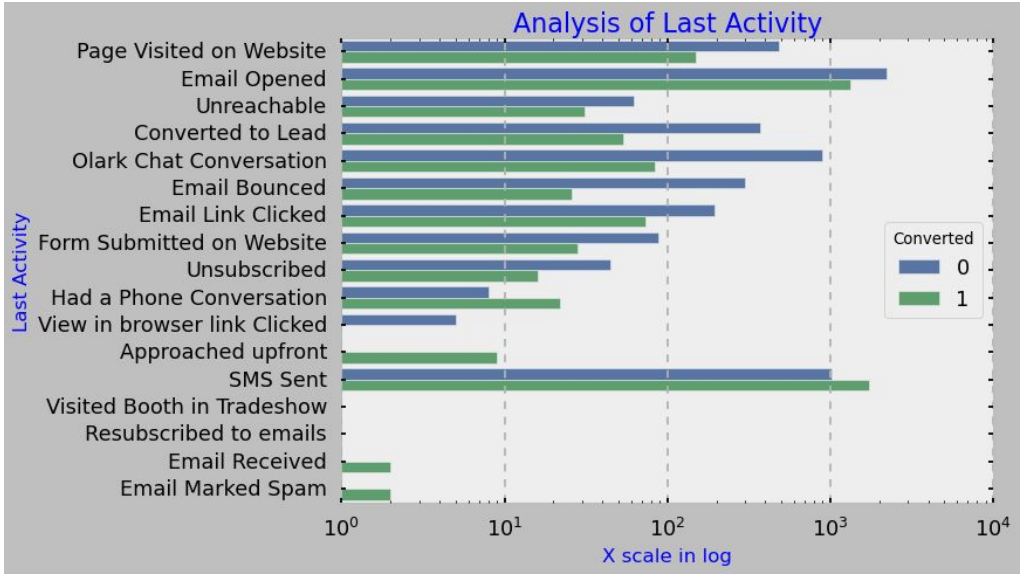
Analysis of Categorical Variables

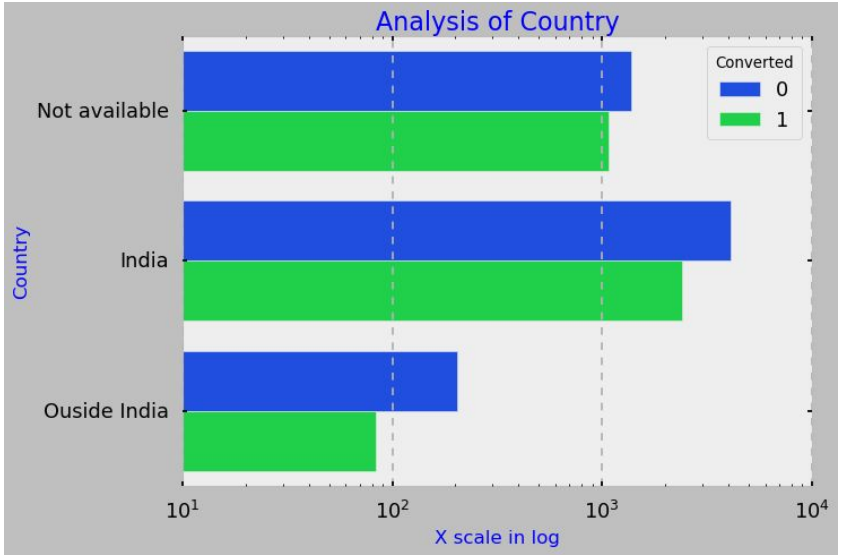


1. Lead Add form has high converted lead, when compared to non converted lead for the same.
2. Landing page submission has the highest converted lead across all converted lead source.



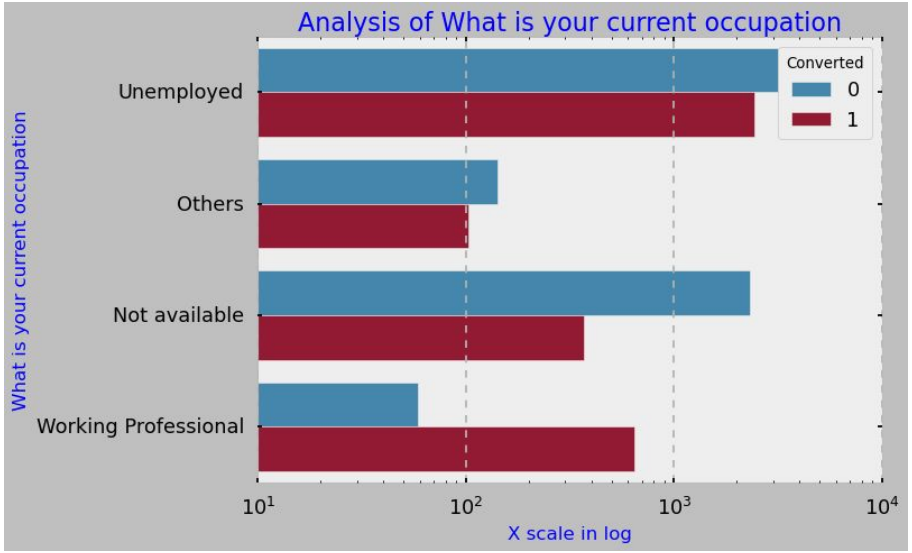
1. Lead Source through Google has the highest converted leads
2. Lead Source through Direct Traffic has the highest non convertead leads

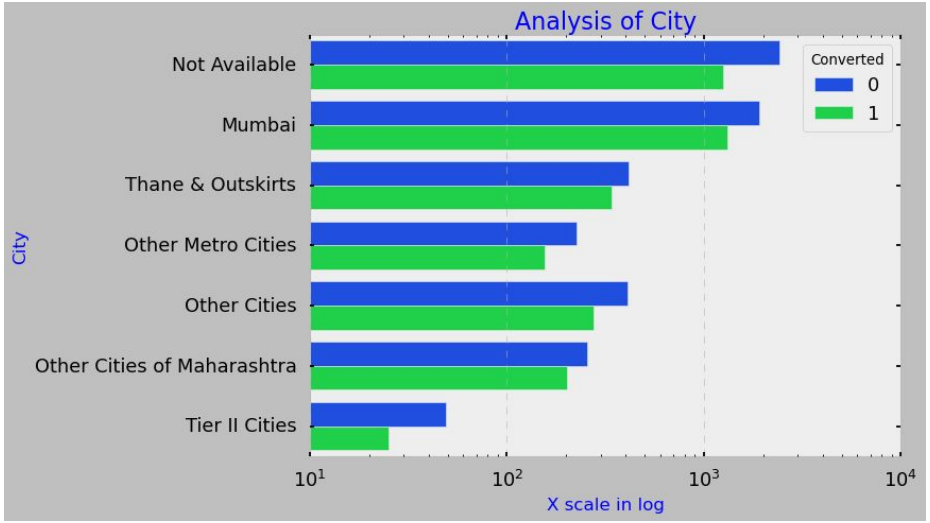




1. Lead Add form has high converted lead, when compared to non converted lead for the same.
2. Landing page submission has the highest converted lead across all converted lead source.

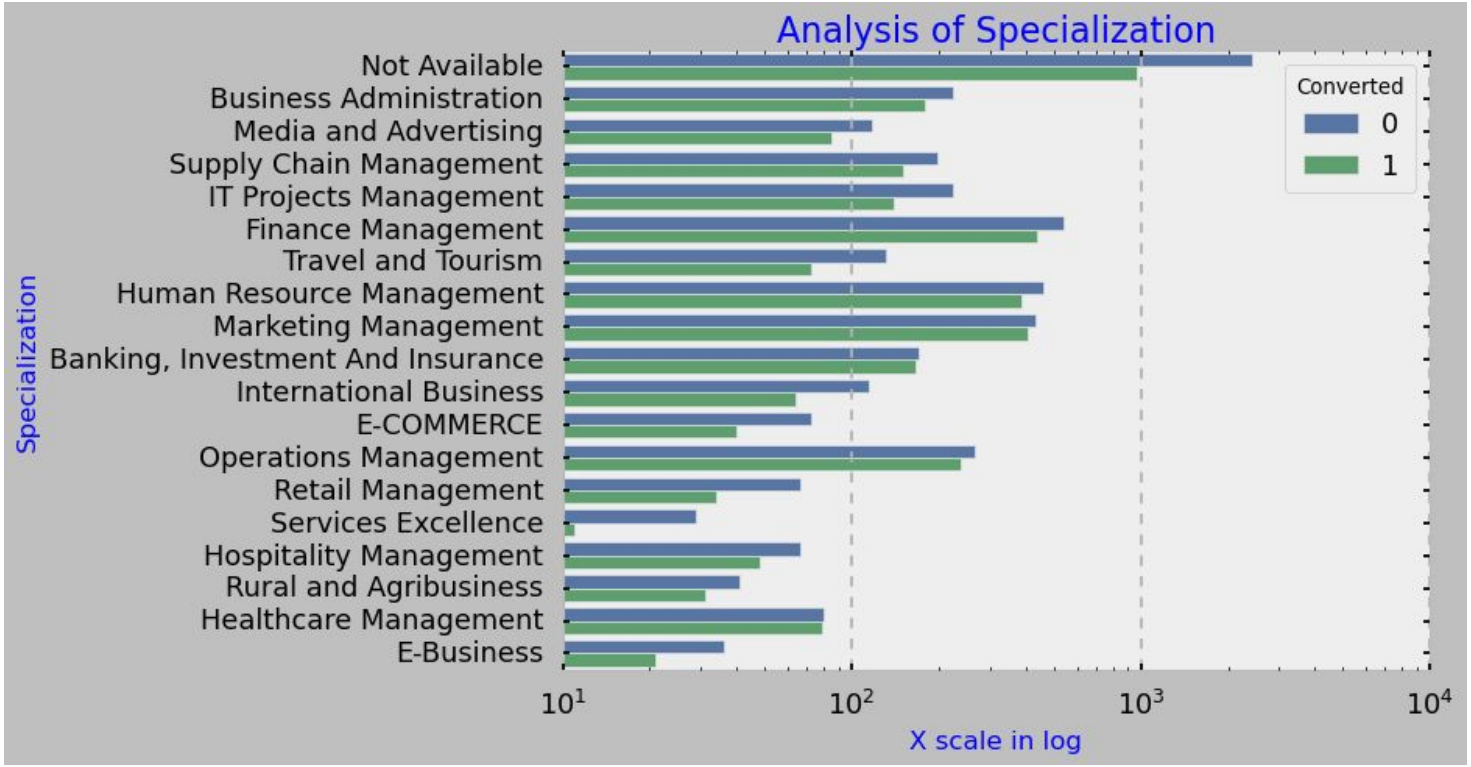
1. Country as India has the highest converted leads
2. Converted leads are least for outside India





1. Unemployed leads have the highest lead conversion count.
2. Working professional have highest lead conversion percentage

1. Mumbai has the highest Converted leads
2. Tier 2 cities have the least converted leads

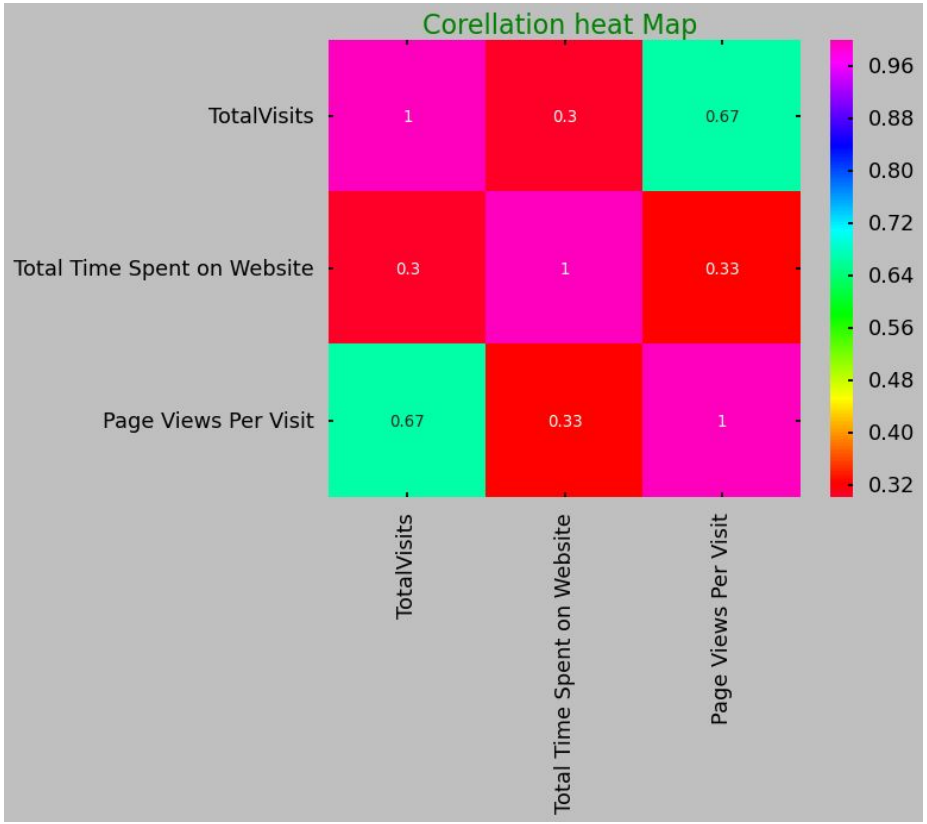


1. Converted leads are maximum where Specialization is not specified.

2. Specialization such as Service Excellence, E-Business and Rural and Agribusiness has the least converted leads

3. Finance Management Specialization has the highest converted leads among specified specialization

Analysis of numerical variables: Correlation Heatmap



1. Total visit and Page Views per visit have the highest correlation.
2. Total time spent on website has weak correlation with Total visit and Page Views Per Visit

Data Preparation

1. **Dummifying Categorical variables** : Converting categorical variables to numerical type where we have each category as column name with values 1 for positive occurrences and 0 for no occurrences. (examples in below screenshots)

   'Lead Origin', 'Lead Source', 'Do Not Email', 'Do Not Call', 'Country', 'Specialization', 'What is your current occupation', 'Search', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'City', 'A free copy of Mastering The Interview'

Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Origin_Quick Add Form
0	0	0	0
0	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0

Lead Source_Google	Lead Source_Olark Chat	Lead Source_Organic Search	Lead Source_Others
0	1	0	0
0	0	1	0
0	0	0	0
0	0	0	0
1	0	0	0

2. **Splitting the data into testing and training the dataset**

Divided the dataset into training data with 70% using which model will be constructed,
Test Data 30% which will be used to test the built model.
Random state chosen as 100

3. **Scaling numerical variables :** numerical variables are scaled to fit in the range [0,1]
Using *MinMaxScaler*

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

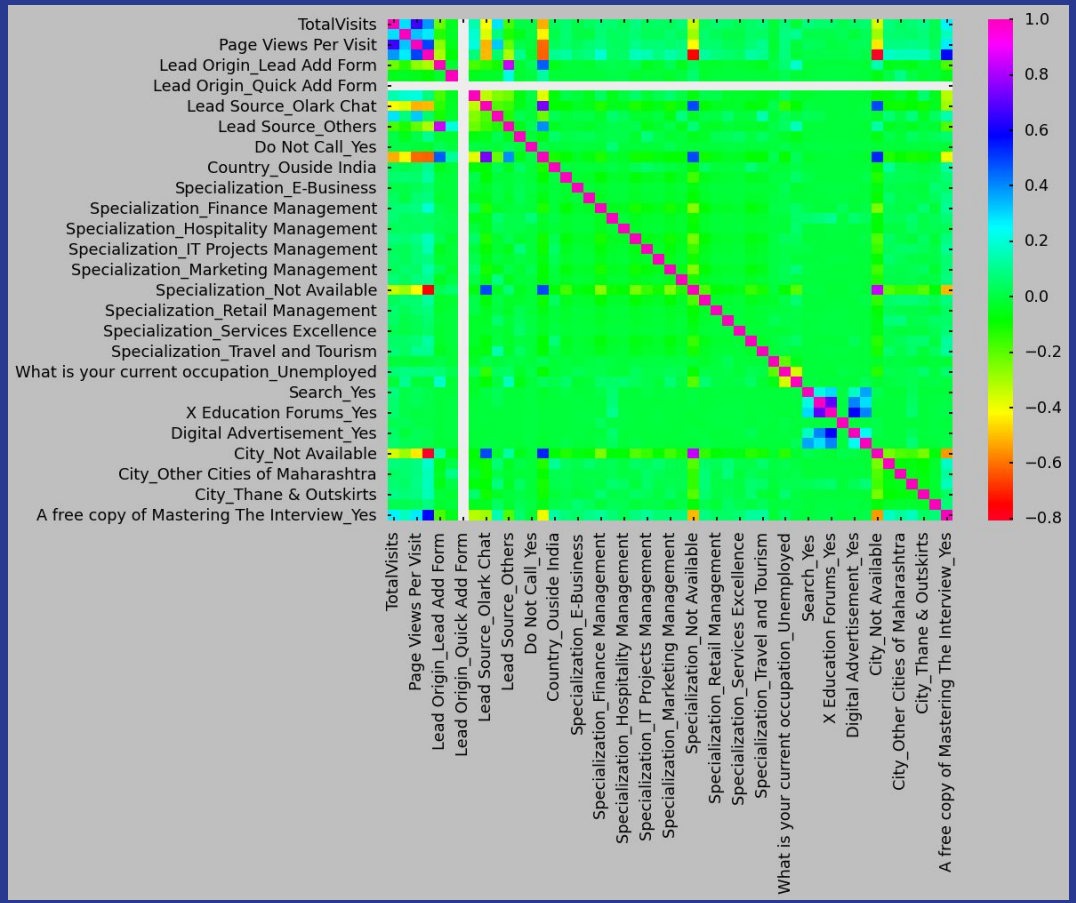
TotalVisits	Total Time Spent on Website	Page Views Per Visit
0.0	0	0.0
5.0	674	2.5
2.0	1532	2.0
1.0	305	1.0
2.0	1428	1.0

Before Conversion

TotalVisits	Total Time Spent on Website	Page Views Per Visit
0.139535	0.606074	0.37500
0.116279	0.196303	0.15625
0.000000	0.000000	0.00000
0.093023	0.704225	0.25000
0.023256	0.066461	0.06250

After Conversion

Checking collinearity using *HeatMap*



1. From the plot we can observe most of the variables have weak or no correlation

2. *total visit and page view per visit* have strongly correlation.

3. *Specialization_Not Available and City_Not Available* have high correlation.

Data Modelling

Variable Selection using RFE

Number of columns after the dummy variable creation went to 50. Used RFE to choose 20 variables based on importance.

Variables supported by RFE:

```
Index(['TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit',
       'Lead Origin_Landing Page Submission', 'Lead Origin_Lead Add Form',
       'Lead Origin_Lead Import', 'Do Not Email_Yes', 'Country_Not available',
       'Specialization_E-Business', 'Specialization_Hospitality Management',
       'Specialization_Not Available', 'Specialization_Retail Management',
       'Specialization_Services Excellence',
       'What is your current occupation_Others',
       'What is your current occupation_Unemployed',
       'What is your current occupation_Working Professional',
       'Newspaper Article_Yes', 'Newspaper_Yes', 'Digital Advertisement_Yes',
       'City_Not Available'],
      dtype='object')
```

```
(9233, 50)
```

Manual Selection of variables for Model

The model is created and then on each iteration one variable is dropped based on rules mentioned below until the desired results are obtained.

1. After running the model **p-value** and **VIF** factor are used to make decision for dropping the variables one by one.

2. Rules applied: ***p-value <=*** *0.05 And* ***VIF*** *< 5*

3. Rules for combination of p-value and VIF values
   - If both p-value and VIF high simultaneously : compare the variables then drop for highest one
   - If one is high and other is low: drop priority will be given to higher p-value
   - if both low : do not drop

(9233, 50)

Logistic model regression results after First trial

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9757	0.166	-11.919	0.000	-2.301	-1.651
TotalVisits	1.9927	0.525	3.795	0.000	0.964	3.022
Total Time Spent on Website	4.5498	0.162	28.090	0.000	4.232	4.867
Page Views Per Visit	-0.7244	0.373	-1.941	0.052	-1.456	0.007
Lead Origin_Landing Page Submission	-1.0301	0.131	-7.870	0.000	-1.287	-0.774
Lead Origin_Lead Add Form	2.6048	0.191	13.672	0.000	2.231	2.978
Lead Origin_Lead Import	-1.9189	0.547	-3.511	0.000	-2.990	-0.848
Do Not Email_Yes	-1.3487	0.160	-8.423	0.000	-1.663	-1.035
Country_Not available	1.0429	0.125	8.338	0.000	0.798	1.288
Specialization_E-Business	-0.5625	0.449	-1.254	0.210	-1.442	0.317
Specialization_Hospitality Management	-0.8191	0.327	-2.505	0.012	-1.460	-0.178
Specialization_Not Available	-0.6422	0.150	-4.283	0.000	-0.936	-0.348
Specialization_Retail Management	-0.3651	0.319	-1.144	0.253	-0.991	0.261
Specialization_Services Excellence	-0.5011	0.572	-0.876	0.381	-1.622	0.620
What is your current occupation_Others	1.2883	0.207	6.217	0.000	0.882	1.694
What is your current occupation_Unemployed	1.2319	0.085	14.543	0.000	1.066	1.398
What is your current occupation_Working Professional	3.5734	0.191	18.694	0.000	3.199	3.948
Newspaper Article_Yes	21.8591	2.11e+04	0.001	0.999	-4.13e+04	4.13e+04
Newspaper_Yes	-24.1794	4.82e+04	-0.001	1.000	-9.45e+04	9.44e+04
Digital Advertisement_Yes	-42.2081	2.98e+04	-0.001	0.999	-5.84e+04	5.84e+04
City_Not Available	-0.5540	0.163	-3.409	0.001	-0.873	-0.235

p-value of **Newspaper_Yes** is very high and hence dropped the variable and re-creating the model

Total 8 trials are run for the model iteratively dropping one variable in each trial and re-creating the model

Trial	Variable dropped	p-value	VIF
1	Newspaper_Yes	0.999	-
2	Digital Advertisement_Yes	0.999	1.20
3	Newspaper Article_Yes	0.588	1.00
4	Specialization_Services Excellence	0.382	1.01
5	Specialization_Retail Management	0.259	1.01
6	Specialization_E-Business	0.218	1.01
7	City_Not Available	0.000	7.46

Model summary in 8th trial

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.1890	0.155	-14.167	0.000	-2.492	-1.886
TotalVisits	2.0227	0.519	3.897	0.000	1.005	3.040
Total Time Spent on Website	4.5293	0.161	28.087	0.000	4.213	4.845
Page Views Per Visit	-0.7094	0.371	-1.913	0.056	-1.436	0.017
Lead Origin_Landing Page Submission	-0.8422	0.117	-7.176	0.000	-1.072	-0.612
Lead Origin_Lead Add Form	2.6376	0.190	13.871	0.000	2.265	3.010
Lead Origin_Lead Import	-1.4922	0.535	-2.791	0.005	-2.540	-0.444
Do Not Email_Yes	-1.3683	0.160	-8.577	0.000	-1.681	-1.056
Country_Not available	1.0073	0.124	8.106	0.000	0.764	1.251
Specialization_Hospitality Management	-0.8048	0.326	-2.469	0.014	-1.444	-0.166
Specialization_Not Available	-0.9440	0.118	-8.016	0.000	-1.175	-0.713
What is your current occupation_Others	1.3213	0.207	6.391	0.000	0.916	1.727
What is your current occupation_Unemployed	1.2396	0.085	14.659	0.000	1.074	1.405
What is your current occupation_Working Professional	3.5508	0.192	18.513	0.000	3.175	3.927

	Features	VIF
2	Page Views Per Visit	4.90
3	Lead Origin_Landing Page Submission	3.57
0	TotalVisits	3.53
11	What is your current occupation_Unemployed	2.87
7	Country_Not available	2.85
9	Specialization_Not Available	2.50
1	Total Time Spent on Website	2.17
4	Lead Origin_Lead Add Form	1.65
12	What is your current occupation_Working Profes...	1.36
6	Do Not Email_Yes	1.11
10	What is your current occupation_Others	1.08
5	Lead Origin_Lead Import	1.03
8	Specialization_Hospitality Management	1.02

***** All p-values and VIFs fall in range. Model achieved in 8th trial***

Desired model achieved in 8th trial

Final attributes:

```
Index(['const', 'TotalVisits', 'Total Time Spent on Website',
       'Page Views Per Visit', 'Lead Origin_Landing Page Submission',
       'Lead Origin_Lead Add Form', 'Lead Origin_Lead Import',
       'Do Not Email_Yes', 'Country_Not available',
       'Specialization_Hospitality Management', 'Specialization_Not Available',
       'What is your current occupation_Others',
       'What is your current occupation_Unemployed',
       'What is your current occupation_Working Professional'],
      dtype='object')
```

*Note**- const is added while creating the model, is not an attribute*

Creating Predictions on train set

Results with cut-off 0.5:

	Converted	Conversion_Prob	Predicted
0	1	0.725063	1
1	0	0.314725	0
2	1	0.852133	1
3	1	0.787025	1
4	0	0.169432	0

Confusion Matrix:

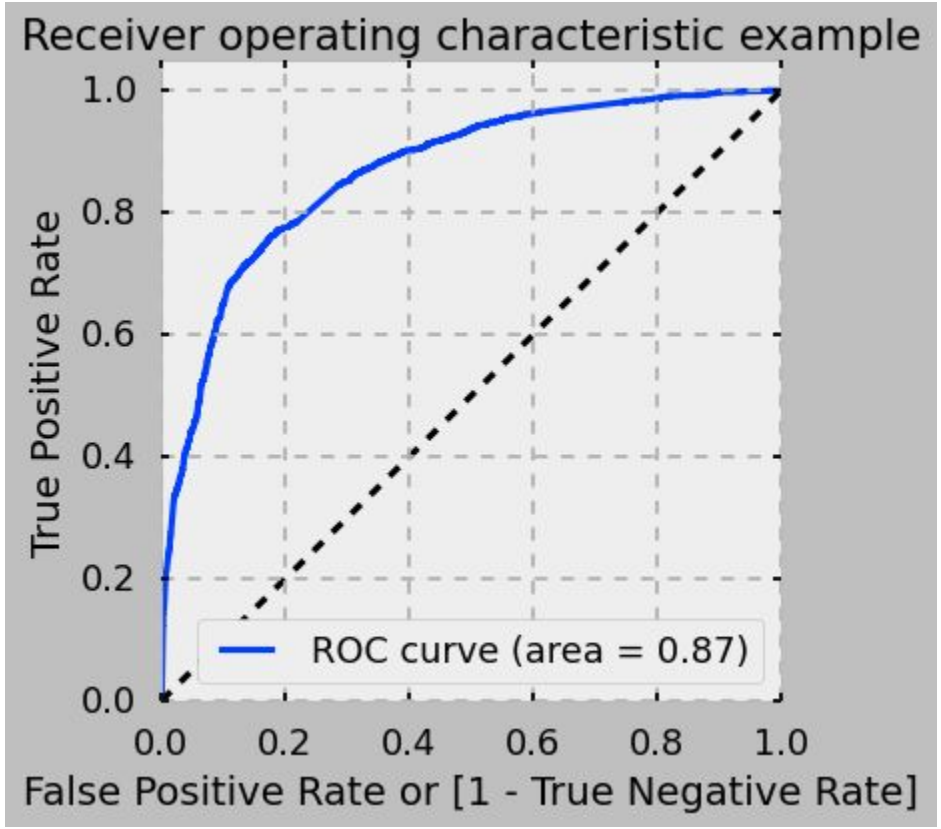
```
array([[3531,  440],
       [ 790, 1702]], dtype=int64)
```

Accuracy: 0.809

Sensitivity: TP/(TP+FN) = 0.682

Specificity: TN/(TN+FP) = 0.889

Plotting ROC curve to choose adequate threshold

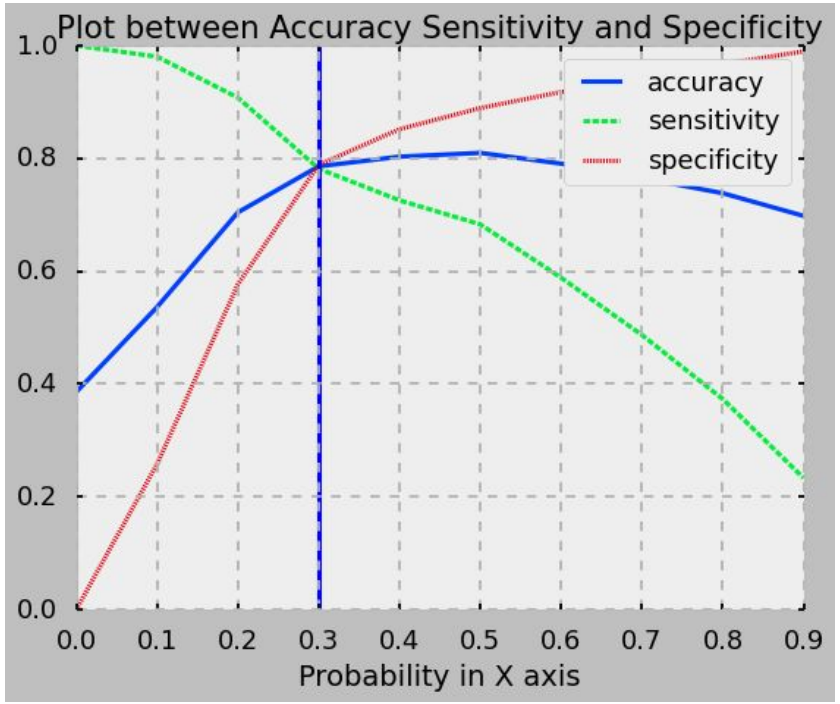


ROC curve has high value of 0.87

Plotting Accuracy Sensitivity and Specificity for different cut-offs

Cut offs : 0.0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9

	prob	accuracy	sensitivity	specificity
0.0	0.0	0.385579	1.000000	0.000000
0.1	0.1	0.536283	0.980337	0.257618
0.2	0.2	0.704162	0.907705	0.576429
0.3	0.3	0.785858	0.781701	0.788466
0.4	0.4	0.803033	0.725522	0.851675
0.5	0.5	0.809686	0.682986	0.889197
0.6	0.6	0.790654	0.588283	0.917653
0.7	0.7	0.765125	0.486758	0.939814
0.8	0.8	0.738202	0.373194	0.967263
0.9	0.9	0.697973	0.233146	0.989675



From the curve above, **0.3 is the optimum point** to take it as a cutoff probability.

Results with cut-off 0.3 as obtained by curve in previous slide

Confusion Matrix:

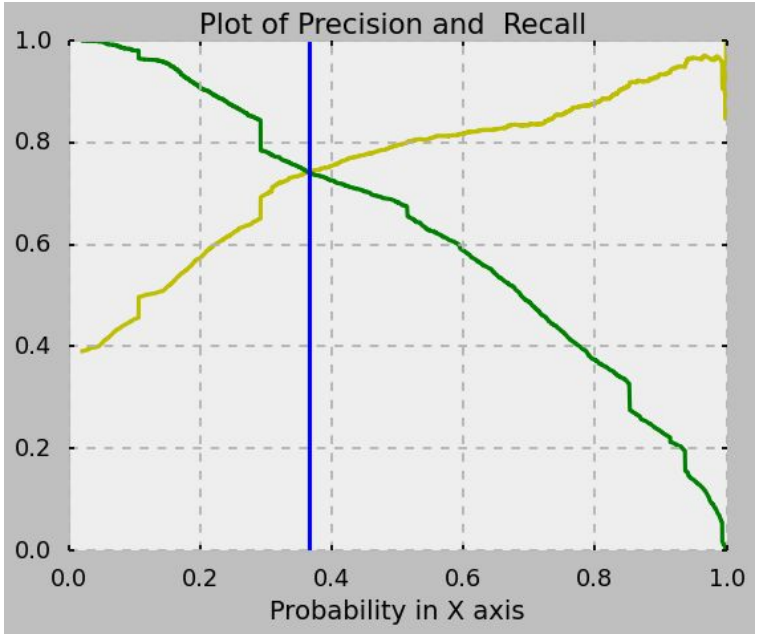
```
array([[3131,  840],
       [ 544, 1948]], dtype=int64)
```

```
Specificity 0.79
sensitivity 0.78
Positive_predictive_value 0.7
Negative_predictive_value 0.85
Precision 0.7
Recall 0.78
```

Precision and Recall

Precision = TP / TP + FP : 0.698
Recall = TP / TP + FN : 0.781

Cutoff for optimum precision and recall is **0.365



Results with cut-off 0.365 as obtained by curve in Precision and Recall

Confusion Matrix:

```
array([[3326,  645],
       [ 647, 1845]], dtype=int64)
```

Accuracy : 0.8

```
Specificity 0.84
sensitivity 0.74
Positive_predictive_value 0.74
Negative_predictive_value 0.84
Precision 0.74
Recall 0.74
```

Prediction on Test data using cut off 0.365

Confusion Matrix:

```
array([[1434,  269],
       [ 266,  801]], dtype=int64)
```

Accuracy : 0.81

```
Specificity 0.84
sensitivity 0.75
Positive_predictive_value 0.75
Negative_predictive_value 0.84
Precision 0.75
Recall 0.75
```

Results and Conclusions

Final Results

Performance metrics of Train Data

1. Accuracy 0.80
2. Specificity 0.84
3. sensitivity 0.74
4. Positive_predictive_value 0.74
5. Negative_predictive_value 0.84
6. Precision 0.74
7. Recall 0.74

Performance metrics of Test Data

1. Accuracy 0.81
2. Specificity 0.84
3. sensitivity 0.75
4. Positive_predictive_value 0.75
5. Negative_predictive_value 0.84
6. Precision 0.75
7. Recall 0.75

Conclusions

We can suggest the below points to X Education to improve its conversion rate

1. Customer Spending more time have higher chance to enroll into the course

2. Working professional has higher chance to enroll into the course and they needs to be targeted

3. Customer with Lead origin as Lead Add form are more likely to enroll the course

4. Customer who are not interested in receiving mails are less likely to enroll

5. Customer with No Specialization or Specialization ad Hospitality and management are less likely to join

6. Customer with Lead origin ad Landing page submission or Lead import are less likely to enroll

7. Customer with occupation as Unemployed or Others would be interested in joining the course