LEAD SCORING CASE STUDY

SUBMITTED BY –

SHILADITYA CHAKRABORTY

PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

OBJECTIVE

X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

The foundational data offered valuable insights into the patterns of potential customer visits, their duration on the site, the channels through which they accessed the platform, and the resulting conversion rate.

METHODOLOGY

1. Data Cleaning –
   - Handling of duplicate values
   - Handling of missing values
   - Dropping features with high percentage of missing values
   - Imputation of missing values, if applicable

2. Exploratory Data Analysis –
   - Performing categorical feature EDA and finding imbalance in data
   - Performing numerical feature EDA and performing outlier analysis

3. Model Building –
   - Feature scaling
   - Dummy variable creation
   - Train-Test Splitting of data
   - Checking correlations between features
   - Model building using RFE and subsequent manual elimination

METHODOLOGY (CONTD.)

4. Prediction on the Train Set –
- Prediction with arbitrary cut-off as 0.5
- ROC Curve
- Optimal Cut-off point
- Key Metrics

5. Prediction on the Test Set –
- Key Metrics

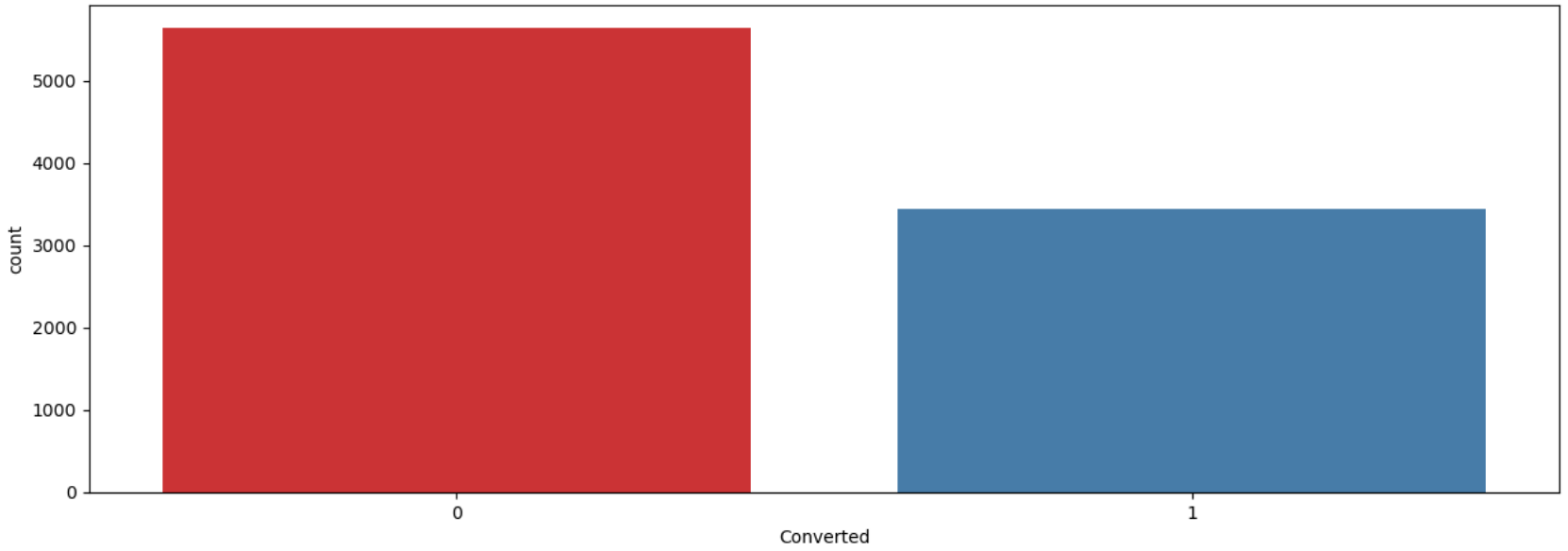
Data Cleaning

1. Initial data inspection – 9240 rows and 37 features
2. Missing Value treatment –
   - 10 columns are dropped with more than 30% missing values
   - 3 columns with more than 20% missing values –
     - The column “Country” is dropped since the data is highly imbalanced
     - For the other 2 columns missing values are imputed
   - Other records with missing values are dropped
3. 11 columns are dropped as the data is highly imbalanced.
4. 2 columns are dropped as the data is unique for each record
5. Final shape of the dataset – 9074 records and 11 features

Exploratory Data Analysis

Distribution of the Target Variable

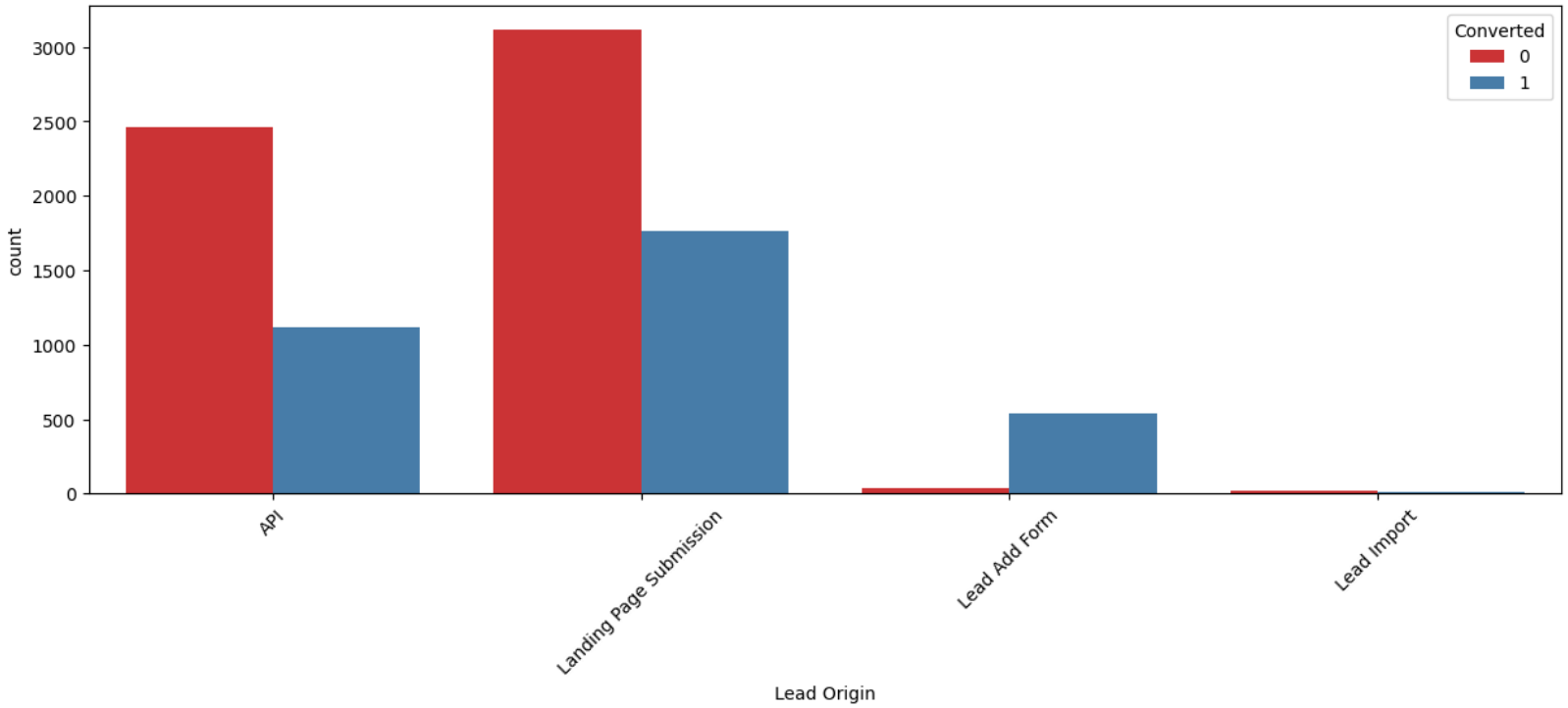
- Conversion rate – 37.85%
- Distribution



Exploratory Data Analysis (contd.)

Categorical Variable Analysis

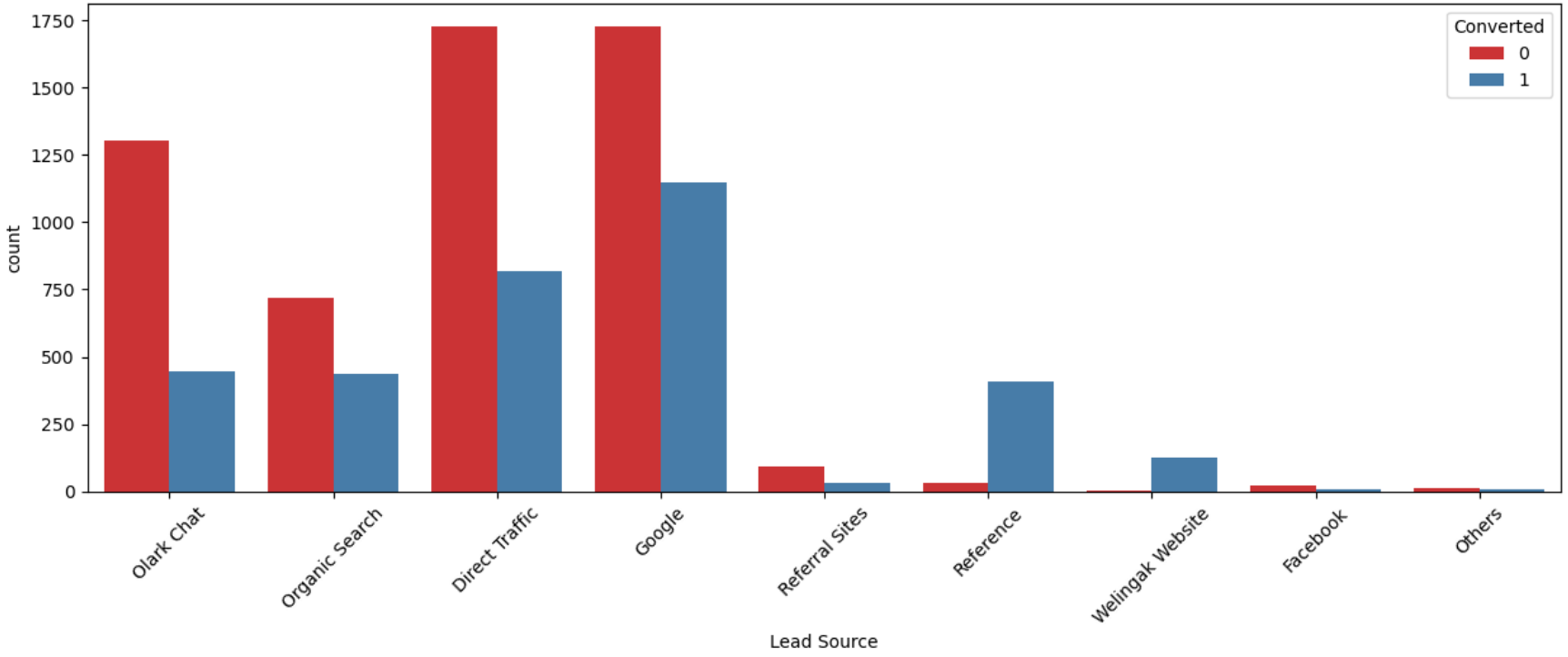
Analysis of 'Lead Origin'



Exploratory Data Analysis (contd.)

Categorical Variable Analysis

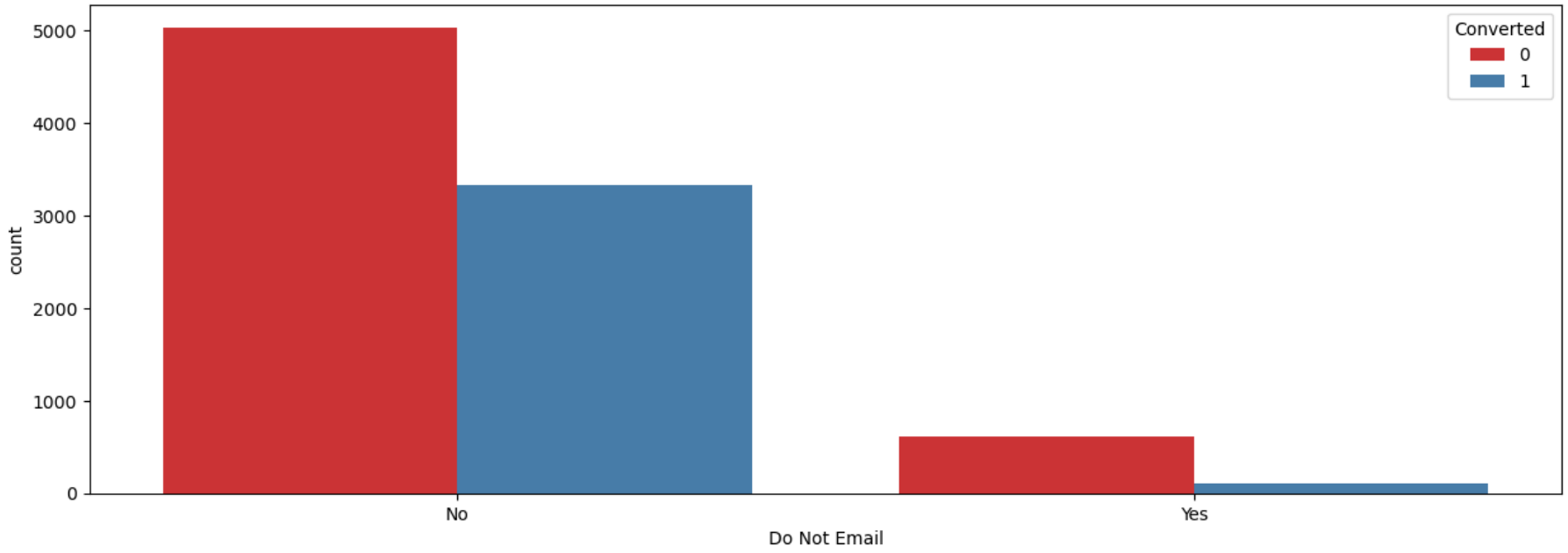
Analysis of ‘Lead Source'



Exploratory Data Analysis (contd.)

Categorical Variable Analysis

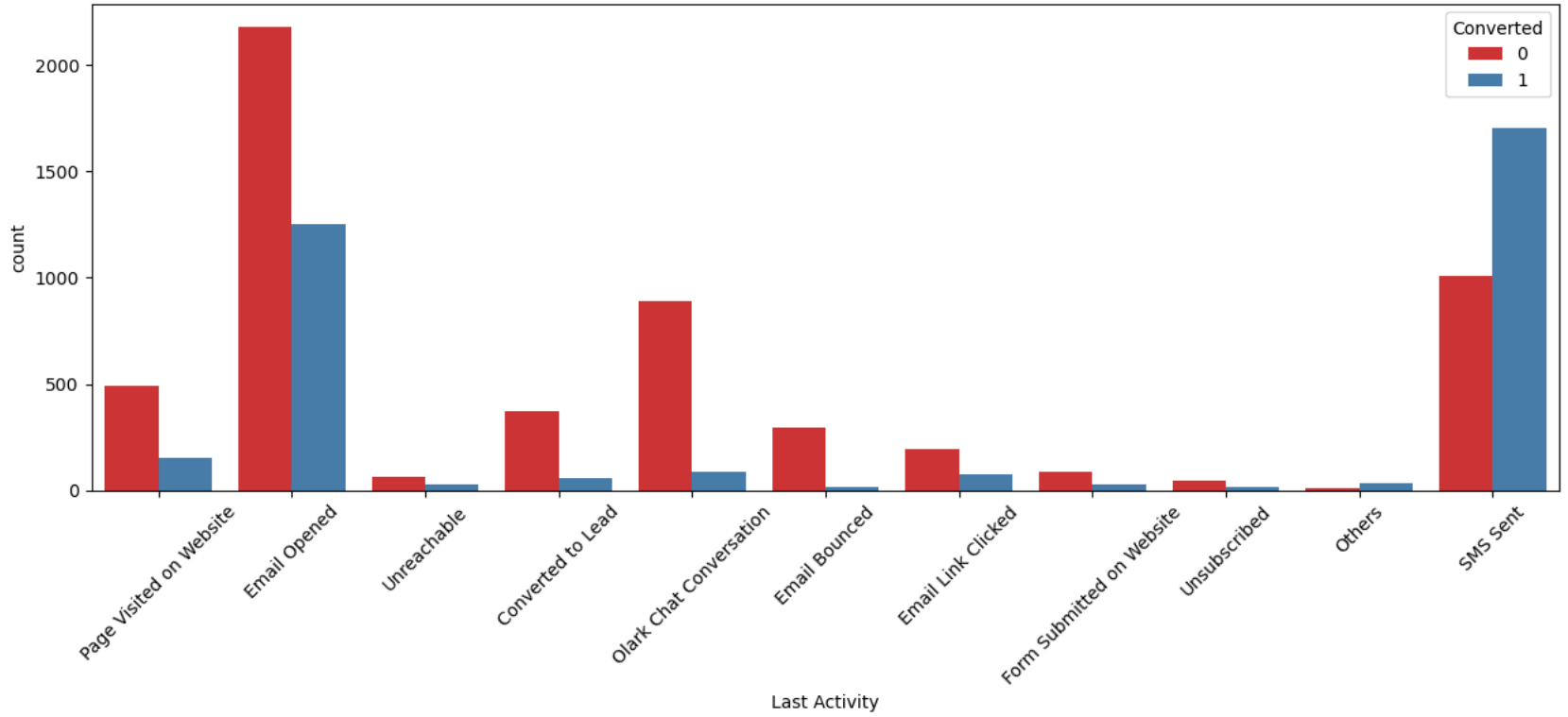
Analysis of ‘Do Not Email'



Exploratory Data Analysis (contd.)

Categorical Variable Analysis

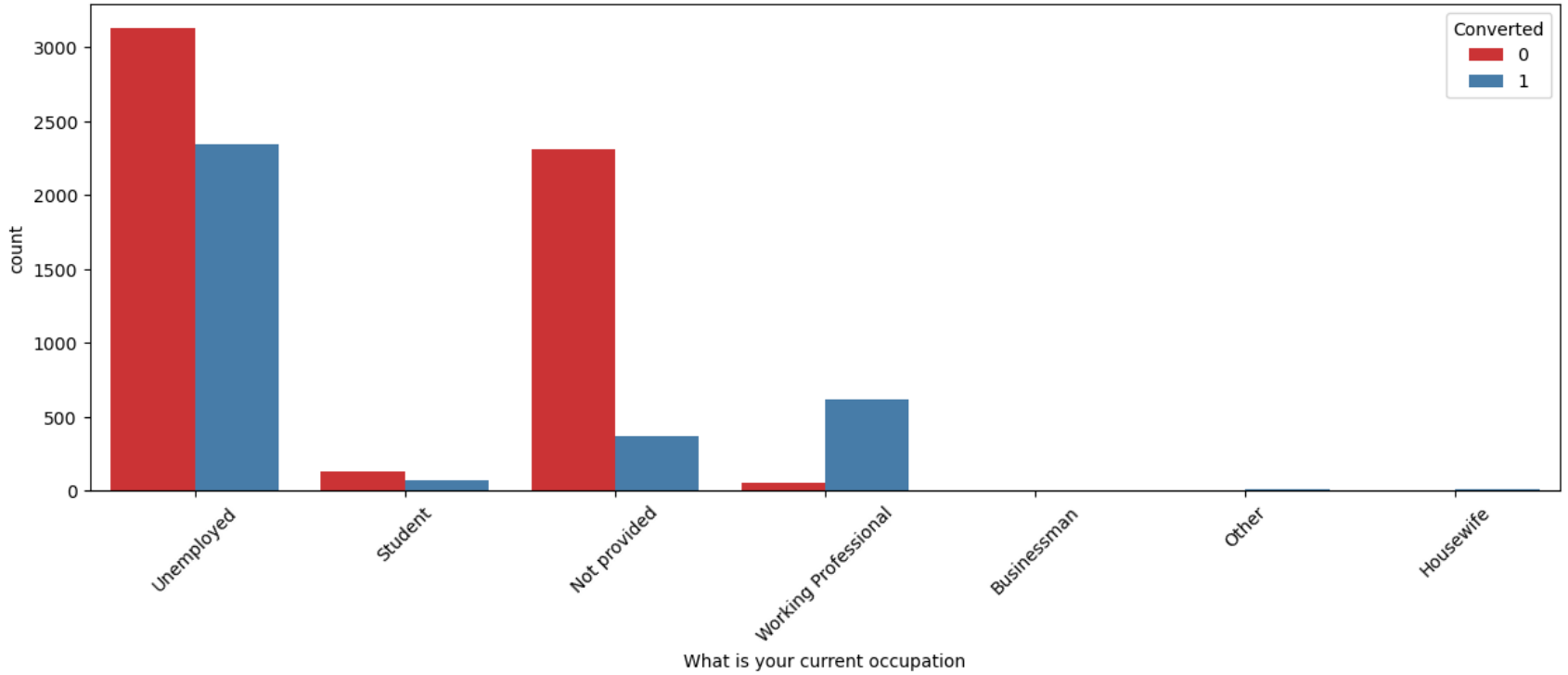
Analysis of ‘Last Activity'



Exploratory Data Analysis (contd.)

Categorical Variable Analysis

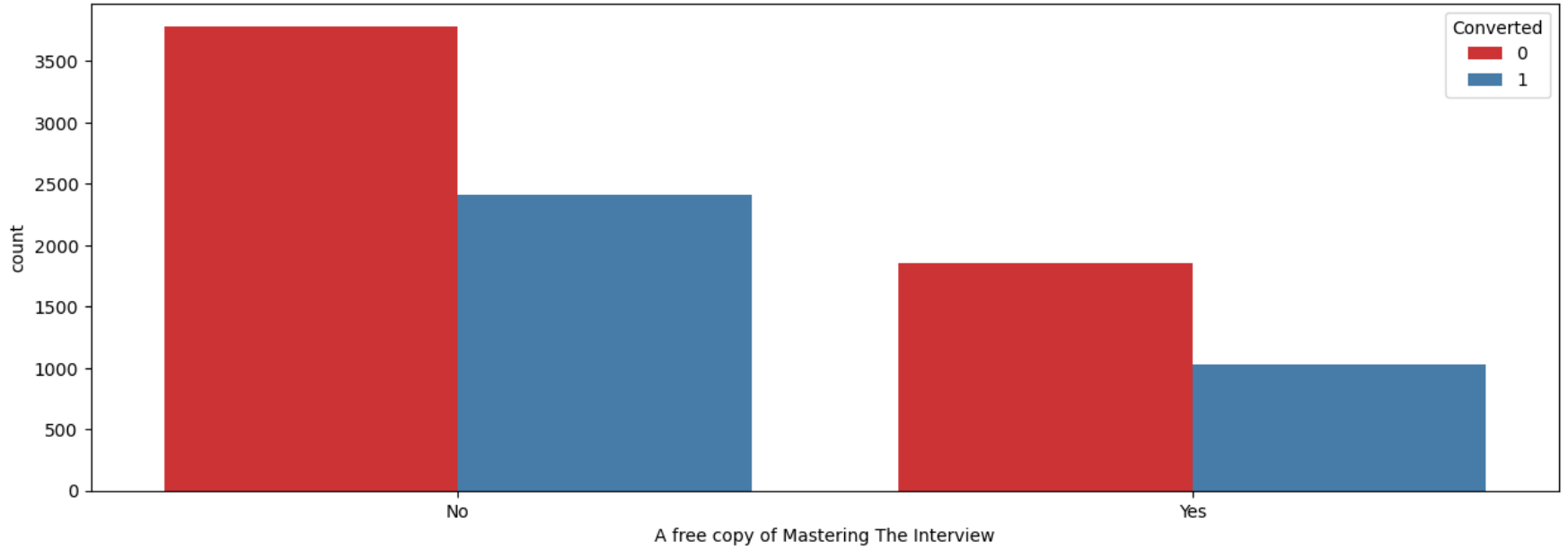
Analysis of ‘What is your current occupation'



Exploratory Data Analysis (contd.)

Categorical Variable Analysis

Analysis of ‘A free copy of Mastering The Interview'



Exploratory Data Analysis (contd.)

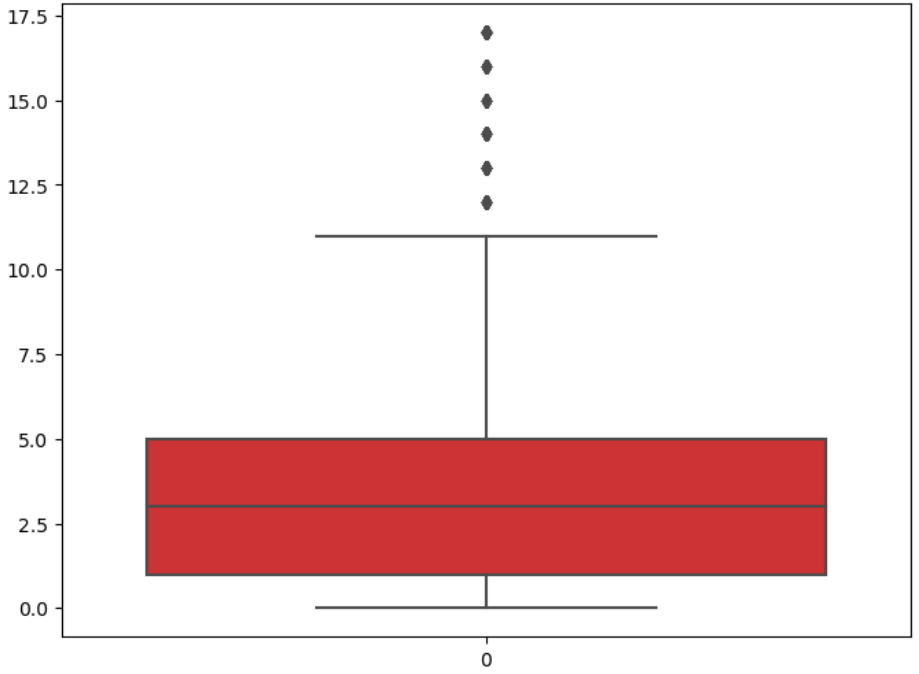
Observations of Categorical Variable Analysis

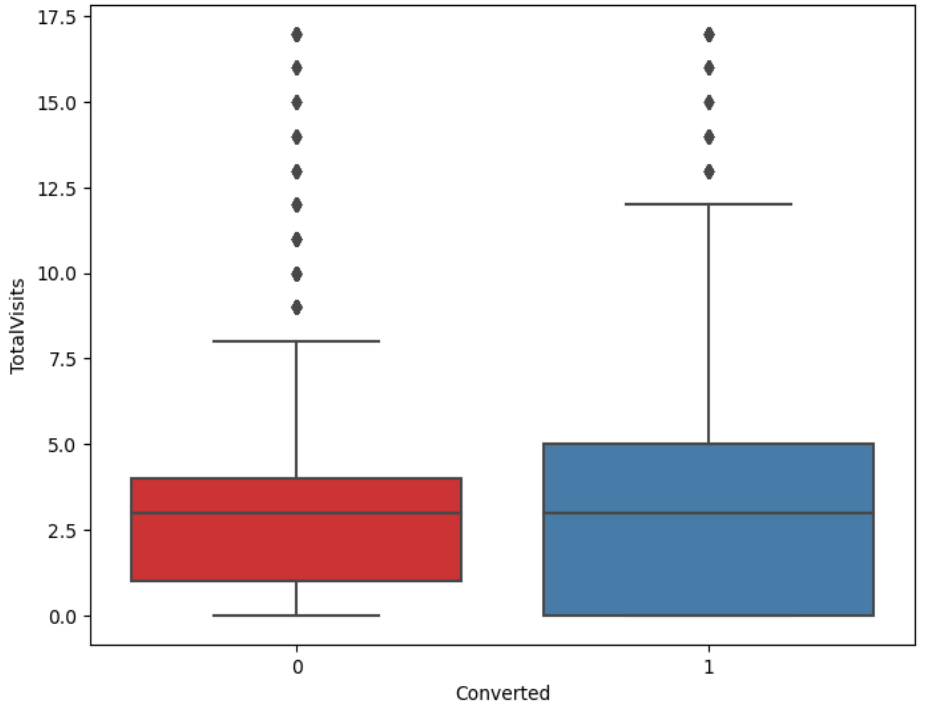
1. ‘Lead Origin’
   - For API and Landing Page Submission, the successful lead conversions are less than the one's that are not converted.
   - For Lead Add Form, the successful conversions are more than the unsuccessful ones.
2. ‘Lead Source’
   - For Reference and Welingak Institute the rate of successful lead conversion is high.
   - For the other categories the rate of successful leas conversion is low.
   - The maximum distribution is of the category "Google" followed by Direct Traffic and Olark Chat
3. ‘Last Activity’
   - There is more lead conversion when SMS is sent to user.
4. 'What is your current occupation’
   - Lead conversion is low for the case when Occupation is not provided by user
   - Most users in the dataset is Unemployed

Exploratory Data Analysis (contd.)

Numerical Variable Analysis

Analysis of ‘TotalVisits’

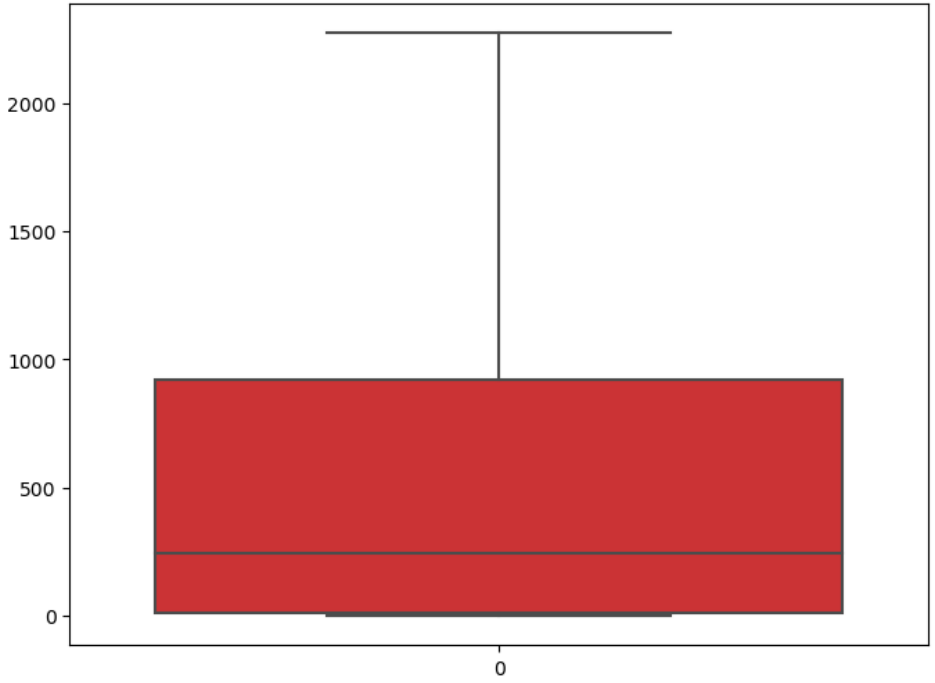


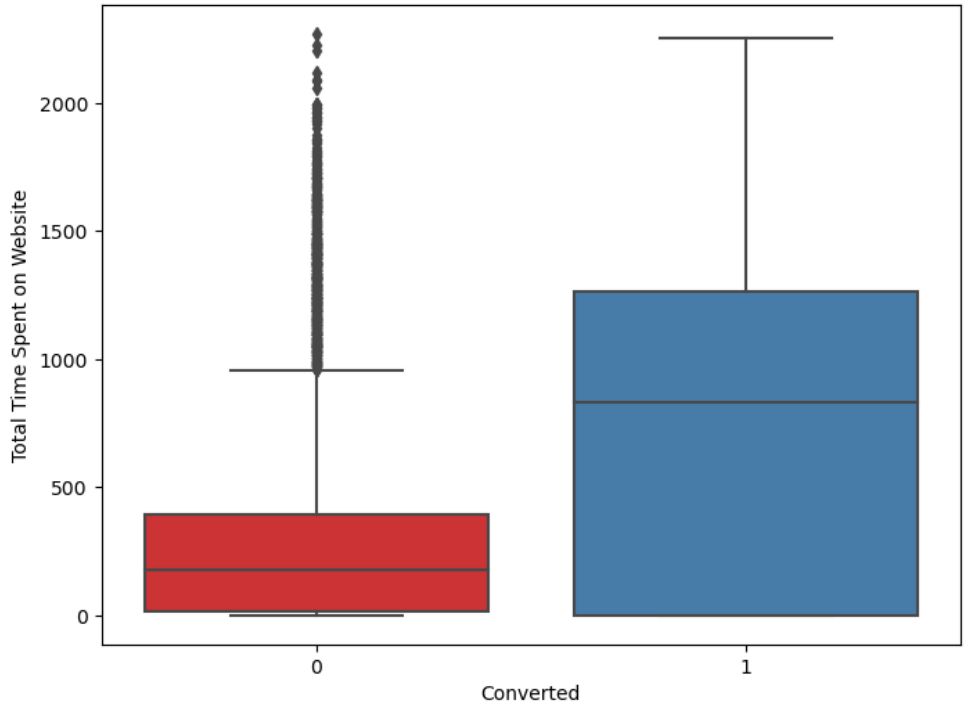


Exploratory Data Analysis (contd.)

Numerical Variable Analysis

Analysis of ‘Total Time Spent on Website’

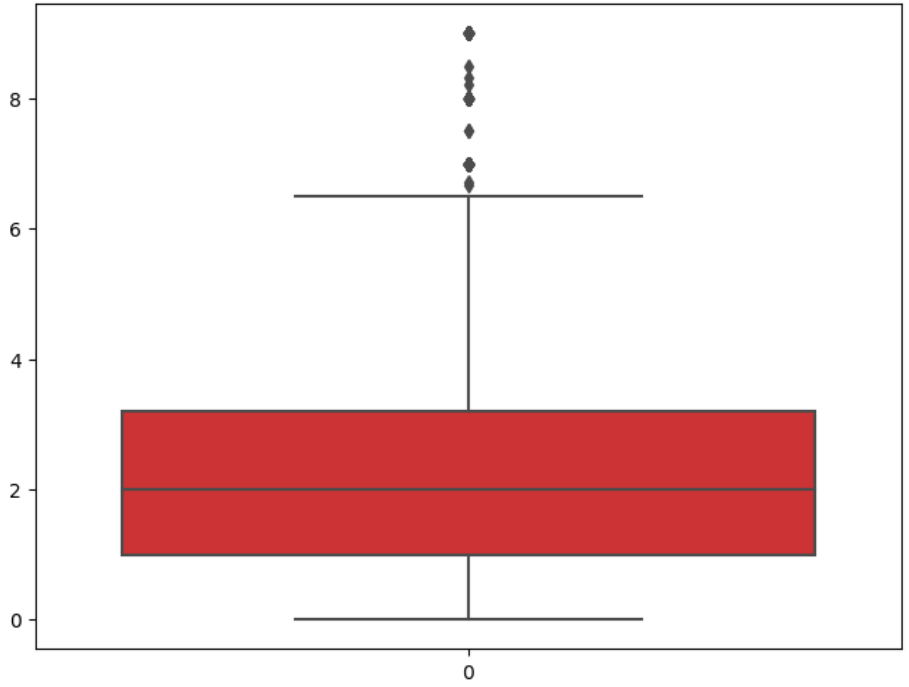


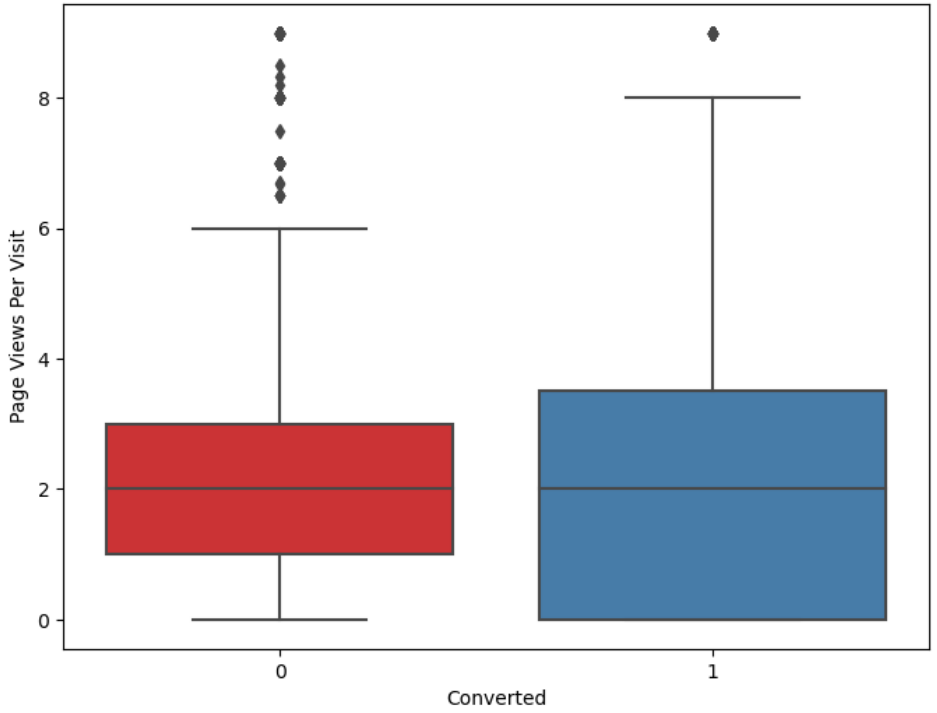


Exploratory Data Analysis (contd.)

Numerical Variable Analysis

Analysis of ‘Page Views Per Visit’



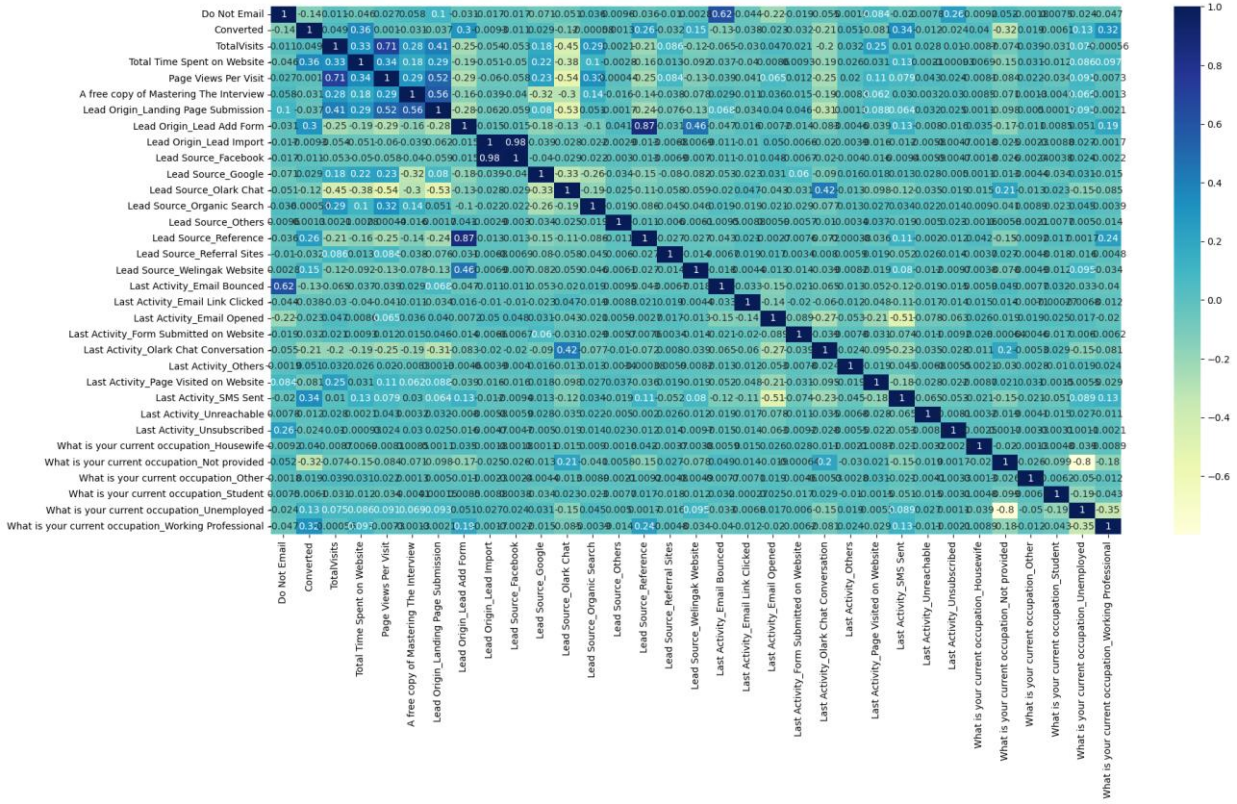


Model Building

1. Feature Scaling – The features of the model are brought to the same scale by Standard scaler before model building
2. Dummy variable creation – Dummy variables are created for the categorical features
3. Train-Test Splitting of data – The data set is split into Training and Test sets with a 7:3 ratio

Model Building

4. Checking correlations between features



Highly correlated attributes create dependency on various independent factors which will give us inappropriate results.

Model Building

5. Model building using RFE and subsequent manual elimination

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.8921	0.089	-21.153	0.000	-2.067	-1.717
Do Not Email	-1.5469	0.193	-8.003	0.000	-1.926	-1.168
Total Time Spent on Website	1.1444	0.041	27.775	0.000	1.064	1.225
Lead Origin_Lead Add Form	3.5908	0.220	16.298	0.000	3.159	4.023
Lead Origin_Lead Import	1.1301	0.461	2.450	0.014	0.226	2.034
Lead Source_Olark Chat	1.3290	0.104	12.776	0.000	1.125	1.533
Lead Source_Welingak Website	2.0245	0.757	2.674	0.008	0.540	3.509
Last Activity_Email Opened	0.9326	0.097	9.643	0.000	0.743	1.122
Last Activity_Others	2.6655	0.474	5.620	0.000	1.736	3.595
Last Activity_SMS Sent	2.0610	0.100	20.635	0.000	1.865	2.257
Last Activity_Unreachable	1.2081	0.310	3.903	0.000	0.601	1.815
Last Activity_Unsubscribed	2.0198	0.475	4.255	0.000	1.089	2.950
What is your current occupation_Not provided	-1.2957	0.088	-14.720	0.000	-1.468	-1.123
What is your current occupation_Working Professional	2.4743	0.188	13.174	0.000	2.106	2.842

	Features	VIF
2	Lead Origin_Lead Add Form	1.61
4	Lead Source_Olark Chat	1.41
11	What is your current occupation_Not provided	1.38
5	Lead Source_Welingak Website	1.33
1	Total Time Spent on Website	1.31
8	Last Activity_SMS Sent	1.29
6	Last Activity_Email Opened	1.26
12	What is your current occupation_Working Profes...	1.19
0	Do Not Email	1.14
10	Last Activity_Unsubscribed	1.08
3	Lead Origin_Lead Import	1.02
7	Last Activity_Others	1.00
9	Last Activity_Unreachable	1.00

Prediction on the Train Set

1. Prediction with arbitrary cut-off as 0.5

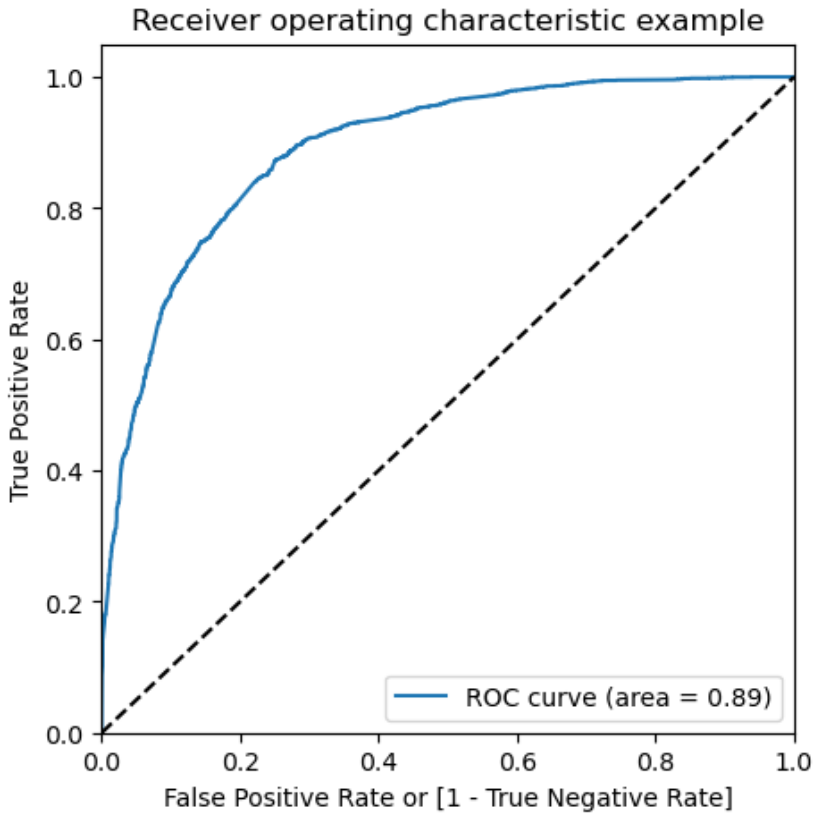
```
[[3464  441]
 [ 745 1701]]
```

Confusion Matrix

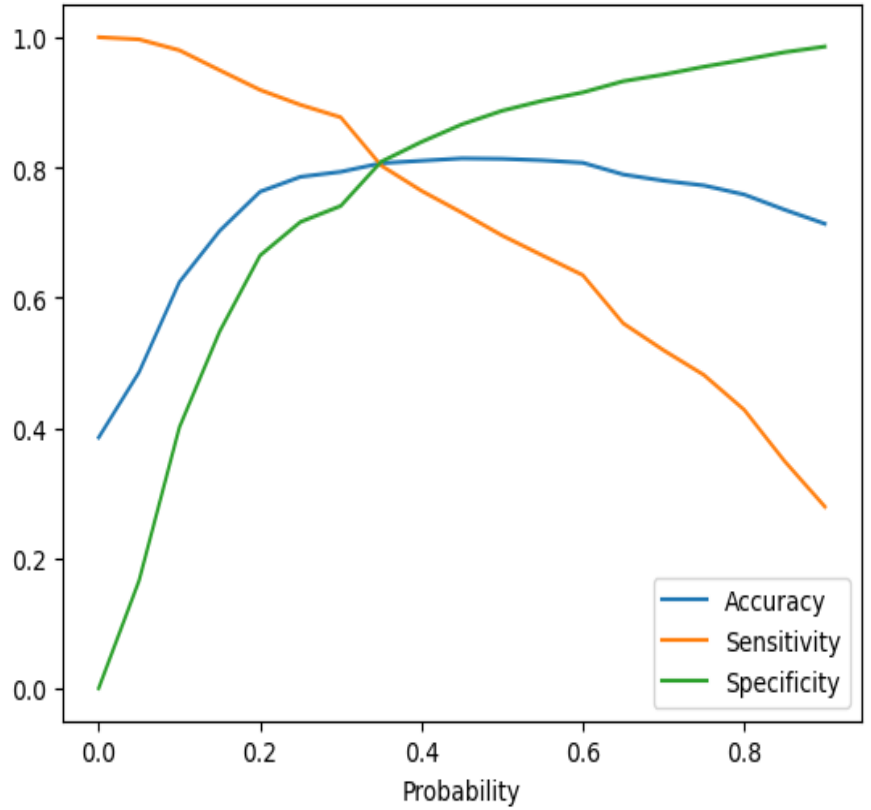
```
Accuracy 0.8132577546843017
Sensitivity 0.6954210956663941
Specificity 0.887067861715749
False Positive Rate 0.11293213828425096
Positive Predictive Value 0.7941176470588235
Negative Predictive Value 0.8229983368971252
```

Key Metrics of Logistic Regression

Prediction on the Train Set



ROC Curve



Cut-off Point

Prediction on the Train Set

```
[[3464  441]
 [ 745 1701]]
```

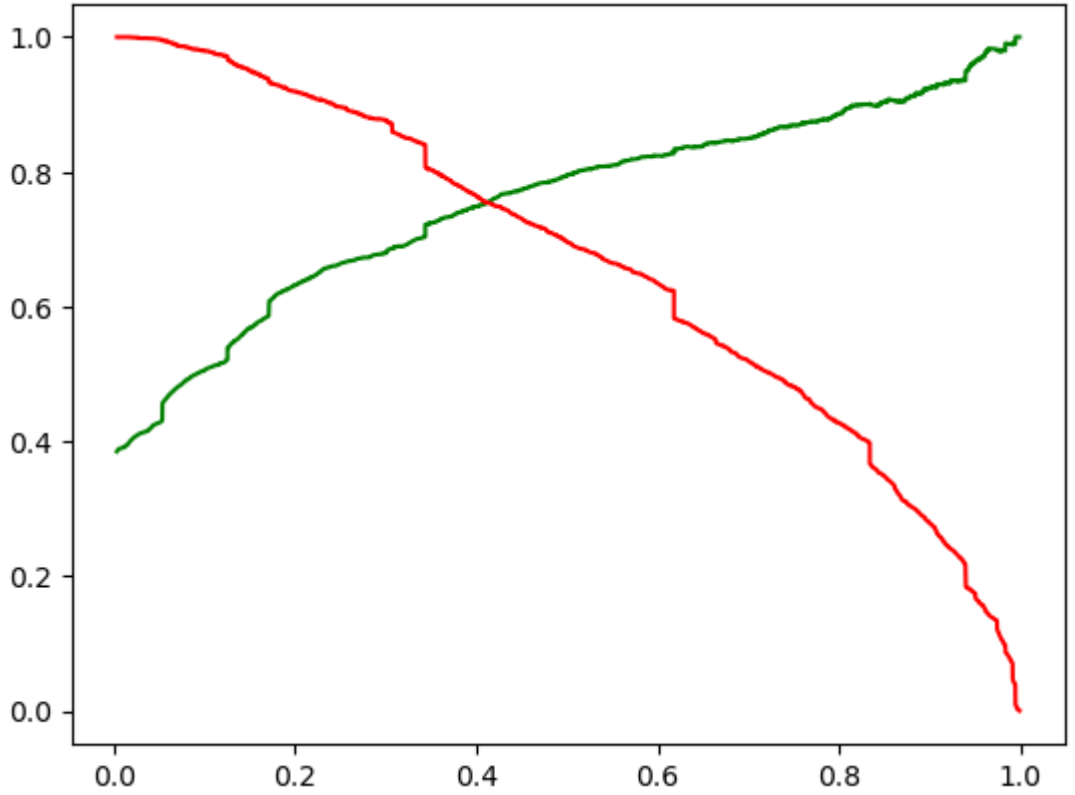
Confusion Matrix

```
Accuracy 0.8132577546843017
Sensitivity 0.6954210956663941
Specificity 0.887067861715749
False Positive Rate 0.11293213828425096
Postitive Predictive Value 0.7941176470588235
Negetive Predictive Value 0.8229983368971252
```

Key Metrics of Logistic Regression

```
Precision 0.7244559203246035
Recall 0.802943581357318
```

Precision and Recall



Precision and Recall Trade-off

Prediction on the Test Set

```
[[1407  327]
 [ 218  771]]
```

Confusion Matrix

```
Accuracy 0.8132577546843017
Sensitivity 0.6954210956663941
Specificity 0.887067861715749
False Positive Rate 0.11293213828425096
Positive Predictive Value 0.7941176470588235
Negative Predictive Value 0.8229983368971252
```

Key Metrics of Logistic Regression

```
Precision 0.7021857923497268
Recall 0.7795753286147624
```

Precision and Recall

CONCLUSION

1. The Test set has optimal key metric values of logistic regression.
2. The most important variables (top 5) are –
   - Lead Origin_Lead Add Form
   - Last Activity_Others
   - What is your current occupation_Working Professional
   - Last Activity_SMS Sent
   - Lead Source_Welingak Website
3. The accuracy and stability of the model is adaptive.