LEAD SCORING CASE STUDY

Using Logistic Regression

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- 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.
- 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.

BUSINESS OBJECTIVE

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.
- Lead X wants us to build a model to give every lead a lead score between 0 -100 . So that they can identify the Hot leads and increase their conversion rate as well.
- The CEO want to achieve a lead conversion rate of 80%.
- They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches.

METHODOLOGY

- Data cleaning and data manipulation
- EDA (Univariate and Bivariate data analysis)
- Feature Scaling, Dummy Variables and encoding of the data
- Classification technique: Logistic Regression
- Validation of the model
- Model presentation
- Conclusions and recommendations

DATA CLEANING

- Check and handle duplicate data.
- Check and handle NA values and missing values.
- Drop columns, if it contains large amount of missing values and not useful for the analysis.
- Imputation of the values, if necessary.
- Check and handle outliers in data.

DATA MANIPULATION

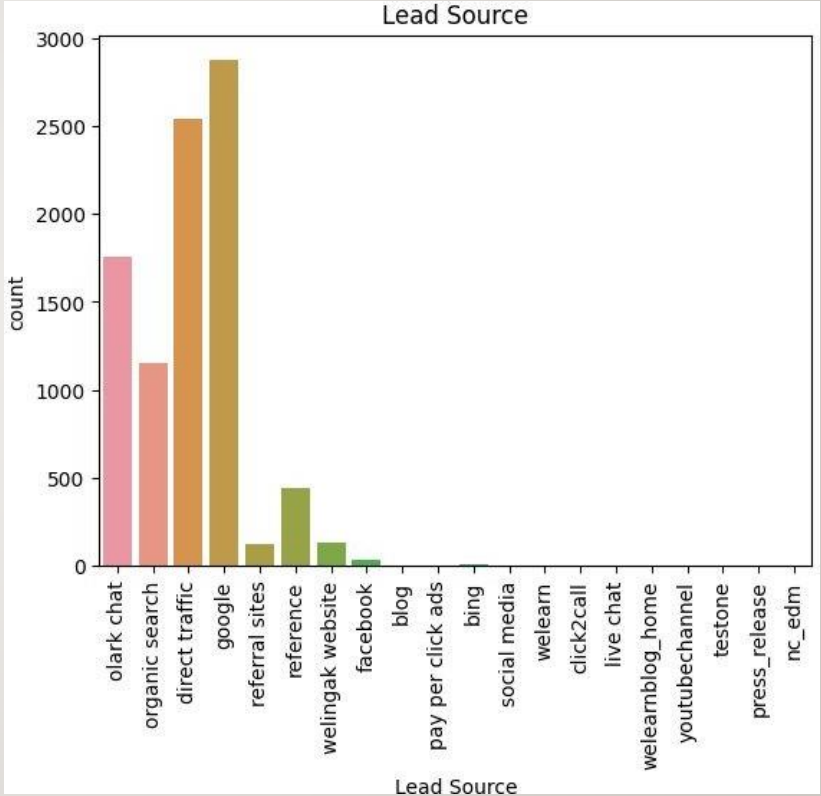
- Total Number of Columns =37, Total Number of Rows =9240
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are:
    - “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement
- Dropping the columns having more than 30% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

Exploratory Data Analysis

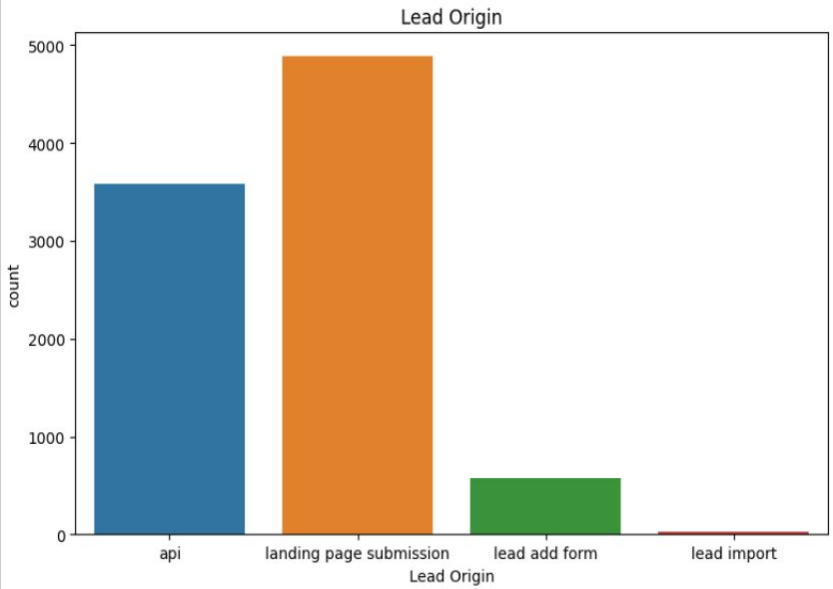
- Univariate data analysis
  - value count
  - distribution of variable

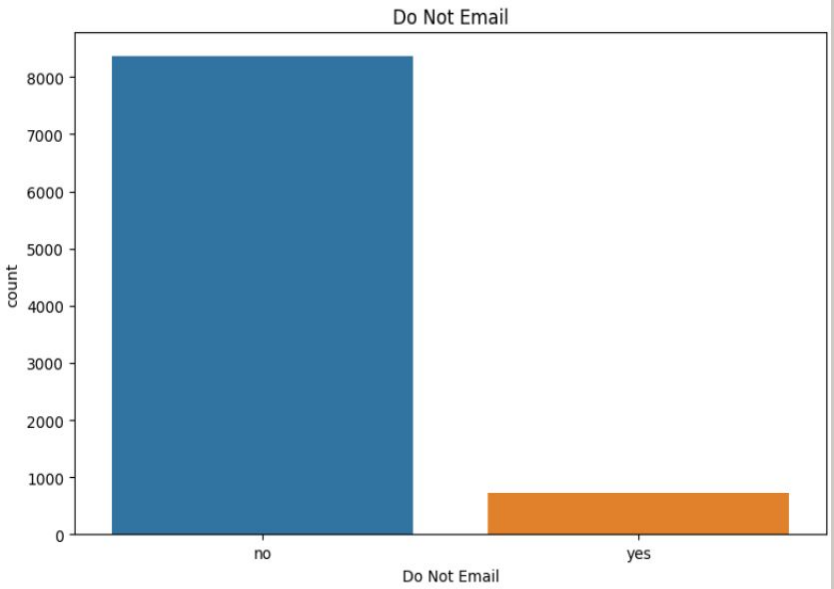
- Bivariate data analysis:
  - correlation coefficients
  - pattern between the variables

- Very high conversion rates for lead sources ‘Reference’ and ‘Welingak Website’.
- Most leads are generated through ‘Direct Traffic’ and ‘Google’.

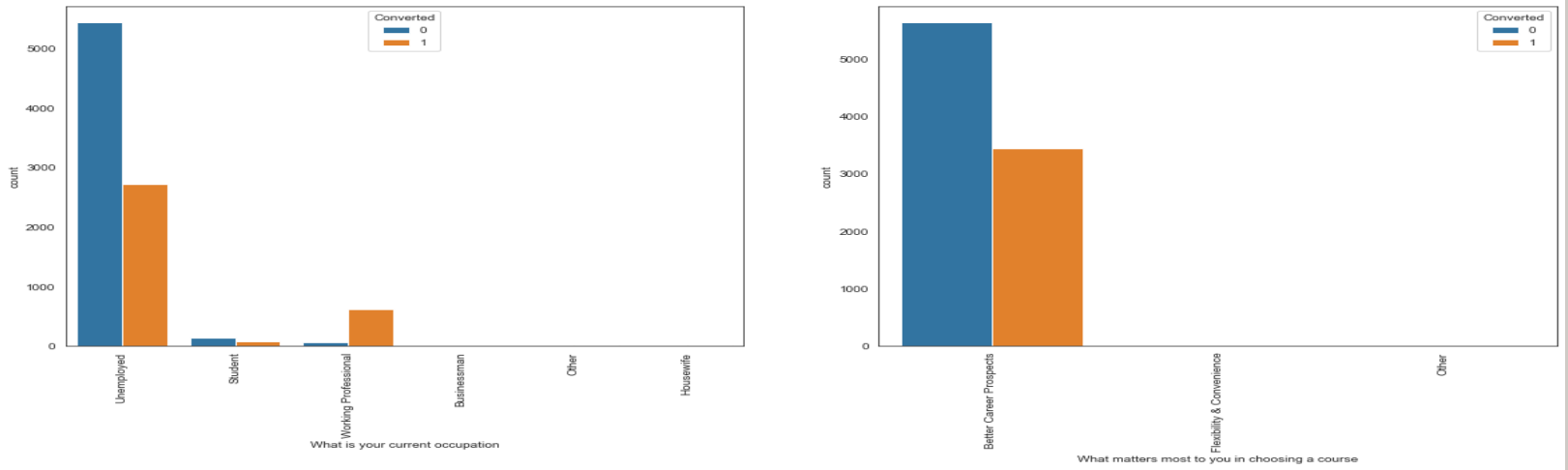


Univariate Analysis:

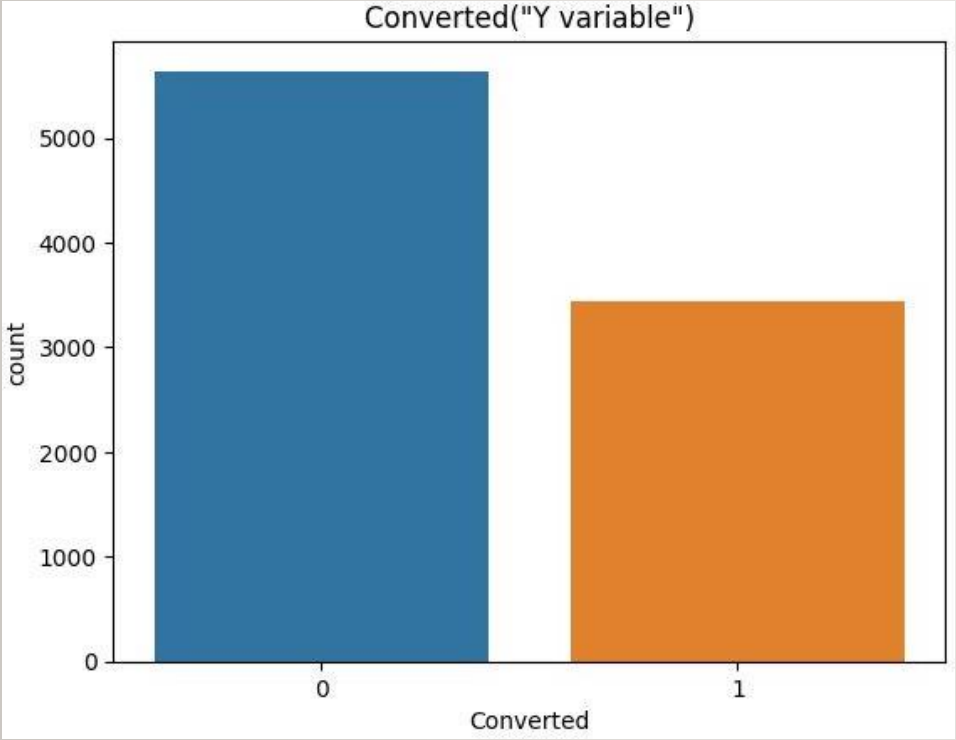


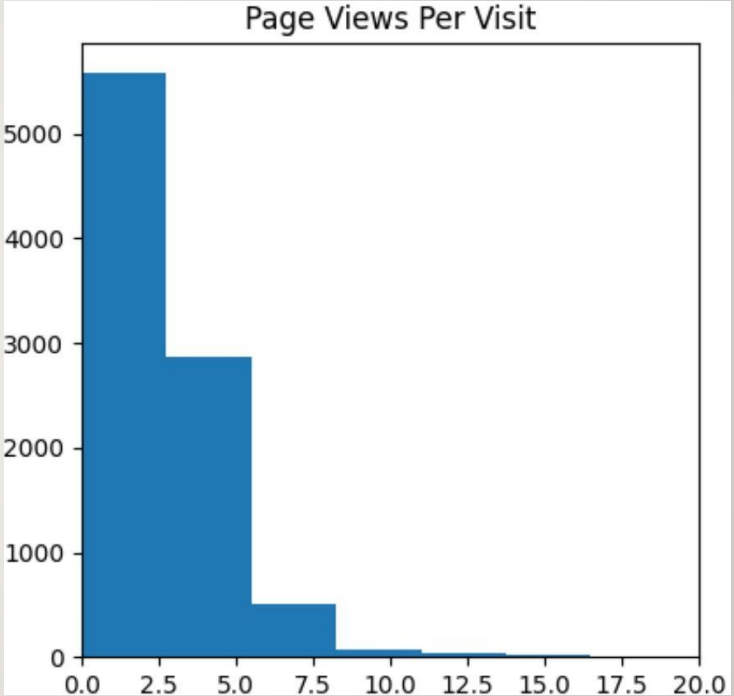


Current occupation:



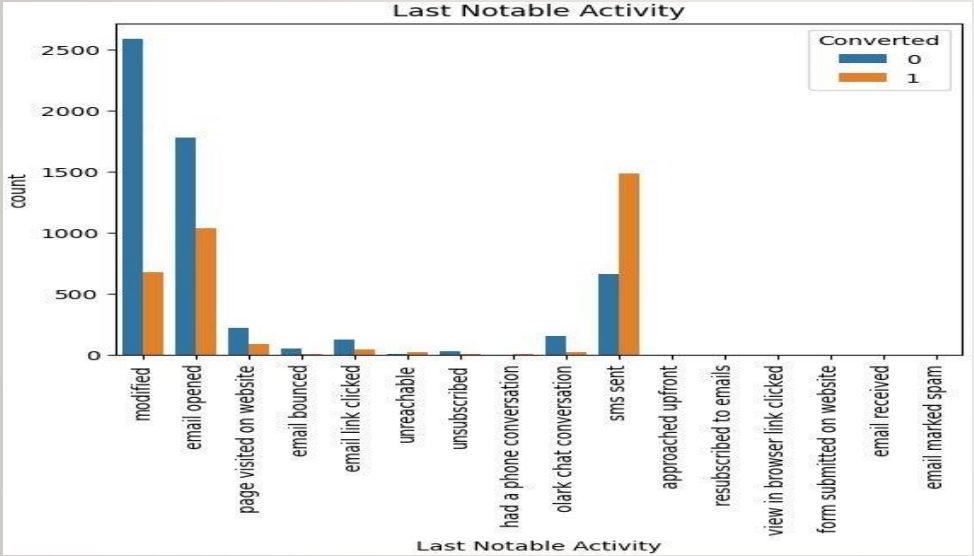
- Working professionals are most likely to get converted


Converted("Y variable")
count
5000
4000
3000
2000
1000
0
0
1
Converted

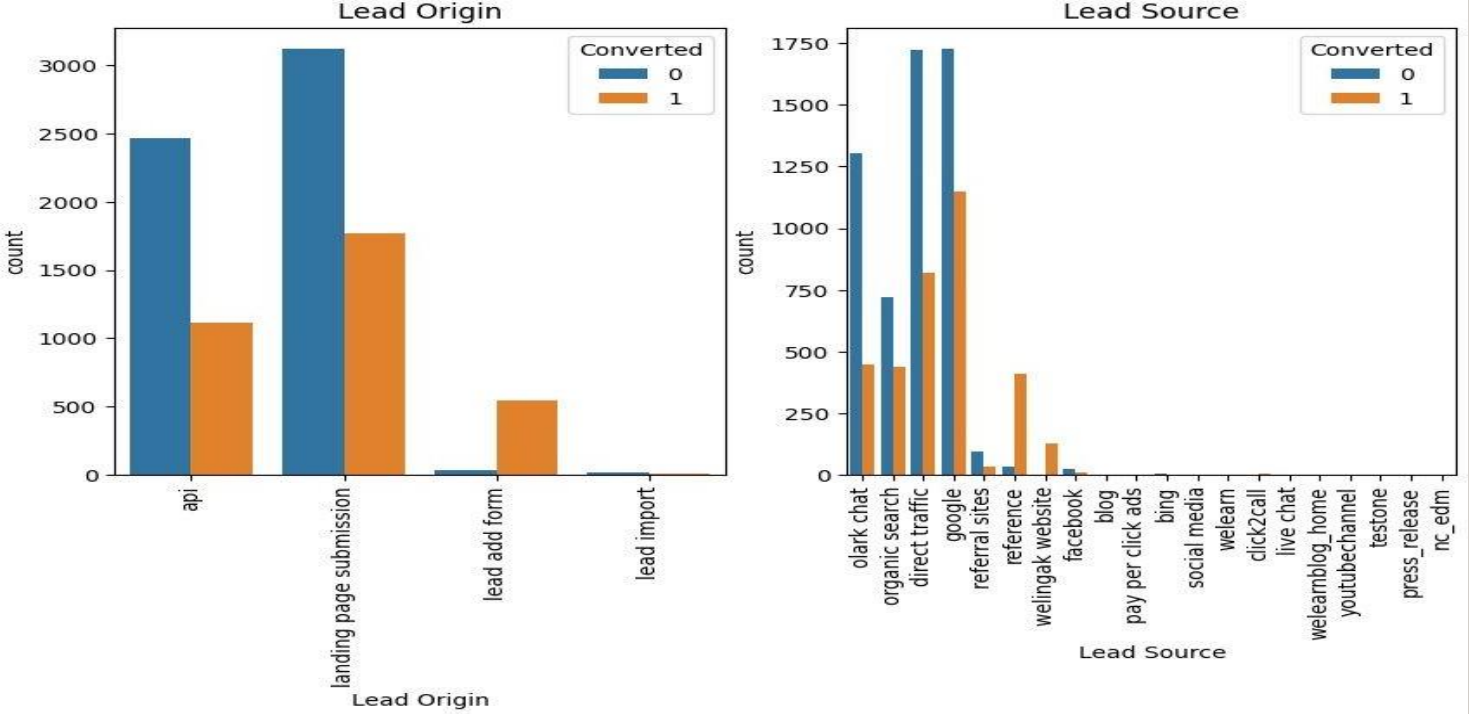

Page Views Per Visit
5000
4000
3000
2000
1000
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

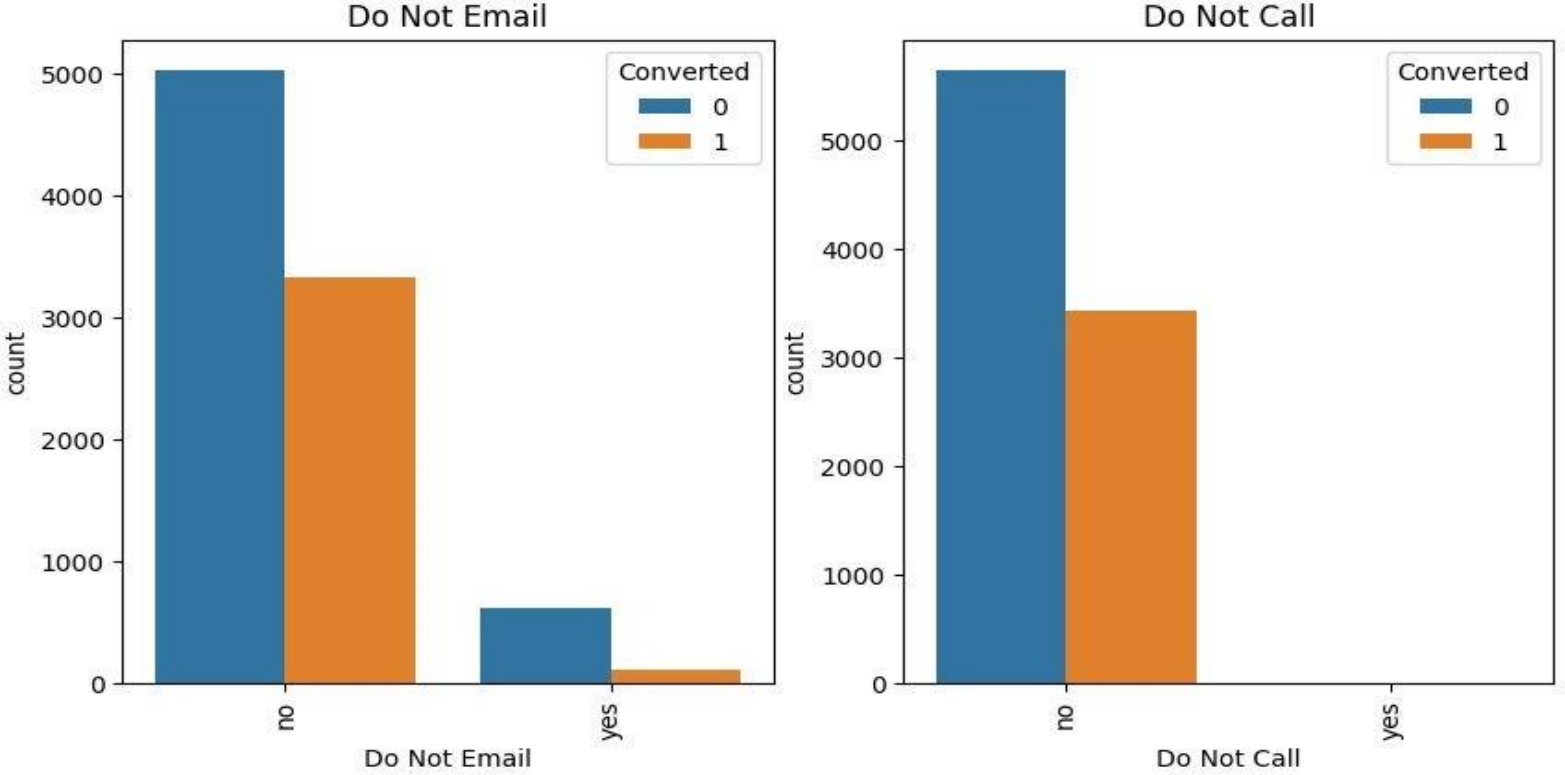
Bi Variate analysis:

Last Notable Activity:

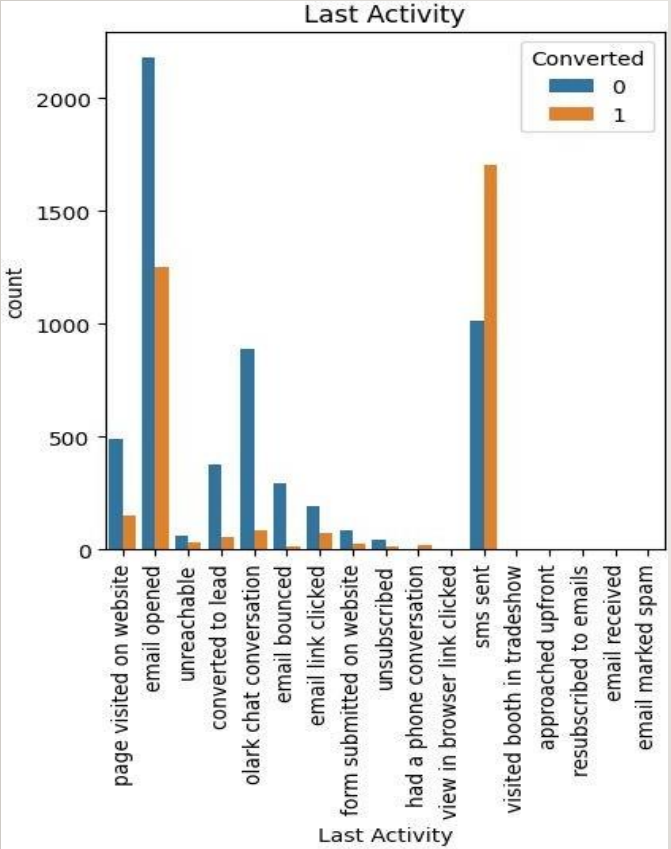


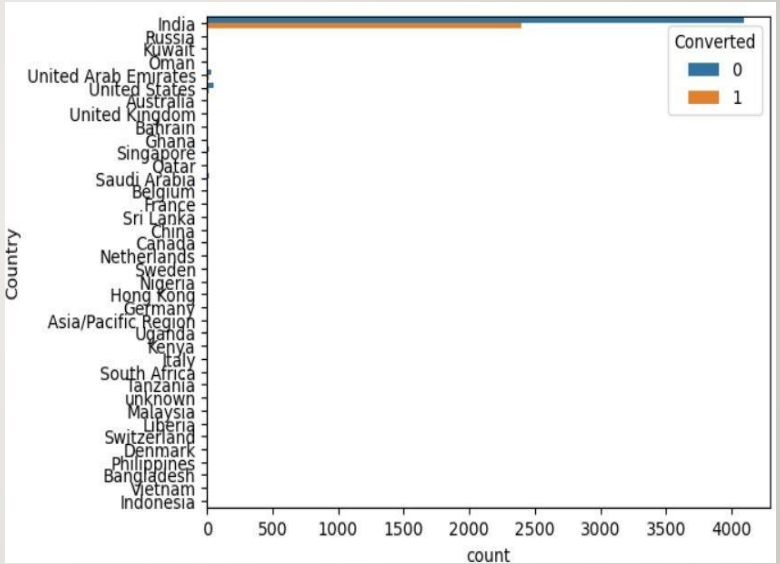
- Highest conversion rate is for the last notable activity '**SMS Sent**'.


Lead Origin
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
api
landing page submission
lead add form
lead import
Lead Origin
Lead Source
Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
olark chat
organic search
direct traffic
google
referral sites
reference
welingak website
facebook
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source

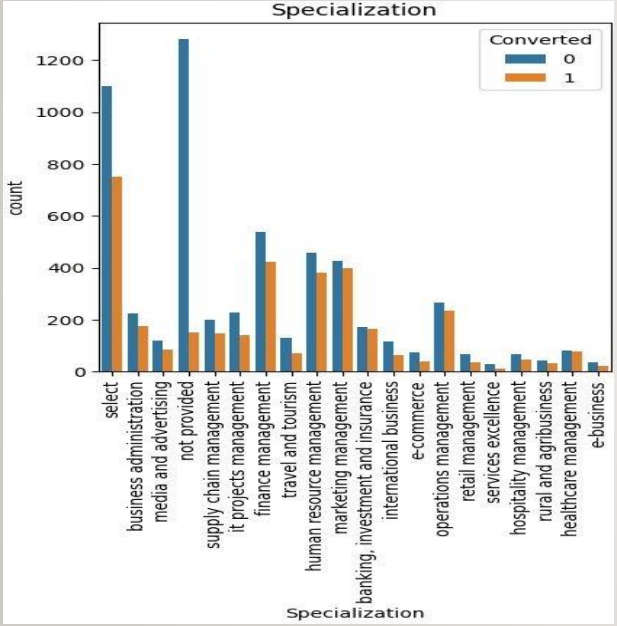

Do Not Email
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Do Not Email
Do Not Call
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Do Not Call

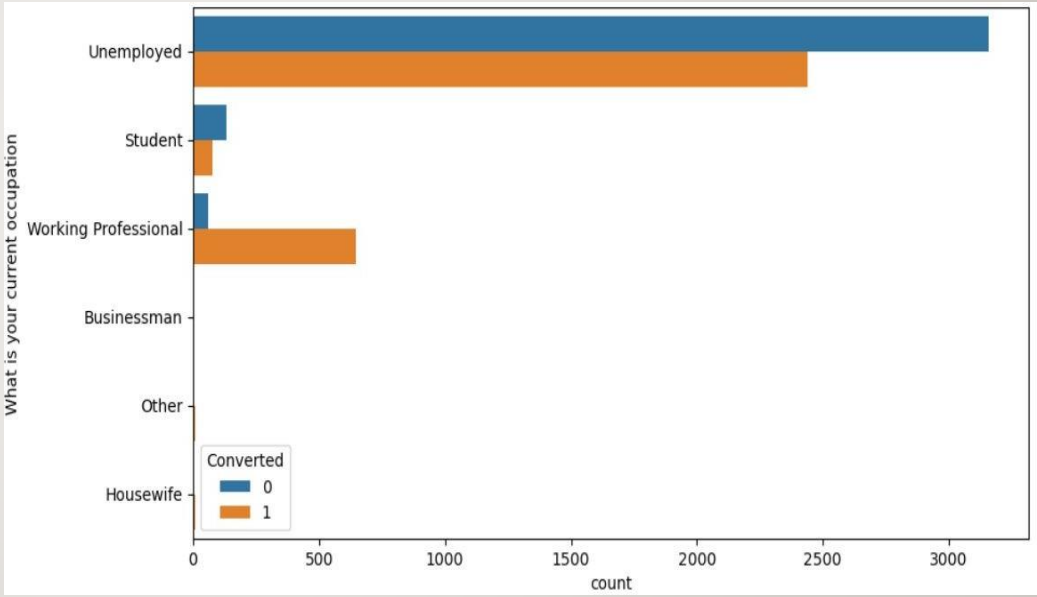
CATEGORICAL VARIABLE RELATION

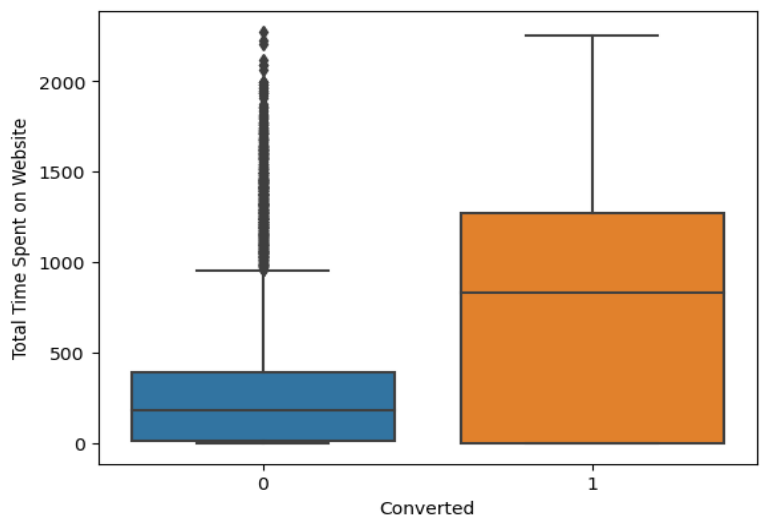


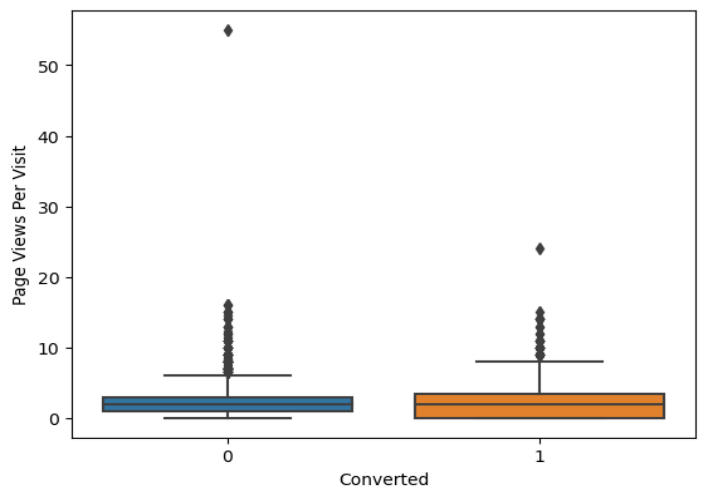


- Leads from HR, Finance & Marketing management specializations are high probability to convert







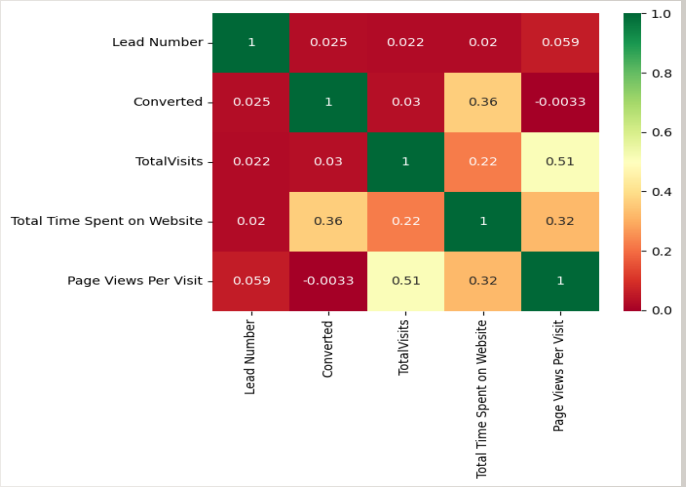


- The Above boxplot shows that total time spent on website

- It shows so many outliers present at page views per visit

Correlation analysis using heat map:

1.There is no high correlation between them



DATA CONVERSION:

- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 9074
- Total Columns for Analysis: 13

Train- Test split:

Train - Test Split

In [41]:
```
y=df['Converted']
x=df.drop('Converted',axis=1)
```

In [42]:
```
# train -test split
x_train, x_test, y_train, y_test=train_test_split(x,y,train_size=0.7,random_state=100)
x_train.shape
```

Out[42]: (6351, 90)

1. We have splitted train data to 70% and test data is 30%

MODEL EVALUTION

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6337
Model Family:	Binomial	**Df Model:**	13
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2594.6
Date:	Mon, 13 Nov 2023	**Deviance:**	5189.3
Time:	08:57:21	**Pearson chi2:**	6.21e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.4031
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8779	0.085	-10.316	0.000	-1.045	-0.711
Do Not Email	-1.5980	0.172	-9.272	0.000	-1.936	-1.260
Total Time Spent on Website	1.1340	0.041	27.976	0.000	1.055	1.213
Lead Origin_Landing Page Submission	-0.4002	0.093	-4.307	0.000	-0.582	-0.218
Lead Origin_Lead Add Form	3.3771	0.230	14.683	0.000	2.926	3.828
Lead Source_Olark Chat	1.2739	0.122	10.420	0.000	1.034	1.514
Lead Source_Welingak Website	1.9632	0.751	2.613	0.009	0.490	3.436
Last Activity_Had a Phone Conversation	2.7372	0.750	3.647	0.000	1.266	4.208
Last Activity_Olark Chat Conversation	-1.3462	0.166	-8.128	0.000	-1.671	-1.022
Last Activity_SMS Sent	1.3137	0.075	17.564	0.000	1.167	1.460
Specialization_Other	-0.4485	0.167	-2.693	0.007	-0.775	-0.122
What is your current occupation_Other	-1.0858	0.113	-9.633	0.000	-1.307	-0.865
What is your current occupation_Working Professional	2.4892	0.186	13.395	0.000	2.125	2.853
Last Notable Activity_Unreachable	2.0674	0.494	4.185	0.000	1.099	3.036

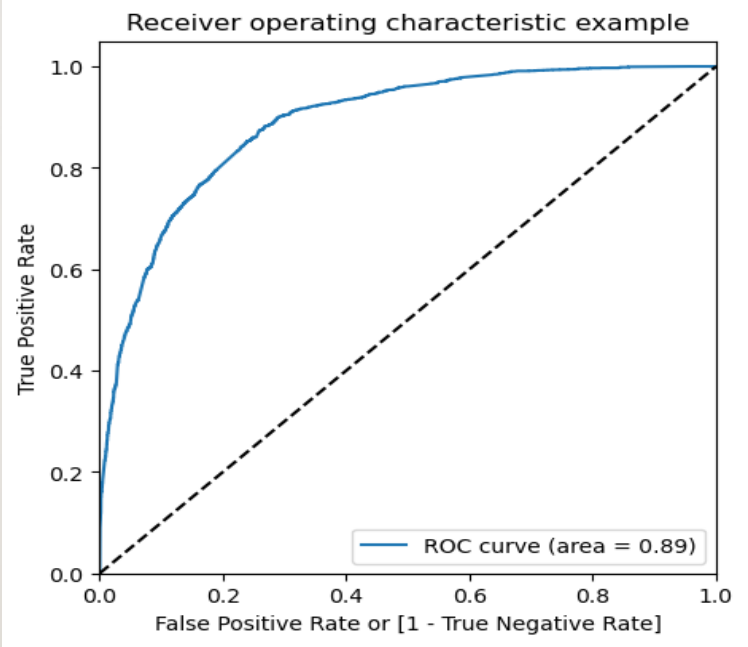
- We can see that p-value is less than 0.05 and vif is less than 0.5 for all variables

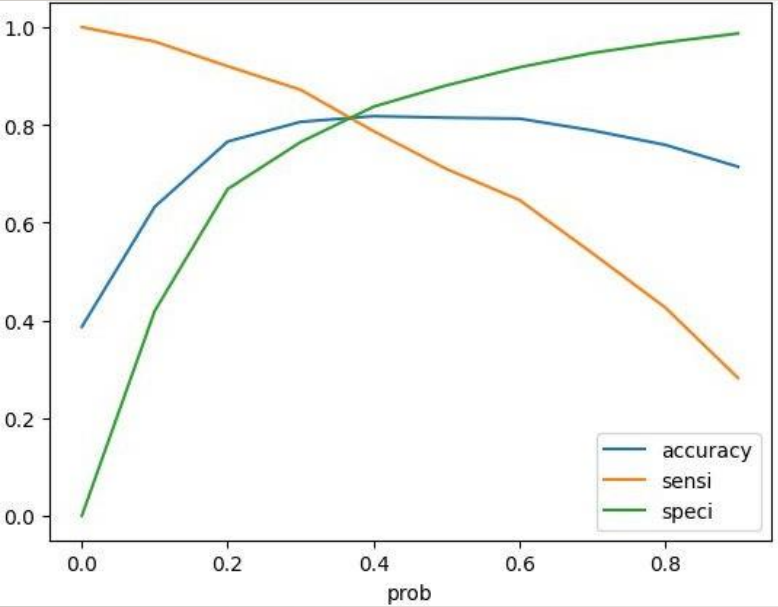
MODEL BUILDING

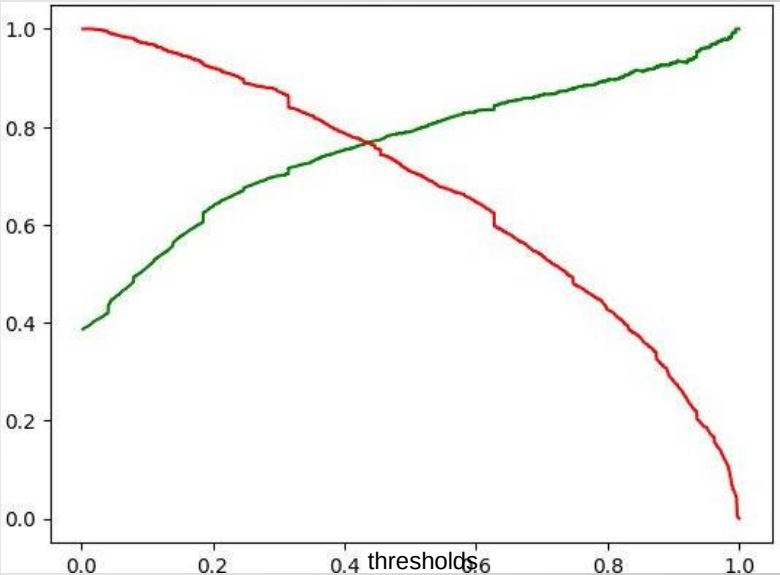
- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- Predictions on test data set
- Overall accuracy 81%

ROC CURVE

- Finding Optimal Cut off Point
- Optimal cut off probability is that
- probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.35.




1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
prob
accuracy
sensi
speci


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.8
1.0
thresholds

METRIC FOR TRAIN SET

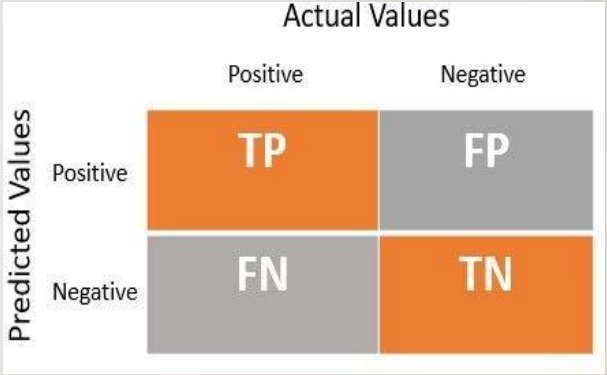
Accuracy : 80%

$$\text{Precision} = \frac{TP}{(TP+FP)} = 79\ \%$$

$$\text{Recall} = \frac{TP}{(TP+FN)} = 70\%$$

$$\text{Sensitivity} = \frac{TP}{(TP+FN)} = 83\ \%$$

$$\text{Specificity} = \frac{TN}{(TN+FP)} = 77\%$$



Confusion Matrix of Train Set	
3035	870
410	2036

METRIC FOR TEST SET

Accuracy : 80%

$$\text{Precision} = \frac{TP}{(TP+FP)} = 70\ \%$$

$$\text{Recall} = \frac{TP}{(TP+FN)} = 83\ \%$$

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} = 83\ \%$$

$$\text{Specificity} = \frac{TN}{(TN + FP)} = 78\ \%$$

Confusion Matrix of Test Set	
1355	379
166	823

Lead score:

```
Logr_model.params.sort_values(ascending=False)
```

```
Lead Origin_Lead Add Form                                3.377075
Last Activity_Had a Phone Conversation                   2.737236
What is your current occupation_Working Professional     2.489167
Last Notable Activity_Unreachable                        2.067432
Lead Source_Welingak Website                             1.963232
Last Activity_SMS Sent                                   1.313687
Lead Source_Olark Chat                                   1.273944
Total Time Spent on Website                              1.134004
Lead Origin_Landing Page Submission                     -0.400195
Specialization_Other                                    -0.448545
const                                                   -0.877930
What is your current occupation_Other                   -1.085831
Last Activity_Olark Chat Conversation                   -1.346210
Do Not Email                                            -1.598011
dtype: float64
```

RECOMMENDATIONS

- By referring to the data visualizations, focus on
  - *Increasing the conversion rates for the categories generating more leads and*
  - *Generating more leads for categories having high conversion rates.*
- Pay attention to the relative importance of the features in the model and their positive or negative impact on the probability of conversion.
- Based on varying business needs, modify the probability threshold value for identifying potential leads.
- Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

- We see max number of leads are generated by google / direct traffic. Max conversion ratio is by reference and welingak website.
- Leads who spent more time on website, more likely to convert.
- Most common last activity is email opened. highest rate = SMS Sent. Max are unemployed. Max conversion with working professional.

THANKYOU