LEAD SCORE_CASE STUDY
CASE STUDY

By Anshul Tare and Rashi Gupta

Problem Statement

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. Moreover, the company also gets leads through past referrals.

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 GOAL

X Educationyou want model to select the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

The company requires build a model 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%.

Steps Envolved

EDA
- Cleaning, outlier Treatment,
- Univariate and bivariate Analysis

Preprocessing
- Scaling using MinMax Scalar
- One-Hot Encoding Creating dummy

Model building
- Feature Elimination using RFE
- Manual Feature Elimination with p-value and VIF

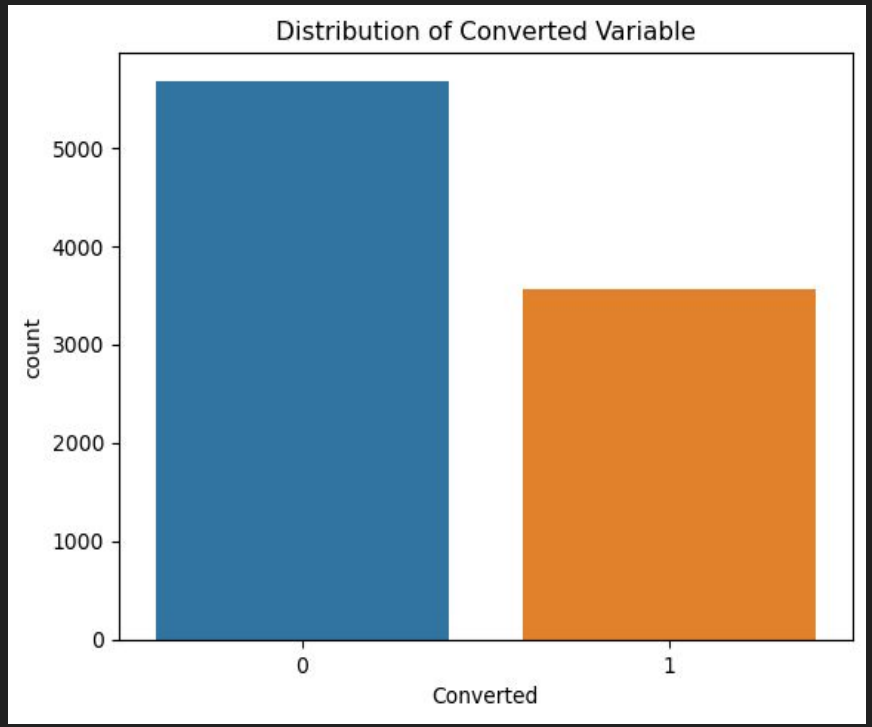
Plotting the ROC Curve
Finding Optimal Cutoff Point
Metrics beyond simply accuracy
Precision and Recall
Making predictions on the test set

EXPLORATORY DATA ANALYSIS (EDA):

UNIVARIATE ANALYSIS:

- As per problem statement 'Converted' is our target variable.
- The target variable, Indicates whether a lead has been successfully converted or not.
  - 0: Not converted into lead.
  - 1: Lead has been successfully Converted.

We can observe from value count and count plot that data is properly balanced with respect to ratio.



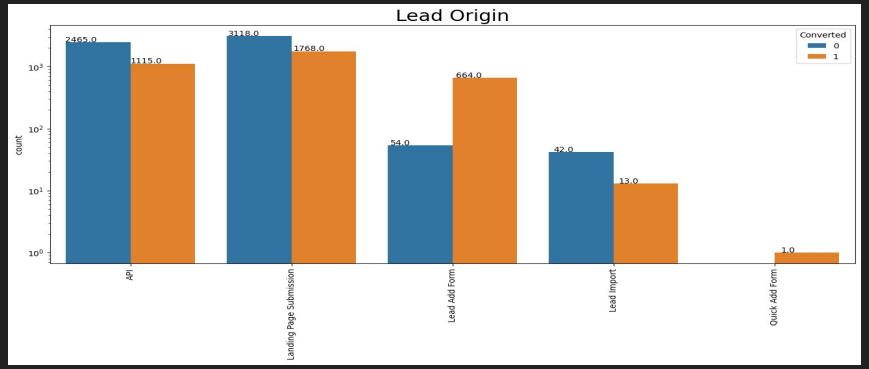
Analyzing Categorical Variables: Lead Origin' Vs 'Converted' 'Lead Source' based on 'Converted'

OBSERVATION: We can observe & calculate from the plot that:

a. Conversion rate for 'API' is ~ 31% and for 'Landing Page Submission' is ~36%.

b. For 'Lead Add Form' number of conversion is more than unsuccessful conversion.

c. Count of 'Lead Import' is lesser.



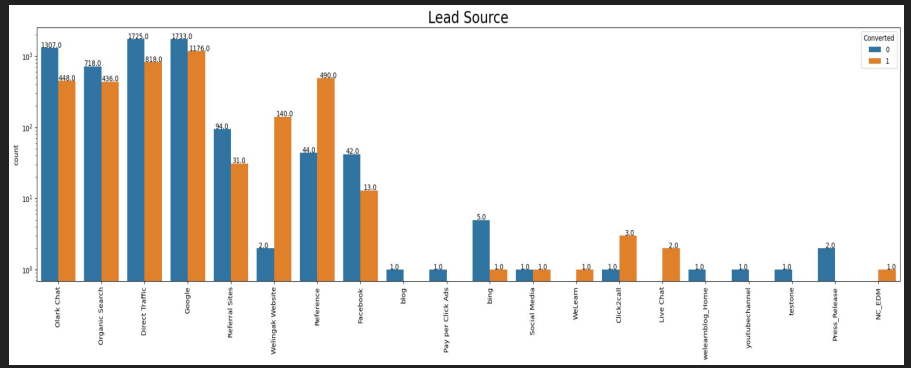
OBSERVATION:

a. Google and Direct traffic generates maximum number of leads.

b. Conversion rate of 'Reference' and 'Welingak Website' leads is high.

RECOMMENDATION:

To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

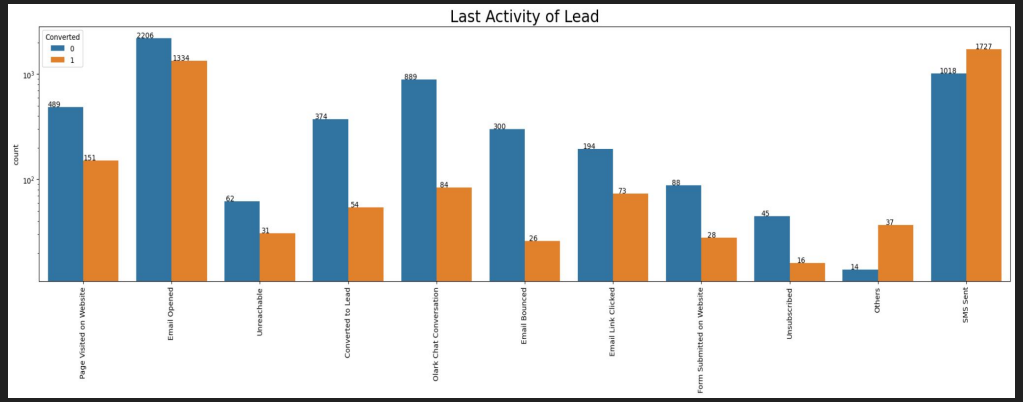


Analyzing Categorical Variables: Last Activity' Vs 'Converted' 'Current Occupation' based on 'Converted'

OBSERVATION:From above plot we can observe :

a. Conversion rate for last activity of 'SMS Sent'is ~63%.

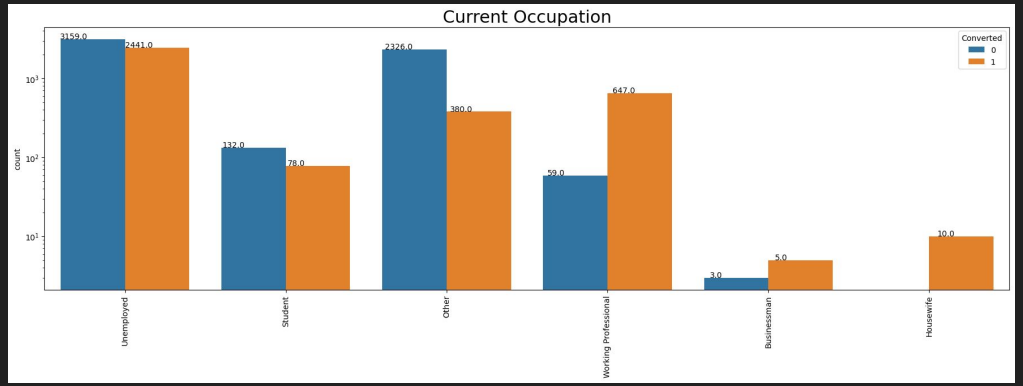
b. Highest last activity of leads is 'Email Opened'.



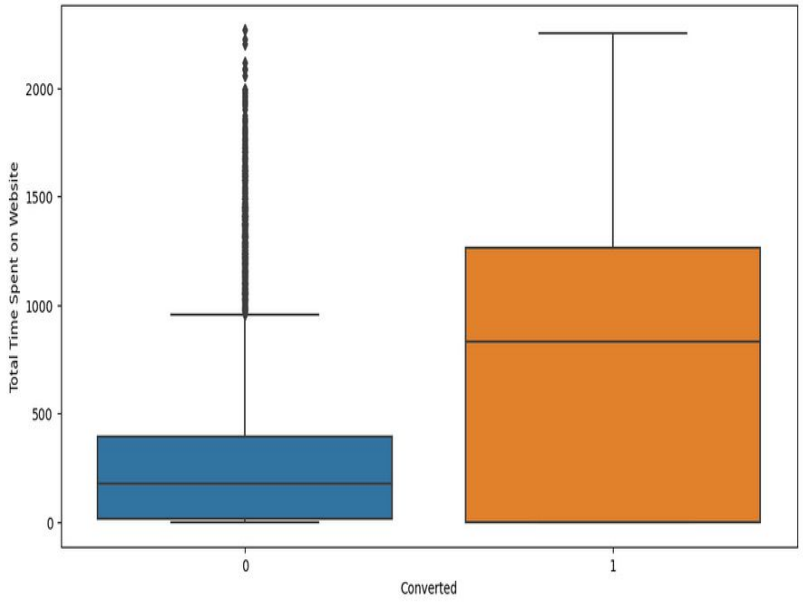
OBSERVATION:

1. 'Unemployed' leads are generating more number of leads and having ~45% conversion rate.

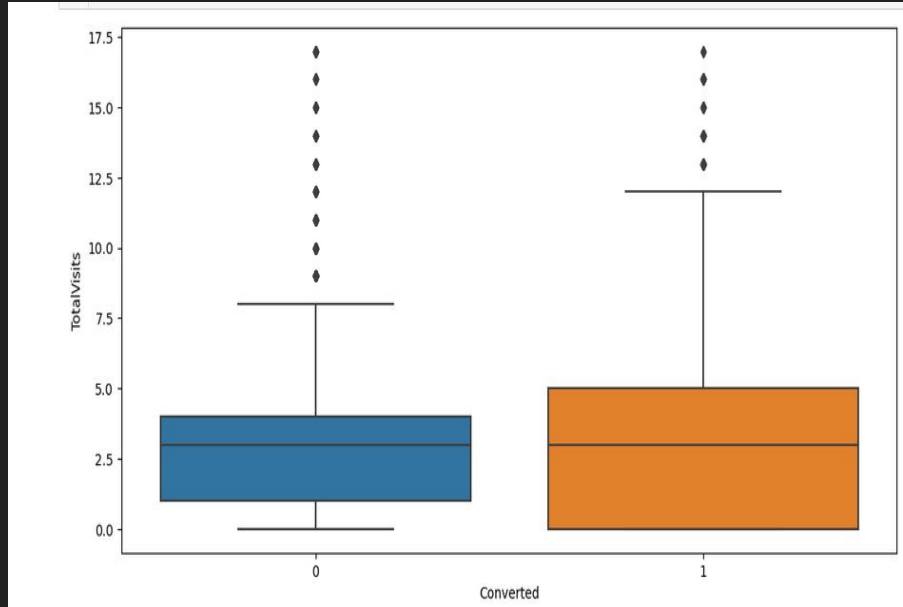
2. Conversion rate is higher for 'Working Professionals' .



"Total Visits" and "Total Time Spent on Website" vs Converted variable to check data distribution

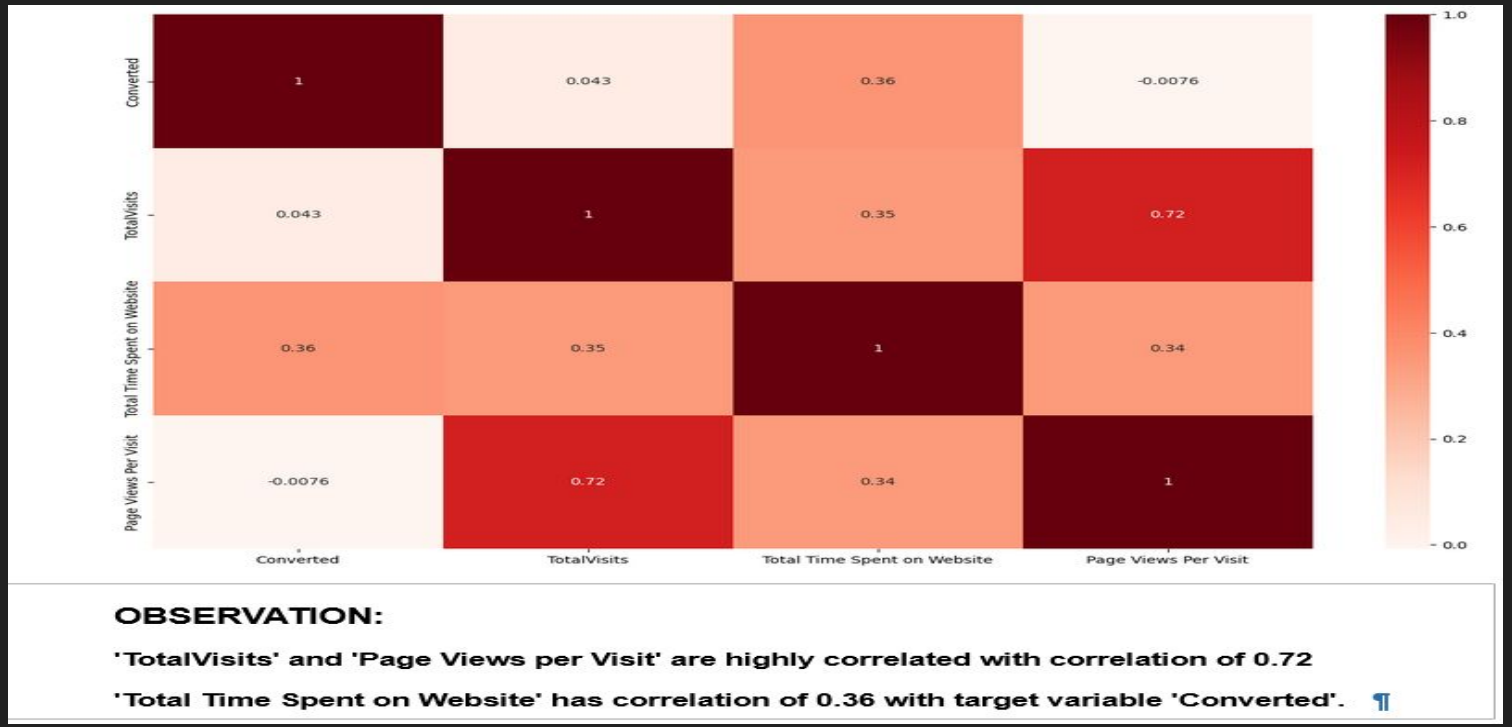


OBSERVATION: Leads spending more time on website are more likely to opt for curses or converted.



OBSERVATION: From above plot we can see that median for converted and non-converted is approx same.

Heatmap to understand the attributes correlation:

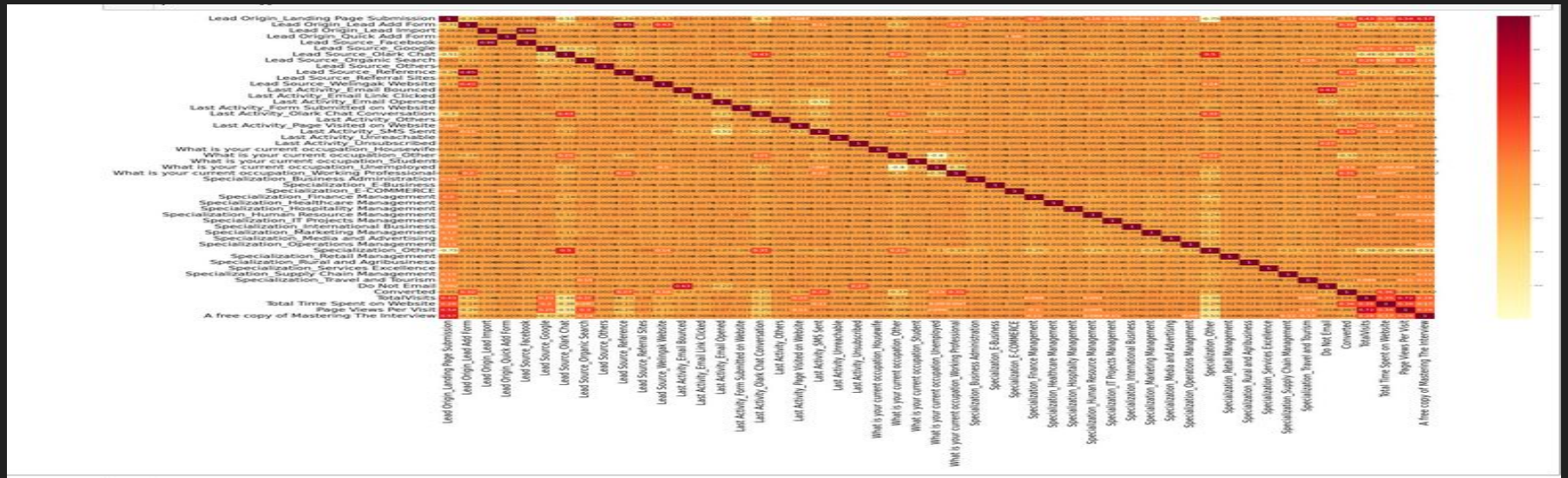


OBSERVATION:

'TotalVisits' and 'Page Views per Visit' are highly correlated with correlation of 0.72

'Total Time Spent on Website' has correlation of 0.36 with target variable 'Converted'. ¶

correlation coefficients to see which variables are highly correlated



In [ ]:

```
# OBSERVATION: The heatmap clearly shows which all variable are multicollinear in nature, and which variable have high
## We will refer this map for building the logistic model so as to validate different correlated values along with VIF .
## Hence from above heatmap we can see that:
## a. 'Lead Source_Facebook' and 'Lead Origin_Lead Import' having higher correlation of 0.98.
## b. 'Do Not Email' and 'Last Activity_Email Bounced' having higher correlation.
## c. 'Lead Origin_Lead Add Form' ,'Lead Source_Welingak Website', 'Last Activity_SMS Sent' and 'What is your current o
## d. 'Lead Origin_Lead Add Form' and 'Lead Source_Referance' having higher correlation of 0.85.
## e. 'TotalVisits' and 'Page Views Per Visit' having correlation of 0.72.
```

Summary from logistic regression model and VIF of final model

Out[651]:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6363
Model:	GLM	Df Residuals:	6349
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2577.6
Date:	Sun, 19 Mar 2023	Deviance:	5155.3
Time:	16:56:46	Pearson chi2:	6.80e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.4052
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3519	0.126	-2.803	0.005	-0.598	-0.106
Lead Origin_Landing Page Submission	-0.8696	0.129	-6.739	0.000	-1.123	-0.617
Lead Origin_Lead Add Form	2.9788	0.211	14.145	0.000	2.566	3.392
Lead Source_Olark Chat	1.1708	0.124	9.422	0.000	0.927	1.414
Lead Source_Welingak Website	3.1537	1.029	3.065	0.002	1.137	5.170
Last Activity_Olark Chat Conversation	-1.2222	0.167	-7.323	0.000	-1.549	-0.895
Last Activity_SMS Sent	1.3824	0.075	18.345	0.000	1.235	1.530
Last Activity_Unsubscribed	1.4457	0.449	3.219	0.001	0.565	2.326
What is your current occupation_Other	-1.1883	0.089	-13.389	0.000	-1.362	-1.014
What is your current occupation_Working Professional	2.3930	0.189	12.657	0.000	2.022	2.764
Specialization_Hospitality Management	-0.9632	0.336	-2.864	0.004	-1.622	-0.304
Specialization_Other	-0.8710	0.124	-7.038	0.000	-1.114	-0.628
Do Not Email	-1.5728	0.180	-8.738	0.000	-1.926	-1.220
Total Time Spent on Website	1.0724	0.040	26.650	0.000	0.993	1.151

Out[652]:

	Features	VIF
10	Specialization_Other	2.18
2	Lead Source_Olark Chat	2.04
0	Lead Origin_Landing Page Submission	1.66
7	What is your current occupation_Other	1.62
1	Lead Origin_Lead Add Form	1.52
5	Last Activity_SMS Sent	1.51
4	Last Activity_Olark Chat Conversation	1.48
3	Lead Source_Welingak Website	1.31
12	Total Time Spent on Website	1.25
8	What is your current occupation_Working Profes...	1.20
11	Do Not Email	1.20
6	Last Activity_Unsubscribed	1.10
9	Specialization_Hospitality Management	1.02

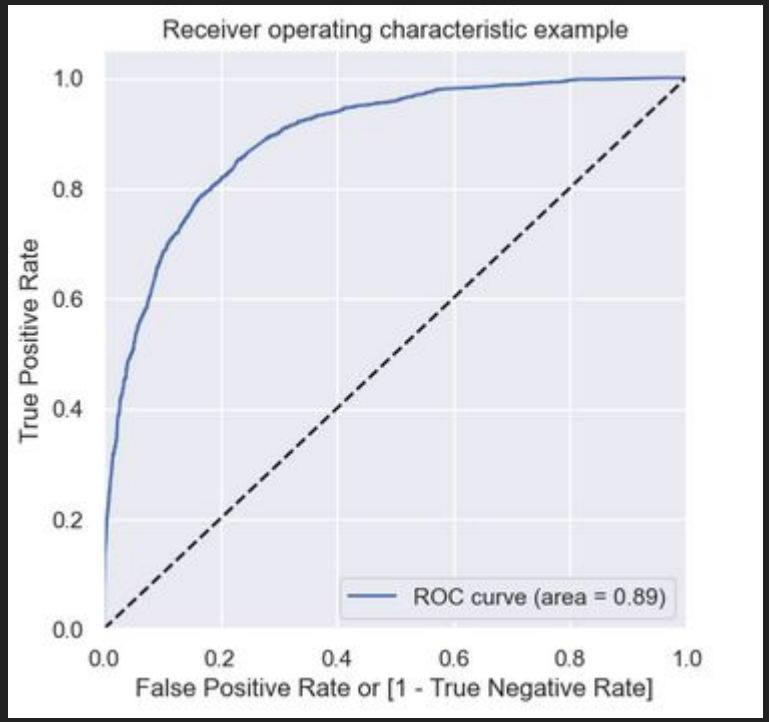
In [ ]:

```
# OBSERVATION:
# From model 'logm3' we can see that P-values of variables are significant and VIF values are below 3 .
# So we need not drop any more variables and we can proceed with making predictions using this model only
    #considering model 'logm3' as final model.
```

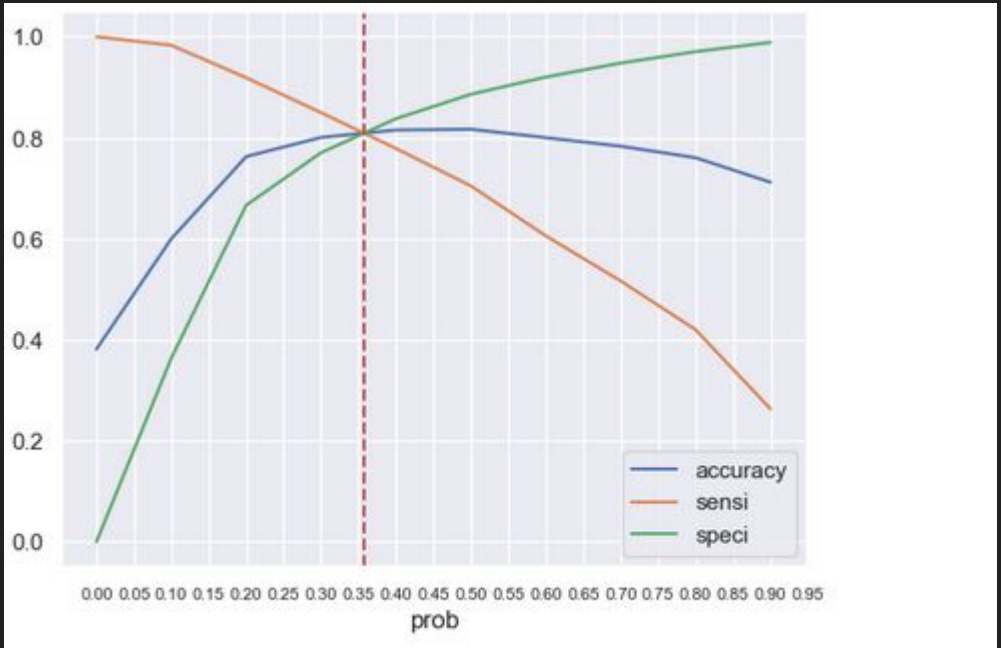
Plotting the ROC Curve.

- ROC curve will help in demonstrating the following:
  a. It will show the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
  b. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
  c. The closer the curve comes to the 45-degree diagonal of the ROC s

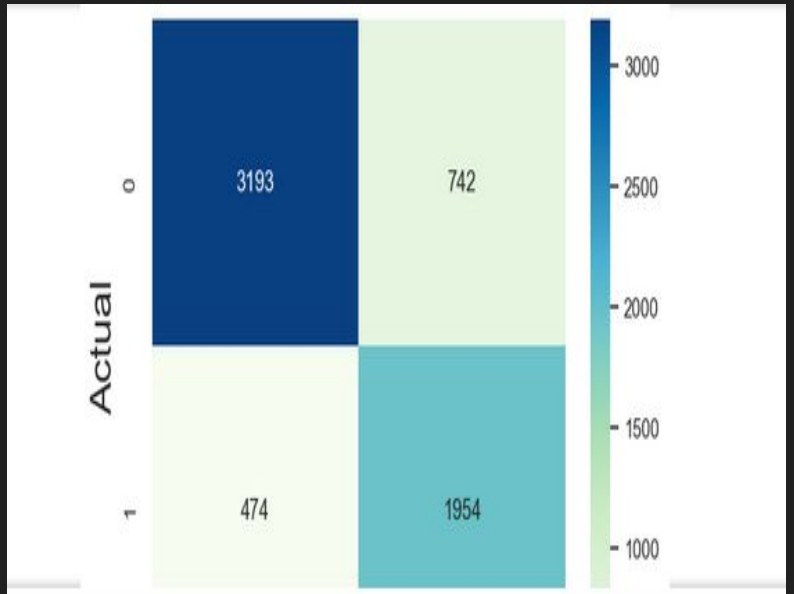
- OBSERVATION: We are getting a good value of 0.89 indicating a good predictive model(As ROC Curve should be a value close to 1)



Model Evaluation (Train and Test)

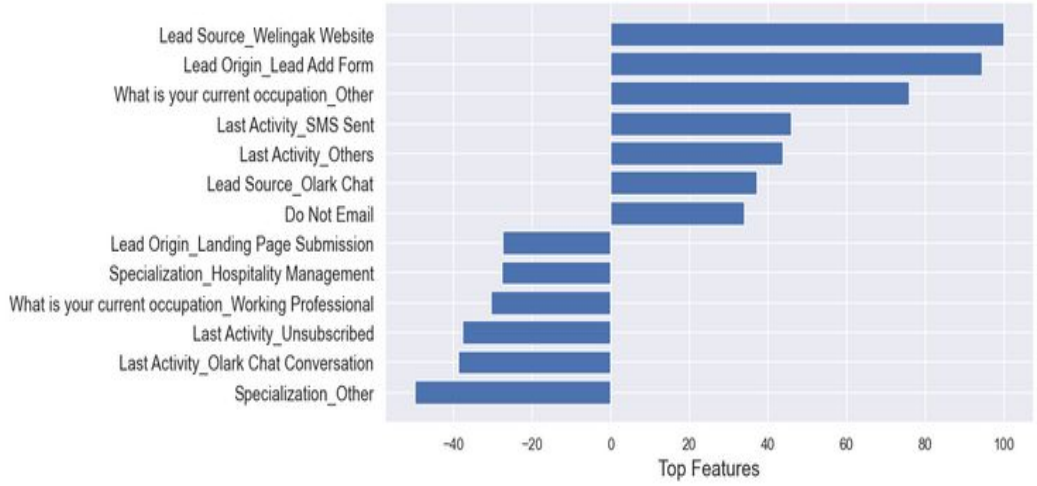



```
1 ## From the curve above, it seems that 0.358 is optimal cutoff point to take.
```




```
In [ ]: 1 # Actual/Predicted        not_converted   converted
        2       # not_converted      3170            765
        3       # converted          463             1965
```

feature variables based on their relative importance



In [ ]:

```
# FINAL MODEL LINE EQUATION:
##
Converted = 0.261843 + 3.15 X Lead Source_Welingak Website
+ 2.98 X Lead Origin_Lead Add Form + 2.39 X What is your current occupation_Working Professional
+ 1.45 X Last Activity_Unsubscribed + 1.38 X Last Activity_SMS Sent + 1.17 X Lead Source_Olark Chat
+ 1.07 X Total Time Spent on Website - 0.87 X Lead Origin_Landing Page Submission - 0.87 X Specialization_Other
- 0.96 X Specialization_Hospitality Management - 1.19 X What is your current occupation_Other
- 1.22 X Last Activity_Olark Chat Conversation
```

```
# FINAL OBSERVATION:
```

```
# Evaluation Metrics for the train Dataset:
# A. Accuracy :0.80
# B. Sensitivity:~0.80
# C. Specificity:0.81
# D. Precision: 0.72
# E. Recall: 0.80
```

```
# Evaluation Metrics for the test Dataset:
# A. Accuracy : 0.80
# B. Sensitivity: ~ 0.80
# C. Specificity: 0.80
# D. Precision: 0.72
# E. Recall: 0.80
```

RECOMMENDATION:

To improve the potential lead conversion rate X-Education will have to mainly focus important features responsible for good conversion rate are :

1. Lead Source_Welingak Website : As conversion rate is higher for those leads who got to know about course from 'Welingak Website',so company can focus on this website to get more number of potential leads.

2. What is your current occupation_Working Professional : The lead whose occupation is 'Working Professional' having higher lead conversion rate ,company should focus on working professionals nad try to get more number of leads.

3. Lead Origin_Lead Add Form: Leads who have engaged through 'Lead Add Form' having higher conversion rate so company can focus on it to get more number of leads cause have a higher chances of getting converted.

4. Last Activity_SMS Sent: Lead whose last activity is sms sent can be potential lead for company.

5. Total Time Spent on website: Leads spending more time on website can be our potential lead.