Lead Scoring Case Study

Approach to Solve the Problem

1. Understanding the Problem Statement

- The objective is to improve the lead conversion rate by identifying potential "Hot Leads.“
- Creating a model to assign a **lead score** indicating the likelihood of a lead converting.
- The goal is to prioritize the sales team's focus on high-potential leads and achieve a conversion rate of **~80**

2. Data Understanding and Preprocessing

Dataset Overview:

```
df.shape
```

```
(9240, 37)
```

3.Handle Missing Values

```
# Checking the columns with percentage of null values in descending order

round(100*(df.isnull().sum()/len(df.index)),2).sort_values(ascending=False).head(15)
```

```
How did you hear about X Education              78.46
Lead Profile                                    74.19
Lead Quality                                    51.59
Asymmetrique Profile Score                      45.65
Asymmetrique Activity Score                     45.65
Asymmetrique Activity Index                     45.65
Asymmetrique Profile Index                      45.65
City                                            39.71
Specialization                                  36.58
Tags                                            36.29
What matters most to you in choosing a course   29.32
What is your current occupation                 29.11
Country                                         26.63
Page Views Per Visit                             1.48
TotalVisits                                      1.48
dtype: float64
```

```
# Dropping the columns as % of missing values are more than 40%

for i in ndf.columns[2:]:
    f = round(ndf[i].isnull().sum()/ len(ndf) *100,2)
    if f >40:
        ndf.drop(columns = i, axis = 1,inplace = True)
```

```
ndf.shape
```

```
(9240, 30)
```

```
#Dropping columns whose normalized value greater than 90%

for i in ndf.columns[2:]:
    x = ndf[i].value_counts(normalize =True)
    if x.iloc[0]>.90:
        ndf.drop(i,axis = 1, inplace = True)
```

```
ndf.shape
```

```
(9240, 15)
```

3. Missing Values Imputation

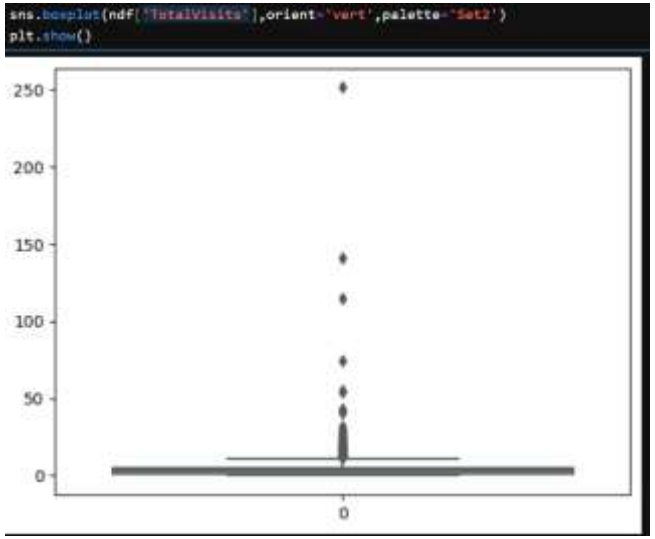
```
#Missing value imputation
for i in ndf.columns[2:]:
    s= ndf[i].isnull().sum()
    if  s != 0:
      print(i,s, '\n')
```

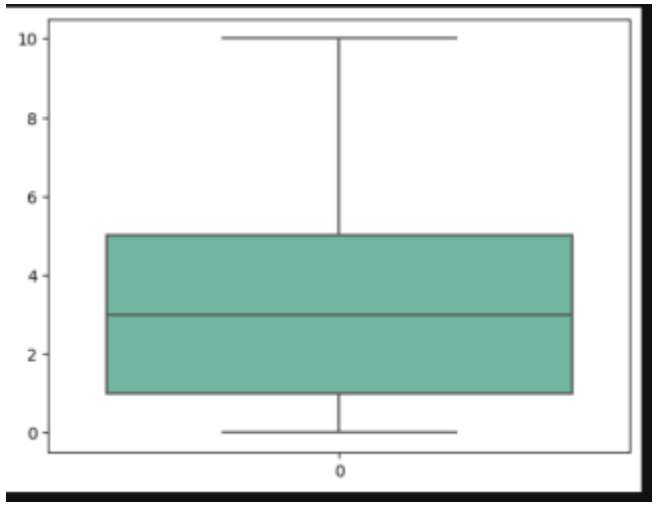
```
TotalVisits 137

Page Views Per Visit 137

Last Activity 103

Specialization 3380

What is your current occupation 2690

Tags 3353

City 3669
```

- Page Views Per Visit - replacing nan values with median
- TotalVisits - replacing nan values with median
- Rest Paramesters with ‘UNKNOWN’

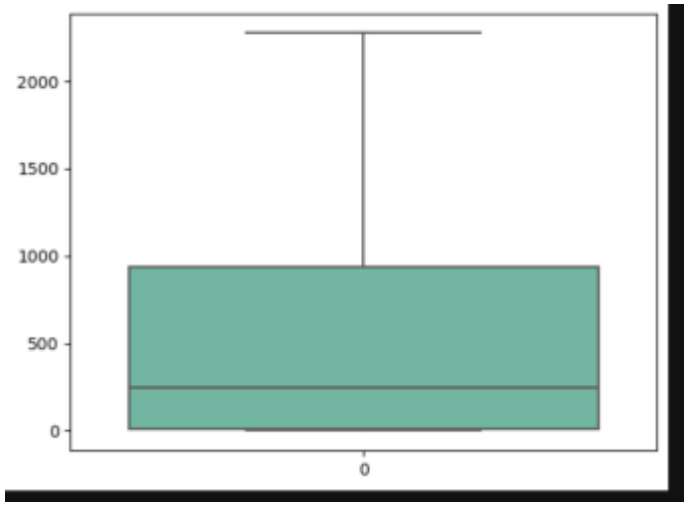
4. Outlier Treatment:

A. 'TotalVisits' - capping between 5th – 95 th Percentile

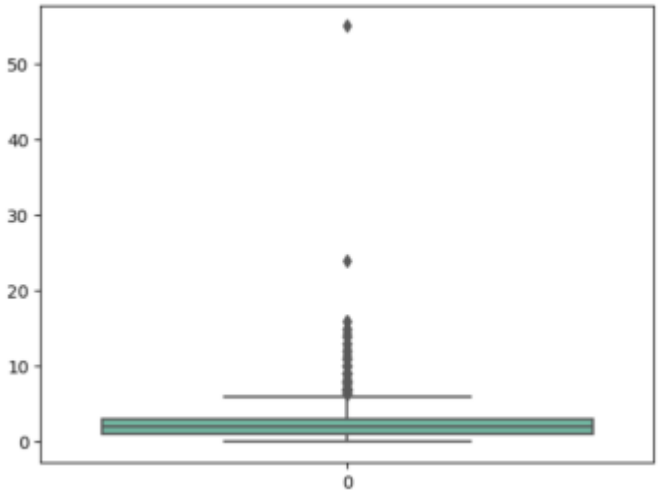


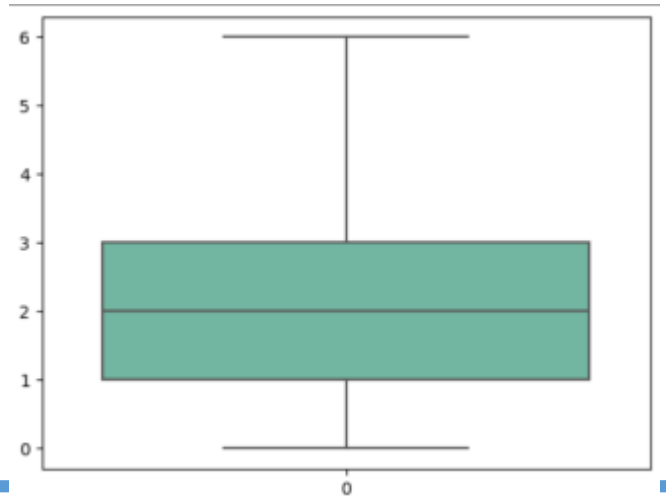


B. Total Time Spent on Website-



C. 'TotalVisits' – capping between 5th – 95 th Percentile





5. Categorising and Grouping variables with negligible counts

A. Lead Source

'bing', 'Click2call', 'Social Media', 'Live Chat','Press_Release','Pay per Click Ads','blog','WeLearn','welearnblog_Home',' youtubechannel','testone','NC_EDM'

B. Last Activity

'Approached upfront','View in browser link Clicked','Email Received','Email Marked Spam','Visited Booth in Tradeshow','Resubscribed to emails'

C. Tags

'Ringing','Busy','Lost to EINS', 'Already a student','switched off', 'opp hangup', 'wrong number given', 'invalid number', 'Diploma holder (Not Eligible)','switched off', 'Not doing further education','Interested in full time MBA','University not recognized', 'Recognition issue (DEC approval)', 'Shall take in the next coming month','Lateral student','Lost to Others','Still Thinking','in touch with EINS','number not provided','Want to take admission but has financial problems','In confusion whether part time or DLP','Interested in Next batch'

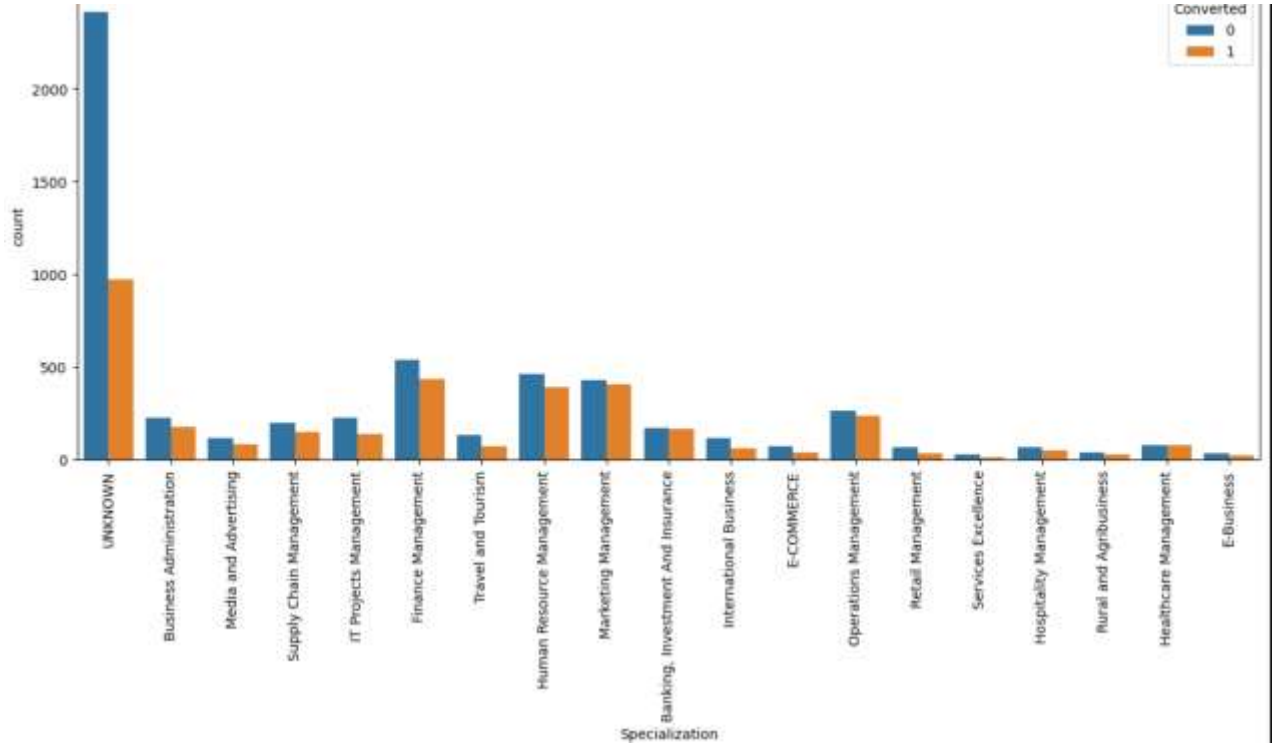
D. City

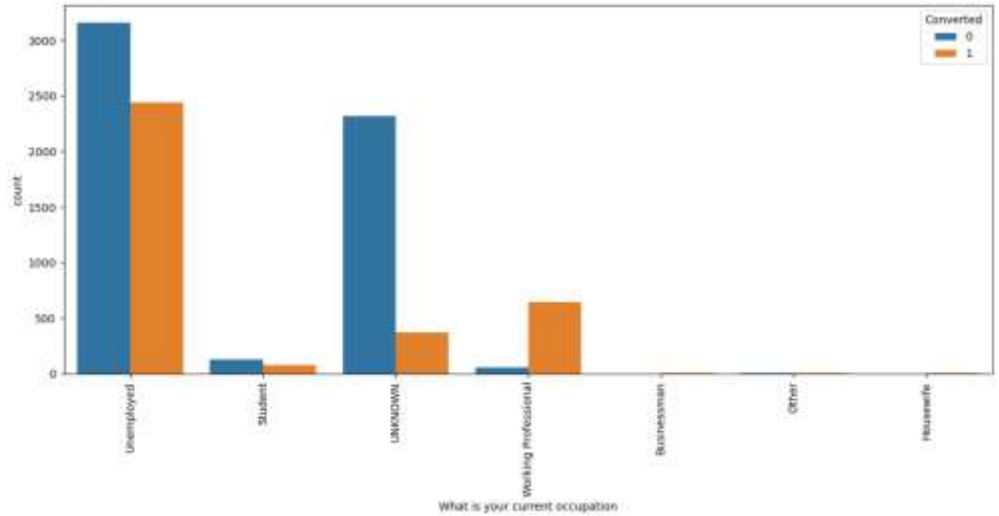
'Other Metro Cities','Tier II Cities'

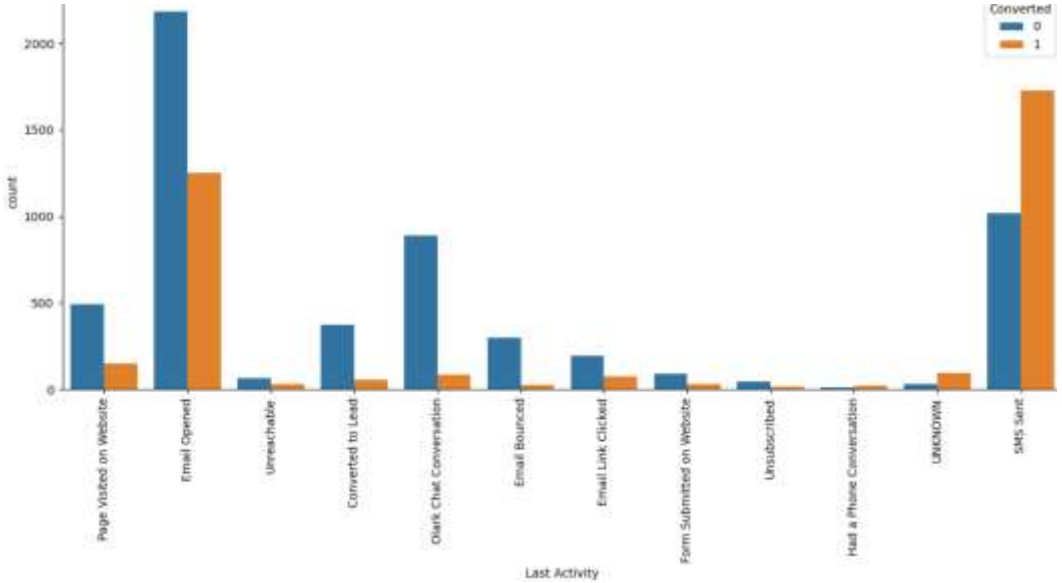
E. Last Notable Activity'

Grouping variabales having counts >10

6. Bi- variate Analysis







7. Final Columns for Analysis

```
#Final Columns
ndf.shape
```

```
(9240, 12)
```

```
ndf.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 12 columns):
 #   Column                         Non-Null Count  Dtype  
---  ------                         --------------  -----  
 0   Lead Origin                    9240 non-null   object 
 1   Lead Source                    9240 non-null   object 
 2   Converted                      9240 non-null   int32  
 3   TotalVisits                    9240 non-null   float64
 4   Total Time Spent on Website    9240 non-null   int64  
 5   Page Views Per Visit           9240 non-null   float64
 6   Last Activity                  9240 non-null   object 
 7   Specialization                 9240 non-null   object 
 8   What is your current occupation  9240 non-null   object 
 9   Tags                           9240 non-null   object 
 10  City                           9240 non-null   object 
 11  Last Notable Activity          9240 non-null   object 
dtypes: float64(2), int32(1), int64(1), object(8)
memory usage: 830.3+ KB
```

8. Feature Selection

```
#Creating Dummy variables

object_columns = ndf.select_dtypes(include='object').columns.tolist()
print(object_columns)
dummies = pd.get_dummies(ndf[object_columns],drop_first=True)
```

```
ndf.shape
```

```
(9240, 69)
```

Using RFE selecting 20 feature columns

```
Index(['Total Time Spent on Website', 'Lead Origin_Landing Page Submission',
       'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat',
       'Lead Source_UNKNOWN', 'Lead Source_Welingak Website',
       'Last Activity_Email Bounced', 'Last Activity_Email Opened',
       'Last Activity_Form Submitted on Website', 'Last Activity_SMS Sent',
       'Specialization_Hospitality Management',
       'What is your current occupation_Working Professional',
       'Tags_Graduation in progress', 'Tags_Interested in other courses',
       'Tags_UNKNOWN', 'Last Notable Activity_Email Opened',
       'Last Notable Activity_Had a Phone Conversation',
       'Last Notable Activity_Modified',
       'Last Notable Activity_Olark Chat Conversation',
       'Last Notable Activity_Others'],
      dtype='object')
```

9. Model Results

A. GLM

Generalized Linear Model Regression Results

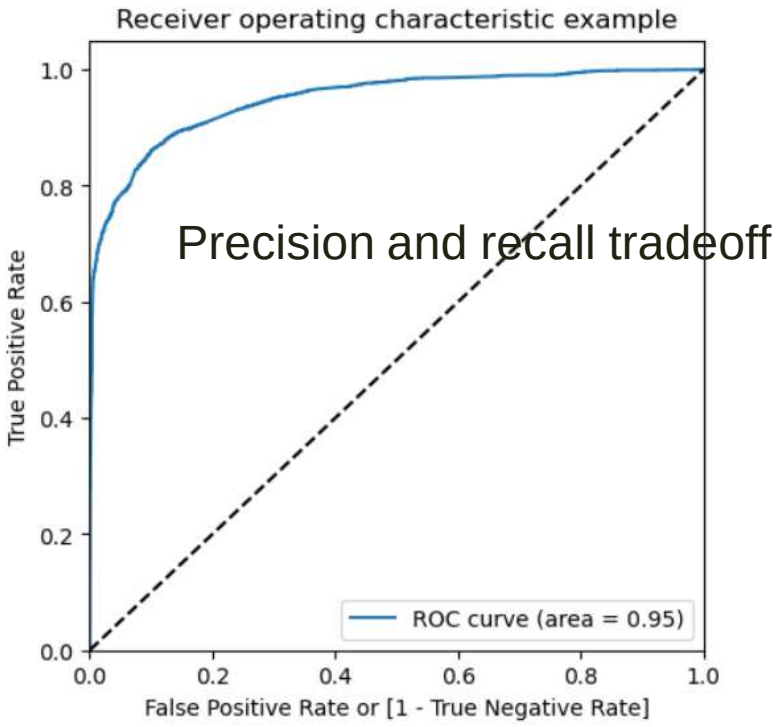
Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6453
Model Family:	Binomial	**Df Model:**	14
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1736.2
Date:	Tue, 21 Jan 2025	**Deviance:**	3472.5
Time:	20:00:32	**Pearson chi2:**	6.35e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.5473
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.6490	0.213	12.427	0.000	2.231	3.067
Total Time Spent on Website	1.0201	0.049	20.985	0.000	0.925	1.115
Lead Origin_Landing Page Submission	-0.7469	0.112	-6.648	0.000	-0.967	-0.527
Lead Origin_Lead Add Form	1.2121	0.277	4.369	0.000	0.668	1.756
Lead Source_Olark Chat	0.9286	0.143	6.512	0.000	0.649	1.208
Lead Source_Welingak Website	4.3540	0.772	5.642	0.000	2.842	5.866
Last Activity_Email Opened	0.7232	0.135	5.363	0.000	0.459	0.988
Last Activity_Form Submitted on Website	1.0688	0.435	2.455	0.014	0.216	1.922
Last Activity_SMS Sent	2.0213	0.134	15.075	0.000	1.758	2.284
What is your current occupation_Working Professional	0.9479	0.305	3.109	0.002	0.350	1.546
Tags_Graduation in progress	-5.0952	0.526	-9.688	0.000	-6.126	-4.064
Tags_Interested in other courses	-6.4348	0.368	-17.473	0.000	-7.157	-5.713
Tags_UNKNOWN	-4.9435	0.175	-28.195	0.000	-5.287	-4.600
Last Notable Activity_Had a Phone Conversation	3.5185	1.386	2.539	0.011	0.803	6.234
Last Notable Activity_Modified	-0.6844	0.110	-6.199	0.000	-0.901	-0.468

B. VIF

	Features	VIF
1	Lead Origin_Landing Page Submission	2.45
5	Last Activity_Email Opened	1.77
7	Last Activity_SMS Sent	1.73
3	Lead Source_Olark Chat	1.60
12	Last Notable Activity_Modified	1.60
2	Lead Origin_Lead Add Form	1.59
0	Total Time Spent on Website	1.24
4	Lead Source_Welingak Website	1.24
8	What is your current occupation_Working Profes...	1.18
10	Tags_Interested in other courses	1.11
6	Last Activity_Form Submitted on Website	1.05
9	Tags_Graduation in progress	1.02
11	Last Notable Activity_Had a Phone Conversation	1.00

10. ROC Curve and Model Evaluation

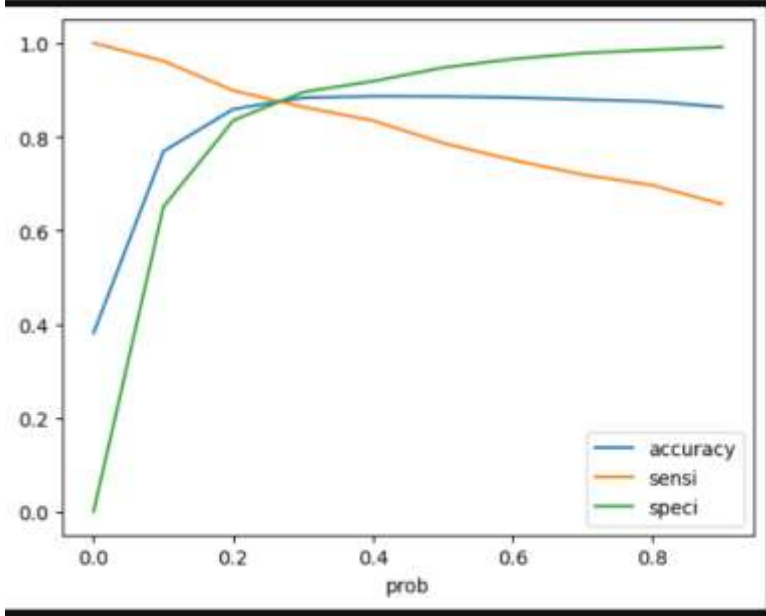


Precision and recall tradeoff

A. Graph showing ROC Curve Area

	prob	accuracy	sensi	speci
0.0	0.0	0.381262	1.000000	0.000000
0.1	0.1	0.769171	0.961882	0.650425
0.2	0.2	0.859307	0.898621	0.835082
0.3	0.3	0.883426	0.863747	0.895552
0.4	0.4	0.886518	0.834550	0.918541
0.5	0.5	0.886209	0.786699	0.947526
0.6	0.6	0.883890	0.750608	0.966017
0.7	0.7	0.879716	0.719384	0.978511
0.8	0.8	0.875387	0.696675	0.985507
0.9	0.9	0.863482	0.656123	0.991254

```
# Let's plot accuracy sensitivity and specificity for various probabilities.
cutoff_df.plot.line(x='prob', y=['accuracy','sensi','speci'])
plt.show()
```

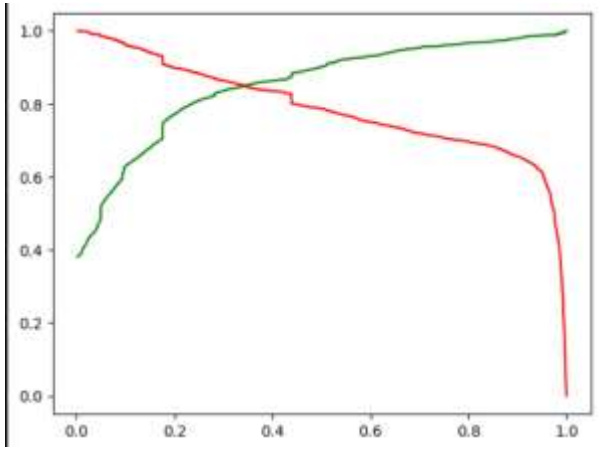


From the curve above, 0.3 is the optimum point to take it as a cutoff probability.

B. Graph showing optimum point

```
Precision :  0.9023255813953488
```

```
Recall :  0.786699107866991
```



C. Precision and recall trade off

11. Results:

Observations:

Running the model on the test data

1. accuracy : 89.36 %
2. sensitivity : 88.32 %
3. specificity : 87.18 %

Running the model on the Train data

1. accuracy : 87.62 %
2. sensitivity : 88.32 %
3. specificity : 87.18 %

```
hot_leads.shape
```

```
(768, 5)
```

```
Lead Source_Welingak Website                              4.354003
Last Notable Activity_Had a Phone Conversation            3.518481
const                                                     2.649043
Last Activity_SMS Sent                                    2.021298
Lead Origin_Lead Add Form                                 1.212099
Last Activity_Form Submitted on Website                   1.068844
Total Time Spent on Website                               1.020128
What is your current occupation_Working Professional      0.947908
Lead Source_Olark Chat                                    0.928604
Last Activity_Email Opened                                0.723211
Last Notable Activity_Modified                           -0.684412
Lead Origin_Landing Page Submission                      -0.746945
Tags_UNKNOWN                                             -4.943507
Tags_Graduation in progress                              -5.095233
Tags_Interested in other courses                         -6.434808
dtype: float64
```

- There are about 768 prospect leads which can be contacted and having high chances of enrolment into a course.

12. Recommendations:

The company should focus on leads coming from the followings given below:

- Lead Source_Welingak Website
- Last Notable Activity_Had a Phone Conversation
- Last Activity_SMS Sent
- Lead Origin_Lead Add Form
- Last Activity_Form Submitted on Website
- Total Time Spent on Website
- What is your current occupation_Working Professional
- Lead Source_Olark Chat
- Last Activity_Email Opened
- Last Notable Activity_Modified
- Lead Origin_Landing Page Submission
- Tags_UNKNOWN
- Tags_Graduation in progress
- Tags_Interested in other courses