# Lead-Scoring Case Study

Shivaprasad Metimath
Madhusudhan Anand

# 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.** Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team **start making calls, writing emails, etc.** Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%. Now, although 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. A typical lead conversion process can be represented using the funnel image on the left

# Goal

There are quite a few goals for this case study.

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. 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.

2. There are some more problems presented by the company which our model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

# Step 1. Data Preparation

```
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
Prospect ID                                      9240 non-null object
Lead Number                                      9240 non-null int64
Lead Origin                                      9240 non-null object
Lead Source                                      9204 non-null object
Do Not Email                                     9240 non-null object
Do Not Call                                      9240 non-null object
Converted                                        9240 non-null int64
TotalVisits                                      9103 non-null float64
Total Time Spent on Website                      9240 non-null int64
Page Views Per Visit                             9103 non-null float64
Last Activity                                    9137 non-null object
Country                                          6779 non-null object
Specialization                                   7802 non-null object
How did you hear about X Education               7033 non-null object
What is your current occupation                  6550 non-null object
What matters most to you in choosing a course    6531 non-null object
Search                                           9240 non-null object
Magazine                                         9240 non-null object
Newspaper Article                                9240 non-null object
X Education Forums                               9240 non-null object
Newspaper                                        9240 non-null object
Digital Advertisement                            9240 non-null object
Through Recommendations                          9240 non-null object
Receive More Updates About Our Courses           9240 non-null object
Tags                                             5887 non-null object
Lead Quality                                     4473 non-null object
Update me on Supply Chain Content                9240 non-null object
Get updates on DM Content                        9240 non-null object
Lead Profile                                     6531 non-null object
City                                             7820 non-null object
Asymmetrique Activity Index                      5022 non-null object
Asymmetrique Profile Index                       5022 non-null object
Asymmetrique Activity Score                      5022 non-null float64
Asymmetrique Profile Score                       5022 non-null float64
I agree to pay the amount through cheque         9240 non-null object
A free copy of Mastering The Interview           9240 non-null object
Last Notable Activity                            9240 non-null object
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB
```

37 Columns, mix of numeric and categorical variables. 9240 Rows

# Handling Missing Data

1. Check for missing values
2. Drop columns with more than 40% missing values
3. Check for rows with missing values
4. Remove rows that have missing values
5. Convert to NaN (values that don’t add meaning like ‘select’)
6. Drop NA
7. Check for corrections (google, Google)
8. Drop columns that won’t add any significance/variance
9. Check data again for completeness

# Before Cleanup

# After Cleanup

```
In [7]: round(100*(lead_score.isnull().sum(axis=0)/len(lead_score)),3)

Out[7]: Prospect ID                                      0.000
        Lead Number                                      0.000
        Lead Origin                                      0.000
        Lead Source                                      0.390
        Do Not Email                                     0.000
        Do Not Call                                      0.000
        Converted                                        0.000
        TotalVisits                                      1.483
        Total Time Spent on Website                      0.000
        Page Views Per Visit                             1.483
        Last Activity                                    1.115
        Country                                         26.634
        Specialization                                  15.563
        How did you hear about X Education              23.885
        What is your current occupation                 29.113
        What matters most to you in choosing a course   29.318
        Search                                           0.000
        Magazine                                         0.000
        Newspaper Article                                0.000
        X Education Forums                               0.000
        Newspaper                                        0.000
        Digital Advertisement                            0.000
        Through Recommendations                          0.000
        Receive More Updates About Our Courses           0.000
        Tags                                            36.288
        Lead Quality                                    51.591
        Update me on Supply Chain Content                0.000
        Get updates on DM Content                        0.000
        Lead Profile                                    29.318
        City                                            15.368
        Asymmetrique Activity Index                     45.649
        Asymmetrique Profile Index                      45.649
        Asymmetrique Activity Score                     45.649
        Asymmetrique Profile Score                      45.649
        I agree to pay the amount through cheque         0.000
        A free copy of Mastering The Interview           0.000
        Last Notable Activity                            0.000
        dtype: float64
```

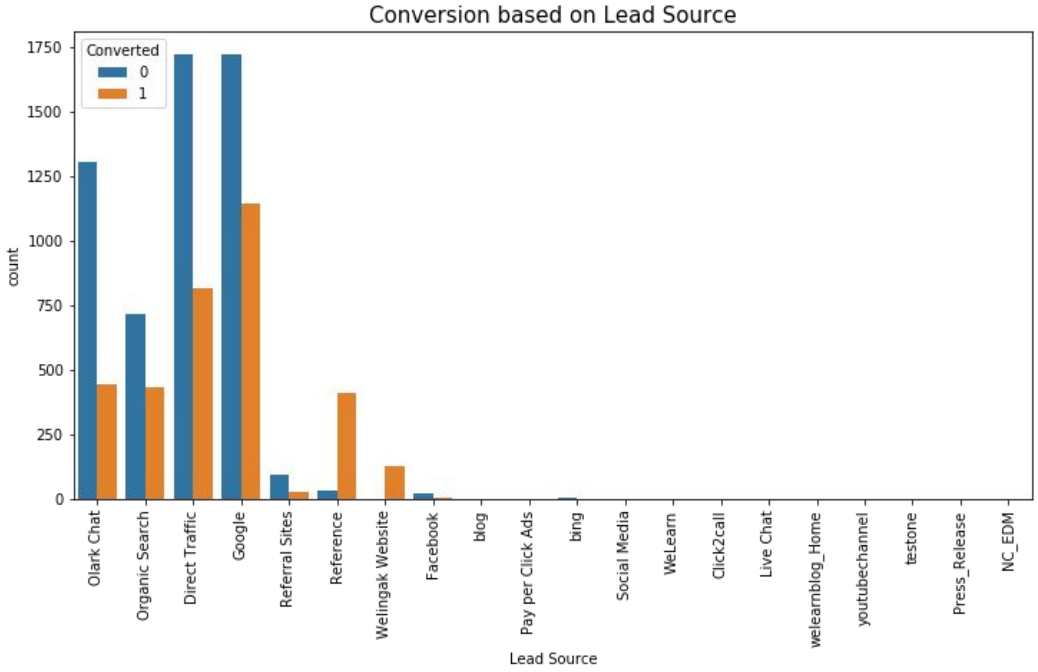
```
In [33]: round(100*(lead_score.isnull().sum(axis =0)/len(lead_score)), 2)

Out[33]: Lead Origin                                0.0
         Lead Source                                0.0
         Do Not Email                               0.0
         Converted                                  0.0
         TotalVisits                                0.0
         Total Time Spent on Website                0.0
         Page Views Per Visit                       0.0
         Last Activity                              0.0
         What is your current occupation            0.0
         A free copy of Mastering The Interview     0.0
         Last Notable Activity                      0.0
         dtype: float64
```

After the above cleaning and replacement the data appears to be in a good condition

hence we are good to proceed with further analysis and model building

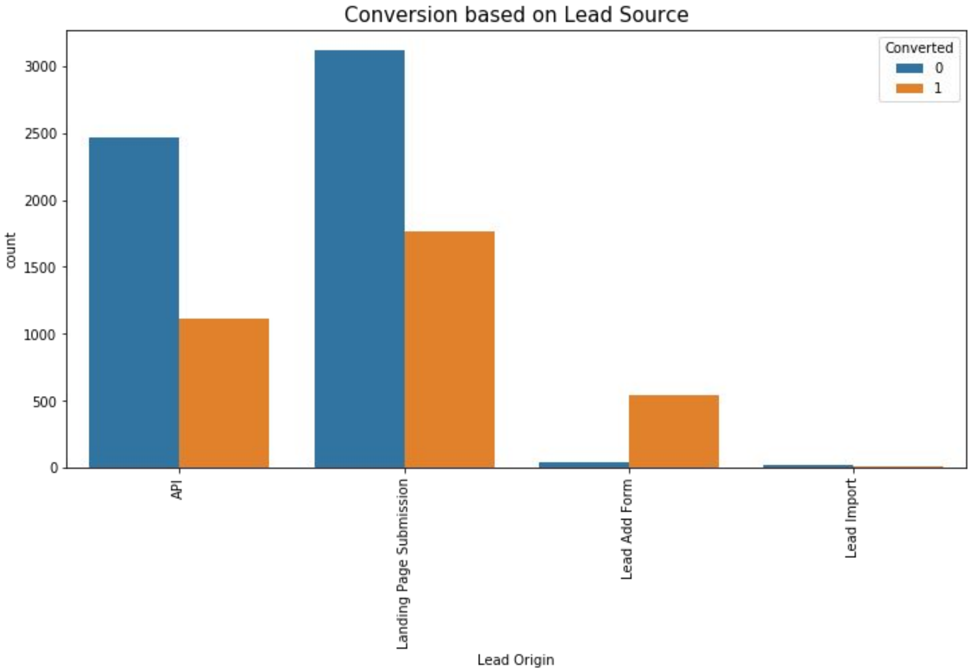
# EDA

# Lead conversion based on Lead Source




Check that **Olark Chat, Direct Traffic and Google** has more conversion. Conversion rate is good at **Organic Search and on Reference**.

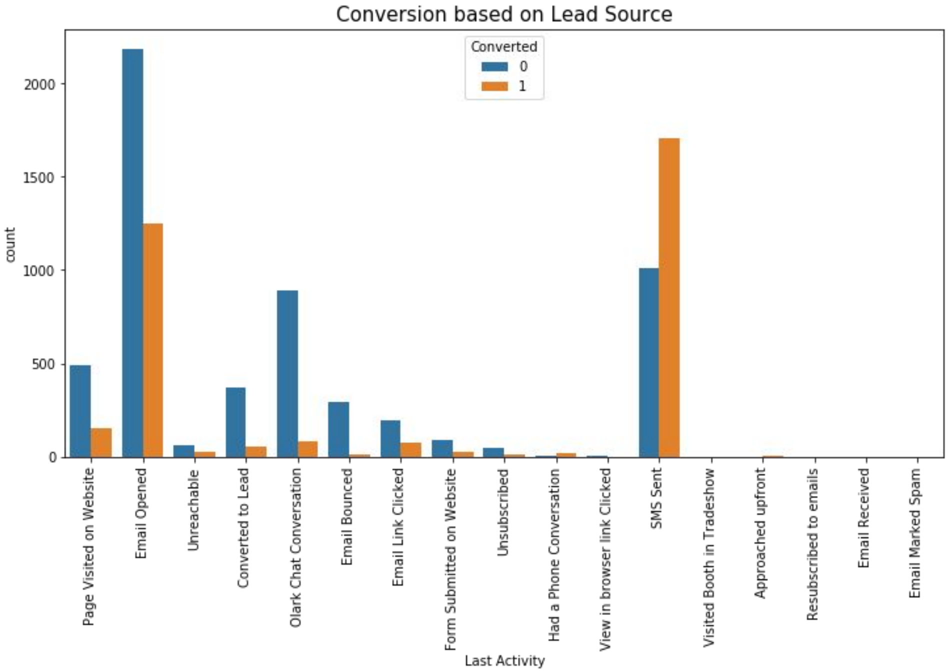
# Conversion based on Lead Source




From the above plot we see that the highest conversion rate based on lead source are:
1. Landing page submission

2. API

# Contributors of high conversion rate




Top 3 contributors with high conversion rate of lead are:

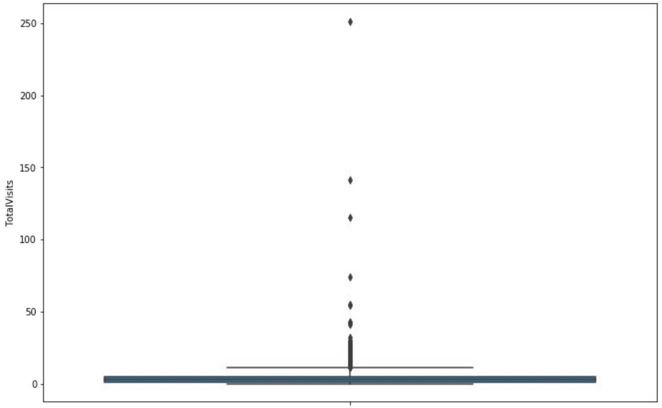
1. SMS sent
2. Email opened
3. Olark chat conversion

# Outlier Treatment

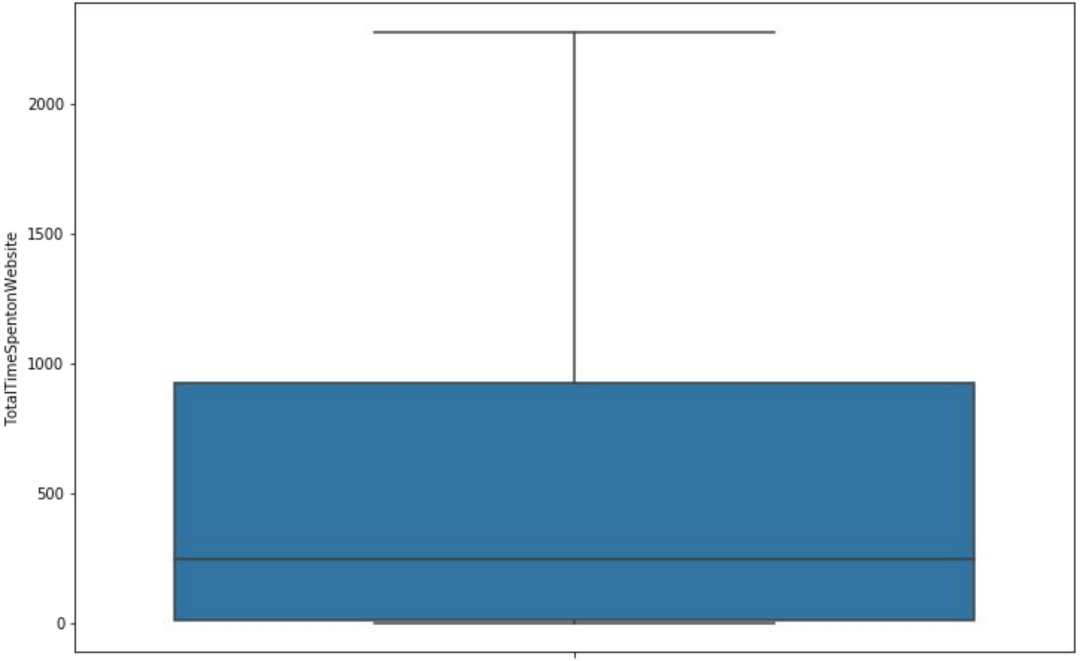
# Finding and removing outliers

| | TotalVisits | TotalTimeSpentonWebsite | PageViewsPerVisit |
|---|---|---|---|
| **count** | 9074.000000 | 9074.000000 | 9074.000000 |
| **mean** | 3.456028 | 482.887481 | 2.370151 |
| **std** | 4.858802 | 545.256560 | 2.160871 |
| **min** | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 1.000000 | 11.000000 | 1.000000 |
| **50%** | 3.000000 | 246.000000 | 2.000000 |
| **75%** | 5.000000 | 922.750000 | 3.200000 |
| **90%** | 7.000000 | 1373.000000 | 5.000000 |
| **95%** | 10.000000 | 1557.000000 | 6.000000 |
| **99%** | 17.000000 | 1839.000000 | 9.000000 |
| **max** | 251.000000 | 2272.000000 | 55.000000 |

Notice when everything has a uniform growth percentile, TotalVisits has 251 as Max from 17 at 99%. Thats a big jump, you can see that in the box plot below.

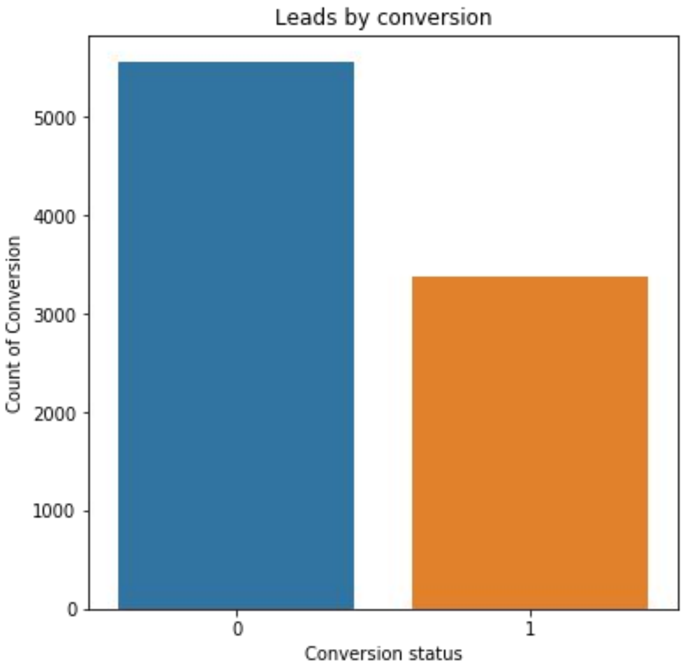


# No Outliers!




We removed that outlier. Now, no outliers!

# Checking after outlier treatment and after fixing missing values




Checking the converted and not converted leads in the data after the data cleaning and outlier removal

# Dummy Variables & Scaling

# Dummies

| | Do Not Email | Converted | TotalVisits | Total Time Spent on Website | Page Views Per Visit | A free copy of Mastering The Interview | Lead Origin_Landing Page Submission | Lead Origin_Lead Add Form | Lead Origin_Lead Import | Lead Source_Direct Traffic | ... | Last Notable Activity_Form Submitted on Website | Last Nota Activity_I a Phone Conversa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 0 | 0.0 | 0 | 0.0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 |
| **1** | 0 | 0 | 5.0 | 674 | 2.5 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 |
| **2** | 0 | 1 | 2.0 | 1532 | 2.0 | 1 | 1 | 0 | 0 | 1 | ... | 0 | 0 |
| **3** | 0 | 0 | 1.0 | 305 | 1.0 | 0 | 1 | 0 | 0 | 1 | ... | 0 | 0 |
| **4** | 0 | 1 | 2.0 | 1428 | 1.0 | 0 | 1 | 0 | 0 | 0 | ... | 0 | 0 |

# Standard Scaler

| | Do Not Email | TotalVisits | Total Time Spent on Website | Page Views Per Visit | A free copy of Mastering The Interview | Lead Origin_Landing Page Submission | Lead Origin_Lead Add Form | Lead Origin_Lead Import | Lead Source_Direct Traffic | Lead Source_Facebook | ... | Last No Activity Submitt Website |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **6676** | 0 | -0.049636 | 1.395668 | 0.395289 | 0 | 1 | 0 | 0 | 1 | 0 | ... | 0 |
| **6138** | 0 | 0.297929 | 0.609686 | 0.926758 | 0 | 1 | 0 | 0 | 0 | 0 | ... | 0 |
| **8650** | 0 | -0.049636 | 1.178657 | 0.395289 | 1 | 1 | 0 | 0 | 1 | 0 | ... | 0 |
| **3423** | 0 | -1.092332 | -0.878390 | -1.199117 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| **6552** | 0 | -1.092332 | -0.878390 | -1.199117 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |

# Model building

# Split Train and Test Data

```
# Selecting only the columns which are selected by RFE
col = X_train.columns[rfe.support_]
col
```

```
Index(['Do Not Email', 'TotalVisits', 'Total Time Spent on Website',
       'Page Views Per Visit', 'Lead Origin_Lead Add Form',
       'Lead Source_Direct Traffic', 'Lead Source_Google',
       'Lead Source_Olark Chat', 'Lead Source_Organic Search',
       'Lead Source_Referral Sites', 'Lead Source_Welingak Website',
       'Last Activity_Converted to Lead', 'Last Activity_Email Bounced',
       'Last Activity_Email Marked Spam',
       'Last Activity_Had a Phone Conversation',
       'Last Activity_Olark Chat Conversation',
       'Last Activity_Resubscribed to emails', 'Last Activity_SMS Sent',
       'Last Activity_Unsubscribed',
       'Last Activity_View in browser link Clicked',
       'What is your current occupation_Housewife',
       'What is your current occupation_Other',
       'What is your current occupation_Student',
       'What is your current occupation_Unemployed',
       'What is your current occupation_Working Professional',
       'Last Notable Activity_Had a Phone Conversation',
       'Last Notable Activity_Modified',
       'Last Notable Activity_Olark Chat Conversation',
       'Last Notable Activity_Page Visited on Website',
       'Last Notable Activity_Unreachable'],
      dtype='object')
```

# Split and First predictions

In [88]:
```
#Lets get the predicted values on the train set
y_train_pred = pd.DataFrame(res.predict(X_train_sm))
y_train_pred = y_train_pred.values.reshape(-1)
```

In [89]:
```
## Creating a dataframe with the actual Converted data and the predicted probabilities
y_train_pred_final = pd.DataFrame({'Converted':y_train.values, 'Convert_Prob':y_train_pred})
y_train_pred_final['LeadID'] = y_train.index

##### lets create a column 'predicted' assigning the value as 1 if prob of conversion is above 0.5 else as 0
y_train_pred_final['predicted'] = y_train_pred_final.Convert_Prob.map(lambda x: 1 if x > 0.5 else 0)
y_train_pred_final.head()
```

Out[89]:

| | Converted | Convert_Prob | LeadID | predicted |
|---|---|---|---|---|
| **0** | 1 | 0.551830 | 6676 | 1 |
| **1** | 1 | 0.734836 | 6138 | 1 |
| **2** | 1 | 0.920502 | 8650 | 1 |
| **3** | 0 | 0.031934 | 3423 | 0 |
| **4** | 0 | 0.144225 | 6552 | 0 |

# First Accuracy

```
#lets look at the accuracy of this model
acc = metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.predicted)
print(acc)
```

```
0.8224463656740314
```

We see that the accuracy of the model is 0.8222

Lets proceed with rmeoving the insignificant variables to model it further precisely

# Fine Tuning

```
: col = col.drop('Last Activity_Email Marked Spam')
  col

: Index(['Do Not Email', 'TotalVisits', 'Total Time Spent on Website',
         'Page Views Per Visit', 'Lead Origin_Lead Add Form',
         'Lead Source_Direct Traffic', 'Lead Source_Google',
         'Lead Source_Olark Chat', 'Lead Source_Organic Search',
         'Lead Source_Referral Sites', 'Lead Source_Welingak Website',
         'Last Activity_Converted to Lead', 'Last Activity_Email Bounced',
         'Last Activity_Had a Phone Conversation',
         'Last Activity_Olark Chat Conversation',
         'Last Activity_Resubscribed to emails', 'Last Activity_SMS Sent',
         'Last Activity_Unsubscribed',
         'Last Activity_View in browser link Clicked',
         'What is your current occupation_Housewife',
         'What is your current occupation_Other',
         'What is your current occupation_Student',
         'What is your current occupation_Unemployed',
         'What is your current occupation_Working Professional',
         'Last Notable Activity_Had a Phone Conversation',
         'Last Notable Activity_Modified',
         'Last Notable Activity_Olark Chat Conversation',
         'Last Notable Activity_Page Visited on Website',
         'Last Notable Activity_Unreachable'],
        dtype='object')
```

# ViF and Final Accuracy

```
vif
```

Out[166]:

| | Features | VIF |
|---|---|---|
| 5 | Lead Source_Google | 1.57 |
| 12 | Last Activity_SMS Sent | 1.53 |
| 1 | TotalVisits | 1.52 |
| 3 | Lead Origin_Lead Add Form | 1.52 |
| 4 | Lead Source_Direct Traffic | 1.47 |
| 13 | What is your current occupation_Other | 1.47 |
| 6 | Lead Source_Organic Search | 1.37 |
| 2 | Total Time Spent on Website | 1.29 |
| 8 | Lead Source_Welingak Website | 1.29 |
| 11 | Last Activity_Olark Chat Conversation | 1.23 |
| 9 | Last Activity_Converted to Lead | 1.17 |
| 14 | What is your current occupation_Working Profes... | 1.16 |
| 0 | Do Not Email | 1.13 |
| 7 | Lead Source_Referral Sites | 1.03 |
| 10 | Last Activity_Had a Phone Conversation | 1.01 |
| 15 | Last Notable Activity_Unreachable | 1.01 |

In [167]:

```
#lets look at the accuracy of this model
acc = metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.predicted)
print(acc)
```

```
0.8179634966378482
```

Now we can clearly observer that the p value and VIF value both are within the limits

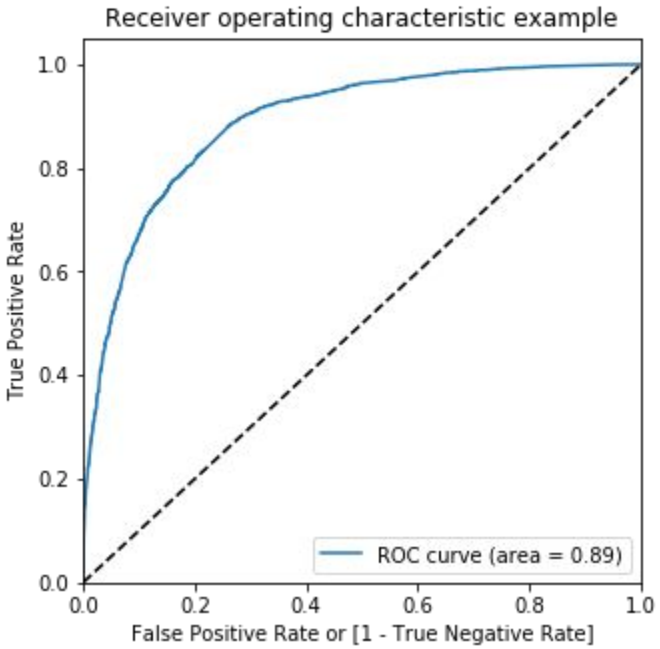
And the accuracy is also well above 0.80

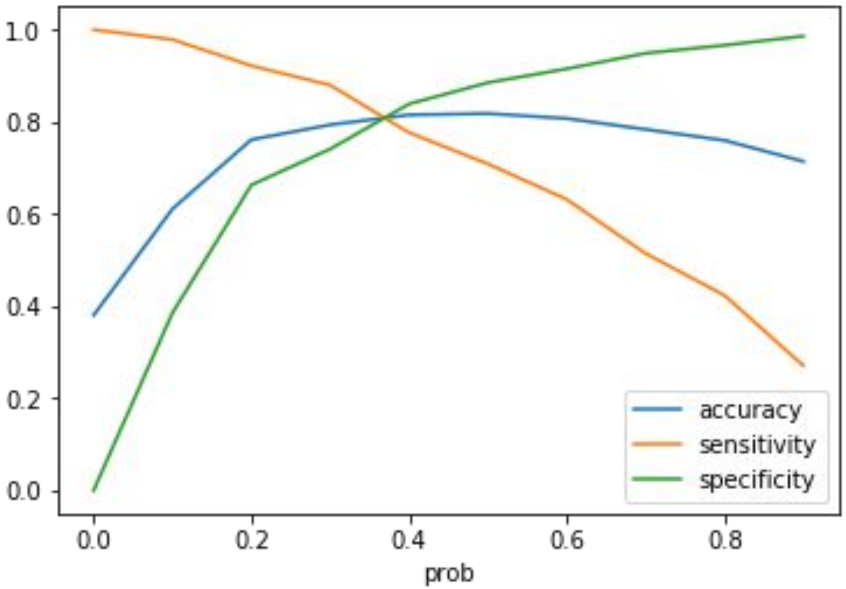
Lets now proceed with testing it on the test data

# 81%

Accuracy

# Roc




From the above graph, we see that at 0.36 all the features are achievable

However once we check the Rate we would be to finalize on it

# Checking other parameters of accuracy

In [189]:
```
from sklearn.metrics import confusion_matrix
#True negative
TN = confusionM_2[0,0]
#False positives
FP = confusionM_2[0,1]
#False negatives
FN = confusionM_2[1,0]
#True Positive
TP = confusionM_2[1,1]
```

In [190]:
```
# Let's check the sensitivity and specificity of our logistic regression model
print("Sensitivity=",(TP / (TP+FN)))

print("Specificity=",(TN / (TN+FP)))
```

```
Sensitivity= 0.7987368421052632
Specificity= 0.8155515370705244
```

In [191]:
```
# Calculate false postive rate - which says how much is showing as converted, when actually not converted
print("false postive rate =",(FP/ (TN+FP)))
```

```
false postive rate = 0.1844484629294756
```

In [192]:
```
# Positive predictive rate
print("Positive predictive rate =",(TP / (TP+FP)))


# Negative predictive rate
print("Negative predictive rate =",(TN / float(TN+ FN)))
```

```
Positive predictive rate = 0.7265415549597856
Negative predictive rate = 0.8685006877579092
```

As we can see that the positive prediction rate is only 0.72

lets stick on to cut off value at 0.5 as we even had the accuracy greater than this

# Hot Leads Final

# Hot Leads predicted along with Lead Id

```
y_pred_final['final_predicted'] = y_pred_final.Convert_Prob.map(lambda x: 1 if x > 0.8 else 0)
y_pred_final.head()
```

| | LeadID | Converted | Convert_Prob | final_predicted | Lead_Score |
|---|---|---|---|---|---|
| 0 | 7625 | 0 | 0.704232 | 0 | 70.42 |
| 1 | 5207 | 1 | 0.376488 | 0 | 37.65 |
| 2 | 2390 | 1 | 0.945249 | 1 | 94.52 |
| 3 | 4362 | 0 | 0.438270 | 0 | 43.83 |
| 4 | 1023 | 0 | 0.276035 | 0 | 27.60 |

Refer to the notebook for lead details and start calling!
All the best!

# Thank you

Shivaprasad Meti and Madhusudhan Anand