Lead Scoring Case Study Logistics Regression

- Shivram J & Pappu Kapgate

Problem Statement

An education company named X Education sells online courses to industry professionals.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. The typical lead conversion rate at X education is around 30%.

X Education needs help in selecting the most promising leads. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

The recommendation posted will provide insights into the various factors that affect the population and the basis of selection of the leads to be pursued.

Implementation Approach

The approach for this analysis was broken down into multiple steps namely.

Step 1 Understanding the data

- Missing values, Duplicate Values, Single Valued columns

Step 2 EDA and Analysis

- Univariate
- Bivariate

Step 3 Data Preparations for modelling

- Handling Outliers
- Dummy variable creation
- Scaling, Data splitting

Step 4 Model Building

- Performing RFE
- Logistic Regression
- Evaluating Model –
  - ROC Curve, Accuracy, Specificity, Sensitivity
  - Precision & Recall
- Making Predictions using Test Sets

Understanding the data & Cleaning

The dataset consists of lead information about Lead source, customer profiles, customer responses and activity information ..

Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	...	Get updates on DM Content	Lead Profile
7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0	0	0.0	...	No	Select
2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5	...	No	Select
8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	...	No	Potential Lead
7cf4-4e39-9de9-19797f9b38cc	660719	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	...	No	Select
3256f628-4-4826-9d63-4a8b88782852	660681	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	...	No	Select

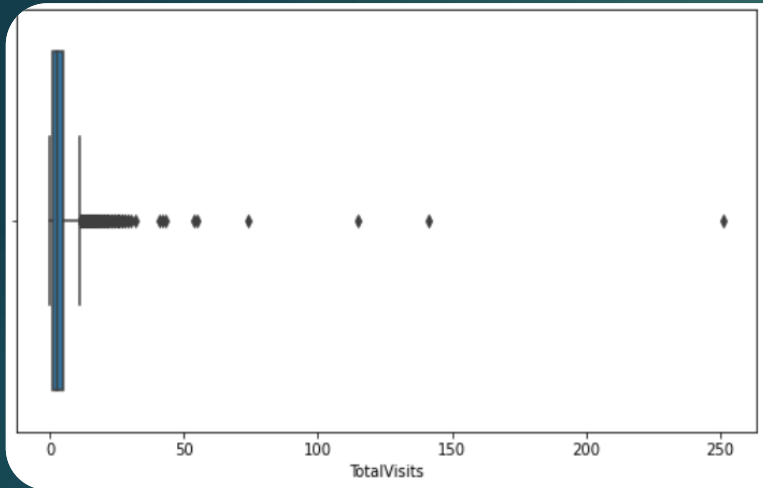
```
Magazine                                    1
Receive More Updates About Our Courses      1
Update me on Supply Chain Content           1
Get updates on DM Content                   1
I agree to pay the amount through cheque    1
dtype: int64
```

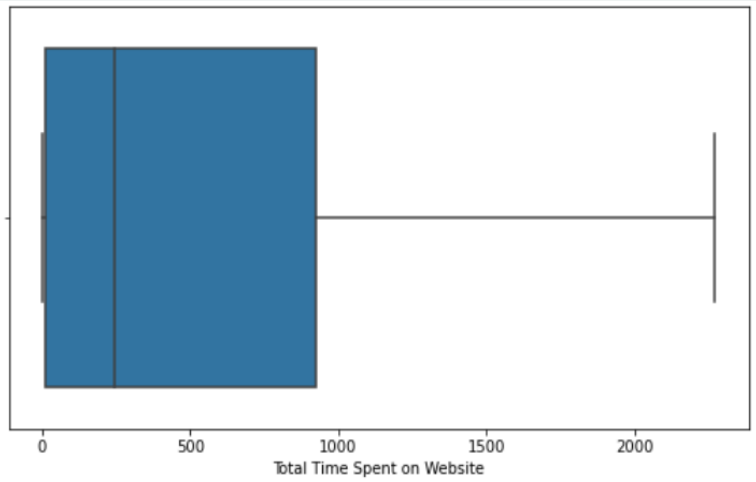
Dropping all columns identified above which have only 1 value in it.

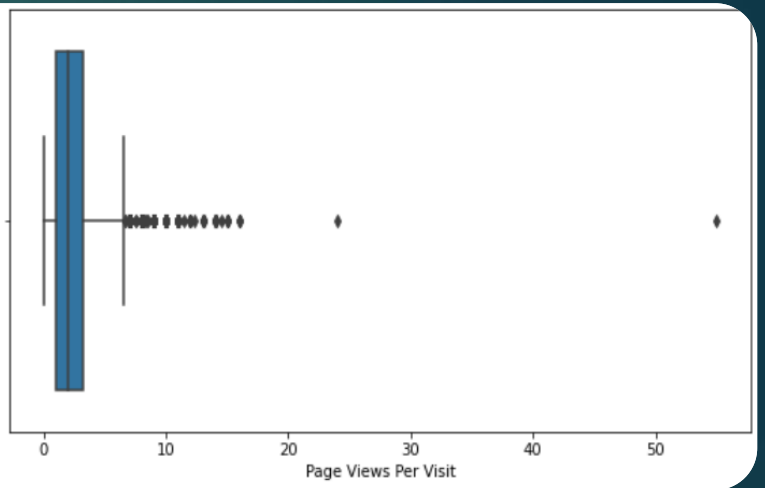
```
List of Null Value columns with % of Null values
Lead Source                                    0.389610
TotalVisits                                    1.482684
Page Views Per Visit                           1.482684
Last Activity                                  1.114719
Country                                       26.634199
Specialization                                36.580087
How did you hear about X Education            78.463203
What is your current occupation               29.112554
What matters most to you in choosing a course 29.318182
Tags                                          36.287879
Lead Quality                                  51.590909
Lead Profile                                  74.188312
City                                          39.707792
Asymmetrique Activity Index                   45.649351
Asymmetrique Profile Index                    45.649351
Asymmetrique Activity Score                   45.649351
Asymmetrique Profile Score                    45.649351
```

.. And there were missing and skewed values in the dataset which had to be handled.

Handling outliers

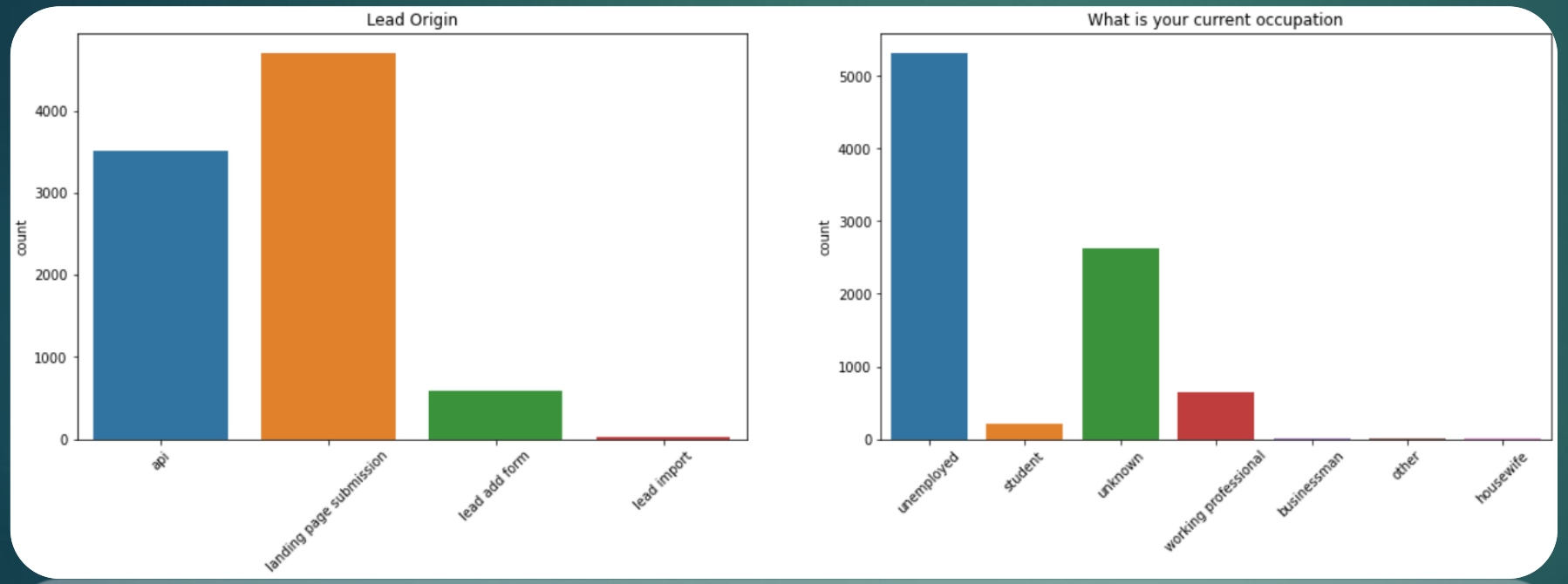






Retaining 99% quantiles of data across all the columns and removed the outliers from the analysis.

EDA – Data Exploration & Analysis



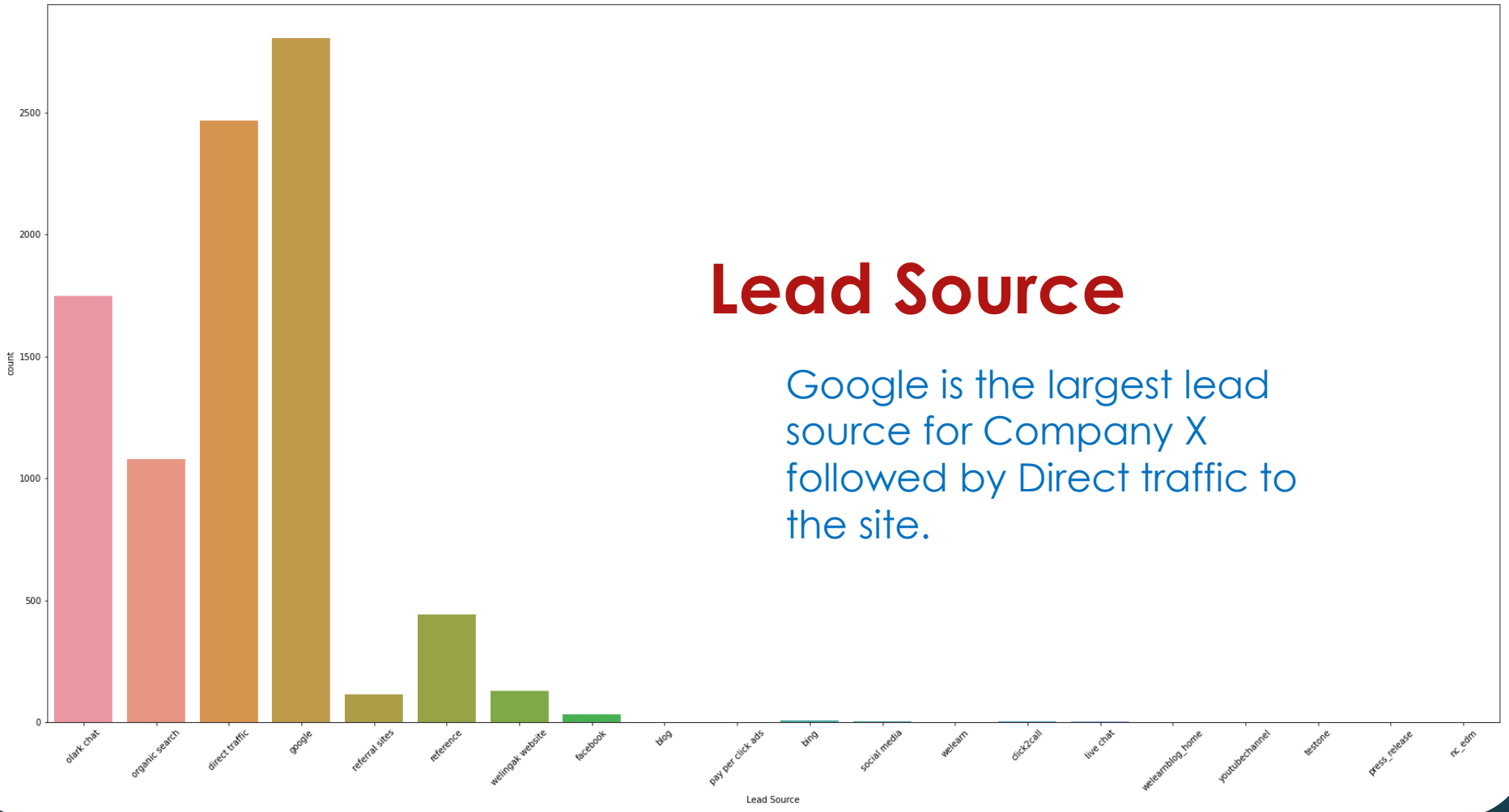
Current Occupation

People who are unemployed contribute to the highest leads for Company X.

Lead Origin

Most number of leads originate from Landing Page followed by APIs.

EDA – Data Exploration & Analysis



Lead Source

Google is the largest lead source for Company X followed by Direct traffic to the site.

EDA – Data Exploration & Analysis

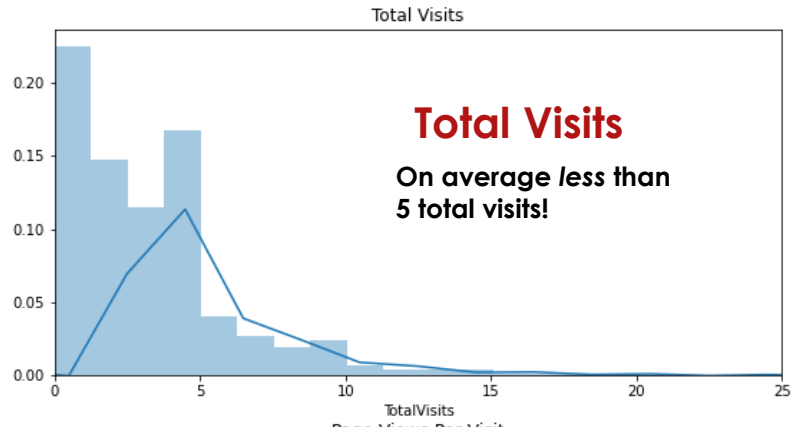
Unknown ? Lot of the leads do not capture this information. In future this should be made as a mandatory field.

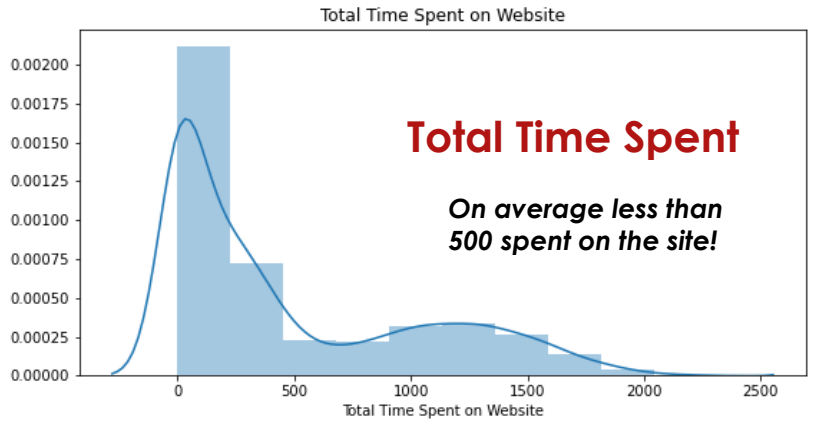
We felt this field may be important for business. Hence, we retained it!

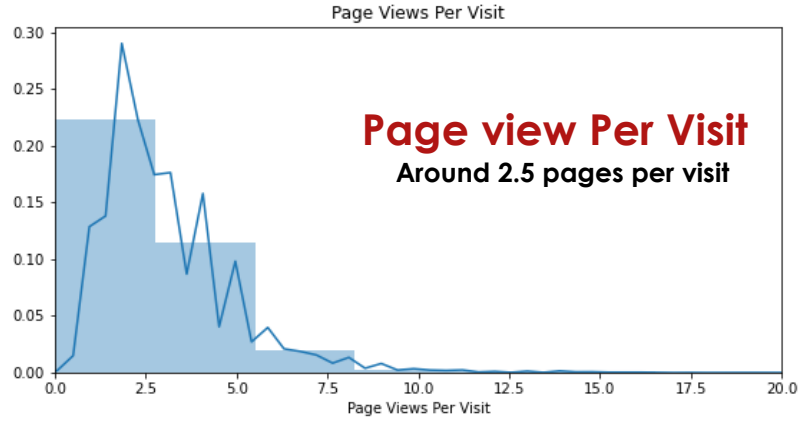
Specialization

Most number of leads are for Financial Management specialization

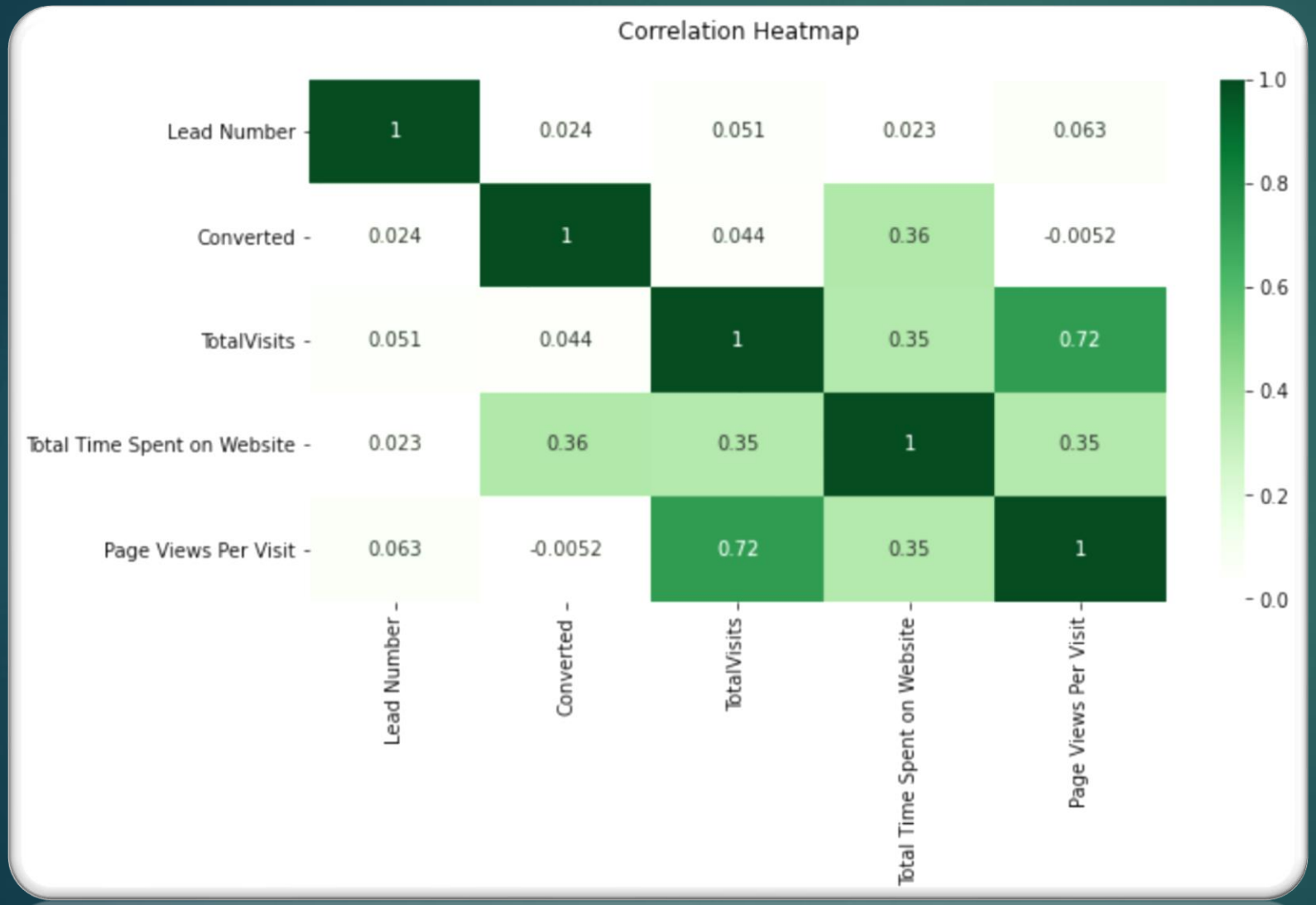
EDA – Data Exploration & Analysis







EDA –BI Variate Analysis



Highly correlated!

Total Visits &
Page view Per Visit

Total Time Spent &
Page view Per Visit

Total Time Spent &
Total Visits

Modelling – Data preparation

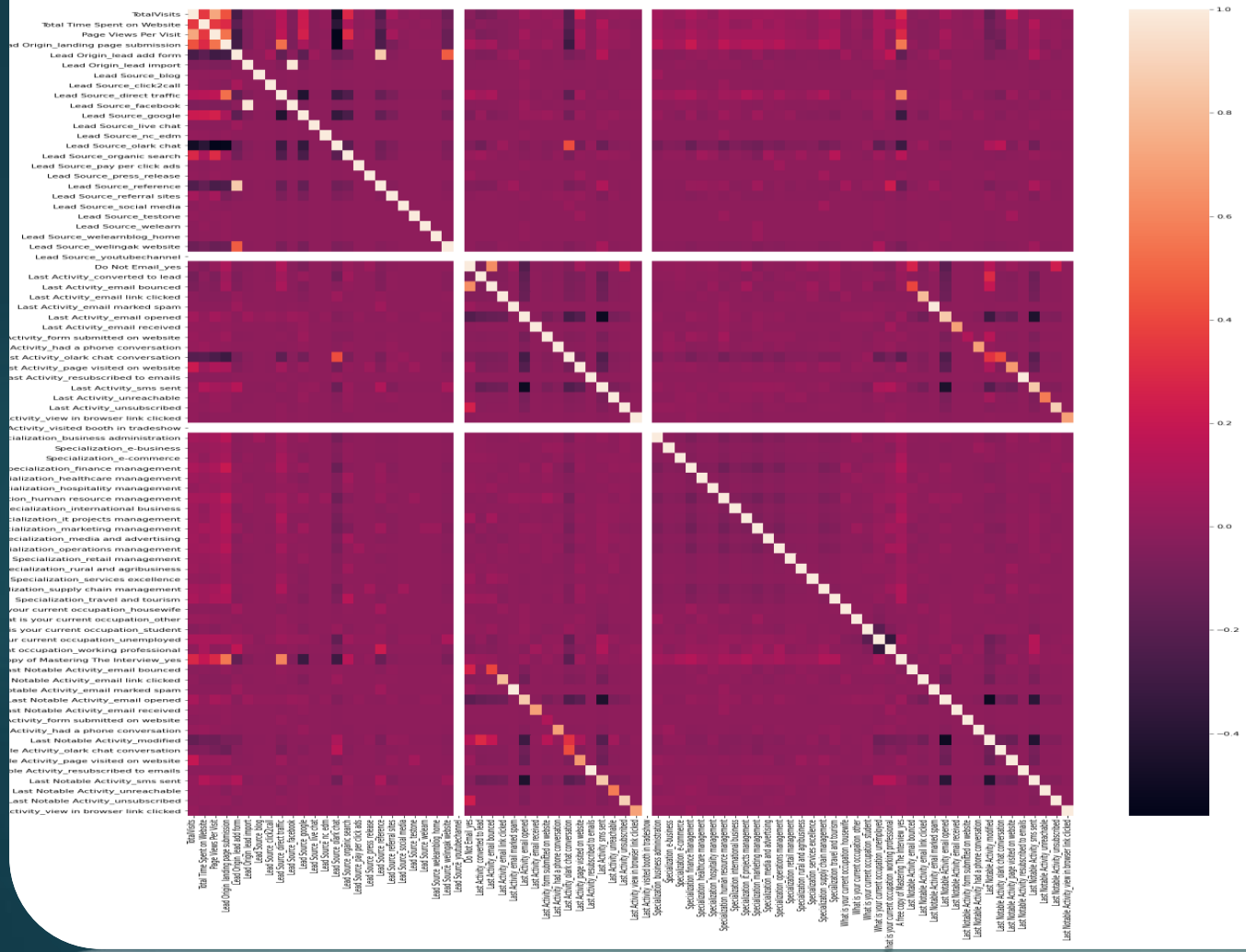
Dummy Variable!

- Lead Origin
- Lead Source
- Do Not Email
- Last Activity
- Specialization
- What is your current occupation
- A free copy of Mastering The Interview
- Last Notable Activity

Standard Scaling!

- Total Visits
- Page Views Per Visit
- Total Time Spent on Website

We have selected the above columns for creating dummy variables and then scaling the numeric columns.



Modelling -

Checking correlation

Modelling – Recursive Feature Elimination (RFE)

Selected 15 variables!

We have selected the following columns after performing RFE.

We wanted the variables which are most relevant in predicting the target variable.

1. TotalVisits
2. Total Time Spent on Website
3. Lead Origin_lead add form
4. Lead Source_direct traffic
5. Lead Source_google
6. Lead Source_organic search
7. Lead Source_referral sites
8. Lead Source_welingak website
9. Do Not Email_yes
10. Last Activity_had a phone conversation
11. Last Activity_sms sent
12. What is your current occupation_housewife
13. What is your current occupation_working professional
14. Last Notable Activity_olark chat conversation
15. Last Notable Activity_unreachable

Modelling – Iteration 1 (1st Model)

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6183
Model:                            GLM   Df Residuals:                     6167
Model Family:                Binomial   Df Model:                           15
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2687.1
Date:                Sat, 06 Feb 2021   Deviance:                       5374.1
Time:                        21:26:24   Pearson chi2:                 6.23e+03
No. Iterations:                     21
Covariance Type:            nonrobust
==========================================================================================================
                                                      coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
const                                            -1.6027      0.077    -20.946      0.000      -1.753      -1.453
TotalVisits                                       1.5622      0.248      6.311      0.000       1.077       2.047
Total Time Spent on Website                       3.8290      0.139     27.616      0.000       3.557       4.101
Lead Origin_lead add form                         2.7980      0.217     12.912      0.000       2.373       3.223
Lead Source_direct traffic                       -1.4971      0.123    -12.147      0.000      -1.739      -1.256
Lead Source_google                               -1.2003      0.119    -10.084      0.000      -1.434      -0.967
Lead Source_organic search                       -1.4438      0.150     -9.627      0.000      -1.738      -1.150
Lead Source_referral sites                       -1.5161      0.353     -4.291      0.000      -2.209      -0.824
Lead Source_welingak website                      1.9542      0.748      2.614      0.009       0.489       3.420
Do Not Email_yes                                 -1.4212      0.171     -8.321      0.000      -1.756      -1.086
Last Activity_had a phone conversation            1.8766      0.636      2.953      0.003       0.631       3.122
Last Activity_sms sent                            1.4912      0.074     20.246      0.000       1.347       1.636
What is your current occupation_housewife        23.3804   1.77e+04      0.001      0.999   -3.47e+04    3.47e+04
What is your current occupation_working professional  2.8212      0.187     15.067      0.000       2.454       3.188
Last Notable Activity_olark chat conversation    -1.4137      0.360     -3.932      0.000      -2.118      -0.709
Last Notable Activity_unreachable                 1.4369      0.535      2.686      0.007       0.388       2.486
==========================================================================================================
                                          Features   VIF
0                                      TotalVisits  3.49
4                                Lead Source_google  2.54
1                        Total Time Spent on Website  2.36
3                         Lead Source_direct traffic  2.24
5                         Lead Source_organic search  1.88
2                          Lead Origin_lead add form  1.48
10                            Last Activity_sms sent  1.47
7                        Lead Source_welingak website  1.32
12  What is your current occupation_working profes...  1.16
8                                     Do Not Email_yes  1.11
6                         Lead Source_referral sites  1.07
9              Last Activity_had a phone conversation  1.01
11          What is your current occupation_housewife  1.01
13      Last Notable Activity_olark chat conversation  1.01
14                  Last Notable Activity_unreachable  1.01
```

Observation!

Occupation_housewife has a high p-value

Total Visits has a high VIF

.. We decide to remove Occupation_housewife and rebuilt the model

Modelling – Iteration 2 (2nd Model)

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6183
Model:                            GLM   Df Residuals:                     6168
Model Family:                Binomial   Df Model:                           14
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2690.9
Date:                Sat, 06 Feb 2021   Deviance:                       5381.9
Time:                        21:26:24   Pearson chi2:                 6.24e+03
No. Iterations:                     7
Covariance Type:            nonrobust
==========================================================================================================
                                                      coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
const                                               -1.6010      0.076    -20.934      0.000      -1.751      -1.451
TotalVisits                                          1.5536      0.247      6.280      0.000       1.069       2.039
Total Time Spent on Website                          3.8242      0.138     27.615      0.000       3.553       4.096
Lead Origin_lead add form                            2.8248      0.216     13.062      0.000       2.401       3.249
Lead Source_direct traffic                          -1.4931      0.123    -12.124      0.000      -1.734      -1.252
Lead Source_google                                  -1.1954      0.119    -10.051      0.000      -1.429      -0.962
Lead Source_organic search                          -1.4305      0.150     -9.554      0.000      -1.724      -1.137
Lead Source_referral sites                          -1.5125      0.353     -4.283      0.000      -2.205      -0.820
Lead Source_welingak website                         1.9277      0.748      2.578      0.010       0.462       3.393
Do Not Email_yes                                    -1.4235      0.171     -8.334      0.000      -1.758      -1.089
Last Activity_had a phone conversation               1.8718      0.636      2.945      0.003       0.626       3.118
Last Activity_sms sent                               1.4872      0.074     20.202      0.000       1.343       1.631
What is your current occupation_working professional 2.8179      0.187     15.052      0.000       2.451       3.185
Last Notable Activity_olark chat conversation       -1.4148      0.359     -3.936      0.000      -2.119      -0.710
Last Notable Activity_unreachable                    1.4321      0.535      2.676      0.007       0.383       2.481
==========================================================================================================
                                          Features   VIF
0                                       TotalVisits  3.49
4                                 Lead Source_google  2.54
1                           Total Time Spent on Website  2.36
3                            Lead Source_direct traffic  2.24
5                            Lead Source_organic search  1.87
2                             Lead Origin_lead add form  1.47
10                               Last Activity_sms sent  1.47
7                          Lead Source_welingak website  1.32
11  What is your current occupation_working profes...  1.16
8                                     Do Not Email_yes  1.11
6                             Lead Source_referral sites  1.07
9                 Last Activity_had a phone conversation  1.01
12        Last Notable Activity_olark chat conversation  1.01
13                    Last Notable Activity_unreachable  1.01
```

Observation!

Total Visits still has a high VIF

.. We removed Total Visits and rebuilt the model

Modelling – Iteration 3 (Final Model)

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6183
Model:                            GLM   Df Residuals:                     6169
Model Family:                Binomial   Df Model:                           13
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2710.7
Date:                Sat, 06 Feb 2021   Deviance:                       5421.5
Time:                        21:26:24   Pearson chi2:                 6.26e+03
No. Iterations:                     7
Covariance Type:            nonrobust
==============================================================================================================
                                                        coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------------
const                                                  -1.5601      0.075    -20.671      0.000      -1.708      -1.412
Total Time Spent on Website                             3.8775      0.138     28.098      0.000       3.607       4.148
Lead Origin_lead add form                               2.7917      0.216     12.945      0.000       2.369       3.214
Lead Source_direct traffic                             -1.1842      0.110    -10.722      0.000      -1.401      -0.968
Lead Source_google                                     -0.8715      0.105     -8.312      0.000      -1.077      -0.666
Lead Source_organic search                             -0.9885      0.130     -7.632      0.000      -1.242      -0.735
Lead Source_referral sites                             -1.0961      0.347     -3.162      0.002      -1.775      -0.417
Lead Source_welingak website                            1.9353      0.747      2.589      0.010       0.470       3.400
Do Not Email_yes                                       -1.4522      0.170     -8.547      0.000      -1.785      -1.119
Last Activity_had a phone conversation                  1.9276      0.641      3.007      0.003       0.671       3.184
Last Activity_sms sent                                  1.4687      0.073     20.038      0.000       1.325       1.612
What is your current occupation_working professional    2.8128      0.186     15.095      0.000       2.448       3.178
Last Notable Activity_olark chat conversation          -1.3440      0.352     -3.824      0.000      -2.033      -0.655
Last Notable Activity_unreachable                       1.4987      0.545      2.751      0.006       0.431       2.566
==============================================================================================================
                                          Features   VIF
0                        Total Time Spent on Website  2.33
3                                Lead Source_google  1.70
2                         Lead Source_direct traffic  1.61
1                          Lead Origin_lead add form  1.47
9                              Last Activity_sms sent  1.47
6                       Lead Source_welingak website  1.32
4                         Lead Source_organic search  1.28
10  What is your current occupation_working profes...  1.16
7                                    Do Not Email_yes  1.10
5                         Lead Source_referral sites  1.02
8              Last Activity_had a phone conversation  1.01
11      Last Notable Activity_olark chat conversation  1.01
12                  Last Notable Activity_unreachable  1.01
```

Observation!

VIF and P value look good from a statistical perspective.

.. And relevant business variables are also present in this model.

Model – Evaluation

```
# Creating confusion matrix
confusion = metrics.confusion_matrix(y_train_pred_final.Converted, y_train_pred_final.Predicted )
confusion
```

```
array([[3759,  105],
       [1488,  831]], dtype=int64)
```

Confusion Matrix !

Accuracy!

```
# Check the overall accuracy
metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.Predicted)
```

```
0.74235807860262
```

```
# Calculating the specificity
TN/(TN+FP)
```

```
0.9728260869565217
```

Specificity!

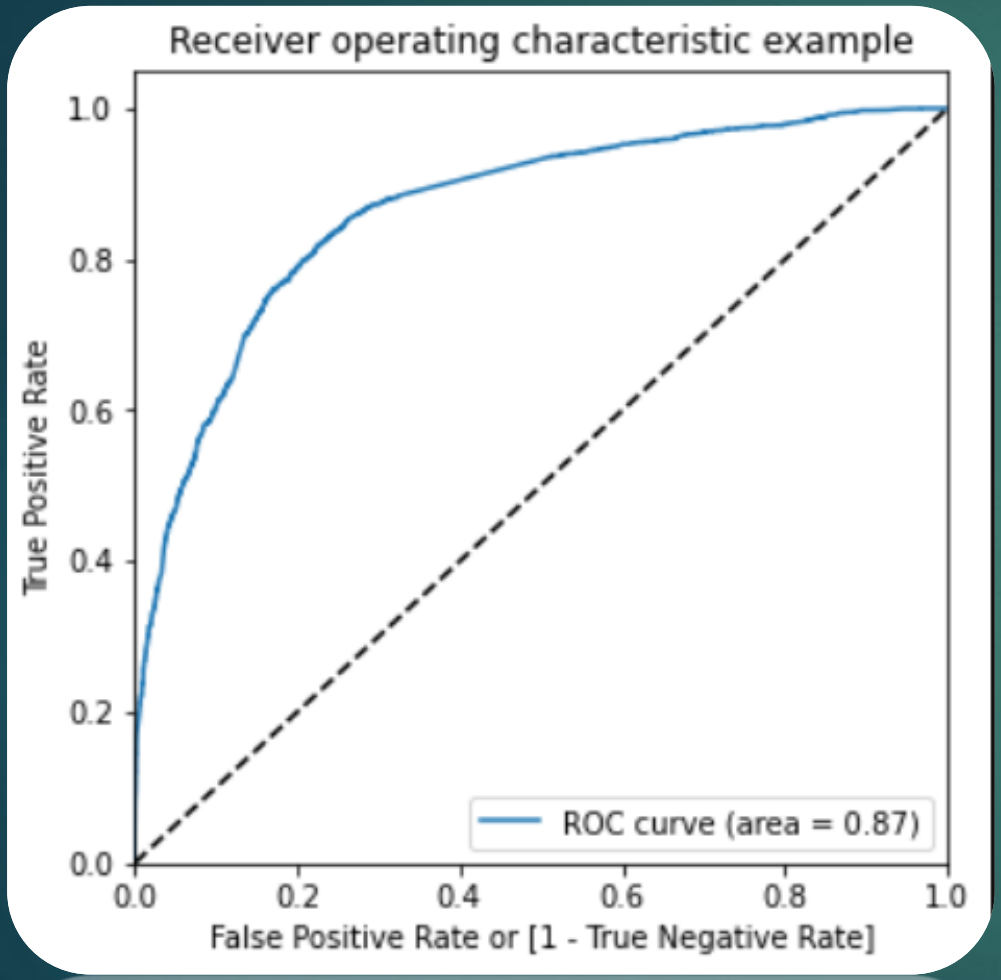
```
# Calculating the sensitivity
TP/(TP+FN)
```

```
0.35834411384217335
```

Sensitivity!

.. We used the mentioned metrics to evaluate the effectiveness of the model…

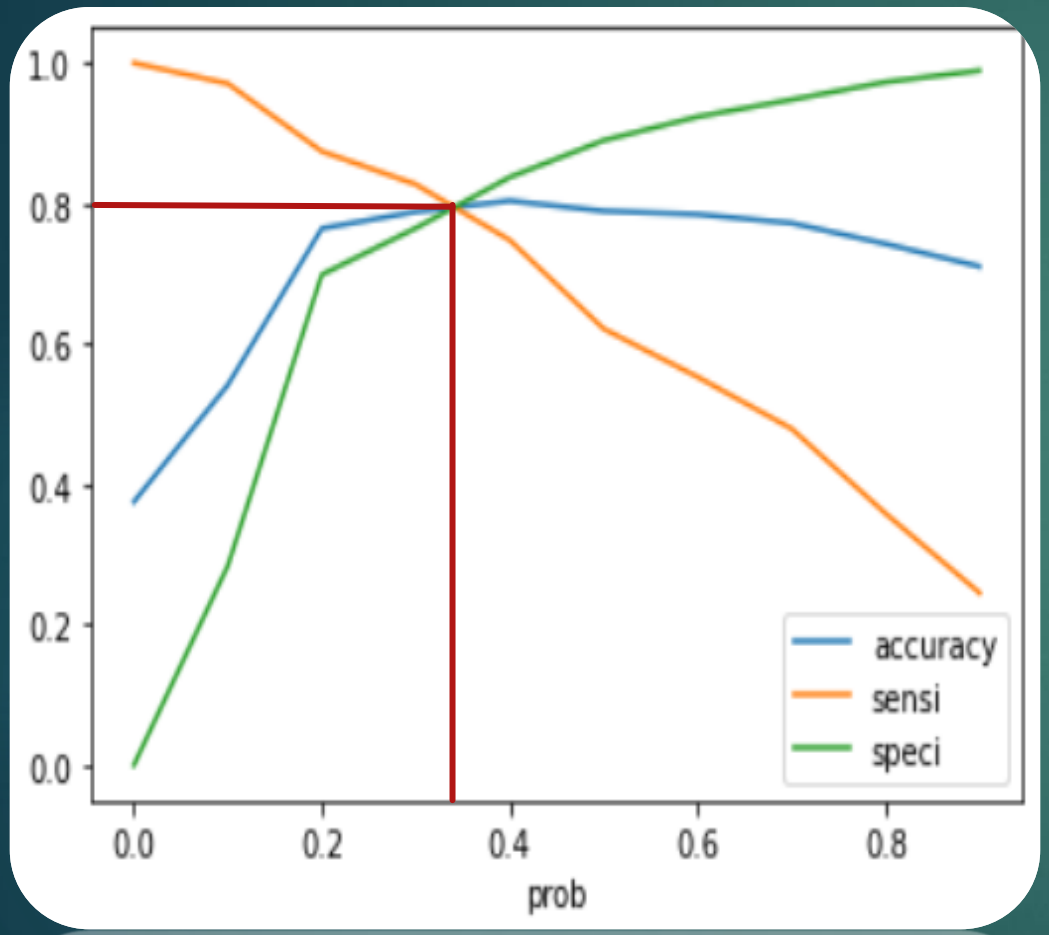
Model – Optimize Cut off (ROC Curve)



The model looks decent from the ROC curve as we see the tradeoff between sensitivity vs specificity.

Area under the curve =0.87 which is a good indication of the build model's effectiveness.

Model – Optimal Threshold



Plotted the different probability [.1-.9] against accuracy, sensitivity and specificity and got the cut off at 0.35

Model – Precision & Recall

```
# Precision = TP / TP + FP
confusion[1,1]/(confusion[0,1]+confusion[1,1])
```

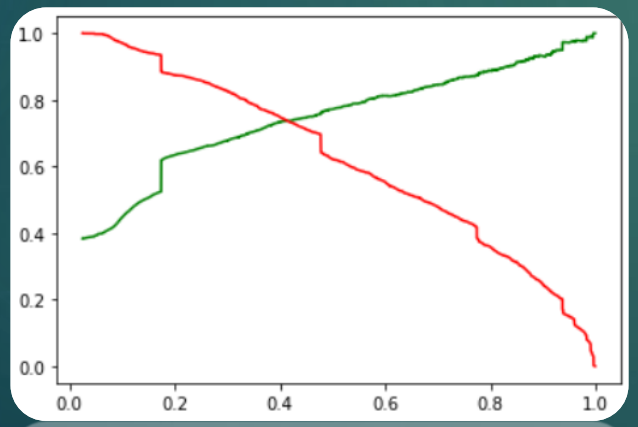
```
0.8878205128205128
```

```
#Recall = TP / TP + FN
confusion[1,1]/(confusion[1,0]+confusion[1,1])
```

```
0.35834411384217335
```

Precision in our case would be the probability of predicting a hot lead which is an actual hot lead!

88%



35%

Recall is the probability of identifying a hot lead correctly !

..We will not be using this for cut off as recall is only 35%

Prediction – On Test Set

```
#Creating confusion Matrix
confusion2 = metrics.confusion_matrix(y_pred_final.Converted, y_pred_final.final_predicted )
confusion2
```

```
array([[1366,  295],
       [ 193,  797]], dtype=int64)
```

Confusion Matrix !

81% Accuracy!

```
# Let's check the overall accuracy.
metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_predicted)
```

```
0.8159185213127121
```

```
# Let us calculate specificity
TN / float(TN+FP)
```

```
0.8372153209109731
```

83% Specificity!

```
# Let's see the sensitivity
TP / float(TP+FN)
```

```
0.7477360931435963
```

74% Sensitivity!

.. We used the mentioned metrics to evaluate the effectiveness of the model…

X Education company must focus mostly on the following variables to achieve the 80% lead conversion rate toward their company. These variables have a high potential to understand the customer profile and whether that customer will potentially buy the courses from the company.

Conclusion

1. **Total time Spent on website.**
2. **Lead Origin : Add format**
3. **Lead Source : [Direct Traffic, Google, Organic Search, Referral Sites , Welingak Website]**
4. **Last notable activity: [Olark Chat Conversation ,Unreachable]**
5. **Do not send Email**
6. **When the Last Activity: [Had a Phone Conversation, SMS Sent]**
7. **When the customer current occupation is working professional**