Lead Scoring Case Study

By- Srinivasaragavan V

Vishal Yadav

Problem Objective

Problem Statement

- An education company named **X Education** sells online courses to industry professionals. 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. The typical lead conversion rate at X education is around 30%.

Goal:

- 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.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Problem Solving Methodology

Data Cleaning and Preparation

Identify the data quality and clean the data on necessity basis.

Handle Select and null values based on converted rate without blindly removing those data points.

Applied Outlier Treatment

Apply Logistic Regression

Split data into train and test. Apply Standard Scaling on data.

Apply Logistic Regression using GLM Model on train data and apply RFE.

Find Variance Inflation Factor and P-values from Successful Model

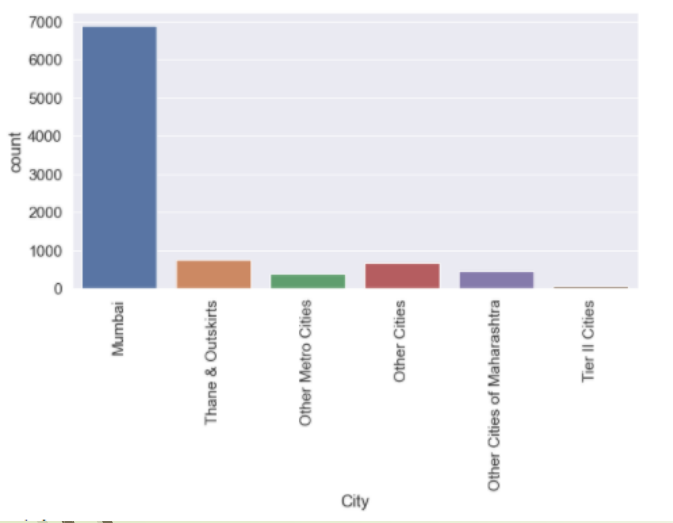
Identifying Influencing features from Model

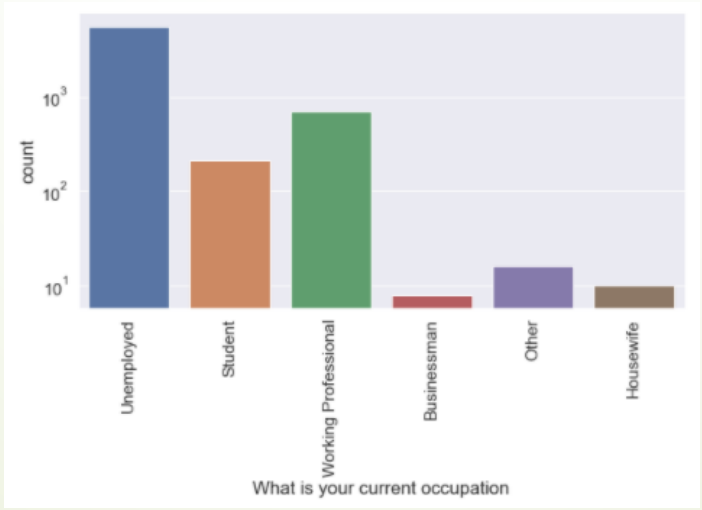
Based on Successful Logistic Regression model, identify influencing variables.

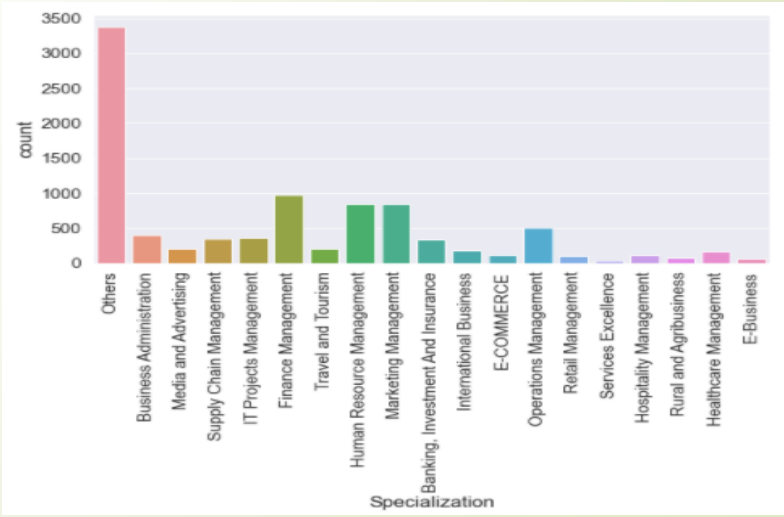
Draw conclusions and Recommendations from model.

Data Cleaning:

- Checked for duplicates values in row.
- Dropping unnecessary columns with only null values, single unique feature, rating columns created by sales team after contacting with customers.
- Conversion of Select to NAN in data frame.
- Dropping attributes high % NA values.
- Imputed values with highest count in particular columns
- Segregated all NA values into others as separate entity.
- Highly skewed columns were dropped.







Exploratory Data Analysis- Categorical variable

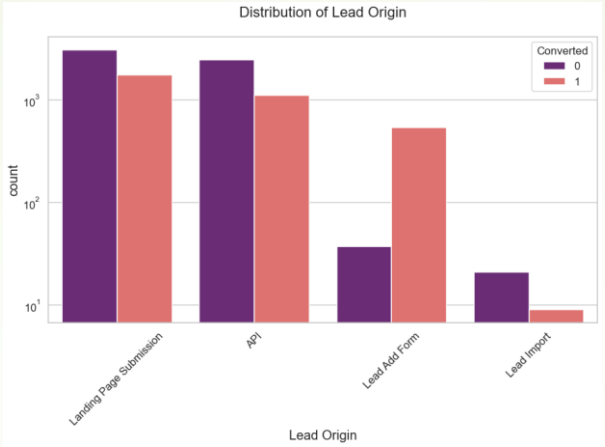
Following are observations :

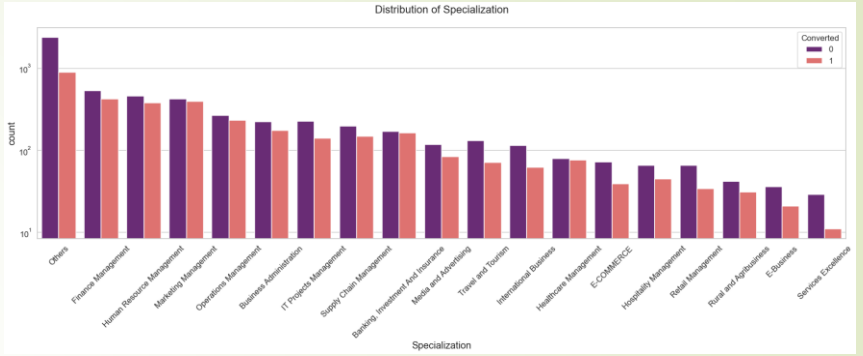
Distribution Of Lead Origin
Landing page submission is comparatively high than the rest of the categories
Lead add form has high certainty in lead conversion

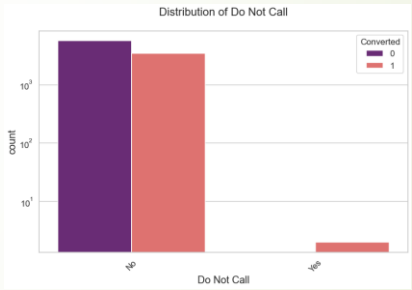
Distribution of Lead Source
Google is the best lead source among all other categories in the lead source
Direct Traffic, Olark Chat and Organic Search are some of the best entities in lead source
The best category for lead conversion is Reference and Welingak Website among all the lead source categories

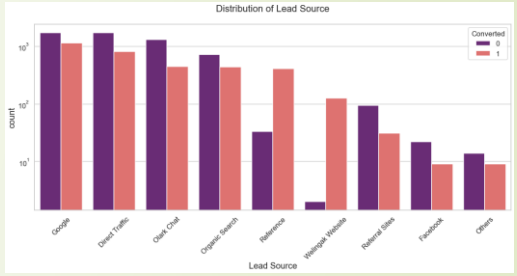
Distribution of Occupation
Working Professionals going for the course have high chances of joining it.

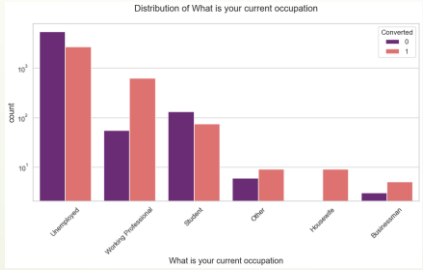
Distribution Of City
Most leads are from Mumbai with around 30% conversion rate.

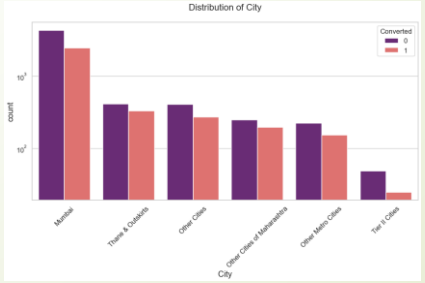


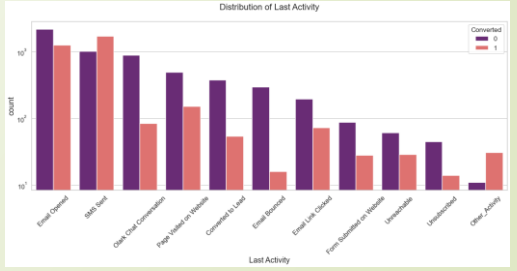












Exploratory Data Analysis- Numerical variable

Following are observations :

Total Visits:

Median for converted and not converted leads are the same.

Nothing conclusive on the basis of Total Visits.

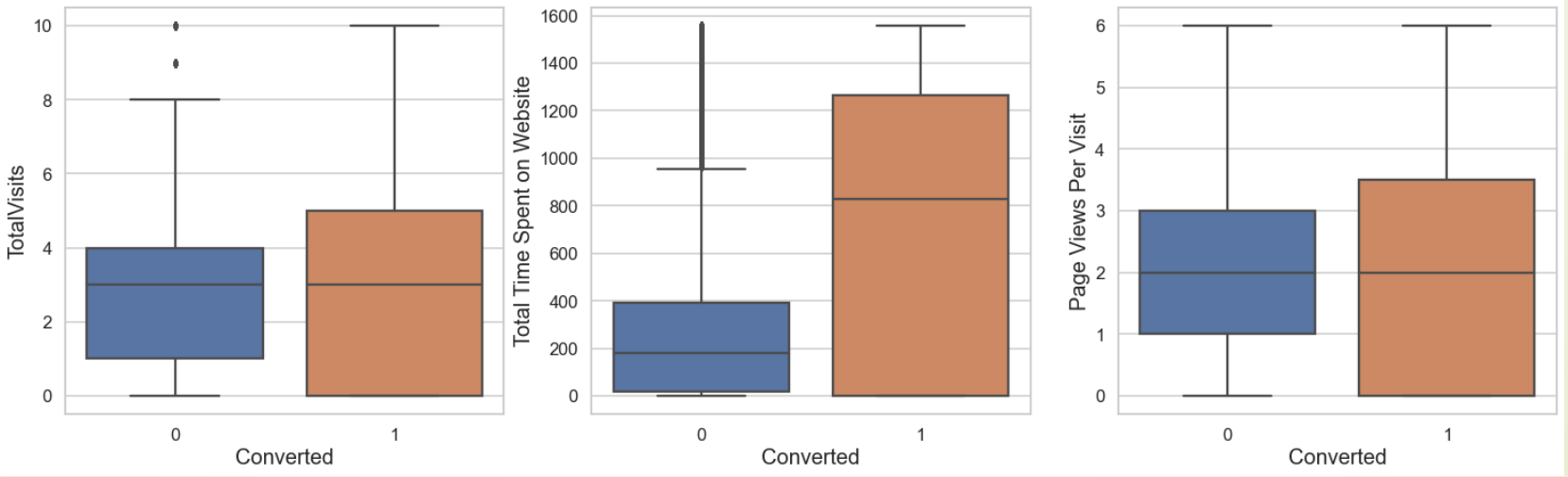
Total Time Spent on Website

Leads spending more time on the website are more likely to be converted.

Website should be made more engaging to make leads spend more time.

Page Views Per Visit

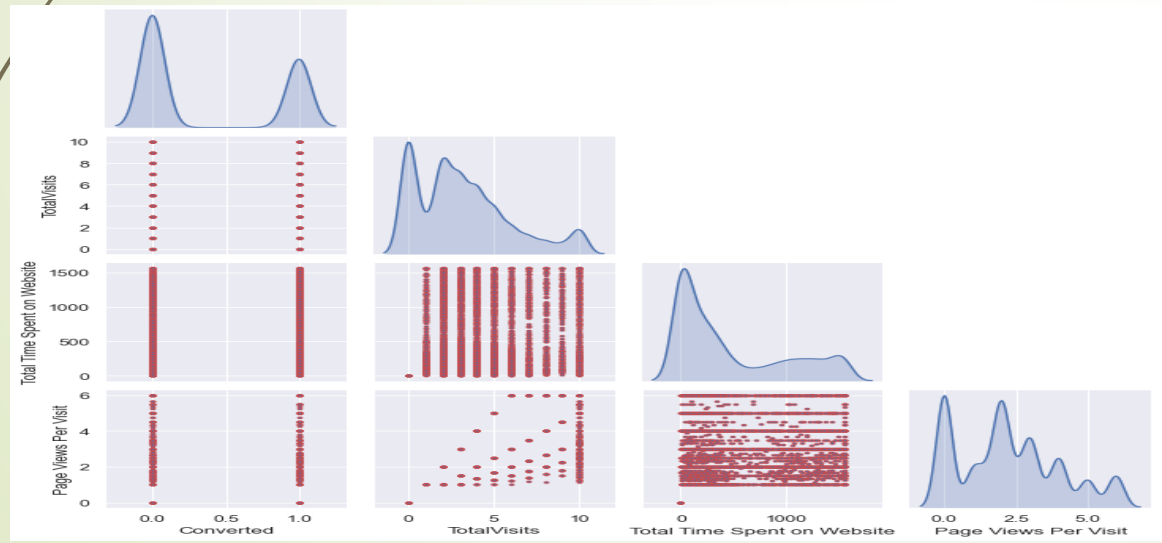
Nothing can be said specifically for lead conversion from Page Views Per Visit

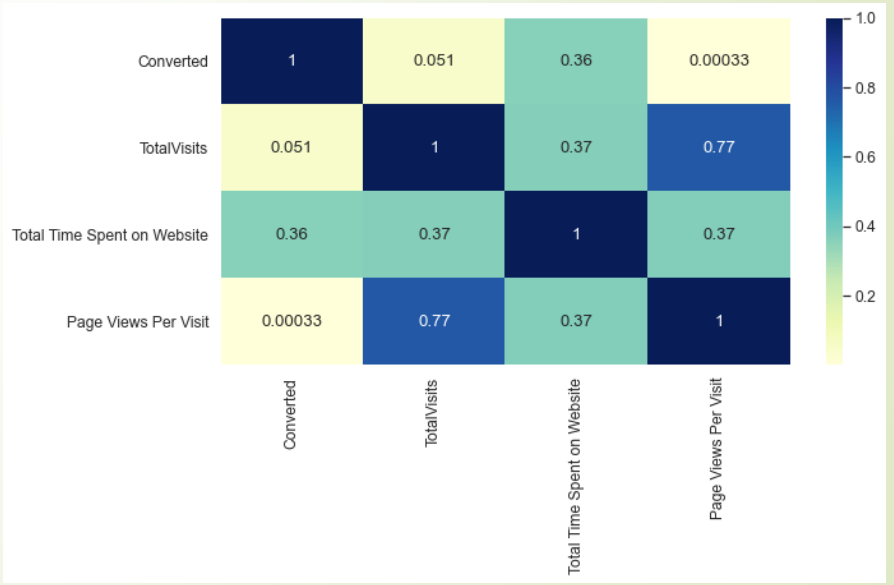


Exploratory Data Analysis- Bi-variate Analysis

Following are observations :
- From pair plot we can observe clearly that our dataset has highly skewed values with lot of random peaks.

With heatmap we draw inference like:
- Total Visits and Page views per Visit has high correlation than other features.
- Total visits and converted has very low correlation i.e. total visit we can derive meaningful lead scoring
- Total visits and Total Time spent on Website have a reasonable correlation result
- There is positive correlation between Total Time Spent on Website and Conversion
- There is almost no correlation in Page Views Per Visit and Total Visits with Conversion





Model Building:

- For model building we need to scale and split data into train and test.
- Using Logistic regression we are building model.
- Variable selection done through RFE (recursive feature elimination) and further manually we removed features with high P value and VIF value.
- Analyzing various parameter for train dataset Specificity, Sensitivity, Accuracy, Precision, Recall for train data.
- Plot the ROC Curve: It demonstrates several things:- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity.

Recursive Feature elimination :

```
col = X_train.columns[rfe.support_]
col
```

```
Index(['Total Time Spent on Website', 'Lead Origin_Landing Page Submission',
       'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat',
       'Lead Source_Reference', 'Lead Source_Welingak Website',
       'Last Activity_Email Bounced', 'Last Activity_Email Opened',
       'Last Activity_Olark Chat Conversation', 'Last Activity_Other_Activity',
       'Last Activity_SMS Sent', 'Last Activity_Unreachable',
       'Specialization_Hospitality Management', 'Specialization_Others',
       'What is your current occupation_Housewife',
       'What is your current occupation_Student',
       'What is your current occupation_Unemployed',
       'What is your current occupation_Working Professional'],
      dtype='object')
```

```
X_train.columns[~rfe.support_]
```

```
Index(['TotalVisits', 'Page Views Per Visit', 'Lead Origin_Lead Import',
       'Lead Source_Facebook', 'Lead Source_Google',
       'Lead Source_Organic Search', 'Lead Source_Others',
       'Lead Source_Referral Sites', 'Last Activity_Email Link Clicked',
       'Last Activity_Form Submitted on Website',
       'Last Activity_Page Visited on Website', 'Last Activity_Unsubscribed',
       'Specialization_Business Administration', 'Specialization_E-Business',
       'Specialization_E-COMMERCE', 'Specialization_Finance Management',
       'Specialization_Healthcare Management',
       'Specialization_Human Resource Management',
       'Specialization_IT Projects Management',
       'Specialization_International Business',
       'Specialization_Marketing Management',
       'Specialization_Media and Advertising',
       'Specialization_Operations Management',
       'Specialization_Retail Management',
       'Specialization_Rural and Agribusiness',
       'Specialization_Services Excellence',
       'Specialization_Supply Chain Management',
       'Specialization_Travel and Tourism',
       'What is your current occupation_Other', 'City_Other Cities',
       'City_Other Cities of Maharashtra', 'City_Other Metro Cities',
       'City_Thane & Outskirts', 'City_Tier II Cities'],
      dtype='object')
```

Logistic Regression Model

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	7259
Model:	GLM	Df Residuals:	7247
Model Family:	Binomial	Df Model:	11
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3038.7
Date:	Mon, 07 Sep 2020	Deviance:	6077.5
Time:	17:50:09	Pearson chi2:	7.97e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9083	0.134	-6.774	0.000	-1.171	-0.645
Total Time Spent on Website	1.1306	0.037	30.420	0.000	1.058	1.203
Lead Origin_Landing Page Submission	-1.0897	0.119	-9.187	0.000	-1.322	-0.857
Lead Origin_Lead Add Form	3.9061	0.212	18.395	0.000	3.490	4.322
Lead Source_Olark Chat	1.1547	0.115	10.082	0.000	0.930	1.379
Last Activity_Email Bounced	-1.4049	0.315	-4.455	0.000	-2.023	-0.787
Last Activity_Email Opened	0.6100	0.097	6.303	0.000	0.420	0.800
Last Activity_Olark Chat Conversation	-1.0969	0.179	-6.131	0.000	-1.448	-0.746
Last Activity_Other_Activity	2.3254	0.483	4.813	0.000	1.378	3.272
Last Activity_SMS Sent	1.7121	0.100	17.166	0.000	1.517	1.908
Specialization_Others	-1.1145	0.115	-9.692	0.000	-1.340	-0.889
What is your current occupation_Working Professional	2.7092	0.182	14.882	0.000	2.352	3.066

Using RFE and manual feature elimination for features having P-Value more than 0.05 and VIF more than 5. We reached a final model with P-Value less than 0.05 and VIF less than 5.

	Features	VIF
1	Lead Origin_Landing Page Submission	2.69
9	Specialization_Others	2.49
5	Last Activity_Email Opened	2.48
8	Last Activity_SMS Sent	2.33
3	Lead Source_Olark Chat	2.22
6	Last Activity_Olark Chat Conversation	1.79
2	Lead Origin_Lead Add Form	1.37
0	Total Time Spent on Website	1.32
10	What is your current occupation_Working Professional	1.18
4	Last Activity_Email Bounced	1.17
7	Last Activity_Other_Activity	1.02

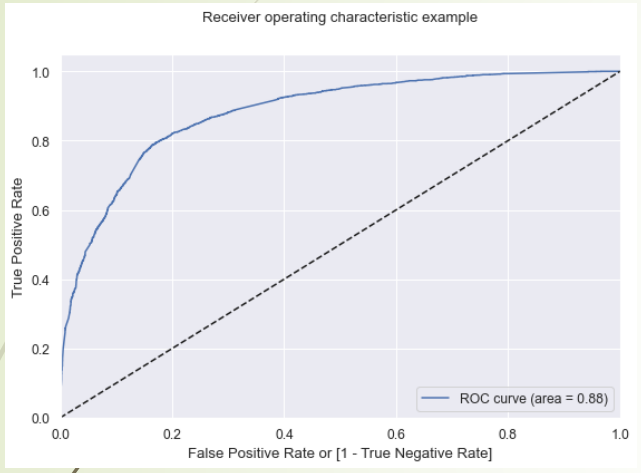
Coefficients

```
round(res9.params.sort_values(ascending=False),4)
```

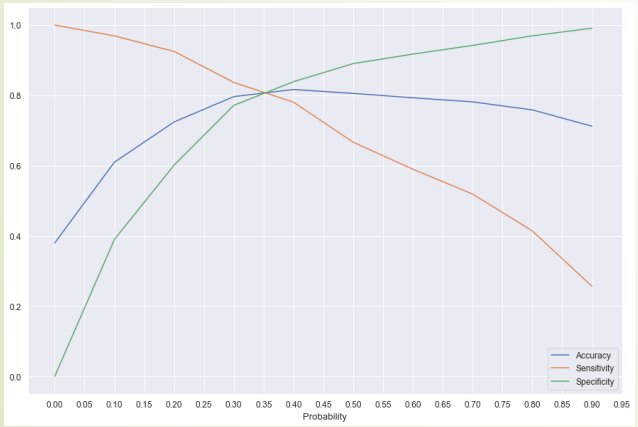
```
Lead Origin_Lead Add Form                              3.9061
What is your current occupation_Working Professional   2.7092
Last Activity_Other_Activity                           2.3254
Last Activity_SMS Sent                                 1.7121
Lead Source_Olark Chat                                 1.1547
Total Time Spent on Website                            1.1306
Last Activity_Email Opened                             0.6100
const                                                 -0.9083
Lead Origin_Landing Page Submission                   -1.0897
Last Activity_Olark Chat Conversation                 -1.0969
Specialization_Others                                 -1.1145
Last Activity_Email Bounced                           -1.4049
dtype: float64
```

Plotting the ROC Curve

ROC Curve



Finding Optimal Cut-off Point

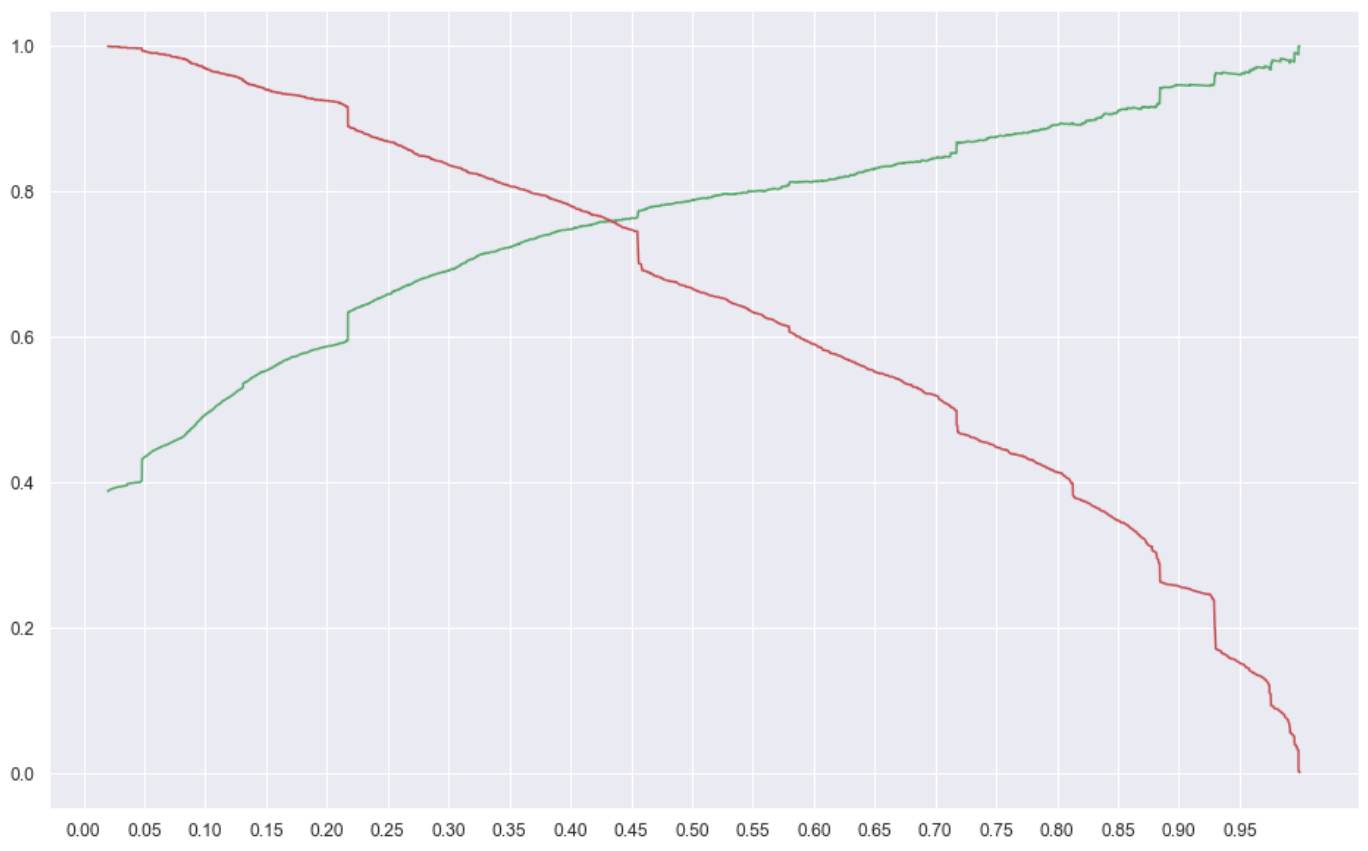


An ROC curve demonstrates several things:

- It shows the trade-off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

Precision and recall trade-off

- As per Precision-Recall Trade-off, the cut-off is around 0.425 (between 0.4 and 0.45) . We can choose the cut-off as 0.47 and use the Precision-Recall-Accuracy metrics to evaluate the model.



Lead scores for varying cut-off probability

	Probability Cut-Off	Projected Leads
0	0.05	7870.0
1	0.10	6739.0
2	0.15	5801.0
3	0.20	5381.0
4	0.25	4487.0
5	0.30	4117.0
6	0.35	3794.0
7	0.40	3537.0
8	0.45	3328.0
9	0.50	2891.0
10	0.55	2704.0
11	0.60	2472.0
12	0.65	2273.0
13	0.70	2103.0
14	0.75	1752.0
15	0.80	1585.0
16	0.85	1298.0
17	0.90	928.0
18	0.95	553.0

Here we examine the Projected lead scored for different cut-off probability to estimate the lead. So, this will be an useful template to change the cut-off based on business needs.

Model Evaluation Statistics

Final Observation:

Comparison of the values obtained for Train & Test:

```
print("Train Data Accuracy    :{} %".format(round((trainaccuracy*100),2)))
print("Train Data Sensitivity :{} %".format(round((trainSensitivity*100),2)))
print("Train Data Specificity :{} %".format(round((trainSpecificity*100),2)))
print("Train Data F1 Score    :{}  ".format(round((trainF1_score_train),2)))
print("Test Data Accuracy     :{} %".format(round((testaccuracy*100),2)))
print("Test Data Sensitivity  :{} %".format(round((testsensitivity*100),2)))
print("Test Data Specificity  :{} %".format(round((testspecificity*100),2)))
print("Test Data F1 Score     :{}  ".format(round((testF1_score),2)))
```

```
Train Data Accuracy    :80.95 %
Train Data Sensitivity :80.73 %
Train Data Specificity :81.08 %
Train Data F1 Score    :0.72
Test Data Accuracy     :81.54 %
Test Data Sensitivity  :80.73 %
Test Data Specificity  :81.08 %
Test Data F1 Score     :0.74
```

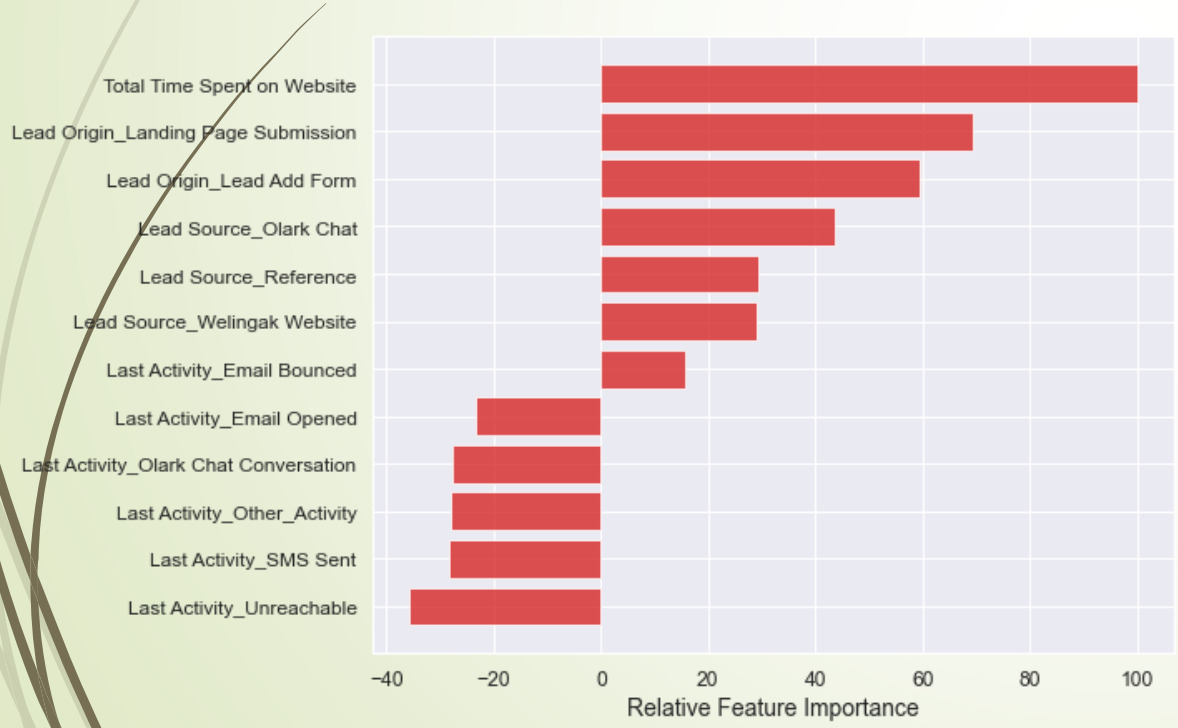
Classification Report

```
print (classification_report(Y_train_pred_final['Converted'], Y_train_pred_final['final_predicted']))
```

	precision	recall	f1-score	support
0	0.87	0.81	0.84	4503
1	0.72	0.81	0.76	2756
accuracy			0.81	7259
macro avg	0.80	0.81	0.80	7259
weighted avg	0.82	0.81	0.81	7259

Relative Feature Importance and Recommendations

Recommendations :
X Education Company needs to focus on following key aspects to improve the overall conversion rate:



- Increase user engagement on Welingak website since this helps in higher conversion
- Focus on Working Professional which has high conversion certainty.
- Get Total Time Spent on Website increased by advertising and user experience which makes the customer engaging in the website. since this helps in higher conversion
- Improve the Olark Chat service since this is affecting the conversion negatively
- Improving Lead add form also improves the lead conversion with high certainty