Lead Scoring Case Study – X Education

BATCH – DS 46,

SUBMISSION DATE: 24TH JAN 2023

BY

Introduction

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

Many professionals who are interested in the courses land on their website and browse for courses. Company also market its courses in other websites. Interested candidates fill up form with their details. These informations considered as potential leads. Companies also receive leads from past referrals.

Sales team for X- Eduation deals with leads and aims to maximize the conversion to paying customers. Sales team calls the potential customers / hot leads and pitch for their courses.

Problem Statement

Company's current lead conversion rate is 30% which is very low. Target for improving lead conversion rate as 80%. And for this purpose company wishes to analyze past data (~9000 records), identify factors influencing lead conversion and build a machine learning model. It would assign score to each lead and the leads with higher score would be potential lead having higher conversion chance. Intention is to contact only potential leads thus saving on the resources and maximizing on the conversion rate.

Data Study

Sample Size: 9240

Feature Variables:
35 variables includes:

- ID columns: Prospect ID and Lead Number
- Candidate online search details
- Profile of the candidate
- Tags added by Sales team post sales conversations

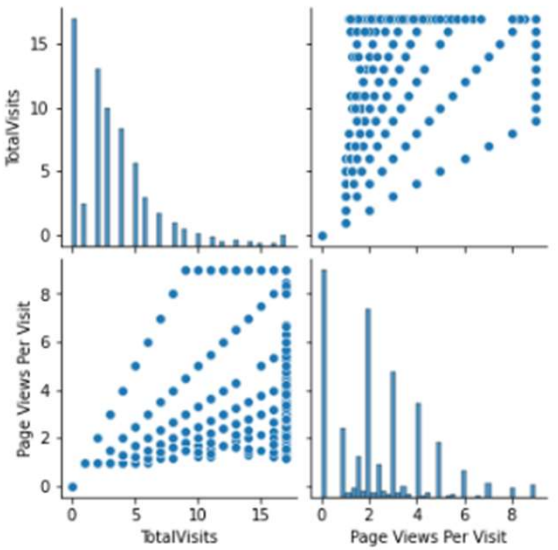
Target Variable:
Converted
1 = Lead converted to paying customer
2 = Not Converted

Analysis & Model building Steps

- Read and Understand the problem statement
- Go through the data dictionary and lead data to understand the same.
- Data Inspection / Data Quality Checks / NULL Inspection
- Data Cleaning
  - Unnecessary columns and columns where imbalance values there can be removed as they will not help for any insights and information can be biased.
  - NULL / Outliers treatment based. Drop column, delete rows or impute where necessary
  - Data specific treatments (like "Select" category should be unknown field.
- Exploratory data analysis. Visualize data for better insights.
- Data Preparation for model building (includes creation of dummy fields/features from categorical variables)
- Model Building by splitting train test data on selected features (via RFE method), scaling, Verify statistical parameters and optimize on the features
- check model performance over test data (confusion matrix, sensitivity, F1-score and etc)
- Prediction and evaluation metrics

Exploratory Data Analysis

Pair-plot clearly indicated that if Page Views per visit increases total view also increases. These two attributes are positively correlated. So looks like **correlation and causation** exists. Due to this reason one of them can be considered for the analysis.

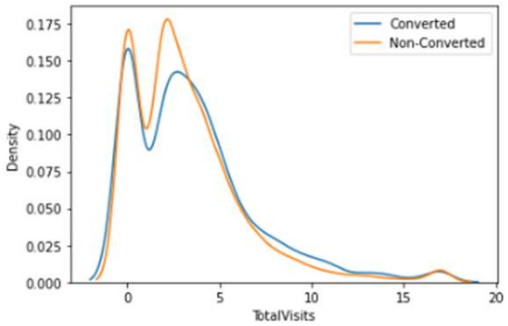


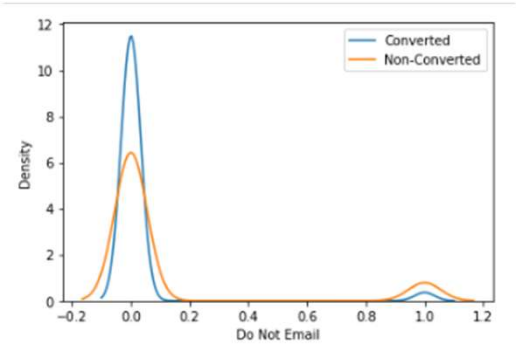
Exploratory Data Analysis (Contd..)

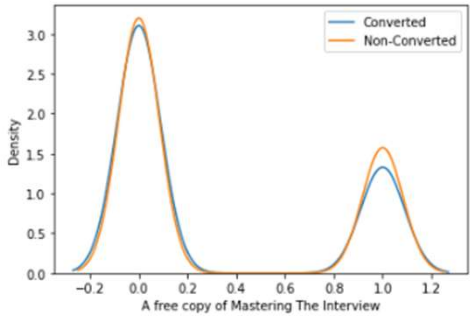
When **Total Visits** are lesser non-conversions are higher. So total visits are more conversion ratio increases.

Customer who has **requested for not sending email** where non-conversions are higher. Conversions are higher when customer has not requested for not sending emails.

Customers who requested for "**free copy of mastering the interview**", non-conversion is slightly higher. So this is not providing significant insights.

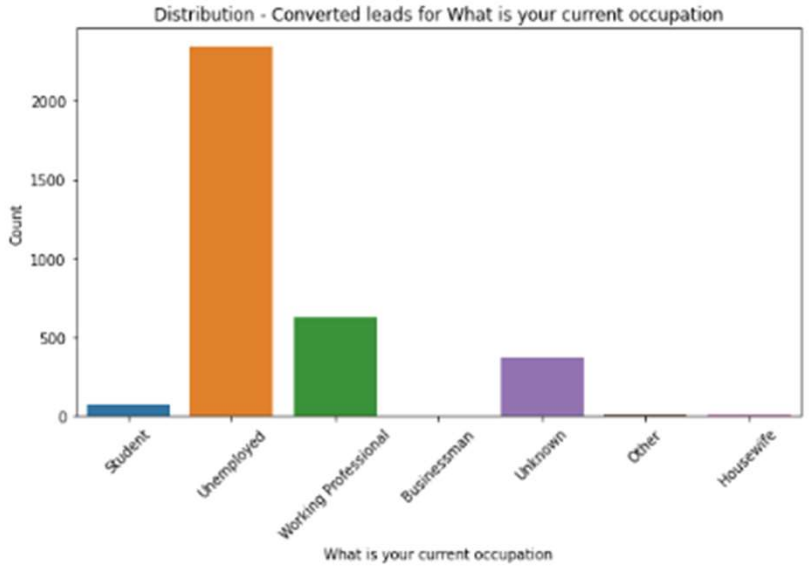


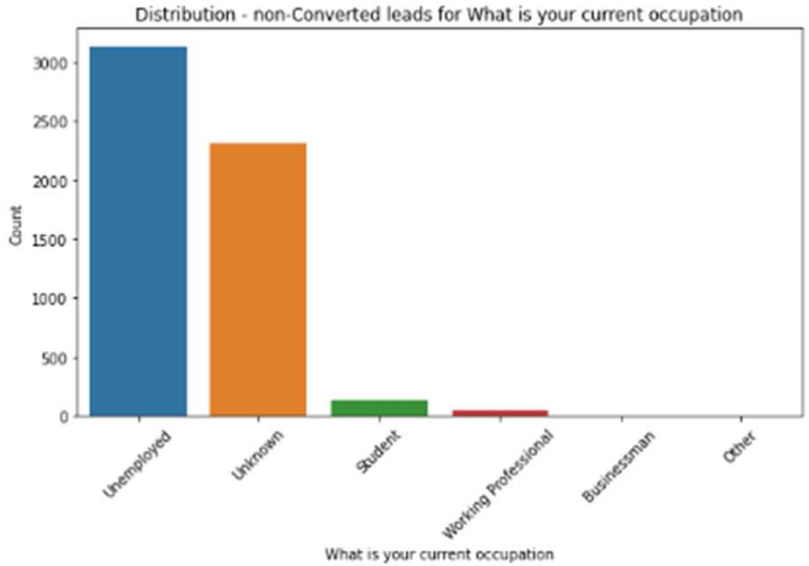




Exploratory Data Analysis (Contd..)

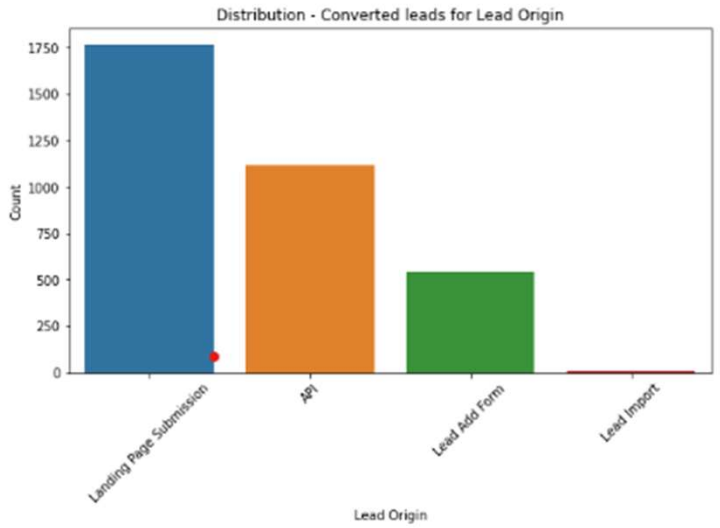
Conversion / non-conversion for **unemployed leads** are always higher. However observed that **working professionals leads** are higher in conversion then non-conversion

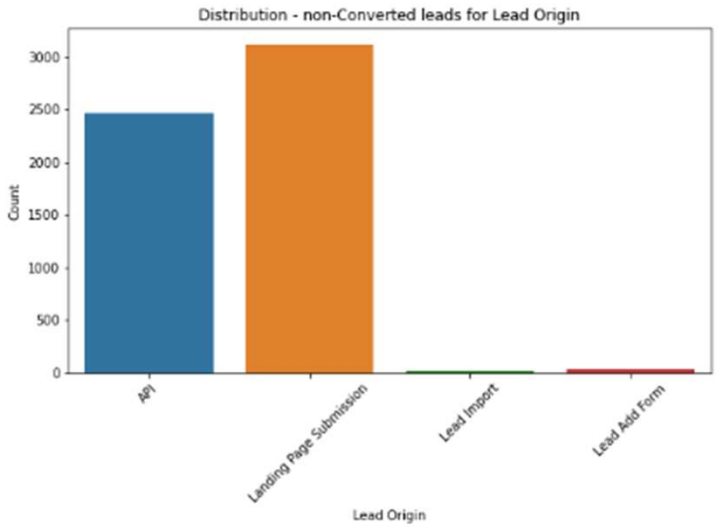




Exploratory Data Analysis (Contd..)

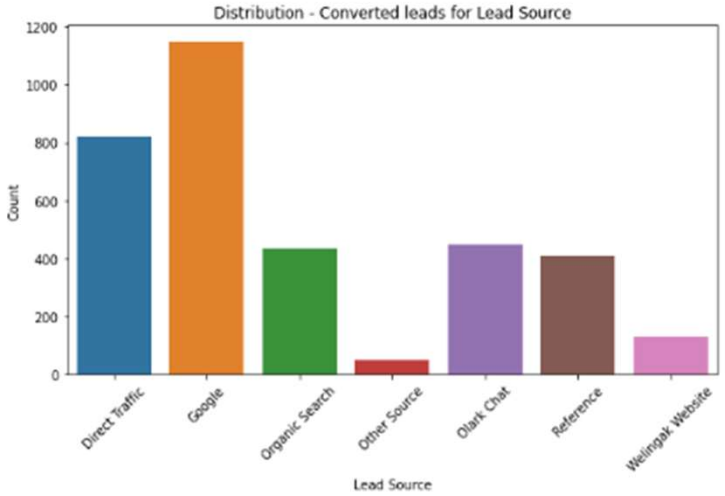
Conversion / non-conversion for **lending page submission and API** are always higher. However observed that **leads who have added forms** are higher in conversion then corresponding non-conversion.

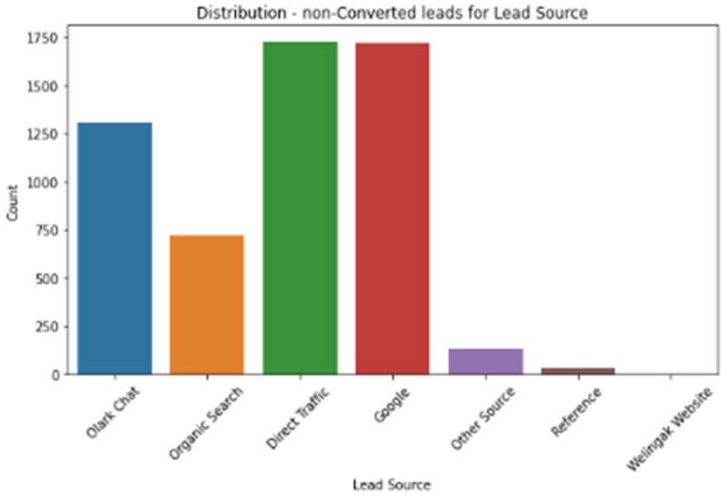




Exploratory Data Analysis (Contd..)

Conversion / non-conversion for **lead source as Google and direct traffic** are always higher. However observed that **lead sources for Welingak website** are higher in conversion then corresponding non-conversion. But the population is relatively lesser so need to review its statistical significance.

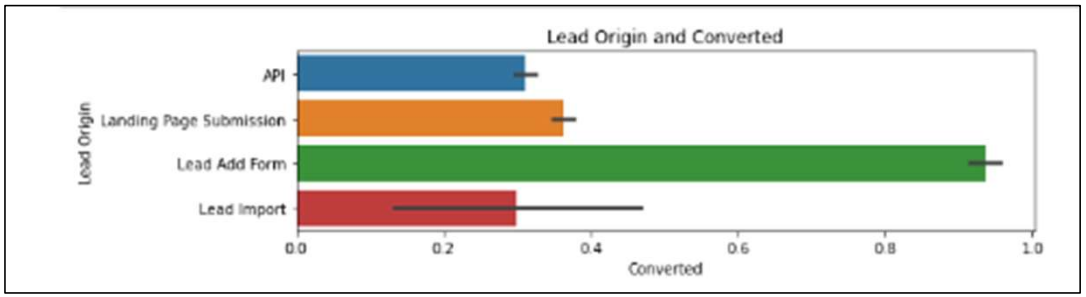


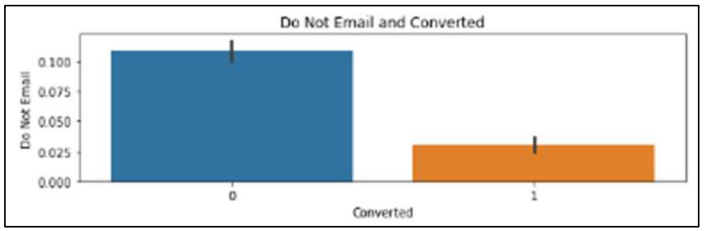


Exploratory Data Analysis (Contd..)

Following Bar-plot indicates that

- if lead origin is because of form addition then chances of conversion is higher.
- If the customer / lead has instructed not to send email which indicates that most likely will not be converted to paying customer



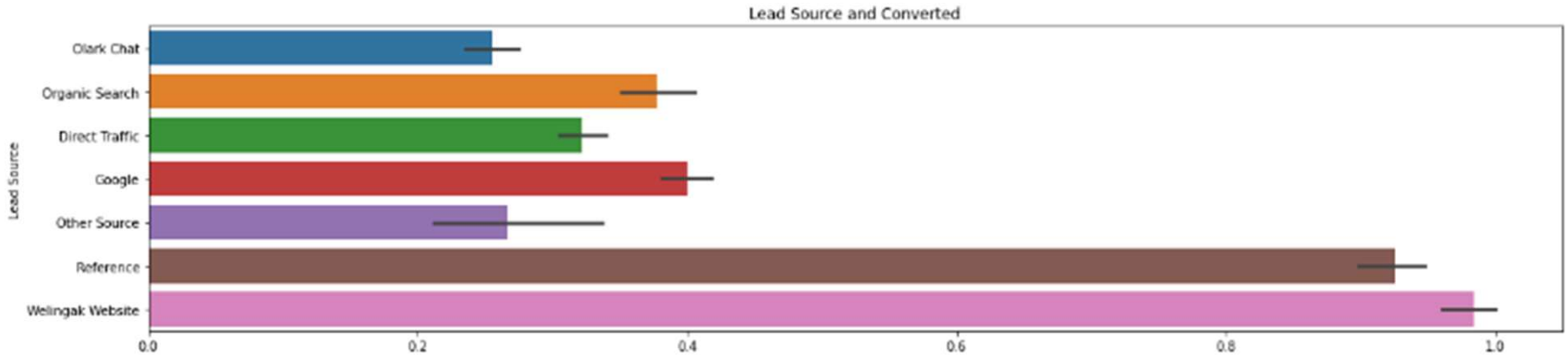


Exploratory Data Analysis (Contd..)

Following Bar-plot indicates that if the lead source is a referral then most likely would be converted.

There is a sign that if lead source is from Welingak website there are chances of conversion

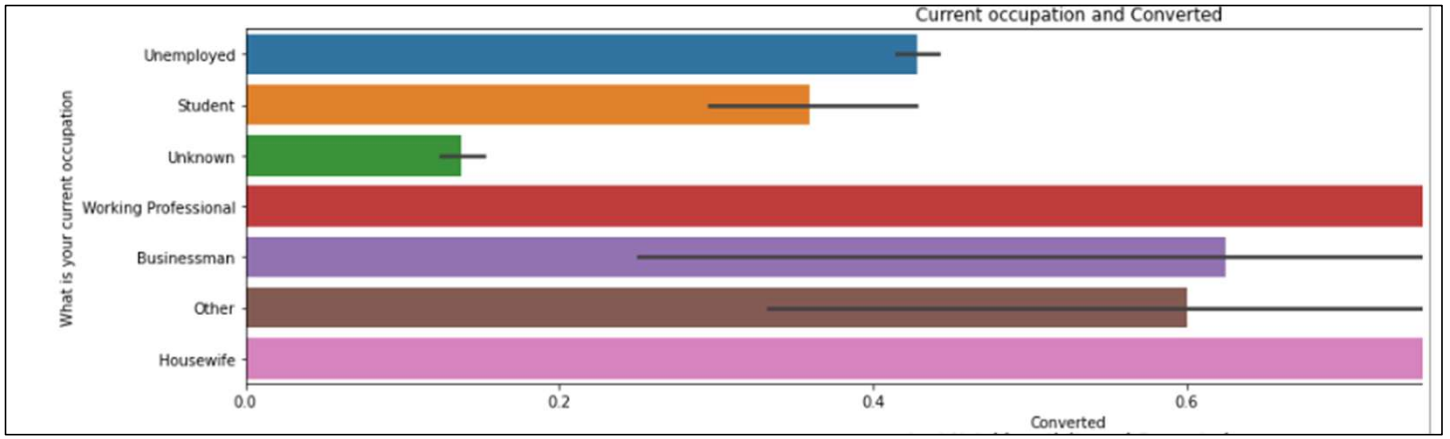
Olark chat or direct traffic has comparatively less impact on the conversion.



Exploratory Data Analysis (Contd..)

Working professional's are most likely will be taking up online classes and become paying customers.

Although in the graph it suggests that house-wife are potential customers but their overall number is very less.



Exploratory Data Analysis (Contd..)

Inference from Heatmap : High correlation is observed between -

- 'Lead Source Reference' & 'Lead_Origin_Lead Add Form' of 0.85
- 'Lead_Origin API' & 'Lead_Origin_Landing Page submission' of -0.87
- 'Specialization_Unknown' & 'Lead Origin Landing Page Submission' of -0.76
- 'Country_India' & 'Lead_Source_Olak_Chart' of -0.74
- 'Page Views per visit' & 'TotalVisits' of 0.71

RFE Analysis – Feature selection

Following 13 features are selected using RFE method:

- Do Not Email
- Total Time Spent on Website
- Lead_Origin_Lead Add Form
- Lead_Source_Olark Chat
- Lead_Source_Welingak Website
- Occupation_Housewife
- Occupation_Unknown
- Occupation_Working Professional
- Last_Ntbl_Activity_Email Link Clicked
- Last_Ntbl_Activity_Email Opened
- Last_Ntbl_Activity_Modified
- Last_Ntbl_Activity_Olark Chat Conversation
- Last_Ntbl_Activity_Page Visited on Website

Final Model

Logical Regression model built on the Initial 13 features received via RFE method. Analyzed p-value and RFE values and eliminated columns like "Occupation_Housewife", "Lead Source Welingkak Website".

On every step we have perform model evaluation by verifying factors like Accuracy, Precision, Sensitivity and precision.

Logical Regression Final model has 11 features on which we did testing as well as prediction.

	Features	VIF
8	Last_Ntbl_Activity_Modified	1.66
4	Occupation_Unknown	1.57
1	Total Time Spent on Website	1.53
7	Last_Ntbl_Activity_Email Opened	1.44
3	Lead_Source_Olark Chat	1.39
5	Occupation_Working Professional	1.15
2	Lead_Origin_Lead Add Form	1.12
0	Do Not Email	1.10
9	Last_Ntbl_Activity_Olark Chat Conversation	1.09
10	Last_Ntbl_Activity_Page Visited on Website	1.06
6	Last_Ntbl_Activity_Email Link Clicked	1.03

Final Model – Regression Summary

All p-values are zero which indicates all these columns are statistically significant.

If we see coefficients following attributes are positively impacting conversion:

- Total Time Spent on Website
- Lead_Origin_Lead Add Form
- Occupation_Working Professional
- Lead_Source_Olark Chat

Following negatively impacted:

- Last_Ntbl_Activity_Olark Chat Conversation
- Do Not Email

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6339
Model Family:	Binomial	**Df Model:**	11
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2592.2
Date:	Sat, 21 Jan 2023	**Deviance:**	5184.4
Time:	20:10:59	**Pearson chi2:**	6.49e+03
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.4035
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.6059	0.084	-7.255	0.000	-0.770	-0.442
Do Not Email	-1.8323	0.177	-10.324	0.000	-2.180	-1.484
Total Time Spent on Website	4.6361	0.167	27.697	0.000	4.308	4.964
Lead_Origin_Lead Add Form	4.1309	0.213	19.400	0.000	3.714	4.548
Lead_Source_Olark Chat	1.2391	0.103	11.984	0.000	1.036	1.442
Occupation_Unknown	-1.1764	0.088	-13.304	0.000	-1.350	-1.003
Occupation_Working Professional	2.4330	0.187	13.003	0.000	2.066	2.800
Last_Ntbl_Activity_Email Link Clicked	-1.7119	0.261	-6.569	0.000	-2.223	-1.201
Last_Ntbl_Activity_Email Opened	-1.3080	0.089	-14.656	0.000	-1.483	-1.133
Last_Ntbl_Activity_Modified	-1.9680	0.093	-21.122	0.000	-2.151	-1.785
Last_Ntbl_Activity_Olark Chat Conversation	-2.4788	0.329	-7.542	0.000	-3.123	-1.835
Last_Ntbl_Activity_Page Visited on Website	-1.5741	0.206	-7.635	0.000	-1.978	-1.170

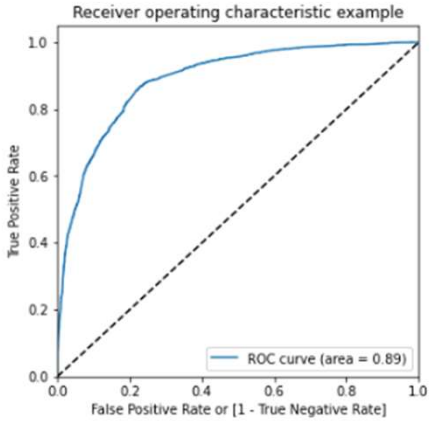
Final Model – Evaluation Parameters

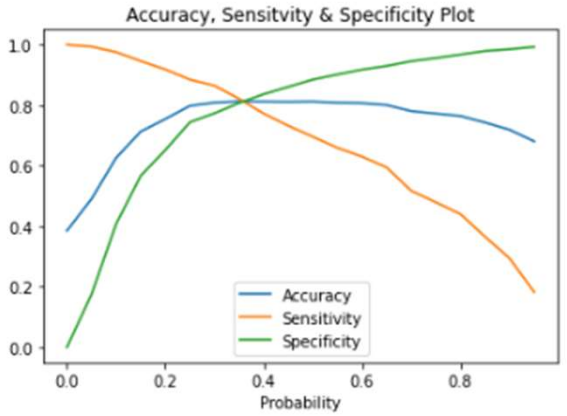
Probability	Accuracy	Sensitivity	Specificity	Precision	Recall
0	0.385136	1	0	0.385136	1
0.05	0.490474	0.993868	0.17516	0.430113	0.993868
0.1	0.626201	0.97547	0.407426	0.50766	0.97547
0.15	0.712959	0.946034	0.566965	0.577778	0.946034
0.2	0.754369	0.918234	0.651729	0.622851	0.918234
0.25	0.798142	0.88471	0.743918	0.683944	0.88471
0.3	0.808534	0.864677	0.773367	0.705	0.864677
0.35	0.812785	0.821341	0.807426	0.727635	0.821341
0.4	0.811841	0.772281	0.83662	0.747527	0.772281
0.45	0.811211	0.732216	0.860691	0.767024	0.732216
0.5	0.811998	0.69583	0.884763	0.790892	0.69583
0.55	0.808219	0.659035	0.901665	0.807615	0.659035
0.6	0.80696	0.630008	0.917798	0.827605	0.630008
0.65	0.800661	0.59444	0.929834	0.841435	0.59444
0.7	0.780507	0.517171	0.945455	0.855886	0.517171
0.75	0.772319	0.47915	0.955954	0.872024	0.47915
0.8	0.764447	0.440311	0.967478	0.894518	0.440311
0.85	0.742875	0.365495	0.979257	0.916923	0.365495
0.9	0.718942	0.293949	0.985147	0.925354	0.293949
0.95	0.68068	0.182339	0.99283	0.940928	0.182339

Final Model – ROC Curve

Accuracy, Sensitivity and specificity cut-off provides threshold point as 0.35

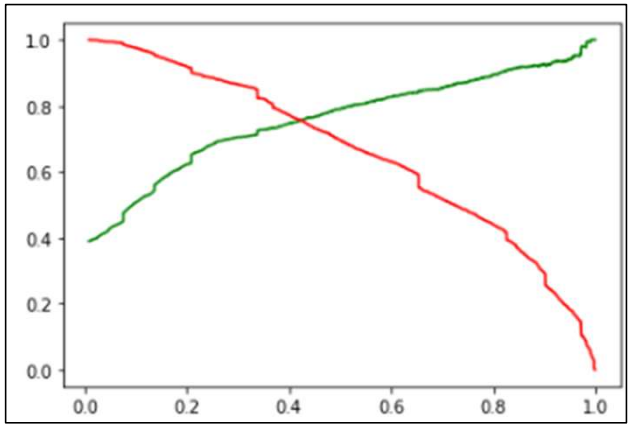
Area under ROC Curve is 0.89 which seems to be indication of good model.





Final Model – Precision & Recall Curve

- Optimal cut-off probability is near to 0.4.
- Precision ~ Recall Statistics remain close to 0.75



Final Model – Evaluation

- Confusion Metrics created with threshold at to 0.35.
- Overall Model accuracy is 80%

	Converted	Converted_Prob	final_predicted	Lead Score
0	0	0.094195	0	9.42
1	0	0.596414	1	59.64
2	0	0.208370	0	20.84
3	1	0.748723	1	74.87
4	1	0.673721	1	67.37

Confusion Metrics		
Actual \ Predicted	**Positive**	**Negative**
Positive	**1392**	**342**
Negative	**201**	**788**

Evaluation Metrics	
Accuracy	**80**
Sensitivity	**80**
Specificity	**80**
Precision	**70**

Final Model – Coefficients & Interpretation

Features	Coefficient	Interpretation / Insights
Total Time Spent on Website	4.6361	Higher time spent indicates higher chance of conversion
Lead Origin – Lead add form	4.1309	Lead filling the forms indicates higher ineterest and thus conversion
Occupation – Working Professional	2.4330	Working professionals most likely would be interested for pursuing online courses.
Lead Source Olark Chart	1.2391	Leads sources from Olark chat has seen higher conversion
Last Notable Olark Chart conversation	-2.4788	Last Notable activity as Olark chat conversation has negative impact.
Last Notable activity Modified	-1.9680	Last notable activity as customer modified information in web page has negative impact.
Do Not Email	-1.8323	Candidates who has not mentioned to stop email relatively higher chance of conversion.

Recommendation

Top features which indicates higher chances of conversion:

1. Total Time spent on website: More time spent on X-Education website and higher chances of conversion
2. Lead originated due to the fact that the lead has filled the form in website indicates lead has interest and most likely will be converted post sales pitch.
3. Working professionals are most likely would be opting for online course and hence higher chance of conversion

Additional Recommendation:

1. Referrals received are highly correlated with lead adding form; thus referrals most likely will be converted.
2. Housewives are very less in number however it is seen that higher chance of conversion. Potential future conversion.