Lead Score Case Study

Submitted By:

Anumeha Sinha

Avinash Kambalapelli

Lakshmi Santosh Kadari

Problem Statement and Goal

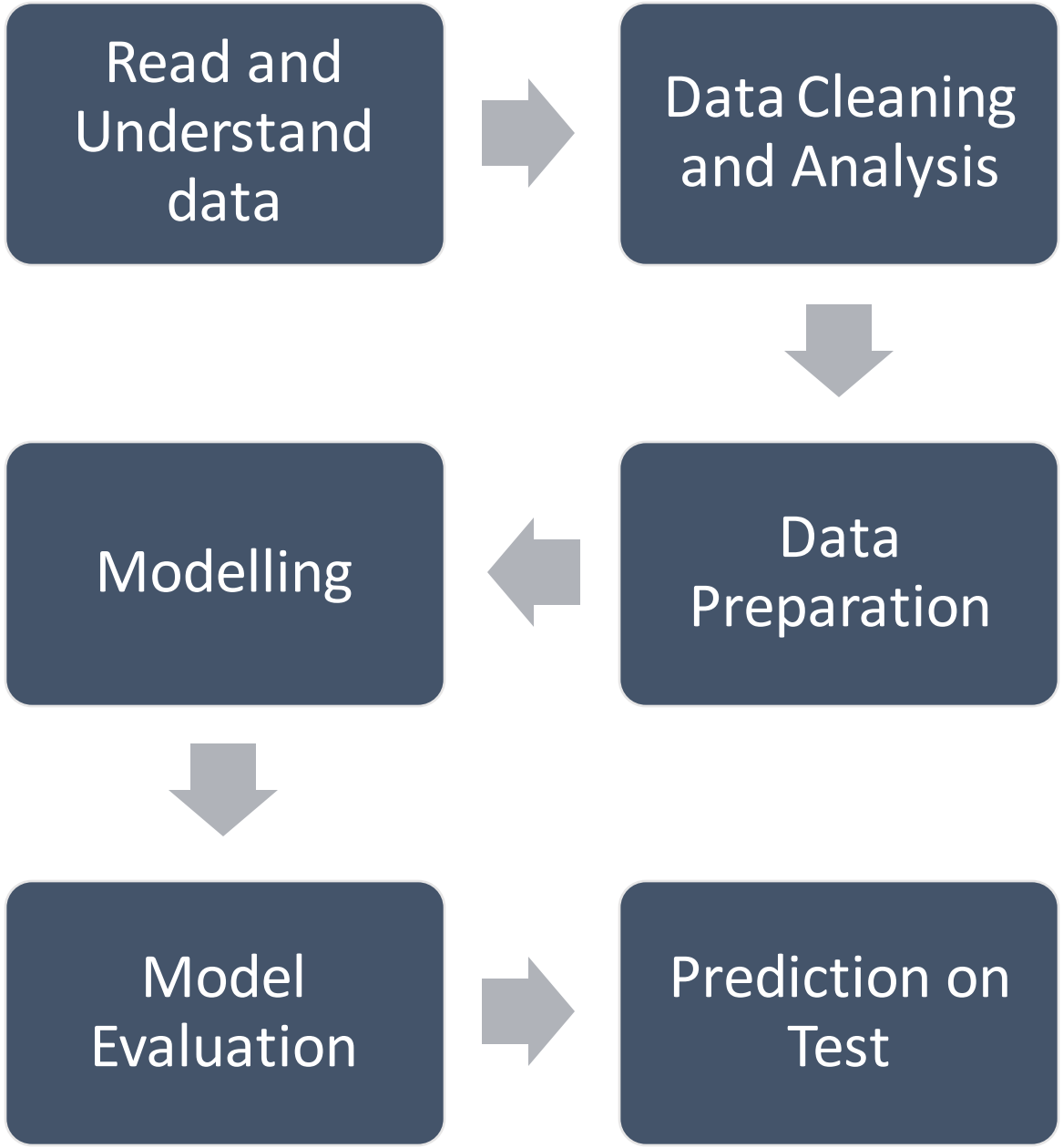
Problem Statement:

- An education company named X Education sells online courses to industry professionals
- Company markets its courses on several websites and search engines like Google
- Many professionals who are interested in the courses land on their website and browse for courses
- Although X Education gets a lot of leads, its lead conversion rate is very poor. In a day out of say 100 they are able to convert only 30(which is 30%)

Goal of Case study:

- Company wants to identify the potential leads(hot leads) and focus on hot leads by contacting them
- 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.

Analysis Approach



Analysis Approach

1. Read and Understand data : Understanding the business problem and analyzing it. Going through the dataset and understanding briefly the columns and rows

2. Data Cleaning and Analysis : There could be different issues within the datasets like missing values, outliers, incorrect format , inconsistent spelling. Fixing these issues in columns and rows by either dropping or filling relevant data ,changing the format or correcting the spelling are few fixes. Univariate , Bivariate and Multivariate analysis is done on feature variable to visualize and understand better. Outlier check is done to find out if values are out of range .Correct measure is taken if outliers are found – based on mean, median and IQR techniques

3. Data Preparation : Dummy variables are created for all the categorical variables. Train-Test split is done on the data (70-30). Divided data is then scaled using MinMaxScaler. Lastly the data is divided for modelling and prediction.

4. Modelling : Recursive Feature Elimination (RFE) is used to select 15 features and logistic regression model is built on these feature variables. Remove the feature one by one if p-value is greater than 0.05 and/or VIF is greater than 5.Same steps are repeated until a stable model is built with good performance matrix

5. Model Evaluation : Confusion matrix is made to find out model's performance using sensitivity, specificity, accuracy, precision and recall. Optimal probability cutoff is 0.38 for train data (calculated using ROC curve and by plotting sensitivity and specificity for various probabilities).

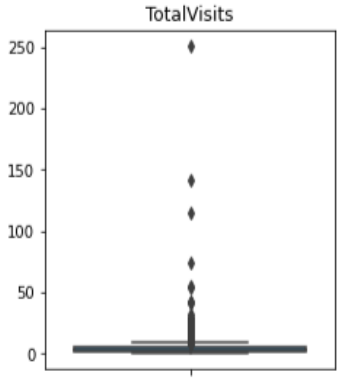
6. Prediction on Test : Prediction is done on test data using the optimal cutoff of 0.42. This is calculated by plotting Precision and Recall tradeoff. The accuracy is 89%, sensitivity is 86% and specificity is 90%.Lead score is assigned based on the probability (probability x100). Based on this score hot lead can be easily extracted.

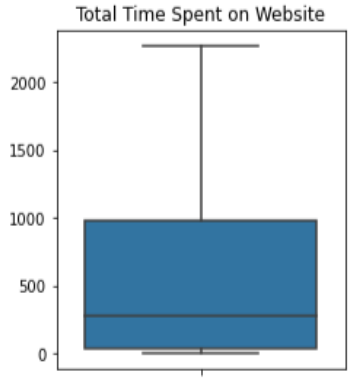
Data Manipulation

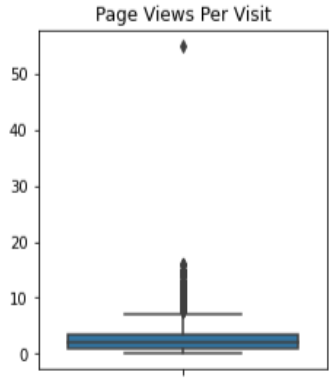
- All the columns which are having value as "Select" has been replaced with null values.
- Drop the columns which have more than 40% missing values
- Drop all categorical columns which are having majority values(>80%) same as its not adding value to the model. Ex: Country Column has 96% of people are from India and hence dropped
- Columns with less percentage of missing values are imputed with the required values- using mean, median or mode
- Replacing the levels of few categorical variables to "Others"
  for better understanding while building models.

EDA (Univariate Analysis – Outlier Check – numerical variables)

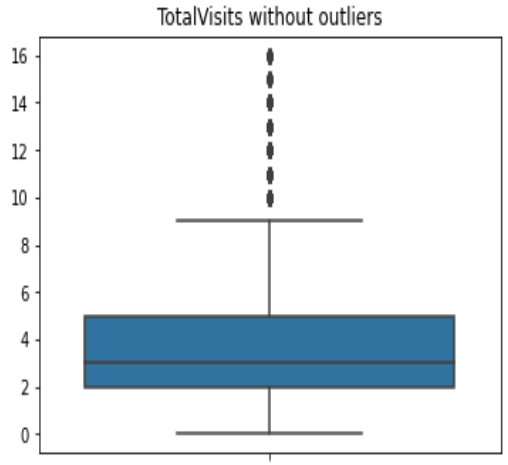
Huge outliers are present in "TotalVisits" and "Page Views Per Visit"

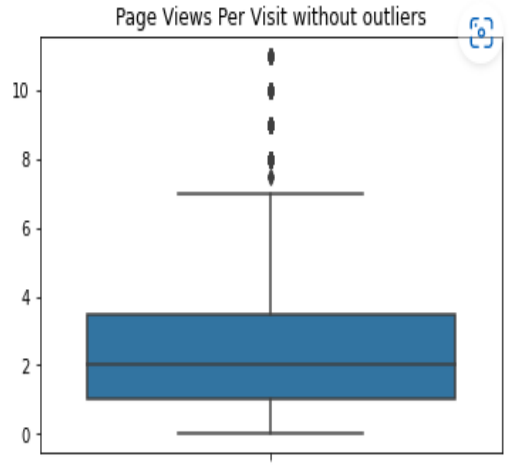






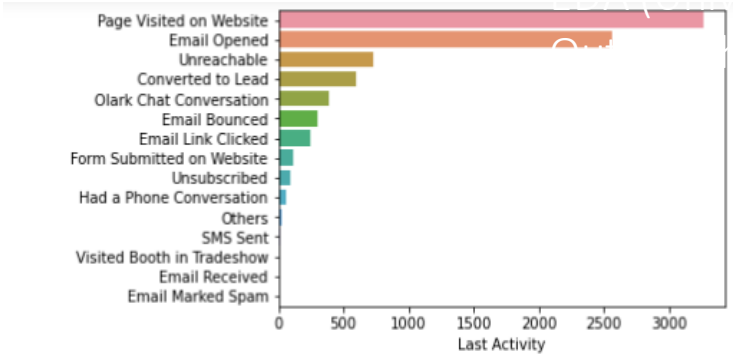
Outliers are treated using IQR technique taking Q1 as 0.2 and Q3 as 0.9

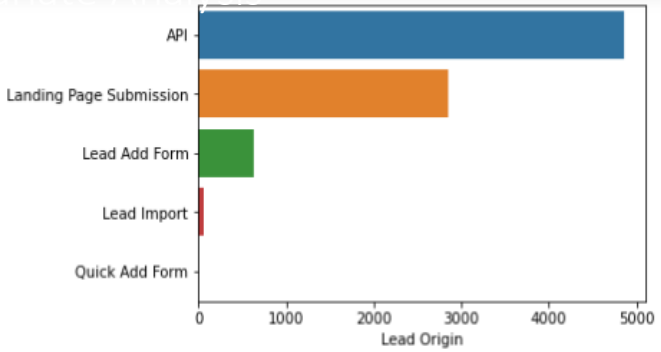


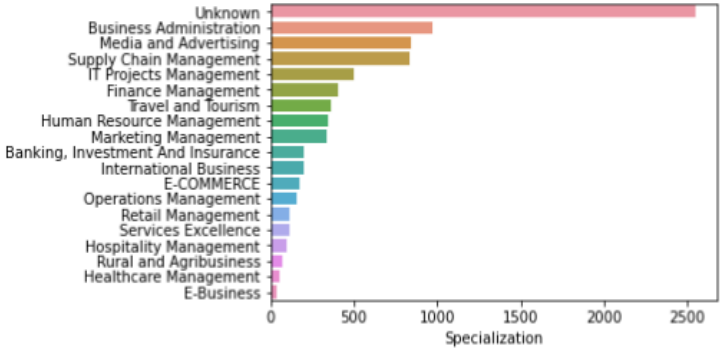


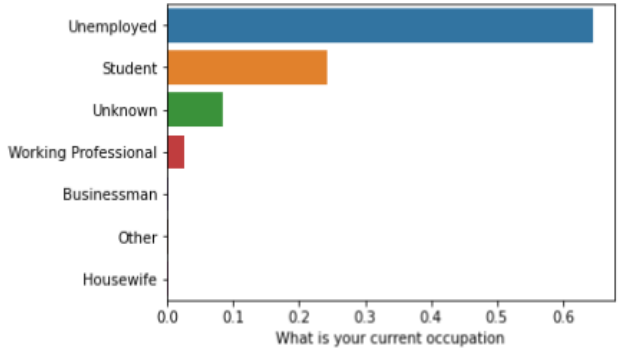
EDA (Univariate Analysis – Categorical variables)

- **Last Activity** - Many have visited pages on the websites
- **Lead Origin** - Most leads are through APIs
- **Specialization** - Most searched and opted specialization is "Business Administration". Mostly people looking for course are not sure about which specialization to opt for.
- **What is your current occupation** - Approx 60% of the people are Unemployed.





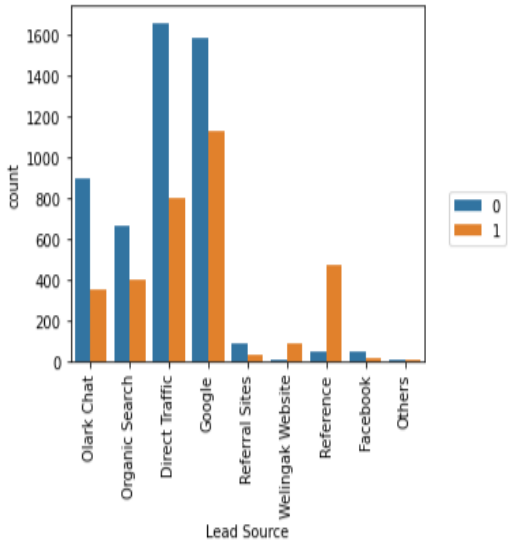


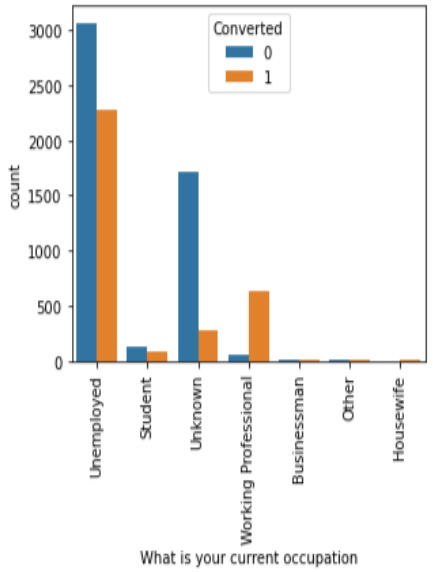


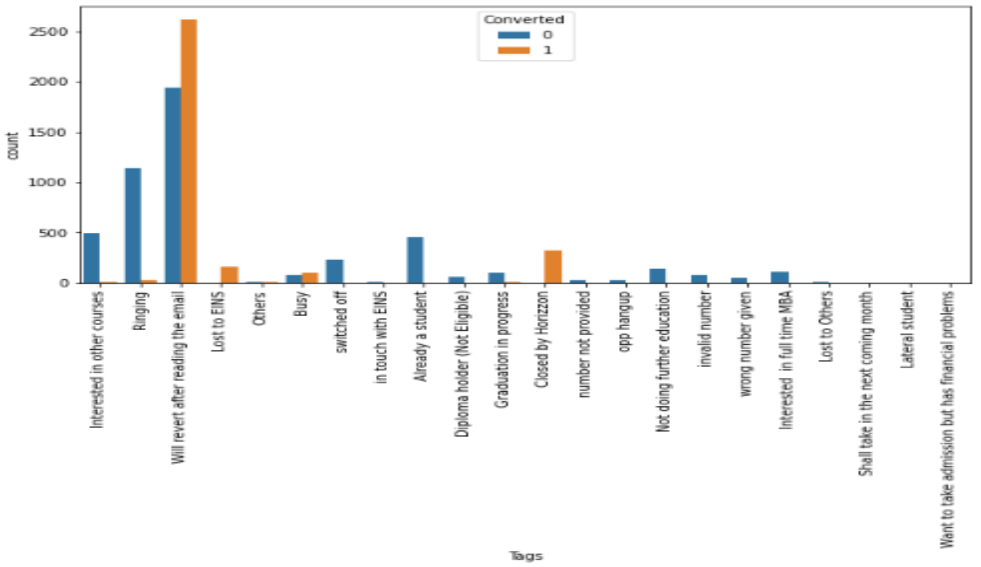
EDA(Bivariate analysis)

Checking the Lead Source and What is your current occupation with respect to "Converted" target variable

Checking Tags with respect to "Converted" target variable







- Most reliable conversion happens when the lead source is Google and Direct Traffic
- Highest conversion is of Unemployed and Working Professional
- Tags with "Will revert after reading the email" has the highest conversions

Data Preparation

- Dummy variable creation is done for categorical variables
- Split the data into Train-Test set (70% of data is train set and rest 30% for test set)
- Scaled using MinMaxScaler. Lastly the data is divided for modelling and prediction

Model Building(Train Data)

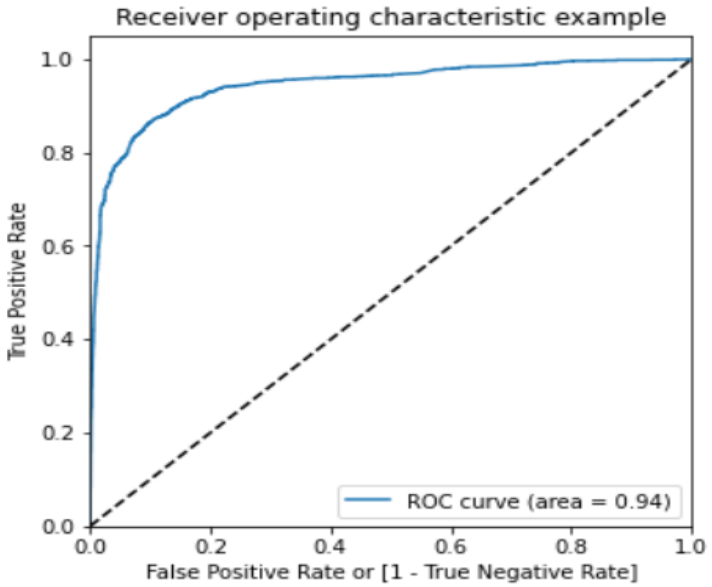
- RFE is used for feature selection with the initial selection of 15 features
- Logistic regression model is built on these selected feature variables
- P-value and VIF are checked and based on the values features are manually removed one by one.
- Remove the feature if p-value is greater than 0.05 and/or VIF is greater than 5
- The above steps are repeated 5 times to get a stable model with good sensitivity, specificity and accuracy

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1820	0.163	-7.266	0.000	-1.501	-0.863
Total Time Spent on Website	4.0593	0.197	20.576	0.000	3.673	4.446
Lead Origin_Landing Page Submission	-1.1425	0.151	-7.559	0.000	-1.439	-0.846
Lead Origin_Lead Add Form	3.0544	0.255	11.985	0.000	2.555	3.554
Last Activity_SMS Sent	1.3586	0.100	13.551	0.000	1.162	1.555
Specialization_Unknown	-1.0029	0.156	-6.442	0.000	-1.308	-0.698
What is your current occupation_Unknown	-3.0261	0.124	-24.462	0.000	-3.269	-2.784
What is your current occupation_Working Professional	1.5234	0.231	6.587	0.000	1.070	1.977
Tags_Interested in other courses	-2.9594	0.414	-7.143	0.000	-3.772	-2.147
Tags_Ringing	-3.3374	0.249	-13.411	0.000	-3.825	-2.850
Tags_Will revert after reading the email	2.4128	0.112	21.470	0.000	2.193	2.633
Last Notable Activity_Modified	-0.6323	0.096	-6.579	0.000	-0.821	-0.444

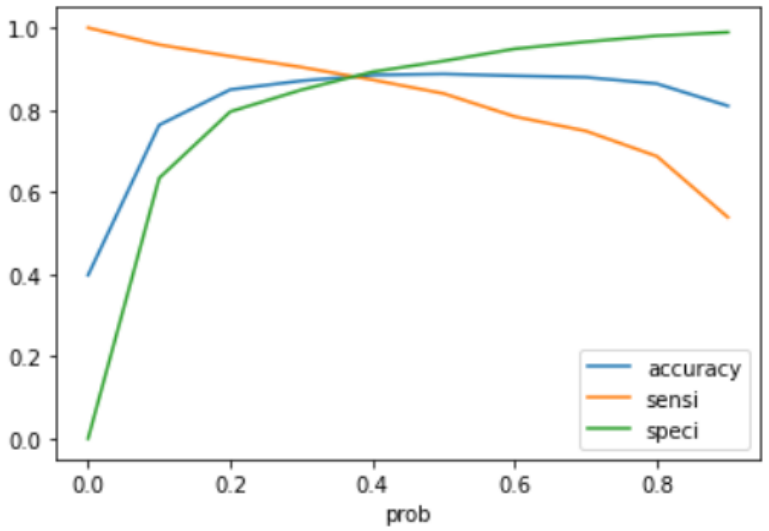
	Features	VIF
9	Tags_Will revert after reading the email	3.85
1	Lead Origin_Landing Page Submission	3.06
5	What is your current occupation_Unknown	2.17
0	Total Time Spent on Website	2.08
4	Specialization_Unknown	2.00
10	Last Notable Activity_Modified	1.71
3	Last Activity_SMS Sent	1.62
8	Tags_Ringing	1.48
2	Lead Origin_Lead Add Form	1.26
6	What is your current occupation_Working Profes...	1.26
7	Tags_Interested in other courses	1.21

Model Evaluation (Train Data)

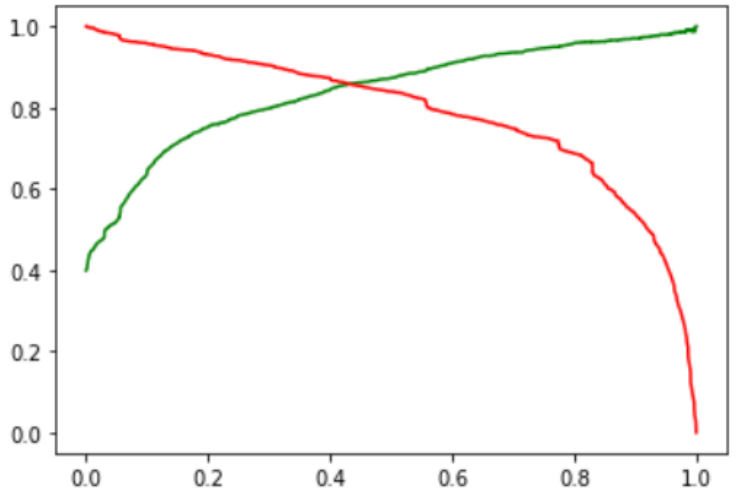
- Optimal probability cut-off is calculated (using ROC curve and by plotting sensitivity and specificity for various probabilities).
- The graph shows approximately **0.38** as the optimal probability cut-off value.
- Confusion Matix: [3205 282
  369 1936]
- Accuracy : 88%, Sensitivity : 87% and Specificity : 88%.
- Precision: 87% and Recall: 83%.



ROC Curve Area: 0.94



Optimal Cut-off value: 0.38

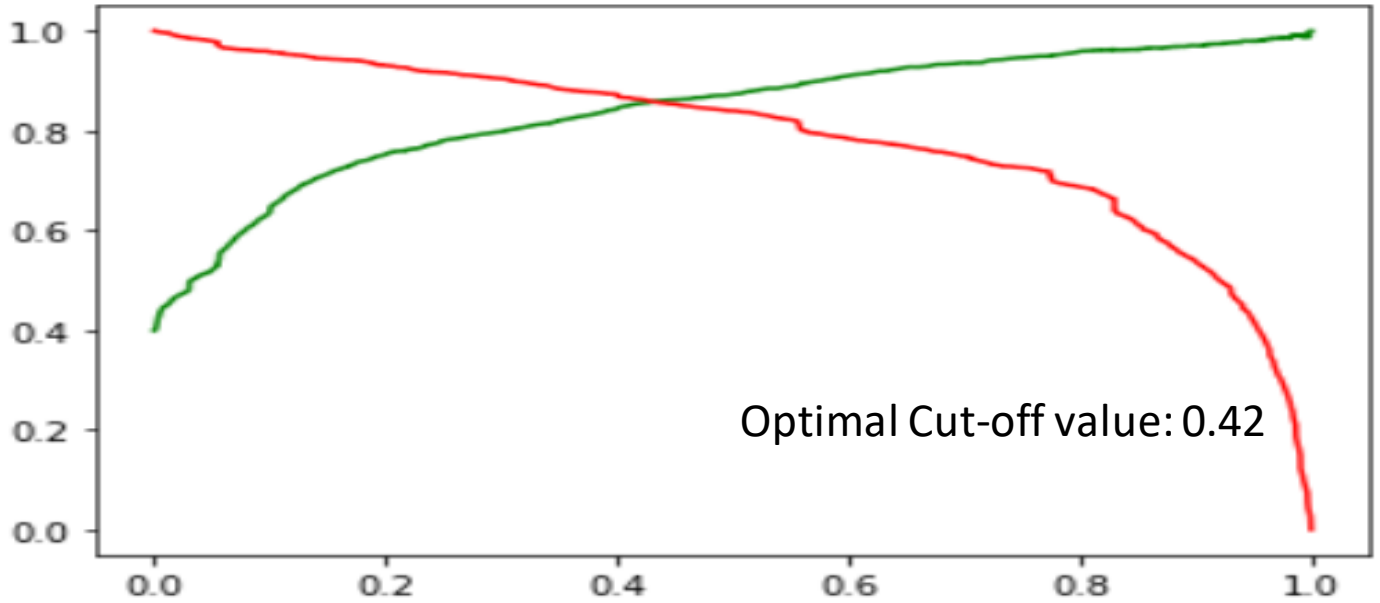


Precision/Recall Trade-off

Prediction(Test Data)

- The graph shows approximately **0.42** as the optimal probability cut-off value.
- Confusion Matix: [1360 136
  130 857]
- Accuracy : 89%, Sensitivity : 86% and Specificity : 90%.
- Precision: 84% and Recall: 89%.

Precision/Recall Trade-off Curve



Conclusion

- Lead score is assigned based on the probability (probability x100). Based on this score hot lead can be easily extracted.
- Lead score is calculated for both train and test dataset. The higher the lead score value there will be higher chance of conversion.

Train Data

	Converted	Conversion_Prob	Predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted	Lead Score
4353	1	0.975233	1	1	1	1	1	1	1	1	1	1	1	1	98
561	0	0.184513	0	1	1	0	0	0	0	0	0	0	0	0	18
1618	0	0.982166	1	1	1	1	1	1	1	1	1	1	1	1	98
647	1	0.656059	1	1	1	1	1	1	1	1	0	0	0	1	66
4435	0	0.314975	0	1	1	1	1	0	0	0	0	0	0	0	31

Test Data

	Converted	Conversion_Prob	final_predicted	Lead Score
0	0	0.056404	0	6
1	1	0.806960	1	81
2	0	0.031357	0	3
3	0	0.122361	0	12
4	0	0.006749	0	1

Conclusion

Significant features for hot leads	Business impact
Total Time Spent on Website	Reaching the people those who are spending **more time on the website.** The X Education company will get more potential buyers
Lead Origin – Filled the form	Reaching out to the people who have **filled the form.** The X Education company will get more potential buyers
Tags - Will revert after reading the email	Contact the leads those who have responded as **"Will revert after reading the email"**, these leads could be the potential buyers.
Current occupation - Working professionals and unemployed	Reaching out to the leads those who are **"Working professionals"** and **"Unemployed"** have higher chances of buying the courses
Last Activity – SMS Sent and Email Opened	Leads actively checking **"Email Opened"** and "**SMS Sent"** should be contacted most