Lead Scoring Business Solution

By Prranesh M B and Mangesh J

Problem Statement

X Education gets a lot of leads. However, its lead conversion rate is very poor. Lets say, if 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.

The ballpark figure of lead conversion rate is considered to be around 80%

Solution Approach

Data Reading and cleaning

- Data is read and initial inspection is performed to understand the data
- Cleaned the data by dropping off the columns that has no contribution to prediction

EDA and Data Preparation

- Perform the univariate and bi-variate analysis
- Test-train split, Scaling, and creation of dummy variables

Data Modelling and evaluation

- Create the model with the finalized variables and evaluate the model using ROC
- Finding of optimal threshold value to classify the probabilities

Predicting on Test data

- Predict the probabilities and evaluation of metrics on the test data

Data understanding

From the provided dataset, Initially there were **9240 records** of the customers with **37 attributes**.

```
Converted:
0    5679
1    3561
Name: Converted, dtype: int64
---------------------------------------
```

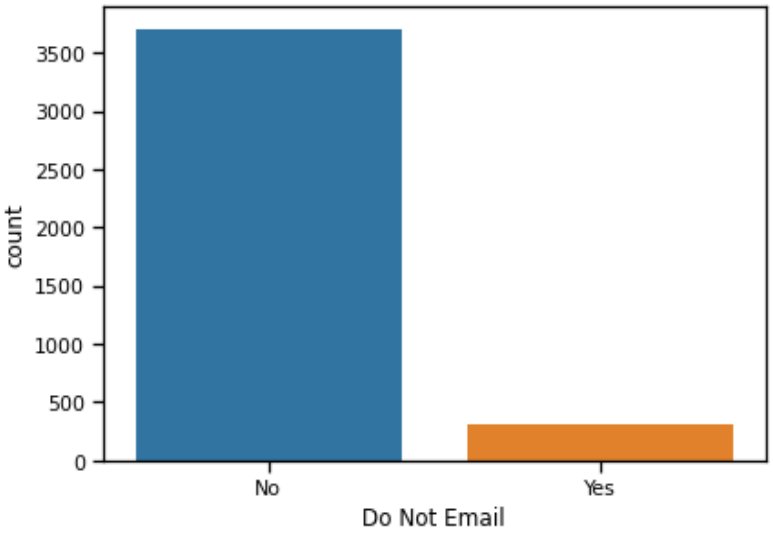
Out of them only 3561 customers were converted, which is approximately 33%.

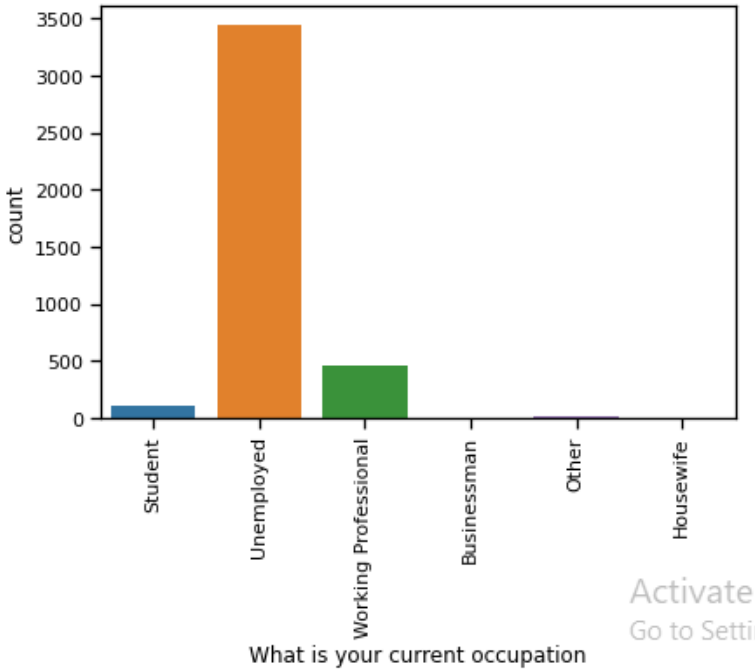
In the data cleaning step, the data size had been reduced to **Rows:** 4020 and **Columns :** 13. The EDA is performed on these attributes of the customer

```
Lead Origin
Lead Source
Do Not Email
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
Last Activity
Country
Specialization
What is your current occupation
A free copy of Mastering The Interview
Last Notable Activity
```

EDA observations

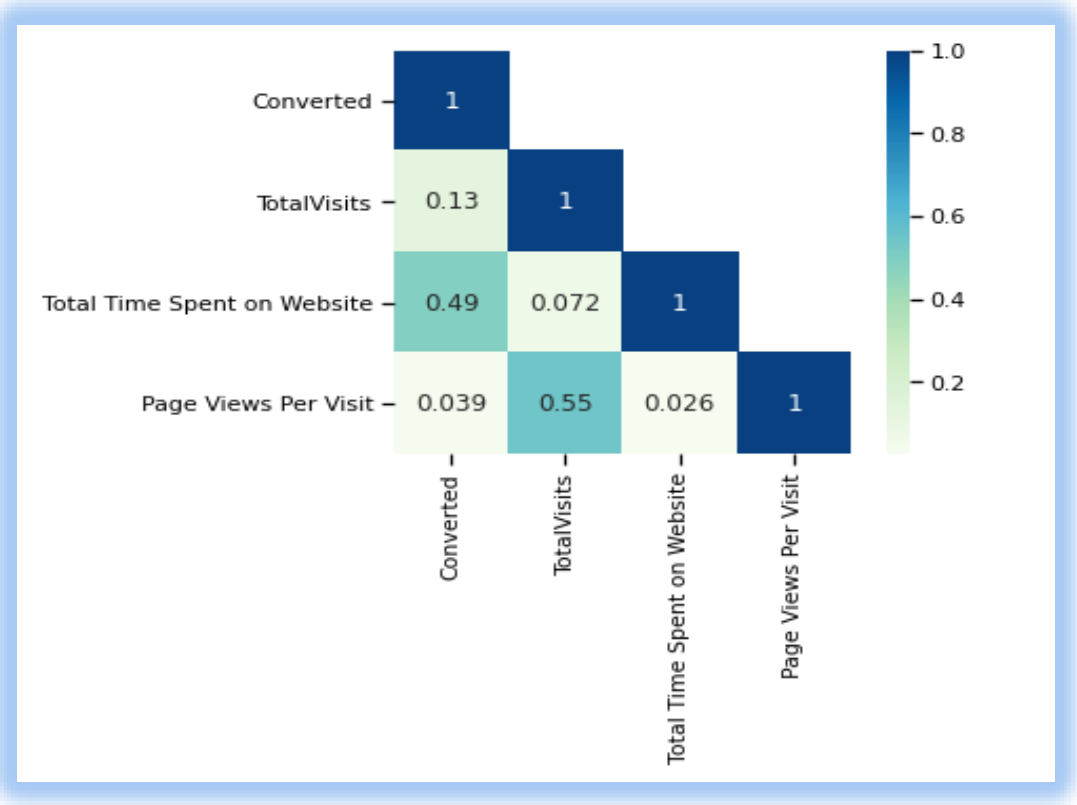
- There were few **outliers** identified, these outliers are capped at **95 percentile** value of respective columns
- From EDA of categorical variables, two predominant observations are identified





Similarly, a predominant observation is identified from numeric variables.

A good correlation between the target Variable (profiles which Are converted) and Total time spent on Website is high



Modelling

Step1 : Creation of dummy variables, Train-test split(train = 70%, test = 30% data), standard scaling of numeric variables

Step2: Using Recursive Feature Elimination, selected 15 variables initially

Step3: Hyper parameter tuning of the model considering the business knowledge

Step4: Evaluation of the model and ROC curve

Step5: Finding the optimal threshold value to classify the probability into potential leads and no leads using perspectives of sensitivity-specificity and Precision-recall

Modelling steps in brief

- After the dummy variables were created and merged to the original dataset. The data size has been increased to

Rows: 4020 Columns: 101

- Test-train split (70:30) is performed and then the scaling is done on the numerical variables "TotalVisits", "Total Time Spent on Website", "Page Views Per Visit“

- Using RFE techniques, from 101 columns, 15 columns we filtered out to construct first model

- Along with consideration of columns that provide Business semantics, further fin-tuning of the model is done by analyzing the statistical summary of the mode and its multi-collinearity (correlation between the variables)

- The final model is attained after build 5 models by hyper-parameter tuning

Below is the statistical summary of the list of columns that are concluded for solving the business problem with which the final model is built. All these variables has VIF value less than 2.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	2814
Model:	GLM	Df Residuals:	2806
Model Family:	Binomial	Df Model:	7
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1316.3
Date:	Sun, 05 Sep 2021	Deviance:	2632.7
Time:	14:49:35	Pearson chi2:	2.89e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1734	0.147	1.179	0.238	-0.115	0.461
Do Not Email	-1.5009	0.234	-6.413	0.000	-1.960	-1.042
Total Time Spent on Website	1.1256	0.052	21.683	0.000	1.024	1.227
Lead Origin_Landing Page Submission	-0.8909	0.154	-5.770	0.000	-1.194	-0.588
Lead Source_Referral Sites	1.6918	0.649	2.608	0.009	0.420	2.963
Last Activity_Had a Phone Conversation	1.9884	0.785	2.534	0.011	0.451	3.526
Last Activity_SMS Sent	0.9699	0.103	9.458	0.000	0.769	1.171
What is your current occupation_Working Professional	3.1250	0.273	11.433	0.000	2.589	3.661

Model Evaluation and ROC

Keeping the threshold to an **arbitrary value (0.5)** to classify the leads, The final model has been evaluated on the following metrics

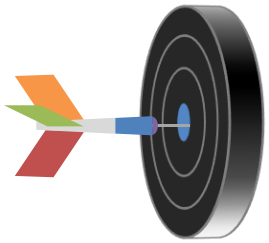
on the train data

Accuracy : 79 %

Sensitivity: 74 %

Specificity: 83%

False positive rate: 16%



Positive predictive values: correctly predicted converted leads/ Total converted leads

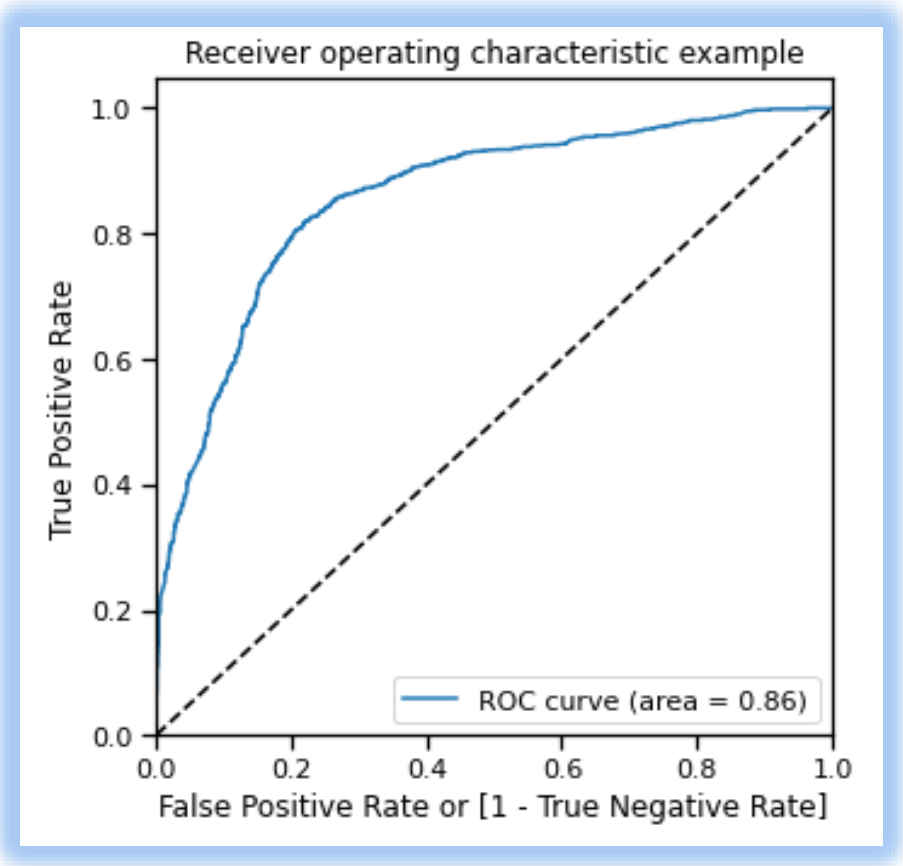
= 80%

Negative predictive values: correctly predicted non-converted leads/ total non-converted leads

= 78%

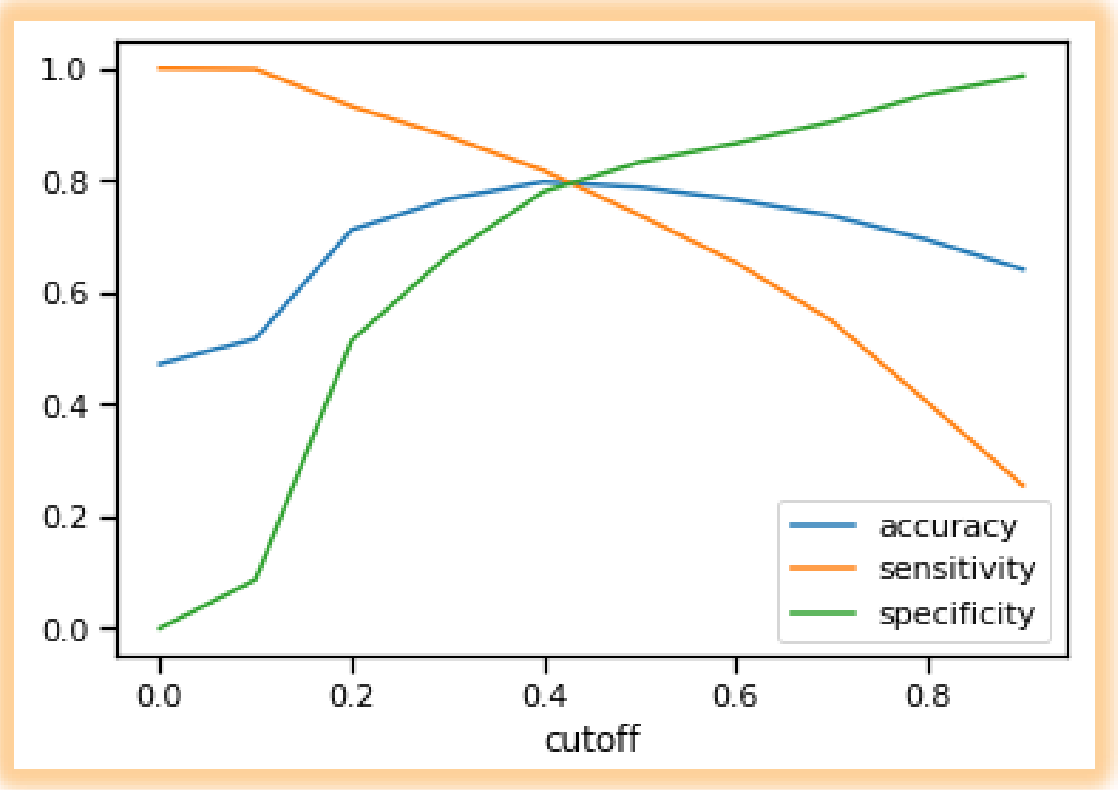
The ROC curve is obtained as below with area under the curve

AUC = 0.86



Optimal probability Cutoff

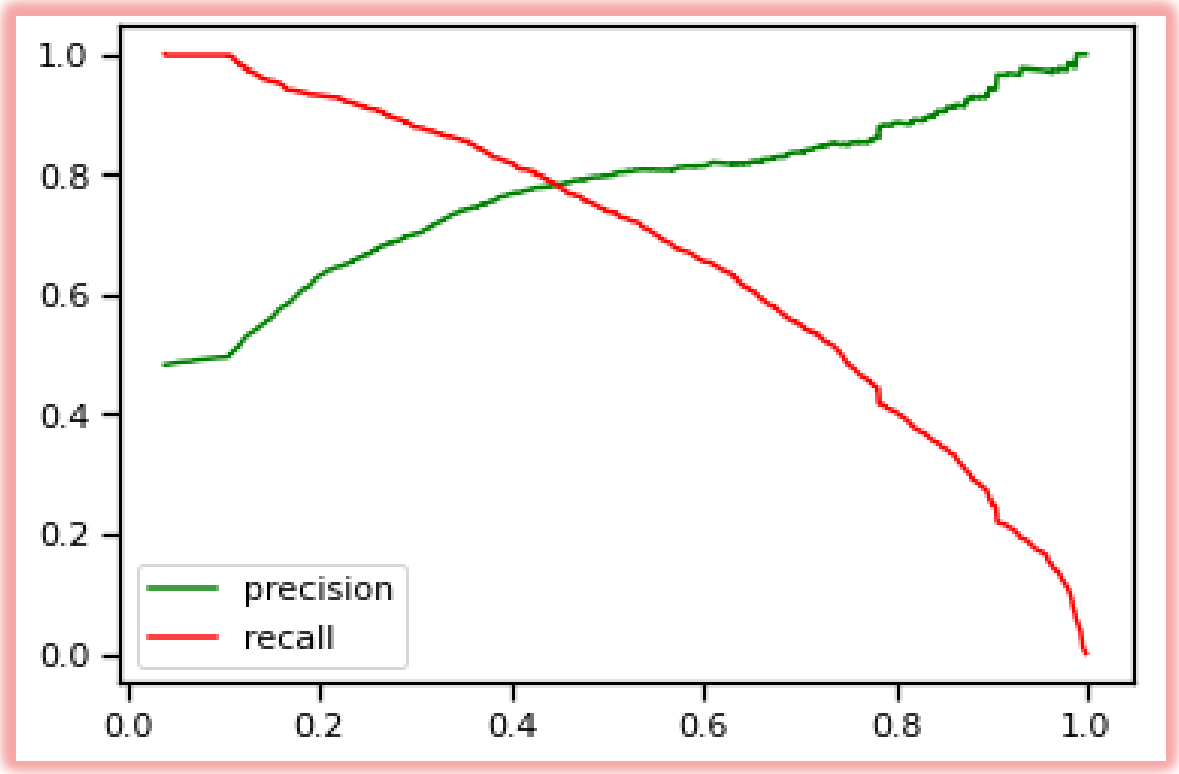
Sensitivity-Specificity Perspective: The optimal cutoff is identified by analyzing the plot of the accuracy, sensitivity and specificity for different cutoff values.



The optimal cutoff is identified to be 0.42

Optimal probability Cutoff

Precision-recall Perspective: Similar to sensitivity-specificity perspective, the optimal cutoff is identified as the balanced point from the lines of precision and recall



The optimal cutoff is here as well identified to be 0.42

Evaluation on optimal cutoff

Considering **0.42** as the optimal cutoff, the evaluation of model is performed on the below metrics using train data

- Accuracy : 80 %
- Sensitivity: 81 %
- Specificity: 79%
- False positive rate: 20%
- Positive predictive values: correctly predicted converted leads/ Total converted leads = 78%
- Negative predictive values: correctly predicted non-converted leads/ total non-converted leads = 81%

- Precision = 78%
- Recall = 81%

Evaluation on test data

The model is finally evaluated to classify the leads on the test data

- Accuracy : 77 %
- Sensitivity: 77 %
- Specificity: 78%
- False positive rate: 22%
- Positive predictive values: correctly predicted converted leads/ Total converted leads = 74%
- Negative predictive values: correctly predicted non-converted leads/ total non-converted leads = 80%

- Precision = 74%
- Recall = 77%

Conclusion

- ✓ The accuracy, sensitivity and specificity were showing promising results on both the train and test dataset
- ✓ The list of variables used for the final model has good business interpretation
- ✓ From the evaluation metrics analyzed, it can be concluded that the model is stable can be used to solve the business requirements with the new dataset as well.

The top 5 predominant features that contribute towards the lead getting converted are

1) **What is your current occupation_Working Professional**
2) **Last Activity_Had a Phone Conversation**
3) **Lead Source_Referral Sites**
4) **Total Time Spent on Website**
5) **Last Activity_SMS Sent**

To conclude, the company can emphasize to push the customer to spend more time on the websites, particularly working professionals by providing phone conversation or sending SMS to explain about their products. This focused marketing aids by increasing the number of converted leads with reduced effort from the sales team.

Thank You