Lead Score Case Study

Group Member

Praveen Raj

Karan Mehta

Problem Statement

- X education company sells online courses to industry professional
- X education's lead conversation rate is 30% i.e. from 100 leads in a day only 30 gets converted
- The company wants to build a model that will involve assigning lead score to the leads and wants to identify the 'Hot leads' i.e the potential leads with higher lead score
- Business Objectives:
  - X education wants to know the most promising leads- 'Hot Leads'
  - For which they want to build a model to identify the Hot Leads

Methodology

- Step1: Reading and Understanding the Data
  - Checking for general structure of the data
  - Check for data types
- Step 2: Data Cleaning and Preparation
  - missing values
  - Missing value treatment including dropping the missing the values
  - Check for outliers
- Step 3: Dummy variables Creation

Methodology

- Step 4: Splitting the data set into train and test set
- Step 5: Scaling for numerical variables and Correlation
- Step 6: Model Building: Classification technique: Logistic Regression
- Step7: Model Evaluation
- Step 8: Finding optimal Cuttoff
- Step9: Making Prediction on test set
  - Precision and Recall
- Conclusion and Recommendation

Step1: Reading and Understanding the Data

- Initially dataset has 9240 rows and 37 columns
- There are quite many categorical variables present in this dataset for which we will need to create dummy variables.
- There are a lot of null values present as well, which needs to be treated accordingly

Step 2: Data Cleaning and Preparation

- **Missing Value treatment**
  - Dropping ‘Lead Quality’ , ‘Asymmetrique Profile Score’, ‘Asymmetrique Activity Score’, ‘Asymmetrique Profile Index’, & ‘Asymmetrique Activity Index’ since there missing value is greater than 30%
  - Assuming that 'City', 'Country' columns will not add much insights in our analysis.Dropping those columns
  - How did you hear about X Education’, 'Lead Profile' have lot of rows as 'Select' in it which might be of giving no insights while building the model/Misleading the model to a new perception. Therefore dropping those variables
  - Dropping variables with only one response No i.e. Do Not Call, Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, Receive More Updates About Our Courses, Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque.

Step 2: Data Cleaning and Preparation

- Dropping 'What matters most to you in choosing a course’ since most response is ‘Better Career Prospects’
- Dropping rows with null values from variable What is your current occupation, TotalVisits, Lead Source, Specialization.
- ‘Prospect ID' & 'Lead Number' - These two columns are unique identifier for each records. Hence this won't add much insights to the analysis. Dropping those columns

Step 3: Dummy Variable Creation

Creating Dummy variable for categorical variables:

- 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

Dropping the original variables after creating the dummy variables

Step 4: Splitting the Data set into Training and Test set

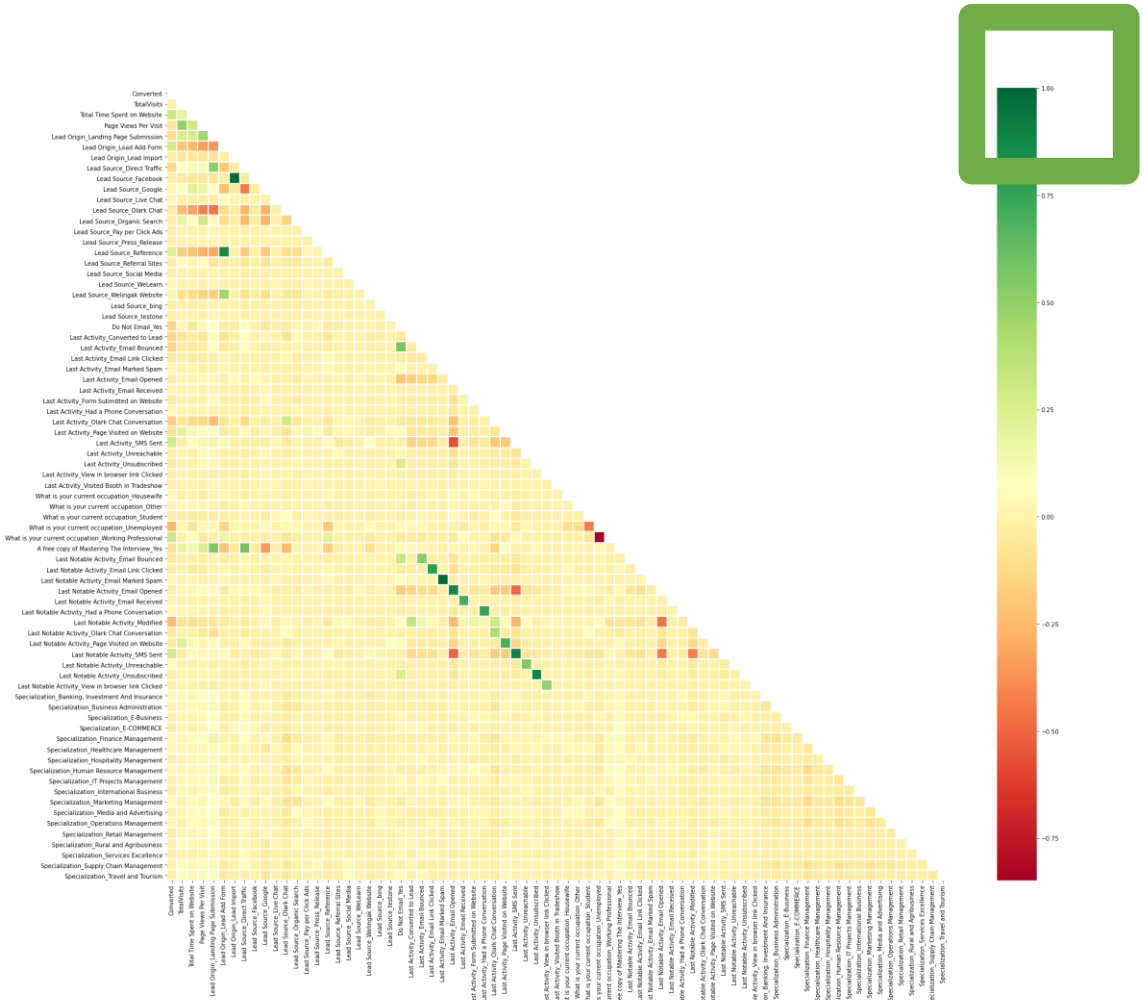
- Splitting the data set into training and test set with converted as target variable
- Splitting the data set into 70:30 ration for training set and test set

Step 5: Scaling of the data

- Scaling the numerical data with the help of minmax scaler
  - TotalVisits
  - Total Time Spent on Website
  - Page Views Per Visit

Step 6: Correlation Analysis

Checking for correlation among variables using corr and heatmap



Step6: Model Building

- Since there are 74 variables, we will use RFE first to get top 15 variables
- By using RFE we get Top 15 variables:
  - TotalVisits
  - Total Time Spent on Website
  - Lead Origin_Lead Add Form
  - Lead Source_Olark Chat
  - Lead Source_Reference
  - Lead Source_Welingak Website
  - Do Not Email_Yes
  - Last Activity_Had a Phone Conversation
  - Last Activity_SMS Sent
  - 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
  - Last Notable Activity_Had a Phone Conversation
  - Last Notable Activity_Unreachable

Step6: Model Building

- Using stat model to analyze the statistics part such that p-value and VIF
- Building the model

Dep. Variable:	Converted	**No. Observations:**	4461
Model:	GLM	**Df Residuals:**	4445
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2072.8
Date:	Tue, 07 Dec 2021	**Deviance:**	4145.5
Time:	20:51:23	**Pearson chi2:**	4.84e+03
No. Iterations:	22		
Covariance Type:	nonrobust		

Step 6: Model Building

- We can infer few variables have p-value greater than 0.05. Those needs to be handled. checking the VIF values

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0061	0.600	-1.677	0.094	-2.182	0.170
TotalVisits	11.3439	2.682	4.230	0.000	6.088	16.600
Total Time Spent on Website	4.4312	0.185	23.924	0.000	4.068	4.794
Lead Origin_Lead Add Form	2.9483	1.191	2.475	0.013	0.614	5.283
Lead Source_Olark Chat	1.4584	0.122	11.962	0.000	1.219	1.697
Lead Source_Reference	1.2994	1.214	1.070	0.285	-1.080	3.679
Lead Source_Welingak Website	3.4159	1.558	2.192	0.028	0.362	6.470
Do Not Email_Yes	-1.5053	0.193	-7.781	0.000	-1.884	-1.126
Last Activity_Had a Phone Conversation	1.0397	0.983	1.058	0.290	-0.887	2.966
Last Activity_SMS Sent	1.1827	0.082	14.362	0.000	1.021	1.344
What is your current occupation_Housewife	22.6492	2.45e+04	0.001	0.999	-4.8e+04	4.8e+04
What is your current occupation_Student	-1.1544	0.630	-1.831	0.067	-2.390	0.081
What is your current occupation_Unemployed	-1.3395	0.594	-2.254	0.024	-2.505	-0.175
What is your current occupation_Working Professional	1.2743	0.623	2.045	0.041	0.053	2.496
Last Notable Activity_Had a Phone Conversation	23.1932	2.08e+04	0.001	0.999	-4.08e+04	4.08e+04
Last Notable Activity_Unreachable	2.7868	0.807	3.453	0.001	1.205	4.369

Step 6: Model Building

- Checking the VIF
- *we can see that VIF values are in decent range except for 3 columns . Considering this , first going to drop the column 'Lead Source_Reference' since it has both high p-value and high VIF value*

Features	VIF	
2	Lead Origin_Lead Add Form	84.19
4	Lead Source_Reference	65.18
5	Lead Source_Welingak Website	20.03
11	What is your current occupation_Unemployed	3.65
7	Last Activity_Had a Phone Conversation	2.44
13	Last Notable Activity_Had a Phone Conversation	2.43
1	Total Time Spent on Website	2.38
0	TotalVisits	1.62
8	Last Activity_SMS Sent	1.59
12	What is your current occupation_Working Profes...	1.56
3	Lead Source_Olark Chat	1.44
6	Do Not Email_Yes	1.09
10	What is your current occupation_Student	1.09
9	What is your current occupation_Housewife	1.01
14	Last Notable Activity_Unreachable	1.01

Step6: Model Building

- Dropping the Lead score reference
- Building the model again

Dep. Variable:	Converted	**No. Observations:**	4461
Model:	GLM	**Df Residuals:**	4446
Model Family:	Binomial	**Df Model:**	14
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2073.2
Date:	Tue, 07 Dec 2021	**Deviance:**	4146.5
Time:	20:51:23	**Pearson chi2:**	4.82e+03
No. Iterations:	22		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0057	0.600	-1.677	0.094	-2.181	0.170
TotalVisits	11.3428	2.682	4.229	0.000	6.086	16.599
Total Time Spent on Website	4.4312	0.185	23.924	0.000	4.068	4.794
Lead Origin_Lead Add Form	4.2084	0.259	16.277	0.000	3.702	4.715
Lead Source_Olark Chat	1.4583	0.122	11.960	0.000	1.219	1.697
Lead Source_Welingak Website	2.1557	1.037	2.079	0.038	0.124	4.188
Do Not Email_Yes	-1.5036	0.193	-7.779	0.000	-1.882	-1.125
Last Activity_Had a Phone Conversation	1.0398	0.983	1.058	0.290	-0.887	2.966
Last Activity_SMS Sent	1.1827	0.082	14.362	0.000	1.021	1.344
What is your current occupation_House wife	22.6511	2.45e+04	0.001	0.999	-4.8e+04	4.8e+04
What is your current occupation_Stude nt	-1.1537	0.630	-1.830	0.067	-2.389	0.082
What is your current occupation_Unem ployed	-1.3401	0.594	-2.255	0.024	-2.505	-0.175
What is your current occupation_Worki ng Professional	1.2748	0.623	2.046	0.041	0.053	2.496
Last Notable Activity_Had a Phone Conversation	23.1934	2.08e+04	0.001	0.999	-4.08e+04	4.08e+04
Last Notable Activity_Unreacha ble	2.7872	0.807	3.454	0.001	1.205	4.369

Step 6: Model Building

- Repeating these steps
- The final VIF check

FEATURES	VIF
What is your current occupation_Unemployed	2.82
Total Time Spent on Website	2.00
TotalVisits	1.54
Last Activity_SMS Sent	1.51
Lead Origin_Lead Add Form	1.45
Lead Source_Olark Chat	1.33
Lead Source_Welingak Website	1.30
Do Not Email_Yes	1.08
What is your current occupation_Student	1.06
Last Activity_Had a Phone Conversation	1.01
Last Notable Activity_Unreachable	1.01

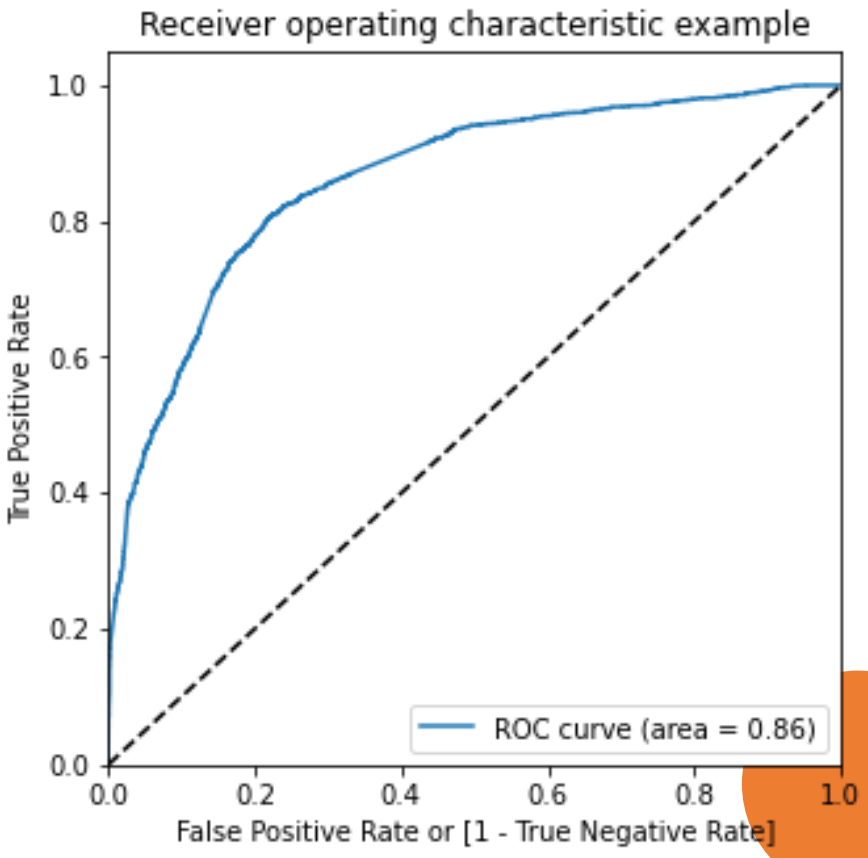
Step7: Model Evaluation

- Running the model on test set and predicting the probability on train set
- Creating a data frame with actual converted flag and predicted probabilities
- Assigning a threshhold values as 0.5 i.e if Conversion_Prob > 0.5 make 1 else 0
- Developing confusion matrix
- Accuracy: 78.8%
- Sensitivity: 73%
- Specificity: 83%

Converted	Conversion_Prob
0	0.300117
0	0.142002
1	0.127629
1	0.291558
1	0.954795

Step8: Finding the Optimal cut-off

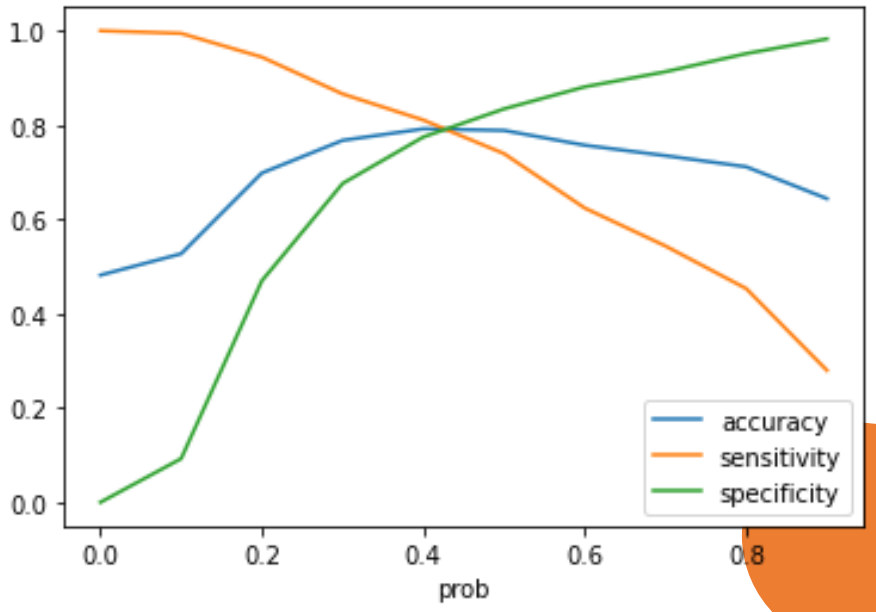
- *The area under ROC curve is 0.86 which seems to be a good model.*



Step8: Finding the Optimal cut-off

- *Plotting plot accuracy , sensitivity , specificity to get optimal cut-off value*

- *We get the optimal cutoff as .42 as all te three line merge at it*



Step8: Finding the Optimal cut-off

- Checking with .42 as cuttoff
  - Accuracy: 79%
  - Sensitivity: 79%
  - Specificity: 78.8%

Overall Accuracy, Sensitivity, Specificity seems to be good and almost similar around 80%

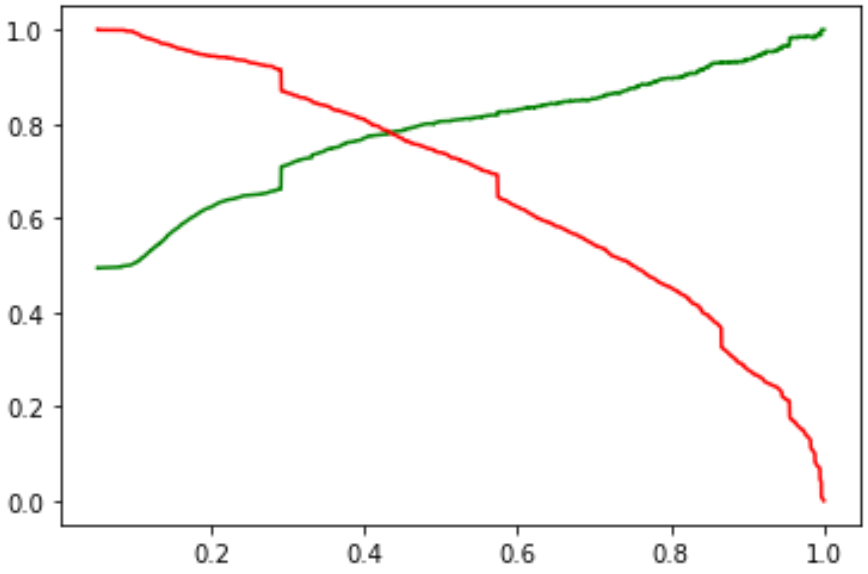
Step 9: Making Predictions on the Test Set

- Running the model on test set
- Using .42 as cut off, We can infer that Accuracy , Sensitivity , Specificity metrics for both train and test sets are almost same. Hence the model performs good with accuracy around 80%
- We can also infer that no Overfitting and Underfitting parameters is done

Step 9: Making Predictions on the Test Set

- *Precision and Recall trade off*

From the plot, we can infer that at Probability 0.44 both Precision and recall curve merges. Choosing this as Optimal Threshold



Step 9: Making Predictions on the Test Set

- By choosing the arbitrary cut-off as 0.44 , the Model evaluation metrics seems overall good around 79%
- We can infer that Accuracy , Precision , Recall metrics for both train and test sets are almost same. Hence the model performs good with accuracy around 78-80%
- We can infer that no Overfitting and Underfitting parameters is done

Summary statistics of Final Model

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 4461
Model:                            GLM   Df Residuals:                     4449
Model Family:                Binomial   Df Model:                           11
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2079.1
Date:                Tue, 07 Dec 2021   Deviance:                       4158.1
Time:                        20:51:25   Pearson chi2:                 4.80e+03
No. Iterations:                     7
Covariance Type:            nonrobust
==============================================================================================
                                                 coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                                          0.2040      0.196      1.043      0.297      -0.179       0.587
TotalVisits                                   11.1489      2.665      4.184      0.000       5.926      16.371
Total Time Spent on Website                    4.4223      0.185     23.899      0.000       4.060       4.785
Lead Origin_Lead Add Form                      4.2051      0.258     16.275      0.000       3.699       4.712
Lead Source_Olark Chat                         1.4526      0.122     11.934      0.000       1.214       1.691
Lead Source_Welingak Website                   2.1526      1.037      2.076      0.038       0.121       4.185
Do Not Email_Yes                              -1.5037      0.193     -7.774      0.000      -1.883      -1.125
Last Activity_Had a Phone Conversation         2.7552      0.802      3.438      0.001       1.184       4.326
Last Activity_SMS Sent                         1.1856      0.082     14.421      0.000       1.024       1.347
What is your current occupation_Student       -2.3578      0.281     -8.392      0.000      -2.908      -1.807
What is your current occupation_Unemployed    -2.5445      0.186    -13.699      0.000      -2.908      -2.180
Last Notable Activity_Unreachable              2.7846      0.807      3.449      0.001       1.202       4.367
==============================================================================================
```

Conclusion

- Features which contribute more towards the probability of a lead getting Converted are:
  - TotalVisits
  - Total Time Spent on Website
  - When the Lead Origin was Lead Add Form
  - When the Lead Source was
    - Olark chat
    - Welingak Website
  - When the Last Activity was
    - Phone Conversation
    - SMS sent
- After Obtaining the list of leads, We must inform them about new courses , offers ,services , job information and extension of higher studies to them
- Conducting Surveys to the leads will help us to determine their intention in joining the Online courses. This will help us in refining the approach better