Lead Scoring Case Study – Logistic Regression Problem

Summary

Problem Statement:

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their websites and browse for courses. There are a lot of leads generated in the initial stage but only a few of them come out as paying customers. The company needs to nurture the potential leads well (i.e., educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

- X Education has appointed you to help them select the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead scores have a higher conversion chance and the customers with lower lead scores have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Step1: Reading and Understanding Data: -

Read and analyse the data. (No of Rows and Columns, names of column)

Data Inspection likes Datatypes and data frame size.

```
xdf.info() #Providing information about dataframe.
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count  Dtype  
---  ------                                         --------------  -----  
 0   Prospect ID                                    9240 non-null   object 
 1   Lead Number                                    9240 non-null   int64  
 2   Lead Origin                                    9240 non-null   object 
 3   Lead Source                                    9204 non-null   object 
 4   Do Not Email                                   9240 non-null   object 
 5   Do Not Call                                    9240 non-null   object 
 6   Converted                                      9240 non-null   int64  
 7   TotalVisits                                    9103 non-null   float64
 8   Total Time Spent on Website                    9240 non-null   int64  
 9   Page Views Per Visit                           9103 non-null   float64
 10  Last Activity                                  9137 non-null   object 
 11  Country                                        6779 non-null   object 
 12  Specialization                                 7802 non-null   object 
 13  How did you hear about X Education             7033 non-null   object 
 14  What is your current occupation                6550 non-null   object 
 15  What matters most to you in choosing a course  6531 non-null   object 
 16  Search                                         9240 non-null   object 
 17  Magazine                                       9240 non-null   object 
 18  Newspaper Article                              9240 non-null   object 
 19  X Education Forums                             9240 non-null   object 
 20  Newspaper                                      9240 non-null   object 
 21  Digital Advertisement                          9240 non-null   object 
 22  Through Recommendations                        9240 non-null   object 
 23  Receive More Updates About Our Courses         9240 non-null   object 
 24  Tags                                           5887 non-null   object 
 25  Lead Quality                                   4473 non-null   object 
 26  Update me on Supply Chain Content              9240 non-null   object 
 27  Get updates on DM Content                      9240 non-null   object 
 28  Lead Profile                                   6531 non-null   object 
 29  City                                           7820 non-null   object 
 30  Asymmetrique Activity Index                    5022 non-null   object 
 31  Asymmetrique Profile Index                     5022 non-null   object 
 32  Asymmetrique Activity Score                    5022 non-null   float64
 33  Asymmetrique Profile Score                     5022 non-null   float64
 34  I agree to pay the amount through cheque       9240 non-null   object 
 35  A free copy of Mastering The Interview         9240 non-null   object 
 36  Last Notable Activity                          9240 non-null   object 
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB
```

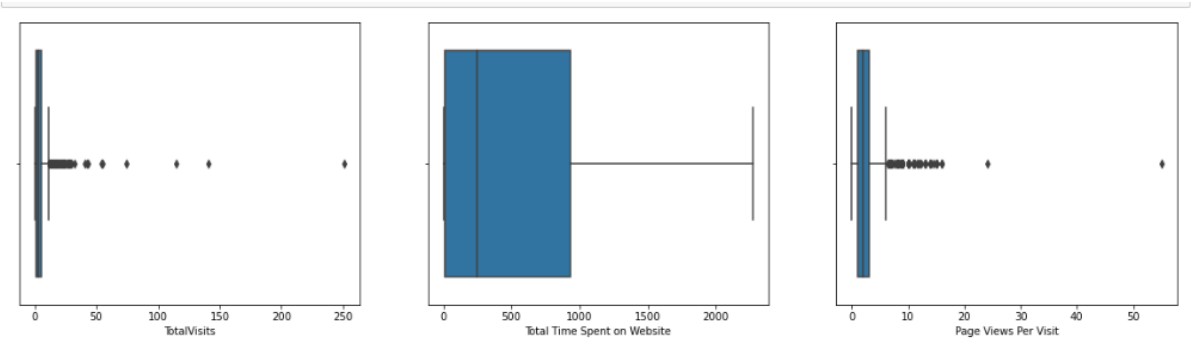
Step2: Data Cleaning: -

Checking Missing Values/Null values

```
null_valuesp.sort_values(ascending=False)
```

```
How did you hear about X Education               78.463203
Lead Profile                                     74.188312
Lead Quality                                     51.590909
Asymmetrique Profile Score                       45.649351
Asymmetrique Activity Score                      45.649351
Asymmetrique Activity Index                      45.649351
Asymmetrique Profile Index                       45.649351
City                                             39.707792
Specialization                                   36.580087
Tags                                             36.287879
What matters most to you in choosing a course    29.318182
What is your current occupation                  29.112554
Country                                          26.634199
Page Views Per Visit                              1.482684
TotalVisits                                       1.482684
Last Activity                                     1.114719
Lead Source                                       0.389610
Receive More Updates About Our Courses            0.000000
I agree to pay the amount through cheque          0.000000
Get updates on DM Content                         0.000000
Update me on Supply Chain Content                 0.000000
A free copy of Mastering The Interview            0.000000
Prospect ID                                       0.000000
Newspaper Article                                 0.000000
Through Recommendations                           0.000000
Digital Advertisement                             0.000000
Newspaper                                         0.000000
X Education Forums                                0.000000
Lead Number                                       0.000000
Magazine                                          0.000000
Search                                            0.000000
Total Time Spent on Website                       0.000000
Converted                                         0.000000
Do Not Call                                       0.000000
Do Not Email                                      0.000000
Lead Origin                                       0.000000
Last Notable Activity                             0.000000
dtype: float64
```

Checking Outliers in features-



We have dropped the variables that had a high percentage of NULL values in them.

Also imputing the missing values as and where required with median values in case of numerical variables.

Creation of new classification variables in case of categorical variables.

The outliers were identified and removed based on the Interquartile distance for some of the continuous variables.

Step3: Performing Exploratory Data Analysis: -

we started with the Exploratory Data Analysis of the data set to get a feel of how the data is

oriented and representing visuals.

We have analysed all features and visualised to check distribution also consolidated variables which represent less count or contributions.

Also, write observations according to feature analysis in comparison of the target variable

Step4: Data Preparation & Transformation: -

Converting some binary variables (Yes/No) to 0/1

Creating Dummy Variables for the categorical variables (one-hot encoded).

Step5: Data Splitting (Train - Test Split): -

we will split the data into 2 parts: -

Train Data (On which model will be built and is 70% of total data)

Test Data (On which build model will be tested and is 30% of total data)

Step6: Feature Rescaling

It is good to have all the variables on the same scale for the model to be easily interpretable.

We can use standardization or normalization so that the units of the coefficients obtained are all on the same scale.

There are two ways we can do rescaling:

- Min-Max scaling (Normalisation): Between 0 and 1
- Standardisation: mean-0, sigma-

We used the Min Max Scaling to scale the original numerical variables.

```
#After standardising numerical variables
xdf1_train[numerical_features_train].head()
```

	TotalVisits	Total Time Spent on Website	Page Views Per Visit
5182	0.090909	0.024412	0.166667
8469	0.181818	0.082113	0.333333
8382	0.272727	0.619174	0.250000
8031	0.272727	0.191300	0.500000
6712	0.272727	0.923657	0.333333

Step7: Feature selection using RFE (Recursive Feature Elimination): -

Using the Recursive Feature Elimination, we went ahead and selected the 20 top important features.

Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.

Finally, we arrived at the 15 most significant variables. The VIFs for these variables were also found to be good.

We then created the data frame having the converted probability values and we had an initial assumption that a probability value of more than 0.5 means 1 else 0.

Based on the above assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model.

We also calculated the 'Sensitivity' and the 'Specificity' matrices to understand how reliable the model is.

Step8: Plotting the ROC Curve - Optimise Cut off: -

We then tried plotting the ROC curve for the features and the curve came out to be pretty decent with an area under the curve of 96% which further solidified of the model.

Step9: Finding the Optimal Threshold - Cut off: -

Then we plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values.

The intersecting point of the graphs was considered the optimal probability cut-off point. The cut-off point was found to be 0.32.

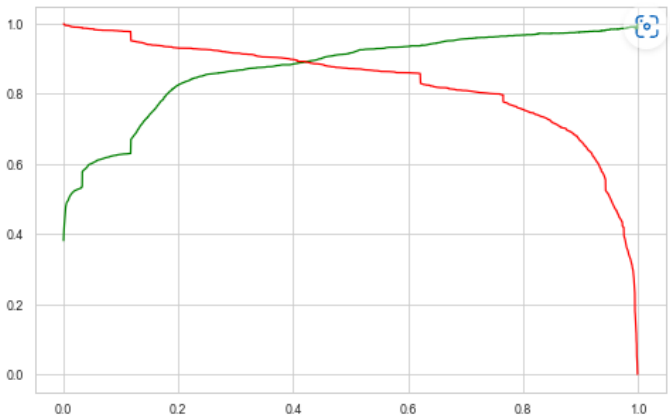
We could also observe the new values of the 'accuracy=91.55%, 'sensitivity=91.33%', and 'specificity=91.69%'.

Also calculated the lead score and figured that the final predicted variables approximately gave a target lead prediction of 80%.

Step10: Computing the Precision and Recall: -

we also found out the Precision and Recall metrics values came out to be 87.16% and 91.33% respectively on the train data set.

Based on the Precision and Recall trade-off, we got a cut-off value of approximately 0.41.



Step11: Making Predictions on Test Set: -

We have applied our final model to the Test dataset after doing feature selection & standardisation like the training dataset.

Also Applied a Probability threshold value of 0.32 on the test dataset to predict the lead score.

calculated the conversion probability based on the Sensitivity and Specificity metrics and found the accuracy value to be 91.78%; Sensitivity=90.81%; Specificity= 92.39%.

Classification report of test dataset: -

	precision	recall	f1-score	support
0	0.94	0.92	0.93	1591
1	0.88	0.91	0.90	1013
accuracy			0.92	2604
macro avg	0.91	0.92	0.91	2604
weighted avg	0.92	0.92	0.92	2604

Sensitivity of the prediction over test data set is 90%

Step12: Conclusion: -

Last, we have concluded all points and features are important for solving the business problem.