X Education, an online course provider, faces challenges with its lead conversion rate, which currently stands at 30%. The company wishes to improve the efficiency of its lead conversion process by identifying the most potential leads, referred to as Hot Leads, by assigning a score to each lead. The target is to increase the lead conversion rate to around 80%, a significant improvement from the current rate.

Several steps were performed, as outlined below (for details, refer to code comments).

Step 1: **Reading and Observing**

- Corrected the datatype of incorrect datatypes
- Identified all missing and null values
- There were columns having select as class. Relabelled then as Missing.
- Dropped redundant data such as Prospect ID & Lead Number

```
#   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   Lead Origin                                   9240 non-null   object
 1   Lead Source                                   9240 non-null   object
 2   Do Not Email                                  9240 non-null   object
 3   Do Not Call                                   9240 non-null   object
 4   Converted                                     9240 non-null   int64
 5   TotalVisits                                   9103 non-null   float64
 6   Total Time Spent on Website                   9240 non-null   int64
 7   Page Views Per Visit                          9103 non-null   float64
 8   Last Activity                                 9240 non-null   object
 9   Country                                       9240 non-null   object
 10  Specialization                                9240 non-null   object
 11  How did you hear about X Education            9240 non-null   object
 12  What is your current occupation               9240 non-null   object
 13  What matters most to you in choosing a course 9240 non-null   object
 14  Search                                        9240 non-null   object
 15  Magazine                                      9240 non-null   object
 16  Newspaper Article                             9240 non-null   object
 17  X Education Forums                            9240 non-null   object
 18  Newspaper                                     9240 non-null   object
 19  Digital Advertisement                         9240 non-null   object
 20  Through Recommendations                       9240 non-null   object
 21  Receive More Updates About Our Courses        9240 non-null   object
 22  Tags                                          9240 non-null   object
 23  Lead Quality                                  9240 non-null   object
 24  Update me on Supply Chain Content             9240 non-null   object
 25  Get updates on DM Content                     9240 non-null   object
 26  Lead Profile                                  9240 non-null   object
 27  City                                          9240 non-null   object
 28  Asymmetrique Activity Index                   9240 non-null   object
 29  Asymmetrique Profile Index                    9240 non-null   object
 30  Asymmetrique Activity Score                   5022 non-null   float64
 31  Asymmetrique Profile Score                    5022 non-null   float64
 32  I agree to pay the amount through cheque      9240 non-null   object
 33  A free copy of Mastering The Interview        9240 non-null   object
 34  Last Notable Activity                         9240 non-null   object
```

Step 2: Data Cleaning & EDA

- Variable are split into categorical and Numerical groups based on datatype
- Variables such as *Asymmetrique Activity Index,Asymmetrique Profile Index,Asymmetrique Activity Score, Asymmetrique Profile Score* with more than 40% missing are dropped
- Identified all 0 variance variables and dropped them
- Merged various classes with very low counts into a single class for various variables
- Categorical data is analysed using univariate and Bi variate analysis and imputing is done based on conditions
- All "Yes" and "No" values are converted into 1s and 0s
- For numerical variables, outliers identified and their treatment is done
- Missing value are imputed with median if variables are there and with mean if no outlier is present.

Step 3: Encoding and standardizing

- All the categorical data is converted into dummy variables.
- All numerical values are standardised using StandardScaler()
- Correlation checks and highly corelated are dropped.

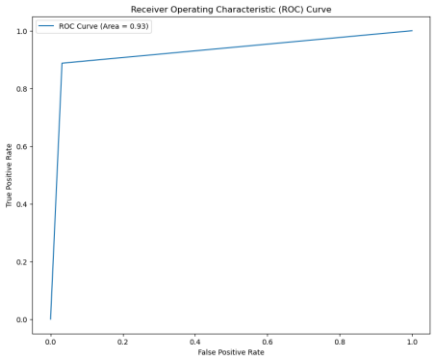
Step 4: Modeling

- Basic model i.e., Minimum viable product (MVP) is trained and Evaluated on training dataset
- Multiple models are trained using different variables to introduce complexity based on automated technique RFE and VIF.
- To further reduce variables, manual feature dropping is performed using the model summary.
- Final model summary

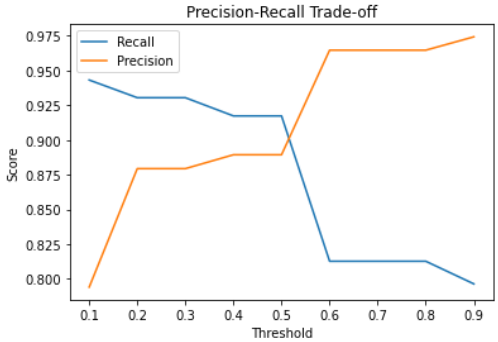
Features	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.4192	0.156	-28.365	0.000	-4.725	-4.114
Lead Source_Welingak Website	3.5935	1.018	3.530	0.000	1.598	5.589
Tags_Busy	3.6010	0.252	14.262	0.000	3.106	4.096
Tags_Closed by Horizzon	9.9022	1.014	9.764	0.000	7.915	11.890
Tags_Lost to EINS	10.6054	1.030	10.296	0.000	8.587	12.624
Tags_Missing	4.4575	0.186	23.947	0.000	4.093	4.822
Tags_Will revert after reading the email	7.2919	0.216	33.698	0.000	6.868	7.716
Tags_switched off	-1.7262	0.728	-2.370	0.018	-3.154	-0.298
Lead Profile_Missing	-2.4753	0.138	-17.993	0.000	-2.745	-2.206
Last Notable Activity_SMS Sent	2.5453	0.124	20.478	0.000	2.302	2.789

Step 5: Evaluation and Threshold calculation

- ROC curve is plotted using the model to know evaluate the performance compared to the MVP. AUC is .91



- Accuracy, recall and precision score are compared with MVP score
- Threshold value is calculated using the Precision-Recall curve.



  - Threshold value come as 0.51

Step 6: Evaluation on Test dataset

- Using threshold prediction are adjusted and the final evaluation is done on the training dataset
- Accuracy, recall, precision is calculated and evaluated against the score of training set.
- True positive rate, false negative which are important in this are calculated.
- Model evaluations are as follows

Training Set	Evaluation Metrics	Test set
.92	Recall Score	.93
.89	Precision Score	.88
.92	Accuracy	.92
.93	Specificity	.92
.08	False Negative rate	.07

Summary

- The model provide Score to each lead on a scale of 0 to 100 higher the number more is the probability of getting into converted.
- Any lead with a Lead score more than 51(threshold) can be considered as Hot Lead.
- The model has an accuracy of 92% and an average false positive rate of 7.5%.