Lead Scoring Case Study Summary

- Business Problem:

X Education an education company that sells online courses to industry professionals, need to select promising leads to increase their target lead conversion rate to 80% in an effective manner

- Dataset provided for Analysis: Leads.csv

- Approach Taken:

Step 1] Data Exploration:

The given dataset contains 9240 records with 37 attributes of 30 Object type, 4 Float64 type, 2 int64 type, and the current conversion rate is 38.45%
There are columns with null values, few have a default word Select as an entry, few have a single value, and few columns are highly skewed

Step 2] Data Cleaning:

Step	Columns	Action
” Select” Value	1] Specialization 2] How did you hear about X Education 3] Lead Profile 4] City	Replace with Null Value
Null Value > 40%	1] Lead Quality 2] Lead Profile 3] Asymmetrique Activity Index 4] Asymmetrique Profile Index 5] Asymmetrique Activity Score 6] Asymmetrique Profile Score	Drop columns
Skewed Columns	1] Country 2] What matters most to you in choosing a course 3]Do Not Call 4] Search 5] Newspaper Article 6] X Education Forums 7] Newspaper 8] Digital Advertisement 9] Through Recommendations	Drop Columns
Columns with Single Value	1] Magazine 2] Receive More Updates About Our Courses 3] Update me on Supply Chain Content 4] Get updates on DM Content 5] I agree to pay the amount through cheque	Drop Columns

Step 3] EDA :

Numerical Columns
Upon performing EDA on Numerical columns TotalVisits, Total Time spent on Website, Page Views Per Visit we observe that TotalVisits, Page Views Per Visit contain outliers, these can be capped using the 99%
Categorical Columns
Columns with categories with less overall percent can be grouped as Others
1] Lead Source
2] Specialization
3] What is your current occupation
4] City
5] A free copy of Mastering The Interview
6] Last Notable Activity

Step 4] Data Preparation

Columns with string binary values can be replaced with 1 or 0
And for categorical columns with more than 2 categories dummy variables can be created
Once done data is split into Train and Test Dataset
And feature scaling is performed using Standardisation

Step 5] Model Building

A logistic regression model can be built on Train data and view the P-Values and VIF values of the attributes and drop the ones with higher P-value and keep iterating the process until the P-value of attributes present is almost 0 and they have less VIF value use the RFE method for this and access the model using the stats model
Once done calculate the confusion matrix and the accuracy of the model and calculate the Sensitivity, Specificity metric to weigh the effectiveness of the model
Plot out a ROC Curve for the same

Step 6] Finding the optimal Cut-off

Calculate at what probability are we getting the most optimal accuracy, sensitivity, and specificity.
Finally use the most optimal model and run it on the Test set and calculate all the metrics

- Challenges Faced:

  - Handling a lot of categorical data
  - Tuning the model to obtain a good sensitivity score
  - Obtaining a final stable model with the variable set provided

- Overall Inferences:

Measure	Train Dataset	Test Dataset
Accuracy	77.5%	77.1%
Sensitivity	78.1%	78%
Specificity	77.1%	77%
Precision	68%	69%
Recall	78%	78%

By following the above steps, a model was built that assigns a lead score to each prospect. The Sales team can run that model on new prospects and check if the score is greater than 30. If yes, then the prospect will have a 78% chance of a conversion. This will save a lot of time for the sales team as they can focus on leads that have a good chance of conversion.