Lead Score Case Study – Summary Report

The Customer (X Education Company) Lead Score Case Study's Summary Steps are as follows:

Step 1: Read and Understand Dataset:

- ✓ After reading leads.csv dataset it was understood that dataset contains 9240 rows and 37 columns

Step 2: Data cleaning:

- ✓ The column variables which had missing values greater than or equal to 40% (i.e., >=40%) had been dropped.
- ✓ Missing values are imputed.
- ✓ Highly Skewed variables/columns are dropped.
- ✓ Percentage of null values in the variables are verified.
- ✓ Finally, we eliminated the skewed, unique identifier variables and cleaned null values from all variables.

At the end of data cleaning, we had 11 variables.

Step 3: Exploratory Data Analysis:

- ✓ For outlier analysis box plots are plotted.
- ✓ Using soft capping method removed outliers.
- ✓ For categorical vs converted column count plots are plotted.

Step 4: Data Pre-Processing:

- ✓ Dummy variables are created.
- ✓ Identified 'Yes or 'No' values columns and converted them to 1's and 0's.

Train-Test Split

- ✓ All independent variables were assigned to X, while the dependent variable (Converted) was assigned to y.
- ✓ With the train set at 70% and the test set at 30%, and the random state at 100%, a train-test split was done.

Scaling

- ✓ Using standard scaler scaled the training data

Looking at correlation

- ✓ After Plotting the heat map dropped the columns with high correlation (>0.7).

Step 5: Model Building:

- ✓ Total 15 variables were considered for RFE and the coarse tuning was done as part of feature selection.
- ✓ Using p-value and VIF done manual tuning and right features for the model were selected.
- ✓ Metrics (Accuracy, sensitivity, specificity, precision, recall, false positive rate) are calculated.

Step 6: Plotting ROC Curve:

- ✓ Area under curve is 0.86 after plotting ROC.

Step 7: Finding optimal probability cut-off:

- ✓ Based on the graph, 0.34 would be the optimal cut-off point. As a result, leads with a probability conversion of greater than 34% are considered promising.
- ✓ Between Precision and Recall plotted Trade-off
- ✓ Calculated Metrics.

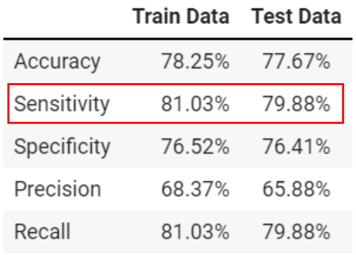
Step 8: Model Evaluation and Model Performance:

- ✓ Performed predictions on test set.
- ✓ Calculated Metrics.

Step 9: Lead score calculation:

- ✓ For entire dataset lead scores are calculated.

Step 10: Final Observation:



	Train Data	Test Data
Accuracy	78.25%	77.67%
Sensitivity	81.03%	79.88%
Specificity	76.52%	76.41%
Precision	68.37%	65.88%
Recall	81.03%	79.88%

Recommendations:

Major indicators that a lead will get converted to a promising lead:

1. **Lead_Source_Welingak website -** A lead generated through the Welingak website is more likely to convert.
2. **Lead_Source_Reference -** A lead that has been referred by a prior client has a higher chance of being converted.
3. **Current Occupation_Working Professional -** Professionals in the workplace are more likely to convert.

Major indicators that a lead will NOT be converted:

1. **Do Not Email -** A lead who has selected 'Do Not Email' is less likely to become a paying customer.

Conclusion:

- ✓ **The model's sensitivity score is around 81 percent, which is close to the CEO's aim of 80 percent.**