Lead Scoring Case Study Summary

Submitted by – ANKIT DATTA & RHITHIK PR.

Problem Statement:

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers.

The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution Summary:

Step 1: - Data Reading and Understanding:

- Read and inspected the data.

Step 2: - Data Cleaning:

- Dropped variables with unique values.
- Replaced 'Select' values with Null values.
- Dropped columns with NULL values exceeding 30%.
- Removed imbalanced and redundant variables, imputed missing values, handled outliers, and corrected case inconsistencies in labels.
- Excluded sales team-generated variables.

Step 3: - Data Transformation:

- Converted binary variables into '0' and '1'.

Step 4: - Dummy Variables Creation:

- Created dummy variables for categorical variables.
- Removed repeated and redundant variables.

Step 5: - Test-Train Split:

- Divided the dataset into test and train sections (70-30%).

Step 6: - Feature Rescaling:

- Used Min Max Scaling for numerical variables.
- Checked correlations among variables using a heatmap.
- Dropped highly correlated dummy variables.

Step 7: - Model Building:

- Utilized Recursive Feature Elimination to select the top 15 important features.
- Recursively examined P-values to select significant variables and dropped insignificant ones.
- Arrived at 11 most significant variables with good VIFs.
- Determined optimal probability cutoff, plotted ROC curve (86% area coverage), and checked model accuracy, sensitivity, and specificity.
- Verified if 80% cases are correctly predicted based on the converted column.
- Assessed precision, recall, accuracy, sensitivity, and specificity for the final model on the train set.
- Established a cutoff value based on Precision and Recall trade-off (approximately 0.3).
- Implemented learnings on the test model, calculated conversion probability, and found accuracy (77.52%), sensitivity (83.01%), and specificity (74.13%).

Step 8: - Conclusion:

- The lead score in the test set meets the CEO's expectation of a target lead conversion rate of around 80%.
- The model's good sensitivity aids in selecting promising leads.
- Features contributing more to the conversion probability include Lead Origin_Lead Add Form, what is your current occupation_Working Professional, and Total Time Spent on Website.