Lead Scoring Case Study Summary

Problem Statement:

An education company named X Education sells online courses to industry professionals. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

X Education company requires to build a model wherein a lead score has to be assigned to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a 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 Methodology:

Data Inspection: Reading and Understanding Data

Data Cleaning:

- Checked the unique values in each Categorical column
- Dropped categorical columns with only one value
- Replaced values of columns having 'select' has one of the categories with nan
- Checked the Null value percentage in columns and dropping columns with NULL values >40%
- Created a separate category level 'Others' for nan values
- Grouped all low frequency values to Others
- Checked for the column outliers and Removing Outliers which are in extreme 5% range of the Column values.
- Dropped Columns that will not add any value to the model or are having multiple correlated columns

Exploratory Data Analysis

- By Univariate Analysis & Bivariate Analysis for Numerical Variables, plotted histogram for categorical variables to get data distribution
- Dropped columns with highly skewed data
- Checked Data Imbalance Ratio
- Plotted Heatmap to show correlation between numerical variables

Data Preparation

- Renamed column names for readability before Dummy Creation
- Mapped binary categorical variables to 1 & 0

Dummy Variable Creation

- Creating dummy variables for categorical variables and dropping the original column.

Test Train Split:

- Assigned feature variable to X & response variable to y
- Split the dataset into train and test in the ratio of 7:3

Feature Scaling

- Scaled the numerical features using Standard Scaling

Feature selection using RFE i.e. Recursive Feature Elimination:

- Selected 15 top features which will be considered in model building

Model Building

- Using the 15 top features select by RFE and logistic regression model, developed the model and fine-tuned by recursively looking for p-values and VIF values and dropped the insignificant features.

Model Evaluation

- Checked Accuracy score to measure the performance of the model, accuracy being the ratio of Total correct instances to the total instances and it came out to be 82.7% for train dataset.
- Created Confusion matrix and checking Sensitivity, Specificity, Precision, Recall and found out Sensitivity=68.9%; Specificity= 90.32%

Plotted the ROC Curve

- It shows the trade-off between sensitivity and specificity.
- ROC curve Area under the curve came out to be 90%

Finding Optimal Cut-off Point

- Optimal cut-off probability is that prob where we get balanced sensitivity and specificity. From the curve we got 0.34 as the optimum point to take it as a cut-off probability.
- Using this optimal point, we found accuracy=81.72%, Sensitivity=79.9% and Specificity= 82.15%.

Precision and Recall & trade-off

- The precision-recall trade-off value was approximately 0.4 with a balance between precision and recall. But Sensitivity, Recall values were 76% approximately but the Business has asked for 80%. Using Sensitivity -Specificity threshold of 0.34 we could get above 80% metrics.

Making predictions on the Test dataset

- Scaled the numerical features of the test dataset using Standard Scaling
- Predicted using final model on Test Dataset
- Checked ROC Curve for test dataset
- Model Evaluation on Test Dataset by accuracy score and created Confusion matrix, Precision, Recall
- Evaluated test dataset based on the Sensitivity and Specificity metrics and found out the accuracy Score=79.8%, Sensitivity=75.99%; Specificity= 82.15%.

Adding Lead Score column to Test dataframe