Lead Scoring Case Study Summary

Problem Statement:

An education company named X Education sells online courses to industry professionals.

The typical lead conversion rate at X education is very poor, only around 30%.

To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.

Business Objective:

X education wants to know most promising leads.

Build a model that assign lead score to every lead and thus help in identifying 'Hot Leads'.

Deployment for future use.

Solution Approach:

Step 1 - Source Data & Data Cleanup:

Read and analyse the data.

We then started data cleanup by dropping columns with high percentage of missing values. We then checked the data imbalance for columns then dropped the columns with high data unbalance.

After this we started with Imputing missing values in Columns with most suited values.

Step 2 - Exploratory Data Analysis:

We then plotted Bivariate plots for the columns / variable with our Target variable "Converted" to find any concurrent relation. We then creating

Dummy variables for Categorical variables and finally Checked for Outliers in the data set and removed them.

Step 3 - Splitting Data into Test-Train Datasets:

We then split the data into Test-Train data sets in the ratio of 7:3 to help us build the model on the training set and finally test the accuracy of the model on the test data set.

Step 4 - Feature Scaling:

We then normalized the range of independent variables using the standard scaling method.

Step 5 - Model Building:

We then started our model building by running the Logistic Regression model on training dataset. Then we performed feature selection using RFE. Then we removed the variables with high p-value and VIF values and iterated the process until all the process has p-value less than 0.05 and Variance Inflation factor values of all the predictor variables are less than 5 indicating no multicollinearity issues.

Step 6 - Model Evaluation:

We then perform model evaluation using Optimal Cutoff Point and Precision and Recall Trade-off.

- Optimal Cutoff point - We observed that 0.3 is the optimum probability cut-off point.
- Precision and Recall trade-off

Step 7 - Applying the best model in Test data:

Lastly, we ran the model on Test dataset to make predictions. The following observations were observed:

- Accuracy = 0.82

- Recall = 0.86
- Precision = 0.71
- Sensitivity of test data = 0.86
- Specificity of test data = 0.799

Indicating a good accurate model.

Conclusion:

- The model performs consistently well on both the train and test datasets, with accuracies around 0.82-0.83, it demonstrates a strong overall predictive capability. Sensitivity and specificity scores above 0.80 reflect the model's ability to effectively identify both positive and negative cases.
- The top three variables in the model that contribute most to the probability of a lead getting converted are Tags_Closed by Horizzon, Tags_Lost to EINS, and Tags_Will revert after reading the email.
- These variables indicates varying degrees of readiness and potential for conversion, influencing the overall likelihood of a lead becoming a customer.