Lead Scoring Case Study Report

Creating a logistic regression model to forecast the likelihood of a lead being successfully converted into an online course for the educational firm X Education.

Business Objective

- ln order to assist X Education in choosing the most promising leads, or "Hot Leads," which are leads with the highest likelihood of becoming paying clients.
- Each lead must be given a score between 0 and 100 so that the business may use it to target potential prospects and create a logistic model.
- To deal with this, a logistic regression model that forecasts the lead conversion odds for every lead can be created.
- The chance of a lead being converted is determined by the threshold value, below which it is not expected to be converted and above which it will be.
- The lead score value for every lead can be obtained by multiplying the lead conversion probability.

Problem Solving

The approach for this project has been to divide the entire case study into various checkpoints to meet each of the sub-goals.

1. Understanding the dataset and data preparation.
2. Applying recursive feature eliminating to identify the best performing feature for building the model.
3. Building the model using RFE. Eliminate all the features with high p-value and VIFs and finalizing the model.
4. Perform various evaluation with various metrics like sensitivity, specificity, precision and recall.
5. Plotting the ROC curve to find the optimal cutoff and deciding the probability threshold to predict the dependent variable for the training dataset. Using the model on test dataset and perform model evaluation for the test set.

Data Preparation and Feature Engineering

- Eliminate any columns with a single unique value. Take "Do Not Call," "Search," "Magazine," "Newspaper," and so forth as examples.
- Eliminating the rows with a high percentage of missing "SELECT" values in a specific column. Columns with a significant percentage of missing information include "Total Visits," "Lead Source," and "Specialization."
- Applying the median and mode to the null data.
- Creating a Univariate analysis on categorical and numerical columns for analyzing and handling the outliers present in the dataset.

- Constructing dummy variables for features that are categorical.
- We apply the MinMaxScaler function on the columns labeled "Total Visits," "Total Time Spent on Website," and "Page Views Per Visit."
- Examined the dataset's correlations using a heatmap and in the form of table.
- To create a more accurate model, we first constructed a model using train and test split and then removed the columns with high p-values and VIF.
- Following that, a ROC curve is plotted to determine the ideal cutoff, and the projected probability and confusion metrix are then displayed using that threshold.