Lead Scoring Case Study Summary

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%. The task is to build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

Summary of Steps followed:

Step 1: Reading and Understanding Data

Read and analyse the data after importing the necessary libraries, check for data imbalance, statistical summary, information and so on.

Step 2: Data Cleaning

First, the null values were checked. Then, we dropped the variables that had high percentage (35%) of NULL values in them. This step also included imputing the missing values as and where required with median values in case of numerical variables and creation of new classification variables in case of categorical variables. The outliers were identified and removed.

Step 3: Exploratory Data Analysis

EDA of the data was done to see how the data is oriented. During this stage, approximately three variables were found with only one value across all rows. These variables were subsequently eliminated.

Step 4: Creating Dummy Variables

We proceeded by generating dummy variables for the categorical variables.

Step 5: Test Train Split:

The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

Step 6: Feature Rescaling

We used Min-Max Scaling to scale the original numerical variables. Subsequently, we utilized the stats model to construct our initial model, providing us with a comprehensive statistical overview of all model parameters.

Step 7: Feature selection using RFE:

Using Recursive Feature Elimination, we proceeded to select the top 20 important features. Utilizing the generated statistics, we recursively examined the P-values to identify the most significant values, retaining them while dropping the insignificant ones. Eventually, we identified the 15 most significant variables. The Variance Inflation Factors (VIFs) for these variables were also deemed satisfactory. Subsequently, we constructed a dataframe containing the converted probability values, assuming that a probability value exceeding 0.5 corresponds to 1, otherwise 0. Based on this assumption, we derived the Confusion Matrix and calculated the overall Accuracy of the model. Additionally, we computed the 'Sensitivity' and 'Specificity' metrics to gauge the reliability of the model.

Step 8: Plotting the ROC Curve

We proceeded to plot the ROC curve for the features, and the resulting curve demonstrated a commendable area coverage of 89%, further enhancing the credibility of the model.

Step 9: Finding the Optimal Cutoff Point

Next, we generated probability graphs for 'Accuracy', 'Sensitivity', and 'Specificity' across different probability values. The point where these graphs intersected was identified as the optimal probability cutoff point, which was determined to be 0.42. With this new cutoff value, we observed that nearly 80% of values were accurately predicted by the model. Furthermore, we calculated updated metrics: 'accuracy=81%', 'sensitivity=79.8%', and 'specificity=81.9%'. Additionally, we calculated the lead score and determined that the final predicted variables yielded an approximate target lead prediction of 80%.

Step 10: Computing the Precision and Recall metrics

Additionally, we calculated the Precision and Recall metrics, which yielded values of 79% and 70.5% respectively on the train dataset. Utilizing the Precision and Recall tradeoff, we determined a cutoff value of approximately 0.42.

Step 11: Making Predictions on Test Set

We applied the insights gained to the test model and computed the conversion probability using the Sensitivity and Specificity metrics. The resulting accuracy value was determined to be 78.45%, with Sensitivity at 77.94% and Specificity at 78.91%.