Lead Scoring – Summary

Khushal and Muqeet

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:

Step 1 - Reading and understanding the Data

This is the first step in data analysis process where we acquire, load, and gain insights into the dataset we are working with. This process is crucial as it forms the foundation for all subsequent data manipulation, analysis, and modelling tasks.

Step 2: Data Cleaning

We systematically addressed variables with high percentages of NULL values by either dropping them or imputing missing data using median values for numerical variables. Additionally, for categorical variables, we created new classification variables as needed. Outliers were identified and effectively removed from the dataset as part of this process.

Step 3: Exploratory Data Analysis

This involves examining the dataset to summarize its main characteristics, often with visual methods. This step helps uncover patterns, spot anomalies, and formulate initial hypotheses. In this step, there were around 3 variables that were identified to have only one value in all rows. These variables were dropped.

Step 4: Data Preparation - Creating Dummies

Creating dummy variables involves transforming categorical variables into a format that can be used for modelling. This step is crucial in machine learning and statistical modelling to handle categorical variables appropriately. Each categorical level becomes a binary variable (0 or 1), making it easier for algorithms to interpret categorical data numerically.

Step 5: Test-Train Split

The dataset is split into two subsets: the training set and the test set. The training set is used to build and train the machine learning model, while the test set is used to evaluate its performance and assess how well it generalizes to new data. This separation helps ensure that the model's performance metrics are reliable and unbiased when applied to unseen data.

Step 6: Rescaling the features with MinMax Scaling

MinMax Scaling is used to standardize the features of the dataset, transforming their numerical values into a consistent scale typically ranging from 0 to 1. This method preserves the relative differences between features, preventing those with larger numerical ranges from disproportionately influencing the model. By ensuring all features contribute equally to the training process, MinMax Scaling promotes more reliable and precise predictions across the entire dataset.

Step 7: Model Building

Model Building involves selecting the best machine learning algorithm based on the task (classification or prediction), testing various models, optimizing their settings for accuracy, and validating them using cross-validation. The goal is to develop a robust model that accurately predicts outcomes on new data, ensuring reliable performance in classifying categories or predicting values.

We utilized the stats model to build our initial model, providing a comprehensive statistical view of all model parameters.

Step 8: Feature Selection Using RFE

We used Recursive Feature Elimination to select the important features. Using the generated statistics, we recursively examined the p-values to retain significant values and discard insignificant ones and we rebuilt up to 7 models and we arrived at the significant variables. The model was stable with significant pvalues and the VIF's for these variables were also found to be good.

Based on the above assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model. We also calculated the '**Sensitivity'** and the '**Specificity'** matrices to understand how reliable the model is. We found the sensitivity to be 70.79% and the Specificity to be 88.15%.

Step 9: Plotting the ROC Curve

We then plotted the ROC curve for the features and the curve came out be pretty decent with an area coverage of 89% which further solidified the of the model.

Step 10: Finding Optimal Cutoff Point

We plotted the probability graph for the '**Accuracy'**, '**Sensitivity'**, and '**Specificity'** for different probability values. The intersecting point of the graphs was considered as the optimal probability cutoff point. The cutoff point was found out to be 0.37 Based on the new value we could observe that close to 80% values were rightly predicted by the model. We could also observe the new values of the 'accuracy = 81.07%, 'sensitivity = 80.62%', 'specificity = 81.36%'. Also calculated the lead score and figured that the final predicted variables approximately gave a target lead prediction of 80.62%.

Step 11: Computing the Precision and Recall metrics

We also found out the Precision and Recall metrics values came out to be 79.22% and 70.79% respectively on the train data set. Based on the Precision and Recall tradeoff, we got a cut off value of approximately 0.4.

Step 12: Making predictions on the test set

We implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 81.06%; Sensitivity=80.39%; Specificity= 81.46%.