Lead Score Case Study Summary

Problem Statement:

X education sells online courses to industry professionals and they want help in finding the promising leads.

The company requires a machine learning model which will predict the probability score of each lead, such that the higher score will have higher conversion chance, and lower score will have lower conversion chance.

The CEO has particularly asked us for the lead conversion rate to be around 80%.

Summary:

Step 1: Reading and Understanding the data.

Extracting, reading and understanding the data.

Step 2: Data Cleaning.

We first took care of the variables that were having ‘Select’ values in them. Then we dropped the variables that had high percentage of values in them. Then remaining variables with few null values were taken care with imputing methods, The outliers were identified in few columns so we removed the outlier up to a certain percentile.

Step 3: EDA.

Now we started with exploratory data analysis wherein we visualized the categorical variables to understand their distribution and their impact/relation with the target variable. We dropped some columns where there was only single value of ‘No’ would have no impact on the target variable.

Step 4: Handling categorical variables.

We then created dummy variables for the categorical variables. But for those columns we that had yes/no values we converted them to 1/0, which was much better than to create dummies for them.

Step 5: Train Test Split:

After preparing the data for model, we started with splitting the data into train and test sets with proportion of 70% for train and 30% for test.

Step 6: Feature Scaling.

There were columns with values that had range very high from what most of the columns had, that is most of data was in 0 to 1 range, but few columns had higher range. Therefore, we used MinMax scaler to scale those data values to required range.

Step 7: Recursive Feature Scaling.

Using RFE, we selected the top 15 features that are more impactful for our model or target variable.

Step 8: Assessing the model

Then using statsmodel, we assessed our model on different metrics like P-value, to understand the variables and their impact on model. We also use VIF values of the features to check the correlation between features and eliminated the ones that were having higher than 5 VIF value.

Step 9: Plotting the ROC curve.

We plotted the ROC curve of these features and we retained an area coverage of 85%, which indicated a very decent model.

Step 10: Finding the Optimal Cut-off value:

Then we calculated the accuracy, sensitivity and specificity which came to be 80%, 76% and 81% respectively. And then plotted the probability graph of the range of cut-offs and their accuracy, sensitivity and specificity, which gave us an optimal cut-off value of 0.35 at the intersection of all three metrics.

Step 11: Computing the precision and recall.

We got the precision and recall values of 72% and 75% on the training set. The precision and recall curve also gave an optimal value of between 0.3 and 0.4.

Step 12: Making prediction on test set.

Finally, we implemented our model on the test set and calculated the conversion rate of each lead. The accuracy value on test set we achieved is 79.9%, which is almost what was expected by the CEO.