Summary Report

This is a problem statement where an education firm 'X Education' wants to improve its lead identification so that those prospective leads can be persuaded and focused upon. As a result, this would help in better lead conversion and saving of efforts by not focussing on not so prospective leads.

So, basically, this is a logistic regression problem, where based on existing lead conversion data, a model can be created which will predict whether a prospective lead will convert (i.e., purchase the courses from X Education) or not.

To prepare this model, I followed the below steps.

Step 1: Reading and understanding the data

Step 2: Cleaning and preparing the data

Step 3: Train-test split and standardizing numeric variables

Step 4: Building and evaluating the model

Step 5: Making predictions on test set

Before proceeding with these steps, I also checked if there was significant class imbalance (like too many yes or no in the target variable). But a 38%-62% split was good enough.

I dealt with the null values and made sure that the cleaned data has no null values and are belonging to appropriate datatypes.

Then, I mapped the Yes/No binary variable as 1/0. I also created dummy variables for other multilevel categorical features.

Then I did the train-test split (70:30) and standardized the numerical variables.

Then I started preparing the model using logistic regression GLM model. I, iteratively, kept removing the features which were having high p-values or high VIFs till the point where my model was having all the significant features with very less VIF values.

I kept checking that the accuracy of the model is not getting compromised significantly due the feature elimination. I checked the ROC curve and the area under curve of 0.87 was encouraging.

Then to achieve a good sensitivity (True Positive Rate), I found out the optimum cut-off to be 0.35.

And, my final model was able to delivery almost 79% accuracy, 78% sensitivity and around 80% specificity on the test data.