Summary

Tracking and handling the customers in an online business is a very important aspect and this case puts forward a similar real time scenario/situation that needs to be handled by an online education company named ‘X Education’ which markets its online courses through various sources.
As the result of marketing, many customers land on the company website. The company has all the records of the activities related to the customer and wants to identify the potential leads in order to focus on them and reduce the effort on non-potential customers.

Proceedings and learnings :
Once all the data was extracted, it was noticed that few columns have high null values and few had ‘Select’ as the value. Both of these would not be of use in the analysis. Hence, they were dropped.

Now, let's think of ‘Country’ and ‘City’ features, does the location of the customers affect lead conversion. I don’t think so. Therefore, they do not have significant usage in the analysis and are dropped. The null value of the remaining features were imputed with the most occurring value of that particular feature. The data available after cleaning is then visualized (EDA) to understand the trend.

The features with No/Yes values were converted to 0/1. Some categorical features had very high levels (say around 12) which would become difficult to handle when dummy variables are created for these. Therefore, the levels of these were reduced by combining the less significant levels as ‘Others’. This helped in getting a better distribution of the levels.

The dummy variables were created and the dataset was split into train and test data. Feature scaling was performed to standardize the numeric data (other than 0/1 features) for better performance and efficiency.

A base model was first built using train data and then feature selection was performed using RFE for automatic feature selection of the top features, let's say, starting with 16. Then p-value and VIF value is observed for each of the features. The features are removed one after the other till the model has features with p-value almost zero (zreo : very significant) and VIF <5 (<5 : least correlation with each other).

Then, predicted probabilities were produced by the model. An optimal cut off value has to be found to assign 0/1 against the probabilities. The intersection of accuracy, sensitivity and specificity values for different probability values in a line graph gives the optimal cut-off.

Then, the test data was picked, transformed and prediction was performed producing the probabilities. Again, 0/1 was assigned against each probability based on the intersection point obtained from recall and precision trade from the training set.

Metrics such as precision, recall, accuracy and F1 score were calculated.

Train set -	Accuracy = 85.14 %	Test set -	Accuracy = 85.93 %
	Precision = 77.35 %		Precision = 82.48 %
	Recall = 86.29 %		Recall = 81.73 %

Final test F1 score = 85.91 %

It can be observed that the train set has good metric percentages and test data prediction also has all the metrics above 80%. This means the model has performed well.