Summary

The solution below has been performed for X Education to enable better targeting and ensuring better conversion of leads to join the platform. The input data provided gives a number of features to perform analysis.

The company needs a model such that a score is assigned to each lead which could be used by the marketing team to improve the conversion. The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Solution:

1) Reading and analysing the Data – The data is read into pandas and review all the attributes based on the data dictionary provided.

2) Data Cleansing -
    - The columns with high null values as well as those which have a very high proportion of single values are dropped from the analysis.
    - Some columns also have values which are equivalent to NULLs, which are converted and then assessed for usage.
    - These columns will not help in the prediction. Eg Country, City, Search, Magazine, Newspaper Article, X Education Forums, Specialization, Tags etc

3) EDA – The data is then visualized through various charts which help us identify the key categorical attributes and values. We also are able to perform outlier analysis for continuous variables.

4) Dummy Variables: Categorical variables are then converted to dummy variables which help convert the categories to binary numeric values. This is an important step for model building.

5) Train-Test split – The data was split in 70% and 30% ration for train and test data respectively.

6) Scaling - For numeric values we used the MinMaxScaler to scale the values to have a better usage in the model creation.

7) Feature selection using RFE: Using an iterative approach of Recursive Feature Elimination, we selected the 20 top important features. Using the pvalue, variables were then dropped until we retained all features with insignificant P-values.
   We also checked the VIFs for the retained features and the VIF scores were also found to be satisfactory.

8) Calculated the Confusion matrix, accuracy , sensitivity and specificity metrics.

9) ROC Curve. We also plotted the ROC curve which indicated the model is pretty good with large area under the curve.

10) Finding the Optimal Cutoff Point Then 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 this cutoff, the predictions were made using the model.

11) Based on the precision and recall values plot, the cut off point was found to be 0.4

12) Predictions on the test set- The model was used on the test data split and then the metrix were again calculated (Accuracy, Sensitivity and Specificity metrics along with Precision and recall). All these aligned well with the train metrics and then it was concluded that the model had performed well.

Overall the Model with score above 80% meets the requirements of the CEO with 80% conversion. This model will enable them to attain conversion of their leads and improve business performance.