X Education is an education company that provides online courses to professionals in various industries. They attract a significant number of potential customers who visit their website to browse the available courses. When these visitors provide their email address or phone number on the website, they are considered as leads. Additionally, X Education also receives leads through referrals from their past customers. The sales team then proceeds to contact the leads through phone calls and emails, among other methods, in an attempt to convert them into paying customers. However, only a fraction of the leads get converted, with the typical conversion rate standing at 40%.

X Education has conducted an analysis to identify the characteristics of leads that are more likely to convert into paying customers. With this knowledge, they aim to increase their conversion rate to 80%.

We have used the following steps to prepare our analysis:

1. Loading and cleaning the data:
We have imported a CSV file called "Leads.csv" and examined its data frame. To enhance the quality of the data, we have dropped the columns that contained over 30% of null values. Additionally, we have performed imputation to fill in the missing values and checked for any duplicate entries in the data set.To address the issue of missing values, we have replaced the null values with the median and mode of their respective columns. This approach helps to retain as much data as possible without losing valuable insights.
Furthermore, we have reduced the dimensionality of the data set by combining the categorical value counts in columns with less than 100 values into a new category named "others". This approach helps to simplify the data and make it more manageable for analysis.

2. EDA:
We have used EDA in order to check the condition of our dataset and it is found that a lot of elements are irrelevant which belong to categorical variables. However numerical variables looks good. It is understandable from our EDA that there are many elements that have very little data and so will be of less relevance to our analysis.

3. Creation of Dummy Variables:
The dummy variables were created for categorical features. For numeric values we used the StandardScaler.

4. Spitting dataset into train and test and Scaling:
We have split the dataset into train and test at 70% and 30% respectively and used stratification so as to achieve equal distribution of classes. Done Scaling for numerical columns.

5. Feature Selection:
Firstly, SelectKBest with chi2 was used to select the top 20 relevant variables.

6. Model Building:

We have used sklearn Logistic Regression Classifier to build our classification algorithm.

7. Model Evaluation:

Confusion matrix was made. Determined the best cut-off value using ROC curve to find the F1 score , Precision and Recall which came to be around 82% and 79%. The area under ROC curve is 0.82 indicating a good predictive model.

8. Final Observation:

Accuracy on train data : 0.831
Accuracy on test data : 0.828
Precision on test data : 0.816
Recall on test data : 0.785
ROC score on test data : 0.818