Summary :

Problem Statement:

- An education company X Education sells online courses to industry professionals.
- although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

The following actions were taken while formulating a solution to the above problem:

Step 1: Reading and understanding the data.

The data was read as a csv file and it was ensured that all the attributes retained original values.

Step 2: Data Cleaning

After initial inspection, it was observed that there were a few columns that had missing values in them. All those attributes that had large number of missing values (more than 45%) were eventually dropped. For attributes with missing values lesser than the set standard, the values were imputed based on the nature of attributes (categorical or numerical). The numerical attributes had outliers in them that we capped between 1st and 99th percentile.

Step 3: Exploratory Data Analysis:

On careful examination of attributes individually, it was observed that certain categorical columns had very high proportion of similar values which caused high imbalances. Such categorical columns were removed.

Step 4: Data Preparation and Modelling:

Post Exploratory Data Analysis, we transformed our attributes in such a way that less important attribute values were clubbed into one. This was done to avoid possible computational issues in the model building phase. Dummy variables were created for all the categorical attributes. The train and test dataset split were performed and necessary variables were scaled to make the data unitless.

Step 5: Feature Selection and model building:

Most important 15 features were selected using Recursive Feature Elimination technique. Various models were built using the attributes that we selected using Recursive feature elimination. On the third iteration, we were able to achieve a model that had all the statistically significant variables and there was no multicollinearity between the variables/attributes.

Step 6: Model Evaluation:

The final model was evaluated based on accuracy score, sensitivity score, specificity score, precision score, recall score and F1 score. All these metrices were achieved after creating a confusion matrix of actual leads and the ones that the model predicted. We attained maximum model efficiency with a decision boundary value of 0.3. The model achieved an accuracy score, sensitivity score, specificity score, precision score of 89%,89% ,87% and 87% respectively. The model seemed to be doing well both on the train dataset and test set with an average AUC value of 0.95.

Step 7: Making predictions on the test dataset:

The model was used to make the predictions on the test dataset and the model performed very similarly to how it performed on the train dataset. The model achieved an accuracy score, sensitivity score, specificity score, and precision score of 88%,89%,87% and 81% respectively.