Q) A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

SUMMARY

We have been given an overview of the lead scoring case study, which includes information on how X Education follows up on leads for programme enrollment from several channels in an effort to turn them into prospective clients.

Around 30%, the conversion rate is now rather low. Thus, our job is to examine the information and develop a hypothesis that can forecast conversion rates for leads up to 80%. To do this, we performed a foundational investigation of the provided data set.

Throughout the whole project, we learned a lot concerning the potential consumers' visitation patterns, length of stay, method of access, and percentage of conversions from the basic data provided.

1. Initial check:

A few null values and not desirable values were incorporated within this dataset, which only contained inrelevant and unadequate data. We thus proceeded with those numbers, removing and replacing others with the mode, mean, and median. In order to prevent the loss of data, certain null values were converted to "others.

2. Exploratory Data Analysis:

We performed a brief EDA to assess the quality of our data. Numerous components of the category variables were shown to be meaningless. The numbers appear to be in range, and there are outliers identified and rectified.

3. Creation of dummy variables:

Dummy variables were created for a few categorical data points, and we took out the "Lead Number" components. The Standard Scaler was used for numerical values.

4. Train-Test split:

The percentage split for train and test data had been 70% and 30%, respectively.

5. Model Building:

First, RFE was used to identify the top 20 important factors. The remaining variables were individually deleted based on their VIF values and p-values. Only columns with a VIF less than 5 and a p-value less than 0 were used for the final model.

6. Model Evaluation:

We developed a confusion matrix to describe the efficiency of an algorithm for classification. The ROC curve was utilised to determine the most accurate, sensitive, and particular threshold values.

7. Predictions:

Predictions were performed on the test data frame using an ideal cut based on 0.37, with an accuracy of 80 percent.

8. Recommdations:

At last, we recommended the sales team contact those potential customers urgently; those lead scores were high, and the chances of enrolment were very high.