Summary

The data provided to us is X education data and they sell online courses to industry professionals. People interested in the course land on the website, browse courses, reach through email or calls. The conversion ratio is **45.50%** as per the data processing done.

We deleted the leastnumber of rows possible to generate better inferences from the data provided to us. We removed the columns having more than 40% missing value since they were of not much use to make conclusions.

Then, imputing missing values was the next step we followed as part of our process.

After finishing the data cleaning step, we started working on the Exploratory Data Analysis (EDA) part where we analyzed the categorical and numerical variables. We tried analyzing the Correlations between different variables. We also checked outliers and Treated outliers for each column respectively.

After completing EDA, we have created dummy variables and convert binary variables into numeric values. Then, dropped the original variables by which we created dummies.

Then, we did the data split into 2 different data sets i.e., *Train and Test data* by using the standard scaler. We used standard scaler method for splitting so that all the variables are on the same scale.

Then, we started with the Recursive Feature Elimination (RFE) approach by keeping the 15 variables, checking the VIF and p- values and rebuilding the model again and again till we get an adequate P-value and decent VIF score. We made total of 5 to reach to our final resultant model.

After building the final model (Model - 5), we calculated Accuracy, specificity and sensitivityand optimal cutoff to decide the churn probabilities. Then, plotted a ROC curve graph to measure the area which came out to be **0.86.** Then, we plotted sensitivity, specificity and probability to calculate the optimal cutoff.

Now, we have to make predictions on the test data set and on the same columns as the

training model. We did the same calculations on the test data same way as training data.

Important columns - Source of Lead, Lead origin, current occupation, Last acitivity, Time spend on the website, Choosing_Course_Better Career Prospects, Last Notable activity and Specialization are the key variables. The conversion ratio can be increased by following the model observation.

Final Result:

- Train Data Set:

Accuracy - 78.94%
Sensitivity - 79.16%
Specificity - 78.76%

- Test Data Set:

Accuracy - 80.22%
Sensitivity - 81.53%
Specificity - 79.16%