Summary

The given case study is of an education company named X Education which sells online courses to industry professionals, the objective is to predict (determine) the potential leads who probably buy this online course. The provided dataset consists of 37 predictive variables, the **logistic linear regression** was applied to predict the probability that the given candidate will select the online course or not.

In exploratory data analysis, the variables have been dropped based on null values present in them and their importance in the model. The variables having a null value of more than 45% are dropped in an earlier stage. The dummy variables have been created for the variable which contains categorical variables in it.

The **logistic linear regression model** is applied by considering 'Converted' as a target variable, 70% of the given data used for model training and the remaining 30% dataset for testing the obtained model. Recursive Feature Elimination (RFE) method is used to find the 20 optimum variables for the model, further the variables are eliminated based on their p-Values and variance influence factors.

The potential factors which are used to predict the leads are lead origin, last activity, lead source, Last Notable Activity, Total Time Spent on Website, current occupation and Specialization of candidates. The optimum predicting probability is determined between on the basis of accuracy, sensitivity and specificity, which obtained to be 0.45 in given problem.

The obtained model is evaluated for accuracy, recall, and precision score, in this problem we need to maximize the recall score. As we have to bring down the false negative value to the lowest, which means we have to accurately target the potential customer if we fail we might lose some business.

The values of accuracy, Precision and Recall for the train and test dataset is as follows:

Parameter	Train Dataset	Test Dataset
Accuracy	79.04%	79.86%
Precision	79.25%	79.95%
Recall	76.41%	77.60%

Also, in given dataset, the logistic regression model selects 13 variables for lead prediction and they are as follows:

- Do Not Email
- Total Time Spent on Website
- Lead Origin_Landing Page Submission
- Lead Source_Olark Chat
- Lead Source_Reference
- Lead Source_Welingak Website
- Last Activity_Email Opened
- Last Activity_Had a Phone Conversation
- Last Activity_SMS Sent
- WIYCO_Working Professional
- LNA_Modified
- LNA_Olark Chat Conversation
- LNA_Unreachable