Summary

X Education offers online courses for industry professionals. Despite receiving numerous leads, their conversion rate is quite low. They have appointed us to identify the most promising leads, specifically those most likely to become paying customers. We received a historical leads dataset containing approximately 9,000 data points. After thoroughly analyzing the requirements, Leads Data Dictionary, and the leads data, our team followed the steps below to develop a model aiming for a target lead conversion rate of around 80%.

1) Cleaning data
    a. Started with loading data and analyzing data for null/missing values. Columns that have more than 35% of missing values, excluded from analysis and model building.
    b. Columns with unique values can be removed, as they do not impact the overall solution.
    c. Data with "select" has been replaced with "Not provided".
    d. Numerical variable missing values are replaced with 0. I.e., Total Time Spent on Website, Page Views Per Visit and total visits.
    e. Since the majority of leads are from India, with only a few from other countries, the data has been grouped into two categories: India and Others.
2) EDA (Univariant/Bi/Correlation)
    a. EDA is performed on numerical, categorical and combination of both to check condition , correlation between numerical variables and outliers in the data.
3) Dummy creation for categorical variables
    a. Dummy variable used to represent categorical data in a numerical format suitable for model building.
    b. MinMaxScaler is a data preprocessing used to normalize numerical features by scaling them to a specific range, typically between 0 and 1.
4) Split the data into Train-Test
    a. The data was split into 70% for training and 30% for testing.
5) Model Building
    a. RFE was done to attain the top 15 relevant variables. Later variables manually removed based on the VIF and P-values i.e., remove variables with VIF > 5 and p-value >0.05
6) Model Evaluation
    a. A confusion metrics , Optimal cut off value (using ROC curve) was used to find the accuracy >80%, sensitivity and specificity.
7) Predictions:
    a. Prediction was done on test data with optimal cut off as .35 with accuracy, sensitivity, specificity of 80%.