Summary

Problem Statement:

X Education gets a lot of leads and 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'. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Approach to solve the problem:

1. **Cleaning the Data:**
   After importing the data we checked for duplicate values, null values which were treated accordingly and some columns have the value as "Select" which were imputed either NaN or Not provided in order not to lose too much data.

2. **EDA: Data Analysis**
   After cleaning the data EDA was performed, we checked in various category what was the conversion rate, we also removed some no variance variables as well, there was only one value as No.

3. **Creating the Dummy variables:**
   As there were categorical values, so each of those dummy variables were created (k-1), to avoid having more dummy variables. And we scaled the data using MinMaxScaler.

4. **Split the data set into Train and Test (70:30):**
   The final dataset was distributed in train and test data in 70% and 30% ratio
   .
5. **Model Building : Using Mix Approach (rfe and manual elimination)**
   RFE was use to get top most relevant feature to create model then manually eliminate the features which has p-value more than significance level, and VIF more than significance level.

6. **Model Building : Using Mix Approach (rfe and manual elimination)**
   RFE was use to get top most relevant feature to create model then manually eliminate the features which has p-value more than significance level, and VIF more than significance level.

7. **Model Evaluation**
   To evaluate our model we plot the ROC curve, calculated the
   Precision, Sensitivity, Specificity for train dataset and test dataset which was up to the mark, we found our optimum cutoff which was 0.38.
   For train data set: Accuracy-81.58%, Sensitivity-79.65%, Specificity-82%
   For test data set: Accuracy-80.85%, Sensitivity-79.65%, Specificity-82.75%

   Conclusion: Accuracy- 80.85%, Sensitivity- 79.65% and Specificity- 82% values of test set which are approximately closer to the respective values calculated using trained set.