SUMMARY

Problem Statement:

An education company named X Education get from several websites and search engines like Google. The typical conversion rate is now 30%. 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 a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The target lead conversion rate is around 80%

Solution Approach of the Assignment:

1. **Importing Libraries and Loading Data**:
   Imported the needed libraries and got the data loaded into Data Frame from csv file provided

2. **Inspecting the Data Frame:**
   Data Frame is inspected to see its shape, datatype of columns, nulls in the columns and not so significant columns if any for our analysis.

3. **Data Cleaning:**
   Value 'Select' was replaced with nulls as it doesn't provide any information. Also nulls in the columns were handles in 3 ways – Dropped the columns with nulls >40%, imputed nulls with "not available" for columns with moderate null-values and dropped records with null-values which were around 1%

4. **EDA:**
   Univariate and Bivariate analysis were done to study the data. It was found that most of categorical variables were irrelevant. Numerical variables are significant and there were no outliers found in data.

5. **Dummy Variables Creation:**
   Dummy variables were created for categorical variables and MinMaxScaling was done for numerical variables.

6. **Test Train split:**
   Data was split into 70% training and 30% test data

7. **Model Building:**
   Top 15 features for our model were identified using RFE. LogisticRegression model was build and refined by dropping variables by studying P-value and VIF. Final model with p-value less than .05 and VIF <5 got selected

8. **Model Evaluation:**
   Using the final model, confusion matrix was made. Optimum cut off value (using ROC curve) was used to identify the accuracy, sensitivity and specificity and they came to be just over 80%

9. **Prediction on Test Data:**
   Prediction was done on the test data set with an optimum cut off as 0.35 and got accuracy, sensitivity and specificity as 81%.

10. **Precision- Recall:**
    This method was used to recheck with identified cut off of 0.42 and got Accuracy as 81%, Precision as 74% and recall as 76% for test dataset.

Observations:

Below are the features that influence the Conversation rate of leads in X education and hence they need to focus on these features:

1. Total Visits
2. Total Time Spent on Website
3. Lead Origin is lead add form
4. When the current occupation is
   a) Working professional.
   b) Other
   c) Student
   d) Unemployed
5. Lead Source is from
   a) welingak website.
   b) Olark chat
6. Do Not Email option is set as yes
7. Last Activity was
   a) Olark chat conversation
   b) SMS sent.