Executive Summary

The objective of this summary was to assign a lead score to a lead which can further be utilized by X Education company to persuade high potential lead prospects. To begin with,

1. We cleaned up the data by loading and inspecting it of
    a. Duplicates
    b. Missing values
    a. Identify Outliers and treat them.
    b. Check for data that is irrelevant to this analysis.
2. Perform exploratory data analysis.
    a. Identify various categories that influence the leads.
    b. Identify the current conversion rate of the leads.
3. Preparing the Data
    a. Creating Dummies for Categorical variable
    b. Converting binary values to numerical
4. Test-Train Split
    a. Feature Scaling
    b. Checking for collinearity
5. Model Building
    a. Ensure the p value for all data attributes is less than <0.05.
    b. Ensure the VIF is less than <2.5.
6. Creating Prediction
    a. Predicting the probabilities on train set
    b. Substituting 0 and 1 with a cut-off as 0.5
7. Model Evaluation
    a. Creating confusion matrix
    b. Derive Accuracy, Sensitivity, Specificity Scores
8. Plotting the ROC curve
    a. Identify optimal cut-off.
    b. Derive scores based on cut-off.
9. Prediction on Test Data Set
    a. Scaling
    b. Confusion Matrix
    c. Derive Accuracy, Sensitivity, Specificity Scores
    d. Calculate Precision & Recall

Based on the model built in Python, it was observed that the conversion rate on the current data is 38.5% and a further analysis into the data helped us understand that the primary reasons for a lead conversion is where the category of Occupation is Unemployed, and the source of this lead conversion is the Landing page and Google.

Through model building we were able to predict the accuracy, sensitivity and specificity scores on the train data set that gave us a score of around 80% which was in line with the Precision and Recall score on the train data set which was also between 70%-80%. These scores were good indicators for us to determine that the model selected would provide us with the insights we needed when applied on the test data set.

Further analysis on the test data set gave us a cut-off score of 0.41 and the overall accuracy score of 75.8% indicating that the model we built was working as expected to serve the purpose and we were able to predict the following with high lead score as being most influential variables.

1. Total time spent.
2. Total number of visits
3. Lead Source with following priorities.
    a. Google
    b. Direct Traffic
    c. Organic Search
    d. Welingak Website
    e. SMS
    f. Olark chat conversations
4. Lead Origin
    a. Lead add format
5. Occupation
    a. Working Professional

Considering these key factors, X Education can significantly increase conversions by persuading these potential buyers and get them to purchase the courses from them.