Summary

This analysis is performed for X Education and to find ways to increase conversion rate and to make professionals to join their courses. The dataset provided gave us a lot of information about how the potentials customers visit the site, the time they spend over there, Then how they reached the site and the conversion rate.

The following technical steps are used:-

1. **Data Cleaning:**
    - First step is to clean the dataset which has outliers, missing values and to do data type correction.
    - Option'Select' has to replace with a null value since it did not give us much information and percentage of null value of each column is calculated and find if any duplicated value exists.
    - Checked for number of unique Categories for all Categorical columns.
    - From that Identified the Highly skewed columns and dropped them.
    - Treated the missing values by imputing the favourable aggregate function like (Mean, Median, and Mode).
    - Detected the Outliers.

2. **Exploratory Data Analysis:**
    - A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seems good but found the outliers
    - Performed Univariate Analysis for both Continuous and Categorical variables.
    - Performed Bivariate Analysis with respect to Target variable.

3. **Dummy Variables:**
    - The dummy variables are created for all the categorical columns.
4. **Scaling:**
    - Used Standard scalar to scale the data for Continuous variables.
5. **Train-Test Split:**
    - The Spit was done at 70% and 30% for train and test the data respectively.
6. **Model Building:**
    - By using RFE with provided 25 variables. It gives top 25 relevant variables. Later the irrelevant features was removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value 0.05 were kept).
7. **Model Evaluation:**

- Logistic Regression Model is decent and accurate enough, when compared to the model derived using PCA, with 78.6 % Accuracy on Test Set, 73.3 % Sensitivity and 82.3 % Specificity.

  We can vary these parameters by varying the cut-off value and thus predict Hot leads based on scenarios like availability of extra resources and vice-versa.

8. **Prediction:**
    - Prediction was done on the test data frame an optimum cut-off as 0.35 with accuracy, sensitivity and Specificity of 80%.
9. **Precision-Recall:**
    - The method was also used to recheck and a cut-off of 0.42.
10. **Conclusion :**

We have noted that the variables that important the most in the potential buyers are:

Top 3 variables in model, that contribute towards lead conversion are:

- Total Time Spent on Website
- Last Notable Activity_SMS Sent
- TotalVisits

X Education Company needs to focus on following key aspects to improve the overall conversion rate:

- Increase user engagement on their website since this helps in higher conversion
- Increase on sending SMS notifications since this helps in higher conversion
- Get Total visits increased by advertising etc. since this helps in higher conversion
- Improve the Olark Chat service since this is affecting the conversion negatively