Summary

The case study was done for X Education to improve the lead conversion rate and get more working professionals to enroll to their courses. The initial data provided gave a wide variety of information on how the potential leads reach the website, their location, specialization, and the conversion rate.

The following are the steps used:

1. **Cleaning the data:**
   The initial data was partially with few null values and found that 'select' option to be replaced with null values as no suitable information is not given. For few columns, the null values are replaced by "not provided" to retain most of the data. There were few columns such as 'country' which found to have too many distinct values and high value counts for the 'India', hence made into only three categories viz., India, not provided and Outside India.

2. **EDA:**
   A detailed EDA was carried out on the data generated by sales team with respect to the target variable "Converted". Also, made visualization plots to show the impact of different categorical variables on the target variable, and some seem to be much irrelevant. The plots graphed on numeric variables were consistent enough to reflect the Lead conversion rate.

3. **Encoding Binary Variables:**
   The binary variables were encoded to fetch a value of "0" or "1".

4. **Dummy Variables:**
   The dummy variables were created on all relevant categorical variables and original columns were dropped to avoid redundancy. Later we used the StandardScaler to standardize the values for model building.

5. **Train-Test Split:**
   The Train-Test Split was done in the ratio 7:3 respectively.

6. **Model Building:**
   Initially, we used automated RFE to select 15 most relevant variables for model building. Then, few variables were removed based on p-value and VIF and finally the variables with p-value<0.05 and VIF < 4 were retained.

7. **Model Evaluation:**
   Confusion matrix was calculated to observe the changes in the model and further evaluation. Next ROC curve has been plotted in order to fix the optimum cut off value to find the accuracy, sensitivity and specificity which resulted between ~72% and ~84%. Precision and Recall metrics were calculated on both train and test data and varied between ~65% and ~85%.

8. **Prediction:**
   Prediction was made on the test data frame using an optimum cut off of 0.285 with accuracy, sensitivity and specificity of 77%, 85.55% and 75.55% respectively.

9. **Lead Score:**
   Lead Score was calculated for each lead which range between 0 and 100. The high the lead score reflect high chance of conversion.

It was found that the variables that mattered the most in the potential buyers are (In descending order):

1) Lead Origin
   a) Lead add form
   b) Lead import
   c) Landing Page Submission
2) What is your current Occupation
   a) Working Professional
   b) not provided
3) Total Time Spent on Website
4) Country
   a) Not provided
5) Specialization
   a) not provided
   b) Hospitality
6) Do not Email

Final Comments:

The model will basically fetch high score leads which will be passed on to sales team and apply multiple channels to connect with the customer through phone, email and SMS.

Marketing team may need to nurture these leads and result in high conversion rate.