Summary

The model building and prediction is being done for company X Education and to find ways to convert potential users. We will further understand and validate the data to reach a conclusion to target the correct group and increase conversion rate. Let us discuss steps followed:

1. **Data Cleaning and EDA:**
   - Converted “Select” value to NULL and then Percentage of null value in each column was checked
   - The columns with more than 45% missing values were dropped and columns having missing values less than 2%, those rows were dropped
   - Remaining Columns having null values were replaced with others or the most common value. Eg, in the country column, since India is the most common occurrence among the non-missing values, we imputed all not provided values with India.
   - Columns where there was single value across a column and where there was high data imbalance (1 value contributes to 85%+ data of the column)were dropped
   - Based on requirement and analysis, any redundant columns were dropped
   - Univariate and bivariate analysis was done – Handling Outliers, Standardizing data and combining variables

2. **Train-Test split & Scaling :**
   - For category columns Dummy Variables were created and binary variables were converted
   - The split was done at 70% and 30% for train and test data respectively.
   - Standard scaler was used on the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']. Train Dataset was fit_transformed whereas Test Data was Transformed

3. **Model Building**
   - RFE was used for feature selection
   - Then RFE was done to attain the top 20 relevant variables
   - Later the rest of the variables were removed manually depending on the VIF values and p-value
   - A confusion matrix was created, and overall accuracy was checked which came out to be 85.89%

4. Model Evaluation

- With Default 0.5 Cutoff Value We get the below Performance Parameters
  - On **Training Data**

    Accuracy = 85.89%
    Specificity = 90.52%
    Recall = 78.36%
    Precision = 83.53%
    False Positive Rate = 9.48%
    Negative Predictive Value = 87.21%

- The optimum cut off value was found using ROC curve. The area under ROC curve was 0.92.
- After Plotting we found that optimum cutoff was **0.35** which gave

  Accuracy = 85.51%
  Specificity = 86.85%
  Recall = 83.33%
  Precision = 79.54%
  False Positive Rate = 13.15%
  Negative Predictive Value = 89.46%

  - Prediction on **Test Data**

    Accuracy: 85.14%
    Specificity: 86.34%
    Recall: 83.10%
    Precision: 78.22%
    False Positive Rate: 13.66%
    Negative Predictive Value: 89.64%

5. Recommendations:

The company should prioritize to make calls to the leads coming from the ***Lead Source_Olark Chat, Last Activity_Olark Chat Conversation, Lead Origin_Lead Add Form and Import*** and those who spent ***more time on the websites*** as they are more likely to get converted. Also, those who are ***specialized in HR Management, Marketing Management and Operations Management*** are more likely to convert.

The Model seems to predict the Conversion Rate very well and we should be able to give the Company confidence in making good calls based on this model.