Summary

This model is built for an education company named X Education who sells online courses to industry professionals who has many professionals who are interested in the courses land on their website and browse for courses.

The motive of the model is to improve the lead conversion rate of X education (It is roughly at 30% now) by identifying all their 'hot leads' from a pool of leads (once they fill up a form with email id) they get from various sources like google search, other references etc.

The analysis and model were created using the steps given as follows:

1. **Data Cleaning:**
   - To clean the dataset, we chose to check for duplicate data which was not present.
   - The data possessed many null values and the option 'Select' which was required to be replaced with a null value since it did not give us much information.
   - Many null values were in 45% range, hence, columns with more than 45% of null values were dropped.
   - 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:**
   - An initial 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 seem good but found outliers and corrected them.
   - Performed Univariate Analysis for both Continuous and Categorical variables.
   - Performed Bivariate Analysis with respect to Target variable, "Converted".

3. **Dummy Variables:**
   - Categorical features were found to be "Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', 'City', 'Last Notable Activity'
   - Dummy variables were created for all the above categorical features.

4. **Scaling:**
   - Used Standard scalar to scale the data for Continuous variables.

5. **Train-Test Split:**
   - The Spit was done at 70:30 ratio, 70% for Train and 30% for Test data respectively.

6. **Model Building:**

- By using RFE, top 20 relevant variables were found out.
- Irrelevant features were removed manually depending on Significance, VIF values and p-values (The variables with VIF < 5 and p-value < 0.05 were kept).

7. **Model Evaluation:**
   - A confusion matrix was made.
   - The ROC curve was plotted to find the area below which was 0.89 denoting good model.
   - Accuracy, sensitivity, and specificity all came to be around 80% which is deployable.
   - Sensitivity, Specificity, Positive Predictive value, Negative predictive value is all found to be satisfactory.
8. **Prediction:**
   - Accuracy, sensitivity, and specificity for various probabilities were plotted and optimum cut-off point which was found to be 0.34.
   - Prediction was done on the test data frame based on optimum cut-off as 0.34 with accuracy, sensitivity and Specificity of roughly 80%.
9. **Precision-Recall:**
   - Trade off curve between Precision and Recall was made to recheck and a cut-off of 0.41 was observed which was matching.
10. **Inference and Recommendation:**
    - A model was created to successfully identify the “Converted” parameter of all the Leads or in other words to find the “Hot Leads”.
    - Target lead conversion rate using this model is around 80%.
      This Model seems to predict the Conversion Rate as desired and decision shall be made in making good calls to get a higher lead conversion rate of 80% from roughly above 30% given in the raw data.

    - **Equation of final model (Which is model 9 in our case here) can be given as below:**
      Converted = -0.0376 + 5.8115 * Lead Source_Welingak Website + 3.3166 * Lead Source_Reference + 2.6083 *What is your current occupation_Working Professional + 2.1751 * Last Activity_Other_Activity + 1.2942 * Last Activity_SMS Sent + 1.0954 * Total Time Spent on Website + 1.0819 * Lead Source_Olark Chat - 0.9004 * Last Notable Activity_Modified - 0.9613 * Last Activity_Olark Chat Conversation - 1.194 * Lead Origin_Landing Page Submission - 1.2025 * Specialization_Others -1.5218 * Do Not Email

    - **Resultant model's Final Features (Most significant to least):**

```
Lead Source_Welingak Website                    5.811465
```

Lead Source_Reference	3.316598
What is your current occupation_Working Professional	2.608292
Last Activity_Other_Activity	2.175096
Last Activity_SMS Sent	1.294180
Total Time Spent on Website	1.095412
Lead Source_Olark Chat	1.081908
const	-0.037565
Last Notable Activity_Modified	-0.900449
Last Activity_Olark Chat Conversation	-0.961276
Lead Origin_Landing Page Submission	-1.193957
Specialization_Others	-1.202474
Do Not Email	-1.521825

- The above model provides clear indication that those leads with Welingak Website and Reference as Sources are very high potential.
- Also, Working Professionals are high potential leads. Calls shall be made to these hot leads.
- Those who has given, “Do not Email” as yes are having negative potential in the model too, along with some of the others.
- Also, using Lead indicator, specific leads to be contacted also can be found and contacted as desired.

End