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 Inspection**

   - ✓ We have 9240 rows and 37 columns in our lead's dataset.

2. **Data Cleaning**

   - ✓ We see that for some columns we have high percentage of missing values. We can drop the columns with missing values greater than 40%.
   - ✓ There are 37% missing values present in the Specialization column. It may be possible that the lead may leave this column blank if he may be a student or not having any specialization or his specialization is not there in the options given. So, we can create another category 'Others' for this.
   - ✓ We have retained 98% of the rows after cleaning the data.

3. **Exploratory Data Analysis**

   - ✓ Quick check was done on % of null value and we dropped columns with more than 40% missing values.
   - ✓ The lead conversion rate is 38%.
   - ✓ API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
   - ✓ Lead Add Form has more than 90% conversion rate but count of lead are not very high.
   - ✓ Lead Import are very less in count.
   - ✓ Google and Direct traffic generate maximum number of leads.
   - ✓ Conversion Rate of reference leads and leads through welling website is high.
   - ✓ Conversion rate for leads with last activity as SMS Sent is almost 60%.
   - ✓ Since India was the most common occurrence among the non-missing values, we imputed all not provided values with India.
   - ✓ Unemployed leads are the most in numbers but has around 30-35% conversion rate.

4. **Data Preparation**

- ✓ 38% lead conversion rate.
- ✓ RFE was used for feature selection
- ✓ Then RFE was done to attain the top 20 relevant variables

5. **Model Building**

- ✓ 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 81.68 %

6. **Model Evaluation**

- ✓ We have got sensitivity of 80% and this was mainly because of the cut-off point of 0.5 that we had arbitrarily chosen. Now, this cut-off point had to be optimized in order to get a decent value of sensitivity and for this we will use the ROC curve.

a) Prediction on Test Data

- ➢ Accuracy: 80.4 %
- ➢ Sensitivity: 80.4 %
- ➢ Specificity: 80.5 %

b) Prediction on Train Data

- ➢ Accuracy: 81.4 %
- ➢ Sensitivity: 81.7 %
- ➢ Specificity: 80.6 %

CONCLUSION

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.