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. **Cleaning data:**

The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to 'NA' so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to 'India', 'Outside India' and 'NA'.

- **EDA:**

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 and no outliers were found.
Quick check was done on % of null value and we dropped columns with more than 45% missing values.

- We also saw that the rows with the null value would cost us a lot of data and they were important columns. So, instead we replaced the NaN values with 'not provided'.
- Since India was the most common occurrence among the non-missing values, we imputed all not provided values with India.
- Then we saw the Number of Values for India were quite high (nearly 97% of the Data), so this column was dropped.
- We also worked on numerical variable, outliers and dummy variables.

1. **Train-Test split & Scaling :**

- The split was done at 70% and 30% for train and test data respectively.
- We will do min-max scaling on the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']

2. **Model Building**

- RFE was used for feature selection.
- Then RFE was done to attain the top 15 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 89.94%.

3. Model Evaluation

- **Sensitivity – Specificity**

  If we go with Sensitivity- Specificity Evaluation. We will get :

  - Accuracy 89.94%
  - Sensitivity 89.20%
  - Specificity 90.39%

- **Precision – Recall:**

  If we go with Precision – Recall Evaluation

  - On **Training Data**

    Accuracy 89.80%
    Precision 89.14%
    Recall 89.54%

  - Prediction on **Test Data**

    Accuracy 89.94%
    Precision 89.20%
    Recall 90.39%

CONCLUSION

TOP VARIABLE CONTRIBUTING TO CONVERSION:

- LEAD SOURCE:
  - Total Visits
  - Total Time Spent on Website
- Lead Origin:
  - Lead Add Form
- Lead source:
  - Direct traffic
  - Google
  - Welingak website
  - Organic search
  - Referral Sites

Last Activity:

- Do Not Email_Yes
- Last Activity_Email Bounced
- Olark chat conversation

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.