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. **EDA:**
   - Quick check was done on % of null value and we dropped columns with more than 40% missing values.
   - We saw the Number of Values for India were quite high, so this column was dropped.
   - We also worked on numerical variables, outliers and dummy variables.

2. **INFERENCE:**
   • Total Visits and Page Views Per Visit have the highest correlation.
   • The target variable i.e, Converted and Page Views Per Visit have negative correlation.

3. **Train-Test split & Scaling :**
   - The split was done at 70% and 30% for train and test data respectively.
   - We will do StandardScaler for scaling on the variables 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit']

4. **Model Building**
   - RFE was used for feature selection.
   - Then RFE was done to attain the top 20 relevant variables.
   - Building Model by dropping the variables with P-Value > 0.05 And VIF> 5
   - Predictions On Test Dataset
   - Overall Accuracy Is 90.0 % & subject to change

5. Model Evaluation

- **Sensitivity – Specificity**

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

  - On **Training Data**

    - The optimum cut off value was found using ROC curve. The area under ROC curve was 0.96.
    - After Plotting we found that optimum cutoff was **0.3** which gave

      Accuracy = 0.901
      Sensitivity = 0.908
      Specificity = 0.897

  - **Test Data**

    - We get

      Accuracy = 0.909
      Sensitivity = 0.913
      Specificity = 0.906

6. 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.

1. Logistic regression model is used to predict the probability of conversion of a customer.
2. Lead Score & conversion rate of final predicted model is over 90% in test data as well as Training Data
3. Overall this model is compatible to adjust with the company's future requirements as well
4. Top 3 Variables that contributes for leads getting converted in the model are:
    1. Tags_Will revert after reading the email
    2. Tags_Closed by Horizzon
    3. Last_Activity_SMS Sent