SUMMARY

The model is being built and predicted for the company X Education in order to uncover strategies to convert potential users.

We will further analyse and validate the data in order to arrive at a conclusion on how to target the relevant group and enhance conversion rate.

Let us go over the measures that were taken:

1. EDA:

• We performed a quick check on the percentage of null values and deleted columns with more than 45% missing values.

• We also discovered that rows with null values would cost us a lot of data, despite the fact that they were crucial columns. So we changed the NaN values with 'not provided' instead.

• Because India was the most common occurrence among the non-missing values, we attributed all missing values to India.

• When we realised that the number of values for India was fairly high (almost 97% of the data), we dropped this column.

• Additionally, we worked on numerical variables, outliers, and dummy variables.

2. Train-Test Splitting and Scaling:

• The train and test data were separated at 70% and 30%, respectively.

• We will use min-max scaling to determine the variables ['TotalVisits,' 'Page Views Per Visit,' and 'Total Time Spent on Website.']

3. Model Building:

• RFE was used for feature selection, followed by RFE to determine the top 15 relevant variables.

• The remaining variables were then manually deleted based on the VIF values and p-value.

• A confusion matrix was built, and total accuracy was determined to be 80.91%.

4. Model Assessment:

• Specificity - Sensitivity

If we use the Sensitivity-Specificity Evaluation method. We will receive:

Using Training Data

- The ROC curve was used to determine the best cut off value. The area beneath the ROC curve was 0.88.
- After plotting, we discovered that the best cutoff was 0.35, which resulted in

The accuracy is 80.91%.

- 79.94% Sensitivity
- 81.50% specificity.

Prediction Based on Test Data:

We obtain:

- 80.02% Precision
- 79.23% Sensitivity
- 80.50% specificity

5.Precision – Recall:

If we use the Precision - Recall Evaluation method,

Using Training Data:

- With a cutoff of 0.35, the precision and recall are 79.29% and 70.22%, respectively.
- In order to enhance the aforesaid percentage, we must modify the cut off value. After charting, we discovered that 0.44 was the best cut off value.
- The accuracy is 81.80%.
- Precision is 75.71%.
- 76.32% recall rate

Prediction Based on Test Data:

o We obtain

- The accuracy rate is 80.57%.
- Precision is 74.87%.
- Recall rate: 73.26%

6. So, if we use Sensitivity-Specificity Evaluation, the ideal cut off value is 0.35, and if we use Precision-Recall Evaluation, the optimal cut off value is 0.44.

CONCLUSION:

THE MOST IMPORTANT VARIABLE IN 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 appears to accurately anticipate the Conversion Rate, and we should be able to provide the Company confidence in making wise decisions based on this model.