SUMMARY

In order to convert potential consumers, business X Education is building models and making predictions. In order to target the right audience and boost conversion rates, we will further analyse and confirm the data. The steps are as follows:

1. EDA:

I. We performed a quick check on the percentage % of null values and dropped columns that had more than 30% missing values.

II. Additionally, we observed that the rows with null values would cost us a significant amount of data and were crucial columns. Instead, we substituted "missing" for the NaN values.

III. In column "Lead Source", The most dominating values are (google, direct traffic, olark chat, organic search, reference, welingak website, referral sites, fakebook). Since the presence of all other categories are very less, we can put those sources to a new category "Other_Sources".

IV. India was the most common occurrence among the non-missing values, so we imputed all not provided values with India.

V. The column was removed after it was discovered that the number of values for India was unusually high (almost 97% of the data).

VI. We also dealt with dummy variables, outliers, and numerical variables.

2. Train-Test split & Scaling:

I. The split was carried out at 70% and 30% for train and test data.

II. Min-max scaling was done on the variables: 'TotalVisits', 'Page Views Per Visit' and 'Total Time Spent on Website'

3. Model Building:

I. We used RFE for feature selection.

II. To obtain the top 15 relevant factors, RFE was then conducted.

III. Later, based on the VIF values and p-value, the remaining variables were manually deleted.

IV. The accuracy of the confusion matrix, which was constructed, was verified, and it was found to be 80%.

4. Model Evaluation:

- **Sensitivity – Specificity**

  If we go with Sensitivity- Specificity Evaluation, the following can be observed:

  - **On Training Data**
    - ✓ The optimum cut off value was found using ROC curve. The area under ROC curve was 0.87.
    - ✓ After Plotting we found that optimum cutoff was 0.35 which gave:
      - Accuracy ~ 80%
      - Sensitivity ~ 80%
      - Specificity ~ 77%.
  - **On Test Data Prediction**
    - ✓ We get:
      - Accuracy ~ 81%
      - Sensitivity ~ 82%
      - Specificity ~ 80%

- **Precision – Recall:**

  If we go with Precision – Recall Evaluation:

  - **On Training Data**
    - ✓ With the cutoff of 0.35 we get the Precision & Recall of ~70% & ~80% respectively.
    - ✓ So, to increase the above percentage we need to change the cut off value. After plotting we found the optimum cut off value of 0.41 which gave:
      - Accuracy ~80%
      - Precision ~73%
      - Recall ~75%
  - **Prediction on Test Data**
    - ✓ We get:
      - Accuracy ~82%
      - Precision ~76%
      - Recall ~79%

Note: If we go with Sensitivity-Specificity Evaluation, the optimal cut off value would be 0.35 & if we go with Precision – Recall Evaluation the optimal cut off value would be 0.41

5. CONCLUSION:

The top variables contributing to conversion are as follows:

- Lead Origin:
    - Lead Add Form
- The total time spent on the Website.
- Page Views Per Visit.
- Total Visits
- When the Last Notable Activity is had a phone conversation
- When the last activity is SMS sent
- When the lead origin is Lead add format
- When their current occupation is as a working professional.

We should be able to give the Company confidence in using this model to make wise decisions because it appears to predict the Conversion Rate quite effectively.