SUMMARY

The model building and prediction is being done for company X Education and to find ways to convert potential users. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

The following are the steps used:

1. EDA

- A quick check is done on the size, shape and null values of the data.
- some of the columns have 'select' filled in as their values. That means that the person (NaN) value.
- Drop the columns which don't give any information.
- Drop the columns with more than 30% null values
- Remaining null values are replaced using fillna for each corresponding column.
- We also worked on numerical variable, outliers and dummy variables.

2. Train-Test split and scaling

- The split was done at 70% and 30% for train and test data.
- Scale the features using the Standard Scaler [features are 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit'].
- Checking the collinearity between the variables using correlation matrix.

3. Model Building

- RFE is used for feature selection and attain the top 15 relevant variables.
- 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 80.91%.

4. Model Evaluation

- Optimal cutoff probability is that probability where we get balanced sensitivity and specificity.
- The optimum cut off value is around 0.35 (found using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.

5. Precision – Recall

- This method was used to recheck and a cutoff of 0.41 is found with precision around 72.55% and recall of 76.51% for test dataset.

6. Comparing the values obtained from both the train set and the test set

- **Train set**

  Accuracy = 80.91%

  Sensitivity = 70.09%

  Specificity = 87.58%

- **Test set**

  Accuracy = 80.65%
  Sensitivity = 76.51%
  Specificity = 83.07%

- All three values are similar in both train and test set

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.

....x x x x x x....