# Summary

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

## Steps Followed

### Step1: Importing and Understanding Data.

- Imported required libraries and Data
- Checked first 5 rows to see the data is getting imported.
- Checked for Non-Null values and data types of each column.
- Checked the quantitative spread of the data.
- Checked the shape and description of the data

### Step2: Data Cleaning

- Checked for duplicates if any
- Dropped the high percentage of Null values more than 40%.
- Checked for number of unique Categories for all Categorical columns.
- From that Identified the Highly skewed columns and dropped them.
- Treated the missing values by imputing the favorable aggregate function like (Mean, Median, and Mode).
- Detected the Outliers and Handled then

### Step3: EDA

- Performed Univariate Analysis for both Continuous and Categorical variables.
- Performed Bivariate Analysis with respect to Target variables.
- We also plotted the correlation matrix to identify the columns which are correlated.

### Step4: Data Preparation, Test-Train Split and Feature Scaling

- Converted all binary variables to 0 and 1.
- The dummy variables are created for all the categorical columns.
- Concatenated the dummy variables to the cleaned data set and dropped the initial columns.
- Used Standard scalar to scale the data for Continuous variables.
- The Spit was done at 70% and 30% for train and test the data respectively.

## Step5: Model Building

- Create a Logistic regression model using the prepared data.
- Use RFE to select the best 15 variables.
- Manual Feature Reduction process was used to build models by dropping variables with p – value > 0.05.
- Total 2 models were built with the 2nd model having P-Value < 0.05 and VIF < 5.
- Created a confusion matrix and find the overall accuracy with a cut-off value of 0.5.

## Step6: Model Evaluation

- ROC curve was plotted for the features and the curve came out be pretty decent with an area coverage of 89% which further solidified the of the model.
- Plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values.
- Prediction was done on the test data frame an optimum cut-off as 0.37 with accuracy, sensitivity and Specificity of around 80%.
- Precision-Recall was also used to recheck and a cut-off of 0.40.

## Step7: Predictions on Test Set

- After finalizing the optimum cutoff and calculating the metrics on train set, we predicted the data on test data set. Below are the observations:
    - **Train Set**:
        - Accuracy = 80.93%
        - Sensitivity = 79.04%
        - Specificity = 82.07%
    - **Test Set**:
        - Accuracy = 81.87%
        - Sensitivity = 79.04%
        - Specificity = 82.07%

## Step8: Conclusion

- Accuracy, Sensitivity and Specificity values of test set are around 82%, 80% and 82% which are approximately closer to the respective values calculated using trained set.
- Conversion rate on the final predicted model is around 78.54%
- There are 535 Hot Leads. They should be targeted as they have a high chance of getting converted.
- We have determined the following features that have the highest positive coefficients, and these features should be given priority in our marketing and sales efforts to increase lead conversion.

  **=>Lead Source_Welingak Website: 3.05**

**=>Lead Origin_Lead Add Form: 2.96**

**=>Current_occupation_Working Professional: 2.18**

- Hence overall this model seems to be good.