# Summary

An education company named X Education sells online courses to industry professionals. Analysis is done to find ways to get more industry professionals to join their courses. We have built a logistic regression model and find some important variables with their coefficient.

**Steps Followed:**

1. **Import Libraries and Read Data**

2. **Data Understanding and Inspection**

3. **Data Cleaning:**
    - Option select has been replaced with a null value since it did not give us much information.
    - Removed columns which had > 40% values.
    - Removed unnecessary columns like ‘Prospect ID','Lead Number','Last Notable Activity','Tags.
    - Removed variables which were having imbalanced data like ‘Do Not Call', 'Country', 'What matters most to you in choosing a course', 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper' etc.
    - Handled Missing value
        - Used Median value for continuous variables
        - Used Mode for categorical variables
    - Handled Outliers
        - Used Upper & Lower bound
    - Fixed Invalid values & Standardising Data
        - Like Replaced "google" with "Google"
        - Grouping low frequency value levels to Others for 'Lead Source' & 'Last Activity' column.

4. **EDA:**
    - Checked Data Imbalance
        - Only 38% leads were successfully converted.
        - Done Univariate, Bivariate and Multivariate analysis.
5. **Data Preparation:**
    - Converted binary variables (Yes/No) to 0/1 for below columns.
        - Do Not Email
        - A free copy of Mastering The Interview

- Created Dummy Variables for below columns & removed original after creation.
    - Lead Origin,Lead Source,Last Activity,Specialization,What is your current occupation etc.

6. **Train-Test split:**
    - The split was done at 70% and 30% for train and test data respectively.

7. **Feature Scaling:**
    - Used MinMaxScaler for data standardization for numeric columns.
    - Dropped variables which were highly correlated.

8. **Model Building:**
    - Used RFE to attain the top 15 relevant variables.
    - Later the rest of the variables were removed manually depending on the p-value & VIF values.( VIF < 5 and p-value < 0.05).
    - Logm2 was selected as the final model with 14 variables for making predictions on the Train & Test set.

9. **Model Evaluation:**
    - With the 0.41 cut off Precision around 75% and Recall around 76%.
    - When we used the precision-recall threshold cut-off of 0.41 the values in True Positive Rate,Sensitivity,Recall have dropped to around 75%, but we need it close to 80% as the Business Objective.
    - 80% for the metrics we were getting with the sensitivity-specificity cut-off threshold of 0.35.
    - So, we went with a sensitivity-specificity view for our Optimal cut-off for final predictions.
    - Lead Score was assigned based on 0.35 cut off for Train data set.

10. **Predictions on Test Data:**
    - Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.
    - Lead Score was assigned based on the 0.35 cut off for the Test data set.

**Conclusion:**

- Top 3 features:
    1. Lead Source_Welingak Website
    2. Lead Source_Reference

3. What is your current occupation_Working Professional

**Recommandation:**

- Focus on features with positive coefficients, Analyze negative coefficients.
- More spend can be done on Welingak Website in terms of advertising etc.
- Engage working professionals with messaging.
- Discounts for providing references that convert to lead, encourage providing more references.
- Develop strategies to attract high-quality leads from top-performing lead sources.
- Review landing page submission process for areas of improvement.