# SUMMARY REPORT

# LEAD SCORING CASE STUDY

## Introduction:

As a Data Scientist, my task was to enhance lead conversion rates of X Education, which currently stand at around 30%. X Education sells online courses to industry professionals and receives a large volume of leads. To address the challenge of converting these leads more efficiently, I built a logistic regression model to predict the probabilities (likelihood) of conversion for each lead and develop a strategy for two distinct phases:

1. Periods of aggressive conversion.
2. Periods when sales targets are met.

## Steps Followed:

1. **Data Understanding and Preparation:**
    - For data understanding, I have used Python Jupyter Notebook, to load the required dataset as a DataFrame and then checking various stats like Info and Descriptive Statistical Summary.
    - Next, I proceeded with Data Cleaning and Pre-processing by Removing the Case Mismatch, and then checking the Data for Duplicates Values and Null Values.

2. **Feature Engineering & Handling Missing Values:**
    - After checking the Null Values, I have dropped Variables (Features) that are having more than 30% of missing values.
    - For the rest of the Missing Values, I have used appropriate methods like Imputation using Mode() or Median() or Mean() and by creating new features like new categories
    - Also, I have created new features by Grouping Data in the categorical columns having many categories to avoid over fitting of the data.

3. **Exploratory Data Analysis of all the Features:**
    - I have treated Outliers in the numerical variables using ***The Percentile Method,*** by Capping Outliers at 99 Percentile.
    - For categorical variables I have visualize the distribution using Countplot and dropping Highly biased variables.
    - And then creating the dummies of all the categorical variables to make them compatible with the logistic regression model.

4. **Model Building:**
    - Firstly, I have scaled the dataset using ***Min-Max Normalization*** to get all the Features at a common scale between 0 and 1.
    - Then, I used ***Logistic Regression*** to model the probability of lead conversion. Logistic regression is well-suited to this classification problem because it predicts the likelihood of binary outcomes (conversion or no conversion).
    - After cleaning and preparing the data, I selected 15 Features, including both numerical variables and categorical dummy variables using ***Recursive Feature Elimination (RFE).***

5. **Model Evaluation:**
    - After selecting 15 Features I have evaluated the model by using ***GLM P-Value's*** and ***Variance Inflation Factor (VIF).***
    - Once the model is stable, I can see, that the model has achieved an ***Accuracy of 81.6%***, an ***F1 score of 74%***, and a ***Precision of 79.6%*** on the test dataset. This means the model can fairly & accurately predict which leads will convert while minimizing false positives.
    - The ***Recall score of 69.2%*** shows that the model captures about 69% of the actual converters, which is acceptable but leaves room for improvement.

## Learnings:

This case-study have provided several important insights, like:

1. **Model-Driven Decision Making**: Logistic regression models, when built correctly, can provide actionable insights into which factors drive business outcomes. In this case, it helped identify key variables (like Total Time Spent on Website) that had the most influence on lead conversion.

2. **Optimizing Sales Strategy**: The ability to dynamically adjust conversion strategies based on model predictions can lead to significant efficiency gains. By raising thresholds and adjusting focus, the company can avoid unnecessary resource expenditures (like excess phone calls) when it's not critical.

3. **Balancing Business and Technical Needs**: By presenting the results in the form of presentation to the Stakeholders required to translate complex technical insights into clear, actionable business strategies. It highlighted the importance of aligning data science outcomes with business goals to ensure maximum impact.

## Conclusion:

In conclusion, the logistic regression model offered valuable insights to driver the Lead Conversion and provided a clear path for improving X Education's conversion strategy. The key takeaway was that data-driven decision-making can significantly enhance both resource optimization and business outcomes.