# SUMMARY

## Business Challenge

X Education, a company specializing in online courses for industry professionals, struggles with a low lead conversion rate of around 30%. To address this issue, we developed a model to score leads, aiming to identify and prioritize those with a higher likelihood of conversion, aligning with the CEO's goal of an 80% conversion rate.

## Objective

To build a logistic regression model to predict the probability of each lead converting into a customer, thereby enabling X Education to focus efforts on the most promising leads.

## Methodology and Steps

We divided the project into various checkpoints to meet sub-goals systematically. The steps included:

1. **Data Preprocessing:**
   - Missing Values and Outliers: We dropped columns with more than 30% missing values and treated missing values and outliers in the remaining numeric columns.
   - Binary Encoding: We assigned binary values (0s and 1s) to variables such as 'Do Not Email', 'Do Not Call', 'Search', and 'Magazine'.
   - Dropping Uninformative Columns: Columns with 'Select' as a predominant value, like 'Lead Profile' and 'How did you hear about X Education', were dropped due to their lack of informative value. Similarly, 'What matters most to you in choosing a course' was dropped as it was nearly homogenous.
2. **Feature Engineering & Data Splitting and Scaling:**
   - Created dummy features for categorical variables through one-hot encoding.
   - Training and Test Sets: Split the data into training (70%) and test (30%) sets.
   - Feature Scaling: Standardized the features and removed those highly correlated with each other to reduce redundancy.
3. **Model Building:**
   - Data Imbalance Check: Noted an initial lead conversion rate of 48.1%.
   - Feature Selection: Performed Recursive Feature Elimination (RFE) on scaled data, selecting 15 key features.
   - Iterative Model Refinement: Built and refined the model iteratively, ensuring that all included features had p-values < 0.05 and no multicollinearity (VIF < 5).
4. **Model Evaluation:**
   - Initial Evaluation: Used a probability cutoff of 0.5 for conversion prediction, achieving 78% accuracy, 83% specificity, and 73% sensitivity.

- Cutoff Optimization: Adjusted the cutoff to 0.42 to balance sensitivity and specificity, maintaining a consistent 78% accuracy on both training and test data.
- ROC Curve Analysis: Plotted the ROC curve, achieving an 86% area under the curve, indicating a robust model.

5. **Final Model Testing:**
    - Prediction on Test Data: Evaluated model performance on test data, confirming the model's effectiveness with 78% accuracy, 79% specificity, and 78% sensitivity.

## Insights and Recommendations

- Platform Engagement: Leads spending more time on the platform are more likely to convert.
- Customer Engagement: Improving customer engagement through calls and references, especially for leads active in sending SMS, can significantly enhance conversion rates.
- Target Demographic: Focusing on working professionals, who constitute the majority of the leads, will likely yield better conversion results.

**Our logistic regression model effectively identifies high-potential leads, providing X Education with a valuable tool to enhance their lead conversion strategy and work towards the CEO's target of an 80% conversion rate.**