# Lead Score Case Study Summary

## Problem Statement

X Education, an esteemed online course provider catering to industry professionals, faces the challenge of optimizing lead conversion from a pool of daily website visitors. With courses promoted across various platforms, including Google, the company attracts professionals who explore offerings by browsing courses, completing forms, or engaging with instructional videos. Leads are identified when individuals provide contact information, and referrals further contribute to the lead pool. Following lead acquisition, the sales team initiates outreach through calls and emails. Despite this process, the typical lead conversion rate hovers around 30%.

In addressing this, X Education seeks assistance in pinpointing the most promising leads—those with the highest likelihood of transforming into paying customers. The objective is to construct a predictive model assigning lead scores, prioritizing leads based on conversion potential. The CEO envisions achieving a target lead conversion rate of approximately 80%.

Recognizing the abundance of leads in the initial stage, X Education acknowledges the importance of nurturing potential leads in the middle stage. This involves strategic education about products, constant communication, and tailored information dissemination to bolster lead conversion rates.

Tasked with this mission, you are enlisted to devise a model that assigns lead scores, differentiating customers by conversion chances. The aim is to empower the sales team with a tool that enables prioritized engagement with leads, aligning with the CEO's ambitious target for lead conversion rates.

## Approach Summary:

Step 1: Data Exploration
Begin by reading and comprehending the data, including an analysis of its structure.
Step 2: Data Cleaning:
Identify and handle variables with high NULL percentages, impute missing values, and manage outliers. Drop variables with only one value across all rows.
Step 3: Data Analysis:
Perform Exploratory Data Analysis (EDA) to gain insights into the dataset. Eliminate variables with a single value in all rows.
Step 4: Dummy Variables:
Create dummy variables for categorical features.
Step 5: Test Train Split:
Split the dataset into training and testing sets with a 70-30 proportion.
Step 6: Feature Rescaling:

Apply Min-Max Scaling to numerical variables and create an initial logistic regression model using stats models.

Step 7: Feature Selection using RFE:

Utilize Recursive Feature Elimination to select the top 15 important features, considering P-values and VIF. Choose 15 most significant variables.

Step 8: Confusion Metrics and Accuracy:

Derive probability values, apply a threshold of 0.5 for binary classification, calculate Confusion Metrics, and determine overall Accuracy. Assess Sensitivity and Specificity matrices.

Step 9: ROC Curve:

Plot the Receiver Operating Characteristic (ROC) curve, achieving an area coverage of 88%.

Step 10: Optimal Cutoff Point:

Find the optimal probability cutoff point (0.33), improving accuracy to 80% and aligning with the target lead prediction of 80%.

Step 11: Precision and Recall Metrics:

Compute Precision and Recall metrics, obtaining values of 81% and 80%, and identify a cutoff value of approximately 0.33 based on the tradeoff.

Step 12: Predictions on Test Set:

Apply learnings to the test model, calculate conversion probability, and achieve an accuracy value of 80.8% with Sensitivity at 82% and Specificity at 79%.

Conclusion:

The model exhibits stability, accuracy, and adaptability to potential future changes. Top features influencing conversion rate include last notable activity, lead origin, and current occupation.