# Summary: X Education Lead Scoring Model Development

## Introduction:

X Education is an online education company facing challenges in converting website visitors into leads and customers. To address this issue, a lead scoring model was developed using logistic regression to predict the likelihood of a website visitor/referral getting converted into a paying customer. The model was trained on historical data of leads and their conversion status, and its performance was evaluated on unseen data.

## Data Preprocessing:

The dataset provided contained both numerical and categorical features, with missing values in some columns. Columns with more than 30% missing values were dropped, and missing values in other columns were imputed using appropriate strategies. Irrelevant columns like 'Prospect ID' and 'Lead Number' were removed. Categorical variables were encoded into dummy variables for modelling purposes.

## Data Analysis and Feature Selection:

Exploratory Data Analysis was performed to gain insights into the distribution and summary statistics of numeric variables, as well as the value counts of categorical variables. Recursive Feature Elimination (RFE) was used to select the most relevant features for the logistic regression model.

## Model Building:

The dataset was split into training and testing sets to build the logistic regression model. Feature scaling was applied to numeric columns to ensure uniformity in scale.

## Logistic Regression and Model Evaluation:

A logistic regression model was built using the selected features from RFE and fitted on the training data. The model's performance was evaluated on the testing data using metrics such as confusion matrix, classification report, ROC-AUC score, accuracy, sensitivity, and specificity. The ROC curve was plotted to decide on the optimal probability threshold for classifying leads as converted or not. Using the optimal threshold, accuracy and sensitivity metrics were calculated.

## Modelling Solution:

The developed logistic regression model showed promising results in predicting the likelihood of lead conversion. The model achieved an area under the ROC curve of 0.87, indicating good discriminatory power. By focusing on leads with higher conversion probabilities, X Education can optimize their sales efforts and improve lead conversion rates. The model can also be further fine-tuned and optimized for better performance.

## Conclusion:

The lead scoring model based on logistic regression offers valuable insights into the conversion probability of website visitors. The sensitivity and accuracy are within the CEO’s ballpark of 80% conversion. This data-driven approach will enable X Education's marketing and sales teams to prioritize leads and tailor their strategies accordingly, resulting in increased lead conversion and improved business performance. Continuous monitoring and refinement of the model will help maintain its relevance and effectiveness over time.