# Summary Report on Lead Scoring Case Study

## Introduction

The Lead Scoring Case Study aims to address the inefficiencies in the lead conversion process for X Education, an online education company. The primary objective is to develop a machine learning model to predict the likelihood of leads converting into paying customers, thereby optimizing the sales team's efforts.

## Problem Statement and Business Context

X Education faces a significant challenge with a low lead conversion rate of approximately 30%. The company generates leads through various channels, including websites, search engines, and referrals. However, the sales team spends considerable time contacting leads, most of which do not convert. The goal is to identify 'Hot Leads' to improve the conversion rate by focusing efforts on the most promising leads. Achieving this objective will enhance the efficiency of the sales process and ultimately improve business performance.

## Analysis Approach

The analysis approach involved several key steps:
1. Data Understanding: The dataset included 36 variables, with the target variable being 'Converted' (0 for non-converted leads, 1 for converted leads). The variables consisted of unique IDs, numeric values, and categorical variables, some of which had default values requiring treatment.
2. Exploratory Data Analysis (EDA): Both univariate and bivariate analyses were conducted to identify trends and relationships between variables and the target variable.
3. Data Preparation: This phase included data cleaning (handling missing values and outliers) and feature engineering (creating new features and standardizing variables). The data was then split into training (70%) and test (30%) datasets.
4. Model Building: Logistic regression was chosen as the appropriate model for this classification problem. Recursive Feature Elimination (RFE) was used to select the top 15 variables, followed by hyperparameter tuning to address multicollinearity and statistical significance.

## Model Evaluation and Performance

The model's performance was evaluated using several metrics:
- *Accuracy*: 81.7% on the training set and 79.7% on the test set.
- *Recall*: 85.0% on the training set and 86.5% on the test set.
- *Precision*: 72.0% on the training set and 67.0% on the test set.

The optimum cut-off threshold for the model was determined to be 0.3, balancing precision and recall to ensure the detection of as many potential 'Hot Leads' as possible. The model's ROC curve and Precision-Recall trade-off curve were used to validate these results.

## Model Interpretation

Key features influencing lead conversion included:
1. Time Spent on Website: Positively correlated with conversion likelihood.
2. Pages Viewed per Visit: Negatively correlated with conversion likelihood.
3. Lead Quality and Origin: Significant categorical variables impacting lead conversion.

## Conclusion

The implementation of the logistic regression model successfully addresses X Education's lead conversion problem by identifying high-potential leads. This approach optimizes the sales team's efforts, increasing efficiency and potentially boosting the lead conversion rate above the current 30%. By focusing on 'Hot Leads,' X Education can improve resource allocation, reduce wasted efforts, and enhance its market position. This strategic improvement will not only benefit the company's operational efficiency but also contribute to sustained growth and competitiveness in the online education sector.

**Learnings**

The case study provided several key insights:
1. Importance of Data Preparation: Proper data cleaning and feature engineering are crucial for building an effective model.
2. Model Selection and Tuning: Logistic regression, combined with feature selection and hyperparameter tuning, can effectively handle classification problems with categorical and numerical data.
3. Evaluation Metrics: A comprehensive evaluation using multiple metrics (accuracy, recall, precision) and validation tools (ROC curve, Precision-Recall curve) ensures robust model performance.
4. Business Impact: Data-driven decision-making significantly enhances business strategies, highlighting the importance of predictive modelling in optimizing sales processes.

This comprehensive approach to lead scoring demonstrates the power of data analytics in transforming business operations and achieving strategic goals.