# Summary: Lead Scoring Case Study

The Lead Scoring Case Study for X-Education, aims to develop a logistic regression model to identify potential leads. The primary goals are to improve lead conversion rates and achieve an 80% enrolment target. The study involves:

1. **Understanding the Problem and Data**

The initial phase focuses on understanding the problem and the data dictionary. The dataset comprises 9,240 columns and 37 rows.

2. **Importing Required Libraries**

The study utilizes various libraries including numpy, pandas, matplotlib.pyplot, seaborn, and several sklearn modules like train_test_split, StandardScaler, and RFE. These libraries facilitate data manipulation, visualization, and model building.

3. **Data Understanding**

The data understanding phase involves reading the data, checking the dataset's information, and performing basic statistical analysis. This helps in gaining insights into the structure and distribution of the data.

4. **Data Quality Checking and Cleaning**

Data quality is ensured by checking for duplicates and replacing 'SELECT' values with null values. Columns with single values and unnecessary data were also removed. Data imputation and outlier treatment were performed to refine the dataset.

5. **Exploratory Data Analysis**

Key findings from the exploratory data analysis include:

- A lead success rate of 62.14%.
- Google and direct traffic sources have the highest lead conversion counts.
- The highest lead conversion rates are from Reference and Welingak websites.
- India, UAE, and the USA are the top countries for lead conversions.
- Better career prospects are the most important factor in choosing a course.

**6. Splitting Data: Test and Train Set**

Dummy variables were created before splitting the data into train and test sets. The dataset is then scaled to ensure all values are between 0 and 1, maintaining a similar scale and avoiding imbalance.

7. **Model Building**

A logistic regression model is built, with the top 15 columns selected by RFE. The model 'logm2' has a p-value below 0.05 and a VIF below 5, indicating a good fit.

### 8. Making Predictions on Train Dataset

Predictions were made on the train set, creating a dataframe with actual and predicted values.

### 9. Finding Optimal Cutoff

Various probability cutoffs are analyzed to find the optimal cutoff, which is determined to be 0.33 based on accuracy, sensitivity, and specificity.

### 10. Prediction on Test Data

Predictions are made on the test dataset with added constants.

### 11. Model Evaluation

The model performs well with a sensitivity of 0.82, specificity of 0.84, precision of 0.75, false positive rate of 0.15, and negative predictive value of 0.89. The sensitivity and accuracy values for both test and train datasets are approximately 84% and 85%, respectively, indicating good model performance.

### 12. Determining Top Features and Recommendations

Key features identified include:

- Focus on the Welingak website for more leads.
- Prioritize the "Lead Add Form" origin for higher conversion likelihood.
- Utilize phone conversations as they can lead to business.
- Target working professionals for higher conversions.

This comprehensive analysis provides a strategic roadmap for X-Education to enhance lead conversions and achieve its enrolment targets.