# Lead Scoring Case Study Summary

## Overview

X Education Institute aimed to optimize its sales strategy by implementing a lead scoring system. The goal was to identify high-potential leads and improve conversion rates through targeted efforts. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.


As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc. ) in order to get a higher lead conversion.

## Data Cleaning:

The data was partially cleaned, addressing null values and replacing "select" options with null values. Some null values were updated to "not provided" to retain data integrity. During dummy variable creation, these null values were removed. Additionally, elements pertaining to location were standardized as "India", "Outside India", and "not provided".

## Exploratory Data Analysis (EDA):

A brief EDA was conducted to assess the dataset's condition. Irrelevant elements within categorical variables were identified, while numeric values showed no outliers.

## Dummy Variables:

Dummy variables were generated, and those containing "not provided" elements were subsequently removed. For numeric values, the MinMaxScaler was applied.

## Train-Test Split:

Data was split into 70% for training and 30% for testing.

## Model Building:

Initially, Recursive Feature Elimination (RFE) was employed to select the top 15 relevant variables. Subsequently, variables were manually removed based on VIF values and p-values, retaining those with VIF < 5 and p-value < 0.05.

## Model Evaluation:

A confusion matrix was constructed, and the optimal cutoff value (determined via ROC curve) was utilized to ascertain accuracy, sensitivity, and specificity, each approximating 80%.

## Prediction:

Predictions were made on the test dataset using an optimal cutoff of 0.35, resulting in 80% accuracy, sensitivity, and specificity.

## Precision-Recall:

Precision-recall analysis was conducted, revealing a cutoff of 0.41 with precision at approximately 73% and recall at around 75% on the test dataset.

## Key Variables Influencing Potential Buyers

Upon analysis, the variables deemed most influential in identifying potential buyers, listed in descending order of importance, are as follows:

1. **Total Time Spent on the Website**
2. **Total Number of Visits**
3. **Lead Source, Prioritized as:**
    - a. Google

- b. Direct Traffic
- c. Organic Search
- d. Welingak Website

4. **Last Activity, Particularly:**
    - a. SMS
    - b. Olark Chat Conversation
5. **Current Occupation as a Working Professional**

## Significance:

Understanding the prominence of these variables allows X Education to capitalize on their potential buyer base effectively. By leveraging these insights, X Education stands a strong chance of persuading nearly all potential buyers to convert and enroll in their courses. This strategic alignment ensures X Education's continued growth and success in the competitive online education market.