Lead Scoring Case Study Summary

This analysis is focused on understanding how X Education can attract more industry professionals to enrol in their courses. The data provided indicates how potential customers are interacting with the website, including key insights such as:

- **How Customers Visit the Site**: This tells us about the different channels through which potential customer's find X Education's website. It could include search engines, email campaigns, SMS's or direct visits. Understanding these sources helps identify where to focus marketing efforts.
- **Time Spent on the Site**: This data shows how long visitors stay on the website. A longer time on the site might indicate that the content is engaging or that visitors are exploring different course offerings.
- **How Visitors Reached the Site**: This part of the data focuses on the entry points—whether they clicked on an ad, found the site through organic search, or followed a link.
- **Conversion Rate**: This metric reveals how many visitors take the desired action, such as signing up for a course, compared to how many just browse the site. A low conversion rate might indicate that visitors are interested but need additional encouragement to enrol.

The entire case study was conducted using the following steps:

1. Data Cleaning:

The data was mostly clean, but a few issues needed to be addressed. Some values were missing, and instead of leaving them blank, they were replaced with "not provided" to avoid losing any information. Entries with "select" as an option were also changed to null values since they didn't provide meaningful data. The location data was grouped into three categories: "India," "Outside India," and "Not Provided" to simplify the information.

2. Exploratory Data Analysis (EDA):

A quick check of the data showed that some of the categories were irrelevant and unnecessary. The numerical data was in good shape, with no outliers or unusual values that could affect the analysis.

3. Dummy Variables:

Categorical data was converted into dummy variables to make it usable for machine learning models. Any "not provided" categories were removed to keep the data clean. For the numerical data, MinMaxScaler was applied to scale the values between 0 and 1, ensuring that the features were on the same scale for better model performance.

4. Train-Test Split:

The dataset was split into two parts: 70% for training the model and 30% for testing its performance. This helps ensure that the model learns from one set of data and is evaluated on a separate, unseen set of data to avoid overfitting.

5. Model Building:

Recursive Feature Elimination (RFE) was used to select the top 15 most important features for the model. Additional features were removed based on their statistical values (VIF and p-value) to keep only the most relevant data that contributed to the model's predictions.

6. Model Evaluation:

The model's performance was evaluated using a confusion matrix, which shows how many predictions were correct and how many were incorrect. Using this, the model's accuracy, was found to be around 80%, indicating that the model performed well and made accurate predictions.

The following can be inferred from the analysis:

- X Education needs to focus on Total Visits, Total Time Spent on Website, Page Views Per Visit as these are the top variables that contribute the most towards the probability of a lead getting converted.
- The top three lead sources are: Direct Traffic, Google & Organic Search. The sales team can look into potential customer who use the above three to get potential conversions.
- The sales team should also focus on working professionals as they have a high probability of getting converted. Avoiding those users whose last activity was a long time ago and those who spend the least amount of time on the websites can help the sales team focus more on those users that have a higher potential conversion.