Summary Report

This project aimed to help X Education improve how they convert leads into paying customers by building a logistic regression model. The model predicts the chances of a lead turning into a customer. Here's how we did it and what we learned along the way.

1. Data Understanding and Preprocessing

First, we analyzed the data, which included information like where the leads came from, how much time they spent on the website, and what actions they took. The target was the "Converted" variable, showing whether the lead became a customer (1 for converted, 0 for not).

Before we could build the model, we had to clean the data. This involved:

- **Handling missing values:** Some columns had missing information or "Select" values that weren't useful. We either removed those or replaced them with appropriate values.
- **Converting categorical variables:** Some columns had text-based data (like "Google" or "Direct Traffic"). We turned these into numbers using one-hot encoding to make them usable in the model.

2. Exploratory Data Analysis (EDA)

Next, we performed EDA to understand how different factors influenced lead conversion. We looked at important variables like "Lead Source," "Total Time Spent on Website," and "Page Views Per Visit." For example, we found that leads who spent more time on the site and viewed more pages were more likely to convert.

3. Building the Logistic Regression Model

After cleaning the data, we used a logistic regression model to predict the probability of lead conversion. Logistic regression was a good choice because it helps in predicting two possible outcomes (like 0 or 1 for conversion).

We used Recursive Feature Elimination (RFE) to choose the top 15 most important variables for the model. This helped simplify the model by keeping only the most relevant variables.

4. Evaluating the Model

Once the model was ready, we tested it on new, unseen data (the test set). We checked how well it performed using metrics like accuracy, precision, and recall. The model gave us a "lead score" (the chance of conversion), which helped us identify leads that were more likely to convert.

5. Insights and Learnings

- **Top contributing factors:** The model showed that "Total Time Spent on Website," "Page Views Per Visit," and "Lead Source" (like Google) were the biggest factors in deciding whether a lead would convert.
- **Handling categorical data:** We learned how important it was to convert text-based data into numbers to make it usable in the model.
- **Model performance:** The model helped focus sales efforts on leads more likely to convert, improving efficiency.

6. Recommendations

- **Aggressive conversion strategy:** During peak periods, X Education should focus on leads with the highest chances of converting and follow up with quick phone calls.
- **Conservative approach:** Once the company hits its targets, the sales team should only focus on leads with very high chances of converting, while lower-priority leads can be handled with automated emails or SMS.

Overall, this project helped us learn how to clean data, build a predictive model, and use it to guide business decisions. The main takeaway is that by focusing on high-potential leads, X Education can improve conversion rates and use resources more effectively.