This analysis focuses on building and evaluating a logistic regression model to identify potential leads for an education company named X Education. The primary objective is to identify "hot leads"—those most likely to convert into paying customers. The target is to achieve a lead conversion rate of around 80%. The model assigns a lead score between 0 and 100, which the company can use to target potential leads more effectively.

Data Preparation and Exploratory Data Analysis (EDA):

In the data preparation phase, various steps were taken to clean and organize the data. Columns were renamed for simplicity, and irrelevant columns were dropped. Categorical variables were cleaned using regular expressions, and missing values in numerical variables were imputed. Some columns contained outliers, such as `TotalVisits` and `Page Views Per Visit`, which were handled using percentiles to provide a more consistent data distribution.

The EDA revealed that the majority of leads come from "Landing Page Submission" and "API," with primary lead sources being "Google" and "direct traffic." Interestingly, most customers were open to receiving communications about the courses, with over 90% saying "No" to "Do Not Email" or "Do Not Call" options. However, most advertising platforms seemed ineffective, as almost 100% of respondents indicated they hadn't seen ads before.

A strong insight from the EDA is that converted leads spent significantly more time on the website than non-converted leads. Moreover, factors like the `Lead Origin_Lead Add Form`, `Lead Source_referral`, and `Last Activity_SMS Sent` were identified as strong indicators of promising leads. Other factors such as occupation and course preferences also played a significant role in predicting lead conversion.

Model Building:

A logistic regression model was constructed, and one key metric used to evaluate its fit was the **Pseudo R-squared (CS)**, which had a value of **0.5487**. This indicates that the model explains about 54.87% of the variance in the dependent variable (`Converted`), suggesting a reasonably good fit.

Most variables in the model had p-values close to zero, indicating that they were statistically significant predictors of conversion. The most significant features included:

- `Do Not Email`: a strong negative predictor.
- `Total time on website`: a strong positive predictor.
- `Lead Origin_Lead Add Form`: another significant positive predictor.
- `Tags_Will revert after reading the email`: strongly associated with a higher likelihood of conversion.
- `Last Notable Activity_SMS Sent`: also a strong positive indicator.

Model Evaluation:

The performance of the model was evaluated using a variety of metrics, including accuracy, sensitivity (true positive rate), and specificity (true negative rate). By experimenting with different probability thresholds, the optimal cutoff point of **0.28** was selected, which provided a balance between accuracy, sensitivity, and specificity. At this threshold, the model achieved:

- **Accuracy**: 0.88 on the training set and 0.87 on the test set.
- **Sensitivity and Specificity**: Both were balanced at 0.88 for the training set and slightly lower for the test set.

This minimal drop in performance from the training to the test set indicates a robust and reliable model. The model's confusion matrix confirmed its ability to distinguish between converted and non-converted leads effectively, showing only a slight misclassification.

Conclusion:

The logistic regression model developed for X Education provides a solid foundation for identifying potential leads likely to convert into paying customers. The model balances both business requirements and statistical rigor, achieving high accuracy and reliability. Its application in lead scoring could significantly improve the company's targeting efforts and overall conversion rates.