In this case study, we were tasked with helping X Education, an online education company, increase its lead conversion rate by identifying the most potential leads or "Hot Leads." To achieve this goal, we built a logistic regression model to assign a lead score between 0 and 100 to each lead, allowing the company to focus its sales efforts on the most promising prospects. The CEO's target conversion rate was around 80%.

We began by analyzing a dataset of around 9000 leads, which included various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, and more. The target variable was "Converted," with 1 indicating a converted lead and 0 indicating otherwise. We also addressed the issue of "Select" levels in categorical variables, treating them as null values.

The logistic regression model we built yielded promising results. We identified several key factors that significantly influence lead conversion:

Tags_Will revert after reading the email: Leads originating from the Tags_Will revert after reading the email have a higher probability of conversion, as indicated by the positive coefficient of 4.55.

Lead Source - Welingak Website: Leads from the Welingak Website also show a strong likelihood of conversion, with a coefficient of 2.3084.

Last Activity - Had a Phone Conversation: Engagement with a phone conversation significantly increases the chance of conversion, as evidenced by the coefficient of 2.9971.

What is your current occupation - Working Professional: Leads from working professionals are more likely to convert, with a coefficient of 2.6210.

Tags - Closed by Horizzon and Tags - Lost to EINS: Specific tags, such as "Closed by Horizzon" and "Lost to EINS," are strong indicators of potential conversion, with coefficients of 8.5018 and 9.4256, respectively.

On the other hand, some factors decrease the likelihood of conversion, such as 'Tags - Ringing' and 'Tags - switched off,' which have negative coefficients.
A VIF analysis was performed, indicating low multicollinearity, reinforcing the model's robustness.

We further assessed the model's performance using various metrics. With a threshold of 0.5, the model achieved an 89% accuracy. The sensitivity (true positive rate) was 79.6%, and the specificity (true negative rate) was 96%, indicating a well-balanced model.

To improve the model's precision and recall, we employed Receiver Operating Characteristic (ROC) curve analysis. The area under the ROC curve (AUC) was 94, signifying a robust model. We found that adjusting the threshold to 0.3 optimized both precision and recall.

To explore the precision-recall trade-off, we experimented with thresholds ranging from 2.5 to 3. A threshold of 2.5 yielded a good balance between precision (92.8%) and recall (79.6%).

We applied this threshold to the test dataset, and the model's performance closely mirrored that of the training set.

In conclusion, the logistic regression model we developed provides a lead score for potential customers, helping X Education identify "Hot Leads" with a higher likelihood of conversion. By focusing their efforts on these leads, the company can work towards achieving the CEO's target conversion rate of 80%.