Lead Score Case Study Summary

The challenge was to increase the lead conversion rate at X Education, which currently stands at 30%, with a target of raising it to 80%. The objective was to develop a logistic regression model to assign a lead score between 0 and 100 to each lead, allowing the company to focus on high-potential leads. A higher score indicates a lead is likely to convert, while a lower score suggests the opposite.

First, we imported the dataset into a pandas DataFrame and addressed null values and 'select' entries. We then conducted Exploratory Data Analysis (EDA) to gain insights into the data. Using Recursive Feature Elimination (RFE), we identified the 16 most significant variables.

We developed a logistic regression model and assessed its accuracy and other metrics. Variables with higher p-values, indicating significant impact on conversion rates, were refined, resulting in our final model. The model achieved an overall accuracy of 85%, with a sensitivity of 75%, specificity of 90%, and a false positive rate of 10%.

A ROC curve analysis helped us determine an optimal cutoff point of 0.35 for predicting lead conversion. We validated this model on both training and test datasets. Key variables such as "lead_origin_lead_add_form," "lead_source_welingak_website," and "lead_source_olark_chat" were found to significantly boost conversion rates. Conversely, variables like "lead_quality_worst," "lead_quality_not_sure," and "lead_quality_might_be" negatively impacted conversion rates.