Lead Scoring Case Study Summary

Problem Statement

X Education seeks a solution to identify the most promising leads, focusing on those likely to convert into paying customers. The task involves creating a model that assigns a lead score, with higher scores indicating better chances of conversion. The company's CEO expects a lead conversion rate target of around 80%.

Data Analysis and Preprocessing

1. Data Inspection:

The analysis began by understanding the significance of each variable and how they relate to the target (lead conversion). Variables like `Specialization`, `Lead Profile`, and `How did you hear about X Education?` were thoroughly examined for their impact.

2. Handling Null Values:

Several columns were found to contain values marked as "select," equivalent to null values. These were treated as nulls and removed if deemed unimportant. Columns like `Prospect ID`, `Lead Number`, `Last Activity`, and others were discarded due to irrelevance or high null values.

3. Standardizing Data:

Certain categorical variables, such as `Country`, `Current Occupation`, and `Lead Origin`, had numerous subcategories, many with very few occurrences. To simplify the analysis, these small subcategories were grouped under "Others".

4. Outliers:

Large differences between the 99th percentile and maximum values were detected in variables like `Total Visits` and `Page Views Per Visit`. Outliers were removed using boxplots, affecting less than 100 rows.

5. Handling Correlations:

High positive or negative correlations were identified using a heat map, leading to feature selection to avoid multicollinearity in the model.

Model Building and Evaluation

- **Model Choice:**

Logistic regression was used for the model, leveraging libraries like `statsmodel` and `Scikit-learn`. The evaluation was based on metrics such as accuracy, sensitivity, specificity, and ROC curve analysis.

- **Key Features:**

Important features contributing to lead conversion included:

- `Current Occupation` (Unemployed, Working Professional, Student)

- `Lead Quality` (Might Be, Not Sure, Worst)

- `Lead Source` (Olark Chat, Reference, Welingak Website)

- `Tags` (Will Revert After Email, Ringing, Interested in Other Courses)

- `Last Notable Activity` (SMS Sent)

- **Performance Metrics:**

For leads with a conversion probability greater than 80%, the model's accuracy was 87.3%, with sensitivity at 69.3% and specificity at 98.2%. The false positive rate stood at 1.8%, and predictive values were favourable (96% for positive, 84% for negative).

The **ROC curve** indicated good model performance, with the optimal cutoff probability for balanced sensitivity and specificity determined at 0.35. The model showed an accuracy of 89.3% with this cutoff.

Results

The model demonstrated strong results, showing an alignment between the training and test sets with an accuracy of 88%. Sensitivity was 70.8%, and specificity was 98.3%, indicating the model's robustness.

Recommendations

1. Focus Areas:

The company should prioritize leads with certain positive and negative contributors. Positive factors include leads sourced from the Welingak website or reference and leads where the person is employed or unemployed. Negative factors include leads tagged as "interested in other courses" or "ringing."

2. Different Scenarios:

- Interns chasing leads: Leads with a probability greater than 0.35 should be targeted, balancing quantity and accuracy.

- High-priority leads: For a more focused approach, leads with a probability above 0.9 should be pursued to minimize time spent on less promising leads while maintaining high productivity.