Case Study Summary: Lead Scoring Model

1. Data Cleaning

- Initial State: The data was mostly clean but had some null values and an unclear 'select' option.

- Fixes: Changed some nulls to 'not provided' and adjusted categories to 'India', 'Outside India', and 'not provided'.

2. Exploratory Data Analysis (EDA)

- Findings:

  - Some irrelevant data in categorical fields.

  - Numeric data looked good with no unusual values.

3. Dummy Variables & Scaling

- Dummies: Created dummy variables and removed those with 'not provided'.

- Scaling: Used MinMaxScaler to standardize numeric values.

4. Train-Test Split

- Ratio: Split data into 70% for training and 30% for testing.

5. Model Building

- Feature Selection:

  - Used RFE to pick the top 15 variables.

  - Removed additional variables based on their impact (VIF and p-values).

6. Model Evaluation

- Metrics:

  - Created a confusion matrix.

  - Used the ROC curve to find the best cut-off.

  - Achieved about 80% accuracy, sensitivity, and specificity.

7. Prediction

- Test Data Results:

  - With a cut-off of 0.35, the model maintained around 80% accuracy, sensitivity, and specificity.

8. Precision-Recall Analysis

- Cut-off Value: Adjusted cut-off to 0.41.

- Metrics:

  - Precision: ~73%

  - Recall: ~75%

Key Takeaways

1. Model Choice: Logistic regression was a good fit for the lead scoring task.

2. Lead Conversion:

   - Many leads are early-stage; only a few become customers.

   - Most leads come from India, particularly Mumbai.

3. Data Collection:

   - Make certain fields mandatory to get useful data.

   - Key details include occupation and specialty.

4. Engagement Insights:

   - More visits and time spent can lead to better conversion rates.

   - Leads focused on career advancement, especially in finance, are more likely to convert.

5. Client Interaction:

   - Email engagement helps; those who open emails are more likely to convert.

   - Sending SMS might also help.

6. Target Segment:

   - Many leads are unemployed, so focusing on this group might be beneficial.

This summary provides an overview of the lead scoring model process and highlights the importance of data handling, model evaluation, and engaging with potential clients.