Lead Scoring Case Study Summary

1. Introduction:

- Objective: Enhance lead conversion process by implementing a data-driven lead scoring model.
- Current Challenge: Low lead conversion rate.

2. Data Overview:

- Dataset: Approximately 9000 data points with attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.
- Target Variable: 'Converted' indicating whether a lead was converted (1) or not (0).

3. Data Preprocessing:

- Handled 'Select' levels in categorical variables.
- Explored and cleaned the dataset for missing values, outliers, and inconsistencies.
- Dropped columns with a high percentage of missing values.
- Imputed or dropped remaining missing values.

4. Exploratory Data Analysis (EDA):

- Analyzed distribution and imbalance in categorical variables.
- Explored univariate and bivariate relationships.
- Handled outliers in numerical variables.
- Heatmap for numerical vs target column.

5. Logistic Regression Model:

5.1 Feature Engineering:

- Created dummy variables for categorical columns.
- Handled dummy variable trap by dropping one level.
- Split the dataset into train and test set (70% & 30% respectively).

5.2 Model Building:

- Utilized Logistic Regression to predict the probability of leads getting converted.
- Iteratively refined the model by removing variables with high p-values and VIF until finding variables with VIF less than 5 and p-value less than 0.05.

6. Model Evaluation:

- Assessed performance using metrics such as confusion matrix , accuracy, recall, precision, sensitivity, and specificity.
- Plotted ROC curve to find the optimal cutoff for predictions.

7. Model Refinement:

- Conducted precision-recall trade-off analysis to find an optimal cutoff.

8. Final Model:

- Achieved a balanced model with improved precision and recall.
- Optimal Cutoff: 0.44.

9. Model Evaluation (Final Output):

- **FINAL TRAIN DATASET OUTPUT:**
  - Accuracy = 0.7845
  - Precision = 0.7840
  - Recall = 0.7771
  - Sensitivity = 0.7771
  - Specificity = 0.8010
- **FINAL TEST DATASET OUTPUT:**
  - Accuracy = 0.7866
  - Precision = 0.7829
  - Recall = 0.7675
  - Sensitivity = 0.7675
  - Specificity = 0.8042

	Features	VIF
9	What is your current occupation_Unemployed	2.82
1	Total Time Spent on Website	2.00
0	TotalVisits	1.54
7	Last Activity_SMS Sent	1.51
2	Lead Origin_Lead Add Form	1.45
3	Lead Source_Olark Chat	1.33
4	Lead Source_Welingak Website	1.30
5	Do Not Email_Yes	1.08
8	What is your current occupation_Student	1.06
6	Last Activity_Had a Phone Conversation	1.01
10	Last Notable Activity_Unreachable	1.01

Recommendations and Areas for Improvement:

1. Due to significant skewness, the 'Country' column was excluded from the model. Balancing the data could provide valuable insights.

2. The 'Specialisation' column with over 36% missing values should be addressed to enhance meaningful insights.

3. Explore advertising on popular Social Media platforms, particularly among the younger demographic, to enhance marketing effectiveness.

4. Replicate successful strategies observed in metropolitan areas like Mumbai and Thane in other similar demographics.

5. Offering introductory course content for free could identify potential 'hot' leads, increasing the likelihood of enrollment.

6. Address the low conversion rate among students by breaking courses into affordable segments.

7. Direct users opting out of email updates to a questionnaire to gather insights for tailored disengagement or targeted approaches.