Summary Report on "Lead Scoring Case Study" Assignment

1. Reading & Understanding the Dataset

- **Objective**: Gain initial insights into the dataset's structure and quality.
- **Rationale**: Employing methods like head(), describe(), and info() was crucial for identifying data issues such as missing values or irrelevant features. This step set the stage for effective data cleaning and preprocessing.

2. Data Cleanup

- **Objective**: Ensure data quality and reliability.
- **Rationale**: Handling missing values, removing duplicates, and standardizing data types to maintain data integrity for an accurate machine learning model.

3. Data Preparation

- **Objective**: Transform the dataset for modelling.
- **Rationale**: Scaling numeric variables, encoding categorical variables, and feature engineering to enhance model suitability and pattern capture.

4. Model Building

- **Objective**: Construct a predictive model for lead scoring.
- **Rationale**: Selecting Logistic Regression for its effectiveness in binary classification and employing Recursive Feature Elimination (RFE) for optimal feature selection.

5. Model Evaluation

- **Objective**: Assess the model's predictive accuracy and reliability.
- **Rationale**: Utilizing metrics like confusion matrix, precision, and recall for a comprehensive evaluation, crucial for real-world applicability in lead identification.

6. Model Performance

- **Outcome**: Demonstrated high predictive accuracy and reliability.
- **Detailed Analysis**:
    - Accuracy: The model achieved a high accuracy rate, indicating its effectiveness in correctly classifying leads as convertible or non-convertible.
    - Precision: Train Data (93.9%) & Test Data (92.3%): The model demonstrates high precision both in training and testing phases. This indicates that a significant majority of the leads it classifies as likely to convert (positive) are indeed correct predictions. High precision ensures efficient use of resources by focusing efforts on leads most likely to result in successful conversions.
    - Recall: **Train Data (87.1%) & Test Data (84.9%)**: The model also shows high recall, meaning it successfully identifies a large proportion of actual convertible leads. This high recall rate indicates the model's effectiveness in

capturing most potential opportunities, thereby reducing the chances of missing out on viable leads.

- Confusion Matrix Analysis: The confusion matrix provided deeper insights into the true positives and false negatives, showcasing the model's strength in minimizing incorrect predictions.
- Business Impact: The model's performance translates into significant business implications, such as optimizing resource allocation for high-potential leads and enhancing overall sales efficiency.

7. Conclusion and Learnings

- **Objective**: Derive actionable business insights from the model.
- **Rationale**: Highlighting the importance of comprehensive data preparation, strategic model selection, and in-depth evaluation, the project underscored the role of data-driven decision-making in business optimization.