### **Introduction**

This report summarizes the approach taken to develop a logistic regression model for lead scoring. The goal was to assign a lead score between 0 and 100, helping the company prioritize potential leads based on their likelihood of conversion. The process involved data understanding, preprocessing, model building, evaluation, and deriving business insights.

### **Approach**

#### **1. Data Understanding & Preparation**

The first step was loading and analyzing the dataset to identify missing values, duplicates, and inconsistencies. The dataset contained a mix of categorical and numerical variables. Data quality checks ensured that all inconsistencies were handled appropriately. Missing values were either imputed or removed, depending on their impact on the model.

Categorical variables were transformed using encoding techniques to ensure compatibility with the logistic regression model. Numerical variables were standardized using StandardScaler() to improve model performance. The target variable, ‘Opportunity Status,’ was mapped to binary values (1 for Won and 0 for Loss) for classification.

#### **2. Exploratory Data Analysis (EDA)**

EDA was performed to identify relationships between independent variables and the target variable. Feature importance analysis helped in selecting the most significant predictors. Correlation analysis and visualizations such as histograms, box plots, and pair plots were used to understand feature distributions.

#### **3. Model Building**

A logistic regression model was implemented using Scikit-Learn. The dataset was split into training (70%) and testing (30%) sets. The logistic regression model was trained on the preprocessed data, and predictions were generated for the test set.

#### **4. Model Evaluation**

The model was evaluated using multiple metrics:

* **Accuracy Score:** Measured the overall correctness of predictions.
* **Confusion Matrix:** Provided insight into false positives and false negatives.
* **Classification Report:** Assessed precision, recall, and F1-score.
* **ROC AUC Score:** Evaluated the model’s ability to distinguish between lead conversions.
* **ROC Curve:** Visualized the trade-off between sensitivity and specificity.

The evaluation showed that the model performed well in distinguishing between high and low-quality leads. However, some misclassifications indicated the potential for improvement through feature engineering or alternative modeling approaches.

### **Learnings and Insights**

1. **Data Quality is Crucial:** Handling missing values, duplicates, and categorical encoding significantly impacted model accuracy.
2. **Feature Selection Matters:** Some features had a higher predictive power, while others added noise. Identifying key features helped in improving the model’s performance.
3. **Standardization Enhances Performance:** Standardizing numerical features prevented issues related to scale differences.
4. **Evaluation Metrics are Key:** Relying solely on accuracy can be misleading. The ROC AUC score provided a more reliable measure of model effectiveness.
5. **Business Implications:** A well-optimized lead scoring model can help businesses allocate resources effectively, increasing conversion rates and optimizing sales efforts.

### **Conclusion**

The logistic regression model developed in this assignment provides a solid foundation for lead scoring. The learnings from this project emphasize the importance of data preparation, feature selection, and model evaluation in predictive modeling. Future improvements could include experimenting with more advanced models such as decision trees or neural networks to enhance accuracy further.