Summary Report: Lead Scoring Assignment

This lead scoring assignment provided valuable hands-on experience in handling real-world data. The objective was to analyze historical data and build a predictive model to classify leads based on their likelihood of conversion. The approach involved key steps: data exploration, preprocessing, feature selection, model building, and evaluation.

Approach and Methodology

Data Exploration & Preprocessing
We started by understanding the dataset using info(), describe(), and head(). Handling missing values was a major challenge columns with over 35% missing data were dropped, while others were imputed using mean, median, or mode, depending on the data type. Duplicate records and irrelevant columns were also removed to improve data quality.

Feature Selection
Since the dataset contained categorical variables, we used dummy variables for each of the required columns to convert them into a machine-learning format. Feature selection was crucial, as too many features can introduce noise and slow down training. We applied Recursive Feature Elimination (RFE) to identify the most important predictors and used the Variance Inflation Factor (VIF) to detect and remove multicollinear features. This helped streamline the dataset, improving accuracy and interpretability.

Splitting and Scaling
After cleaning and transforming the data, we split it into training and testing sets using train_test_split() to ensure unbiased evaluation. Since numerical features had varying scales, we applied MinMax Scaling to normalize the data which gets the data in the range 0 to 1, ensuring that no feature dominated the learning process and making training more stable and efficient.

Model Building and Evaluation
We implemented Logistic Regression as our initial model for predicting lead conversion. We used Generalized Linear Models(GLM) and the response variable follows a binomial distribution. To assess performance, we used key metrics:

1. Accuracy - Measures the overall correctness of predictions.
2. Precision - Indicates how many predicted positives were correct.
3. Recall - Measures how many actual positives were correctly identified.
4. Specificity - Assesses how well the model identifies actual negatives.
5. Sensitivity - Assesses how well the model identifies actual positives.

While the model performed well, we saw a tradeoff between precision and recall. For business applications, precision is often prioritized to focus on high-quality leads, but this may reduce recall, meaning potential customers can be missed. Getting the correct balance was one of the most important understandings from this project.

Key Learnings

- Data Cleaning is Essential - Proper handling of missing values, creating dummy variables, and removing unnecessary features improved model accuracy and efficiency.
- Feature Selection Matters - Using RFE and VIF helped to eliminate redundant features, making the model more streamlined and interpret-able.
- Precision vs. Recall is a Business Decision - A high precision model focuses on quality leads, while a high-recall model ensures more leads are captured. The choice depends on specific business goals.
- Accuracy is Not the Only Metric - In the beginning, we thought accuracy was the most important measure. However, precision and recall gave deeper insights, especially when dealing with not properly balanced datasets like lead scoring.