**Lead Score Case Study: Logistic Regression**

**Objective**

The objective of this case study is to develop a logistic regression model to predict which leads are most likely to convert into customers. This model will enable the sales team to prioritize high-potential leads, especially during periods of intensive conversion efforts.

**Technical Steps**

**1. Data Understanding and Preprocessing**

* **Data Cleaning**:
  + **Dropping Irrelevant Columns**: Removed non-informative columns that do not contribute to the model’s predictive power.
  + **Handling Missing Values**: Eliminated features with significant missing data and applied imputation techniques for other features with fewer missing values. Rows with minimal impact were dropped as necessary.
  + **Categorical Feature Processing**: Identified and converted categorical features into numerical representations (dummy variables) to make them compatible with the logistic regression model.
* **Exploratory Data Analysis (EDA)**:
  + **Correlation Analysis**: Performed correlation analysis to identify relationships between numerical features and the target variable, visualized through a heatmap.
  + **Outlier Detection**: Utilized box plots to detect potential outliers, ensuring that only valid data points were retained.
  + **Categorical Features Analysis**: Analysed the distribution and impact of categorical features on lead conversion rates through bar charts.

**2. Model Building**

* **Feature Selection**:
  + **Recursive Feature Elimination (RFE)**: Applied RFE to systematically select the most relevant features, reducing the model’s complexity and addressing multicollinearity issues.
  + **Multicollinearity Handling**: Calculated statistical measures to identify and remove features with high multicollinearity, ensuring a more robust model.
* **Model Training**:
  + Built and trained the logistic regression model using the selected features. The model’s performance was assessed through accuracy and other key metrics.
* **Cutoff Selection**:
  + Determined the optimal probability cutoff by analysing the balance between sensitivity and specificity, enhancing the model’s ability to identify leads that are likely to convert.
* **Model Testing**:
  + Validated the model on a separate test dataset, ensuring consistency in performance metrics and confirming the model’s reliability for real-world application.