Lead Scoring: Process Flow & Observations

Objective: Develop a predictive model to identify potential customers likely to convert.

Dataset: Includes lead demographics, interactions, and conversion status.

Goal: Improve lead prioritization to enhance business decision-making.

Approach: Use machine learning to analyze historical data and predict conversions.

Data Exploration & Preprocessing

Data Loading & Inspection: Checked dataset shape, types, and missing values.

Handling Missing Values: Used imputation techniques or removed columns with excessive missing data.

Categorical Encoding: Converted categorical variables into numerical form using Label Encoding and One-Hot Encoding.

Feature Scaling: Applied MinMaxScaler to normalize numerical features for better model performance.

Feature Engineering & Selection

Feature Selection: Identified important variables affecting lead conversion.

Recursive Feature Elimination (RFE): Used to rank and select the most relevant features.

Multicollinearity Check: Used Variance Inflation Factor (VIF) to eliminate redundant features.

Final Feature Set: Selected optimal features to balance model complexity and accuracy.

Model Building & Training

Data Splitting: Divided dataset into training (70%) and testing (30%) sets.

Algorithm Used: Logistic Regression was chosen for its interpretability and efficiency.

Hyperparameter Tuning: Adjusted model parameters to improve performance.

Training Process: Model was trained on selected features, learning patterns from historical data.

Evaluation Metrics & Results

Performance Metrics Used: Accuracy, Precision, Recall, and F1-score.

ROC-AUC Score: Evaluated model's ability to distinguish between lead conversion and non-conversion.

Confusion Matrix: Analyzed true positives, false positives, and misclassification rates.

Precision-Recall Curve: Assessed model effectiveness in handling imbalanced data.

Observations & Conclusion

- **Key Insights**: Certain features like lead source and previous interactions play a crucial role in conversion.
- **Model Performance**: The model performed well but has scope for improvement with more data.
- **Challenges Faced**: Data imbalance and multicollinearity required careful handling.
- **Future Scope**: Exploring ensemble models like Random Forest or XGBoost for better predictive power.
- **Business Impact**: Helps prioritize high-value leads, improving efficiency and conversion rates.