Summary Report: Logistic Regression Model for Lead Scoring

Objective:

The goal was to build a logistic regression model to assign a lead score between 0 and 100, helping X Education prioritize potential leads with a higher likelihood of conversion. The target was to improve the company's lead conversion rate from 30% to approximately 80%.

Steps Followed:

1. Data Understanding and Cleaning:

- The dataset contained 9000 entries with multiple features such as `Lead Source`, `Total Time Spent on Website`, `Total Visits`, `Last Activity`, etc.

- The target variable, `Converted`, indicated lead conversion (1 = converted, 0 = not converted).

- Data preprocessing included handling missing values, correcting data inconsistencies, and ensuring uniformity in categorical variables.

2. Exploratory Data Analysis (EDA):

- Analyzed patterns and distributions of key features.

- Assessed relationships between independent variables and the target variable.

- Identified highly correlated variables and potential outliers.

3. Feature Engineering:

- Categorical variables were encoded using techniques such as dummy variables.

- Normalization and scaling were applied to numerical variables for uniformity.

4. Data Splitting:

- The dataset was divided into training and test sets (70:30 split) to evaluate model performance.

5. Model Building:

- A logistic regression model was chosen due to its simplicity, interpretability, and effectiveness in binary classification tasks.

- Feature selection was performed to retain only significant predictors, using techniques like p-values, VIF (Variance Inflation Factor).

6. Model Evaluation:

- The model's performance was evaluated using metrics such as:

- Accuracy: To measure overall prediction correctness.
- Precision and Recall: To assess the balance between false positives and false negatives.
- ROC-AUC Curve: To evaluate model discrimination capability.

7. Lead Scoring:

- Predicted probabilities from the logistic regression model were scaled to a 0-100 range to generate lead scores.
- Leads with higher scores were identified as "hot leads" for focused sales efforts.

8. Insights and Learnings:

- Features such as `Total Time Spent on Website`, `Lead Source`, and `Last Activity` significantly impacted conversion likelihood.
- Handling imbalanced data (if applicable) improved model fairness and reliability.
- Logistic regression proved effective due to its interpretability and alignment with the business context.

9. Recommendations:

- Use lead scores to prioritize sales team efforts.
- Monitor and regularly retrain the model to adapt to evolving data patterns.
- Consider integrating other advanced models, like decision trees or ensemble methods, for further improvements.