Lead Scoring Summary Report

To tackle this business challenge, a structured approach was implemented, covering data understanding, exploratory data analysis (EDA), data preprocessing, model development, evaluation, and recommendation formulation.

The initial phase involved data exploration and preparation. EDA uncovered critical insights: leads who spent more time on the website and those classified as **"High in Relevance"** had significantly greater conversion rates. **Google and Direct Traffic** emerged as the most prominent lead sources. Correlation analysis confirmed a strong positive relationship between website engagement and conversion likelihood. Data cleaning steps included addressing missing values through imputation and categorization. The preprocessing phase involved feature engineering, eliminating less relevant variables, handling outliers, encoding categorical data, and scaling numerical features using **MinMaxScaler** to ensure balanced feature contributions. The dataset was divided into **80% training** and **20% testing** sets to support a robust model-building process.

The model was developed iteratively. **Recursive Feature Elimination (RFE)** with **Logistic Regression** was used to select the top 15 features, simplifying the model while maintaining interpretability. Multiple **Logistic Regression** models were then constructed and refined. **Variance Inflation Factor (VIF) analysis** played a crucial role in detecting and resolving multicollinearity, leading to the removal of redundant features and improved model stability. The final model incorporated key predictors such as **Lead Origin, Lead Source, Do Not Email, Total Time Spent on Website, Page Views Per Visit, Tags, Lead Quality, and Last Notable Activity**.

For model evaluation, metrics such as **accuracy, AUC, sensitivity, specificity, and precision-recall curves** were utilized. A probability threshold of **0.4** was determined to strike a balance between sensitivity and specificity, optimizing the identification of actual converters while controlling false positives. The final model achieved an **AUC of approximately 0.91** and an **accuracy of around 84.10%** on the training data, demonstrating strong predictive capability.

Based on these insights, key recommendations include:

- **Enhancing website engagement** to attract high-intent leads.
- **Standardizing lead quality scoring** to improve prioritization.

- **Optimizing marketing efforts** towards high-performing channels like **Google and Direct Traffic**.
- **Personalizing communication strategies** based on lead behavior and interests.

These strategies aim to help **X Education** effectively target high-potential leads, enhance sales efficiency, allocate marketing resources optimally, and achieve the goal of **80% lead conversion**.

This project provided valuable practical experience in applying **machine learning** to real-world business scenarios. It reinforced the significance of **thorough data exploration and EDA** in extracting actionable insights. **Feature engineering and iterative model refinement** were key in building an effective and interpretable predictive model. Additionally, aligning **model evaluation metrics with business objectives** was crucial to ensuring the model delivered tangible value, ultimately improving lead conversion rates for **X Education**.