Lead Scoring Summary Report

To address this business problem, we followed a structured methodology encompassing data understanding, exploratory data analysis (EDA), data preprocessing, model building, evaluation, and recommendation generation.

The initial phase focused on **data exploration and preparation**. EDA revealed key insights: leads spending **more time on the website** and those categorized as **"High in Relevance"** exhibited significantly higher conversion rates. **Google and Direct Traffic** were identified as primary lead sources. Correlation analysis highlighted the positive relationship between website engagement metrics and conversion. Data cleaning involved handling missing values through imputation and strategic categorization. Preprocessing steps included feature engineering, dropping less informative variables, outlier treatment, encoding categorical features, and scaling numerical features using MinMaxScaler to ensure equitable feature contribution. The dataset was split into training (80%) and testing (20%) sets for robust model development and evaluation.

Model building was an iterative process. Recursive Feature Elimination (RFE) with Logistic Regression was initially used to select the top 15 features, reducing dimensionality and enhancing interpretability. Subsequently, multiple Logistic Regression models were built and refined. Variance Inflation Factor (VIF) analysis was crucial in identifying and mitigating multicollinearity, leading to the removal of redundant features and improving model stability. The final model incorporated a refined set of key predictors including Lead Origin, Lead Source, Do Not Email, Total Time Spent on Website, Page Views Per Visit, Tags, Lead Quality, and Last Notable Activity.

Model evaluation We utilize metrics such as accuracy, AUC, sensitivity, specificity, and precision-recall curves. An optimal probability threshold of 0.4 was determined based on the balance between sensitivity and specificity, maximizing the identification of true positives while controlling false positives. The final model demonstrated strong performance with an AUC of approximately 0.91 and an accuracy of around 84.10% on the training data, indicating excellent discriminative ability.

Based on these findings, key **recommendations** are prioritizing website engagement enhancements to capture higher intent leads, refining and standardizing lead quality scoring for better lead prioritization; optimizing marketing efforts towards high-yield channels like Google and Direct Traffic; and personalizing communication based on lead behavior and specialization interests. These recommendations aim to enable X Education to effectively target "hot leads," improve sales efficiency, optimize marketing resource allocation, and ultimately achieve the target of an 80% lead conversion rate.

The assignment provided valuable **learnings** in the practical application of machine learning to solve real-world business challenges. It highlighted the critical importance of thorough data understanding and EDA in uncovering actionable insights. Feature engineering and iterative model refinement proved essential for building a robust and interpretable predictive model. Furthermore, the project emphasized the need to align model evaluation metrics and business objectives to ensure the model effectively addresses the specific needs of X Education and drives tangible business value through improved lead conversion.