SUMMARY REPORT

Problem Statement:

X Education sought assistance in identifying the most promising leads with a high likelihood of converting into paying customers. The objective was to construct a lead scoring model that assigns scores to leads, indicating their potential for conversion. The CEO set a target lead conversion rate of approximately 80%.

Approach:

To address this challenge, Wefollowed a structured approach encompassing four stages: data comprehension and refinement, exploratory data analysis, feature engineering, and model construction and assessment.

Part 1: Data Comprehension and Refinement

In this initial phase, we began by comprehending the dataset's architecture, encompassing variables and their significance. The dataset contained details regarding leads and their interactions with X Education. We Performed data refinement tasks such as managing missing data, eliminating irrelevant columns, and rectifying data disparities.

Part 2: Exploratory Data Analysis (EDA)

EDA entailed acquiring insights into the dataset via visualisations and statistical summaries. We scrutinised variable distributions, detected patterns, and investigated correlations between attributes and the conversion target variable. Noteworthy findings from EDA included the significance of specific features like total visits, website engagement duration, and page views in predicting conversion.

Part 3: Feature Engineering

Feature engineering centred on enhancing the model's predictive capacity by transforming and establishing novel features. Activities encompassed categorical variable encoding, numerical feature scaling, and outlier handling. Furthermore, We partitioned the dataset into training and testing sets for model training and assessment.

Part 4: Model Construction and Assessment

In this stage, Weopted for the logistic regression algorithm to build the lead scoring model. Logistic regression is well-suited for binary classification tasks and offers interpretability. Wetrained the model on the training set, fine-tuned hyperparameters through cross-validation, and assessed its performance on the test set.

The model achieved an accuracy of approximately 79.05% on the test set, closely approaching the CEO's 80% target. Assessment metrics, including sensitivity, specificity, precision, and the precision-recall curve, provided insights into the model's performance. The model displayed promising results in correctly identifying converted leads, although there was room for enhancing recall.

Learnings:

Throughout this assignment, we Acquired several key insights:

1. Data Preprocessing: Effective data cleaning and preprocessing are pivotal stages in any data analysis venture. Managing missing data, rectifying inconsistencies, and transforming variables are imperative for precise modelling.

2. Exploratory Data Analysis: EDA plays a pivotal role in comprehending the dataset, uncovering patterns, and revealing associations between variables. It aids in feature selection and offers insights for modelling decisions.

3. Feature Engineering: The transformation and creation of pertinent features can significantly impact model performance. Scaling numerical attributes, encoding categorical variables, and handling outliers contribute to constructing robust models.

4. Model Selection and Assessment: The selection of an appropriate model and the evaluation of its performance using pertinent metrics are crucial for achieving desired outcomes. Logistic regression, in this instance, struck a balance between interpretability and predictive prowess.

5. Interpretability and Explainability: Logistic regression models offer interpretability, enabling the comprehension of feature effects on the target variable. This facilitates actionable insights and recommendations based on model coefficients.

6. Precision-Recall Trade-off: The precision-recall curve presents a valuable visualisation to grasp the trade-off between precision and recall at various threshold levels. It assists in selecting an optimal threshold aligned with the desired equilibrium between precision and recall.

Conclusion:

In summary, this project revolved around constructing a lead scoring model for X Education to identify potential customers with a higher likelihood of conversion.
