Summary Report: Leveraging Logistic Regression for Lead Conversion

In tackling the assignment for X Education, the primary aim was to construct a logistic regression model capable of assigning lead scores, facilitating the identification of leads with the highest conversion potential. The journey unfolded in several key phases, each contributing to a deeper understanding of the dataset and the formulation of a robust predictive model.

The initial steps involved meticulous **data preprocessing**, addressing missing values and handling the 'Select' level in categorical variables. This stage was pivotal in ensuring the dataset's integrity and establishing a solid foundation for subsequent analyses. Following this, Exploratory Data Analysis (EDA) provided critical insights into the dataset's characteristics, uncovering patterns and correlations. EDA illuminated the prevalence of 'Unknown' values, underscoring the need for thoughtful handling to prevent biases in subsequent modelling.

Feature engineering emerged as a crucial step in enhancing the model's predictive capabilities. Guided by the CEO's target lead conversion rate of 80%, relevant variables were selected for inclusion. The logistic regression model was chosen for its interpretability and suitability for binary classification problems. The model aimed to assign lead scores between 0 and 100, with higher scores indicating a greater likelihood of conversion. Setting a class threshold at **0.36** balanced precision and recall, aligning the model's objectives with the business target.

Model evaluation, a critical aspect of the process, was conducted rigorously on both training and testing sets. Key performance metrics such as accuracy, precision, recall and specificity were computed. Achieving consistent performance on both sets instilled confidence in the model's generalization capability, affirming its reliability in real-world scenarios.

Interpreting the model coefficients provided valuable insights into the features influencing lead conversion. **Lead Origin_Lead Add Form** and **Lead Source_Welingak Website** emerged as **strong positive influencers**, emphasizing the importance of specific engagement indicators. Conversely, **Do Not Email** and **What is your current occupation_not provided** were notable **negative influencers**, highlighting the significance of lead disengagement indicators.

The assignment offered significant learnings, emphasizing the importance of data preprocessing, exploratory data analysis, and the selection of appropriate modelling techniques. The logistic regression model presented not only a solution for immediate lead prioritization but also a broader understanding of predictive modelling in marketing and sales contexts.

Looking ahead, recommendations include periodic model reassessment, dataset updates, and exploration of advanced modelling techniques for continuous refinement. These insights will not only enhance lead conversion rates for X Education but also contribute to a broader understanding of predictive modelling role in driving business success.

In conclusion, the logistic regression model stands as a powerful tool for X Education, providing a systematic approach to lead prioritization. The journey outlined in this report contributes not only to immediate business objectives but also lays the groundwork for ongoing refinements and optimizations in the dynamic landscape of lead conversion.