Project Summary Report

Problem Statement:

X Education, an education company, offers online courses to industry professionals. The company attracts potential leads through its marketing efforts on websites and search engines like Google. Once these professionals land on X Education's website, they might engage with the content by browsing courses, filling out forms, or watching videos. When a professional fills out a form with their email address or phone number, they become a lead. Additionally, leads are also acquired through past referrals. The sales team then contacts these leads via calls and emails. However, the company's lead conversion rate is relatively low, averaging around 30%. To enhance this process, X Education aims to identify high-potential leads, known as 'Hot Leads', to improve the conversion rate and optimise the sales team's efforts.

Objective:

X Education has tasked us with developing a model to identify the most promising leads, thereby increasing the lead conversion rate to approximately 80%. By focusing on the most likely leads, the sales team can allocate their efforts more efficiently, increasing overall conversion rates.

Data Description:

The dataset provided contains approximately 9,000 data points with various attributes, including 'Lead Source', 'Total Time Spent on Website', 'Total Visits', 'Last Activity', and more. The target variable is 'Converted', indicating whether a lead was converted (1) or not (0). Our goal was to analyse these attributes, pre-process the data, and build a predictive model to assign a lead score to each lead, reflecting their conversion probability.

Data Exploration and Cleaning:

Initial data exploration involves identifying the percentage of null values in each column and understanding data completeness. Key features such as 'Lead Quality', 'Lead Profile', 'Tags', 'Current Occupation', and 'Lead Source' reveal significant imbalances and potential data quality issues. For instance, 'Lead Quality' contains ambiguous values, while 'Lead Profile' is dominated by a single category ('Select'). These insights guide decisions on retaining, dropping, or transforming these features.

Data Pre-processing:

Data pre-processing includes handling missing values, converting categorical variables into numerical formats, and potentially dropping columns with high null percentages. Feature engineering may involve creating new features from existing ones and scaling numerical features to ensure uniform contribution to the logistic regression model.

Model Training:

The dataset is split into training and testing sets. A logistic regression model is trained on the training dataset to predict lead conversion likelihood. The model's performance is evaluated using metrics such as accuracy, precision, recall, and the ROC-AUC score. The Generalised Linear Model Regression Results show a significant intercept (constant) with a coefficient of -2.5983, a standard error of 0.093, and a p-value of 0.000. The VIF values are within a stable range, indicating no multicollinearity issues.

Model Evaluation:

The model is evaluated through confusion matrices, classification reports, and ROC curve visualisation. The area under the ROC curve is 0.90, indicating a strong model. The optimum cut-off probability for specificity, sensitivity, and accuracy is determined to be 0.36. The precision of the model is 0.696, recall is 0.821, sensitivity is 0.821, and specificity is 0.796. These metrics demonstrate the model's ability to accurately predict lead conversions.

Conclusion:

This logistic regression model provides a systematic approach to predicting lead conversion, from data exploration and cleaning to feature analysis, data pre-processing, model training, and evaluation. Detailed analysis of key features and their impact ensures a robust predictive model, aiding in effective lead management and decision-making. The strong performance metrics confirm the model's reliability and effectiveness in predicting lead conversions, ultimately helping X Education achieve its target conversion rate.