CASE STUDY - LEAD SCORING

Executive Summary

This analysis was conducted for X Education with the objective of optimizing strategies to attract more industry professionals to enroll in their courses. The initial dataset provided valuable insights into the behaviour of potential customers, encompassing their site visits, duration spent, referral sources, and the overall conversion rate.

The analytical approach involved several key steps:

Data Cleaning:

Initial data cleaning was done, and the 'select' option was replaced with null for enhanced interpretability. Null values were standardized to 'not provided' to minimize data loss. Categorical variables were further refined, consolidating elements for improved clarity.

Exploratory Data Analysis (EDA):

A brief EDA was performed to assess the data's condition. Irrelevant elements in categorical variables were identified, while numeric values exhibited sound characteristics with no discernible outliers.

Dummy Variables:

Dummy variables were created, with subsequent removal of dummies containing 'not provided' elements. For numeric values, the MinMaxScaler was applied.

Train-Test Split:

The dataset was partitioned into training (70%) and testing (30%) sets.

Model Building:

Recursive Feature Elimination (RFE) was employed to select the top 15 relevant variables. Subsequently, manual removal of non-essential variables based on VIF values and p-values was conducted (retaining variables with VIF < 5 and p-value < 0.05) using GLM model.
First we have selected 0.50 as an optimal cutoff for our calculation.

Model Evaluation:

A confusion matrix was generated, and an optimal cut-off value (determined through ROC curve analysis) and Precision Recall curve which came out to be 0.35 and yielded an accuracy, sensitivity, and specificity of approximately 90%.

Prediction:

Predictions were made on the test dataset using an optimal cut-off of 0.35, resulting in an Accuracy of 93.09%, sensitivity of 91.52%, and specificity of 92.20%.

Precision-Recall Analysis:

A precision-recall approach was employed, revealing a cut-off of 0.35 with precision around 88.93% and recall around 93.36% on the test dataset.

Key Insights:

The variables identified as most influential in predicting potential buyers, ranked in descending order of importance, are:

- Tags
- Lead Source
- Total Time Spent on Website
- Last Notable Activity
- Last Activity.
- Current occupation as a working professional.

The final formula for Log Reg model is:

In (p/(1-p)) = -4.8191 + 3.5918 * Total Time Spent on Website + 4.0472 * Lead Source_Welingak Website - 2.1308 * Last Activity_Email Bounced -2.7601 * What is your current occupation_Not Known + 2.8564 * Tags_Busy + 8.9133 * Tags_Closed by Horizzon +9.6557 * Tags_Lost to EINS - 1.7546 * Tags_Ringing + 4.2063 * Tags_Unknown - 6.7144 * Tags_Will revert after reading the email - 2.1073 * Tags_switched off + 2.6838 * Last Notable Activity_SMS Sent

In conclusion, focusing efforts on these key variables presents an opportunity for X Education to significantly enhance its conversion rates by effectively engaging and persuading potential buyers.