Summary for Lead scoring case study

The objective of this case study is to identify the most promising leads for selling the company's online courses. To achieve this, a comprehensive dataset with multiple attributes was provided, aiming to discern patterns and construct a logistic regression model to generate a potential customer list.

The initial phase involved thorough data exploration to understand its structure, data types, presence of null values, and redundancy. Subsequently, data cleaning and preparation commenced, involving the removal of columns with over 40% null values, imputation of columns with fewer null values using mean/mode, and categorization of undefined values as 'unknown'.

Once the data was prepared, exploratory data analysis (EDA) was conducted to explore the relationship between features and the target variable 'converted'. Correlation metrics were examined to identify strong correlations that could influence the model's predictability. Recursive Feature Elimination (RFE) was employed for feature selection.

The model preparation phase included creating dummy variables, splitting the dataset into training and testing sets (70:30 ratio), and scaling numerical variables. Model assessment was carried out using Stats models, with iterations performed until Variance Inflation Factor (VIF) was below 5 and p-values were below 0.05 (significance level) for all features. This led to the final model.

Model evaluation was conducted using the area under the curve (AUC) or Receiver Operating Characteristic (ROC), yielding a value of 0.89, indicating strong performance. Determining the optimal cut-off for enhanced predictability involved assessing various metrics such as sensitivity, specificity, and accuracy, resulting in a final probability cut-off value of 0.35.

Precision-recall trade-off plots were utilized to validate the model's legitimacy. Finally, the trained model was applied to the testing dataset, achieving an accuracy of 80% and a recall value of 80%