Summary Report

Problem statement

X Education, an educational institution, aimed to boost its lead conversion rate from 30% to a target of 80%. To achieve this, they sought to create a model that could assign lead scores to evaluate the likelihood of a lead becoming a customer.

Approach

To determine lead scores, a Logistic Regression model was employed, utilizing the provided metadata for each lead.

Below are the steps followed to tackle this issue:

Data Reading and Understanding

The process involved importing the dataset and inspecting the various aspects of the data set including :

- Number of rows and columns.
- Data types of each columns.
- Statistical summary of all the numerical column.
- Presence of null values and duplicates.

Data Cleaning

The data cleaning process involved :

- Removing columns with more than 40% missing values.
- Checking for null values and imputing them with appropriate methods.
- Treating outliers.
- Fixing invalid data.
- Grouping low frequency values.
- Dropped columns that don’t add any value.

- Renaming columns for better readability.
- Mapping binary categorical values.

Exploratory Data Analysis (EDA)

The EDA involved :

- Checking for data imbalances.
- Performing univariate and bivariate analysis for categorical and numerical variables.
- Correlations between the numerical variables.
- Identifying the variables that had a significant effect on the target variable.

Data Preparation

The data preparation involved :

- Creating dummy features for categorical variables.
- Splitting the data into train and test sets - 70:30 ratio.
- Feature scaling using Standardization.
- Dropping highly correlated columns.

Model Building

The model building process involved :

- Using Recursive Feature Elimination(RFE) to reduce number of variables from 45 to 15.
- Using Manual Feature Reduction process to build models by dropping variables with p – value > 0.05.
- Two models were built before reaching the final model, which was stable with p-values < 0.05 and VIF < 5.
- Model 3 (lrm3) was selected as the final model with 13 variables.

Model Evaluation

- The final model, Model 3 was used to make predictions on the train and test sets.
- Confusion matrix was made and cut off point of 0.35 was selected based on Accuracy, Sensitivity and Specificity plot. This cut off gave accuracy, specificity and precision all around 80%. Whereas precision recall view gave less performance metrics around 75%.

- Plotted ROC curve to find the Area Under the Curve, which returned a value of 0.88.
- Lead score was assigned to train data using 0.35 as cut off.

Making Predictions on Test Data

The process involved :

- Scaling and predicting using the final model.
- Evaluating relevant metrics for train & test – values returned were around 80%.
- Plotting ROC curve to check the AUC – value returned was 0.87.
- Assigning lead score to the test data.

Learnings gathered

Key takeaways include addressing missing values and outliers, handling data imbalances, ensuring feature scaling, avoiding multicollinearity and finding an optimal probability cut-off for balanced Accuracy, Sensitivity and Specificity. The project offered hands-on experience in data tasks and emphasized the crucial role of choosing the right evaluation metrics aligned with business objectives while recognizing metric trade-offs.