SUMMARY

Introduction:

The primary objective of this analysis was to assist X Education in attracting more industry professionals to enroll in their courses. By leveraging available data, we aimed to identify key factors influencing potential customers' decisions and optimize strategies for converting leads into enrolled students. Through meticulous data cleaning, exploratory data analysis (EDA), model building, and evaluation, we have obtained accurate and actionable insights.

Data Cleaning:

The initial dataset was with minimal missing values. We appropriately handled null values, replacing them as necessary. The null values were taken care. Dummy variables were created which was later removed. Geographical data on customers was categorized into 'India,' 'Outside India,' and 'not provided' for clearer analysis.

EDA:

An analysis was conducted to understand the dataset's condition. We carefully examined categorical variables, identifying and removing irrelevant elements to improve data quality. Numeric variables were in good shape, without significant outliers impacting the analysis.

Dummy Variables:

To prepare the data for modelling, we created dummy variables for categorical features, allowing inclusion in predictive models. Dummy variables with 'not provided' elements were removed to avoid introducing bias. Numeric values were scaled using the Min-Max Scaler to ensure consistency and comparability during model building.

Train-Test Split:

For effective model evaluation, we split the dataset into a training set (70%) and a testing set (30%). This allowed us to develop the model using training data and evaluate its effectiveness on unseen data.

Model Building:

We used Recursive Feature Elimination (RFE) to obtain the top 15 variables and manually removed additional variables based on VIF and p-values. This rigorous approach prevented multi-collinearity issues and enhanced the model's predictive power.

Model Evaluation:

The model's performance was evaluated using a confusion matrix to assess accuracy, sensitivity, and specificity. The Receiver Operating Characteristic (ROC) curve determined an optimal cut-off value, achieving an 81% accuracy, sensitivity, and specificity.

Prediction:

Predictions were made on the test dataset using the optimized model with a cut-off value of 0.35, resulting in a robust accuracy, Precision, and Recall of 81.09%, 76% & 77% respectively.

Precision-Recall Analysis:

A precision-recall analysis validated the model, identifying a cut-off value of 0.41 that achieved an impressive precision of around 79% and a recall of around 71% on the test data.

Key Variables Influencing Potential Buyers:

The analysis revealed essential variables significantly influencing potential buyers' decisions. In descending order of importance, these are:

1. Total time spent on the Website.
2. Total number of visits.
3. Lead source, with Google, Direct traffic, Organic search, and Welingak website being prominent sources.
4. Last activity, particularly SMS and Olark chat conversation.
5. Lead origin, particularly in the Lead add format.
6. Current occupation as a working professional.