IDENTIFYING HOT LEADS FOR X EDUCATION

SUMMARY

OBJECTIVE: IMPROVE LEAD-TO-SALE CONVERSION RATE BY IDENTIFYING 'HOT LEADS'

Business Objectives:

1. Identify the most promising leads, or "Hot Leads" for X Education.

2. Develop a model for lead scoring to improve lead conversion.

3. Deploy the model for future use in targeting potential leads.

APPROACH FOR LEAD SCORING CASE STUDY:

- **Reading and Understanding : Firstly I imported all the Libraries which are necessary for my analysis. Imported the Lead Converted Data from the csv file into Jupiter Notebook, Reading and understanding the Data.**
- **Data Handling and Cleaning : Later I cleaned the Data, At firstly Removed the features which are having null values more than 30 percent, Later I gone through each and every feature understand the feature is it important or not for analysis and get rid of Outliers from the data which will affect the analysis. Finally the Data is cleaned and Ready for analysis and Model Building.**
- **Exploratory Data Analysis : The data analysis step, particularly univariate and bivariate analysis, is critical in**

understanding the underlying patterns and relationships in the data. This step provides insights that guide the model-building process, ensuring that the resulting model is well-informed and robust.

- **Univariate analysis involves examining each variable in the dataset individually. It helps in understanding the distribution, central tendency, and variability of the data.**
- **Bivariate analysis examines the relationship between two variables. It is essential for understanding interactions and dependencies in the data.**

❖ **Preparing the Data for Modelling : Data preparation is a crucial step in building any predictive model. It ensures that the data is clean, consistent, and ready for analysis, which ultimately leads to more accurate and reliable models.**

- **Most machine learning algorithms require numerical input. Categorical data needs to be converted into a numerical format.**
- **Creating dummy variables (one-hot encoding), label encoding, or using techniques like target encoding.**
- **Dummy variables represent categorical data as binary vectors, making it possible for algorithms to interpret categorical data.**
- **For a categorical variable with n categories, create n-1 binary columns, each representing the presence (1) or absence (0) of a category**

❖ **Building Model :**

- **Normalization/Standardization: Scaling numerical features for equal contribution.**
- **Train-Test Split: Dividing data into training (70%) and testing sets (30%) to evaluate model performance**
- **Purpose: Ensure the model generalizes well to unseen data and prevents overfitting.**
- **Instantiate the Model: Creating a logistic regression model instance.**
- **Instantiate the Model: Creating a logistic regression model instance.**

❖ **Model Evaluation : Evaluating the model using accuracy score, precision score, recall score, specificity , sensitivity.**

- **Accuracy is the ratio of correctly predicted instances to the total instances.**
- **Precision (also called Positive Predictive Value) is the ratio of correctly predicted positive observations to the total predicted positives.**
- **Sensitivity (Recall) is the ratio of correctly predicted positive observations to the all observations in actual class.**
- **Specificity is the ratio of correctly predicted negative observations to the all observations in actual negative class.**

- **The ROC curve is a graphical representation of a classifier's performance across different thresholds, plotting the True Positive Rate (Sensitivity) against the False Positive Rate (1 - Specificity).**
- **The threshold is the probability cut-off point that determines the classification of a predicted probability.**

❖ **Predictions on Test Data:**

- **Test data evaluation involves assessing the performance of a machine learning model using a separate dataset that was not used during training. This helps determine how well the model generalizes to new, unseen data.**
- **To ensure that the model is not overfitting to the training data and can accurately predict outcomes for new inputs.**
- **Test the model on which we have trained above.**
- **Ensure all preprocessing steps applied to the training data (e.g., scaling, encoding) are also applied to the test data to maintain consistency.**
- **Use the trained model to predict outcomes on the test data.**
- **Obtain predicted probabilities and class labels.**
- **Compare the predicted outcomes to the actual outcomes in the test set.**

- **Calculate performance metrics such as accuracy, precision, recall, specificity, and ROC-AUC score.**