SUMMARY

Problem Statement and Objectives:

To solve case study for X Education company, which specializes in selling online industry-oriented courses. Company identified need to optimize its lead conversion process to focus its sales team on most promising leads, enhancing their conversion rate beyond typical 30% toward desired target of 80%. This required developing a logistic regression model to predict lead score between 0 and 100 for each lead. Score was intended to indicate probability of conversion, with high scores denoting 'hot leads' and low scores suggesting 'cold leads.' Case study further posed strategic challenges related to adapting the model's use during periods of increased resources or diminished lead urgency.

Methodological Steps:

➔ **Data Preparation and Cleaning:**

- Data was initially explored to identify missing values, outliers, and data types requiring conversion. Data was cleansed by removing missing values and addressing outliers where necessary.
- Categorical variables were encoded appropriately to facilitate modelling. Dummy variables were created for categorical features, and continuous features were scaled to standardize their range.

➔ **Exploratory Data Analysis (EDA):**

- An in-depth exploration of the dataset was conducted to understand lead distribution, feature correlations, and potential predictors of conversion.
- Visualization tools- histograms, box plots, heat map and correlation metrics, were used to illustrate relationships between features.

➔ **Model Building - Logistic Regression:**

- Primary model used was logistic regression. It was chosen for its interpretability and suitability for binary classification problems like lead conversion prediction.
- Feature selection was performed to identify most important predictors, reducing multicollinearity and enhancing model performance.

    - Model performance metrics, including precision, recall, F1-score, and area under receiver operating characteristic curve (AUC-ROC), were evaluated. The aim was to optimize model performance while avoiding overfitting.

- **Hyperparameter Tuning:**
    - Grid search techniques were applied to fine-tune the model's parameters, improving its ability to differentiate between hot and cold leads.
    - Strategies to handle imbalanced datasets, such as adjusting class weights, were explored to ensure an equitable prediction performance across both classes.
- **Post-Modelling Strategies:**
    - The model was tested on various scenarios, including aggressive lead conversion periods (where nearly all potential leads are targeted) and times when minimizing call efforts was desired. Strategies were proposed to adjust model predictions and company actions during these phases, such as varying threshold levels for lead classification.

Learnings Gathered:

- **Data Quality and Preprocessing:** Data cleaning and preprocessing were foundational to model accuracy. Handling missing data and normalizing categorical variables played a crucial role in preparing data for modelling.
- **Feature Importance Analysis:** Identifying key predictors of lead conversion helped optimize model and provided actionable insights for business decision-making.
- **Model Interpretation:** Logistic regression's transparency allows for easy explanation of how various features influenced conversion probabilities. This interpretability was beneficial in conveying insights to business stakeholders.
- **Balancing Precision and Recall:** Optimizing for desired lead conversion rate involved careful tuning of precision and recall to align with business goals.
- **Strategic Application of Models:** Customizing the application of the model for different business contexts demonstrated the importance of adaptable machine learning solutions in dynamic operational environments.