SUMMARY REPORT

Introduction:

X Education, an online education company, faced a challenge with a low lead conversion rate of approximately 30%. To address this, the company aimed to enhance the efficiency of its lead conversion process by leveraging a predictive model. The goal was to assign lead scores between 0 and 100 to identify the most potential leads (hot leads) for targeted communication. The desired outcome was to achieve a lead conversion rate of around 80%.

Approach:

1. **Understanding the Problem:** The lead conversion rate of company is poor (around 30%). The business goal is to increase lead conversion rate to around 80%.
2. **Understanding the Data:** To know data structure, variables and their meanings. Handling missing values and ensuring data quality.
3. **Exploratory Data Analysis :**
    - Examining distribution of data, identify outliers, correlations and treatment;
    - Check data imbalance;
    - Identifying the trends that might be associated to lead conversion.
4. **Data Preprocessing :**
    - Categorical variables were encoded, and numerical features were scaled to ensure that the model could effectively learn from the data.
    - The dataset was split into training and testing sets to train the model on one subset and evaluate its performance on another
5. **Buiding Logistic Regression Model:**
    - Logistic regression was chosen as the predictive modeling technique due to its suitability for binary classification problems, such as predicting hot leads (1) or non-hot leads (0).
    - Hyperparameter tuning was performed to optimize the logistic regression model's performance. Using RFE for feature selection, tuning by checking model performance and removing all insignificant features.
    - Making predictions by choosing optimal cut-off value.
6. **Model Evaluation:**
    - The model's performance was assessed on the testing set using metrics such as accuracy, precision, recall.
    - The Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) were considered to evaluate the model's ability to discriminate between hot and non-hot leads
7. **Lead Scoring:**
    - After training and validating the model, lead scores were assigned to each lead in the dataset. A higher score indicated a higher likelihood of conversion.

8. **Communication Strategy:**
    - A communication strategy was devised based on lead scores. The sales team could prioritize leads with higher scores, focusing efforts on those more likely to convert

Results and Learnings:

1. **Model Performance:**
    - The logistic regression model demonstrated promising results, achieving an accuracy of 81.4% on the testing set.
    - Key metrics like precision, recall score were considered to balance false positives and false negatives.
2. **Lead Scoring Insights:**
    - The lead scores provided valuable insights into the potential of each lead. Leads with higher scores were indeed more likely to convert, aligning with the company's objective of targeting hot leads.
    - The implementation of lead scores allowed the sales team to focus on high-priority leads, optimizing their communication efforts. This approach led to a more efficient and targeted conversion strategy
3. **Further Improvements:**
    - Continuous monitoring and updates to the model were recommended to adapt to changing trends
    - Tuning the model using key metrics like precision and recall to meet different business goals.