Lead Scoring For X Education Summary Report

Problem Statement:

- X Education, an online course provider, seeks to improve their lead conversion rate, which is currently around 30%. They aim to identify leads most likely to convert into paying customers. The goal is to develop a logistic regression model to assign a lead score between 0 and 100, targeting a conversion rate close to 80%.

Goals:

1. Data Understanding and Cleaning: Analyze and clean the dataset to handle missing values and irrelevant columns.

2. Feature Engineering: Prepare the data by creating dummy variables and scaling numerical features.

3. Model Building: Develop a logistic regression model using Recursive Feature Elimination (RFE) to select significant features.

4. Model Evaluation: Evaluate the model's performance using various metrics and optimize the probability cutoff for predictions.

5. Prediction on Test Set: Validate the model's performance on a test set.

Data Cleaning and Preparation:

1. **Dropped columns with more than 3000 missing values and those with irrelevant or redundant data.**
2. **Created dummy variables for categorical features.**
3. **Scaled numerical features using `MinMaxScaler`.**

Model Building:

1. **Used RFE to select 15 important features.**
2. **Iteratively refined the logistic regression model by removing features with high p-values and VIFs, ensuring a robust model.**

Model Evaluation:

1. **The ROC curve showed an AUC of 0.86, indicating good model performance.**
2. **Determined the optimal cutoff point at 0.42 to balance sensitivity and specificity.**

Making Predictions on the Test Set:

1. **Scaled the test set features.**
2. **Selected the same columns as the training set.**
3. **Made predictions and evaluated the model on the test set.**

Conclusion and Next Steps:

Conclusion:

The logistic regression model successfully identifies leads likely to convert with an AUC of 0.86. By optimizing the probability cutoff, the model balances sensitivity and specificity, achieving a significant improvement over the baseline conversion rate.

Next Steps:

1. **Model Deployment: Implement the model in the company's CRM system to score new leads in real-time.**
2. **Model Monitoring: Continuously monitor the model's performance and retrain it periodically with new data.**
3. **Feature Expansion: Explore additional features and advanced models (e.g., random forests, gradient boosting) to further enhance prediction accuracy.**