Summary of Analysis for X Education

Steps Taken:

1. **Data Cleaning**:
    - Initial data with 9240 records in leads.csv file has 37 columns which include 30 categorical and 7 numerical columns are available.
    - The option "select" was replaced with NaN values as it lacked meaningful information.
    - Dropped the columns having more than 35% missing value.
    - Dropped the columns having same values such as **'Magazine','Receive More Updates About Our Courses','I agree to pay the amount through cheque','Get updates on DM Content','Update me on Supply Chain Content'** or having all unique value such as '**Prospect ID'**.
2. **Exploratory Data Analysis (EDA)**:
    - A quick EDA was performed to assess the data's condition. It was found that many elements in the categorical variables were irrelevant.
    - The numeric values appeared clean, with no significant outliers.
3. **Creating Dummy Variables**:
    - Dummy variables were created for categorical data. Dummies with "not provided" elements were removed.
    - The MinMaxScaler was used to scale numeric values.
4. **Train-Test Split**:
    - The data was split into 70% training and 30% testing sets.
5. **Model Building**:
    - Recursive Feature Elimination (RFE) was performed to select the top 15 relevant variables.
    - Variables were manually removed based on Variance Inflation Factor (VIF) values and p-values (keeping those with VIF < 5 and p-value < 0.05).
6. **Model Evaluation**:
    - A confusion matrix was created to evaluate the model.
    - The optimal cutoff value (determined using the ROC curve) was used to calculate accuracy around 81%, sensitivity around 68%, and specificity around 88%.
7. **Prediction**:
    - Predictions were made on the test dataset with an optimal cutoff of 0.35, achieving accuracy, sensitivity, and specificity of 80%.
8. **Precision-Recall Analysis**:

- The precision-recall method was also employed, identifying a cutoff of 0.41 with precision around 71% and recall around 81% on the test dataset.

Key Variables Influencing Potential Buyers:

- Prioritize contacting leads with higher Lead Scores first. Assign dedicated support to small batches of these hot leads to increase conversion chances.
- The majority of leads are generated through Google and direct traffic channels. This indicates that these sources are highly effective in attracting potential customers to our website.
- Since the majority of leads are currently unemployed, it would be beneficial to prioritize and focus more on these unemployed leads.
- A high number of total visits and extended time spent on the platform can increase the likelihood of converting a lead.