Summary:

X Education has a problem with a low lead conversion rate, currently at around 30%, while the CEO's target is 80%. To address this issue, a predictive lead scoring model was developed. The process involved data cleaning, exploratory data analysis (EDA), data preparation, model building, and evaluation. Here are the key steps:

Data Cleaning:

- Columns with more than 35% missing values were dropped.
- Categorical columns were handled by creating new categories, imputing high-frequency values, or dropping them if necessary.
- Numerical categorical data were imputed with the mode, and columns with only one unique response were dropped.
- Binary categorical values were mapped.

EDA:

- Data imbalance was identified, with only 38.5% of leads converting.
- Univariate and bivariate analysis was conducted for both categorical and numerical variables to gain insights.
- Variables like 'Lead Origin,' and 'Lead Source' were found to be valuable in predicting lead conversion.
- Time spent on the website had a positive impact on lead conversion.

Data Preparation:

- Dummy features were created for categorical variables.
- The data was split into 80:20 train-test ratio.
- Standard scaler was applied for numerical variables.
- Highly correlated columns were dropped.

Model Building:

- Recursive Feature Elimination (RFE) was used to reduce the number of variables from 48 to 15.
- Manual feature reduction was performed by dropping variables with a p-value > 0.05.
- Three models were built before the final Model 4, which had stable p-values < 0.05 and no multicollinearity (VIF < 5).
- "logm5" was selected as the final model with 12 variables for making predictions on the train and test sets.

Model Evaluation:

- A confusion matrix was created, and a cut-off point of 0.42 was selected based on accuracy, sensitivity, and specificity.
- This cut-off provided balanced accuracy, specificity, and precision, all around 80%.
- Sensitivity-specificity view was chosen over precision-recall view to achieve the CEO's goal of an 80% conversion rate.

Making Predictions on Test Data:

- Predictions were made on the test data using the final model.
- Evaluation metrics for both train and test data were close to 80%.
- Lead scores were assigned to the test data.

Recommendations:

Logistic regression model shows the below mentioned variables as the ones which can ascertain if a lead is hot lead and can be successfully converted:

- Lead Origin_Lead Add Form
- Last Activity_Had a Phone Conversation
- Last Notable Activity_Unreachable
- Lead Source_Welingak Website
- Lead Source_Olark Chat
- Total Time Spent on Website
- Last Activity_SMS Sent

However, despite the statistical significance, ‘Last Notable Activity_Unreachable’ can be removed from the variables as it does not aid in converting a lead without contact information.

Further, variables which when pursued may result in as cold leads are

- Do Not Email_Yes
- What is your current occupation_Student
- What is your current occupation_Unemployed

Essentially, currently unemployed and students have less chance of converting and can be categorized as cold leads.