Approach to solving the Lead Scoring Case study:

1. **Understand the problem statement and identify the objectives and goals of the project.**

The objective of the case study is to help X Education, a company that markets its courses on several websites and search engines, to improve its lead conversion rate. The current lead conversion rate is 30%, but the company wishes to increase it to around 80%. To achieve this goal, the company wants to identify a set of "Hot Leads" that are most likely to convert into paying customers. The task is to build a model that assigns a lead score to each lead, such that leads with a higher score have a higher conversion chance and leads with a lower score have a lower conversion chance. The ultimate goal is to increase the lead conversion rate by focusing on communicating with the potential leads.

Collect and explore the data: Gather and organize the data, and then perform an initial exploration to understand its characteristics and relationships.

2. **Prepare the data: Clean, transform, and preprocess the data so that it can be used for analysis.**

The dataset was then cleaned to deal with missing values- several columns with missing values greater than 35% were removed as imputing these values may not yield accurate results. Further columns with “Select” were imputed as either NaN as these are most likely values where the input was not made during tabulation of the dataset.

The rows with missing values were removed where the percentage of missing values was less than 1% of the data points in the column.

Further, dummies were created for columns with categorical values. As the logistic regression model can only process numeric values.

Exploratory Data Analysis was then performed to look at information that is communicated by the data. These have been commented in the python notebook.

3. **Model the data: Select and apply appropriate statistical and machine learning models to the prepared data.**

The logistic regression model was made using the Statistics model library to filter down the number of columns using Recursive Feature Elimination(RFE) and Variance Inflation Factor(VIF). Finally, the model with 18 features was developed.

4. **Evaluate the model: Assess the performance of the model using appropriate metrics and techniques.**

The developed model has an accuracy of 81.78% on training dataset and an accuracy of 81.52% on the test set.

5. **Communicate results: Summarize and present the findings and insights in a clear and meaningful way to stakeholders.**

The lead score were assigned based on the probability scores obtained by the model. We would recommend using a threshold score of 40 for classifying the lead as a Hot Lead. With this the model has an accuracy of around 81%.