Summary Report:

This project aimed to build a logistic regression model to assign lead scores to potential customers at X Education, helping the company focus on "hot leads" and increase the lead conversion rate from 30% to 80%. Below is a summary of the steps followed and key learnings:

Step 1: Data Cleaning
I started by cleaning the dataset, by handling missing values where first I converted 'Select' levels to nulls, and then checked the percentage of missing data in each column. For columns where missing data was minimal, such as 'Lead Source', 'Total Visits', and 'Last Activity', I dropped those rows. Columns with more than 40% missing data were removed entirely, like 'What matters most to you in choosing a course', which was also highly skewed. Irrelevant columns like 'Prospect ID' and 'Lead Number' were dropped. For columns such as 'Country', where most values were "India", I replaced the missing values with the mode. Other missing values in fields like 'Specialization', 'Tags', and 'Current Occupation' were filled in as "Unknown". Lastly, I removed columns with only one unique value, as they weren't helpful for modeling.

Step 2: Exploratory Data Analysis (EDA)
Next, I explored the data through univariate and bivariate analysis. I identified outliers in variables like 'Total Visits' and 'Page Views Per Visit' and removed them to improve model accuracy. Some columns, such as 'Do Not Email', 'Country', and 'Tags', were deemed irrelevant or not useful for the model and were dropped.

Step 3: Creating Dummy Variables
Since many of the variables were categorical, I created dummy variables to convert these into a format that could be used in the logistic regression model.

Step 4: Train-Test Split
I split the dataset into training and testing sets, with 70% used for training and 30% for testing.

Step 5: Scaling
Features were standardized using fit transform to ensure consistency in feature scaling.

Step 6: Model Building
To build the model, I first used Recursive Feature Elimination (RFE) to select the top 15 features. After that, I manually refined the model by removing features with high p-values and high Variance Inflation Factors (VIF) to avoid multicollinearity. I continued this process until all p-values were below 0.05 and VIFs were under 5.

Step 7: Model Evaluation
Initially, I used a cutoff probability of 0.5 for classification, but after analyzing the ROC curve, I optimized the cutoff to 0.35. This improved the accuracy, which reached 80.8% on the training set and 81% on the test set. I further fine-tuned the model by analyzing the precision-recall tradeoff, adjusting the threshold to 0.43 to balance precision and recall.

Step 8: Lead Scoring
Finally, I calculated lead scores by multiplying the predicted probabilities by 100. A higher score indicated that the lead was more likely to convert, helping the sales team prioritize those leads.

Overall, this assignment taught me the importance of rigorous feature selection, model tuning, and understanding the tradeoff between precision and recall to optimize a model for business needs.