Leads Scoring Case Study Summary:

X Education, an online course provider, needed a model to assign a lead score to each potential customer, indicating the likelihood of conversion. The target lead conversion rate was around 80%.

The solution involved several steps:

1. **Data Reading and Understanding**: The data was read and analyzed.
2. **Data Cleaning**: Variables with more than 30% null values were dropped. Missing values were imputed where required, and new classification variables were created for categorical variables.
3. **Data Analysis**: Outliers for Time Visits and Page Views per Website were identified and removed. More than 10 variables with major data imbalance were dropped.

   Insights from the visualisation of the categorical variables are :

   - **People who are not ordering a free copy of the book** seems to be converting more than the ones dropping as observed
   - People who are **identified** as **Potential leads** and followed up seems to be converting as a customer to the website
   - **Working Professionals** seem to be in the higher conversion ratio compared to the other categories of the people

   - People who have **looked upon online for the website** and those who are a **student of Some School** seems to show more interest in logging in for the course.
   - **People who have interacted with a SMS being sent** in showing an interest seems to be enrolling more in the courses. They have a higher conversion ratio.
   - **People who have been referred** have a higher chance of converting into a customer

4. **Creating Dummy Variables**: Dummy data was created for categorical variables.

5. **Train-Test Split**: The dataset was divided into test and train sections in a 70-30% ratio.
6. **Feature Rescaling**: Standard Scaling was used to scale the original numerical variables. An initial model was created using the stats model.
7. **Feature Selection using RFE**: Recursive Feature Elimination identified the top 15 crucial features. The 11 most significant variables with Variance Inflation Factors (VIFs) below 2 were determined.
8. **Plotting the ROC Curve**: The ROC curve for the features had an area coverage of 82%.
9. **Finding the Optimal Cutoff Point**: The optimal probability cutoff point was found to be 0.48, leading to accuracy=75%, sensitivity=73.4%, specificity=77%. The final predicted variables gave a target lead prediction of approximately 80%.
10. **Computing Precision and Recall**: Precision and Recall metrics values came out to be 79% and 70.5% respectively on the train dataset, leading to a cut-off value of approximately 0.42 based on the Precision and Recall trade-off.
11. **Making Predictions on Test Set :** Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 76%; Sensitivity=73%; Specificity= 73%.