Lead Scoring Case Study Summary

Problem Statement:

X Education is an education company that offers online courses for industry professionals. The company attracts many visitors to its website through various marketing channels. The company faces a problem: its lead conversion rate is very low. Out of 100 leads, only 30 become customers on average. To solve this problem, X Education wants to identify the most potential leads, also known as 'Hot Leads'. The company has hired you to help them with this task. Your job is to build a model that can assign a lead score to each lead based on various factors, such as their demographics, behavior, preferences, etc. The higher the lead score, the more likely the lead is to convert. The lower the lead score, the less likely the lead is to convert. The company's CEO has set a target of achieving an 80% lead conversion rate with this model.

Solution Summary:

1. **Step 1: Reading and Understanding Data**
   - The data was loaded and thoroughly reviewed for analysis.
2. **Step 2: Data Cleaning**
   - Variables with unique values were removed.
   - Columns with 'Select' values (indicating no selection) were converted to null.
   - Columns with more than 52% missing data were dropped, except for "Lead Quality", which was kept and imputed with "Not Sure" due to its significance.
   - The data was cleaned by eliminating skewed or duplicate variables, filling missing values for numerical features, and creating new categories for categorical variables.
   - Inconsistent labels (upper vs. lowercase) were standardized.
   - Sales team-generated features were removed to avoid confusion.
3. **Step 3: Data Transformation**
   - Binary variables were converted to '0' and '1'.
4. **Step 4: Creating Dummy Variables**
   - Dummy variables were created for categorical features, and redundant ones were removed.
5. **Step 5: Test-Train Split**
   - The dataset was split into training and testing sets with a 70-30% distribution.
6. **Step 6: Feature Rescaling**

- All variables were scaled using Standard Scaling.
- A heatmap was plotted to check for correlations between variables.

7. **Step 7: Model Building**
    - **Feature Selection**: Recursive Feature Elimination (RFE) was used to select the 15 most important features.
    - Insignificant features were removed based on P-values, leaving 12 key features with good Variance Inflation Factors (VIF).
    - The optimal probability cutoff was determined by evaluating accuracy, sensitivity, and specificity.
    - An ROC curve was plotted, achieving 95% area coverage, confirming the model's effectiveness.
    - The model was evaluated to ensure it correctly predicted 80% of converted cases.
    - Precision, recall, accuracy, sensitivity, and specificity were calculated on the training set.
    - A cutoff value of 0.25 was chosen based on the precision-recall trade-off.
    - The model was tested on the test set, producing a conversion rate of 90.78%, with sensitivity at 84.12% and specificity at 94.58%.
8. **Step 8: Conclusion**
    - The model's conversion rate on the test set was 84%, surpassing the CEO’s target of 80%.
    - The model’s strong sensitivity helps identify the most promising leads.
    - The top features influencing lead conversion are:
        - Tags_Lost to EINS
        - Tags_Closed by Horizzon
        - Tags_Will revert after reading the email