Lead scoring case study - summary report

Problem statement:

X Education, an online education company catering to industry professionals, attracts potential customers through various marketing channels such as websites, search engines like Google, and referrals. Upon visiting the website, these individuals may browse courses, fill out forms, or watch videos. Those who provide contact information are classified as leads, which are then pursued by the sales team through calls and emails. Despite acquiring numerous leads, X Education's lead conversion rate remains low, hovering around 30%. To improve efficiency, the company seeks to identify "Hot Leads" – the most promising prospects – to prioritize sales efforts. By focusing on these high-potential leads, X Education aims to increase its lead conversion rate and optimize its sales process.

Procedure summary:

1. **Reading and understanding the data**
   - Read and understand the dataset.
2. **Exploratory Data Analysis (EDA):**
   - Checked for unique values in each column and replaced "select" values with null values across all columns.
   - Identified missing values and dropped columns with more than 40% null values.
   - Analyzed categorical columns individually and took necessary actions.
   - Analyzed numerical columns, including correlation checks and outlier treatment.
3. **Preparing data for modeling:**
   - We started by creating the dummy variables of the categorical columns.
   - Split the data into training and test sets.
   - Rescaled variables in both sets using a standard scaler.
4. **Training the model:**
   - Used the RFE approach to select the best 15 features for model building.
   - Performed 4 iterations to create an efficient model with minimal p-value and VIF.
   - Derived probabilities on the training data and calculated overall accuracy, confusion matrix, sensitivity, specificity, false positive rate, positive predictive value, negative predictive value, ROC curve, and precision-recall curve.

5. Prediction and evaluation on the test data:
  ○ Scaled the test data using the previously fitted scaler (transform instead of fit_transform).
  ○ Tested the model on the test set.
  ○ Calculated various metrics like accuracy, sensitivity, specificity, etc.