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**
    - We began by thoroughly reviewing and understanding the dataset provided to gain insights into its structure and content.
2. **Exploratory Data Analysis (EDA):**
    - During this phase, we conducted a series of analyses to gain deeper insights into the dataset.
    - We checked for unique values in each column to identify any potential data anomalies or inconsistencies.
    - Identified "select" values and replaced them with null values across all columns to ensure data uniformity.
    - Implemented a strategy to handle missing values by dropping columns with more than 40% null values, thus preserving data integrity.
    - Analyzed categorical columns individually to understand their distribution and took necessary actions to preprocess the data effectively.
    - Conducted correlation checks and outlier treatment for numerical columns to identify any significant relationships and address outliers that could affect model performance.

3. **Preparing data for modeling:**
    - Created dummy variables for categorical columns to convert them into a format suitable for machine learning algorithms.
    - Split the dataset into training and test sets to facilitate model training and evaluation.
    - Applied feature scaling using a standard scaler to ensure that all features contribute equally to the model's predictions.
4. **Training the model:**
    - Utilized the Recursive Feature Elimination (RFE) approach to select the best 15 features for model building.
    - Conducted multiple iterations to refine the model and ensure its efficiency, minimizing p-values and Variance Inflation Factor (VIF).
    - Ultimately, we identified the 11 most significant variables, and their Variance Inflation Factors (VIFs) were deemed satisfactory.
    - In our final model, we determined the optimal probability cutoff by identifying key points and evaluating accuracy, sensitivity, and specificity.
    - Derived probabilities on the training data and calculated various performance metrics such as 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:**
    - Applied the previously fitted scaler to scale the test data, maintaining consistency with the preprocessing steps applied to the training data.
    - Evaluated the model's performance on the test set to validate its effectiveness in predicting lead conversion.
    - Calculated various evaluation metrics such as accuracy, sensitivity, specificity, etc., to assess the model's performance comprehensively.