Report of my overall approach to solve this problem

Understanding the Problem

X Education, an online course provider for industry professionals, faced a challenge: despite acquiring numerous leads daily, only about 30% converted into paying customers. To enhance efficiency, they aimed to identify 'Hot Leads'—those most likely to convert—allowing the sales team to focus their efforts more effectively.
Jovian

Data Overview

The dataset comprised approximately 9,000 entries, each representing a lead with attributes like Lead Source, Total Time Spent on Website, Total Visits, Last Activity, and more. The target variable, 'Converted,' indicated whether a lead became a customer (1) or not (0). Notably, some categorical variables included a 'Select' level, which was treated as a null value during preprocessing.
Jovian

Approach and Methodology

1. **Data Cleaning and Preprocessing**:
    - **Handling Missing Values**: Columns with a significant percentage of missing values were dropped, while others were imputed appropriately.
    - **Categorical Variables**: 'Select' levels were replaced with NaN and subsequently managed. Dummy variables were created for categorical features to facilitate modeling.
    - **Feature Scaling**: Continuous variables were scaled using MinMaxScaler to ensure uniformity.
2. **Exploratory Data Analysis (EDA)**:
    - Analyzed the distribution of leads across various sources, activities, and other features.
    - Identified trends and patterns, such as the correlation between time spent on the website and conversion rates.
3. **Model Building**:
    - **Logistic Regression**: Chosen for its interpretability and effectiveness in binary classification tasks.
    - **Feature Selection**: Recursive Feature Elimination (RFE) was employed to identify the most significant predictors.
    - **Model Training**: The dataset was split into training and testing sets. The model was trained on the training set, and hyperparameters were tuned to optimize performance.
4. **Model Evaluation**:
    - **Confusion Matrix**: Assessed true positives, true negatives, false positives, and false negatives.
    - **ROC-AUC Curve**: Evaluated the model's ability to distinguish between classes.

- **Precision-Recall Analysis**: Ensured the model maintained a balance between precision and recall, crucial for minimizing false positives and negatives.

Key Learnings

- **Data Quality's Impact**: The presence of placeholder values like 'Select' underscored the importance of meticulous data cleaning to ensure model accuracy.
- **Feature Engineering Significance**: Transforming categorical variables and scaling continuous ones were pivotal steps that enhanced the model's predictive capabilities.
- **Model Selection and Evaluation**: Logistic regression, combined with RFE, provided a robust framework for identifying significant predictors. Evaluating the model using multiple metrics ensured a comprehensive understanding of its performance.
- **Business Implications**: Assigning lead scores enabled the sales team to prioritize efforts, potentially increasing the conversion rate from 30% to the targeted 80%.

Conclusion

By implementing a logistic regression model, X Education can now assign a lead score to each prospect, effectively identifying 'Hot Leads.' This data-driven approach allows the sales team to concentrate on high-potential leads, thereby improving efficiency and conversion rates. The project highlighted the critical roles of data preprocessing, feature engineering, and rigorous model evaluation in developing a solution with tangible business benefits.

I hope this summary provides a clear overview of the lead scoring case study. If you have any questions or need further details on any aspect, feel free to ask!