Summary Report: Lead Conversion Case Study

Objective: The objective of this analysis is to assist X Education in identifying the most promising leads ("Hot Leads") that are more likely to convert into paying customers. This will help the sales team focus their efforts more efficiently, improving the lead conversion rate from the current 30% to the desired 80% during targeted campaigns.

Data Overview: The dataset contains approximately 9,000 records of leads with various features such as "Lead Source," "Total Time Spent on Website," "Last Activity," and the target variable "Converted." The target variable indicates whether a lead was converted (1) or not (0). Several categorical variables in the dataset contain a "Select" category, which was treated as missing data.

Approach:

1. **Data Cleaning:**
    - Replaced "Select" values with NaN and dropped rows with missing values to ensure data quality.
    - Encoded categorical variables using one-hot encoding to prepare them for modeling.
2. **Feature Engineering and Selection:**
    - Standardized numerical features using StandardScaler.
    - Used Recursive Feature Elimination (RFE) with logistic regression to identify the top 10 most relevant features for predicting lead conversion.
3. **Modeling:**
    - Split the dataset into training (70%) and test (30%) sets.
    - Built a logistic regression model to predict the probability of lead conversion.
4. **Evaluation:**
    - Evaluated model performance using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
    - Analyzed the model's coefficients to identify the most impactful features for lead conversion.

Results:

1. **Top Variables Contributing to Lead Conversion:**
    - The top three numerical variables contributing to lead conversion were:
        - "Total Time Spent on Website"
        - "Page Views per Visit"
        - "Lead Score"
    - The top three categorical/dummy variables were:
        - "Lead Source_Google"

        - "Last Activity_Email Opened"
        - "Specialization_Management"

2. **Model Performance:**
    - The logistic regression model achieved a high ROC-AUC score, indicating good discriminative ability.
    - Precision and recall values were analyzed at different thresholds to adjust strategies based on business needs.

Recommendations:

1. **Aggressive Conversion Strategy:** During the 2-month internship period, the goal is to maximize lead conversion. To achieve this:
    - Lower the decision threshold to prioritize recall, ensuring more potential leads are identified.
    - Focus heavily on leads with higher probabilities of conversion, as predicted by the model.
    - Use targeted communication strategies for top-performing categorical segments (e.g., leads from Google or those who opened emails).
2. **Conservative Strategy:** When the company's quarterly targets are met early, focus should shift to minimizing unnecessary calls. To achieve this:
    - Raise the decision threshold to prioritize precision, ensuring calls are made only to the most promising leads.
    - Rely more on automation and digital communication (e.g., emails, webinars) for less likely leads.

Key Learnings:

- Data quality is crucial for building effective predictive models. Handling missing values and encoding variables significantly impacted model performance.
- Feature importance analysis provided actionable insights for optimizing marketing and sales strategies.
- Adjusting thresholds based on business priorities (recall vs. precision) is a powerful way to align model predictions with organizational goals.

Conclusion: This analysis has equipped X Education with a robust logistic regression model to predict lead conversion probabilities. By focusing on high-probability leads and tailoring strategies based on business scenarios, the company can significantly improve its conversion rate while optimizing sales team efforts.