Case Study Summary: Lead Scoring for X Education

Problem Statement

X Education, an online education company, seeks to improve its lead conversion rate from 30% to around 80%. The company generates leads through various channels, including website visits, form submissions, video views, and referrals. These leads are then contacted by the sales team. To enhance the efficiency of the sales team, X Education aims to identify "Hot Leads" — those most likely to convert into paying customers.

Our Approach

1. **Data Preprocessing**:

- We loaded and inspected the dataset, which contained around 9000 data points with various features.
- We identified and handled missing values by dropping columns with a large number of missing values and replacing 'Select' values with NaN.
- We dropped rows with any remaining missing values to ensure data quality.

2. **Exploratory Data Analysis (EDA)**:

- We visualized the distribution of the target variable `Converted` and key categorical features.
- We analyzed the relationship between numerical features (TotalVisits, Total Time Spent on Website, Page Views Per Visit) and the target variable.

3. **Feature Engineering**:

- We one-hot encoded categorical variables to convert them into numerical format.
- We standardized numerical features using `StandardScaler` to ensure they contribute equally to the model.

4. **Model Building**:

- We split the dataset into training and testing sets.
- We built and trained a logistic regression model to predict the likelihood of lead conversion.
- We evaluated the model using metrics such as confusion matrix, accuracy, and ROC AUC score, and plotted the ROC curve.

5. **Hyperparameter Tuning**:

- We used `GridSearchCV` to find the best parameters for the logistic regression model and trained the final model with these parameters.

6. **Lead Scoring**:

- We calculated lead scores based on model predictions and scaled them to a 0-100 range.

Key Findings

1. **Top Three Variables Contributing to Lead Conversion**:

- Lead Profile: Potential Lead
- Total Time Spent on Website
- Last Notable Activity: Email Opened

2. **Top 3 Categorical/Dummy Variables to Focus On**:

- Lead Profile: Potential Lead
- Last Notable Activity: Email Opened
- Lead Source: Direct Traffic

Outcomes

- Our logistic regression model demonstrated an accuracy of approximately 75.96% and a ROC AUC score of 0.8346.
- We identified the top contributing variables and categorical features that significantly influence lead conversion.
- We assigned lead scores to each lead, allowing the sales team to prioritize high-potential leads.

Recommendations for X Education

1. **Aggressive Lead Conversion with Interns**:

- During the two-month period when interns are available, identify high-probability leads using the model.
- Segment leads by engagement level and assign interns to follow up with high-score leads.

- Provide interns with tailored communication scripts based on the lead's activity and interests.
- Continuously monitor conversion rates and adjust strategies based on interns' effectiveness.

2. **Minimizing Useless Phone Calls**:

- When the company reaches its quarterly targets, focus only on leads with the highest scores (e.g., top 10-20%).
- Use automated email follow-ups for medium to low-score leads, reserving phone calls for the highest priority leads.
- Regularly review and refine the communication strategy based on ongoing performance and feedback.
- Be flexible and adjust the strategy dynamically based on changing business needs.

By implementing these strategies, X Education can effectively manage their lead conversion process, maximizing the efficiency of their sales team and achieving higher conversion rates.