Lead Scoring Model Summary Report

Objective

The goal of this model is to help X Education identify and prioritize the most promising leads, known as "Hot Leads," by predicting the likelihood of conversion. With a current conversion rate of 38%, the goal is to significantly improve this rate by building a predictive model that assigns a lead score (0-100) to each potential customer, using logistic regression.

Data Cleaning

Key actions taken during data cleaning are:

1. Missing Values: Features with over 30% missing data were dropped to maintain model integrity.
2. Null Value Imputation: For fields like City, missing values were imputed with the mode. For Lead Source, where missing values were significant, imputation with 'Unknown' was used to avoid bias.
3. Handling ‘Select’ Values: Categorical fields with 'Select' were imputed with ‘Not selected’ to maintain consistency.
4. Renaming Fields: Some fields were renamed for improved usability.
5. Data Grouping: Less common categories in features with many unique values were grouped to reduce complexity.

Exploratory Data Analysis (EDA)

Key findings from EDA include:

1. **Lead Conversion:** The lead conversion rate was 37%.
2. **Insights from EDA:**
    - Leads from 'Other Leads' (e.g., Lead Add Form) had a high conversion rate despite low volume.
    - 'Reference' source leads had a very high conversion rate.
    - Leads from the United States had a lower conversion rate, suggesting regional influences.
    - Specialization information indicated higher conversion rates.
    - Working professionals and businessmen were more likely to convert.
3. **Outlier Handling:** Outliers in Total Visits and Time Spent on Website beyond the 99th percentile were removed to improve model accuracy.

Model Building

1. **Data Splitting:** The dataset was split 70-30 into training and testing data.
2. **Feature Selection:**
    - Recursive Feature Elimination (RFE) reduced the feature set from 41 to 17.
    - Manual feature elimination was used, resulting in 12 key features.
    - Key features include "current_occupation_Working Professional", “Last Activity_SMS Sent” and "Lead Origin_Other Leads."
3. **Model Training:** Logistic regression was chosen and trained on the selected features.

Model Evaluation

The model’s performance was evaluated using various metrics:

1. **Confusion Matrix and Cutoff Selection:** An optimal cutoff of 0.33 was selected to balance sensitivity and specificity.
2. **Metrics:**
    - Accuracy: 81% on training data.
    - Sensitivity: 84%, indicating good detection of converting leads.
    - Specificity: 77%, showing the model's ability to avoid false positives.
3. **Lead Scoring:** Lead scores were calculated for the training dataset based on predicted probabilities.

Model Prediction

1. **Test Data Prediction:** The model predicted lead conversion with **80% accuracy** on the test dataset, achieving **83% sensitivity** and **78% specificity**.
2. **Lead Scoring:** A lead score was assigned to each lead in the test set using the predicted probability, helping prioritize high-conversion leads for follow-up by the sales team.

Conclusion

The logistic regression model effectively predicts lead conversion, with high accuracy, sensitivity, and specificity. By using lead score, X Education can focus sales efforts on the most promising leads and improve the overall conversion rate.