Summary Report: Lead Scoring Model for X Education

Problem Statement

X Education faced a critical challenge: a high volume of leads but a low conversion rate (~30%). The inefficiency in prioritizing promising leads led to wasted sales efforts and missed opportunities. The company tasked us with developing a model to assign lead scores, enabling the sales team to focus on high-potential leads and improve the overall conversion rate to approximately 80%

Data Understanding and Preparation

The dataset contained 9,240 records with 37 features, including numerical and categorical variables. The target variable, Converted, indicated whether a lead converted (1) or not (0). Key challenges included:

- **Missing Values:** Several columns had significant missing values, such as "Lead Quality" and "Tags."
- **Imbalanced Classes:** Only 36% of the leads were converted.
- **Redundant Features:** Some features, like those with high missing percentages, were irrelevant.

Steps Taken:

1. Dropped columns with excessive missing values and low relevance
2. Imputed missing values using the median for numerical features and the mode for categorical features
3. Encoded categorical variables using one-hot encoding and standardized numerical features
4. Split the dataset into training (80%) and testing (20%) sets for model evaluation.

Exploratory Data Analysis (EDA) Key insights from EDA included:

1. Strong correlation between Total Time Spent on Website and conversion.
2. Minor differences in Total Visits and Page Views Per Visit between converted and non-converted leads.

3. Visualizations such as bar charts, box plots, and a heatmap revealed patterns and relationships among features.

Model Development

We chose Logistic Regression for its simplicity and interpretability, which are crucial for explaining the results to business stakeholders

Key Metrics:

- **Accuracy:** 80.5%
- **Precision:** 72.6%
- **Recall:** 79.4%
- **F1-Score:** 75.8%
- **AUC-ROC:** 87% (indicating excellent discriminatory power).

The model's coefficients provided valuable insights into feature importance. For example, Lead Origin_Lead Add Form and Total Time Spent on Website positively influenced conversion, while Do Not Email_Yes had a negative impact. The intercept of -0.19 suggested a low baseline probability of conversion without feature influence.

Threshold Optimization and Lead Scoring

To align with business goals, we analyzed various classification thresholds. A threshold of 0.5 was chosen as its balanced precision (72.6%) and recall (79.4%), minimizing false positives while capturing most potential converters.

Lead scores (0–100) were generated for all leads. Based on these scores, leads were segmented into:

- **Hot Leads (≥ 80):** 2,165 leads – high priority
- **Warm Leads (50–79):** 1,644 leads – nurture with targeted content
- **Cold Leads (< 50):** 5,431 leads – low priority

Business Recommendations

1. **Focus on Hot Leads:** Allocate sales resources to immediately follow up on high priority leads.
2. **Nurture Warm Leads:** Use automated campaigns and educational content to move them to the "Hot" category.
3. **Minimize Effort on Cold Leads:** Engage with automated re-targeting campaigns to save resources.

Conclusion The lead scoring model provides a data-driven solution to improve sales efficiency and conversion rates. With an AUC-ROC of 87% and actionable segmentation, X Education can better allocate resources and achieve its target conversion rate. Continuous monitoring and potential integration of advanced models like Random Forest can further enhance results.