

Summary Report: Lead Scoring Analysis

The assignment focused on enhancing the lead conversion process for X Education, a company experiencing a low lead conversion rate of approximately 30%. The primary objective was to develop a model to assign lead scores, which would enable the company to prioritize high-potential leads, thereby increasing the efficiency of their sales team and improving the overall conversion rate.

Approach

1. Data Acquisition and Exploration:

- Data Sources: We used a dataset containing around 9000 records with attributes such as Lead Source, Total Time Spent on Website, Total Visits, and Last Activity. We also referred to a data dictionary to understand the column definitions and the possible levels of categorical variables.

- Initial Data Review: We examined the dataset for missing values and anomalies. Summary statistics and data distribution were analyzed to understand the dataset's structure and quality.

2. Data Cleaning and Preprocessing:

- Handling Missing Values: We replaced 'Select' values (indicative of missing data) with NaN. For numeric columns, missing values were filled with the median of the respective columns, while categorical columns had missing values replaced with 'Unknown'.

- Feature Encoding: Categorical variables were transformed into a numerical format using one-hot encoding, which converts categorical values into binary columns. This transformation is essential for machine learning algorithms that require numerical input.

3. Model Building:

- Logistic Regression: We employed a logistic regression model to predict lead conversion. The model was trained on 70% of the data and validated on the remaining 30%. Logistic regression was chosen for its effectiveness in binary classification problems and interpretability.

- Model Evaluation: The model's performance was assessed using classification metrics such as precision, recall, and F1-score, along with the ROC AUC score. The classification report and ROC AUC score provided insights into the model's ability to correctly classify leads and distinguish between converted and non-converted leads.

4. Feature Importance Analysis:

- Key Variables: We identified the top three variables contributing most to lead conversion probability: Total Time Spent on Website, Total Visits, and Last Activity_Email Opened. These variables had the highest coefficients in the logistic regression model, indicating their significant influence on conversion likelihood.

- Categorical Variables: Important categorical variables included Last Activity_Email Opened, Lead Source_Google, and Lead Source_Organic Search. These features had a considerable impact on the model's predictions and should be prioritized in lead management strategies.

5. Lead Scoring and Strategic Recommendations:

- Lead Scoring: Each lead was assigned a score between 0 and 100 based on the model's prediction probabilities. Higher scores indicate a higher likelihood of conversion, allowing the sales team to focus on leads with greater potential.

- Intern Period Strategy: During periods with available interns, the strategy should involve aggressively following up with high-score leads to maximize conversions. Interns should be tasked with contacting these high-potential leads to ensure no opportunity is missed.

- Target Achievement Strategy: When the company meets its quarterly targets ahead of schedule, the focus should shift to minimizing contact with low-score leads. This approach conserves resources and reduces the likelihood of wasted effort on leads with a low probability of conversion.

Learnings

- Data Quality is Crucial: Effective data cleaning and preprocessing are fundamental to building a robust predictive model. Handling missing values and encoding categorical variables properly can significantly impact model performance.

- Feature Importance Drives Strategy: Understanding which features have the most influence on lead conversion helps in formulating targeted sales strategies and optimizing lead management processes.

- Model Evaluation: Evaluating models with appropriate metrics ensures that the model is not only accurate but also practical for real-world applications. The ROC AUC score and classification report are valuable tools for assessing model performance.

- Strategic Application: Tailoring sales strategies based on model predictions can lead to more efficient resource use and improved conversion rates. Strategic follow-ups and prioritization of high-potential leads can make a substantial difference in achieving business goals.

In conclusion, the analysis provided actionable insights that can help X Education enhance its lead conversion process, optimize resource allocation, and ultimately improve the effectiveness of its sales team.