Summary Report: Lead Scoring Using Logistic Regression

This report outlines the development of a lead scoring model using logistic regression to help prioritize potential leads based on their likelihood of conversion. The objective was to assign a score (0–100) for each lead, enabling the company to optimize resource allocation and improve marketing effectiveness.

Problem Statement

The company aimed to create a system for scoring leads, where higher scores indicate a higher probability of conversion. This scoring model would allow the sales and marketing teams to focus efforts on "hot leads" and maximize conversion rates. The solution was required to be flexible for future dataset updates and new business requirements.

Data Insights and Preprocessing

The dataset included information on lead attributes such as *Lead Source*, *Total Time Spent on Website*, and *Lead Quality*. To prepare the data:

- Missing values were imputed using the mean for numerical features and the mode for categorical features.
- Categorical variables were transformed using one-hot encoding, while numerical features were scaled for normalization.
- No data rows were dropped, ensuring minimal information loss.

Model Development

A logistic regression model was chosen due to its suitability for binary classification tasks. The dataset was split into training (80%) and testing (20%) sets. The model predicted conversion probabilities, which were converted into lead scores. Hyperparameters like `max_iter` were adjusted to ensure convergence, and the implementation details are documented in the Jupyter notebook.

Evaluation and Results

The model's performance was evaluated using multiple metrics:

- **Precision:** 0.73 — reflects the accuracy of positive predictions.
- **Recall:** 0.49 — indicates the model's ability to identify actual conversions.
- **F1-Score:** 0.59 — balances precision and recall.
- **ROC-AUC:** 0.75 — demonstrates reasonable overall classification performance.

The confusion matrix provided insights into true positives, true negatives, false positives, and false negatives, revealing areas for further improvement.

Business Implications

The lead scores allow the company to:

1. Prioritize high-scoring leads (>80) for aggressive targeting.
2. Reduce effort on low-scoring leads (<30), saving resources.
3. Optimize marketing budgets by focusing on leads with higher conversion potential.

Future Considerations

- **Adapting to New Data:** The model can incorporate new features or lead sources through retraining and preprocessing updates.
- **Improving Imbalanced Data Handling:** Techniques like oversampling or SMOTE can address potential class imbalance issues.
- **Model Exploration:** Advanced models like Random Forest or XGBoost may provide better performance if needed.

Conclusion

The logistic regression model provides a robust framework for predicting lead conversion probabilities, aligning closely with the company's business goals. By enabling targeted marketing strategies and efficient resource allocation, the lead scoring system is expected to significantly improve conversion rates while maintaining flexibility for future needs.

4o