Assignment Summary

X Education implemented a logistic regression lead scoring model using data from its sources. The dataset, consisting of 6773 rows, underwent preprocessing where columns with significant missing values were removed. Categorical variables like 'Lead Origin,' 'Lead Source,' and others were encoded, and numerical features were scaled using min-max scaling.

Feature selection involved reducing smaller categories within a column to larger ones. Through Recursive Feature Elimination (RFE), the model retained essential features such as 'TotalVisits,' 'Total Time Spent on Website,' and various lead-related activities. Notably, the model's interpretability led to the selection of logistic regression over other algorithms like decision trees and random forests.

During model evaluation, key metrics such as accuracy, precision, recall, and specificity were considered. The model showcased an accuracy of 79%, indicating its ability to correctly identify leads 79% of the time. A precision of 80% signified that 80% of the leads predicted as promising were genuinely interested, optimizing resource allocation. The recall value of 73% meant capturing 73% of all truly interested leads, essential for maximizing conversion opportunities. A specificity of 79% highlighted the model's ability to filter out non-promising leads effectively.

To optimize the model, a cutoff parameter of 0.4 was determined by finding the intersection between the accuracy curve, specificity, and selectivity. Challenges arose due to limitations in the dataset, with many predictor variables having significant null values and 'Select' entries. Overfitting issues were addressed by including additional dummy variables. The model's limitations were recognized, prompting consideration of more complex models like decision trees or random forests to balance interpretability and accuracy.

While the model's implementation and its impact on X Education's business conversion rates are yet to be realized, projections indicate significant improvements. If the conversion rate, previously at 38.5%, increases to at least 50% due to the new lead scoring model's precision and accuracy, it would result in substantial revenue growth. This enhanced revenue can be reinvested to further boost marketing, sales initiatives, and overall customer experience, ensuring sustainable business growth.

Looking forward, X Education plans to explore advanced techniques and continuous data refinement, recognizing the importance of adapting the lead scoring system to evolving market dynamics. The combination of precise data analysis, strategic feature selection, and the integration of advanced modeling approaches positions the lead scoring system as a pivotal tool in X Education's pursuit of optimized business outcomes.