Summary

The objective of this assignment was to analyze the lead conversion data for X Education and develop a predictive model using Logistic Regression to identify the most promising leads, thereby improving the company's lead conversion rate. The project involved data preprocessing, feature engineering, and the application of logistic regression to predict lead conversion probabilities. This report summarizes the approach taken and the key learnings gathered throughout the assignment.

Data Preprocessing

The initial step involved importing and exploring the dataset. This phase included handling missing values, identifying and removing outliers. The dataset contained several categorical variables, which were converted into dummy variables using the pd.get_dummies() function. This transformation allowed the inclusion of categorical features in the logistic regression model.

A significant part of data preprocessing was scaling the features using StandardScaler from sklearn.preprocessing. Scaling ensured that all features contributed equally to the model's performance and prevented bias toward features with larger values. The training data was scaled using fit_transform, while the test data was transformed using transform to prevent data leakage.

Feature Engineering

Feature engineering involved creating and selecting variables that would most effectively predict lead conversion. We performed feature selection using Recursive Feature Elimination (RFE). This method helped identify the top features contributing to lead conversion, which were then used in the final model. The selection process involved comparing the results of RFE with and without statistical validation using statsmodels, leading to the refinement of the final feature set.

Model Development and Evaluation

Logistic regression is well-suited for binary classification problems like predicting lead conversion. After feature selection, the model was trained on the scaled training data using LogisticRegression from sklearn. The model's performance was evaluated using a confusion matrix, accuracy, sensitivity, and specificity metrics. These metrics provided insights into the model's ability to correctly identify both converted and non-converted leads.

Variables such as "Total Time Spent on Website" and specific categorical features like "Lead Source_Welingak Website" were identified as significant contributors to lead conversion. The model also highlighted the importance of focusing on certain categorical variables, which could inform the company's future marketing and sales strategies.

Learnings and Insights

- Importance of Data Preprocessing: Proper handling of missing values, outliers, and scaling is crucial for building an effective predictive model.
- Feature Selection: Recursive Feature Elimination proved to be a valuable tool for identifying the most relevant features.
- Model Evaluation: The use of evaluation metrics like accuracy, sensitivity, and specificity provided a comprehensive understanding of the model's performance.
- Practical Application of Logistic Regression: Logistic regression, coupled with effective data preprocessing and feature engineering, can serve as a powerful tool for binary classification problems, especially in business contexts like lead conversion.

In conclusion, this assignment demonstrated the practical application of machine learning techniques in improving business processes. By identifying potential leads more accurately, X Education can enhance its sales strategy and improve its overall conversion rate.