Project Summary

The assignment from X Education involved developing a model to identify and prioritize leads most likely to convert into paying customers. The CEO emphasized a target lead conversion rate of approximately 80%. Our goal was to create a scoring system that accurately reflects the likelihood of conversion for each lead, assigning higher scores to leads with a greater probability of converting.

To tackle this, we opted for a Logistic Regression Model due to its effectiveness in binary classification problems and its ability to provide probability scores. The first step in our process was basic data cleaning. We began by addressing missing data: columns with more than 3,400 null values were removed due to their lack of utility, while rows with missing values in other columns were also eliminated. This step was crucial in ensuring the quality and completeness of our dataset.

After cleaning the data, we moved on to feature engineering. Categorical variables were transformed into dummy variables to make them suitable for the logistic regression model. This step involved creating binary columns for each category within the categorical features, thereby allowing the model to interpret these variables effectively.

Subsequently, we scaled the features using MinMaxScaler. Scaling is important in logistic regression as it ensures that all features contribute equally to the model by normalizing their range. With the data prepared, we split it into training and testing sets, allocating 30% of the data for testing and using the remaining 70% for training the model.

Feature selection was the next critical phase. We employed Recursive Feature Elimination (RFE) to select the top 15 features from our dataset. RFE helps in identifying the most significant features by recursively removing less important ones and assessing model performance. Following initial feature selection, we constructed our first logistic regression model and began the process of refinement. This involved iteratively removing variables that exhibited high Variance Inflation Factor (VIF) values and those with insignificant p-values. The aim was to reduce multicollinearity and enhance the model's interpretability and performance.

Upon finalizing the model, we validated it using a random cutoff value to assess its initial performance. To determine a more reliable cutoff value, we utilized the Receiver Operating Characteristic (ROC) curve. The ROC curve helped us evaluate the trade-offs between the true positive rate and the false positive rate, guiding us to select an optimal threshold for classification.

Further evaluation involved plotting the Precision-Recall curve, which provided insight into the model's performance with respect to precision and recall. The results were satisfactory, indicating that the model was effective in distinguishing between leads likely to convert and those less likely to do so.

Finally, we applied the refined model to our test dataset to obtain various performance metrics, including sensitivity, specificity, precision, recall, and accuracy. These metrics offered a comprehensive evaluation of the model's effectiveness in predicting lead conversion.

In summary, through a structured approach involving data cleaning, feature engineering, model building, and rigorous validation, we developed a logistic regression model that aligns with X Education's conversion goals. The model's ability to score leads effectively ensures that those with higher scores have a greater likelihood of converting into paying customers, meeting the CEO's target conversion rate of 80%.