Summary of Logistic Regression Model Development

The process of developing a logistic regression model began with importing necessary libraries and loading the 'Leads' dataset for analysis. Initial data exploration was conducted using methods such as `.head()`, `.info()`, and `.describe()`, which provided insights into the dataset's structure and basic metrics.

A crucial step involved checking for null values within the dataset. Columns with more than 3,000 null entries were removed to enhance data quality and ensure model effectiveness. Additionally, a thorough review of the data dictionary led to the identification and removal of several columns deemed irrelevant for model building. This included the elimination of variables that exhibited no variance, as they would not contribute meaningfully to the predictive capacity of the model.

During the data cleaning phase, certain columns included a “Select” option that needed attention. Specifically, the 'Specialization' column contained selectable values with sufficient variance and was processed during the creation of dummy variables. In contrast, two columns with the “Select” option that were filled with null values were completely discarded from the dataset to maintain integrity.

Following the cleaning process, the data was divided into training and testing sets to prepare for model building. To further understand relationships between variables, correlation analysis was performed using heatmaps, which visualized the strength of associations among features.

The model development commenced with Recursive Feature Elimination (RFE), selecting 15 features that demonstrated the most predictive power. After building the initial model, Variance Inflation Factors (VIFs) and p-values were evaluated to assess multicollinearity and the significance of each predictor variable. This iterative process continued until a final model was established, characterized by acceptable VIFs and statistically significant p-values.

Model evaluation began with a default cut-off value of 0.5 to classify outcomes. The performance of the model on the training set was analysed, focusing on sensitivity and specificity to understand its predictive accuracy. An ROC curve was plotted to visualize the trade-off between true positive rates and false positive rates. Based on this analysis, an optimized cut-off of 0.42 was determined, balancing sensitivity and specificity more effectively.

Subsequently, predictions were made on the test set, and performance metrics, including precision and recall, were calculated to further gauge model effectiveness. The final model exhibited robust performance across all evaluation metrics, indicating its reliability in predicting the binary outcome based on the selected features.

In conclusion, the logistic regression model was successfully developed through a comprehensive series of steps, including data cleaning, feature selection, model evaluation, and performance optimization. The model achieved favourable metrics, confirming its utility for predicting outcomes in the context of the 'Leads' dataset. Future enhancements may involve exploring additional modelling techniques or incorporating more complex algorithms to potentially improve predictive accuracy further.