SUMMARY

This is the whole path of how I did the analysis:-

1. Cleaning Data:

- The data was mostly clean except for a few null values. Some null values were changed to "not given" to avoid losing data although these were later removed when making dummies.

2. EDA:

- A quick exploratory data analysis (EDA) was done to check the data's condition. Many elements in the categorical variables were irrelevant. Numeric values seemed good with no outliers found.

3. Dummy Variables:

- Dummy variables were created, and dummies with "not given" elements were removed.

4. Train-Test Split:

- The split was done at 70% for training and 30% for testing.

5. Model Building:

- RFE was done to get the top 15 relevant variables. Other variables were removed manually based on VIF values and p-values (kept variables with VIF < 5 and p-value < 0.05).

6. Model Evaluation:

- A confusion matrix was created. The optimum cut-off value (using the ROC curve) was found to be around 0.35.

7. Prediction:

- Predictions were made on the test data frame, with accuracy, sensitivity, and specificity of 80%.

8. Precision-Recall:

- This method was also used, and a cut-off of 0.41 was found with Precision around 61% and Recall around 72% on the test data frame.