Solution summary

Step 1: reading and understanding Data

Analysis and got some familiarity about our dataset. The data dictionary provided was also of big help for it.

Step 2: data cleaning

We dropped the variables that had high percentage of null values in them, imputing the missing values as and creation of new classification variables and the outliers were identified and removed.

Step3:data analysis

Then exploratory data analysis of the data set was done and the redundant variables were removed. Univariate and bivariate analysis of all the variables were performed. Heatmap, boxplots, countplts and violinplots were used for visualization.

Step4: creating dummy variables

We went on with creating the dummy data for the categorical variables concatenate them to the original dataframe and then removed the original variables of those dummy variables.

Step5: Test train split was done 70 -30%

Ste6: feature rescaling of numeric variables and used statsmodel for 1st model and we got a sense of the statistical view of all the parameters of our model.

Step7: feature selection using rfe and selected 18 top important features and analysed their p values and found out their vifs. According to the highest p values and the highest vifs we removed the features. We generally took vifs less than 3 and 0.05 pvalues as acceptable. We arrived at 18 variables as important variables. We then created the data frame having converted probability values and we had an initial assumption that the probability value of more than 0.5 means 1 else 0. We calculated the confusion metrics, overall accuracy of the model, sensitivity and specificity to get the sense of reliability of the model. Our final mode was the 6th model which we reuilt.

Step8: plotting the roc curve

Curve became descent.

Step9: finding the optimal cutoff point

Probability graph for accuracy, sensitivity and specificity for various probability values were plotted they all three curves intersected at the 0.34 point so we took that as the optimal cutoff value.

Step 10: computing the precision and recall metrics

The confusion matrix came out to be

```
[[3607, 896],

[ 508, 2248]]
```

Train Data Accuracy :80.49 %

Train Data Sensitivity :81.57 %

Train Data Specificity :80.1 %

Train Data F1 Score :0.72

Precision : 79.10%

Recall : 66.07%

Step11: making predictions on test set

Confusion matrix came out to be

```
 [[928, 208],

 [155, 524]]
```

- Test Data Accuracy : 80.0 %
- Test Data Sensitivity : 81.57 %
- Test Data Specificity : 80.1 %
- Test Data F1 Score : 0.74
- Precision : 71.58%
- Recall: 77.17%