FindDefault
Solution by K.Nithin Sai

DESIGN CHOICES:

1) Outlier Handling:
    i) Used boxplots to figure out the outliers that exist in the dataset.
    ii) Capped the outliers to the max or min value depending on the position of the outlier to not lose any information.
2) Class Imbalance:
    i) Huge difference between the two target classes.
    ii) Upsampled the minority class with sampling_strategy of 40% instead of 100% to try to mimic real world data.
3) Dimensionality reduction:
    i) PCA with 90% variance gives 22 features needed to keep 90% variance.

4) Train test split:
    i) Used a 67%train -33% test split to create validation and then a 67%train -33% test split to create a test set.
5) Model used:
    i) Used SVC classifier as the estimator with hyperparameters as follows:
        1. C: 1
        2. Kernel: rbf
        3. Degree: 3
        4. Gamma: scale

Performance Metrics:

Validation Set:

1) Accuracy Score: 0.996
2) Confusion_matrix:

[[93610, 129],
[371, 37244]]

3) Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	93739
1	1.00	0.99	0.99	37615
accuracy			1.00	131354
macro avg	1.00	0.99	1.00	131354
weighted avg	1.00	1.00	1.00	131354

Test Set

1) Accuracy Score: 0.995
2) Confusion Matrix:
   [[30927, 49],
   [133, 12238]]

3) Classification report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	30976
1	1.00	0.99	0.99	12371
accuracy			1.00	43347
macro avg	1.00	0.99	1.00	43347
weighted avg	1.00	1.00	1.00	43347