Presentation 3

Credit Card Evaluation

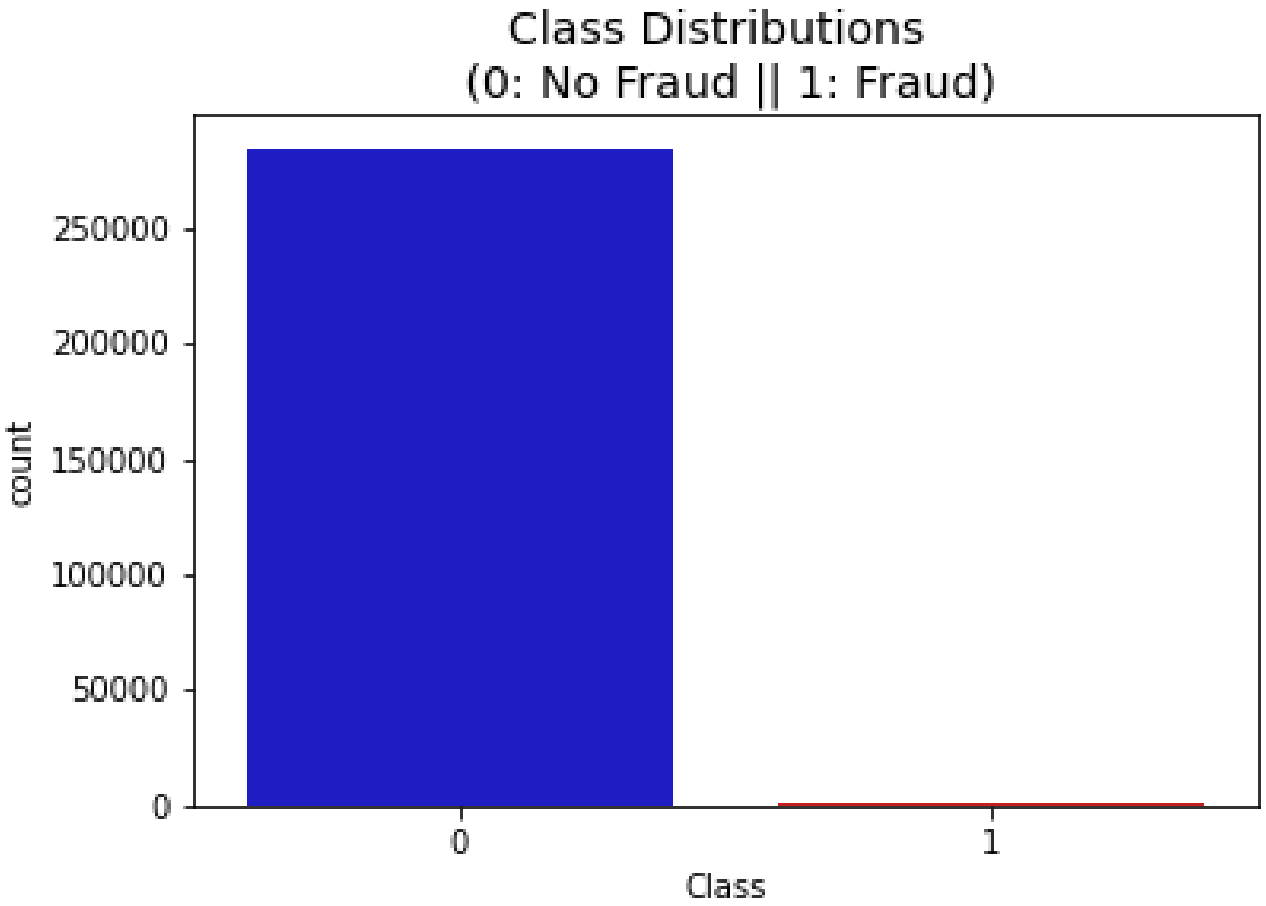
Dataset Statistics:

	Time	V2	V8	V9	V17	V22	V24	Amount	Class
count	284807	284807	284807	284807	284807	284807	284807	284807	284807
mean	94813.86	0	0	0	0	0	0	88.35	0
std	47488.15	1.65	1.19	1.1	0.85	0.73	0.61	250.12	0.04
min	0	-72.72	-73.22	-13.43	-25.16	-10.93	-2.84	0	0
25%	54201.5	-0.6	-0.21	-0.64	-0.48	-0.54	-0.35	5.6	0
50%	84692	0.07	0.02	-0.05	-0.07	0.01	0.04	22	0
75%	139320.5	0.8	0.33	0.6	0.4	0.53	0.44	77.16	0
max	172792	22.06	20.01	15.59	9.25	10.5	4.58	25691.16	1

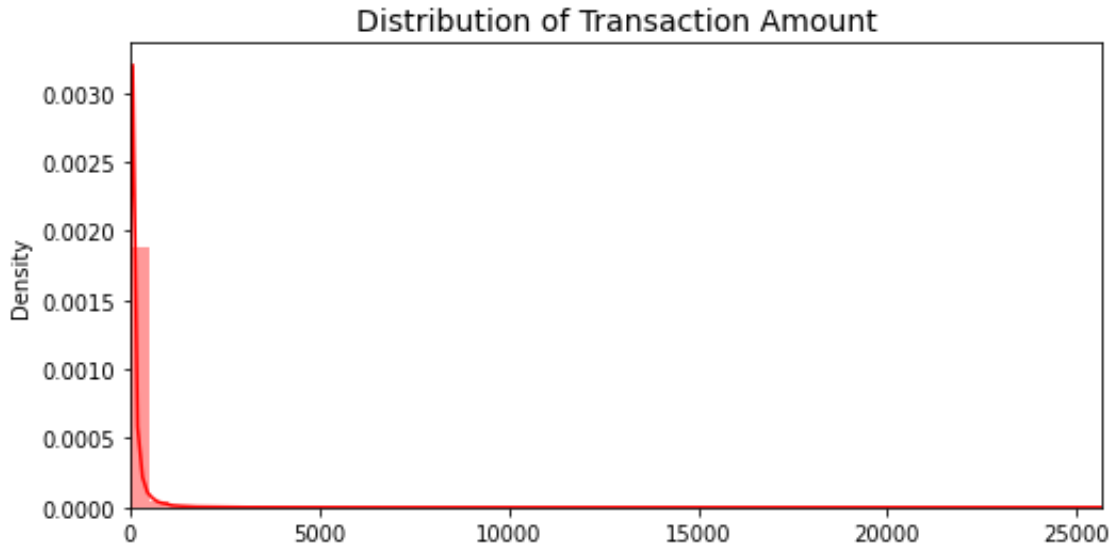
Dataset summary

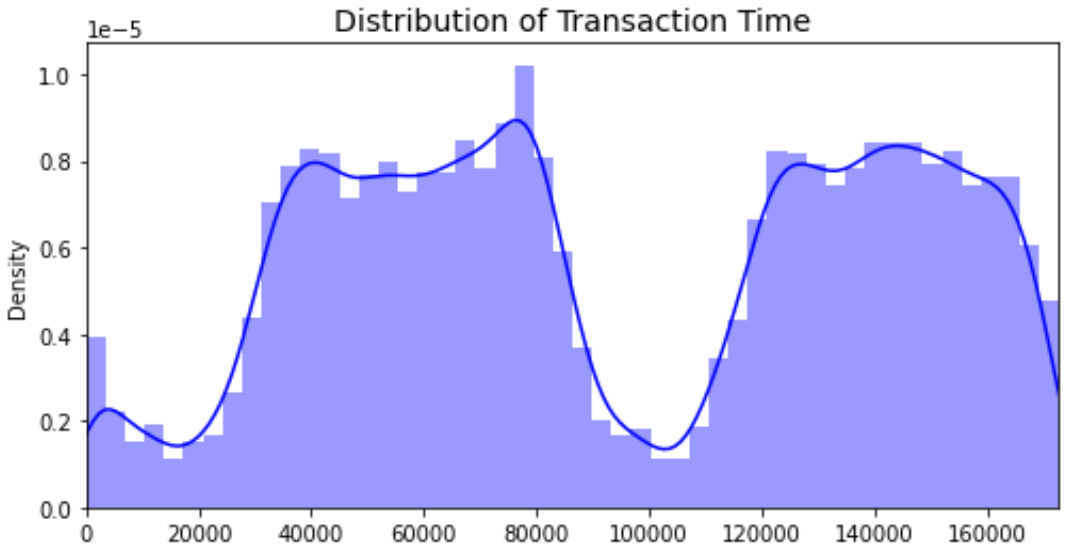
- Number of Columns: 30 + labels
- Number of Nans: 0
- Number of Genuine data: 284,315
- Number of Fraud data: 492
- Genuine data's percentage: 99.83 %
- Fraud data's percentage: 0.17 %

Imbalanced Dataset



Distribution of Time and Amount



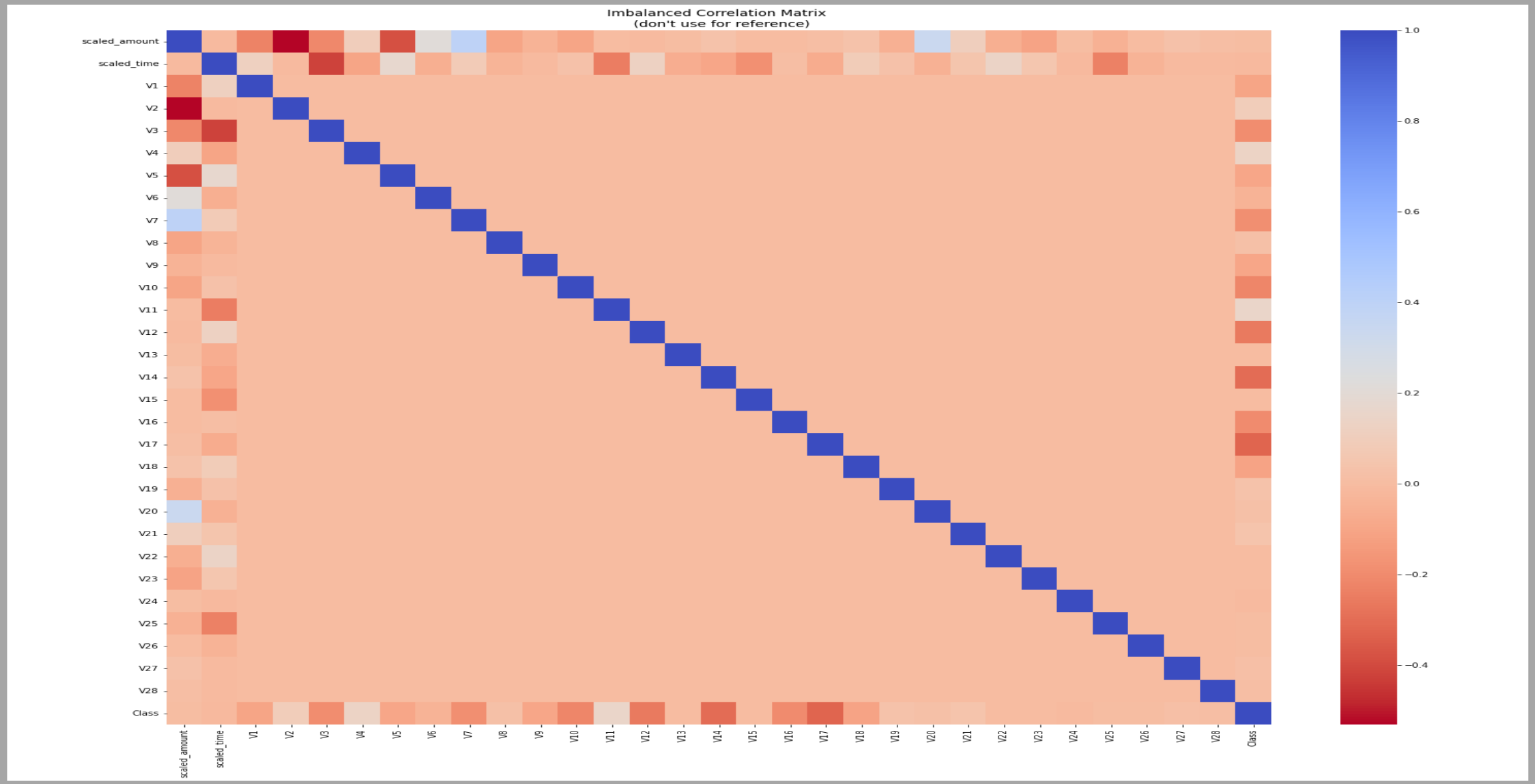


Methods of Dealing with imbalanced datasets

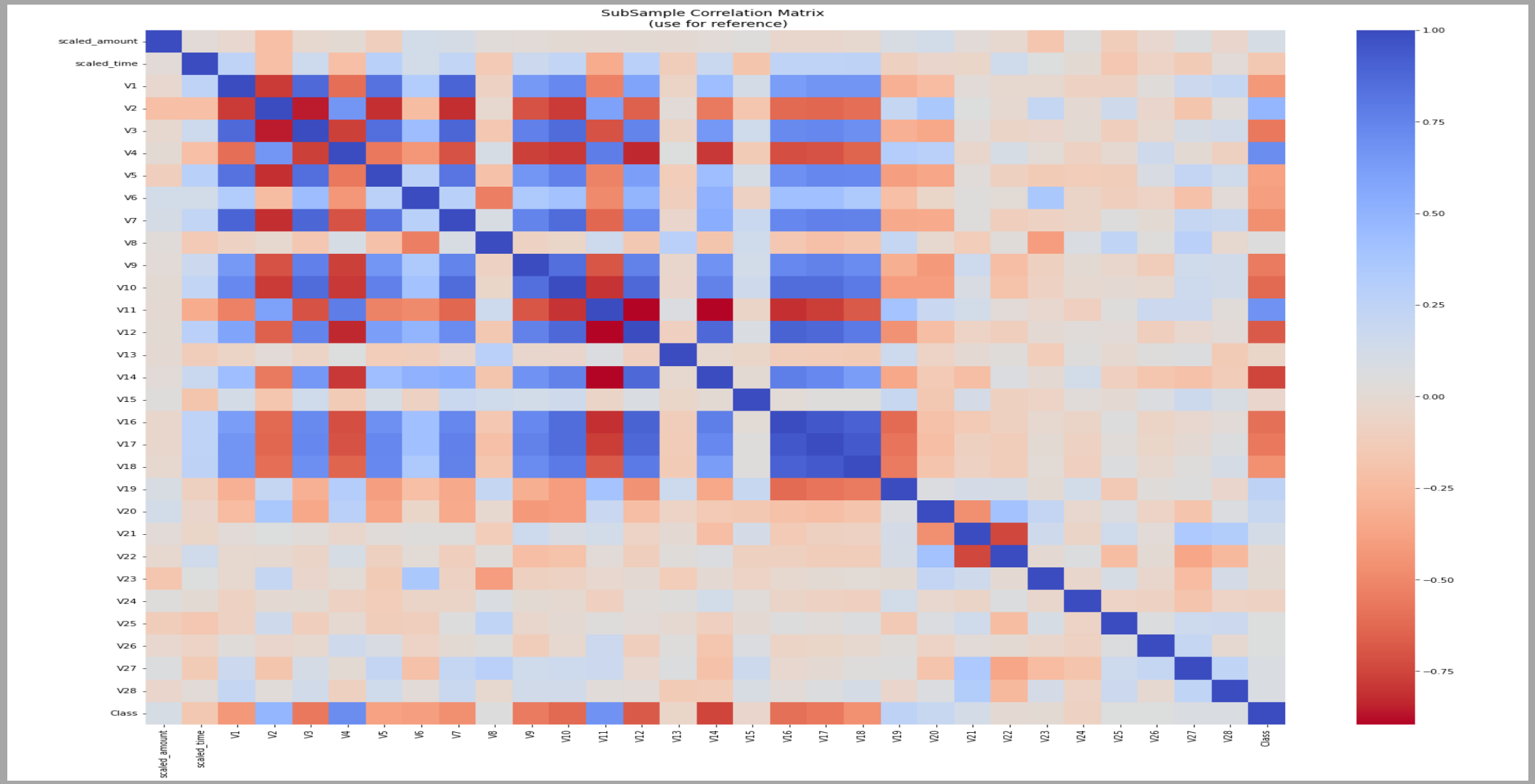
- Under Sampling
- Over Sampling

Imbalanced Dataset

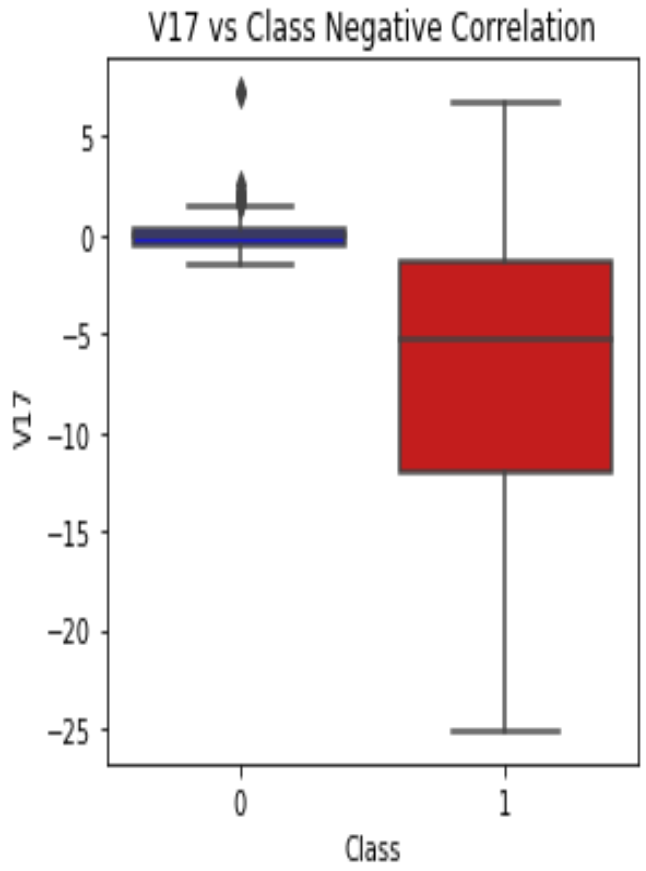
Correlation Matrix (imbalanced data)

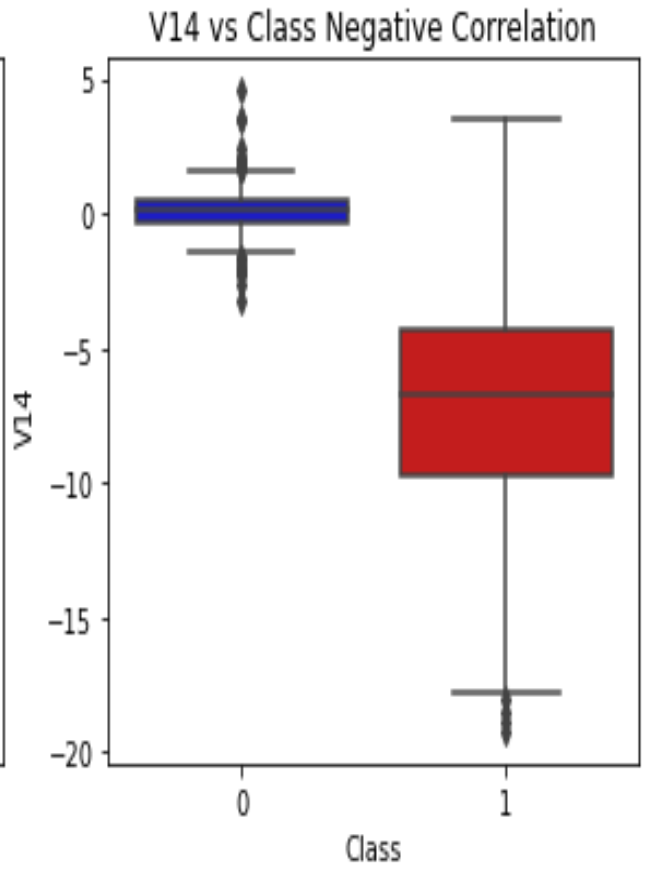

Imbalanced Correlation Matrix
(don't use for reference)
scaled_amount
scaled_time
V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
V11
V12
V13
V14
V15
V16
V17
V18
V19
V20
V21
V22
V23
V24
V25
V26
V27
V28
Class
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

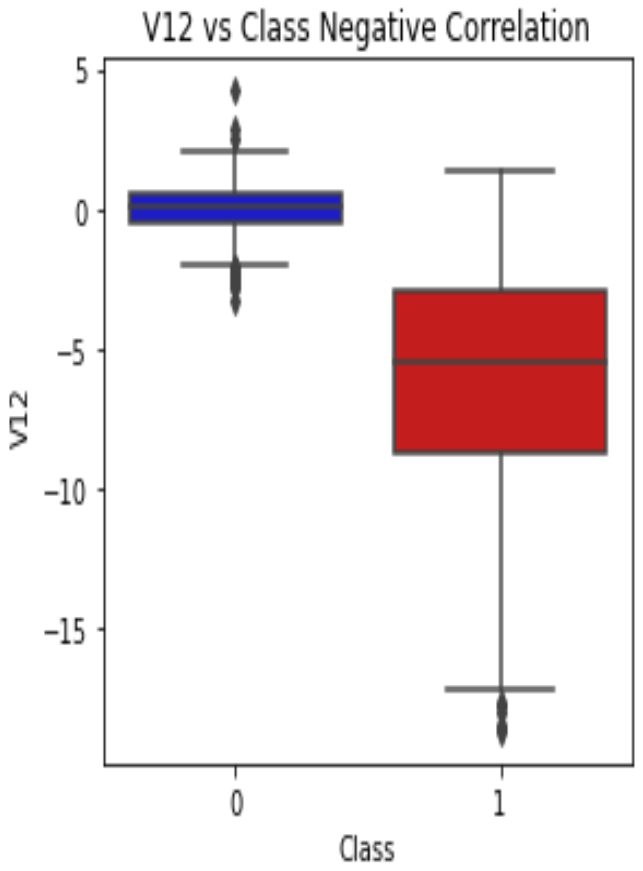
Correlation Matrix (under sample dataset)

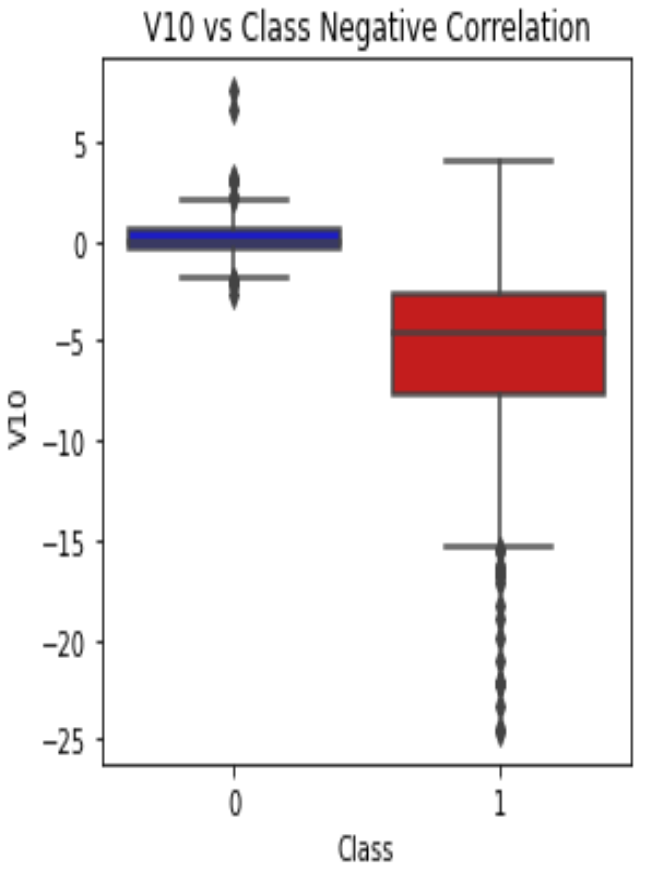

SubSample Correlation Matrix
(use for reference)
scaled_amount
scaled_time
V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
V11
V12
V13
V14
V15
V16
V17
V18
V19
V20
V21
V22
V23
V24
V25
V26
V27
V28
Class
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Negative Correlations:

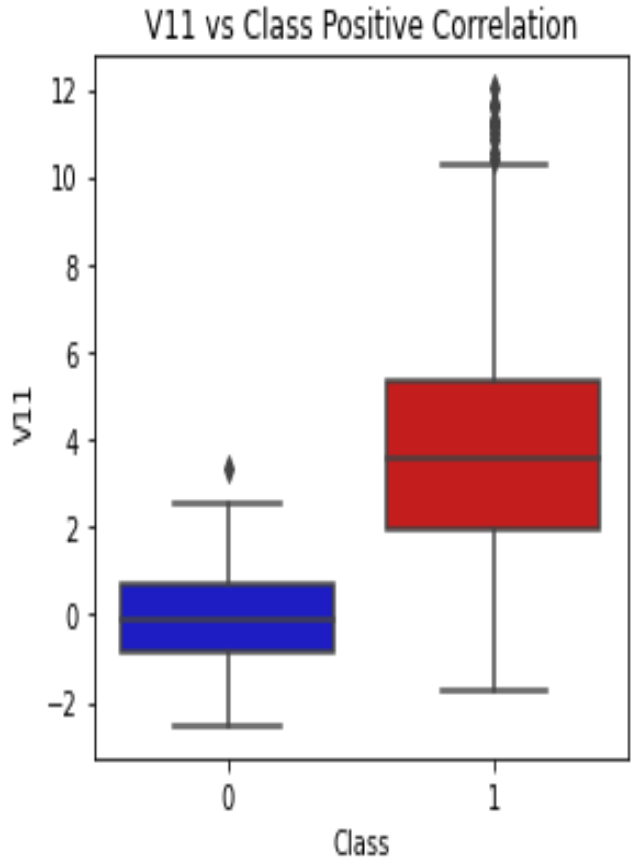


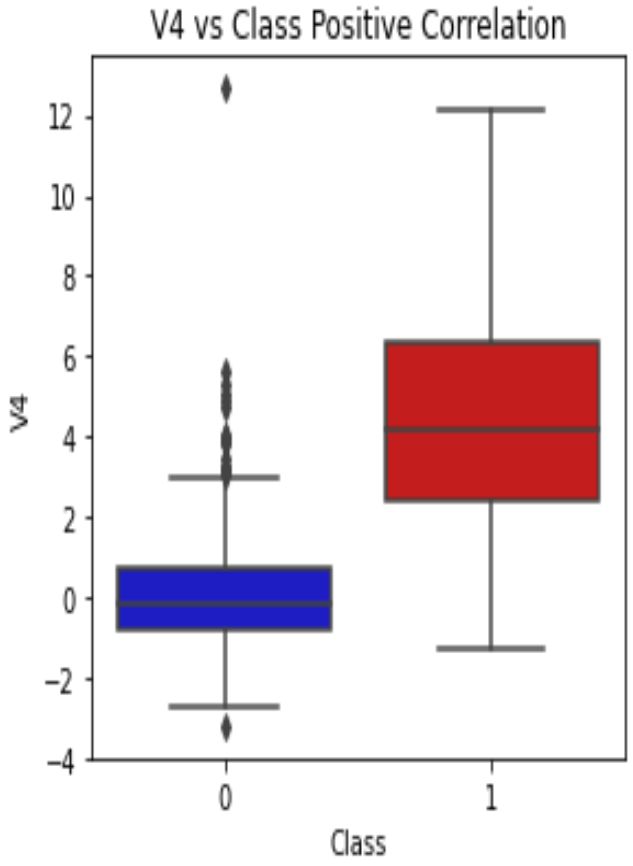


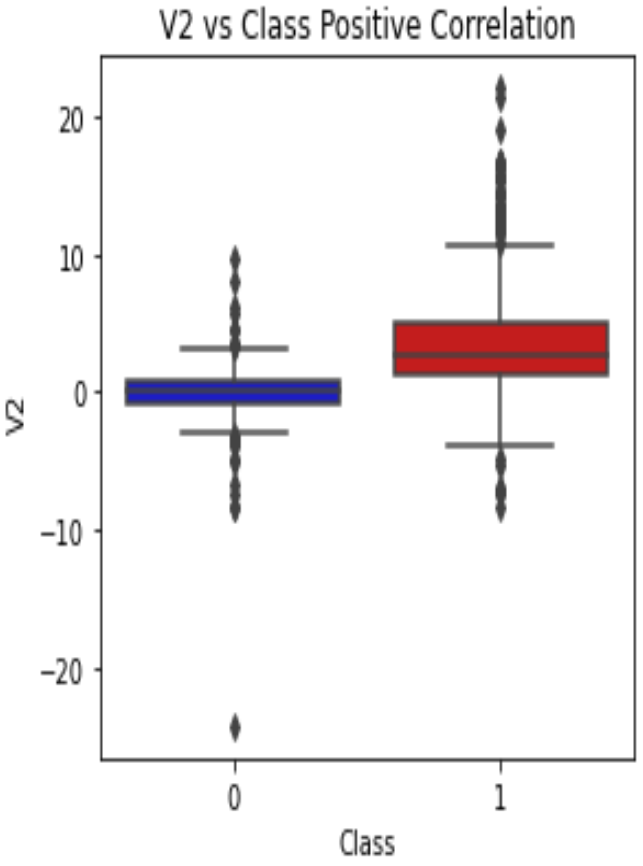


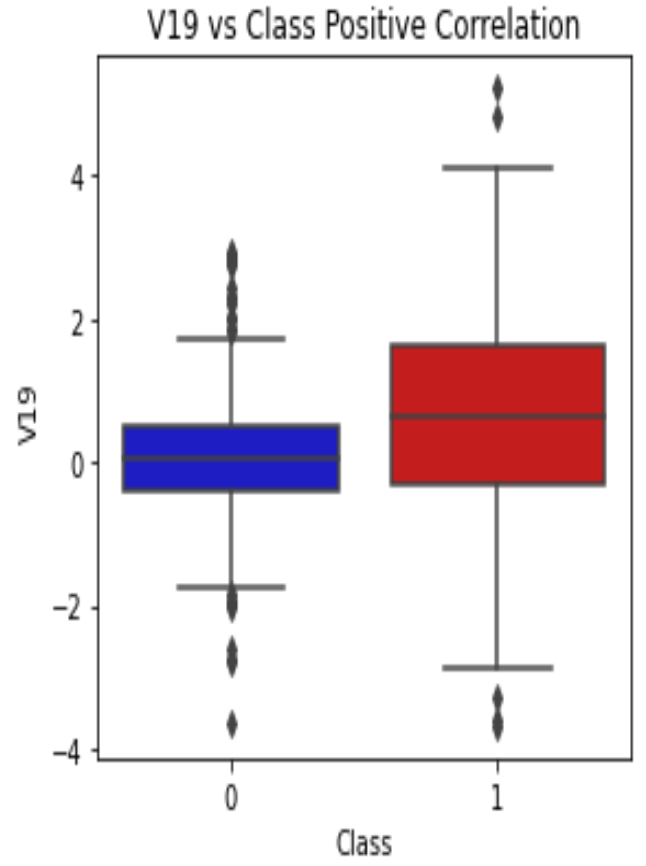


Positive Correlations:







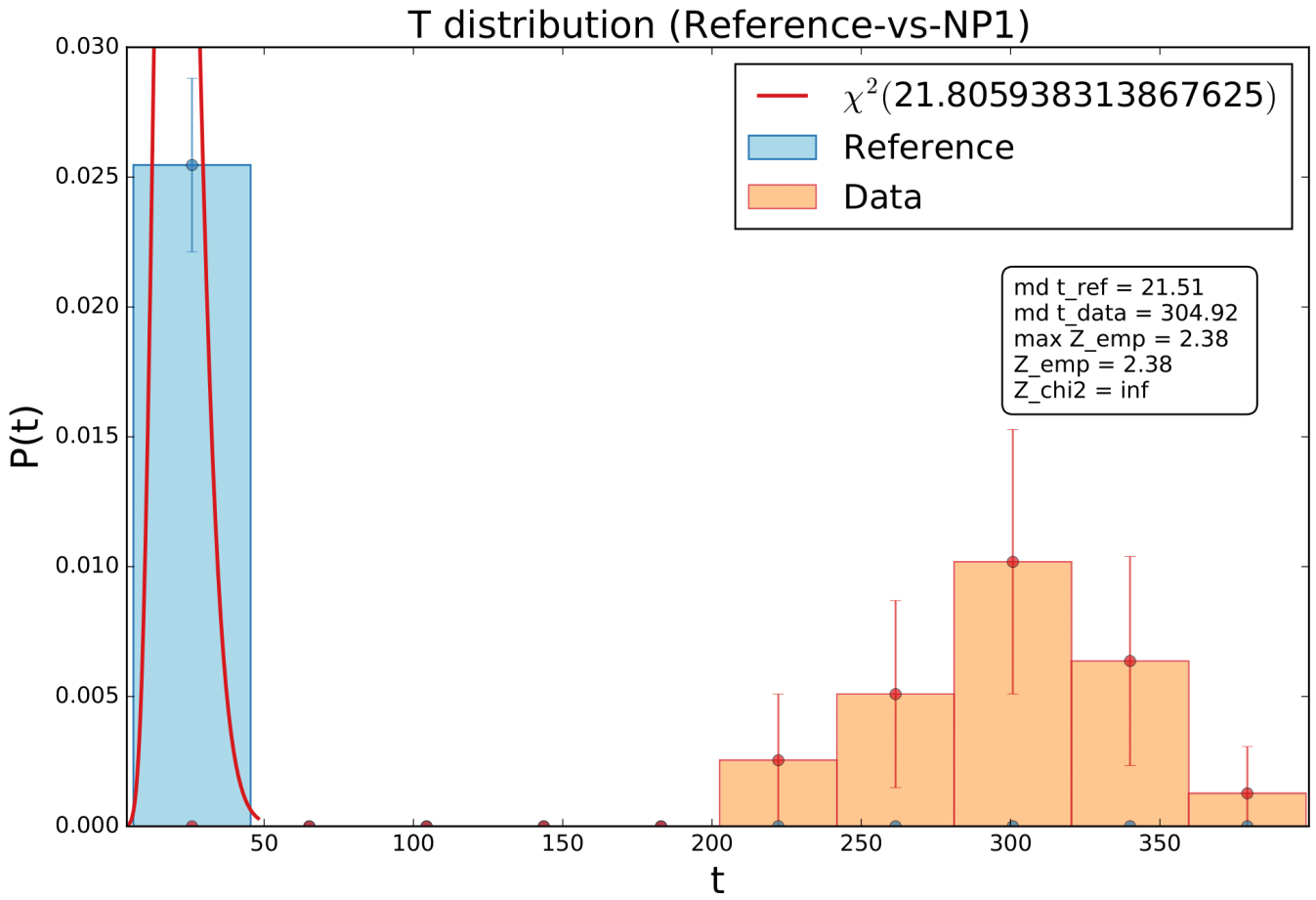


Falkon Library's results on credit card dataset (2D)

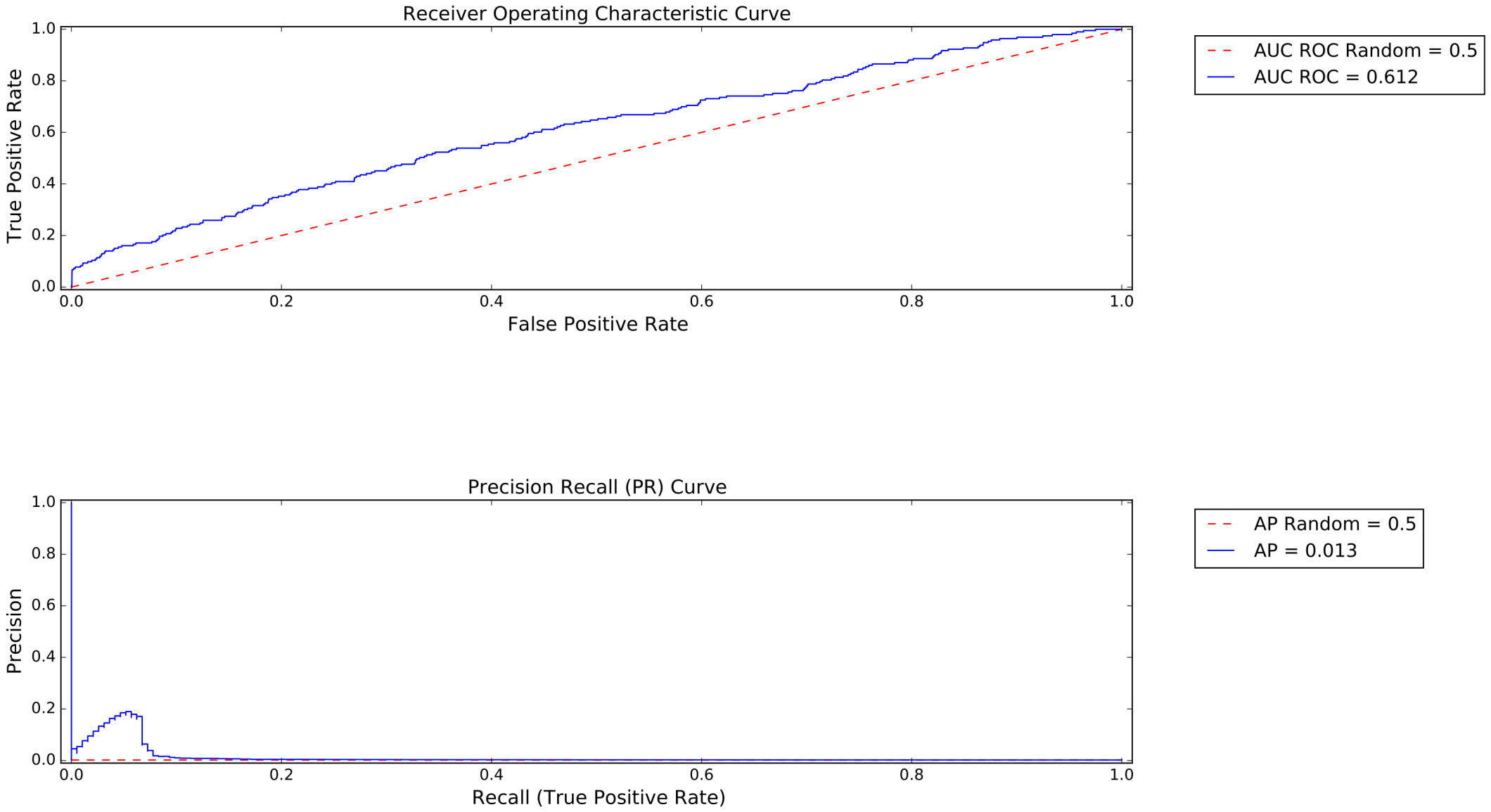
Parameters

- N_0 = 100000 (reference sample)
- N0 = 170 (expected background)
- Ns = 20
- Weight = N0 / N_0
- M = 3000 (centers)
- Lambda = 1e-10 (regularization parameter)
- Falkon sigma = 2.9 (90th percentile pairwise distance for 2D training dataset)
- Degrees of Freedom = 21.8
- KS test statistic = 0.06

T-Distribution



Metric Plots


Receiver Operating Characteristic Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC ROC Random = 0.5
AUC ROC = 0.612
Precision Recall (PR) Curve
Precision
Recall (True Positive Rate)
AP Random = 0.5
AP = 0.013

Computation time

- Reference training (over 100 sample toys): 4h 05m 47s
- Signal detection training (over 20 sample toys): 1h 29m 44s