고려대학교
빅데이터 연구회

KU-BIG

Outlier Detection

유현우 박정진 정희정 송예은 심정은 양수형

목 차

Ⅰ Outlier / Novelty / Anomaly

Ⅱ Evaluate Measure

Ⅲ EDA

PART. I Outlier / Novelty / Anomaly

1. Outlier
   Novelty
   Anomaly

2. Outlier
   (Anomaly)
   Detection

3. Novelty
   Detection

1) Outlier / Novelty / Anomaly

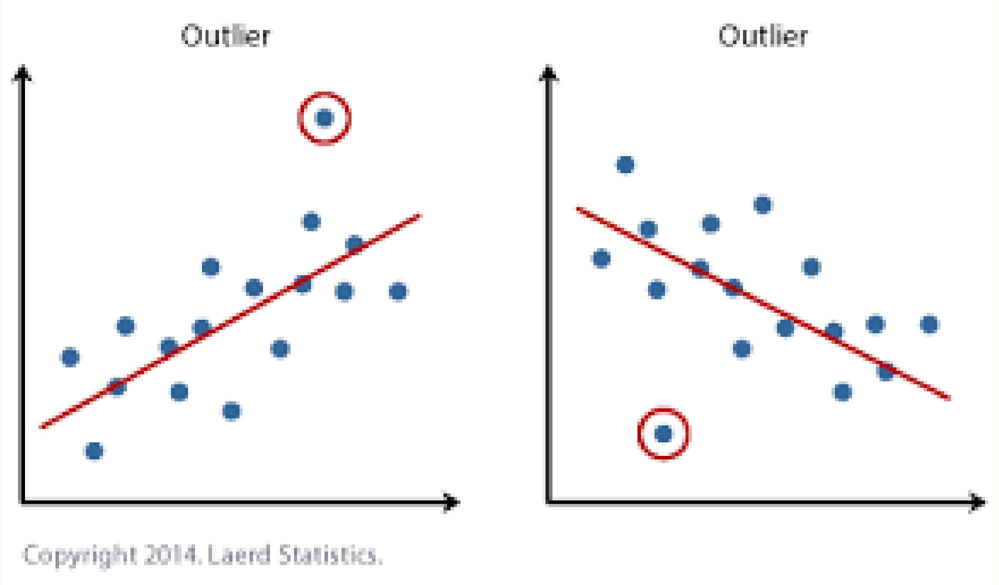
Outliers are also referred to as abnormalities, discordants, deviants, or **anomalies** in the data mining and statistics literature.

(Source: "Outlier Analysis" (Springer), Charu Aggarwal, 2017, http://charuaggarwal.net/outlierbook.pdf)

Outlier = Anomaly

2) Outlier(Anomaly) detection

The training data **contains outliers** which are defined as observations that are far from the others.



Copyright 2014. Laerd Statistics.

(Source: https://scikit-learn.org/stable/modules/outlier_detection.html)

2) Outlier(Anomaly) detection

i) Unsupervised anomaly detection

ii) Supervised anomaly detection

iii) Semi-Supervised

Credit Card Fraud

Credit card fraud is the

Theft and fraud committed

fraudulent source of funds

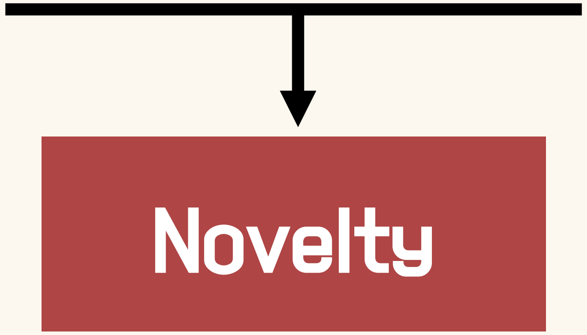
transaction to obtain goo

ving, or to obtain ur

ds from a

3) Novelty detection

The training data **is not polluted by outliers** and we are interested in detecting whether a new observation is an outlier.



Novelty

(Source: https://scikit-learn.org/stable/modules/outlier_detection.html)

PART. Ⅱ Evaluate Measure

1

Metric

For

Out-of-Distribution Detection

2

Better metric

For

Class-imbalanced data

1) Metric for Out-of-Distribution Detection

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

a: TP (True Positive)

b: FN (False Negative)

c: FP (False Positive)

d: TN (True Negative)

$$\text{Accuracy} = \frac{a + d}{a + b + c + d} = \frac{TP + TN}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + TN}$$

(Source: JunGeol Baek, 2019 1st semester Data mining chapter 3. pp.79.)

1) Metric for Out-of-Distribution Detection

		PREDICTED CLASS	
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

a: TP (True Positive)

b: FN (False Negative)

c: FP (False Positive)

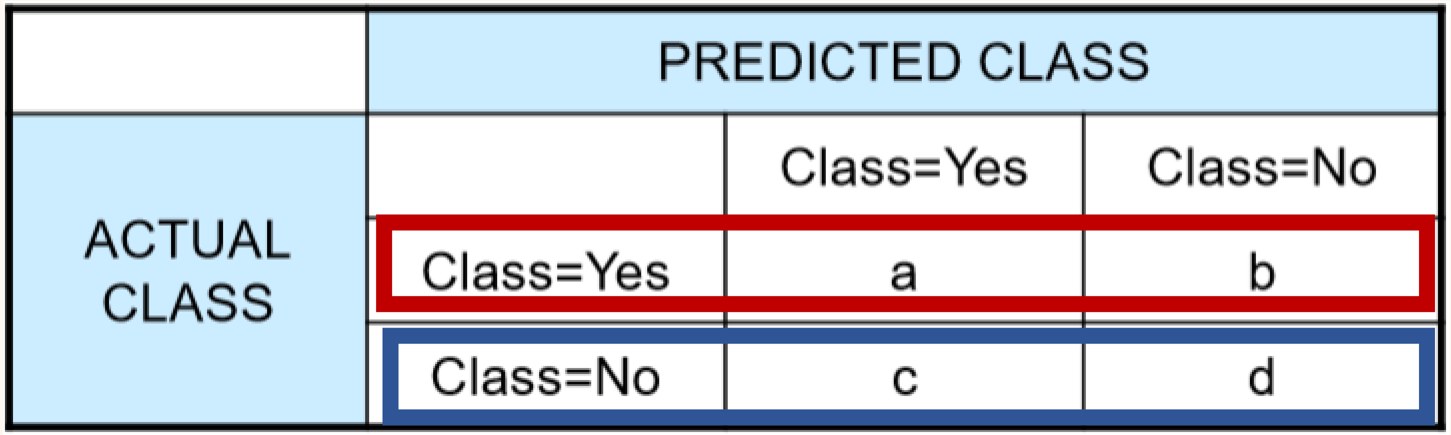
d: TN (True Negative)

$$\boxed{\text{Precision} = \frac{a}{a+c} = \frac{TP}{\text{TP}+\text{FP}}}$$

$$\boxed{\text{Recall} = \frac{a}{a+b} = \frac{TP}{\text{TP}+FN}}$$

(Source: JunGeol Baek, 2019 1st semester Data mining chapter 3. pp.93.)

1) Metric for Out-of-Distribution Detection



		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

a: TP (True Positive)

b: FN (False Negative)

c: FP (False Positive)

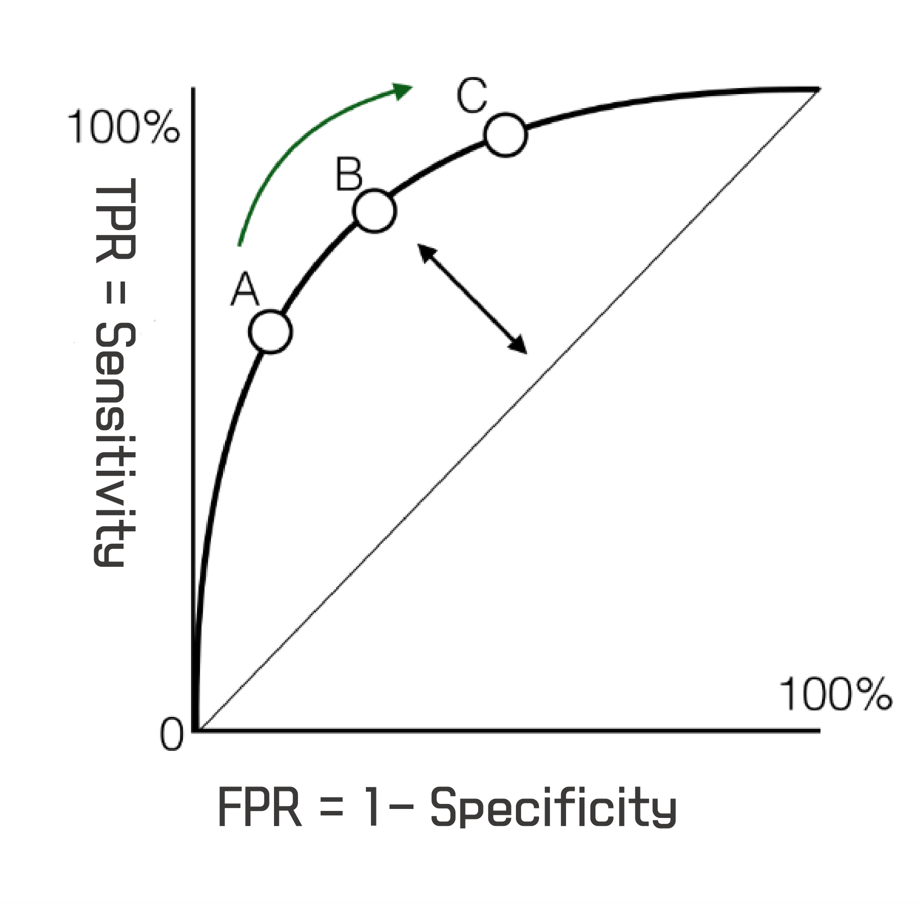
d: TN (True Negative)

$$\mathrm{TPR} = \frac{a}{a+c} = \frac{TP}{\mathrm{TP}+\mathrm{FP}}$$

$$\mathrm{FPR} = \frac{c}{\mathrm{c}+d} = \frac{FP}{\mathrm{FP}+\mathrm{T}N}$$

(Source: JunGeol Baek, 2019 1st semester Data mining chapter 3. pp.84.)

1) Metric for Out-of-Distribution Detection



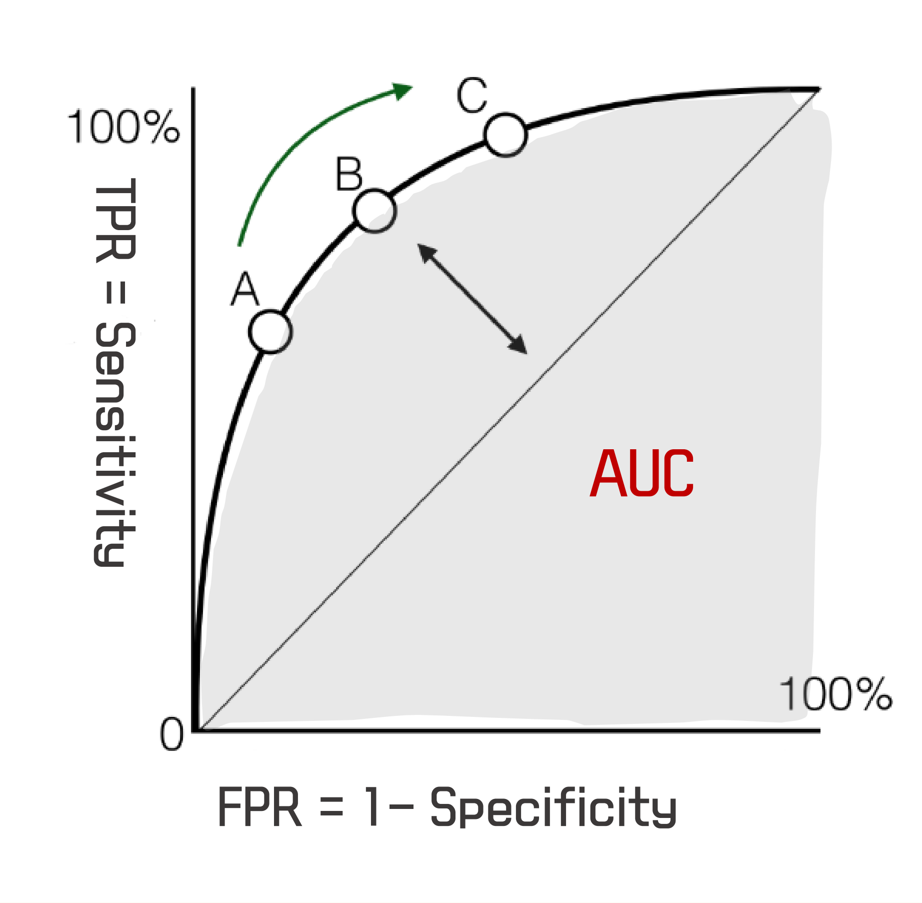
X축 : FPR = FP / (FP + TN)

Y축 : TPR = TP /(TP + FN)

Diagonal line = Random Guessing

Area under ROC curve = AUC

1) Metric for Out-of-Distribution Detection

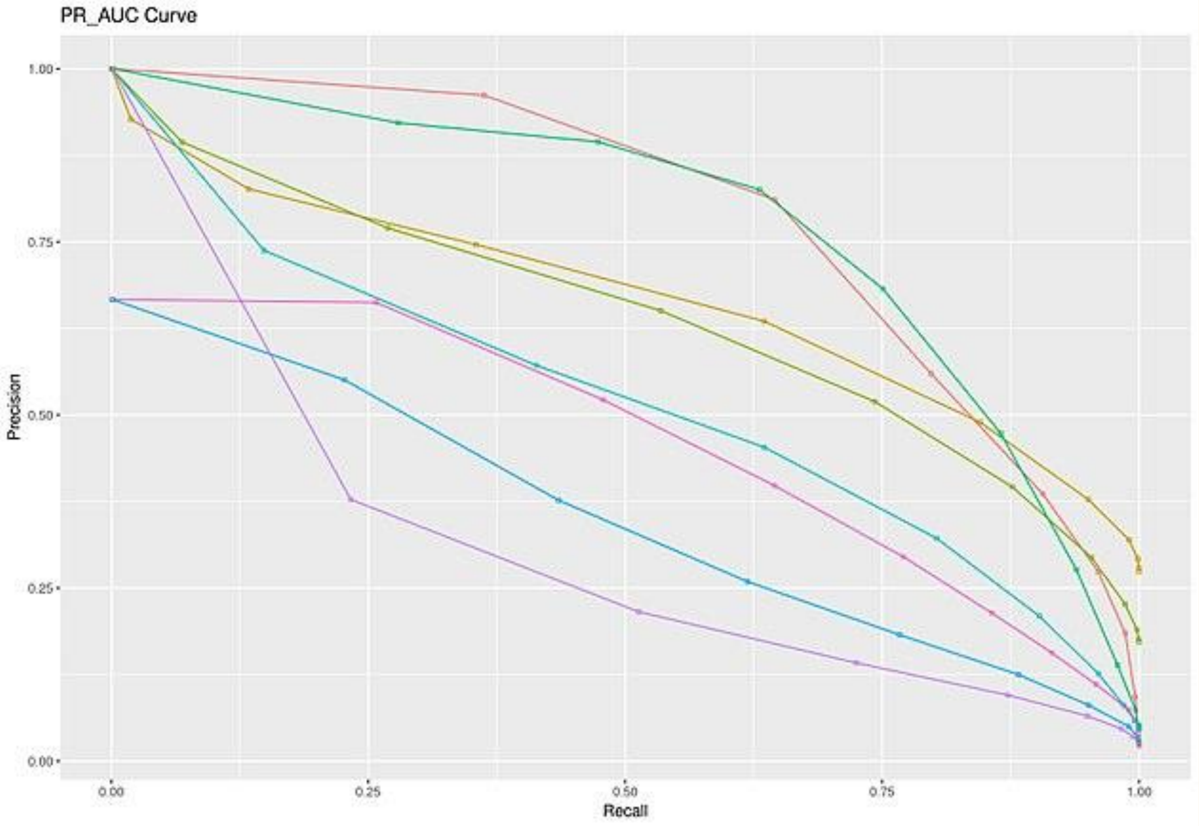


Area under ROC curve = AUC

AUC Range : [0, 1]

100% 맞는 예측 모델일 경우 AUC = 1.

2) Better metric for class-imbalanced data



X축 : Recall : TP / (TP + FN)

Y축 : Precision : TP /(TP + FP)

In Case of imbalanced-Data

⇒ Precision이 FPR에 비해 False Positive를 더 민감하게 잡아낼 수 있다.

⇒ Imbalanced data에서 효과적인 metric!

PART. Ⅲ EDA

1

Description of Data

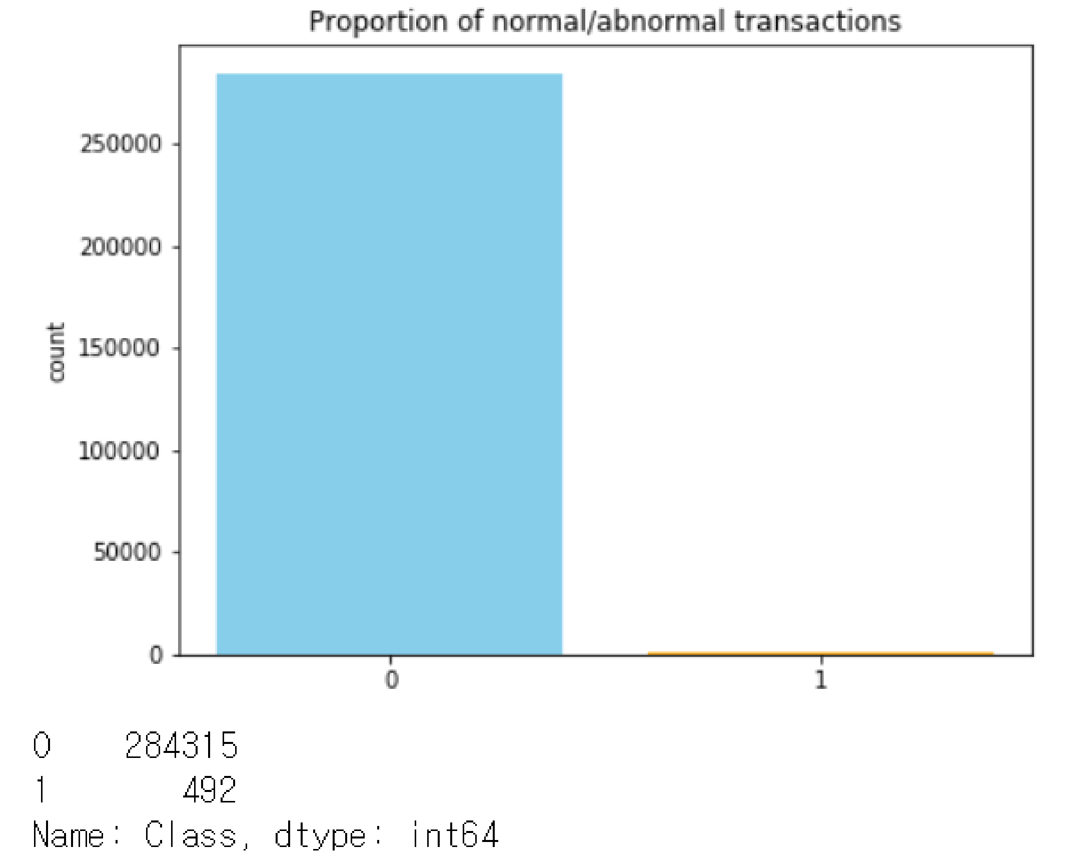
2

Non-linear Relations

1) Description of Data

Credit Card Dataset : (From Kaggle)

- Highly unbalanced data
- 492 frauds out of 284,807 transactions



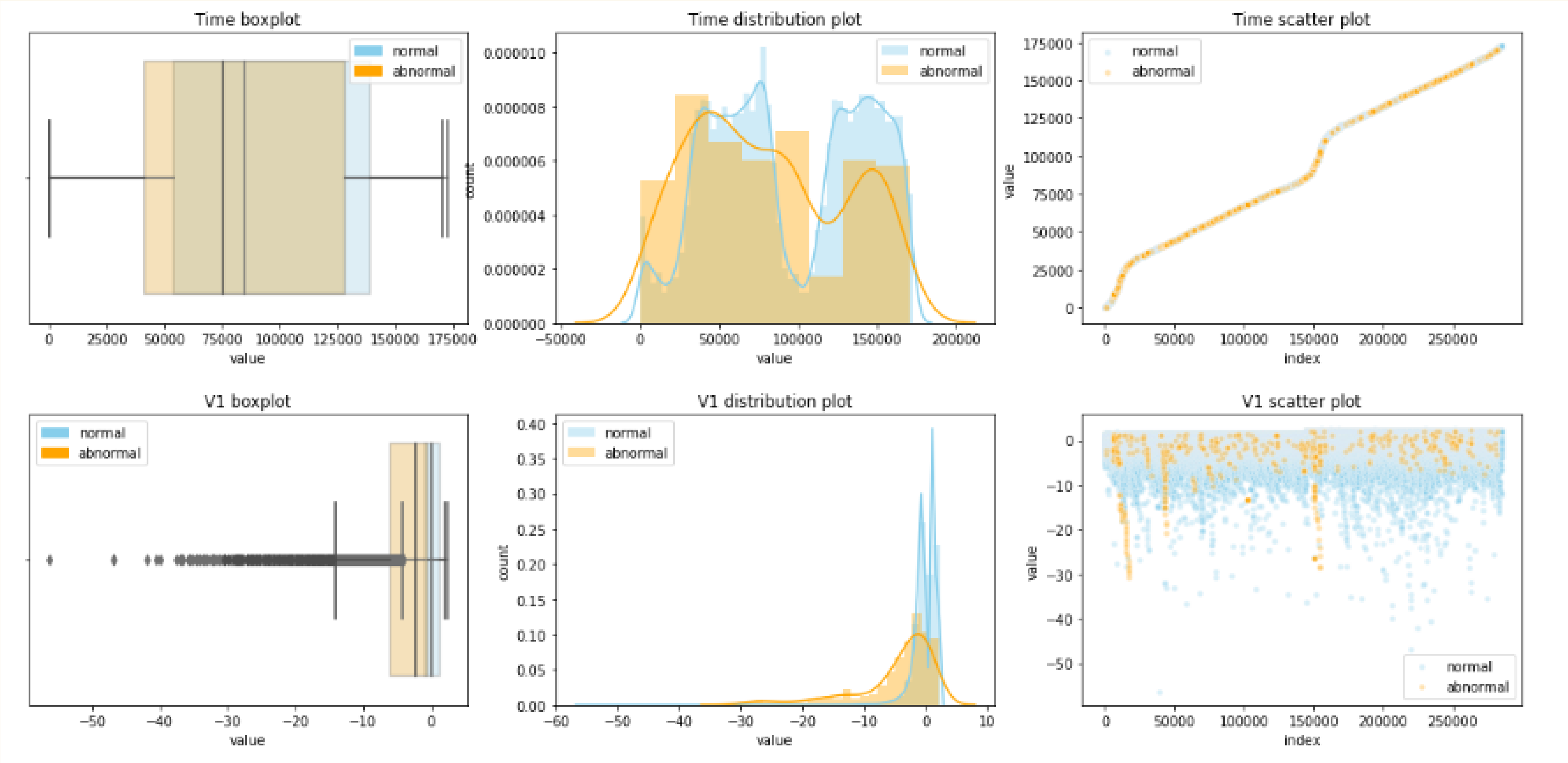
```
0    284315
1       492
Name: Class, dtype: int64
```

1) Description of Data

Input variable :

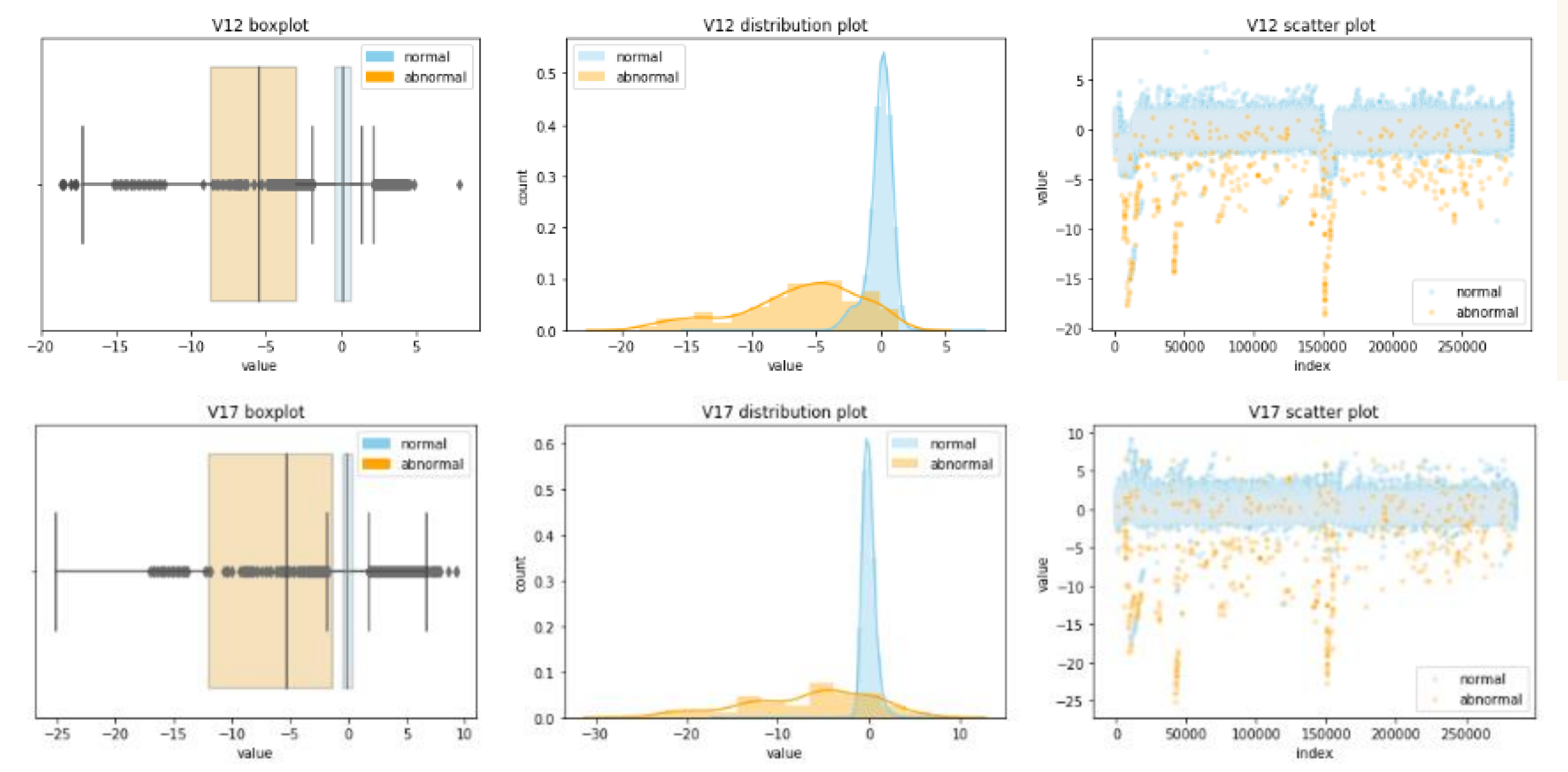
- 28 input variable V, result of PCA transformation.
- Time : seconds (about 48 hours)
- Amount : transaction amount.

1) Description of Data



Plot of Time and V1

1) Description of Data



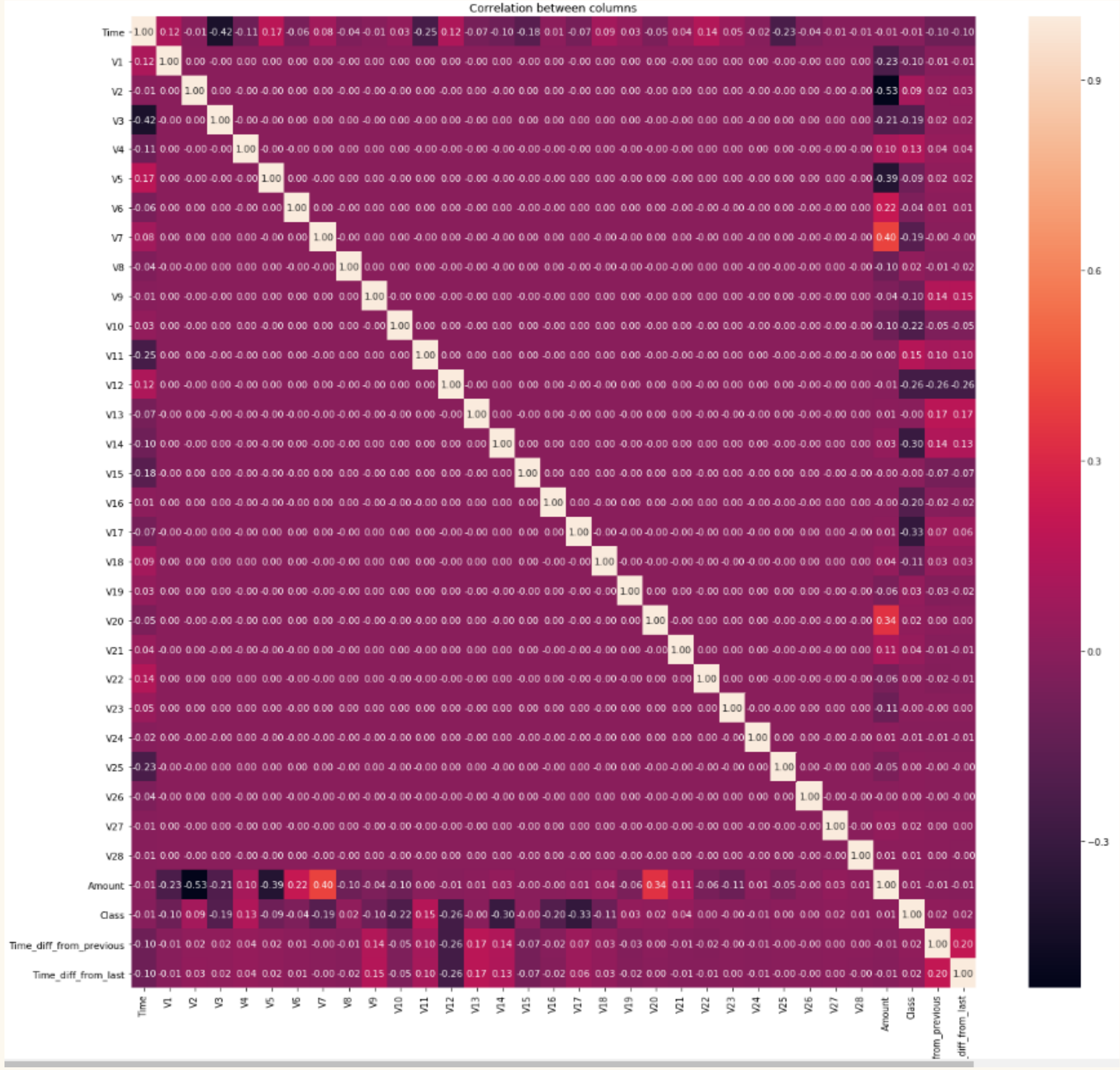
Plot of V12 and V17

2) Linear Relations

Correlation Matrix

- V′s : Principal Components

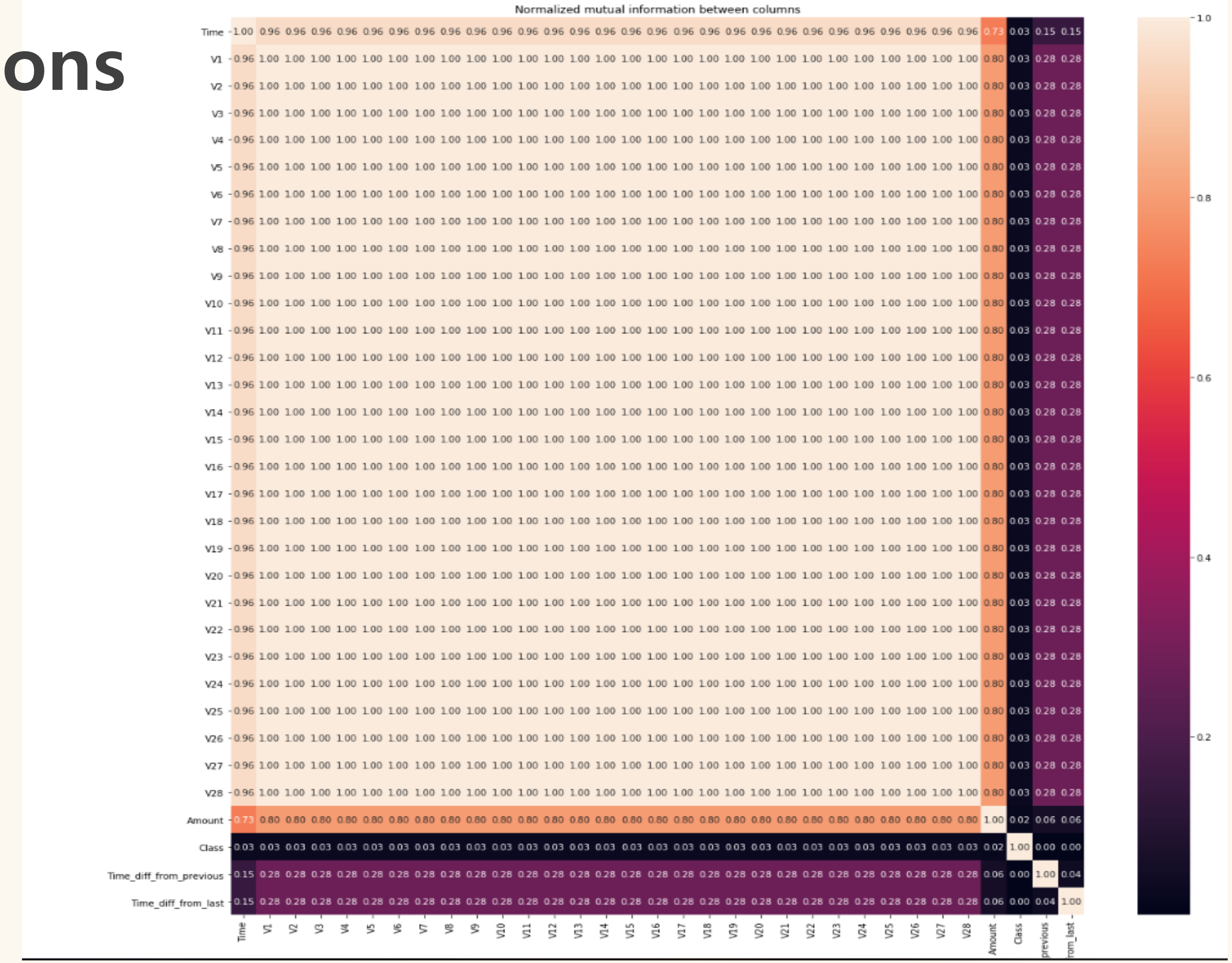
⇒Linearly independent



2) Nonlinear Relations

Mutual Information Matrix

- All V's are nonlinearly dependent.
- All Vs and amount are nonlinearly dependent.



Normalized mutual information between columns

	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	Amount	Class	previous	rom_last
Time	1.00	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.73	0.03	0.15	0.15
V1	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V2	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V3	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V4	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V5	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V6	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V7	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V8	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V9	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V10	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V11	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V12	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V13	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V14	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V15	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V16	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V17	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V18	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V19	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V20	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V21	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V22	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V23	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V24	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V25	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V26	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V27	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
V28	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.03	0.28	0.28
Amount	0.73	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	1.00	0.02	0.06	0.06
Class	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02	1.00	0.00	0.00
Time_diff_from_previous	0.15	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.06	0.00	1.00	0.04
Time_diff_from_last	0.15	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.06	0.00	0.04	1.00

Do you have any question?

Thank you for your attention.