Classification

재무상태표와 2년 후 파산 여부

4조 김민회 남승지 신예진 오태환 조민주 주일찬

Ⅰ

목차

0

조원소개

조원소개

남승지

데이터 전처리,

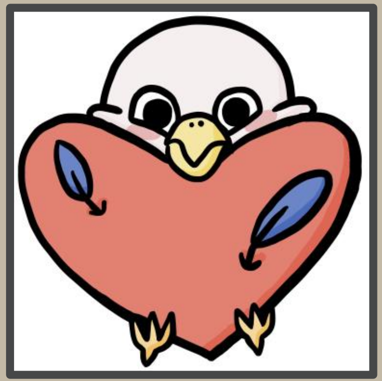
시각화,

ppt 제작

오태환

데이터 전처리,

모델링



조민주

Feature Importance 출력,

모델링

김민회

변수 조사,

변수 선택

신예진

변수 선택,

ppt제작 및 발표

주일찬

변수 조사,

변수 선택

0

도입

데이터 설명

도입

Short-term liabilities

Total assets

Cost of products sold

inventory

Depreciation

net profit

Receivables

sales

Gross profit

Profit on sales

Depreciation

equity

Current assets

Total liabilities

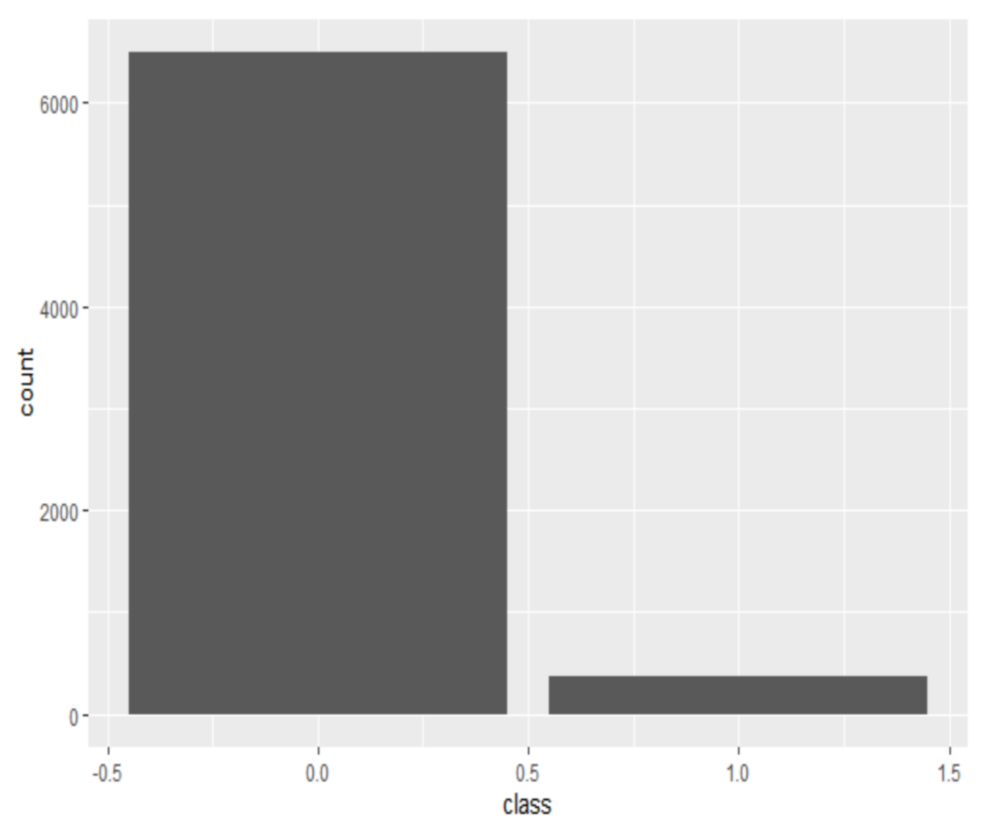
Operating expenses

Fixed assets

도입

[6855개 회사의 64개 독립변수(Attr) + class]

Polish companies bankruptcy Data



Class: 파산기업 1, 정상기업 0

X1 net profit / total assets
X2 total liabilities / total assets
X3 working capital / total assets
X4 current assets / short-term liabilities
.
.
.

X63 sales / short-term liabilities
X64 sales / fixed assets

Attr59	Attr60	Attr61	Attr62	Attr63
0.25454	13.632	3.693	69.389	5.2602
0	?	37.886	0	?
0.12538	?	2.5649	98.95	3.6887
0	8.9302	10.287	40.355	9.0448

총 6132개
이걸 쓸 수 있을까?

Attr33	28
Attr47	57
Attr52	60
Attr32	72
Attr21	112
Attr41	142
Attr24	149
Attr53	162
Attr54	162
Attr28	162
Attr64	162
Attr45	418
Attr60	420
Attr27	462
Attr37	3100

1

Preprocessing

1. **Correlation (0.70) 높은 변수들 제거**
2. Outlier 관측치 제거
3. NA Imputation
4. Skew된 변수 transformation
5. Scaling

설명변수 선택

Correlation이 높은 변수 제거 – 기준: 0.7

각 Attr 별 Correlation이 높은 다른 Atrr 추출

기준: 0.7

Atrr 1	1 ,7, 11, 14, 22, 35
Atrr 2	2, 10, 25, 38, 51
Atrr 3	3
Atrr 4	4, 40, 46
…	…
Atrr 63	33, 63
Atrr 64	28, 53, 54, 64

추출 결과 토대로

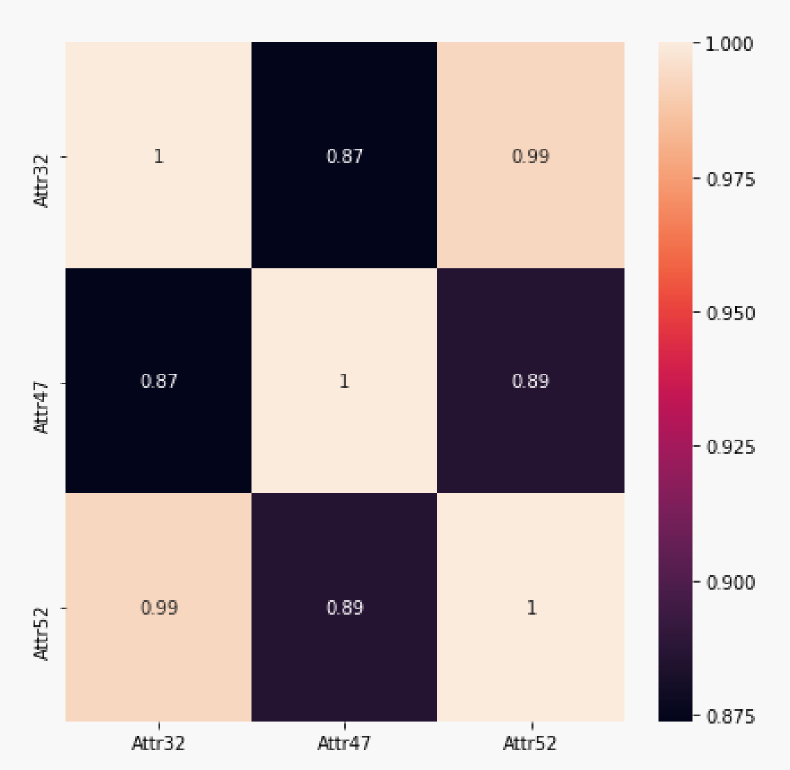
Attr 분류

8, 17, 50
3
2, 10, 25, 38, 51
1, 7, 11, 14, 22, 35, 48
…
32, 47, 52
33, 63

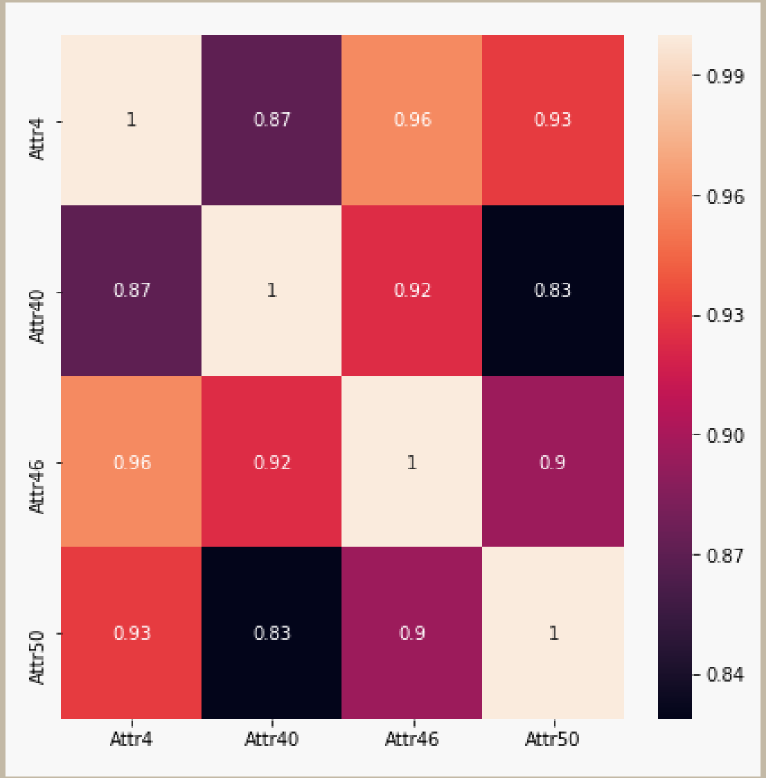
설명변수 선택

Correlation이 높은 변수 제거 – 기준: 0.7

Example



X32, X47, X52 –> X47



X4, X40, X46, X50 –> X46

Correlation 바탕으로

각 그룹에서 대표 항목 선정

총 37개 변수 제거

X1, X5, X6, X9, X10, X15, X17, X18, X19, X20, X21, X26, X27, X29, X41, X42, X45, X46, X47, X54, X55, X57, X59, X60, X61, X63, X64

1

Preprocessing

1. Correlation (0.70) 높은 변수들 제거
2. Outlier 관측치 제거
3. NA Imputation
4. Skew된 변수 transformation
5. Scaling

Outlier

Outlier 제거

outlier 기준: 각 변수에서 가장 작은/큰 값 3개

회사	Attr
6818	Attr17, Attr18, Attr29, Attr42, Attr46, Attr47
1594	Attr21, Attr42, Attr61, Attr63, Attr64
2100	Attr17, Attr19, Attr26, Attr46
4680	Attr17, Attr26, Attr46, Attr63
4995	Attr18, Attr29, Attr60, Attr63
2556	Attr19, Attr42, Attr60
4120	Attr5, Attr45, Attr60
5305	Attr19, Attr21, Attr45
5936	Attr10, Attr19, Attr42
5811	Attr19, Attr26, Attr42

outlier가 공통적인 data 조사

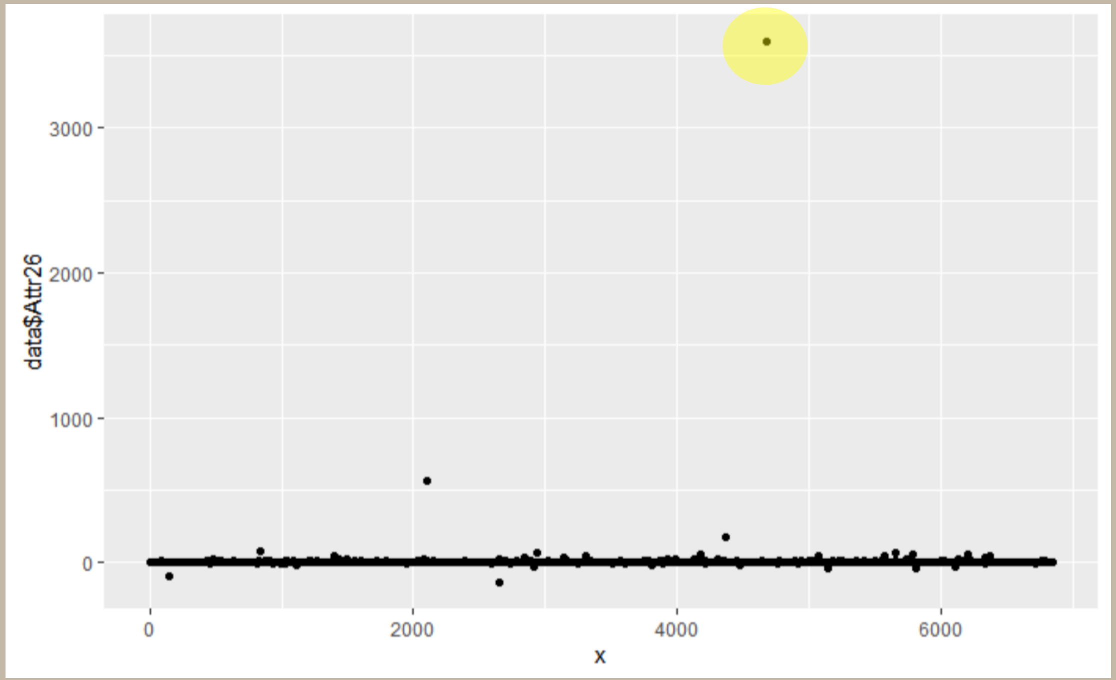
세 개 이상의 Attr 에서 outlier를 가지는 회사 총 10개

(R표기)

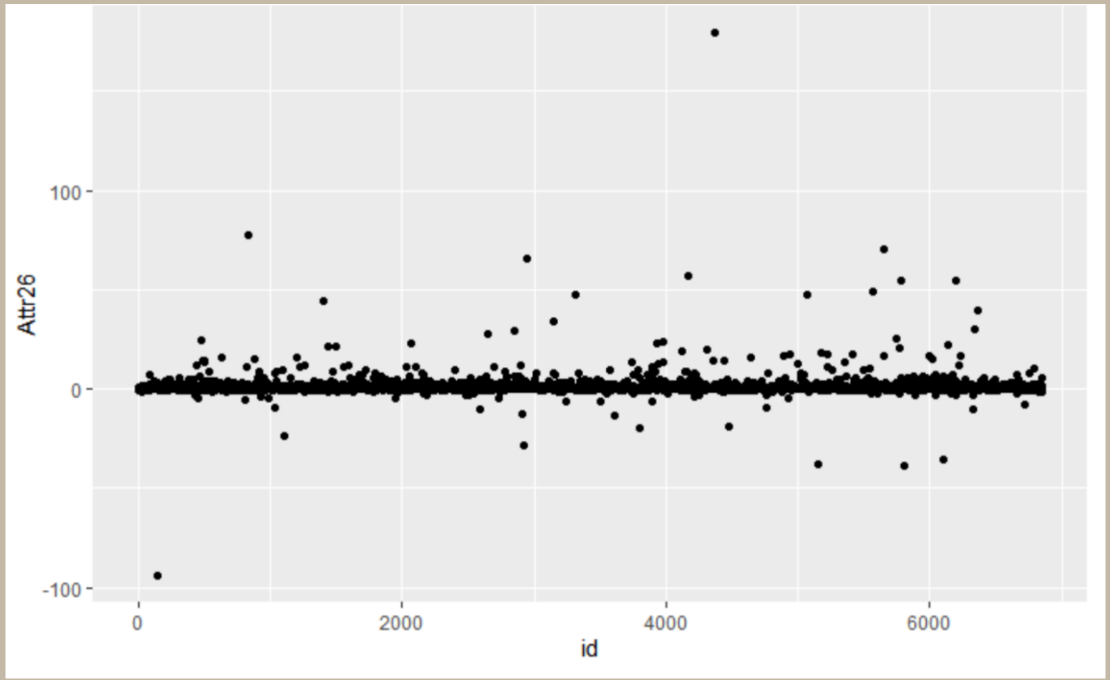
Outlier

Outlier 제거

Plot을 통해 눈에 띄는 outlier 제거



제거 전



제거 후

999,3134,3528,2556,2655,6425,6233,4942,4612,1935,2064,6818,1594,2100,4680,4995,4120,5305,5936,5811

1

Preprocessing

1. Correlation (0.70) 높은 변수들 제거
2. Outlier 관측치 제거
3. NA Imputation
4. Skew된 변수 transformation
5. Scaling

NA Imputation

결측치 제거 & 채우기

NA가 자료의 50%에 가까운

X37 제거

Attr 33: 28	Attr 32: 72	Attr 24: 149	Attr 28: 162	Attr 60: 420
Attr 47: 57	Attr 21: 112	Attr 53: 162	Attr 64: 162	Attr 27: 462
Attr 52: 60	Attr 41: 142	Attr 54: 162	Attr 45: 418	~~Attr 37: 3100~~

NA Imputation

R mice package

pmm

(predictive mean matching)

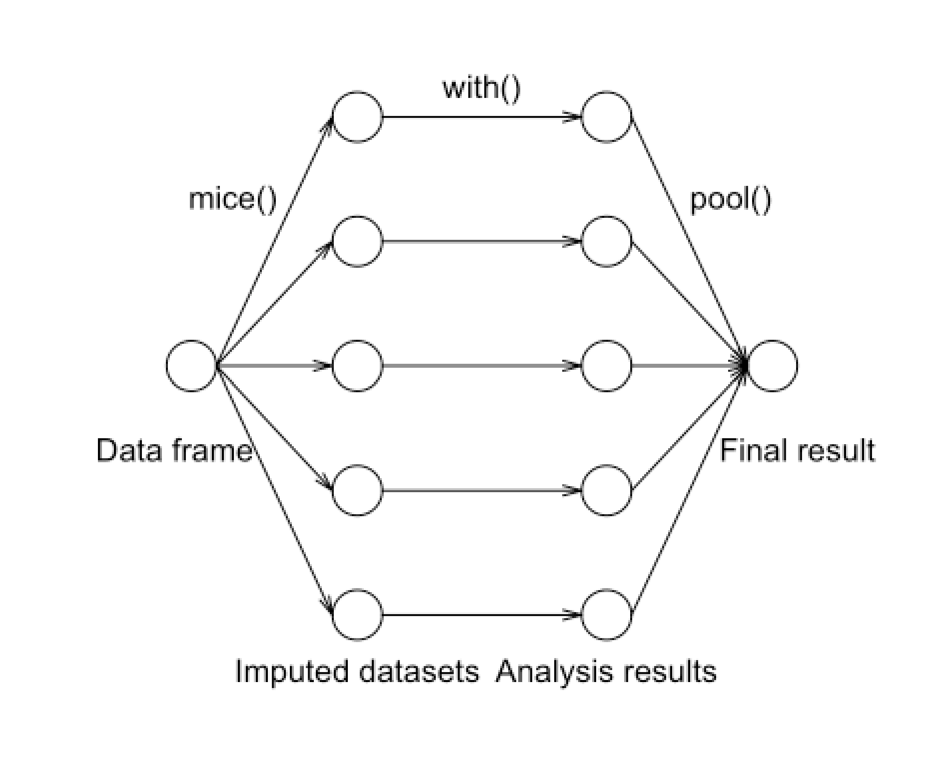
Attr17	Attr21	Attr26	Attr27	Attr29	Attr37
3.32760	1.08530	0.0426070	0.136910	4.5645	1.080400
NA	NA	NA	NA	1.2218	NA
2.09540	0.50040	-0.7660500	-37.147000	2.9728	9.219200
5.65630	0.97491	1.0241000	1.020900	5.4536	NA
3.69070	1.44160	0.9648700	173.050000	7.2973	2.391600
2.12720	1.07240	0.3717400	0.000000	4.0682	NA
3.93820	0.96884	0.5059800	2.946700	5.0540	NA
2.67870	1.38160	0.3628200	1.823200	4.9982	1.671600
1.18240	1.52750	0.1300600	0.481360	4.7943	1.412900
1.00450	0.85659	-0.1178900	-5.466000	3.2373	NA

Attr17	Attr21	Attr26	Attr27	Attr29	Attr37
0.84934	0.7755600	0.7763700	4.4688e+01	1.2218	2.8985e+00
2.09540	0.5004000	-0.7660500	-3.7147e+01	2.9728	9.2192e+00
1.14790	0.9817000	0.2280400	3.3677e+00	2.2221	2.0512e+01
0.88354	0.9373600	0.0589920	3.6921e-01	3.3389	9.6395e-01
4.58490	1.2555000	1.3392000	9.9514e+00	2.2750	8.6046e-01
2.36050	0.8473300	0.6975400	1.2663e+00	4.0935	2.2543e+00
10.80800	1.0812000	3.8298000	3.3689e+02	3.0075	5.4193e+00
22.38300	0.9927300	3.2823000	1.9795e+01	3.8042	3.3541e+02
1.10810	0.9737100	0.1318700	7.4779e-02	3.5765	1.7933e+00
10.29600	0.9195000	0.9125400	1.1177e+02	3.1390	1.0774e+03

NA Imputation

결측치 제거 & 채우기

R mice package: pmm (predictive mean matching)



Bayesian Linear Regression

1. Estimate a linear regression model
2. Draw randomly from the posterior predictive distribution of β^ and produce a new set of coefficients $\beta*$
3. Calculate predicted values for **observed and missing** *Y*
4. For each case where *Y* is missing, find the closest predicted values among cases where *Y* is observed.
5. Draw randomly one of these three close cases and impute the missing value *Yi* with the observed value of this close case.
6. Steps 1-5 are repeated several times.

https://statisticsglobe.com/predictive-mean-matching-imputation-method/

1

Preprocessing

1. Correlation (0.70) 높은 변수들 제거
2. Outlier 관측치 제거
3. NA Imputation
4. Skewed된 변수 transformation
5. Scaling

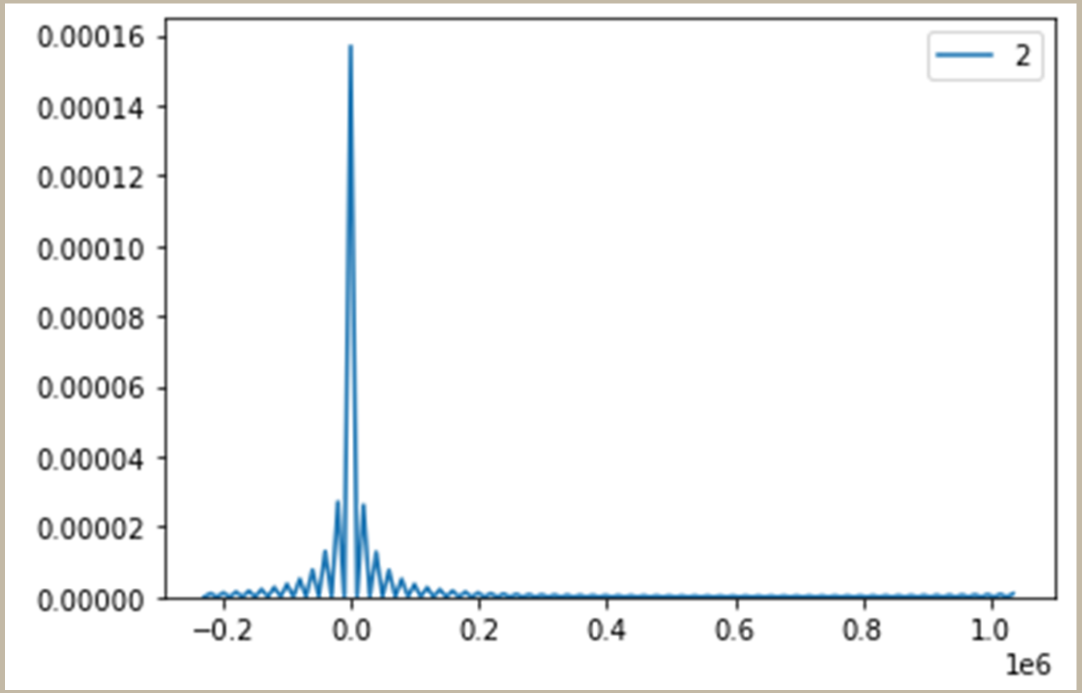
skewed 자료처리

log transformation

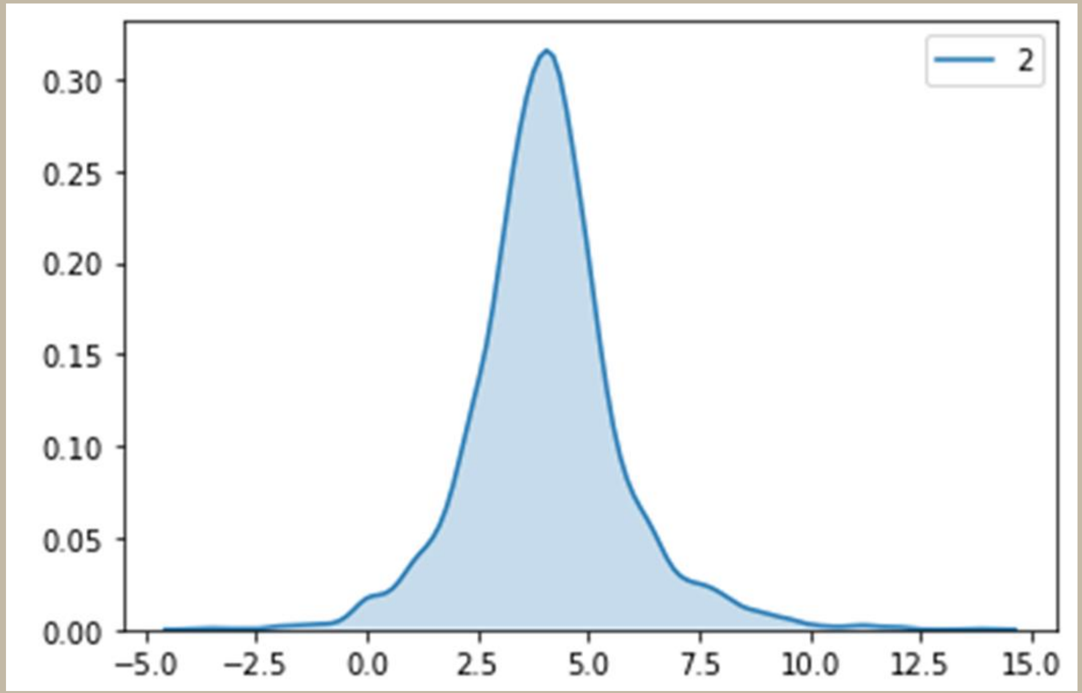
abs(skew_score) 〉1인 경우 → log transformation

최솟값이 음수이거나 0인 경우 → 적절한 값을 더해 양수로 만든 후 log transformation

Example X19



변환 전



변환 후

1

Preprocessing

1. Correlation (0.70) 높은 변수들 제거
2. Outlier 관측치 제거
3. NA Imputation
4. Skew된 변수 transformation
5. Scaling

Scaling

Standard Scaler 이용

평균이 0과 표준편차가 1이 되도록 변환

Attr4	Attr5	Attr6	Attr7
1.8368	34.382	-0.026711	-0.020067
?	29.678	-1.139300	0.760520
1.4678	34.555	0.000000	-0.440760
4.5944	117.65	0.251540	0.148750
2.5745	-26.928	0.617540	0.282690
...	...	...	...
1.015	-14.334	0.260950	0.333530
0.48654	-205.37	-0.120150	-0.143300
1.1855	-5.3824	0.015922	0.008700
3.2569	92.092	0.000000	0.073160

변환 전

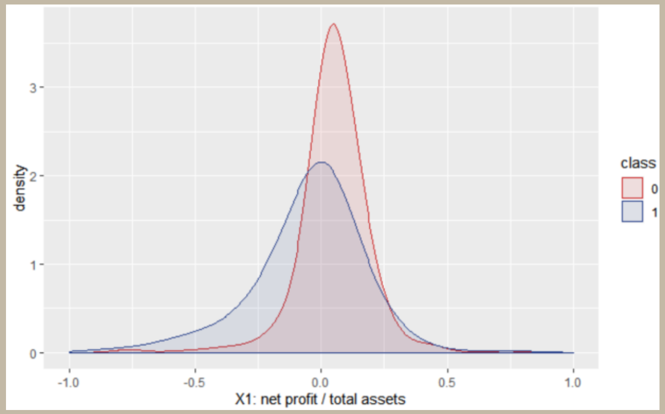
Attr4	Attr5	Attr6	Attr7
-0.021269	-0.009106	-0.007562	-0.149733
-0.022566	-0.009449	-0.206432	1.266020
-0.022380	-0.009094	-0.002787	-0.912745
-0.012970	-0.003043	0.042174	0.156451
-0.019049	-0.013570	0.107595	0.399378
...	...	...	...
-0.023742	-0.012653	0.043856	0.491587
-0.025333	-0.026562	-0.024264	-0.373241
-0.023229	-0.012001	0.000059	-0.097558
-0.016995	-0.004904	-0.002787	0.019353

변환 후

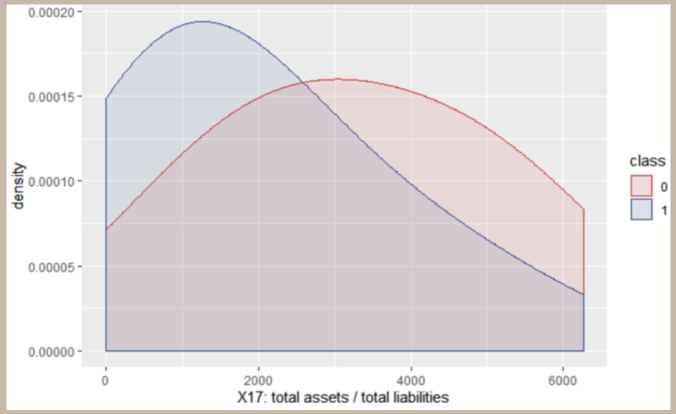
2

Feature Extraction

Feature Extraction 1

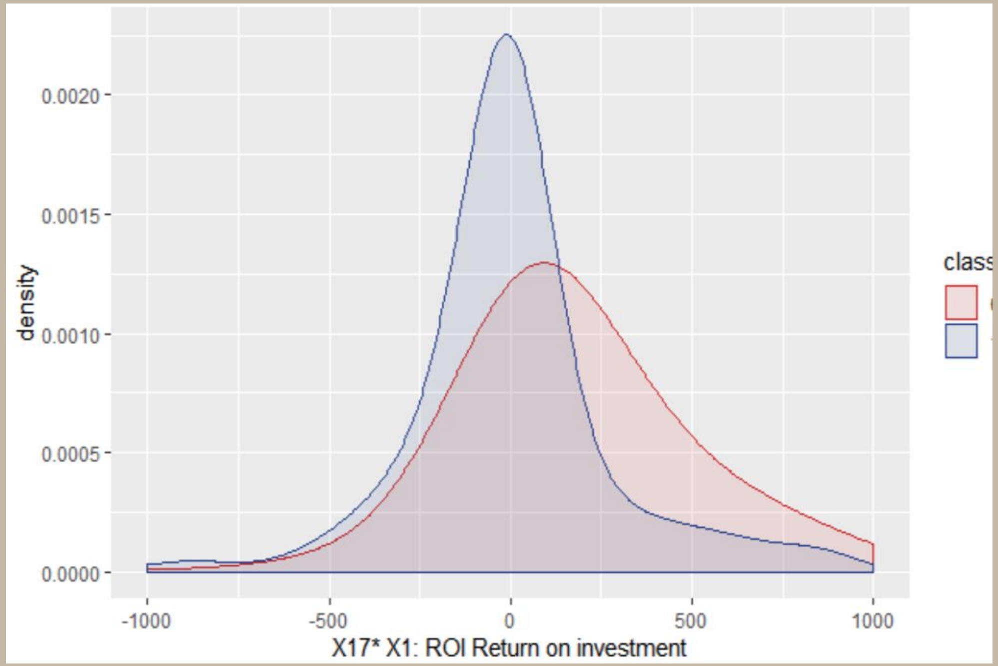


X1: net profit/total assets

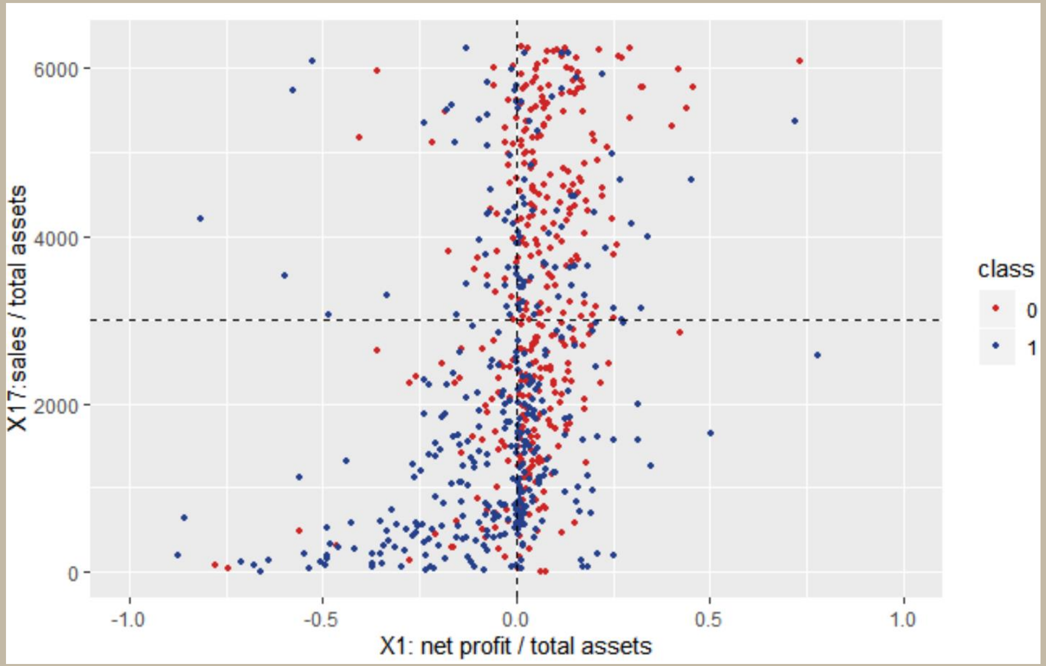


X17: total assets/total liabilities

$$X1*X17 = \frac{\text{Net profit}}{\text{Total liabilities}}$$

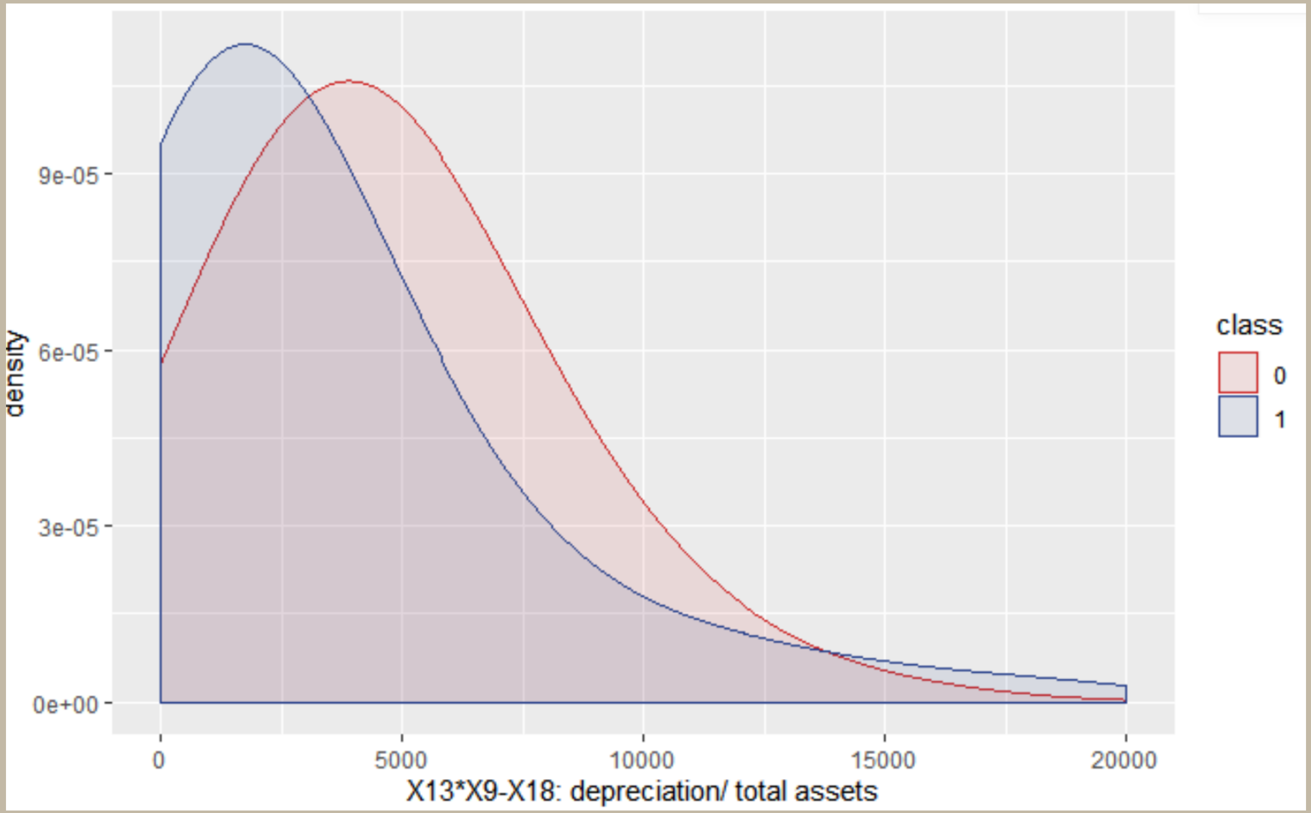


ROI: net profit/total liabilities



ROI: net profit/total liabilities

Feature Extraction 2



$$\frac{\text{Depreciation}}{\text{Total assets}}$$

PCA

변수 37개 제거

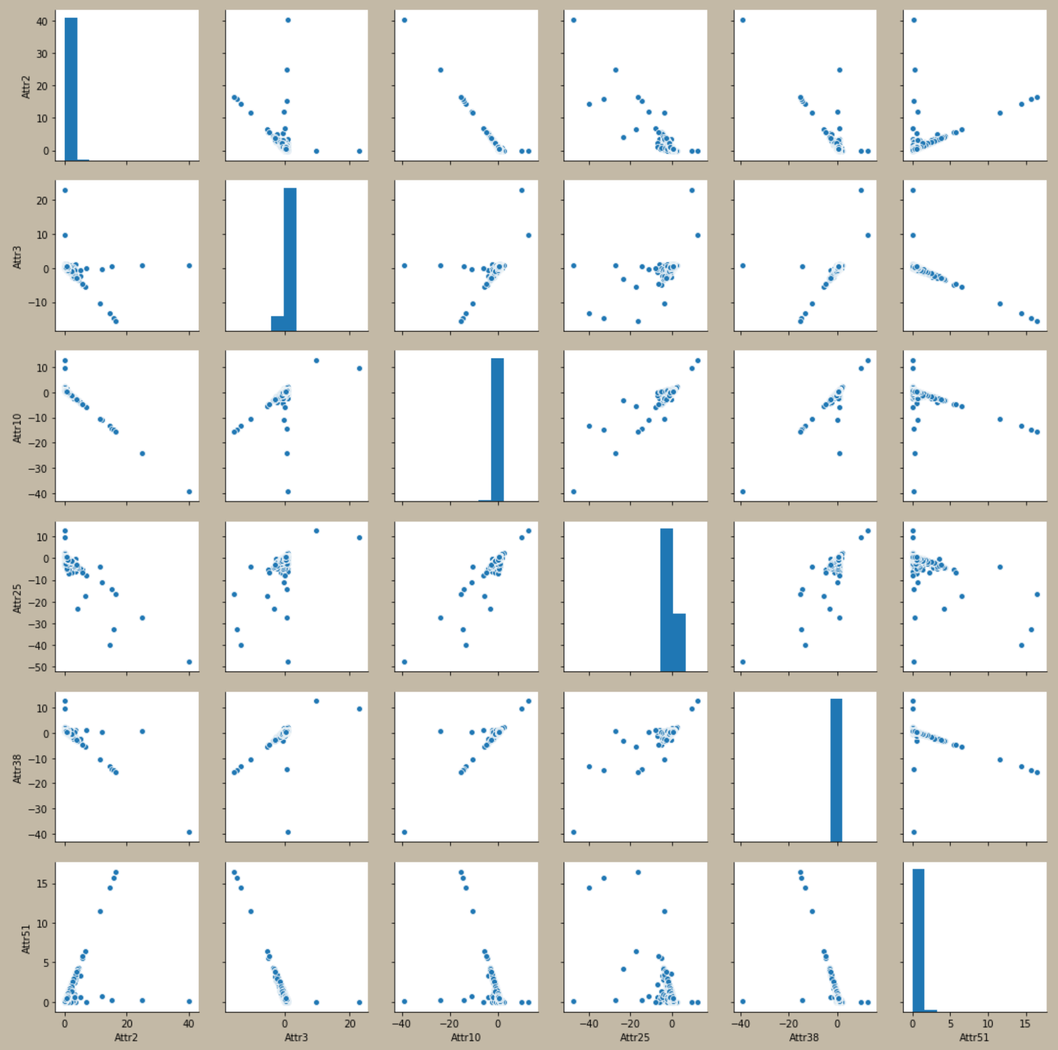
→ 변수들 간 correlation을

PCA 차원축소로

처리할 수 있지 않을까?

→ PCA 차원축소는 변수해석에 어려움이 있지만, 같은 의미의 변수들을 1차원으로 축소하면 해석이 가능할 것 같다.

1, 7, 11, 14, 22, 35, 48 변수의 correlation plot



PCA

X19 gross profit / sales

X23 net profit / sales

X30 (total liabilities – cash) / sales

X31 (gross profit + interest) / sales

X39 profit on sales / sales

X43 rotation receivables + inventory turnover in days

X44 (receivables * 365) / sales

X49 EBITDA (profit on operating activities – depreciation) / sales

X56 (sales –cost of products sold) / sales

X58 total costs /total sales

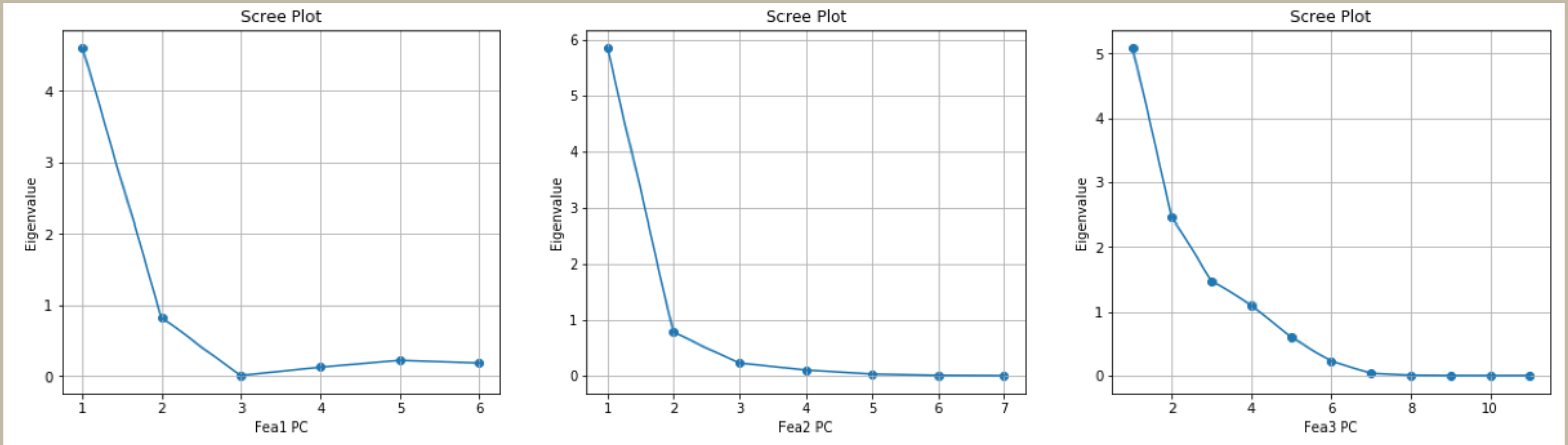
X62 (short–term liabilities *365) / sales

변수들의 의미를 봤을 때, 거의 기업의 현금유동성과 관련이 있는 변수였다.

현금 유동성을 설명하는 새로운 PCA 변수생성!

PCA

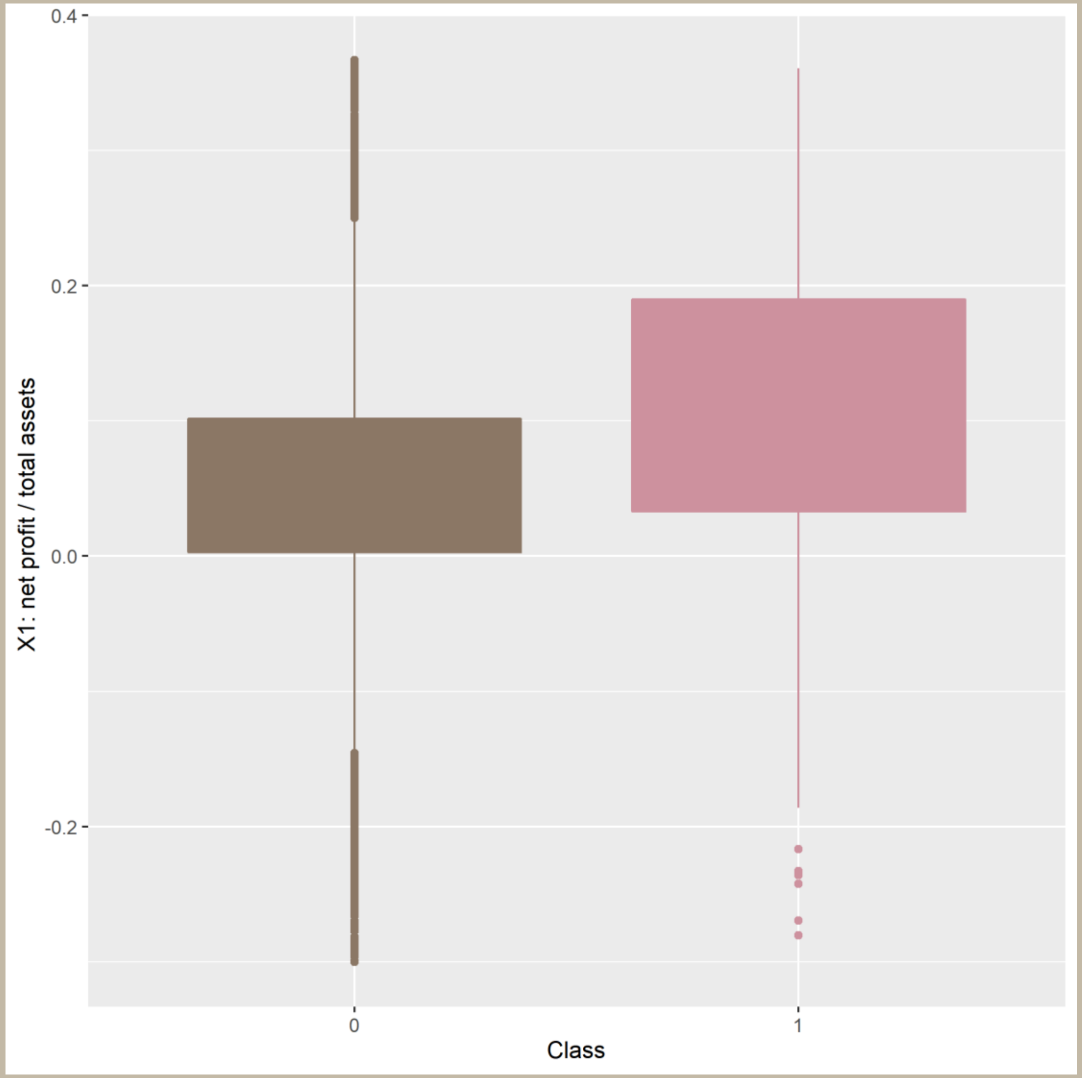
Attr2, Attr3, Attr10, Attr25, Attr38, Attr51	PCA1
Attr1, Attr7, Attr11, Attr14, Attr22, Attr35, Attr48	PCA2
Attr19, Attr23, Attr30, Attr31, Attr39, Attr43, Attr44, Attr49, Attr56, Attr58, Attr62	PCA3



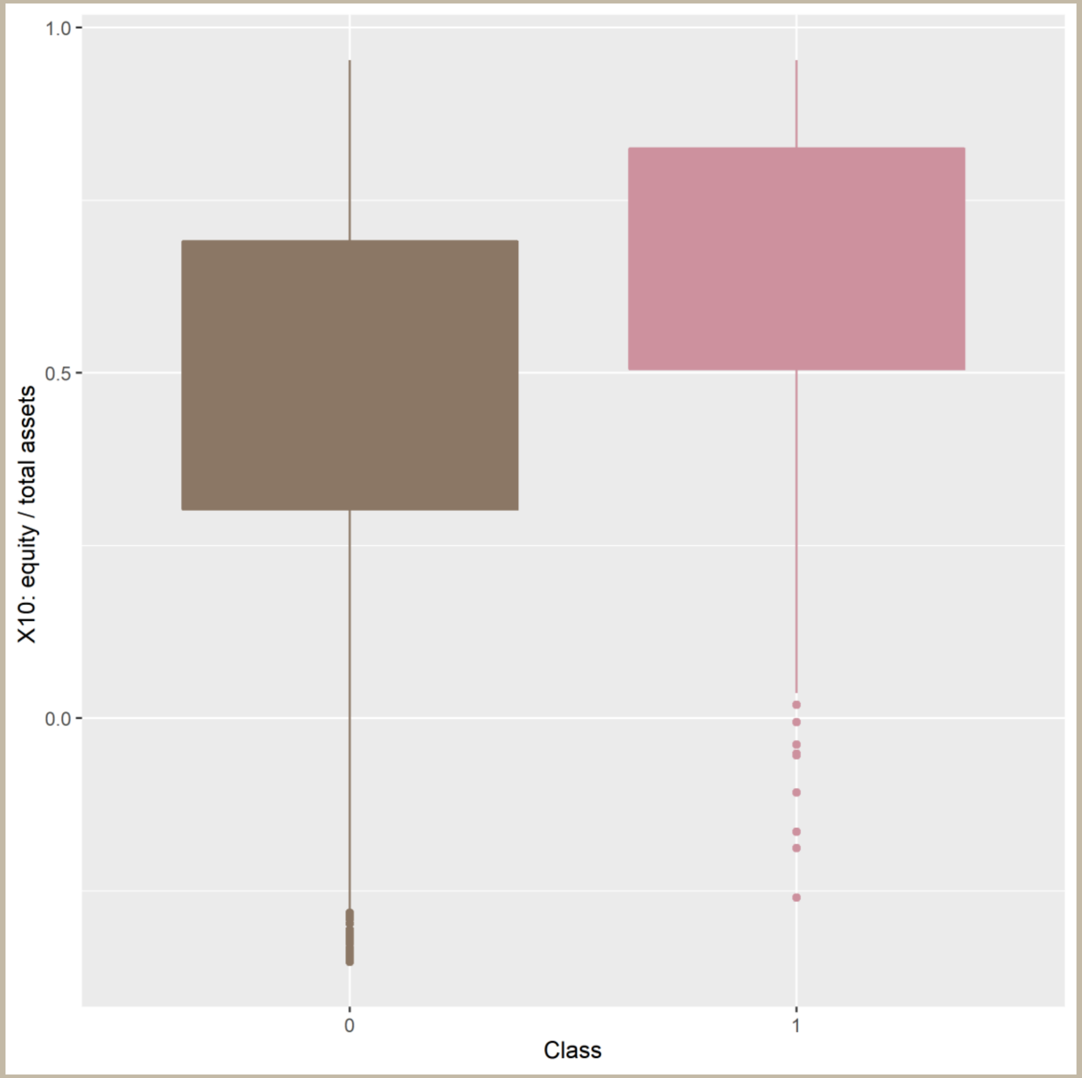
3

Visualization

시각화

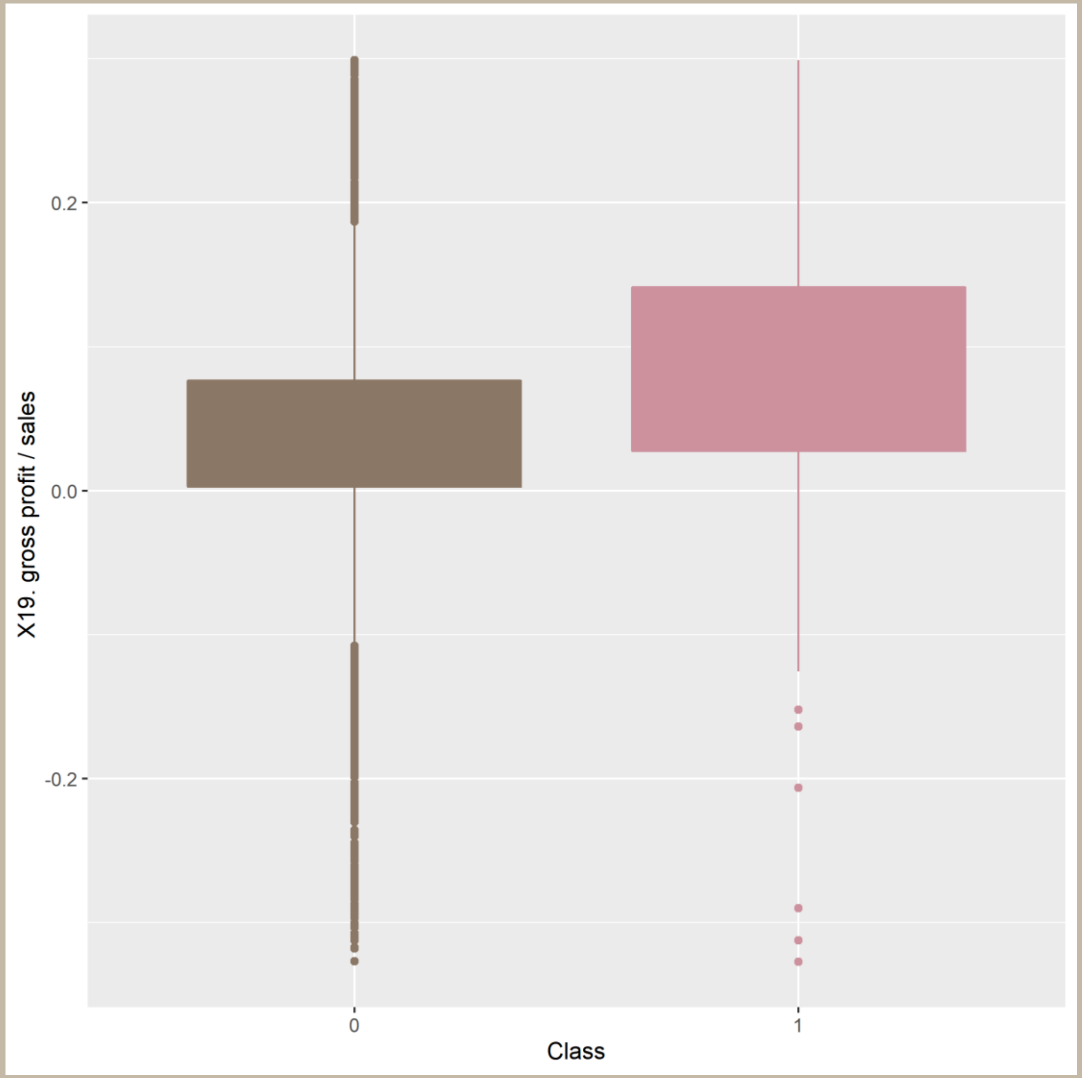


0.4
0.2
0.0
-0.2
X1: net profit / total assets
0
1
Class

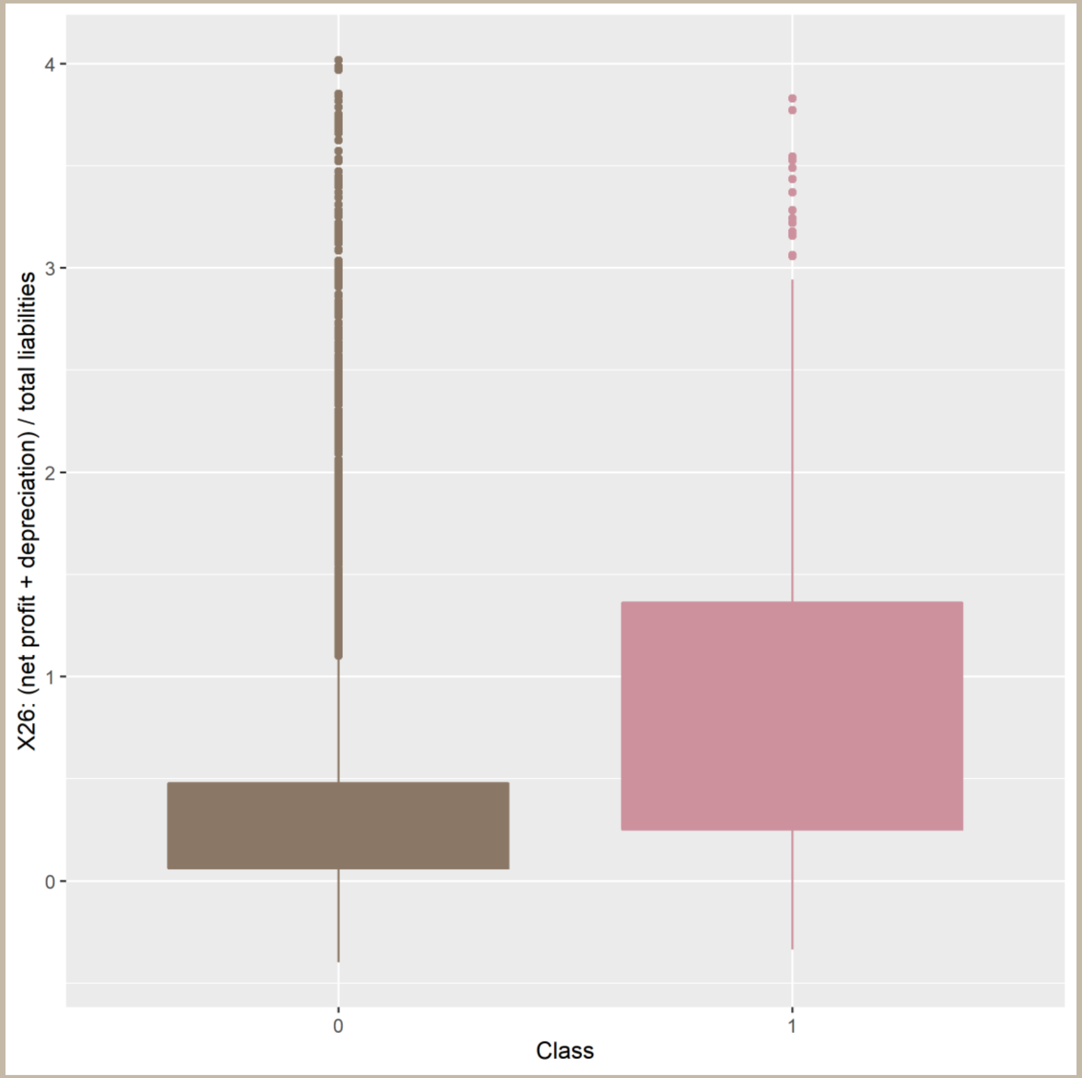


1.0
0.5
0.0
X10: equity / total assets
0
1
Class

시각화

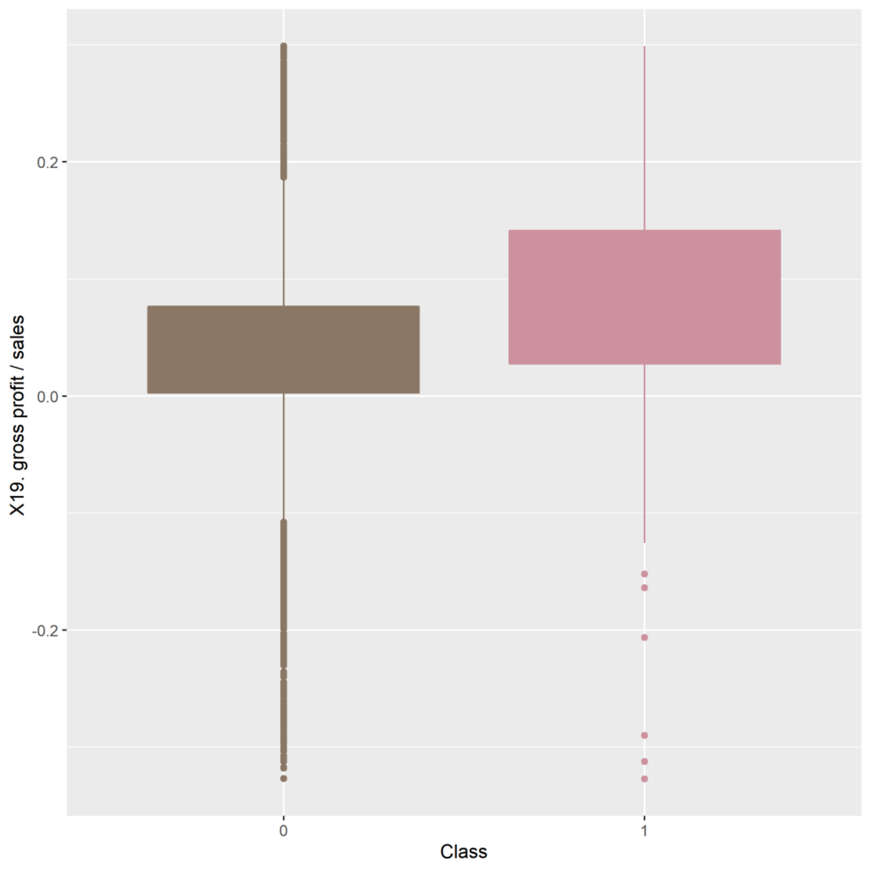


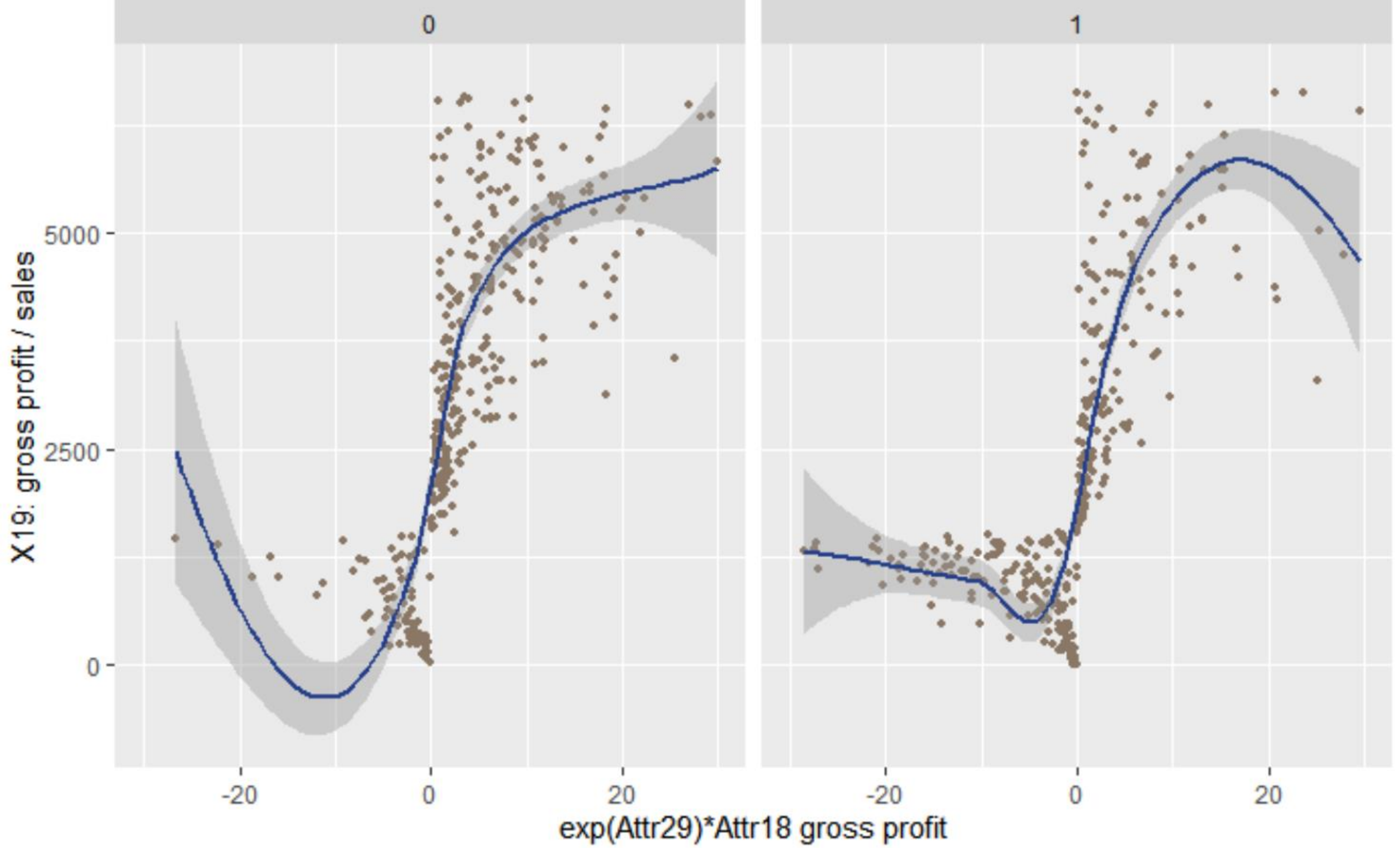
0.2
0.0
-0.2
X19. gross profit / sales
0
1
Class



4
3
2
1
0
X26: (net profit + depreciation) / total liabilities
0
1
Class

시각화

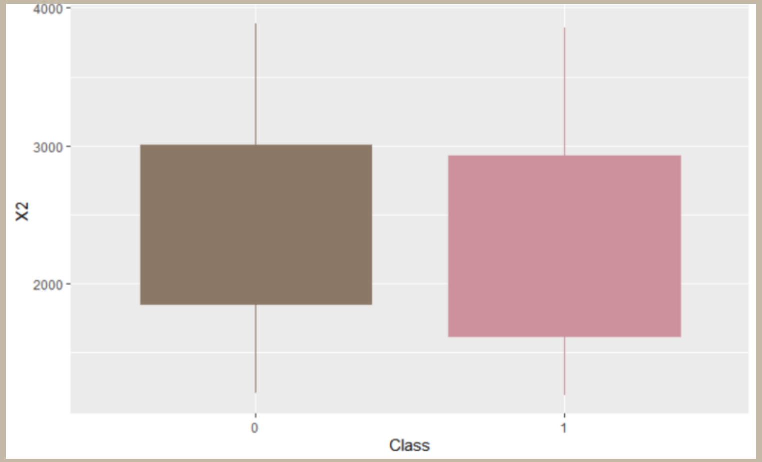




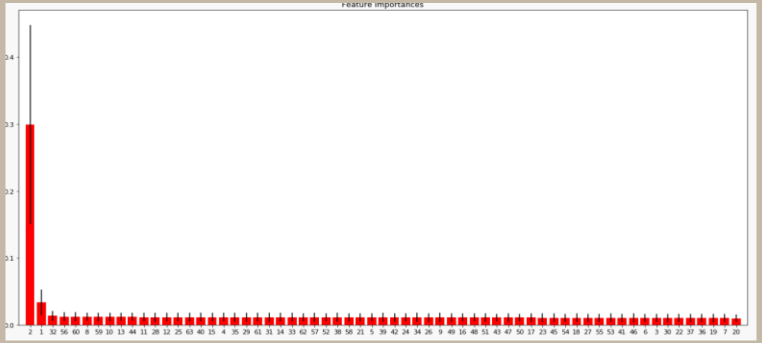
X19: gross profit / sales 변수
부도기업이 정상기업보다 높다?

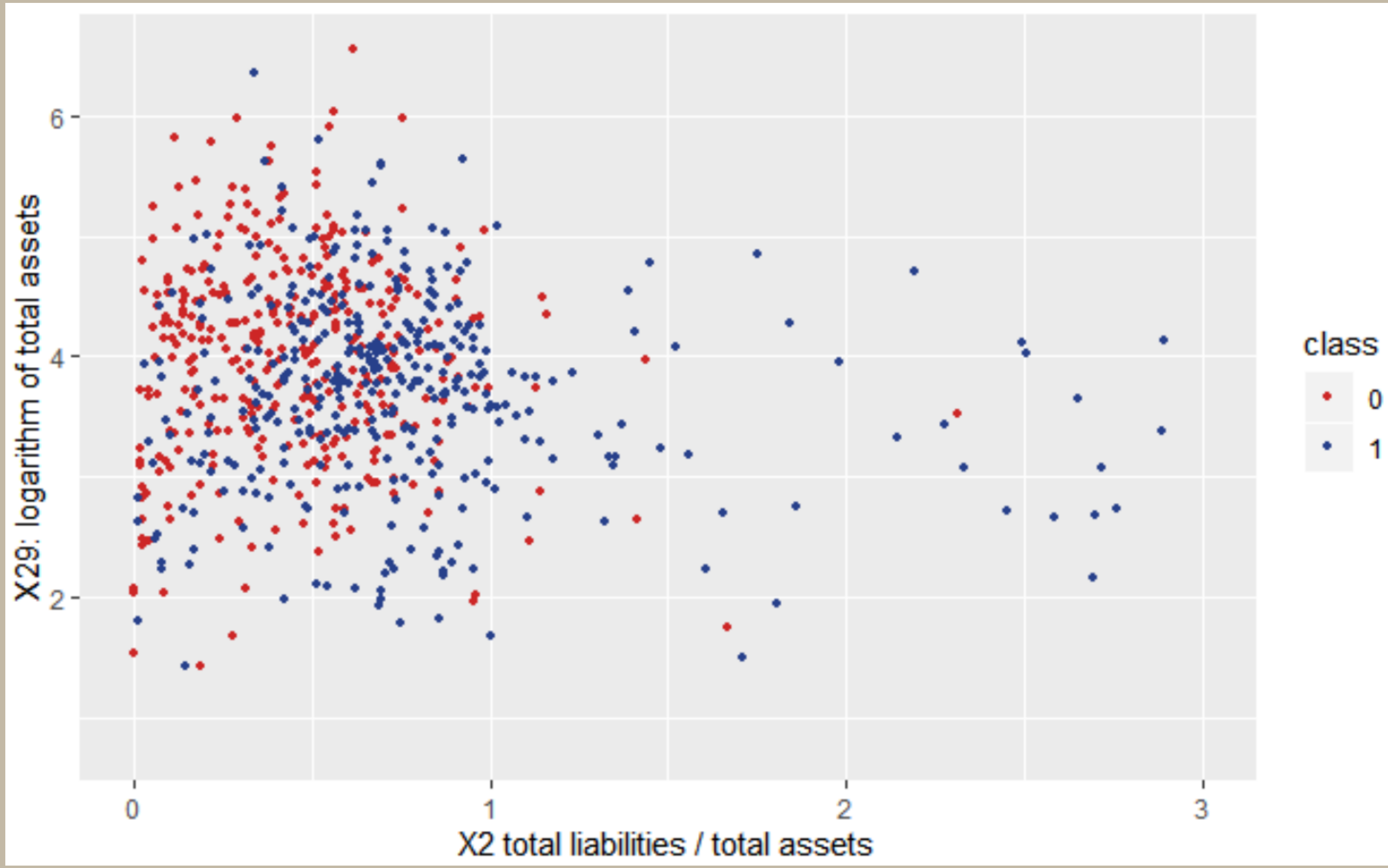
Gross profit 과 비교
- 0을 기준으로 변화가 커진 것이 보임
- 부도기업 gross fit이 – 값으로 더 많이 나타난다.

시각화



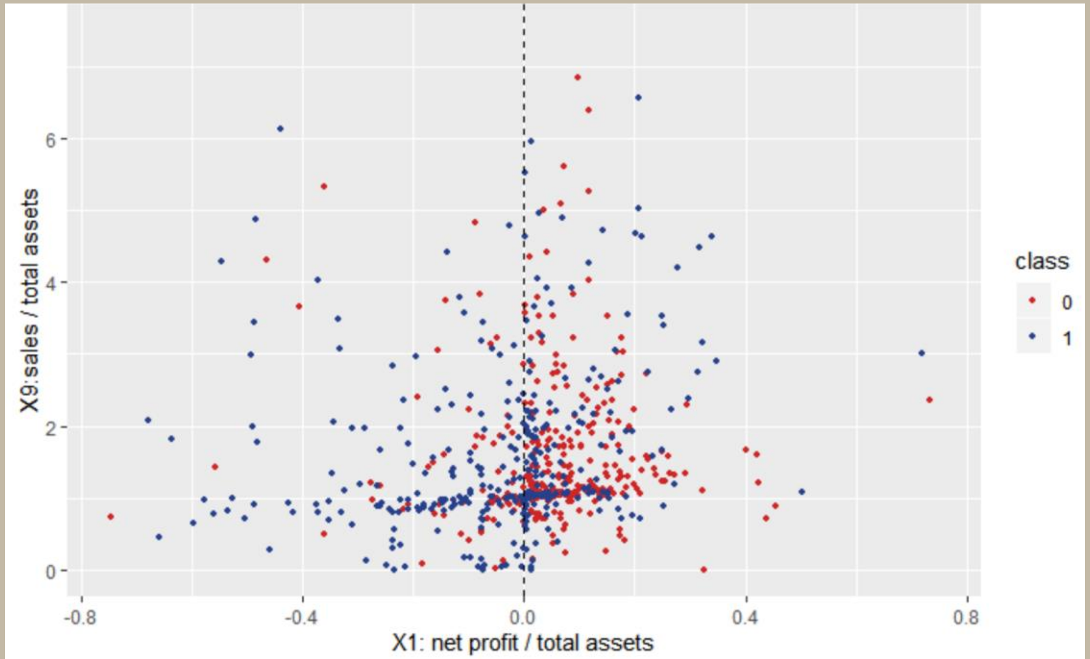
X2와 X19의 boxplot





X2와 X19의 scatter plot

시각화



X1: net profit / total assets
X9: sales / total assets

X1: net profit / total assets
X10: equity / total assets

4

Modeling

우리의 고민

1. Class가 불균형한 데이터를 어떻게 모델에 잘 반영할까?
2. 변수가 많고 변수의미가 어려워, 어떻게 모델 해석을 잘 할 수 있을까?
3. 다중공산성이 높은 변수들 제거 vs PCA/FA 축소
4. NA Imputation 이후 correlation이 증가하는 경우는 어떻게 처리할까?

앞으로의 계획

1. 모델 선택: Logistic Regression, SVM, Random Forest …

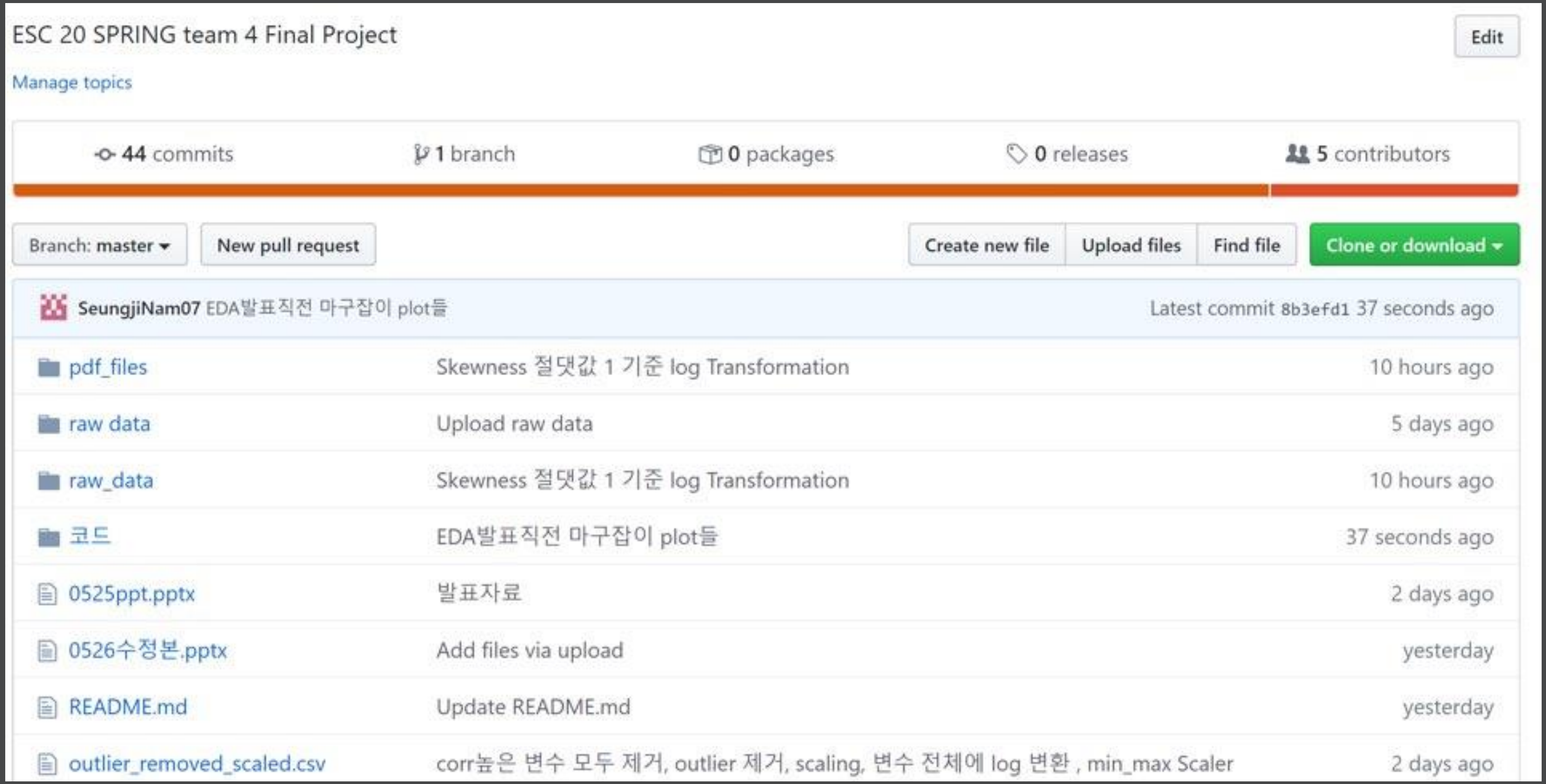
2. 변수 선택: EDA과정에서 고민했던 여러가지를 다 담아 내보자 …

   : 모델의 성능을 올려가면서, 변수를 선택해보자 …

- MICE, KNN, MEDIAN, outlier 뺀 MEAN의 4가지 방법으로 feature extraction
- 이후 skewed된 값은 변환하고, outlier 제거
- PCA, FA: correlation 높은 변수들 중심으로 변수 합치기
- 각 모델들에 대해 모델링 – F1 score, AUC ROC 비교

4조의 git

https://github.com/SeungjiNam07/ESC20SPRING_team4



참조

- https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data
- Naimputation: https://statisticsglobe.com/predictive-mean-matching-imputation-method/

감사합니다:)

THANK YOU