

YonseiESC 2020 SPRING

FINAL PROJECT

EDA

곽현지 서경덕 손동규 유은영 정규형

목차

도입

- 1) 데이터 설명
- 2) insight

변수 제거 및 선정






- 1) 변수의 의미
- 2) DFL
- 3) NA
- 4) Outlier
- 5) Visualization

향후 계획

- 1) Model Selection
- 2) K-CV

도입

1) 데이터 설명

자동 저장 <input checked="" type="radio"/> 컴     FS_2y_before_Bankruptcy_train 									
파일 홈 삽입 그리기 페이지 레이아웃 수식 데이터 검토 보기 도움말									
A1	×	✓	<i>fx</i>	Attr1					
	A	B	C	D	E	F	G	H	
1	Attr1	Attr2	Attr3	Attr4	Attr5	Attr6	Attr7	Attr8	,
2	-0.02456	0.30051	0.10733	1.8368	34.382	-0.02671	-0.02007	2.2519	
3	0.63918	0	1 ?		29.678	-1.1393	0.76052	?	
4	-0.44076	0.47723	0.1926	1.4678	34.555	0	-0.44076	1.0954	
5	0.11952	0.17679	0.63546	4.5944	117.65	0.25154	0.14875	4.5826	
6	0.2304	0.27095	0.23762	2.5745	-26.928	0.61754	0.28269	2.6907	
7	0.15255	0.47011	0.018859	1.0406	-39.998	0	0.15255	1.1272	
8	0.036332	0.25392	0.42075	3.0936	52.16	-0.01365	0.037068	2.9382	
9	0.11767	0.37332	0.26743	2.3229	18.308	0.14871	0.14571	1.6787	
10	0.034426	0.84574	0.00035	1.0006	-64.815	-0.00648	0.043455	0.18039	
11	-0.13918	0.99555	-0.01232	0.98712	-98.196	-0.14454	-0.13918	0.004474	
12	0.003006	0.82265	-0.04681	0.88192	-114.66	-0.0289	0.003006	0.17533	
13	0.056009	0.48351	0.047054	1.1005	-30.895	0.056009	0.056009	1.0682	

재무상태표와 2년 후 파산 여부 2 years before Bankruptcy

- 6855 rows × 65 columns. (Train data)
- Target 'class' (파산 기업 1, 정상 기업 0)
- 6855개 기업 중 361개 부도기업, 6494개 정상기업. 약 1: 18의 비율 고려
- '?'의 결측치가 있는 데이터

도입

2) insight: 여신 여부를 결정하기 위해 어떤 변수를 중점으로 봐야 하는가?

FS_2y_before_Bankruptcy_names - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

X1 net profit / total assets

X2 total liabilities / total assets

X3 working capital / total assets

X4 current assets / short-term liabilities

X5 [(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365

X6 retained earnings / total assets

X7 EBIT / total assets

X8 book value of equity / total liabilities

X9 sales / total assets

X10 equity / total assets

X11 (gross profit + extraordinary items + financial expenses) / total assets

X12 gross profit / short-term liabilities

X13 (gross profit + depreciation) / sales

X14 (gross profit + interest) / total assets

X15 (total liabilities * 365) / (gross profit + depreciation)

X16 (gross profit + depreciation) / total liabilities

X17 total assets / total liabilities

X18 gross profit / total assets

X19 gross profit / sales

X20 (inventory * 365) / sales

➔ 실제로 기업신용평가모형에 사용하는 재무비율들!

도입

2) insight: 실제로 사용되는 재무비율

수익성

기업의 경영성과를 측정하는 비율. 자산 이용의 효율성, 이익창출능력 등에 대한 평가지표로 활용

안전성

자산 & 자본의 관계 비율로, 단기채무지불능력인 유동성과 경기대응 능력인 안전성을 측정하는 지표

활동성

자산 & 자본의 회전율로, 기업에 투하된 자본이 얼마나 활발하게 운용되었는가를 나타내는 비율

예시)

총자산 순이익률(Net income to total assets)

당기순이익의 총자산에 대한 비율. ROA(Return on Assets)
기업의 계획과 실적간 차이 분석을 통한 경영활동 평가에 활용

$$\blacksquare \text{ 총자산 순이익률} = \frac{\text{당기순이익}}{\text{총자산(평균)}} \times 100$$

$$= X1 \text{ net profit} / \text{total assets}$$

당좌비율(Quick ratio)

유동자산 중 현금화 속도가 늦거나 현금화의 불확실성이 큰 재고자산 등을 제외한 당좌자산을 유동부채에 대응시킴으로써 단기채무에 대한 기업의 지급능력을 파악하는데 이용된다.

$$\blacksquare \text{ 당좌비율} = \frac{\text{당좌자산}}{\text{유동부채}} \times 100$$

$$= X46 \text{ (current assets - inventory)} / \text{short-term liabilities}$$

총자산회전율(Total assets turnover)

총자산이 1년 동안 몇 번 회전하였는가를 나타내는 비율.
기업에 투하한 총자산의 운용효율을 총괄적으로 표시하는 지표

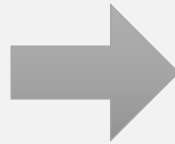
$$\blacksquare \text{ 총자산회전율} = \frac{\text{매출액}}{\text{총자산(평균)}}$$

$$= X9 \text{ sales} / \text{total assets}$$

도입

2) insight: 새로운 변수를 만들 수 있는가?

2	depreciation	감가상각
3	constant capital	고정자본
4	fixed assets	고정자산
5	financial expenses	금융비용
6	short-term securities	단기증권
7	sales	매출액
8	cost of products sold	매출원가
9	profit on sales	매출이익
10	gross profit	매출총이익
11	gross profit (in 3 years)	매출총이익(3년이내)
12	retained earnings	유보이익 / 이익잉여금
13	receivables	수취채권
14	net profit	순이익
15	operating expenses	영업비용
16	profit on operating activities	영업이익



$$\text{Sales} = X9 * \exp(X29)$$

$$\text{EBIT} = X7 * \exp(X29)$$

EBIT 대 매출액(Earnings before interest, tax to sales)

이자 및 세전순이익은 기업의 경영성과를 평가함에 있어 타인 자금을 사용하는 기업이 금융비용발생으로 인해 타인자금을 사용하지 않는 기업보다 수익성이 낮게 평가되는 것을 방지하기 위해 도입된 개념이다.

$$\blacksquare \text{ EBIT 대 매출액} = \frac{\text{세전순이익} + \text{이자비용}}{\text{매출액}} \times 100$$

새로운 변수를 만들거나, 숨어있는 변수를 찾아낼 수 있음

변수 제거 및 선정

1) 변수의 의미

1. 자체로는 의미가 없는 변수 제거 X29: logarithm of total asset, X58: total costs /total sales

2. 의미상 유사한 변수 제거(상관관계 높을 것으로 예측)

1) 역수

X2 & X17 , X15 & X16, X20 & X60, X32 & X52,
X44 & X61 , X62 & X63

X2 total liabilities / total assets

X17 total assets / total liabilities

X62 (short-term liabilities *365) / sales

X63 sales / short-term liabilities

2) 계산을 통해 만들 수 있는 변수

X43(= X44 + (1/X47)), X7 & X48, X39 & X42,
X40 & X46

X43 rotation receivables+inventory turnover in days↵

X44 (receivables*365)/sales: 매출채권회전일수(rotation receivables)↵

X47 (inventory*365)/cost of products sold: 재고자산회전율(inventory turnover in days)의 역수↵

변수 제거 및 선정

2) DFL

DFL?

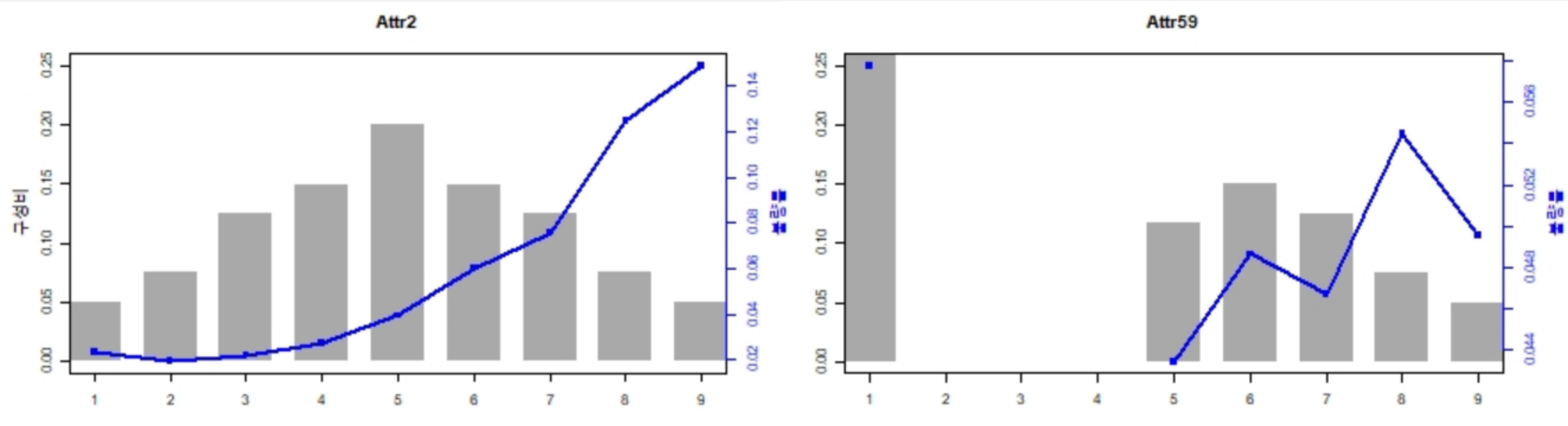


DFL Default Frequency Line 

: 부도율 그래프를 이용한 변수 제거 방법.
변수(재무비율) 별로 9개 구간으로 나눈다.
각 차주의 구성비는 **종 모양**(평균 부분에 가장 많고, 각 구간 끝 부분에 적은 모양)이고, 각 구간 별 부도율을 산출한 후, 부도율(꺾은선 그래프)은 **단조 증가, 단조 감소**인 변수가 좋은 변수.

변수 제거 및 선정

2) DFL



변수별로 DFL을 그려 설명력이 낮은 변수를 drop.
예시에서는 좌측이 좋은 변수.

→ X8, X24, X32, X33, X34, X41, X45, X53, X59, X64 변수를 제거함

변수 제거 및 선정

3) NA

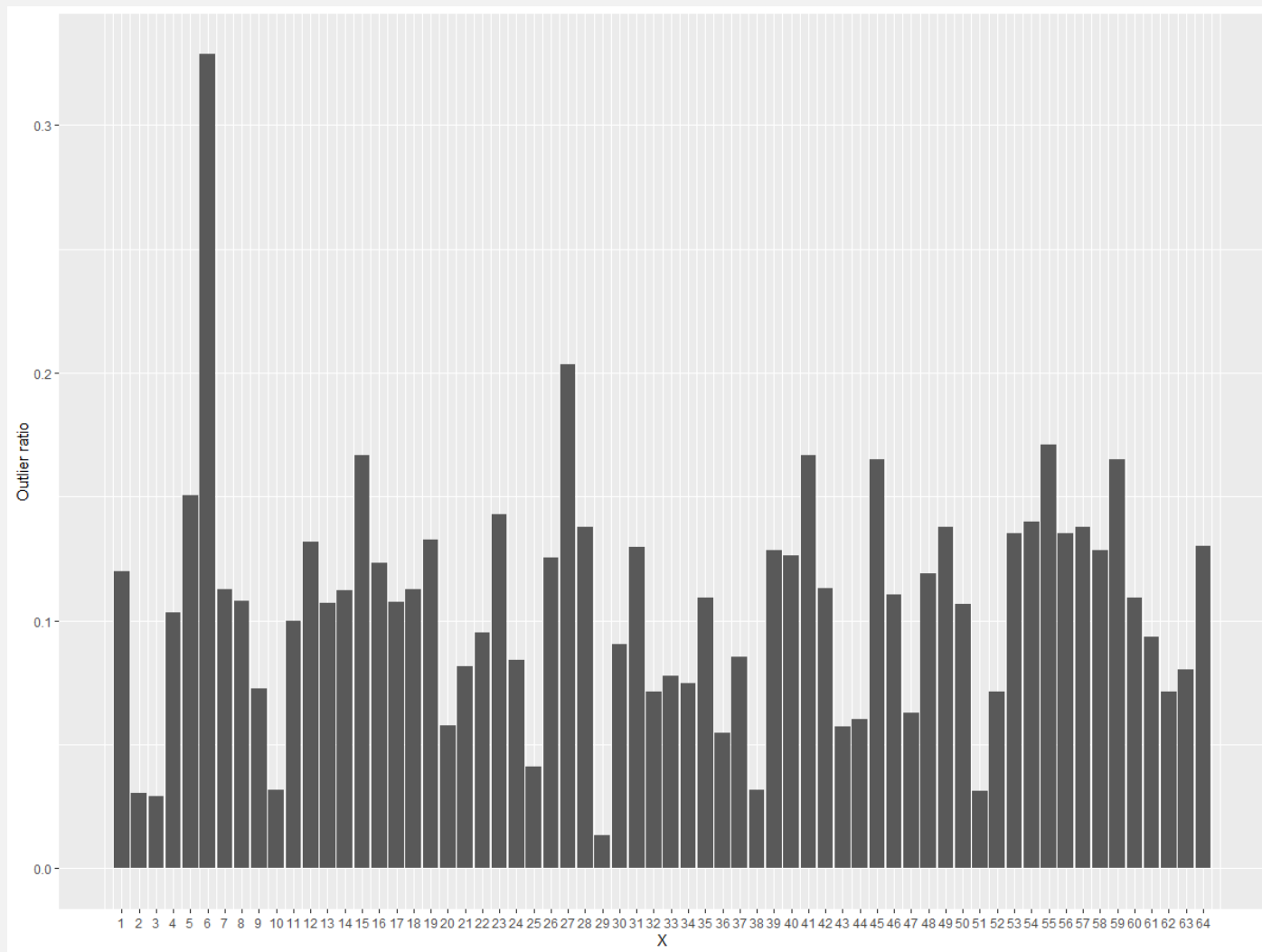
앞의 과정을 통해 변수를 제거하니 NA 있는 행 720개, 이중 부도기업 191개.

NA가 많은 변수 중 다른 변수와의 상관관계가 높은 변수를 제거

→X21, X27, X28, X37, X54 제거

변수 제거 및 선정

4) Outlier



Outlier가 변수별로 약 10%-20%라서
빼지 않기로 판단

변수 제거 및 선정

5) Visualization



최종 선정 : X1, X2, X4, X5, X6, X13, X15, X20, X26, X31, X36, X39, X47, X55, X57

향후 계획

1. EDA Feedback

2. Model selection

- 1) 선정 변수에서 NA 제외한 데이터 -> 설명력을 위해서 변수 자체를 선정
- 2) 전체 data에서 Iterative Imputation으로 NA 채운 완전한 데이터 -> FA analysis
두 자료를 통해 분석 tool 선정

3. K-CV

K-CV로 비교하고, 무엇으로 분석을 할 것인가를 논의를 통해서 선정
선정된 분석 방법론을 통해서 score값 비교
→ 최적의 방법론 선정 + 변수들과 부도의 상관관계 규명 노력

참고

역할배분٩(๑•̀๑•́๑)و

곽현지 변수 정보 정리, 변수 제거, EDA 발표
서경덕 data visualization. test set 위한 coding
손동규(조장) 전체 참여, 방향 제시
유은영 변수 정보 정리, 변수 제거, PPT 제작
정규형 data visualization. test set 위한 coding



Q&A