

Trillion (1조)

UBION Bankruptcy Firm Report

재무비율을 이용한 상장기업 부도 예측 알고리즘 비교 연구

정기호 이주희 신문혁 윤영주

Trillion (1조)

Contents Title

01. 주제 설정 배경

04. Feature Selecting

02. 데이터 출처 및 부도 기업의 정의

05. 모델링 시행

03. 모집단 선정

06. 모델 정확도 검정 및 최종 모델 선정

Distributing Roles



UBION – Trillion

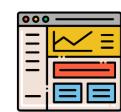
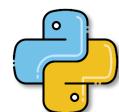


Time Table



UBION – Trillion

#	Task	To do list	Effort	2021.11.15 ~ 2021.12.22					
				1주차	2주차	3주차	4주차	5주차	6주차
1	논문	논문 리뷰 및 발표	1 w	<div style="width: 20px; background-color: #007bff;"></div>					
2	Data	데이터 수집 및 전처리	2 w		<div style="width: 40px; background-color: #007bff;"></div>	<div style="width: 40px; background-color: #007bff;"></div>			
3	Model	모델링 시행 및 평가	3 w			<div style="width: 60px; background-color: #007bff;"></div>			
4	분석	결과 분석 및 정리	1 w				<div style="width: 20px; background-color: #007bff;"></div>		
5	PPT	PPT 제작 및 발표 준비	2 w					<div style="width: 40px; background-color: #007bff;"></div>	



순서도



UBION – Trillion

step 01

부도 정의 및 모집단 선정

외부 감사 의견 코드 (DS, DU, DI)

OR

상장 폐지 (04)

관리 종목 지정 (03, 06)

부도 정의에 맞는 모집단 선정

03 : 제조업

10 : 정보통신업

07 : 운송업

step 02

Feature Selection

변수 유의성 검정

Bartlett – test (등분산)

Z-test, Welch's T-test

Wrapper Feature selection

Wrapper Feature selection + Filter

Embedded Feature Selection

step 03

Sampling & Modeling

Sampling

Under Sampling

Over Sampling

Modeling

Logistic Regression, SGD

KNN, SVM, Decision Tree

Random Forest, Purning

step 04

Result + Insight

Intermediate result

Validation

- Recall + Accuracy
- ROC, AUC

Top 1% : Hyperparameter

Final Result

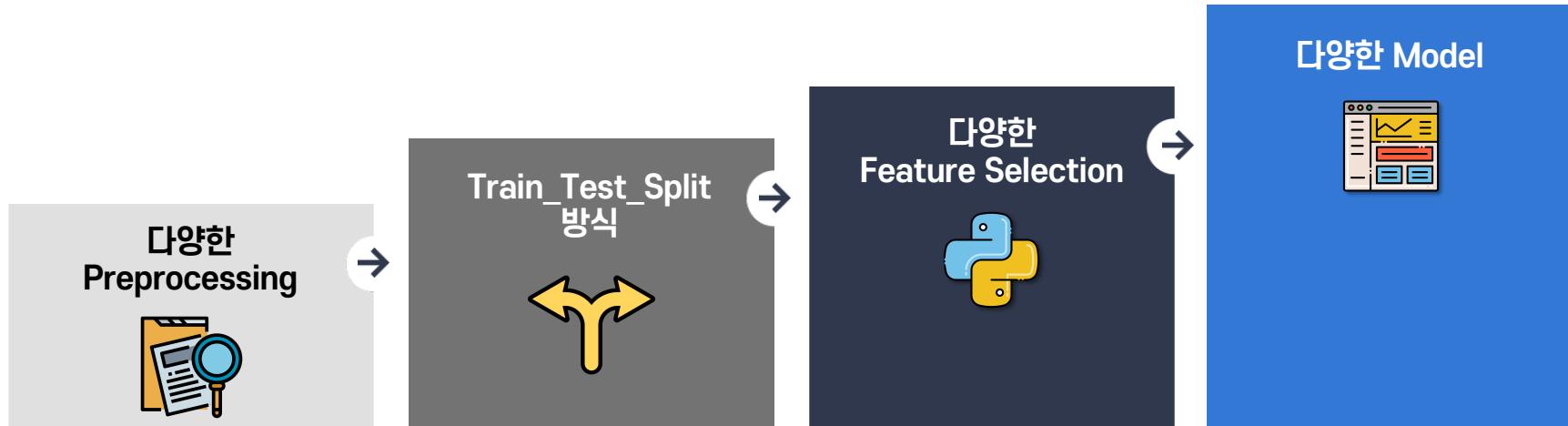
Insight + How to use

1. 주제 선정 배경 및 활용 방안



04. 연구의 목적

UBION – Trillion



다양한 경우의 수로 모델링하여 최적의 조합을 도출한다.
도출된 결과를 비교하여 부동 예측에 중요한 Feature에 대해 고찰한다.

01. 주제 선정 배경 및 활용 방안

- 01. 주제선정배경
- 02. 활용방안
- 03. 선행연구

1. 주제 선정 배경 및 활용 방안

01. 주제 선정 배경



UBION – Trillion



국내은행 리스크관리 강화 필요

이 병 윤 (선임연구위원, 3705-6343)

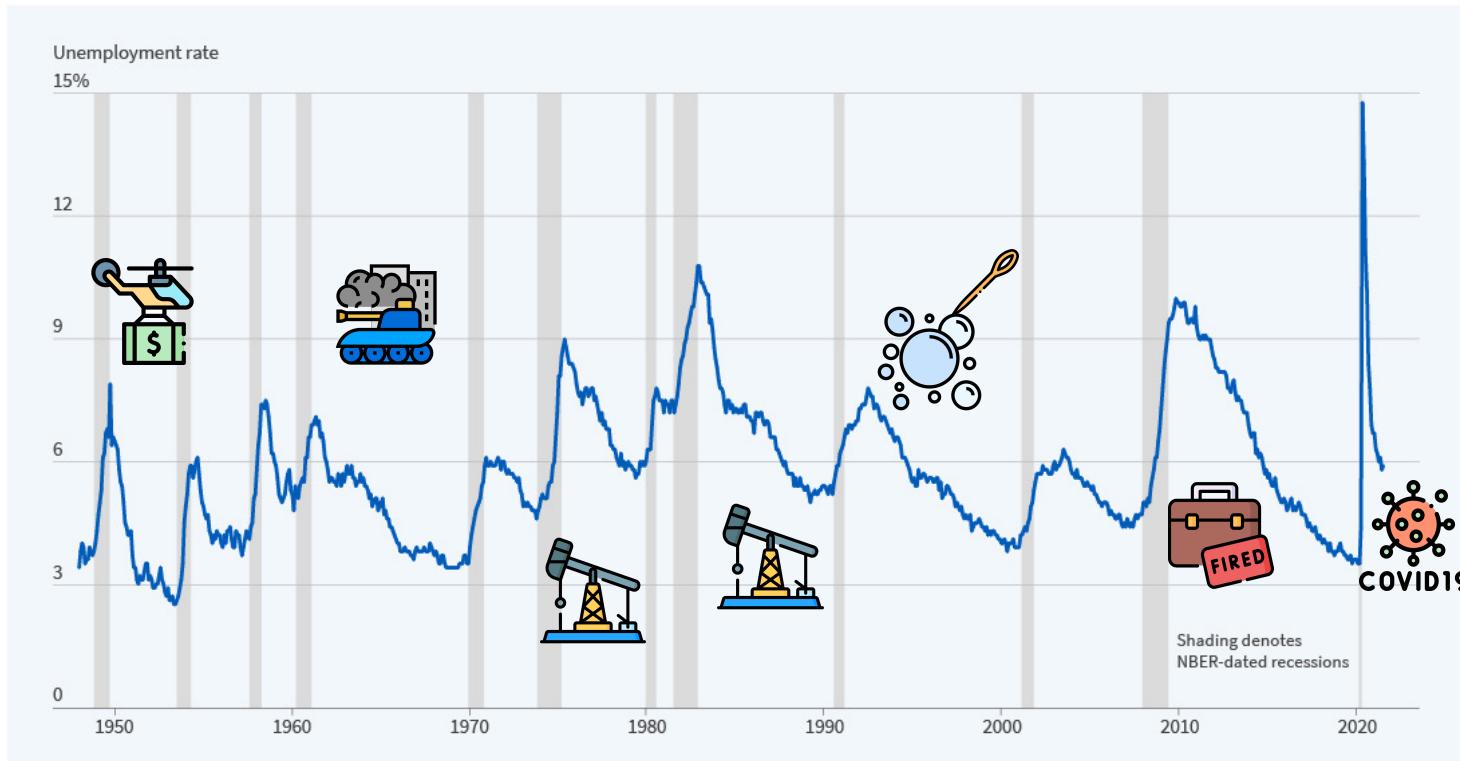
최근 국내은행은 수익성 및 건전성이 모두 크게 개선되는 등 좋은 모습을 보이고 있으나 잠재 리스크가 큰 것으로 보여 주의가 필요함. 국내은행의 대출은 증가세가 매우 높은데 실물경제와는 괴리가 있어 작은 시장 충격에도 부실화할 가능성이 있음. 또 수치상 건전성은 좋은 편이나 중소기업·소상공인에 대한 대출만기 연장 및 원리금 상환유예 프로그램이 진행 중이고 이자보상비율 100% 미만 한계기업의 비중이 높아 유의할 필요가 있음. 따라서 은행과 감독당국은 은행자산의 잠재 부실 규모를 추정하고 경제상황 변화에 따른 부실화 가능성 등에 대한 스트레스테스트를 시행하여 향후 나타날 수 있는 리스크에 미리 대비할 필요가 있음.

1. 주제 선정 배경 및 활용 방안



01. 주제 선정 배경

UBION – Trillion



1. 주제 선정 배경 및 활용 방안



02. 활용 방안

UBION – Trillion

빅데이터를 활용한 기업 여신 자동 심사 시스템

다년간 축적된 B2B여신 데이터를 활용한 기업 여신 자동 심사



1. 주제 선정 배경 및 활용 방안

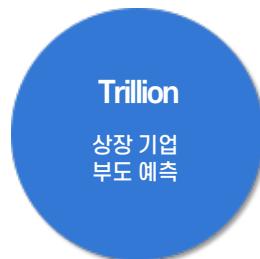
03. 선행 연구



UBION – Trillion

639 경영학연구 제43권 제3호 2014년 6월(pp. 639~669)

통계적 논리



국내 적합성



재무비율을 이용한 부도예측에 대한 연구:
한국의 외부감사대상기업을 대상으로*

박종원 (교신저자)
서울시립대학교 경영대학 교수
(parkjw@seisu.ac.kr)
안성만 (주저자)
농협은행 어신정책부
(asm1230@nonghyup.com)

The Journal of FINANCE

VOL. XXIII

SEPTEMBER 1968

No. 4

FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND
THE PREDICTION OF CORPORATE BANKRUPTCY

EDWARD I. ALTMAN*

1. 주제 선정 배경 및 활용 방안

03. 선행 연구



UBION – Trillion

The Journal of FINANCE

VOL. XXIII

SEPTEMBER 1968

No. 4

FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND
THE PREDICTION OF CORPORATE BANKRUPTCY

EDWARD I. ALTMAN*



X1 : 순운전 총자산비율

TS2000

순운전자본비율

$\frac{\text{유동자산} - \text{유동부채}}{\text{총자산}}$



X2 : 유보액 총자산비율

TS2000

유보액대비율

$\frac{\text{유보액}}{\text{총자산}}$



X3 : Cash Flow 총자산비율 (안정성)

TS2000

CASH FLOW 대 총자본비율

$\frac{(\text{전기 CF} + \text{당기 CF})/2}{\text{총자산}}$



X5 : 총자산 회전률

TS2000

총자본회전률 (TS2000)

$\frac{\text{매출액}}{(\text{전기 총자산} + \text{당기 총자산})/2}$

$$(I) \quad Z = .012X_1 + .014X_2 + .033X_3 + .006X_4 + .999X_5$$

where X_1 = Working capital/Total assets

X_2 = Retained Earnings/Total assets

X_3 = Earnings before interest and taxes/Total assets

~~X_4 = Market value equity/Book value of total debt~~

X_5 = Sales/Total assets

Z = Overall Index



02. 부도 기업 정의

01.부도기업정의

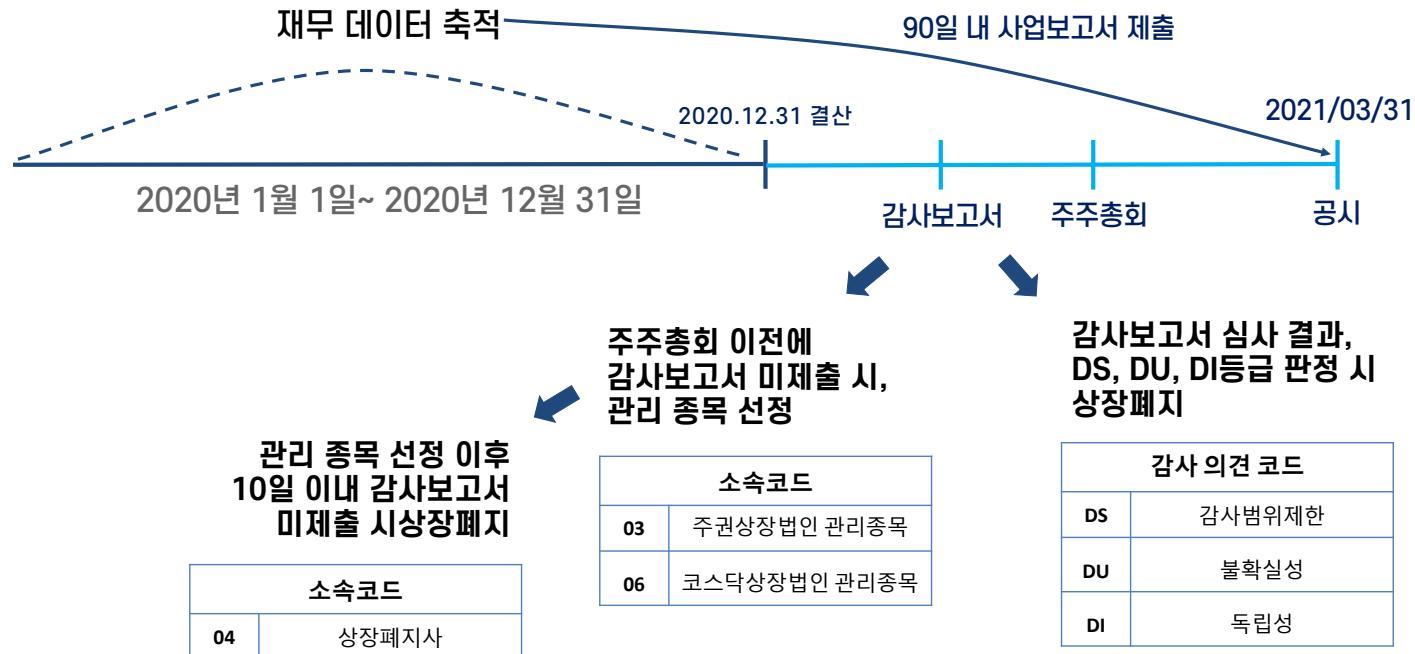
02.데이터출처

2. 부도 기업 정의

01. 부도 기업 정의



UBION – Trillion



2. 부도 기업 정의



01. 부도 기업 정의

UBION – Trillion

데이터 처리



방법 1	정상	정상	부도	정상	부도
------	----	----	----	----	----

방법 2	정상	정상	부도	X	X
------	----	----	----	---	---

방법 3	정상	정상	부도	부도	부도
------	----	----	----	----	----

2. 부도 기업 정의

02. 데이터 출처



UBION – Trillion

The screenshot displays the KOCOinfo website's homepage with three main service modules:

- KOCOinfo Service 01 — TS2000**
Total Solution 2000
기업의 각종 정보를 사용자 선택에 따라 다양한 형태의 분석자료로 추출 및 다운로드 할 수 있습니다.
[바로 실행](#) [자세히 보기](#)
- KOCOinfo Service 02 — COINS**
Corporate Information Solution
기업이 공시하는 보고서 내용을 기초로 기업내용 전반에 걸친 각종 관련 자료를 상세하게 파악할 수 있습니다.
[바로 실행](#) [자세히 보기](#)
- 시범 서비스 — KOCOinfo Service 03 —**
공시보고서 색인 검색
공시보고서의 목차 카테고리별 자동 생성한 색인어 검색으로 쉽고 빠르게 원문 내용을 파악할 수 있습니다.
[바로 실행](#) [자세히 보기](#)

2. 부도 기업 정의

02. 데이터 출처



UBION – Trillion

조건실행

회사(2,760사) 항목(171개)

선택 조건

항목	설명
선택 조건	선택 조건
회사명	회사명
거래소 코드	거래소 코드
회계년도	회계년도
산업 코드	산업 코드
소속 코드	소속 코드
상장 일	상장 일
상장 폐지 일	상장 폐지 일
감사 의견 코드	감사 의견 코드
총자본 증가율	총자본 증가율
유형자산 증가율	유형자산 증가율
유동자산 증가율	유동자산 증가율
영업이익 증가율	영업이익 증가율
경상이익 증가율 (2007년 미전 발생)	경상이익 증가율 (2007년 미전 발생)
순이익 증가율	순이익 증가율
재고자산 증가율	재고자산 증가율
자기자본 증가율	자기자본 증가율
매출액 증가율	매출액 증가율
종업원 1인당 부가가치 증가율	종업원 1인당 부가가치 증가율
종업원 수 증가율	종업원 수 증가율
이상은 유니버설 표기입니다.	이상은 유니버설 표기입니다.

자료 기간

당기(보고서제출 최근 1년) 자료 기간 2007년 01월 ~ 2020년 12월

상장일 기준

상장일 이전 자료 포함 불포함

추가옵션 >

조건 복원 조건 저장 Submit

회사 선택 조건:

2007년 01월 ~ 2020년 12월 기간 내의 상장기업 대상

자료 선택 조건:

회사명, 거래소 코드, 회계년도, 산업코드, 소속코드, 상장일, 상장폐지일, 감사의견코드

+

166개 재무비율, 정보 산출

2. 부도 기업 정의

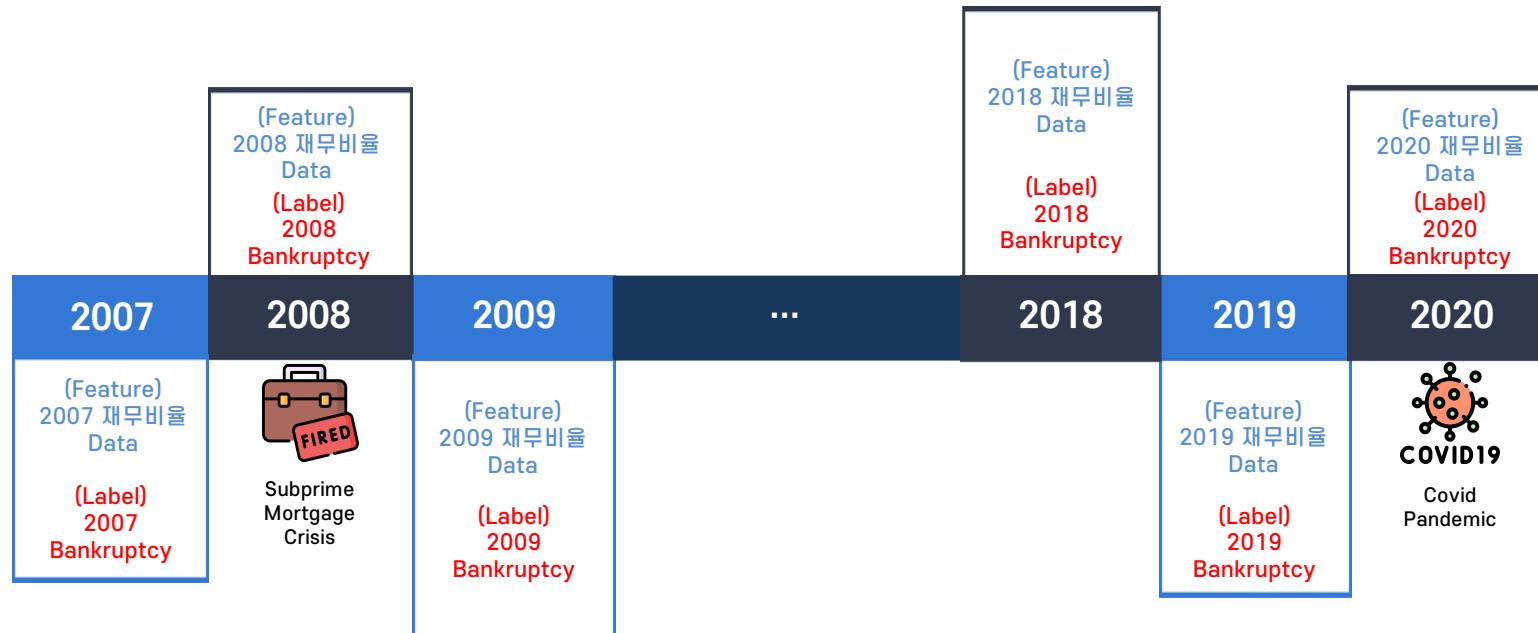


02. 데이터 출처

UBION – Trillion

● 회사 선택 조건:

2007년 01월 ~ 2020년 12월 기간 내의 상장기업 대상



2. 부도 기업 정의



02. 데이터 출처

UBION – Trillion

● 회사 선택 조건:

2007년 01월 ~ 2020년 12월 기간 내의 상장기업 대상

2007	2008	2009	...	2018	2019	2020	2021
(Feature) 2007 재무비율 Data (Label) 2006 Bankruptcy	(Feature) 2008 재무비율 Data (Label) 2007 Bankruptcy	(Feature) 2009 재무비율 Data (Label) 2008 Bankruptcy	...	(Feature) 2018 재무비율 Data (Label) 2017 Bankruptcy	(Feature) 2019 재무비율 Data (Label) 2018 Bankruptcy	(Feature) 2020 재무비율 Data (Label) 2019 Bankruptcy	(Label) 2020 Bankruptcy

2007: (Feature) 2007 재무비율 Data, (Label) 2006 Bankruptcy. Subprime Mortgage Crisis icon.

2008: (Feature) 2008 재무비율 Data, (Label) 2007 Bankruptcy. Subprime Mortgage Crisis icon.

2009: (Feature) 2009 재무비율 Data, (Label) 2008 Bankruptcy.

...

2018: (Feature) 2018 재무비율 Data, (Label) 2017 Bankruptcy.

2019: (Feature) 2019 재무비율 Data, (Label) 2018 Bankruptcy. COVID19 icon.

2020: (Feature) 2020 재무비율 Data, (Label) 2019 Bankruptcy. COVID19 icon.

2021: (Label) 2020 Bankruptcy.

2. 부도 기업 정의



02. 데이터 출처

UBION – Trillion

● 회사 선택 조건:

2007년 01월 ~ 2020년 12월 기간 내의 상장기업 대상

2007	2008	2009	...	2018	2019	2020	2021
(Feature) 2007 재무비율 Data (Label) 2008 Bankruptcy	(Feature) 2008 재무비율 Data (Label) 2009 Bankruptcy	(Feature) 2009 재무비율 Data (Label) 2010 Bankruptcy	...	(Feature) 2018 재무비율 Data (Label) 2019 Bankruptcy	(Feature) 2019 재무비율 Data (Label) 2020 Bankruptcy	(Feature) 2020 재무비율 Data  COVID19 Covid Pandemic	

2. 부도 기업 정의



02. 데이터 출처

UBION – Trillion

● 회사 선택 조건:

2007년 01월 ~ 2020년 12월 기간 내의 상장기업 대상

2007	2008	2009	...	2018	2019
<p>(Feature) 2007 재무비율 Data</p> <p>(Label) 2008 Bankruptcy</p>	<p>(Feature) 2008 재무비율 Data</p> <p>(Label) 2009 Bankruptcy</p> <p></p> <p>Subprime Mortgage Crisis</p>	<p>(Feature) 2009 재무비율 Data</p> <p>(Label) 2010 Bankruptcy</p>		<p>(Feature) 2018 재무비율 Data</p> <p>(Label) 2019 Bankruptcy</p>	<p>(Feature) 2019 재무비율 Data</p> <p>(Label) 2020 Bankruptcy</p>

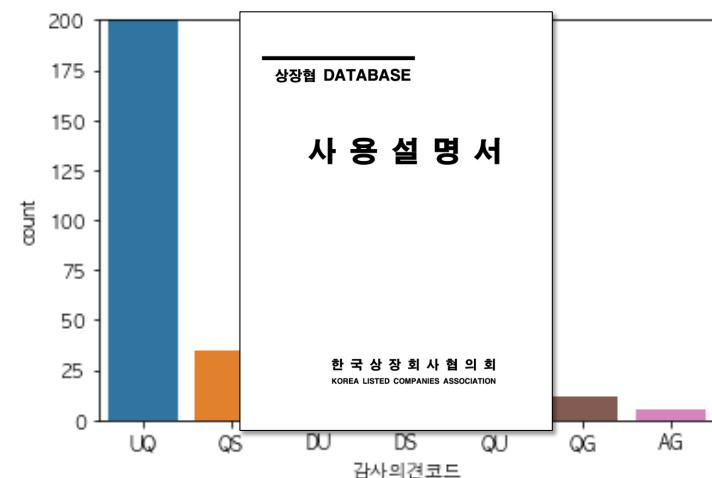
2. 부도 기업 정의

02. 데이터 출처



UBION – Trillion

감사 의견 코드					
감사의견	코드	내 용	감사의견	코드	내 용
적정의견	UQ		부적정의견	AG	GAAP 위반
한정의견	QQ QS QU QA QG QC	한정의견 감사범위제한 불확실성 계속기업전체 GAAP위반 계속성변경	의견거절	DS DU DI	감사범위제한 불확실성 독립성
				NS	보고서미제출



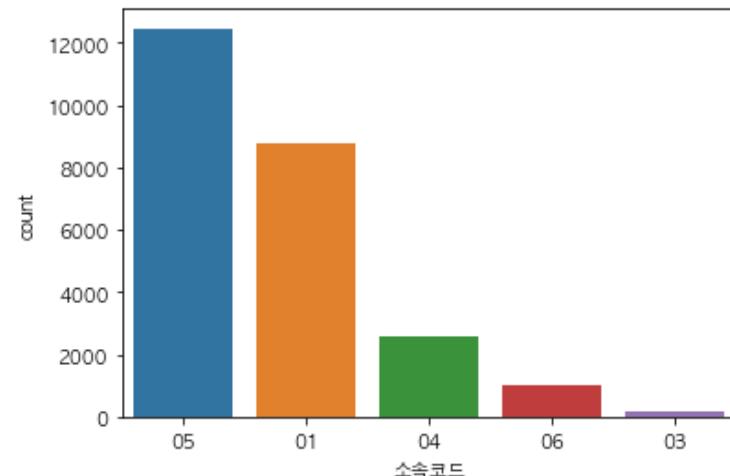
2. 부도 기업 정의

02. 데이터 출처



UBION – Trillion

소속 코드	
01	주권상장법인
03	주권상장법인 관리종목
04	상장폐지사
05	코스닥상장법인
06	코스닥상장법인 관리종목
07	외부감사대상법인
08	코넥스



03. 모집단 선정

- 01. 산업군선정
- 02. 전처리(Basic)시행
- 03. 전처리(Basic)시행 후 Data 나누기
- 04. Train&Test 분리

3. 모집단 선정

01. 산업군 선정



UBION – Trillion

산업 대분류	데이터 count	정상 count	부도 count
제조업 (03)	16079	15,907	172
출판, 영상, 방송통신 및 정보서비스업 (10)	2731	2,688	43
도매 및 소매업 (07)	2146	2,113	33
전문, 과학 및 기술 서비스업 (13)	1627	1620	7
건설업 (06)	854	839	15
운수업 (08)	371	370	1
사업시설 관리 및 사업지원 서비스업 (14)	255	248	7
전기, 가스, 증기 및 수도사업 (04)	166	166	0
교육 서비스업 (16)	145	143	2
예술, 스포츠 및 여가관련 서비스업 (18)	107	103	4

업종(산업분류)	중소기업	중견기업
제조업(C)	상시근로자수 300인 미만 또는 자본금 80억원 이하	상시근로자수 300인 이상 그리고 자본금 80억원 초과
광업(B) 건설업(F) 운수업(H)	상시근로자수 300인 미만 또는 자본금 30억원 이하	상시근로자수 300인 이상 그리고 자본금 30억원 초과
출판, 영상, 방송통신 및 정보서비스(J) 사업시설관리 및 사업지원서비스(N) 전문, 과학 및 기술 서비스업(M) 보건업 및 사회복지사업 (Q)	상시근로자수 300인 미만 또는 매출액 300억원 이하	상시근로자수 300인 이상 그리고 매출액 300억원 초과

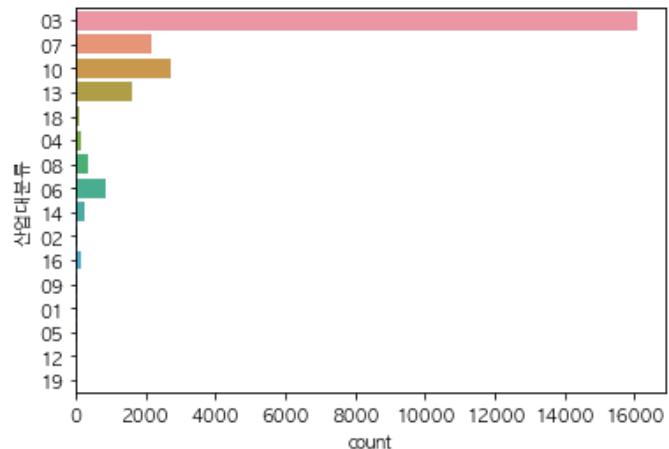
3. 모집단 선정



01. 산업군 선정

UBION – Trillion

산업 대분류	데이터 count	정상 count	부도 count
제조업 (03)	16079	15,907	172
출판, 영상, 방송통신 및 정보서비스업 (10)	2731	2,688	43
도매 및 소매업 (07)	2146	2,113	33
전문, 과학 및 기술 서비스업 (13)	1627	1620	7
건설업 (06)	854	839	15
운수업 (08)	371	370	1
사업시설 관리 및 사업지원 서비스업 (14)	255	248	7
전기, 가스, 증기 및 수도사업 (04)	166	166	0
교육 서비스업 (16)	145	143	2
예술, 스포츠 및 여가관련 서비스업 (18)	107	103	4



3. 모집단 선정

02. 전처리 (Basic) 시행

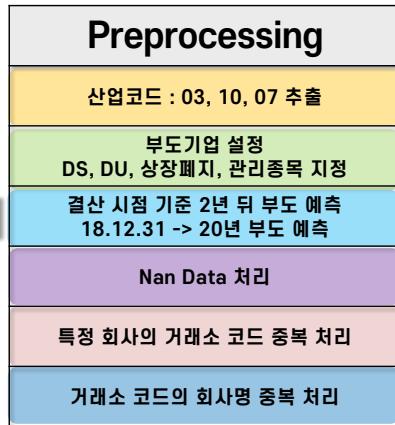


UBION – Trillion

Raw Data (TS2000)

회사명	거래소 코드	회계년도	산업코드	소속 코드	상장일	매지 경 운 전 고 드	당 시 의 건 강 고 드	총자본 증가율	유형자 산증가 율	PCR(Price cash-flow ratio)(# 기)	PCR(Price cash-flow ratio)(# 기)	PSR(Price sales ratio)(# 기)	PSR(Price sales ratio)(# 기)	기업가치 (EV)(백 만원)	EBITDA(백 만원)	
0 (주)CMG 제작	58820	2007/12	32102.0	5.0	2001/08/31	Nan	UQ	120.08	215.25	...	0.0	149.75	20.92	69645.33	-1288.45	
1 (주)CMG 제작	58820	2008/12	32102.0	5.0	2001/08/31	Nan	UQ	56.55	55.67	...	0.0	0.0	9.13	1.49	-454.50	-3419.79
2 (주)CMG 제작	58820	2009/12	32102.0	5.0	2001/08/31	Nan	UQ	-33.92	-8.50	...	0.0	0.0	8.40	1.72	26517.58	-7100.69
3 (주)CMG 제작	58820	2010/12	32102.0	5.0	2001/08/31	Nan	UQ	5.66	-11.92	...	0.0	0.0	3.22	1.04	31200.74	-5598.06
4 (주)CMG 제작	58820	2011/12	32102.0	5.0	2001/08/31	Nan	UQ	14.18	7.86	...	0.0	0.0	3.14	1.05	37758.74	2038.64

전처리 시행



Basic Preprocessing Data

회사명	처 기 부 도 부 부	회계년 도	산업코드	소 속 코드	상장일	당 시 의 건 강 고 드	총자본 증가율	유형자 산증가 율	PCR(Price cash-flow ratio)(# 기)	PCR(Price cash-flow ratio)(# 기)	PSR(Price sales ratio)(# 기)	PSR(Price sales ratio)(# 기)	기업가치 (EV)(백 만원)	EBITDA(백 만원)	EBITDA 증가율 (%)	EBITDA 금액비율 (%)	
0 (주)CMG 제작	58820	0	2007.0	032102	5.0	2001/08/31	UQ	120.08	215.25	...	0.00	149.75	20.92	69645.33	-1288.45	-71.74	0.00
1 (주)CMG 제작	58820	0	2008.0	032102	5.0	2001/08/31	UQ	56.55	55.67	...	0.00	9.13	1.49	-454.50	-3419.79	-31.64	-4.58
2 (주)CMG 제작	58820	0	2009.0	032102	5.0	2001/08/31	UQ	-33.92	-8.50	...	0.00	8.40	1.72	26517.58	-7100.69	-69.98	-6.80
3 (주)CMG 제작	58820	0	2010.0	032102	5.0	2001/08/31	UQ	5.66	-11.92	...	0.00	3.22	1.04	31200.74	-5598.06	-49.32	-5.51
4 (주)CMG 제작	58820	0	2011.0	032102	5.0	2001/08/31	UQ	14.18	7.86	...	0.00	3.14	1.05	37758.74	2038.64	11.48	11.24

- 상장회사 재무비율 정보 Data



시설자금 대출의 평균 거치기간 2년

2007년 이전 재무 데이터 Nan값 처리

인수기업, 피인수기업 데이터 처리

회사명 변경 사유로 인한 데이터 처리

3. 모집단 선정

03. 전처리(Basic) 시행 후 Data 버전 분리



UBION – Trillion

회사명	기준년도	상장일	회사의 종기별 증가율	당기기 금액(억 원)	PBR(Price cash-flow ratio)(회 기)	PSR(Price sales ratio)(회 기)	PSR(Price sales ratio)(전 기)	기업가치 (EV)(백 만원)	EBITDA(백 만원)	EBITDA 증가율 (%)	EBITDA/ 급부채 (배)				
0 (주)CMG	2007.0	032102	5.0	2001/08/31	UQ	120.08	215.25	—	0.00	149.75	20.92	69545.33	-1288.45	-7174	0.00
1 (주)CMG	2008.0	032102	5.0	2001/08/31	UQ	56.55	55.67	—	0.00	9.13	1.49	-454.50	-3419.79	-3184	-4.56
2 (주)CMG	2009.0	032102	5.0	2001/08/31	UQ	-33.92	-8.50	—	0.00	8.40	1.72	26517.58	-7100.69	-69.98	-6.80
3 (주)CMG	2010.0	032102	5.0	2001/08/31	UQ	5.66	-11.92	—	0.00	3.22	1.04	31200.74	-5598.06	-49.32	-5.51
4 (주)CMG	2011.0	032102	5.0	2001/08/31	UQ	14.18	7.86	—	0.00	3.14	1.05	37758.74	2038.64	11.48	11.24

Basic Preprocessing Data

Binning

- 데이터의 극단치를 구간화로 조정
 - 표준편차 기준 ($\pm 3\sigma$, $\pm 2\sigma$, $\pm 1\sigma$, μ)
 - 총 8개의 구간으로 Data 구간화

Firm Size

- 상시근로자수 300인 미만인 기업
 - 산업코드 03, 10, 07의 중소기업 기준
 - 300인 이상 중견기업을 제외

Differential

- Feature의 전기 대비 변화율을 이용
 - 동일 기수의 값으로 구성된 재무비율
 - 전기, 현재 기수의 값으로 구성된 재무비율

날짜	기준년도	상장일	회사의 종기별 증가율	당기기 금액(억 원)	PBR(Price cash-flow ratio)(회 기)	PSR(Price sales ratio)(회 기)	PSR(Price sales ratio)(전 기)	기업가치 (EV)(백 만원)	EBITDA(백 만원)	EBITDA 증가율 (%)	EBITDA/ 급부채 (배)	구간화된 기준 부채 액			
												기준 부채 액부	회계년 도	산업코드	수 적 률
0 (주)CMG	2008.0	030803	3.0	0	120.08	215.25	—	0.00	149.75	20.92	69545.33	-1288.45	-7174	0.00	0
1 (주)CMG	2009.0	030803	3.0	0	56.55	55.67	—	0.00	9.13	1.49	-454.50	-3419.79	-3184	-4.56	1
2 (주)CMG	2010.0	030803	3.0	0	-33.92	-8.50	—	0.00	8.40	1.72	26517.58	-7100.69	-69.98	-6.80	2
3 (주)CMG	2011.0	030803	3.0	0	5.66	-11.92	—	0.00	3.22	1.04	31200.74	-5598.06	-49.32	-5.51	3
4 (주)CMG	2012.0	030803	3.0	0	14.18	7.86	—	0.00	3.14	1.05	37758.74	2038.64	11.48	11.24	4

날짜	기준년도	상장일	회사의 종기별 증가율	당기기 금액(억 원)	PBR(Price cash-flow ratio)(회 기)	PSR(Price sales ratio)(회 기)	PSR(Price sales ratio)(전 기)	기업가치 (EV)(백 만원)	EBITDA(백 만원)	EBITDA 증가율 (%)	EBITDA/ 급부채 (배)	구간화된 기준 부채 액			
												기준 부채 액부	회계년 도	산업코드	수 적 률
0 (주)CMG	2008.0	030803	3.0	0	120.08	215.25	—	0.00	149.75	20.92	69545.33	-1288.45	-7174	0.00	0
1 (주)CMG	2009.0	030803	3.0	0	56.55	55.67	—	0.00	9.13	1.49	-454.50	-3419.79	-3184	-4.56	1
2 (주)CMG	2010.0	030803	3.0	0	-33.92	-8.50	—	0.00	8.40	1.72	26517.58	-7100.69	-69.98	-6.80	2
3 (주)CMG	2011.0	030803	3.0	0	5.66	-11.92	—	0.00	3.22	1.04	31200.74	-5598.06	-49.32	-5.51	3
4 (주)CMG	2012.0	030803	3.0	0	14.18	7.86	—	0.00	3.14	1.05	37758.74	2038.64	11.48	11.24	4

날짜	기준년도	상장일	회사의 종기별 증가율	당기기 금액(억 원)	PBR(Price cash-flow ratio)(회 기)	PSR(Price sales ratio)(회 기)	PSR(Price sales ratio)(전 기)	기업가치 (EV)(백 만원)	EBITDA(백 만원)	EBITDA 증가율 (%)	EBITDA/ 급부채 (배)	구간화된 기준 부채 액			
												기준 부채 액부	회계년 도	산업코드	수 적 률
0 (주)CMG	2008.0	030803	3.0	0	120.08	215.25	—	0.00	149.75	20.92	69545.33	-1288.45	-7174	0.00	0
1 (주)CMG	2009.0	030803	3.0	0	56.55	55.67	—	0.00	9.13	1.49	-454.50	-3419.79	-3184	-4.56	1
2 (주)CMG	2010.0	030803	3.0	0	-33.92	-8.50	—	0.00	8.40	1.72	26517.58	-7100.69	-69.98	-6.80	2
3 (주)CMG	2011.0	030803	3.0	0	5.66	-11.92	—	0.00	3.22	1.04	31200.74	-5598.06	-49.32	-5.51	3
4 (주)CMG	2012.0	030803	3.0	0	14.18	7.86	—	0.00	3.14	1.05	37758.74	2038.64	11.48	11.24	4

날짜	기준년도	상장일	회사의 종기별 증가율	당기기 금액(억 원)	PBR(Price cash-flow ratio)(회 기)	PSR(Price sales ratio)(회 기)	PSR(Price sales ratio)(전 기)	기업가치 (EV)(백 만원)	EBITDA(백 만원)	EBITDA 증가율 (%)	EBITDA/ 급부채 (배)	구간화된 기준 부채 액			
												기준 부채 액부	회계년 도	산업코드	수 적 률
0 (주)CMG	2008.0	030803	3.0	0	120.08	215.25	—	0.00	149.75	20.92	69545.33	-1288.45	-7174	0.00	0
1 (주)CMG	2009.0	030803	3.0	0	56.55	55.67	—	0.00	9.13	1.49	-454.50	-3419.79	-3184	-4.56	1
2 (주)CMG	2010.0	030803	3.0	0	-33.92	-8.50	—	0.00	8.40	1.72	26517.58	-7100.69	-69.98	-6.80	2
3 (주)CMG	2011.0	030803	3.0	0	5.66	-11.92	—	0.00	3.22	1.04	31200.74	-5598.06	-49.32	-5.51	3
4 (주)CMG	2012.0	030803	3.0	0	14.18	7.86	—	0.00	3.14	1.05	37758.74	2038.64	11.48	11.24	4

3. 모집단 선정

03. Train & Test 분리



UBION – Trillion

Feature

총자본 증가율	유형자 산증가 율	PCR(Price cash-flow ratio)(최 저)	PSR(Price sales ratio)(최 고)	PSR(Price sales ratio)(최 저)	기업가치 (EV)(백만 원)	EBITDA(백 만원)	EBITDA/ 매출액 (%)	EBITDA/ 금융비용 (배)	차 기 부 여 부
120.08	215.25	...	0.00	149.75	20.92	69545.33	-1288.45	-71.74	0.00	0
56.55	55.67	...	0.00	9.13	1.49	-454.50	-3419.79	-31.64	-4.56	0
-33.92	-8.50	...	0.00	8.40	1.72	26517.58	-7100.69	-69.98	-6.80	0
5.66	-11.92	...	0.00	3.22	1.04	31200.74	-5598.06	-49.32	-5.51	0
14.18	7.86	...	0.00	3.14	1.05	37758.74	2038.64	11.48	11.24	0

Label

연도 기준 Split

Before 2018년

After 2018년

Random Split

Shuffle = True

Basic Data

Train Set
부도 data / 176개
Test Set
부도 data / 38개

Train Set
부도 data / 176개
Test Set
부도 data / 38개

Binning

부도 data / 176개
부도 data / 38개

부도 data / 176개
부도 data / 38개

Firm Size

부도 data / 163개
부도 data / 35개

부도 data / 163개
부도 data / 35개

Differential

부도 data / 132개
부도 data / 37개

부도 data / 139개
부도 data / 30개

3. 모집단 선정

03. Train & Test 분리



UBION – Trillion

연도 기준 Split

Random Split



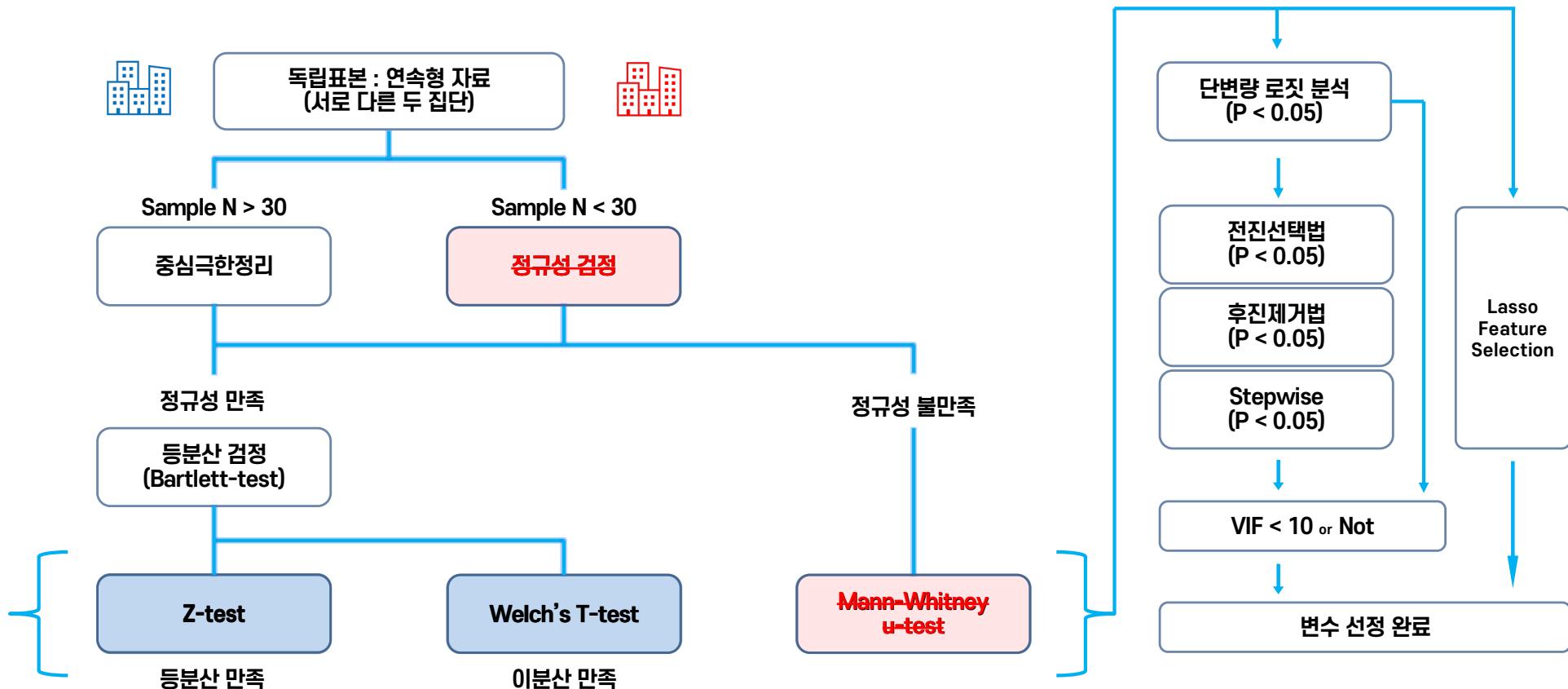
04. Feature Selection

- 01. 순서도
- 02. 유의성 검정 방식

4. Feature Selection



01. 순서도



4. Feature Selection

02. 유의성 검정 방식

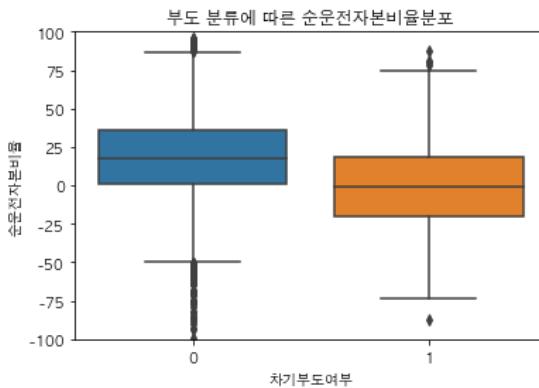


UBION – Trillion

등분산 검정 (Bartlett-test)

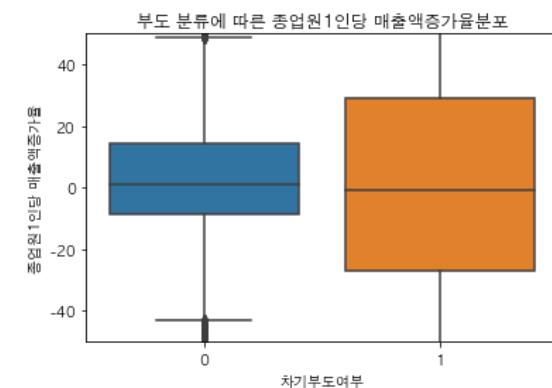
H0 : 2개 모집단의 등분산 총족
H1 : 2개 모집단의 등분산 총족 X

Bartlett, M. S. (1937). Properties of Sufficiency and Statistical Tests. Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, Vol. 160, No.901, pp. 268-282.



Bartlett 통계치, P값 : [35.56, 4.1e -9]

T-test 통계치, P값 : [10.69, 1.3e -26]



Bartlett 통계치, P값 : [0.033, 0.855]

Welch's 통계치, P값 : [-1.166, 0.244]

Z-test

H0 : 2개 모집단의 평균의 동질성 총족
H1 : 2개 모집단의 평균의 동질성 총족 X

Welch's T-test

H0 : 2개 모집단의 평균의 동질성 총족
H1 : 2개 모집단의 평균의 동질성 총족 X

기본 가정

- 모집단의 정규성 총족
- 등분산성 총족

기본 가정

- 모집단의 정규성 총족
- 등분산성 총족 X
- 소표본에 적합

예시

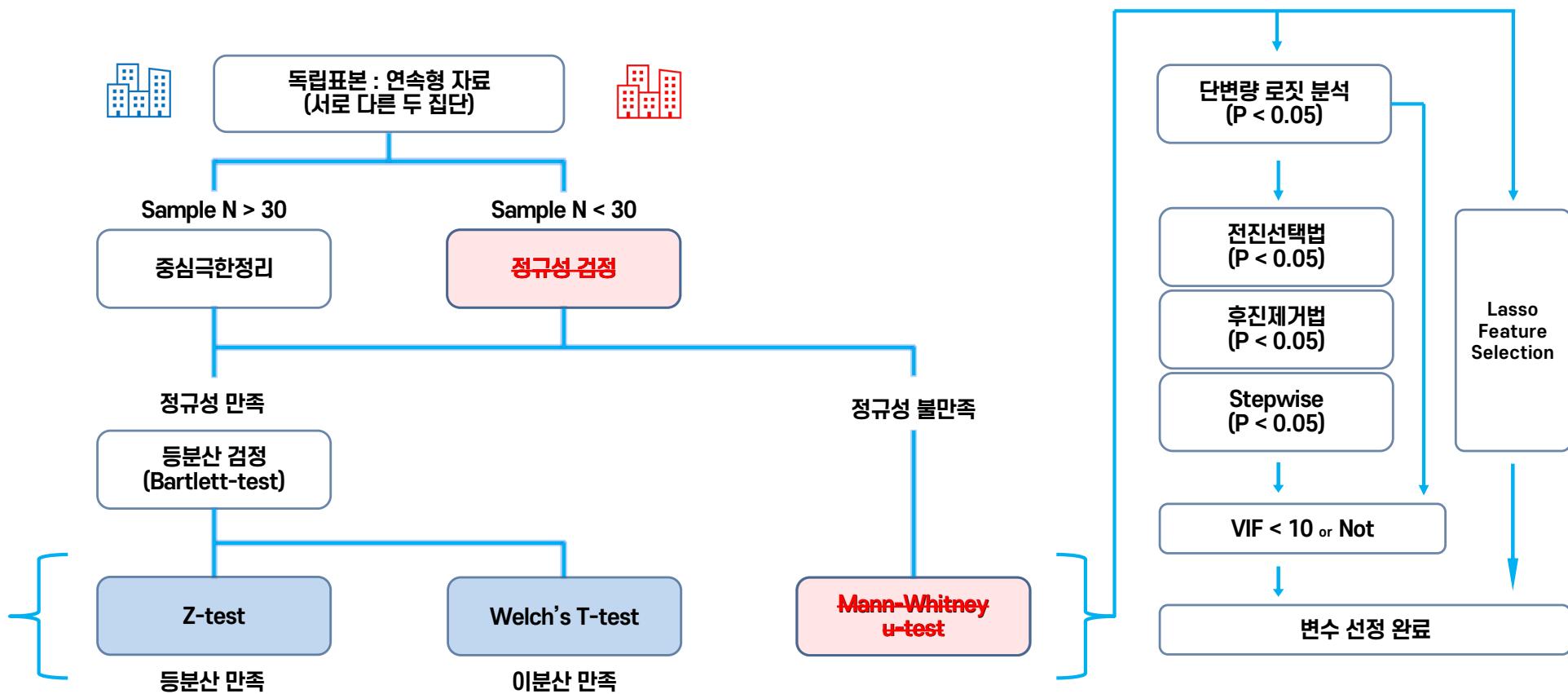
등분산	이분산
평균의 차이 O	순운전자본비율
평균의 차이 X	1인당 매출액 증가율

4. Feature Selection

01. 순서도



UBION – Trillion



4. Feature Selection

02. 유의성 검정 방식



UBION – Trillion

단변량 로짓 분석 (P < 0.05)

feature(설명변수)의 개수	single(단순)	feature 1개 (ex. 단순선형회귀모형)
	multiple(다중)	feature 2개 이상 (ex. 다중선형회귀모형)
label(종속변수)의 개수	univariate(단변량)	label 변수가 1개 (단변량 로짓분석, T or F)
	bivariate(이변량)	label 변수가 2개 (이변량 로짓분석)
	multivariate(다변량)	label 변수가 2개 이상 (다변량 로짓분석, 신용등급 분류)

종속변수(label) : 명목형 변수

목표변수 Y가 특정 범주에 속할 확률을 도출

이 중 단변량 로지스틱은 명목형인 종속변수가 1개인 로지스틱 회귀를 의미

Logit Regression Results

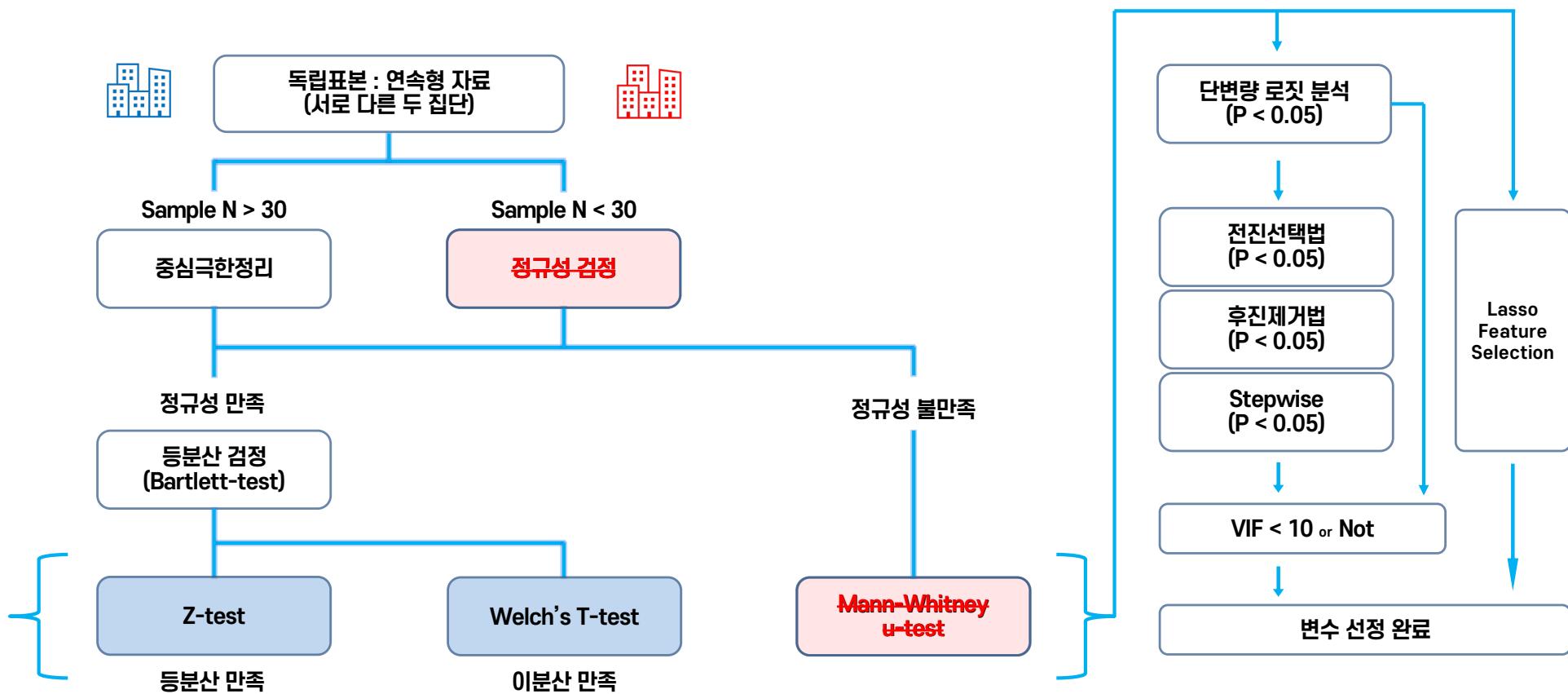
Dep. Variable:	부도여부	No. Observations:	15162			
Model:	Logit	Df Residuals:	15160			
Method:	MLE	Df Model:	1			
Date:	Mon, 13 Dec 2021	Pseudo R-squ.:	0.04140			
Time:	19:31:13	Log-Likelihood:	-919.53			
converged:	True	LL-Null:	-959.24			
Covariance Type:	nonrobust	LLR p-value:	5.029e-19			
	coef	std err	z	P> z	[0.025	0.975]
const	-4.2436	0.077	-55.260	0.000	-4.394	-4.093
순운전자본비율	-0.0200	0.002	-9.162	0.000	-0.024	-0.016

4. Feature Selection

01. 순서도



UBION – Trillion



4. Feature Selection

02. 유의성 검정 방식



UBION – Trillion

Wrapper Feature를 조정하며 모형을 형성하고 예측 성능을 참고하여 Feature를 선택

전진선택법
(P < 0.05)

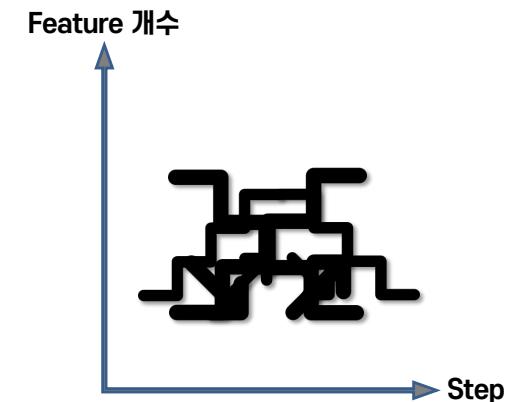
가장 단순한 null model에서 출발하여 변수를 하나씩 추가하여 회귀식을 적합시키는 방식
Feature가 선택되면 제거되지 않음

후진제거법
(P < 0.05)

모든 Feature가 포함된 model에서 출발하여 변수를 하나씩 제거하여 회귀식을 적합시키는 방식
Feature가 제거되면 추가되지 않음

Stepwise
(P < 0.05)

가장 단순한 null model에서 출발하여 변수를 하나씩 추가하고 제거하여 회귀식을 적합시키는 방식
Feature가 선택되어도 제거가 될 수 있으며 전진과 후진제거를 반복한다.



Embedded 예측 모형 최적화 과정(회귀계수 추정 과정)에서 각 Feature가 선택되는 방식

Lasso
Feature
Selection

LASSO (Least Absolute Shrinkage and Selection Operation)

$$(\hat{\alpha}, \hat{\beta}) = \arg \min \left\{ \sum_{i=1}^N \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \sum_j |\beta_j| \leq t.$$

OLS + 회귀계수의 절댓값이 일정 수준을 초과X (규제)
규제로 회귀 계수를 최소화하기에 특정 Feature의 계수는 0으로 수렴될 수 있음
→ 이를 이용하여 Feature Selection을 진행

Filter feature간 상관성을 기반으로 추출

VIF < 10 or Not

$$VIF_i = Var(\hat{\beta}_i) = \frac{\sigma_e^2}{(n-1)Var(X_i)} \cdot \frac{1}{1-R_i^2}$$

(R_i^2 : 독립변수 X_i 를 다른 독립변수들로 선형회귀한 성능(결정계수))

다중공선성 (Feature간 상관관계가 높은 경우)을 측정
VIF가 10이상인 경우 해당 변수의 다중공선성이 있다고 판단

05. 모델링 시행

- 01. 사용하는모델
- 02. 샘플링
- 03. 모델설명
- 04. 중간결과도출

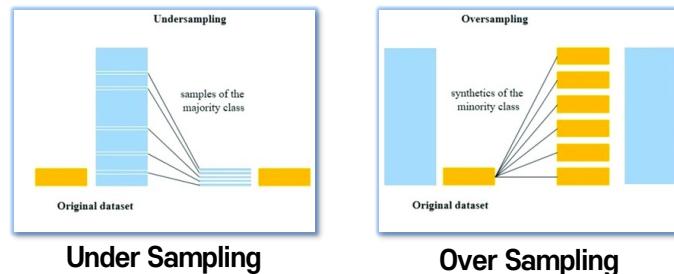
5. 모델링 시행



01. 사용하는 모델

UBION – Trillion

Step 1. Sampling 시행 (train set only)



Step 2. Modeling 시행

 Logistic Regression 다중 회귀의 로짓 변환으로 Data 분류 분류 & 예측 가능	 SGD Classifier Cost 최소값을 확률적으로 도출 분류 & 예측 가능	 K-Neighbors Classifier 기존 Data와 거리 기반 분류 분류 & 예측 가능	 Support Vector Machine Hyper-plane(초평면) 기반 Data 분류 분류 & 예측 가능	 Decision Tree Classifier 의사결정 규칙을 나무 형태로 도표화 분류 & 예측 가능	 Random Forest Classifier 의사결정 나무 기반 배깅 (Bagging) 시행 분류 & 예측 가능	 Pruning 의사결정 나무 기반 깊이, 노드 조정 분류 & 예측 가능
---	--	---	---	---	--	---

5. 모델링 시행

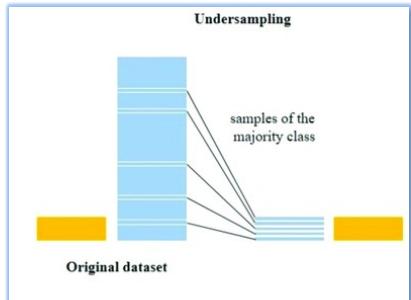
02. 샘플링



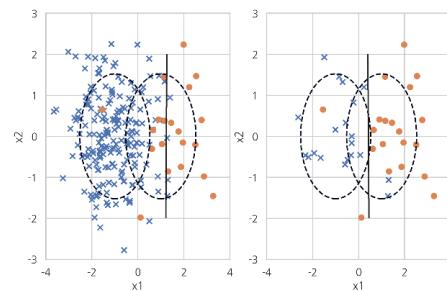
UBION – Trillion

Sampling 목적 Overfitting 문제 해결 + Accuracy & Recall (Precision) 문제 해결

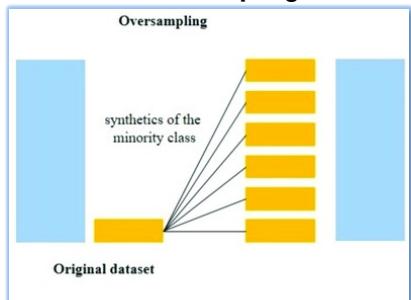
Under Sampling



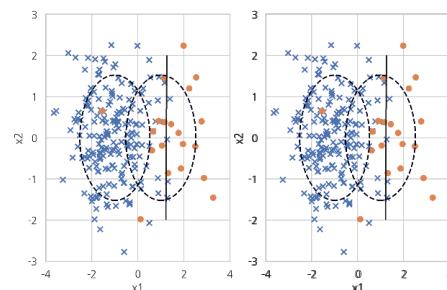
Random Under Sampling



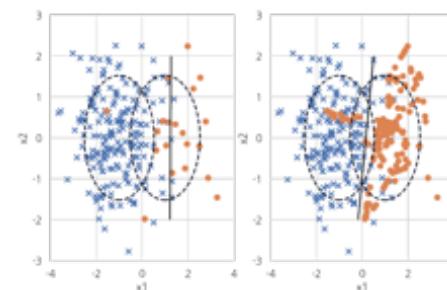
Over Sampling



Random Over Sampling



SMOTE



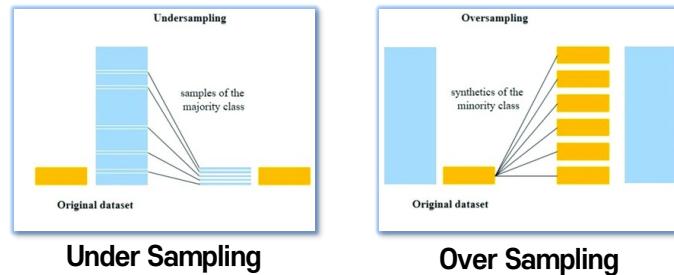
5. 모델링 시행



01. 사용하는 모델

UBION – Trillion

Step 1. Sampling 시행 (train set only)



Step 2. Modeling 시행

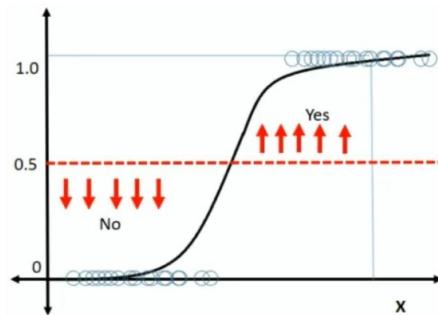
 Logistic Regression 다중 회귀의 로짓 변환으로 Data 분류 분류 & 예측 가능	 SGD Classifier Cost 최소값을 확률적으로 도출 분류 & 예측 가능	 K-Neighbors Classifier 기존 Data와 거리 기반 분류 분류 & 예측 가능	 Support Vector Machine Hyper-plane(초평면) 기반 Data 분류 분류 & 예측 가능	 Decision Tree Classifier 의사결정 규칙을 나무 형태로 도표화 분류 & 예측 가능	 Random Forest Classifier 의사결정 나무 기반 배깅 (Bagging) 시행 분류 & 예측 가능	 Pruning 의사결정 나무 기반 깊이, 노드 조정 분류 & 예측 가능
---	--	---	---	---	--	---

5. 모델링 시행

03. 모델 설명



UBION – Trillion



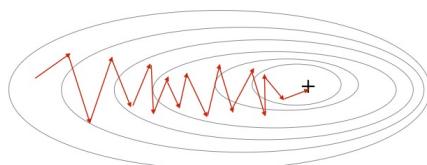
Logistic Regression

선형회귀를 이용하여 데이터의 특정 Label에 속할 확률을 예측하고 분류하는 Supervised Algorithm

장점 : 선형회귀를 사용하면 확률이 음과 양의 방향으로 무한대까지 뻗어 나간다. X축을 학습시간이라고 하 고 y를 시험 합격 확률이라고 가정하면 선형회귀에서는 특정값 이하의 학습시간은 확률이 0이 안된다, 이런 문제를 해결한 것이 로지스틱 회귀이다.

단점

로지스틱 회귀는 많은 사람들에게 사용되고 있지만 교호 작용을 따로 추가해야 한다는 것과 같이 표현적인 제한이 있고 다른 모델들이 더 예측 성능이 좋을 수 있습니다.



SGD Classifier

예측모델은 손실을 줄여 예측하게 될 확률을 높이는데 SGD는 손실을 줄이기 위한 최적값을 찾기 위해 무작위로 아무 값이나 넣어보면서 최적값을 구하는 분류보델이다,

장점: batch 단위로 loss를 계산하기 때문에 loss function을 여러 번 빨리 계산할 수 있으며 Local minima에 쉽게 빠지지 않고 global minima를 찾을 확률이 높다

단점 : 학습률이 낮으면 곧장 최적화하지 못하고 지그재그로 이동하게 되면서 지역 최솟값이 갇혀 빠져나오지 못하는 경우가 있고, 학습률이 높으면 최적화 자체를 실패할 수 있는 문제가 있다.

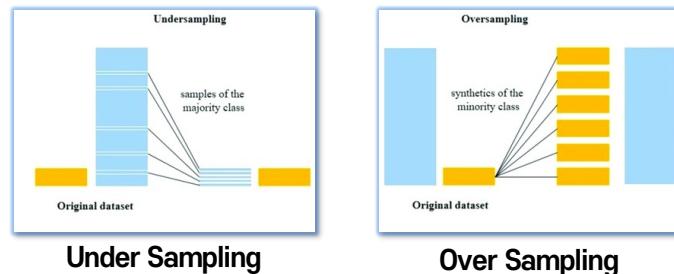
5. 모델링 시행



01. 사용하는 모델

UBION – Trillion

Step 1. Sampling 시행 (train set only)



Step 2. Modeling 시행

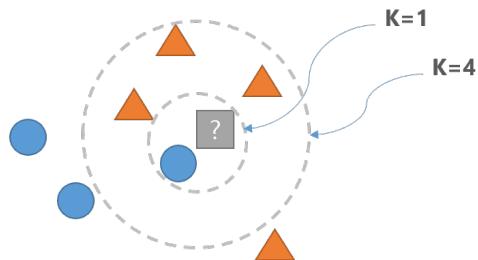
 Logistic Regression 다중 회귀의 로짓 변환으로 Data 분류 분류 & 예측 가능	 SGD Classifier Cost 최소값을 확률적으로 도출 분류 & 예측 가능	 K-Neighbors Classifier 기존 Data와 거리 기반 분류 분류 & 예측 가능	 Support Vector Machine Hyper-plane(초평면) 기반 Data 분류 분류 & 예측 가능	 Decision Tree Classifier 의사결정 규칙을 나무 형태로 도표화 분류 & 예측 가능	 Random Forest Classifier 의사결정 나무 기반 배깅 (Bagging) 시행 분류 & 예측 가능	 Pruning 의사결정 나무 기반 깊이, 노드 조정 분류 & 예측 가능
---	--	---	---	---	--	---

5. 모델링 시행

03. 모델 설명



UBION – Trillion

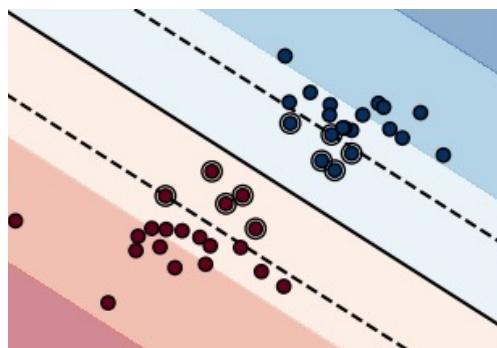


K-Neighbors
Classifier

새로운 데이터를 입력받았을 때 가장 가까이 있는 것이 무엇이냐를 중심으로 새로운 데이터를 분류하는 알고리즘이다.

장점 : 단순하고 효율적이며 훈련 단계가 빠르다, 기저 데이터 분포에 대한 가정을 하지 않는다

단점 : 주변 데이터의 개수 즉 k를 무엇으로 지정하느냐에 따라 분류예측 성능이 크게 좌우됨.



Support
Vector Machine

Support vector와 hyperplane(초평면)을 이용해서 분류를 수행하게 되는 알고리즘

장점 : 신경망보다 사용이 간결함. 범주나 수치 예측 문제에 사용이 가능, 오류 데이터에 대한 영향이 없다. 과적합 되는 경우가 적다.

단점 : 의사결정나무처럼 직관적인 해석이 불가능하다 (어떤 이유로 데이터들이 분류됐는지 알 수가 없다), 학습속도가 느림, 최적의 모델을 찾기 위해서 커널과 모델에서 다양한 테스트가 필요하다. 따라서 여러 연산이 필요하고 입력 데이터 셋이 많을 경우에 학습 속도가 느립니다.

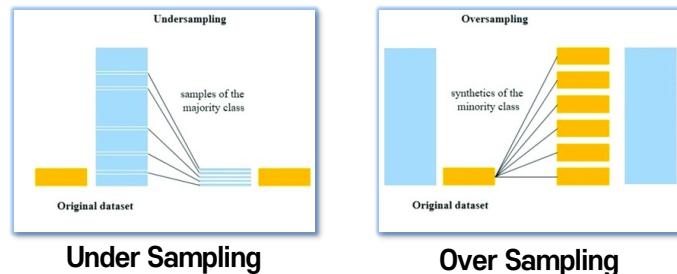
5. 모델링 시행



01. 사용하는 모델

UBION – Trillion

Step 1. Sampling 시행 (train set only)



Under Sampling

Over Sampling

Step 2. Modeling 시행

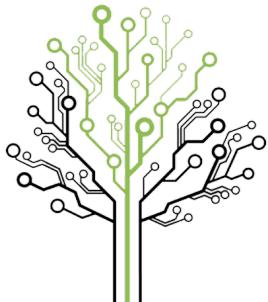
 Logistic Regression 다중 회귀의 로짓 변환으로 Data 분류 분류 & 예측 가능	 SGD Classifier Cost 최소값을 확률적으로 도출 분류 & 예측 가능	 K-Neighbors Classifier 기존 Data와 거리 기반 분류 분류 & 예측 가능	 Support Vector Machine Hyper-plane(초평면) 기반 Data 분류 분류 & 예측 가능	 Decision Tree Classifier 의사결정 규칙을 나무 형태로 도표화 분류 & 예측 가능	 Random Forest Classifier 의사결정 나무 기반 배깅 (Bagging) 시행 분류 & 예측 가능	 Pruning 의사결정 나무 기반 깊이, 노드 조정 분류 & 예측 가능
---	--	---	---	---	--	---

5. 모델링 시행

03. 모델 설명



UBION – Trillion

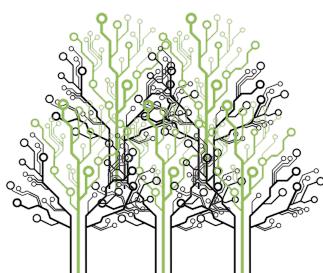


Decision Tree
Classifier

Gini-index 기준으로 노드 형성 + 정보획득률로 Feature importance 도출

결과 해석의 직관성 + 비모수, 이상치 등에 강한 성능

Overfitting, 소수 데이터에 취약하며 Data를 나누는 Cut-off 지점의 1종, 2종 오류 발생



Random Forest
Classifier

Random한 트리를 형성하기에 Feature가 무작위로 주입된다.
따라서 각 트리의 예측이 비상관성을 가지며 Overfitting에도 강함

랜덤하게 Decision Tree를 형성하고 Voting을 사용하여 최종 모델을 선택한다.

개별 트리에 대한 분석이 어렵고 훈련시 메모리의 소모가 크다. 또한, 비정형 데이터에 약하다.



Pruning

깊이와 노드를 조절하여 지나친 세분화를 막아 Overfitting을 막아줄 수 있다.

Decision Tree 내 노드, 깊이를 조절하여 빠른 계산이 가능하며 Overfitting에 강하다.

정보 손실이 발생 + Underfitting에 약하다.

5. 모델링 시행

04. 중간 결과 도출



UBION – Trillion

DataFrame

Basic Preprocessing

	날짜	주가	거래량	시가	고가	저가	전일비	전일대비%	시가총액	PER	EPS	PERxEPS	PERxEPSx주가	PERxEPSx전일비	PERxEPSx전일대비%	PERxEPSx시가총액	PERxEPSx전일비x시가총액	PERxEPSx전일대비%x시가총액
0	2018-01-01	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%
1	2018-01-02	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%
2	2018-01-03	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%
3	2018-01-04	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%
4	2018-01-05	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%

Basic + Binning

	날짜	주가	거래량	시가	고가	저가	전일비	전일대비%	시가총액	PER	EPS	PERxEPS	PERxEPSx주가	PERxEPSx전일비	PERxEPSx전일대비%	PERxEPSx시가총액	PERxEPSx전일비x시가총액	PERxEPSx전일대비%x시가총액
0	2018-01-01	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%
1	2018-01-02	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%
2	2018-01-03	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%
3	2018-01-04	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%
4	2018-01-05	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%

Basic + Firm Size

	날짜	주가	거래량	시가	고가	저가	전일비	전일대비%	시가총액	PER	EPS	PERxEPS	PERxEPSx주가	PERxEPSx전일비	PERxEPSx전일대비%	PERxEPSx시가총액	PERxEPSx전일비x시가총액	PERxEPSx전일대비%x시가총액
0	2018-01-01	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%
1	2018-01-02	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%
2	2018-01-03	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%
3	2018-01-04	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%
4	2018-01-05	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%

Basic + Differential

	날짜	주가	거래량	시가	고가	저가	전일비	전일대비%	시가총액	PER	EPS	PERxEPS	PERxEPSx주가	PERxEPSx전일비	PERxEPSx전일대비%	PERxEPSx시가총액	PERxEPSx전일비x시가총액	PERxEPSx전일대비%x시가총액
0	2018-01-01	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%
1	2018-01-02	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%
2	2018-01-03	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%
3	2018-01-04	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%
4	2018-01-05	50000.0	20000.0	50000.0	50000.0	50000.0	-10000.0	-2.0%	10000000000.0	10.0	1.0	10.0	10000000000.0	-1000000000.0	-2.0%	10000000000.0	-10000000000.0	-2.0%

Train & Test Split

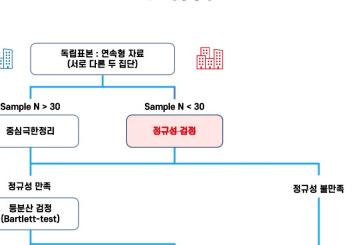
연도 기준 (2018)

Random train_test_split

1624 Cases

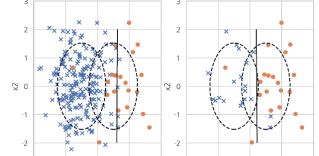
Feature Selection

T-test

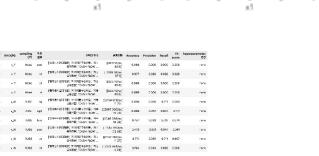
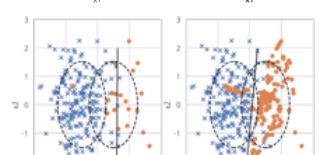
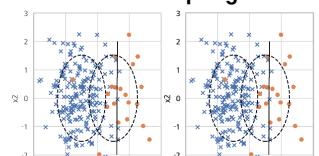


Sampling or Not

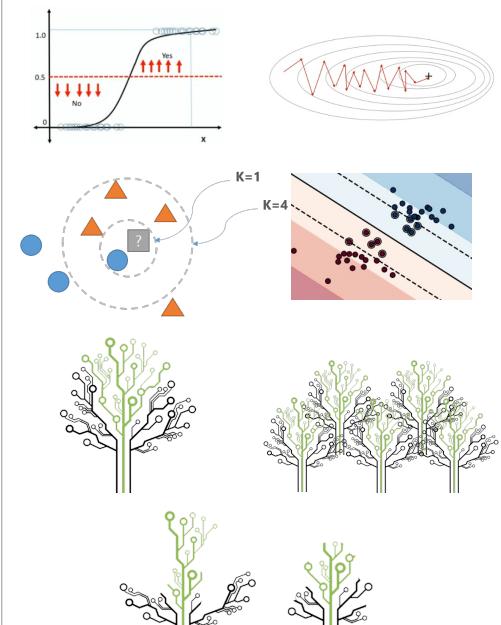
Under Sampling



Over Sampling



Modeling



5. 모델링 시행



04. 중간 결과 도출

UBION – Trillion

	데이터프레임	train_test_split(Y/R)	t-test(Y/N)	변수선택 법	VIF(Y/N)	sampling 방식	적용 모델	선택된 변수	훈동행렬	Accuracy	Precision	Recall	F1-score	hyperparameter 변경	
108	basic_final_df		Y	t_Y	forward	v_Y	None	svm	['총자본사업이익률', '자기자본구성비율', '자본금회전률', 'CASH FLOW ...	[[3237 0]\n [38 0]]	0.988	0.000	0.000	0.000	none
109	basic_final_df		Y	t_Y	forward	v_Y	None	dt	['총자본사업이익률', '자기자본구성비율', '자본금회전률', 'CASH FLOW ...	[[3199 38]\n [37 1]]	0.977	0.026	0.026	0.026	none
110	basic_final_df		Y	t_Y	forward	v_Y	None	pr	['총자본사업이익률', '자기자본구성비율', '자본금회전률', 'CASH FLOW ...	[[3237 0]\n [38 0]]	0.988	0.000	0.000	0.000	none
111	basic_final_df		Y	t_Y	forward	v_Y	None	rf	['총자본사업이익률', '자기자본구성비율', '자본금회전률', 'CASH FLOW ...	[[3237 0]\n [38 0]]	0.988	0.000	0.000	0.000	none
112	basic_final_df		Y	t_Y	stepwise	v_N	RUSE	lg	['총자본사업이익률', '자기자본구성비율', '자본금회전률', 'CASH FLOW ...	[[2724 513]\n [11 27]]	0.840	0.050	0.711	0.093	none
113	basic_final_df		Y	t_Y	stepwise	v_N	RUSE	sgd	['총자본사업이익률', '자기자본구성비율', '자본금회전률', 'CASH FLOW ...	[[2887 350]\n [15 23]]	0.889	0.062	0.605	0.112	none
114	basic_final_df		Y	t_Y	stepwise	v_N	RUSE	knn	['총자본사업이익률', '자기자본구성비율', '자본금회전률', 'CASH FLOW ...	[[2548 689]\n [10 28]]	0.787	0.039	0.737	0.074	none
115	basic_final_df		Y	t_Y	stepwise	v_N	RUSE	svm	['총자본사업이익률', '자기자본구성비율', '자본금회전률', 'CASH FLOW ...	[[1534 1703]\n [2 36]]	0.479	0.021	0.947	0.041	none
116	basic_final_df		Y	t_Y	stepwise	v_N	RUSE	dt	['총자본사업이익률', '자기자본구성비율', '자본금회전률', 'CASH FLOW ...	[[2497 740]\n [11 27]]	0.771	0.035	0.711	0.067	none
117	basic_final_df		Y	t_Y	stepwise	v_N	RUSE	pr	['총자본사업이익률', '자기자본구성비율', '자본금회전률', 'CASH FLOW ...	[[2273 964]\n [4 34]]	0.704	0.034	0.895	0.066	none

06. 모델 정확도 검정 및 최종 모델 선정

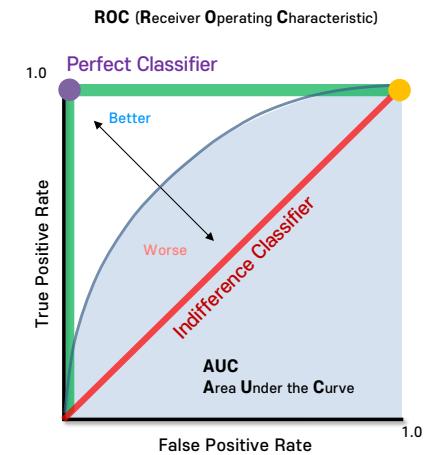
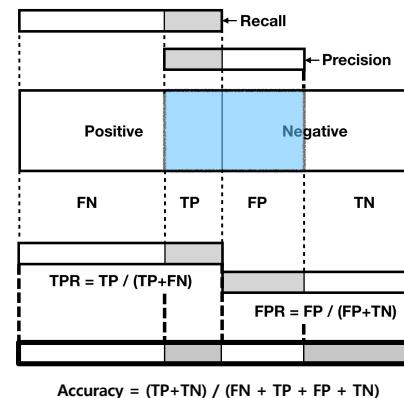
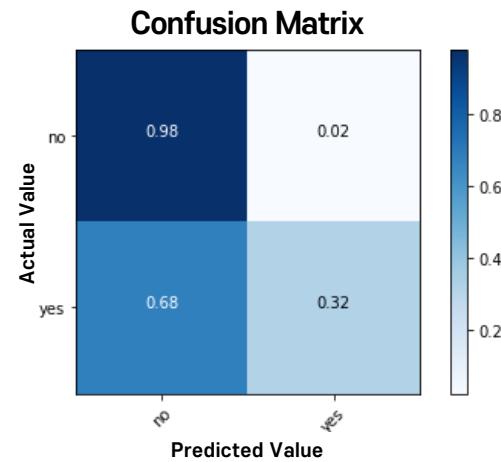
- 01. 정확도검정방식선택
- 02. Hyper Parameter 대상선정및시행
- 03.최종모델선정
- 04.한계점및향후과제

6. 모델 정확도 검정 및 최종 모델 선정



01. 정확도 검정 방식 채택

UBION – Trillion

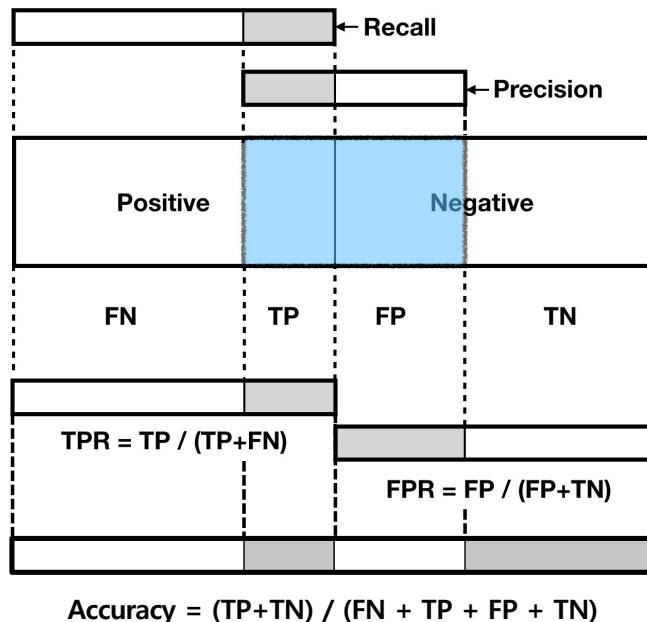


6. 모델 정확도 검정 및 최종 모델 선정



01. 정확도 검정 방식 채택

UBION – Trillion



1. Accuracy

- A. $(TP + TN) / FN + TP + FP + TN$
- B. 직관적이나, 데이터의 불균형에 부적합 -> Precision & Recall 사용

2. Precision

- A. TP / FP

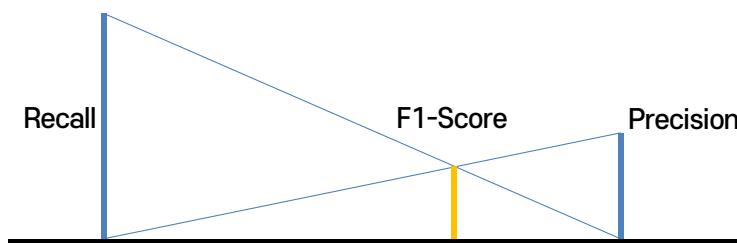
상호보완적이나, Trade – off

3. Recall

- A. TP / FN

4. F1-Score

- A. Precision & Recall의 쓸림을 방지하기 위한 지표
- B. Precision & Recall의 조화 평균 값

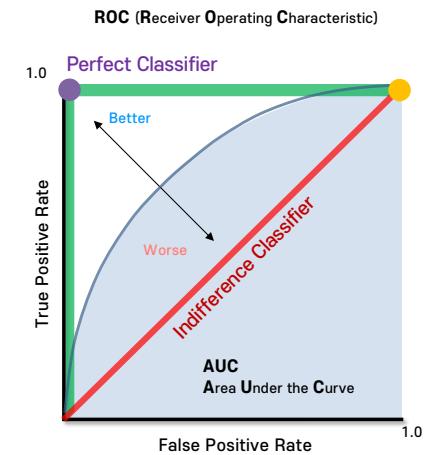
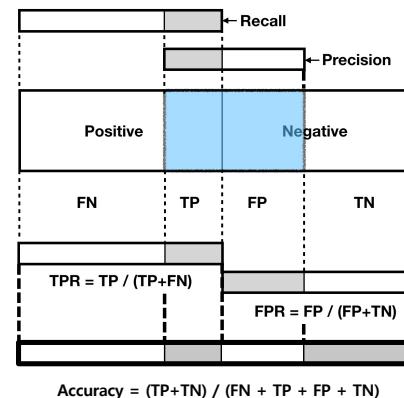
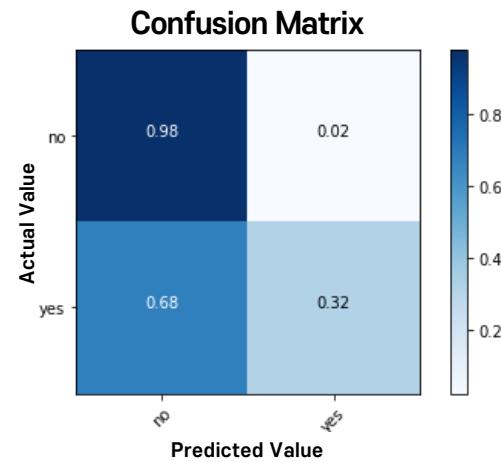


6. 모델 정확도 검정 및 최종 모델 선정



01. 정확도 검정 방식 채택

UBION – Trillion

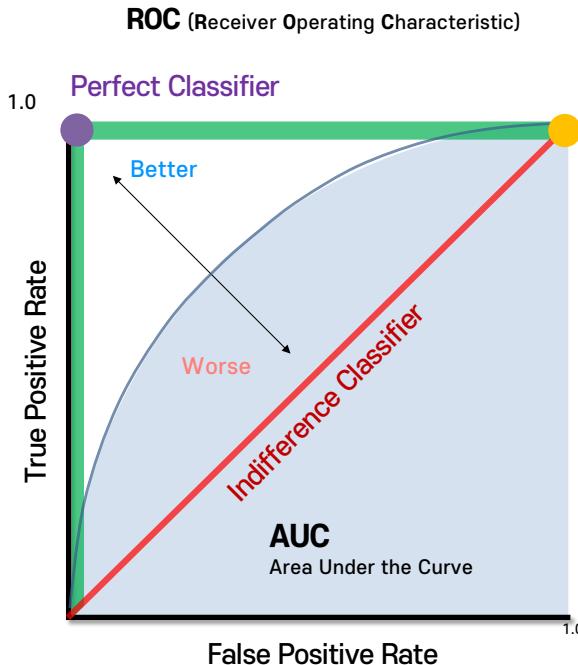


6. 모델 정확도 검정 및 최종 모델 선정



01. 정확도 검정 방식 채택

UBION – Trillion



이진 분류의 성능을 TRP, FPR로 표현할 수 있는 그래프

AUC ROC 곡선 밑 영역의 크기를 표현한 지표
무차별적 분류인 0.5부터 완벽한 분류를 시행하는 1 사이의 값으로 도출됨

X축 FPR(False Positive Rate) 틀린것을 맞았다고 잘못 예측한 수치
($FPR = FP / (FP + TN)$)

Y축 TPR(True Positive Rate) 맞은것을 맞았다고 잘 예측한 수치
($TPR = TP / (TP + FN)$) [=Recall]

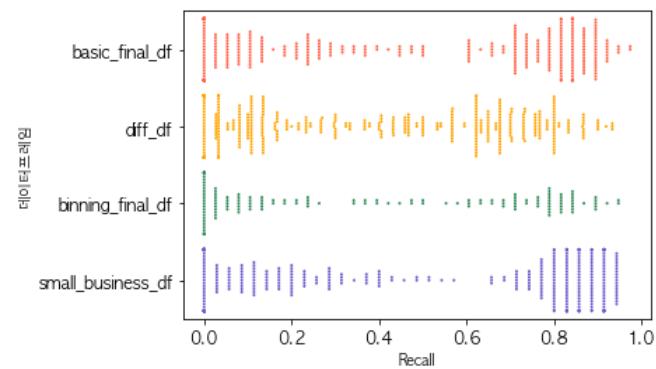
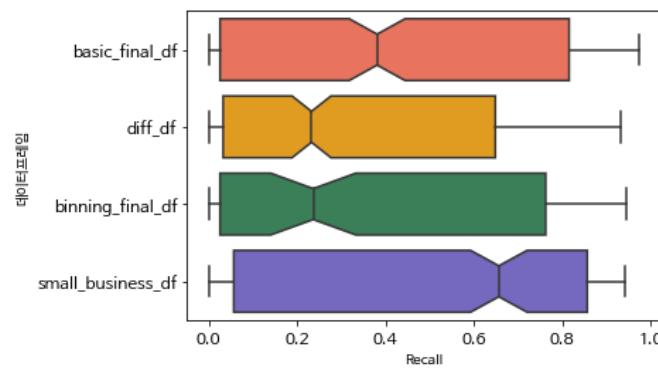
6. 모델 정확도 검정 및 최종 모델 선정



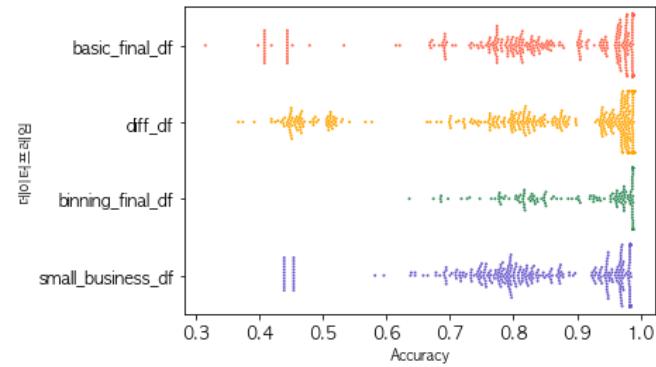
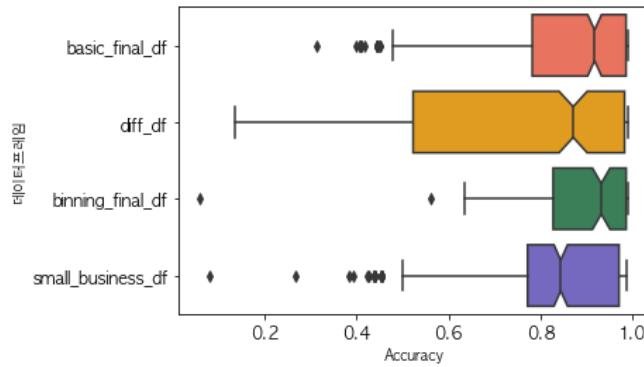
02. Hyper-Parameter 대상 선정 및 실행

UBION – Trillion

Recall



Accuracy



6. 모델 정확도 검정 및 최종 모델 선정

02. Hyper-Parameter 대상 선정 및 실행

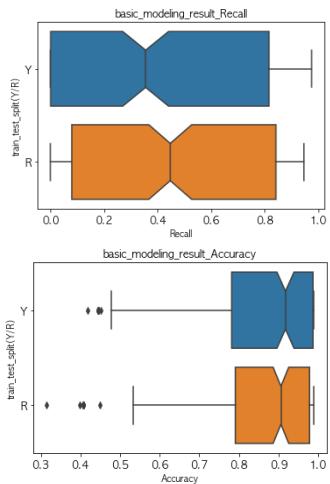


UBION – Trillion

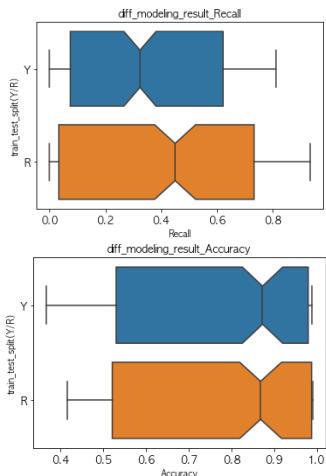
Recall

Accuracy

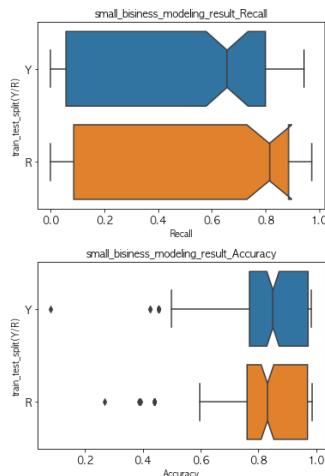
Basic



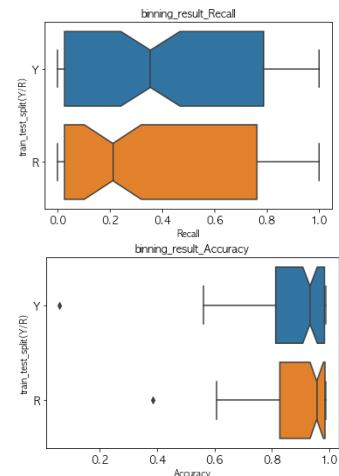
Diff



Small



Binning



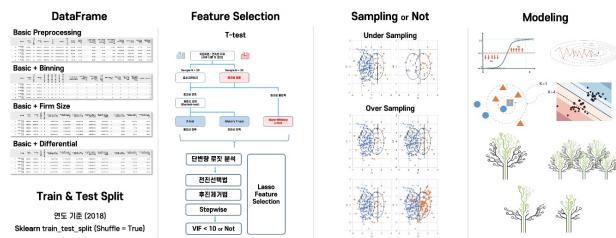
6. 모델 정확도 검정 및 최종 모델 선정



UBION – Trillion

02. Hyper-Parameter 대상 선정 및 실행

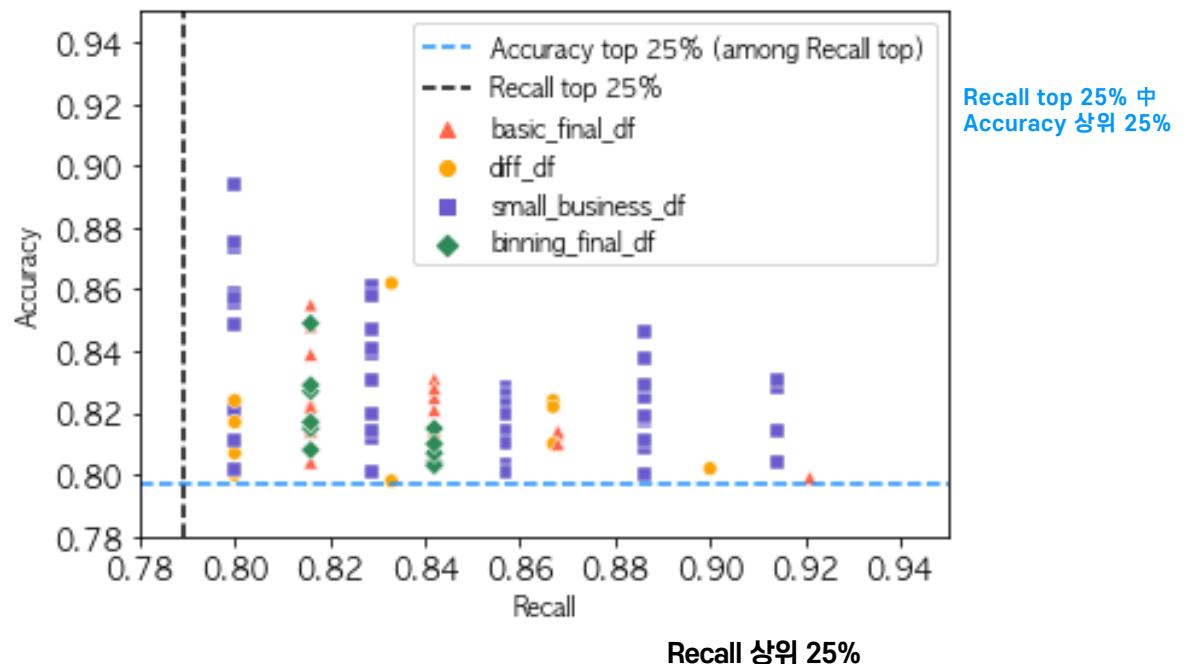
모델링 과정



중간 결과 DataFrame

데이터프레임	train_test_split(Y/N)	L	반복계수	VIF(VIF)	sampling	제작	모형	선택된 변수	총변량수	Accuracy	Precision	Recall	H ₁ -score	hyperparameter
108	basic_final_DF	Y	1 st	forward	v _c	Note	sum	["총생산자원비", "기초생산자원비", "부수생산자원비", "CASH_FLOW"]	[1237 0/1] (38 0)	0.988	0.000	0.000	0.000	None
109	basic_final_DF	Y	1 st	forward	v _c	Note	dt	["총생산자원비", "기초생산자원비", "부수생산자원비", "CASH_FLOW"]	[1237 0/1] (38 0)	0.977	0.026	0.026	0.026	None
110	basic_final_DF	Y	1 st	forward	v _c	Note	pr	["총생산자원비", "기초생산자원비", "부수생산자원비", "CASH_FLOW"]	[1237 0/1] (38 0)	0.988	0.000	0.000	0.000	None
111	basic_final_DF	Y	1 st	forward	v _c	Note	rf	["총생산자원비", "기초생산자원비", "부수생산자원비", "CASH_FLOW"]	[1237 0/1] (38 0)	0.988	0.000	0.000	0.000	None
112	basic_final_DF	Y	1 st	stepwise	v _c	RUSE	lg	["총생산자원비", "기초생산자원비", "부수생산자원비", "CASH_FLOW"]	[2374 5/10] (10 5)	0.840	0.050	0.711	0.093	None
113	basic_final_DF	Y	1 st	stepwise	v _c	RUSE	egd	["총생산자원비", "기초생산자원비", "부수생산자원비", "CASH_FLOW"]	[2897 8/9] (9 8)	0.889	0.062	0.603	0.172	None
114	basic_final_DF	Y	1 st	stepwise	v _c	RUSE	krr	["총생산자원비", "기초생산자원비", "부수생산자원비", "CASH_FLOW"]	[1649 10/8] (8 10)	0.787	0.039	0.737	0.074	None
115	basic_final_DF	Y	1 st	stepwise	v _c	RUSE	svm	["총생산자원비", "기초생산자원비", "부수생산자원비", "CASH_FLOW"]	[1654 7/10] (10 7)	0.419	0.021	0.947	0.041	None
116	basic_final_DF	Y	1 st	stepwise	v _c	RUSE	dt	["총생산자원비", "기초생산자원비", "부수생산자원비", "CASH_FLOW"]	[1247 4/6] (6 4)	0.771	0.035	0.711	0.067	None
117	basic_final_DF	Y	1 st	stepwise	v _c	RUSE	pr	["총생산자원비", "기초생산자원비", "부수생산자원비", "CASH_FLOW"]	[2235 9/4] (4 9)	0.704	0.034	0.695	0.065	None

중간결과 Scatter (Accuracy, Recall)



6. 모델 정확도 검정 및 최종 모델 선정



02. Hyper-Parameter 대상 선정 및 실행

	데이터프레임	train_test_split(Y/R)	t-test(Y/N)	변수선택 법	VIF(Y/N)	sampling 방식	적용 모델	선택된 변수	훈동행렬	Accuracy	Precision	Recall	F1-score	hyperparameter 변경	ROC_Score
0	small_business_df	R	t_Y	forward	v_N	RUSE	lg	['총자본사업이익률', '자기자본구성비율', '총업원수', '단기차입금 대 총차입금...']	[[1819 333]\n [4 31]]	0.846	0.085	0.886	0.155	none	0.933643
1	small_business_df	R	t_Y	stepwise	v_N	RUSE	lg	['자기자본구성비율', '총업원수', '단기차입금 대 총차입금비율', 'CASH F...']	[[1766 386]\n [5 30]]	0.821	0.072	0.857	0.133	none	0.931851
2	small_business_df	R	t_Y	stepwise	v_N	SMOTE	lg	['자기자본구성비율', '총업원수', '단기차입금 대 총차입금비율', 'CASH F...']	[[1782 370]\n [4 31]]	0.829	0.077	0.886	0.142	none	0.929036
3	small_business_df	R	t_Y	forward	v_N	ROSE	lg	['총자본사업이익률', '자기자본구성비율', '총업원수', '단기차입금 대 총차입금...']	[[1778 374]\n [3 32]]	0.828	0.079	0.914	0.145	none	0.927430
4	small_business_df	R	t_Y	forward	v_N	SMOTE	lg	['총자본사업이익률', '자기자본구성비율', '총업원수', '단기차입금 대 총차입금...']	[[1802 350]\n [4 31]]	0.838	0.081	0.886	0.149	none	0.926633
...
87	diff_df	R	t_Y	forward	v_Y	SMOTE	pr	['growth_rate_자기자본순이익률', 'growth_rate_유보액대비율', ...]	[[2338 519]\n [6 24]]	0.818	0.044	0.800	0.084	none	0.824746
88	basic_final_df	R	t_Y	Lasso	v_Y	RUSE	pr	['자기자본증가율', '총업원1인당 인건비증가율', '자기자본순이익률', '경영자본...']	[[2779 465]\n [8 30]]	0.856	0.061	0.789	0.113	none	0.824356
89	diff_df	R	t_Y	None	v_N	SMOTE	pr	['growth_rate_수자비율', 'growth_rate_사내유보 대 자기자본비...']	[[2353 504]\n [6 24]]	0.823	0.045	0.800	0.086	none	0.823906
90	diff_df	R	t_Y	Lasso	v_N	SMOTE	pr	['growth_rate_사내유보율', 'growth_rate_사내유보 대 자기자본...']	[[2334 523]\n [6 24]]	0.817	0.044	0.800	0.083	none	0.799142
91	diff_df	R	t_Y	Lasso	v_Y	SMOTE	pr	['growth_rate_사내유보율', 'growth_rate_사내유보 대 자기자본...']	[[2354 503]\n [6 24]]	0.824	0.046	0.800	0.086	none	0.798530

6. 모델 정확도 검정 및 최종 모델 선정



02. Hyper-Parameter 대상 선정 및 실행

UBION – Trillion

Model

데이터프레임	train_test_split(Y/R)	t-test(Y/N)	변수선택 법	VIF(Y/N)	sampling 방식	적용 모델	선택된 변수	훈동행렬	Accuracy	Precision	Recall	F1-score	ROC_accuracy	hyperparameter 변경	cut-off
binning_final_df	R	t_Y	None	v_N	RUSE	svm	['유형자산증가율', '영업이익증가율', '순이익증가율', '재고자산증가율', '총...	[[2782 462]]n [6 32]]	0.857	0.065	0.842	0.120	0.915	{'C': 100, 'degree': 1, 'gamma': 0.0001, 'kernel...}	Y
small_business_df	R	t_Y	Lasso	v_Y	RUSE	rf	['자기자본증가율', '총입원인당 인건비증가율', '총자본이익률', '자기자본순...	[[1817 335]]n [3 32]]	0.845	0.087	0.914	0.159	0.918	{'max_depth': 8, 'min_samples_leaf': 1, 'min_s...}	Y
small_business_df	R	t_Y	None	v_Y	SMOTE	pr	['자기자본증가율', '총입원인당 인건비증가율', '자기자본순이익률', '경영자본...	[[1776 376]]n [5 30]]	0.826	0.074	0.857	0.136	0.889	0	Y
small_business_df	R	t_Y	forward	v_N	RUSE	lg	['총자본사업이익률', '자기자본구성비율', '총입원수', '단기차입금 대 총차입금...', '총자본사업이익률', '자기자본구성비율', '총입원수', '단기차입금 대 총차입금...', '순운전자본회전율', '경영자본회전율', ...]	[[1813 339]]n [3 32]]	0.844	0.086	0.914	0.158	0.925	{'C': 1}	Y
basic_final_df	R	t_Y	forward	v_N	RUSE	knn	['총자본사업이익률', '자기자본구성비율', '순운전자본회전율', '경영자본회전율', ...]	[[2675 569]]n [4 34]]	0.825	0.056	0.895	0.106	0.889	{'metric': 'euclidean', 'n_neighbors': 7, 'wei...}	Y

DataFrame

데이터프레임	train_test_split(Y/R)	t-test(Y/N)	변수선택 법	VIF(Y/N)	sampling 방식	적용 모델	선택된 변수	훈동행렬	Accuracy	Precision	Recall	F1-score	ROC_accuracy	hyperparameter 변경	cut-off
small_business_df	R	t_Y	stepwise	v_N	RUSE	lg	['자기자본구성비율', '총업원수', '단기차입금 대 총차입금비율', 'CASH F...	[[1804 348]]n [2 33]]	0.840	0.087	0.943	0.159	0.930	{'C': 0.001}	Y
diff_df	R	t_Y	None	v_N	RUSE	rf	['growth_rate_수지비율', 'growth_rate_사내유보 대 자기자본비...	[[2386 471]]n [3 27]]	0.836	0.054	0.900	0.102	0.919	{'max_depth': 8, 'min_samples_leaf': 3, 'min_s...}	Y
binning_final_df	R	t_Y	None	v_N	RUSE	svm	['유형자산증가율', '영업이익증가율', '순이익증가율', '재고자산증가율', '총...	[[2612 632]]n [5 33]]	0.806	0.050	0.868	0.094	0.908	{'C': 10, 'degree': 3, 'gamma': 0.001, 'kernel...}	Y
basic_final_df	R	t_Y	None	v_Y	RUSE	rf	['자기자본증가율', '총입원인당 인건비증가율', '자기자본영업이익률', '자기자...	[[2731 513]]n [6 32]]	0.842	0.059	0.842	0.110	0.913	{'max_depth': 6, 'min_samples_leaf': 1, 'min_s...}	Y

6. 모델 정확도 검정 및 최종 모델 선정

02. Hyper-Parameter 대상 선정 및 실행



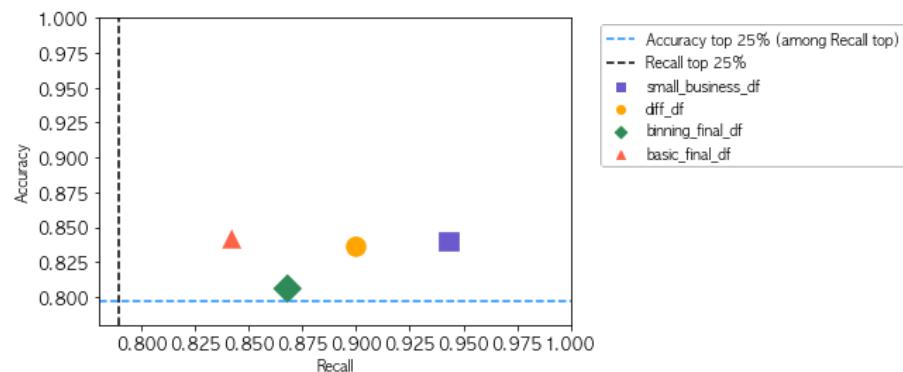
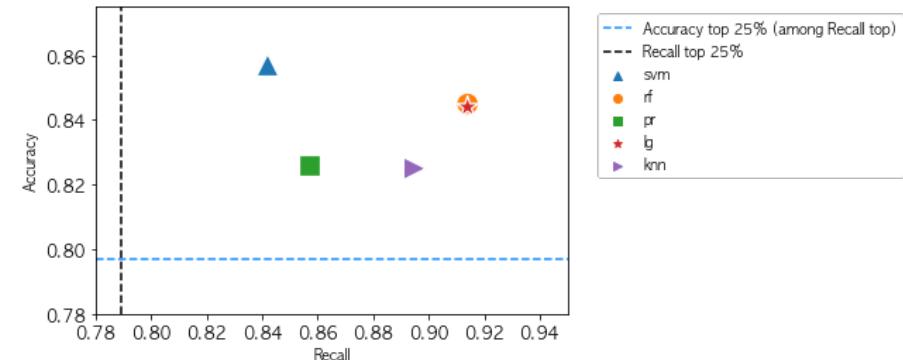
UBION – Trillion

Model

데이터프레임	train_test_split(Y/R)	test(Y/N)	변수선택법	VIF(Y/N)	sampling 방식	적용 모델	선택된 변수	총동행렬	Accuracy	Precision	Recall	F1-score	ROC_accuracy	hyperparameter 설정	cut-off
binning_final_df	R	t,Y	None	v,N	RUSSE	svm	["유형신사업기부", "영업여부", "수익", "임대면적", "재고잔고", "경기장", "..."]	[12782 46210 [8 92]]	0.857	0.065	0.842	0.120	0.915	{'C': 100, 'degree': 1, 'gamma': 0.001, 'kernel': 'rbf'}	Y
small_business_df	R	t,Y	Lasso	v,Y	RUSSE	rf	["유형신사업기부", "영업여부", "수익", "임대면적", "재고잔고", "경기장", "..."]	[11817 33510 [3 32]]	0.845	0.087	0.914	0.159	0.918	{'max_depth': 8, 'min_samples_leaf': 1, 'min_samples_split': 5}	Y
small_business_df	R	t,Y	None	v,Y	SMOTE	pr	["유형신사업기부", "영업여부", "수익", "임대면적", "재고잔고", "경기장", "..."]	[11776 37610 [5 30]]	0.826	0.074	0.857	0.136	0.889	0	Y
small_business_df	R	t,Y	forward	v,N	RUSSE	lg	["유형신사업기부", "영업여부", "수익", "임대면적", "재고잔고", "경기장", "..."]	[11813 33910 [3 32]]	0.844	0.086	0.914	0.158	0.925	{'C': 1}	Y
basic_final_df	R	t,Y	forward	v,N	RUSSE	knn	["총자본사업이익률", "기기장운영수익률", "총판매수익률", "영업대차지영업..."]	[12675 56910 [4 34]]	0.825	0.056	0.895	0.106	0.889	{'metric': 'euclidean', 'n_neighbors': 7, 'weights': 'uniform'}	Y

DataFrame

데이터프레임	train_test_split(Y/R)	test(Y/N)	변수선택법	VIF(Y/N)	sampling 방식	적용 모델	선택된 변수	총동행렬	Accuracy	Precision	Recall	F1-score	ROC_accuracy	hyperparameter 설정	cut-off
small_business_df	R	t,Y	stepwise	v,N	RUSSE	lg	["유형신사업기부", "영업여부", "수익", "임대면적", "재고잔고", "경기장", "..."]	[11904 34810 [2 33]]	0.840	0.087	0.943	0.159	0.930	{'C': 0.001}	Y
diff_df	R	t,Y	None	v,N	RUSSE	rf	["growth_rate_성장률", "growth_rate_사업", "유보 대 차지비율", "..."]	[12388 4710 [3 27]]	0.838	0.054	0.900	0.102	0.919	{'max_depth': 8, 'min_samples_leaf': 3, 'min_samples_split': 5}	Y
binning_final_df	R	t,Y	None	v,N	RUSSE	svm	["유형신사업기부", "영업여부", "수익", "임대면적", "재고잔고", "경기장", "..."]	[12612 63210 [5 33]]	0.806	0.050	0.868	0.094	0.908	{'C': 10, 'degree': 3, 'gamma': 0.001, 'kernel': 'rbf'}	Y
basic_final_df	R	t,Y	None	v,Y	RUSSE	rf	["유형신사업기부", "영업여부", "수익", "임대면적", "재고잔고", "경기장", "..."]	[12731 61310 [6 32]]	0.842	0.059	0.842	0.110	0.913	{'max_depth': 6, 'min_samples_leaf': 1, 'min_samples_split': 5}	Y



6. 모델 정확도 검정 및 최종 모델 선정

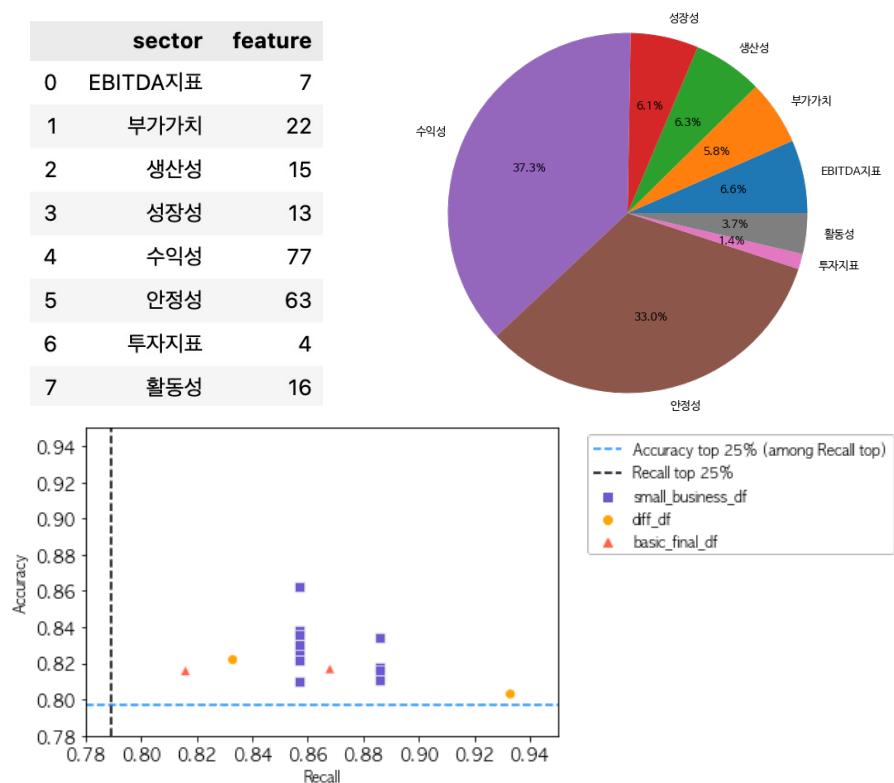
02. Hyper-Parameter 대상 선정 및 실행



UBION – Trillion

데이터프레임	train_test_split(Y/N)	test(Y/N)	변수선택 법	VIF(Y/N)	sampling 방식	적 용 모 델	선택된 변수		총동행률	Accuracy	Precision	Recall	F1-score	hyperparameter 변경	ROC_Score
							['총자본사업이익률', '거기자본구성 비율', '총부원수', '리기자금 대 충당금...']	[1819 333]n [4 31]							
0 small_business_df	R	t_Y	forward	v_N	RUSE	lg	['총자본사업이익률', '거기자본구성 비율', '총부원수', '리기자금 대 충당금...']	[1819 333]n [4 31]	0.846	0.085	0.886	0.155	none	0.933643	
1 small_business_df	R	t_Y	stepwise	v_N	RUSE	lg	['거기자본구성비율', '총업입수', '단기자금 대 충당금...']	[1766 386]n [5 30]	0.821	0.072	0.857	0.133	none	0.931851	
2 small_business_df	R	t_Y	stepwise	v_N	SMOTE	lg	['거기자본구성비율', '총업입수', '단기자금 대 충당금...']	[1782 370]n [4 31]	0.829	0.077	0.886	0.142	none	0.929036	
3 small_business_df	R	t_Y	forward	v_N	ROSE	lg	['총자본사업이익률', '거기자본구성 비율', '총부원수', '리기자금 대 충당금...']	[1778 374]n [3 32]	0.828	0.079	0.914	0.145	none	0.927430	
4 small_business_df	R	t_Y	forward	v_N	SMOTE	lg	['총자본사업이익률', '거기자본구성 비율', '총부원수', '리기자금 대 충당금...']	[1802 350]n [4 31]	0.838	0.081	0.886	0.149	none	0.926633	
5 small_business_df	R	t_Y	stepwise	v_N	ROSE	lg	['거기자본구성비율', '총업입수', '단기자금 대 충당금...']	[1774 378]n [4 31]	0.825	0.076	0.886	0.140	none	0.924907	
6 small_business_df	R	t_Y	Lasso	v_N	SMOTE	lg	['거기자본증가율', '총영업이익률', '자...']	[1786 365]n [4 32]	0.831	0.080	0.914	0.148	none	0.920101	
7 small_business_df	R	t_Y	None	v_N	SMOTE	lg	['거기자본증가율', '총영업이익률', '자...']	[1784 393]n [4 31]	0.818	0.073	0.886	0.135	none	0.919530	
8 small_business_df	R	t_Y	Lasso	v_N	RUSE	lg	['거기자본증가율', '총영업이익률', '자...']	[1780 392]n [4 31]	0.819	0.073	0.886	0.135	none	0.919198	
9 diff_df	R	t_Y	None	v_N	RUSE	rf	['growth_rate_수익률', 'growth_rate_내수영업 대...']	[2312 545]n [4 26]	0.810	0.046	0.867	0.087	none	0.917758	
10 small_business_df	R	t_Y	Lasso	v_Y	RUSE	rf	['growth_rate_수익률', '총영업이익률', '자...']	[2637 404]n [3 32]	0.814	0.073	0.914	0.136	none	0.916470	
11 basic_final_df	R	t_Y	None	v_Y	RUSE	rf	['건비증가율', '거기자본영업이익률', '자...']	[2637 607]n [5 33]	0.814	0.052	0.868	0.097	none	0.912872	
12 small_business_df	R	t_Y	None	v_Y	RUSE	rf	['거기자본증가율', '총영업이익률', '자...']	[1743 409]n [4 31]	0.811	0.070	0.886	0.131	none	0.912354	
13 diff_df	R	t_Y	Lasso	v_Y	RUSE	rf	['growth_rate_수익률', 'growth_rate_내수영업 대...']	[2347 510]n [4 26]	0.822	0.049	0.867	0.092	none	0.911376	
14 small_business_df	R	t_Y	None	v_N	ROSE	lg	['거기자본증가율', '총영업이익률', '자...']	[1852 300]n [6 29]	0.860	0.088	0.829	0.159	none	0.909891	
15 basic_final_df	R	t_Y	stepwise	v_N	RUSE	rf	['거기자본구성비율', '순운전자본회전율', '전...']	[2676 568]n [6 32]	0.825	0.053	0.842	0.100	none	0.906167	

sector	feature
0 EBITDA지표	7
1 부가가치	22
2 생산성	15
3 성장성	13
4 수익성	77
5 안정성	63
6 투자지표	4
7 활동성	16



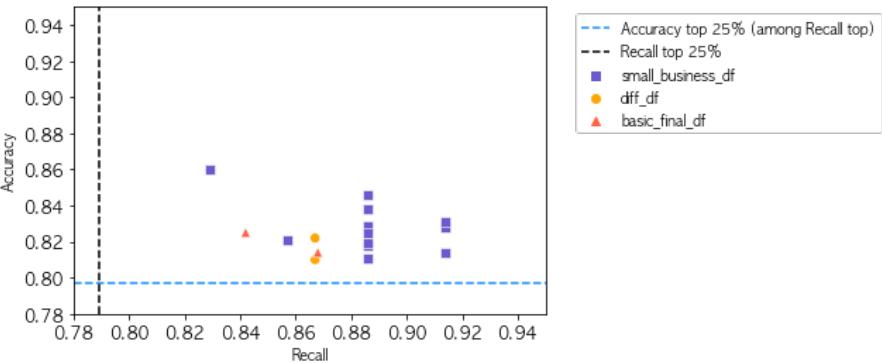
6. 모델 정확도 검정 및 최종 모델 선정

02. Hyper-Parameter 대상 선정 및 실행



UBION – Trillion

데이터프레임	train_test_split(Y/R)	test(Y/N)	변수선택 법	VIF(Y/N)	sampling 방식	적용 모델	선택된 변수		총총행렬	Accuracy	Precision	Recall	F1-score	ROC_AUC	hyperparameter 설정	cut-off
							[[1792 360][n [4 31]]]	['총지분사업이익률', '자기 자본구성비율', '총영업수', '단기사업금 대 총자산률...']								
0 small_business_df	R	t_Y	forward	v_N	RUSE	lg	[[1768 384][n [5 30]]]	['자자자본구성비율', '총영 업수', '단기사업금 대 총자 산률', 'CASH...']	[[1768 373][n [5 30]]]	0.822	0.072	0.857	0.134	0.915	{'C': 0.1}	N
1 small_business_df	R	t_Y	stepwise	v_N	RUSE	lg	[[1792 350][n [5 30]]]	['자자자본구성비율', '총영 업수', '단기사업금 대 총자 산률', 'CASH...']	[[1792 352][n [5 30]]]	0.838	0.079	0.857	0.145	0.928	{'C': 0.1}	N
2 small_business_df	R	t_Y	stepwise	v_N	SMOTE	lg	[[1792 350][n [5 30]]]	['총지분사업이익률', '자기 자본구성비율', '총영 업수', '단기사업금 대 총자 산률', 'CASH...']	[[1792 352][n [5 30]]]	0.827	0.074	0.857	0.137	0.925	{'C': 0.01}	N
3 small_business_df	R	t_Y	forward	v_N	ROSE	lg	[[1779 373][n [5 30]]]	['총지분사업이익률', '자기 자본구성비율', '총영 업수', '단기사업금 대 총자 산률', 'CASH...']	[[1792 352][n [5 30]]]	0.836	0.078	0.857	0.144	0.928	{'C': 0.01}	N
4 small_business_df	R	t_Y	forward	v_N	SMOTE	lg	[[1792 352][n [5 30]]]	['총지분사업이익률', '자기 자본구성비율', '총영 업수', '단기사업금 대 총자 산률', 'CASH...']	[[1785 367][n [5 30]]]	0.830	0.076	0.857	0.139	0.926	{'C': 0.01}	N
5 small_business_df	R	t_Y	stepwise	v_N	ROSE	lg	[[1785 367][n [5 30]]]	['자자자본구성비율', '총영 업수', '단기사업금 대 총자 산률', 'CASH...']	[[1802 350][n [5 30]]]	0.862	0.092	0.857	0.166	0.919	{'C': 0.1}	N
6 small_business_df	R	t_Y	Lasso	v_N	SMOTE	lg	[[1855 297][n [5 30]]]	['자자자본구성비율', '총영 업수', '단기사업금 대 총자 산률', '자자본...']	[[1759 393][n [4 31]]]	0.818	0.073	0.866	0.135	0.916	{'C': 0.001}	N
7 small_business_df	R	t_Y	None	v_N	SMOTE	lg	[[1759 393][n [4 31]]]	['자자자본구성비율', '총영 업수', '단기사업금 대 총자 산률', '자자본...']	[[1754 390][n [4 31]]]	0.816	0.072	0.866	0.134	0.918	{'C': 0.001}	N
8 small_business_df	R	t_Y	Lasso	v_N	RUSE	lg	[[1754 390][n [4 31]]]	['자자자본구성비율', '총영 업수', '단기사업금 대 총자 산률', '자자본...']	[[2369 568][n [2 28]]]	0.803	0.047	0.933	0.089	0.919	{'max_depth': 4, 'growth_rate': 0.001, 'min_samples_leaf': 4, 'min_samples_split': 2}	N
9 diff_df	R	t_Y	None	v_N	RUSE	rf	[[1887 485][n [3 32]]]	['자자자본구성비율', '총영 업수', '단기사업금 대 총자 산률', '자자본...']	[[2347 510][n [6 25]]]	0.777	0.062	0.914	0.116	0.923	{'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 1}	N
10 small_business_df	R	t_Y	Lasso	v_Y	RUSE	rf	[[2650 594][n [5 33]]]	['자자자본구성비율', '총영 업수', '단기사업금 대 총자 산률', '자자본...']	[[1743 409][n [4 31]]]	0.817	0.053	0.868	0.099	0.911	{'max_depth': 3, 'min_samples_leaf': 1, 'min_samples_split': 1}	N
11 basic_final_df	R	t_Y	None	v_Y	RUSE	rf	[[2650 594][n [5 33]]]	['자자자본구성비율', '총영 업수', '단기사업금 대 총자 산률', '자자본...']	[[2347 510][n [6 25]]]	0.811	0.070	0.866	0.131	0.916	{'max_depth': 8, 'min_samples_leaf': 1, 'min_samples_split': 1}	N
12 small_business_df	R	t_Y	None	v_Y	RUSE	rf	[[2347 510][n [6 25]]]	['growth_rate': 0.001, 'max_depth': 12, 'min_samples_leaf': 1, 'min_samples_split': 1]	[[1741 411][n [5 30]]]	0.810	0.068	0.857	0.126	0.910	{'C': 0.001}	N
13 diff_df	R	t_Y	Lasso	v_Y	RUSE	rf	[[2347 510][n [6 25]]]	['자자자본구성비율', '총영 업수', '단기사업금 대 총자 산률', '자자본...']	[[2347 510][n [6 25]]]	0.822	0.047	0.833	0.088	0.891	{'max_depth': 12, 'min_samples_leaf': 1, 'min_samples_split': 1}	N
14 small_business_df	R	t_Y	None	v_N	ROSE	lg	[[1741 411][n [5 30]]]	['자자자본구성비율', '총영 업수', '단기사업금 대 총자 산률', '자자본...']	[[2646 598][n [7 31]]]	0.816	0.049	0.816	0.093	0.915	{'max_depth': 8, 'min_samples_leaf': 6, 'min_samples_split': 1}	N
15 basic_final_df	R	t_Y	stepwise	v_N	RUSE	rf	[[2646 598][n [7 31]]]	['전기본회전율', '경쟁자회사 진율', '이자부담률', ...]	[[2646 598][n [7 31]]]	0.816	0.049	0.816	0.093	0.915	{'max_depth': 8, 'min_samples_leaf': 6, 'min_samples_split': 1}	N



6. 모델 정확도 검정 및 최종 모델 선정



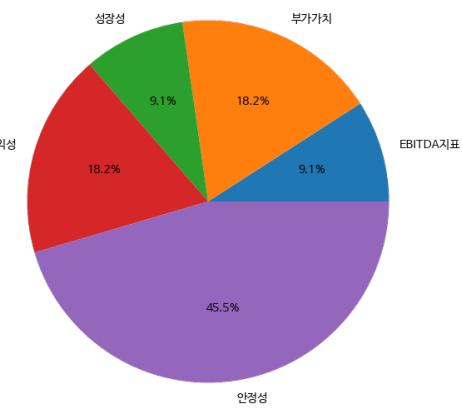
03. 최종 모델 선정

UBION – Trillion

5	small_business_df	R	t_Y	stepwise	v_N	ROSE	lg	['자기자본구성비율', '종업원수', '단기차입금 대 총차입금비율', 'CASH F...']	[[1785 367]\n [5 30]]	0.830	0.076	0.857	0.139	0.926	{'C': 0.01}	N
---	-------------------	---	-----	----------	-----	------	----	---	--------------------------	-------	-------	-------	-------	-------	-------------	---

Logit Regression Results						
Dep. Variable:	차기부도여부	No. Observations:	19870 <th data-cs="3" data-kind="parent"></th> <th data-kind="ghost"></th> <th data-kind="ghost"></th>			
Model:	Logit	Df Residuals:	19859			
Method:	MLE	Df Model:	10			
Date:	Tue, 21 Dec 2021	Pseudo R-squ.:	0.4265			
Time:	04:52:18	Log-Likelihood:	-7898.4			
converged:	True	LL-Null:	-13773.			
Covariance Type:	nonrobust	LLR p-value:	0.000			
	coef	std err	z	P> z	[0.025	0.975]
자기자본구성비율	-0.0088	0.001	-13.917	0.000	-0.010	-0.008
종업원수	-0.0053	0.000	-14.084	0.000	-0.006	-0.005
단기차입금 대 총차입금비율	0.0103	0.001	17.873	0.000	0.009	0.011
CASH FLOW 대 부채비율	-0.0025	0.000	-5.750	0.000	-0.003	-0.002
EBITDA(백만원)	-7.62e-05	3.91e-06	-19.469	0.000	-8.39e-05	-6.85e-05
재고자산 대 유동자산비율	-0.0171	0.001	-13.007	0.000	-0.020	-0.015
인건비(백만원)	-0.0002	1.43e-05	-14.069	0.000	-0.000	-0.000
자기자본증가율	0.0006	0.000	4.407	0.000	0.000	0.001
타인자본구성비율	0.0167	0.001	14.660	0.000	0.014	0.019
총자본순이익률	-0.1935	0.015	-13.078	0.000	-0.223	-0.165
기업순이익률	0.1670	0.015	11.220	0.000	0.138	0.196

	feature	sector
0	자기자본구성비율	안정성
1	종업원수	부가가치
2	단기차입금 대 총차입금비율	안정성
3	CASH FLOW 대 부채비율	안정성
4	EBITDA(백만원)	EBITDA지표
5	재고자산 대 유동자산비율	안정성
6	인건비(백만원)	부가가치
7	자기자본증가율	성장성
8	타인자본구성비율	안정성
9	총자본순이익률	수익성
10	기업순이익률	수익성



6. 모델 정확도 검정 및 최종 모델 선정

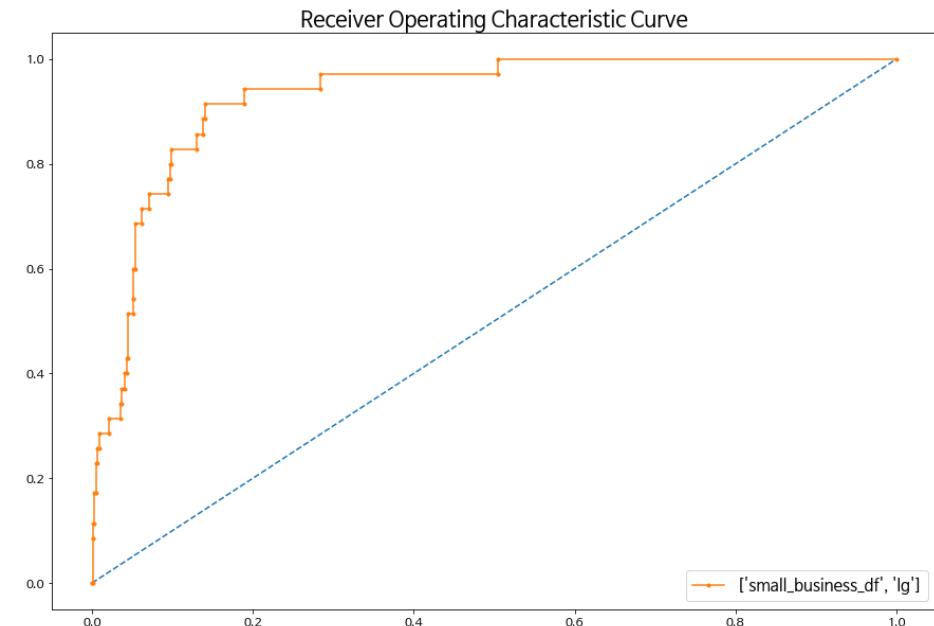


03. 최종 모델 선정

UBION – Trillion

5	small_business_df	R	t_Y	stepwise	v_N	ROSE	lg	['자기자본구성비율', '종업원수', '단기차입금 대 총차입금비율', 'CASH F...']	[[1785 367]\n[5 30]]	0.830	0.076	0.857	0.139	0.926	{'C': 0.01}	N
---	-------------------	---	-----	----------	-----	------	----	---	-----------------------------	-------	-------	-------	-------	-------	-------------	---

Logit Regression Results						
Dep. Variable:	차기부도여부	No. Observations:	19870 <th data-cs="3" data-kind="parent"></th> <th data-kind="ghost"></th> <th data-kind="ghost"></th>			
Model:	Logit	Df Residuals:	19859			
Method:	MLE	Df Model:	10			
Date:	Tue, 21 Dec 2021	Pseudo R-squ.:	0.4265			
Time:	04:52:18	Log-Likelihood:	-7898.4			
converged:	True	LL-Null:	-13773.			
Covariance Type:	nonrobust	LLR p-value:	0.000			
	coef	std err	z	P> z	[0.025	0.975]
자기자본구성비율	-0.0088	0.001	-13.917	0.000	-0.010	-0.008
종업원수	-0.0053	0.000	-14.084	0.000	-0.006	-0.005
단기차입금 대 총차입금비율	0.0103	0.001	17.873	0.000	0.009	0.011
CASH FLOW 대 부채비율	-0.0025	0.000	-5.750	0.000	-0.003	-0.002
EBITDA(백만원)	-7.62e-05	3.91e-06	-19.469	0.000	-8.39e-05	-6.85e-05
재고자산 대 유동자산비율	-0.0171	0.001	-13.007	0.000	-0.020	-0.015
인건비(백만원)	-0.0002	1.43e-05	-14.069	0.000	-0.000	-0.000
자기자본증가율	0.0006	0.000	4.407	0.000	0.000	0.001
타인자본구성비율	0.0167	0.001	14.660	0.000	0.014	0.019
총자본순이익률	-0.1935	0.015	-13.078	0.000	-0.223	-0.165
기업순이익률	0.1670	0.015	11.220	0.000	0.138	0.196



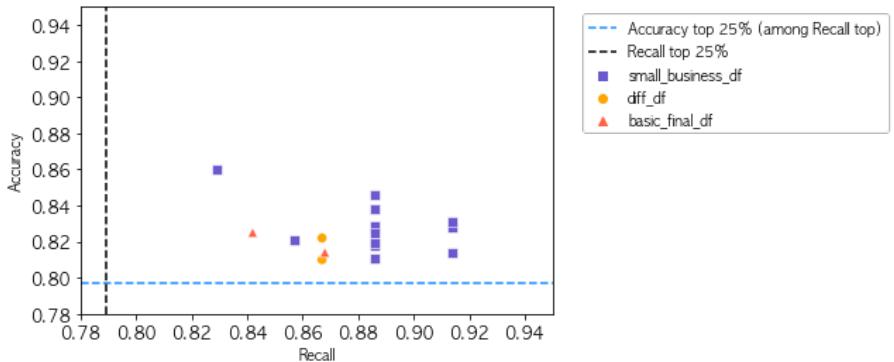
6. 모델 정확도 검정 및 최종 모델 선정



02. Hyper-Parameter 대상 선정 및 실행

UBION – Trillion

	데이터프레임	train_test_split(Y R)	t-test(Y N)	변수선택법	VIF(Y N)	sampling 방식	적용 모델	선택된 변수	총동행렬	Accuracy	Precision	Recall	F1-score	ROC_AUC	hyperparameter 설정
0	small_business_df	R	t_Y	forward	v_N	RUSE	lg	['자기부분기여율', '기기...']	[1792 360][n [4 31]]	0.834	0.079	0.886	0.146	0.924	{'C': 0.1} N
1	small_business_df	R	t_Y	stepwise	v_N	RUSE	lg	['자기부분기여율', '총입...']	[1768 384][n [5 30]]	0.822	0.072	0.857	0.134	0.915	{'C': 0.1} N
2	small_business_df	R	t_Y	stepwise	v_N	SMOTE	lg	['자기부분기여율', '총입...']	[1782 367][n [5 30]]	0.838	0.079	0.857	0.145	0.928	{'C': 0.1} N
3	small_business_df	R	t_Y	forward	v_N	ROSE	lg	['자기부분기여율', '기기...']	[1770 373][n [5 30]]	0.827	0.074	0.857	0.137	0.925	{'C': 0.01} N
4	small_business_df	R	t_Y	forward	v_N	SMOTE	lg	['자기부분기여율', '기기...']	[1799 353][n [5 30]]	0.836	0.078	0.857	0.144	0.928	{'C': 0.01} N
5	small_business_df	R	t_Y	stepwise	v_N	ROSE	lg	['자기부분기여율', '총입...']	[1785 367][n [5 30]]	0.830	0.076	0.857	0.139	0.926	{'C': 0.01} N
6	small_business_df	R	t_Y	Lasso	v_N	SMOTE	lg	['자기부분기여율', '총입...']	[1855 297][n [5 30]]	0.862	0.092	0.857	0.166	0.919	{'C': 0.1} N
7	small_business_df	R	t_Y	None	v_N	SMOTE	lg	['자기부분기여율', '기기...']	[1759 393][n [4 31]]	0.818	0.073	0.886	0.135	0.916	{'C': 0.001} N
8	small_business_df	R	t_Y	Lasso	v_N	RUSE	lg	['자기부분기여율', '기기...']	[1754 398][n [4 31]]	0.816	0.072	0.886	0.134	0.918	{'C': 0.001} N
9	diff_df	R	t_Y	None	v_N	RUSE	rf	['growth_rate', '수익률...', 'growth_rate_사내보유...']	[1289 455][n [2 28]]	0.803	0.047	0.933	0.089	0.919	{'max_depth': 4, 'min_samples_leaf': 4, 'min_s...'} N
10	small_business_df	R	t_Y	Lasso	v_Y	RUSE	rf	['자기부분기여율', '총입원...']	[1667 485][n [3 32]]	0.777	0.062	0.914	0.116	0.923	{'max_depth': 10, 'min_samples_leaf': 1, 'min_s...'} N
11	basic_final_df	R	t_Y	None	v_Y	RUSE	rf	['자기부분기여율', '기기...']	[2650 594][n [5 30]]	0.817	0.053	0.868	0.099	0.911	{'max_depth': 3, 'min_samples_leaf': 1, 'min_s...'} N
12	small_business_df	R	t_Y	None	v_Y	RUSE	rf	['자기부분기여율', '기기...']	[1743 500][n [4 31]]	0.811	0.070	0.886	0.131	0.916	{'max_depth': 8, 'min_samples_leaf': 1, 'min_s...'} N
13	diff_df	R	t_Y	Lasso	v_Y	RUSE	rf	['growth_rate', '사내보유...', 'growth_rate_사내보유...']	[1347 510][n [5 25]]	0.822	0.047	0.833	0.088	0.891	{'max_depth': 12, 'min_samples_leaf': 1, 'min_s...'} N
14	small_business_df	R	t_Y	None	v_N	ROSE	lg	['자기부분기여율', '기기...']	[1741 413][n [5 30]]	0.810	0.068	0.857	0.126	0.910	{'C': 0.001} N
15	basic_final_df	R	t_Y	stepwise	v_N	RUSE	rf	['자기부분기여율', '수익...']	[2646 589][n [7 31]]	0.816	0.049	0.816	0.093	0.915	{'max_depth': 8, 'min_samples_leaf': 6, 'min_s...'} N



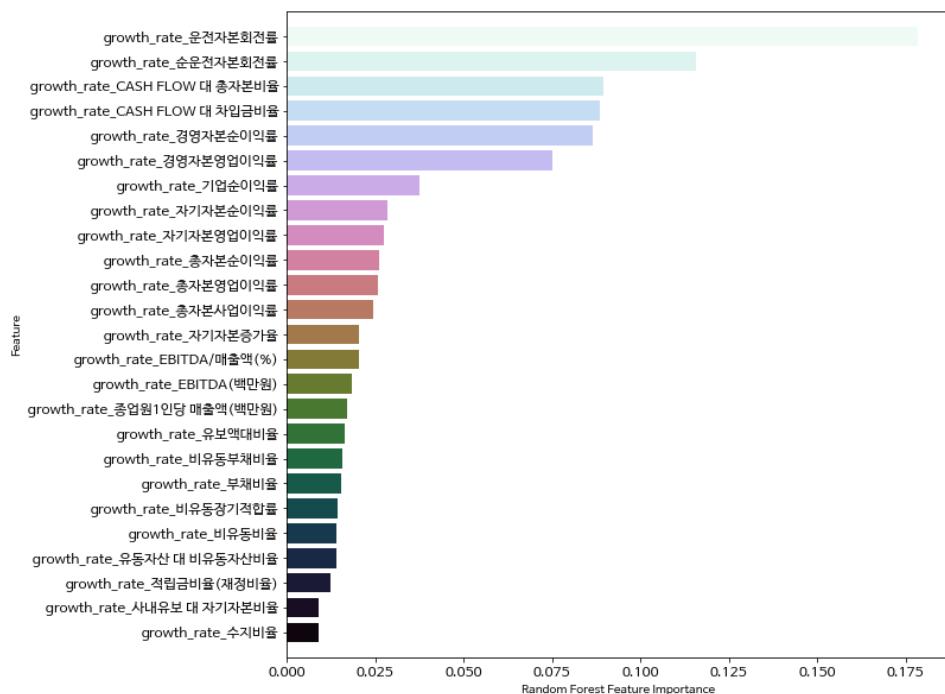
6. 모델 정확도 검정 및 최종 모델 선정

03. 최종 모델 선정

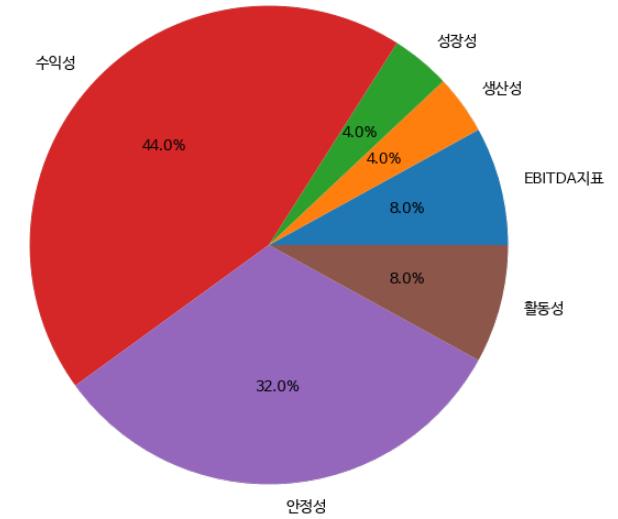


UBION – Trillion

9	diff_df	R	t_Y	None	v_N	RUSE	rf	['growth_rate_수지비율', 'growth_rate_사내유보 대 자기자본비율'] [[2289 568]ln [2 28]]	0.803	0.047	0.933	0.089	0.919	{'max_depth': 4, 'min_samples_leaf': 4, 'min_s...}	N
---	---------	---	-----	------	-----	------	----	---	-------	-------	-------	-------	-------	--	---



feature	sector
0	수지비율
1	수익성
2	수익성
3	수익성
4	수익성
5	수익성
6	수익성
7	수익성
8	수익성
9	수익성
10	수익성
11	수익성
12	수익성
13	수익성
14	수익성
15	수익성
16	수익성
17	수익성
18	수익성
19	수익성
20	수익성
21	수익성
22	수익성
23	수익성
24	수익성



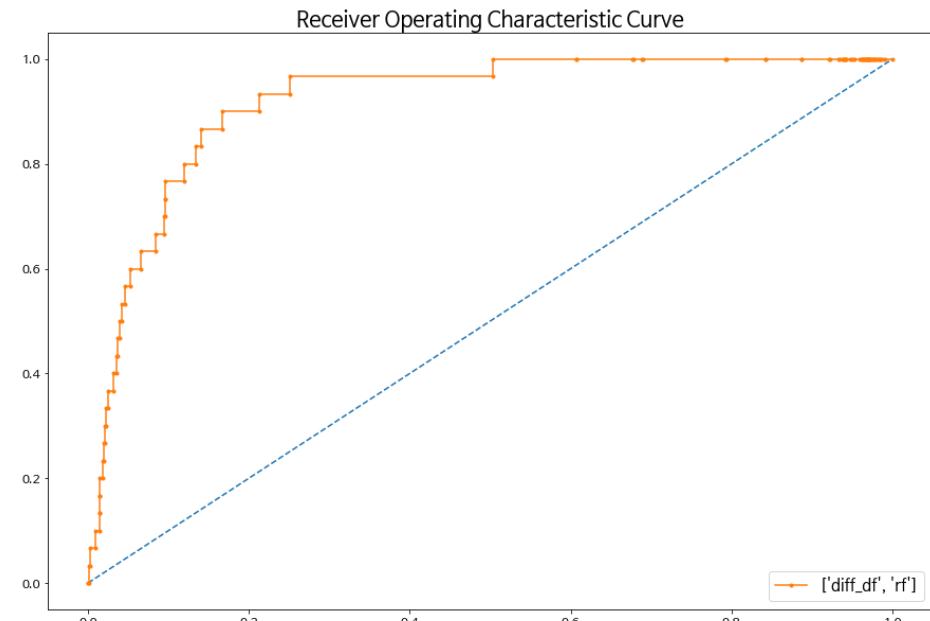
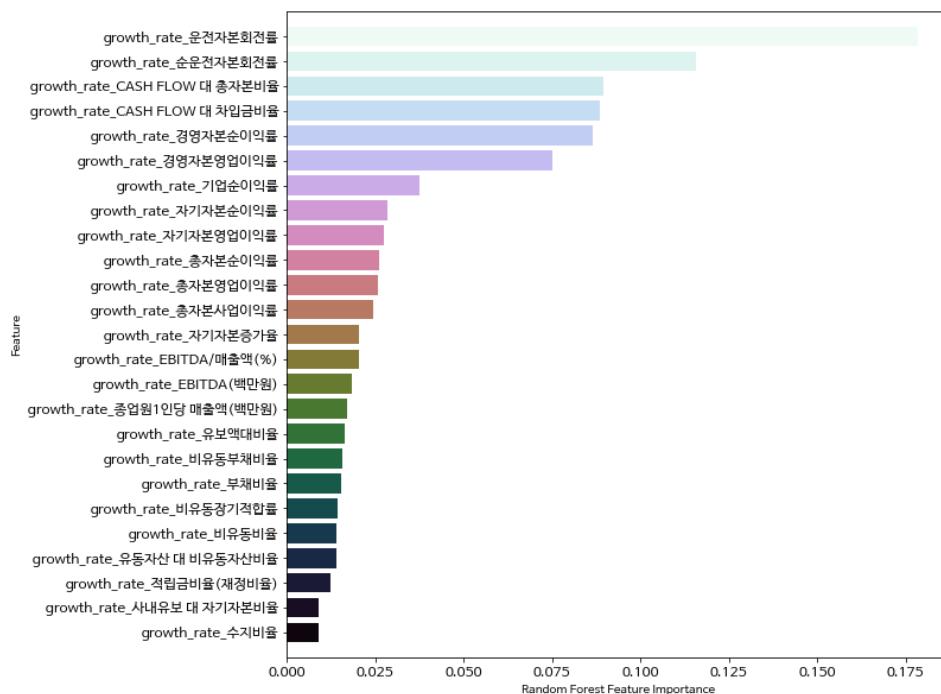
6. 모델 정확도 검정 및 최종 모델 선정



03. 최종 모델 선정

UBION – Trillion

9	diff_df	R	t_Y	None	v_N	RUSE	rf	['growth_rate_수지비율', 'growth_rate_사내유보 대 자기자본비... [[2289 568]n [2 28]]]	0.803	0.047	0.933	0.089	0.919	{'max_depth': 4, 'min_samples_leaf': 4, 'min_s...}	N
---	---------	---	-----	------	-----	------	----	---	-------	-------	-------	-------	-------	--	---



6. 모델 정확도 검정 및 최종 모델 선정



04. 의의 및 한계점

UBION – Trillion

의의



빅데이터 시대를 맞이하여 다양한 데이터를 활용한 부도 예측 알고리즘 비교 연구로서 의미가 있음

Train, Test Split이 기간 기준이 아닌 Random Split을 적용할 경우도 유의한 결과가 도출됨

재무비율의 변화율을 반영한 differential data에서 예측 성능 향상이 가능

상시 종업원 수 300인 이하의 중소기업 Data의 성능이 좋은 편

부도를 결정하는 Feature는 안정성 및 수익성 지표가 높은 편

감사 의견 코드를 활용하여 부도를 새롭게 정의한 점

한계



재무비율만 이용한 탓에 알트만 X4를 확인하기 어려웠음 향후 재무비율 + 재무정보를 같이 활용한 분석의 필요성 존재

결산 기준 2년 후를 예측하여 대출 기간이 짧은 운전자금 대출에 대한 리스크 대비에 미흡함 부도 예측 시점의 다양화 필요

상장 기업만 대상으로 진행했기에, 선행 논문에 비해 오히려 Data 규모가 작음 외부감사대상기업 + 개인사업자 등 다양한 Data 분석 필요성 존재

향후 과제





참고문헌



- [1] 박종원, 안성만(2014), “재무비율을 이용한 부도예측에 대한 연구 : 한국의 외부감사기업을 대상으로”, 경영학연구 43(3), 2014.6, 639-669(31 pages)
- [2] 이병윤(2021), “국내은행 리스크관리 강화 필요”, 금융 포커스 30권 19호
- [3] 한국상장회사협의회 (2008), “상장협 DATABASE 사용설명서”, 동 협회
- [4] Altman, E.I.(1968), “Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy,”. Journal of Finance 23, 589- 609.
- [5] Bartlett, M. S. (1937). Properties of Sufficiency and Statistical Tests. Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, Vol. 160, No.901, 268-282.
- [6] 법인세법 시행규칙 제15조 제3항 관련 업종별자산의 기준내용연수 및 내용연수 범위표
- [7] “주권상장”, 한국거래소, 2021년 11월 22일 접속, <https://listing.krx.co.kr/contents/LST/04/04010101/LST04010101.jsp>
- [8] “중견기업의 정의”, 기업금융나들목, 2021년 12월 15일 접속, <https://www.smefn.or.kr:4443/gs04/quide/enterdefine.jsp>
- [9] “TS2000”, “KOCOinfo, 2021년 11월 22일 접속, <http://www.kocoinfo.com>
- [10] “Federal Funds Effective Rate”, FRED, Accessed Dec 18, 2021, <https://fred.stlouisfed.org/series/DFF>
- [11] “Tuning the hyper-parameters of an estimator”, Scikit-learn, Access Dec 18, 2021, https://scikit-learn.org/stable/modules/grid_search.html

분석 도구



Trillion (1조)

재무비율을 이용한 상장기업 부도 예측 알고리즘 비교 연구

UBION Bankruptcy Financial Report

정기호 이주희 신문혁 윤영주

Trillion (1조)

재무비율을 이용한 상장기업 부도 예측 알고리즘 비교 연구



정기호 이주희 신문혁 윤영주