

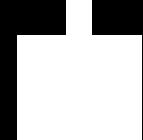
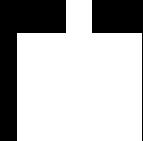
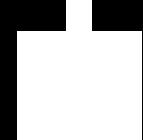
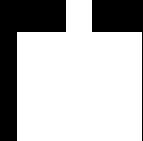
빅데이터 기반 머신러닝을 활용한 Kosdaq 관리종목 지정 예측

- 재무데이터와 뉴스 및 채용플랫폼 데이터 활용 -

컴파스 (Company Profiler Squad) 팀



CONTENTS

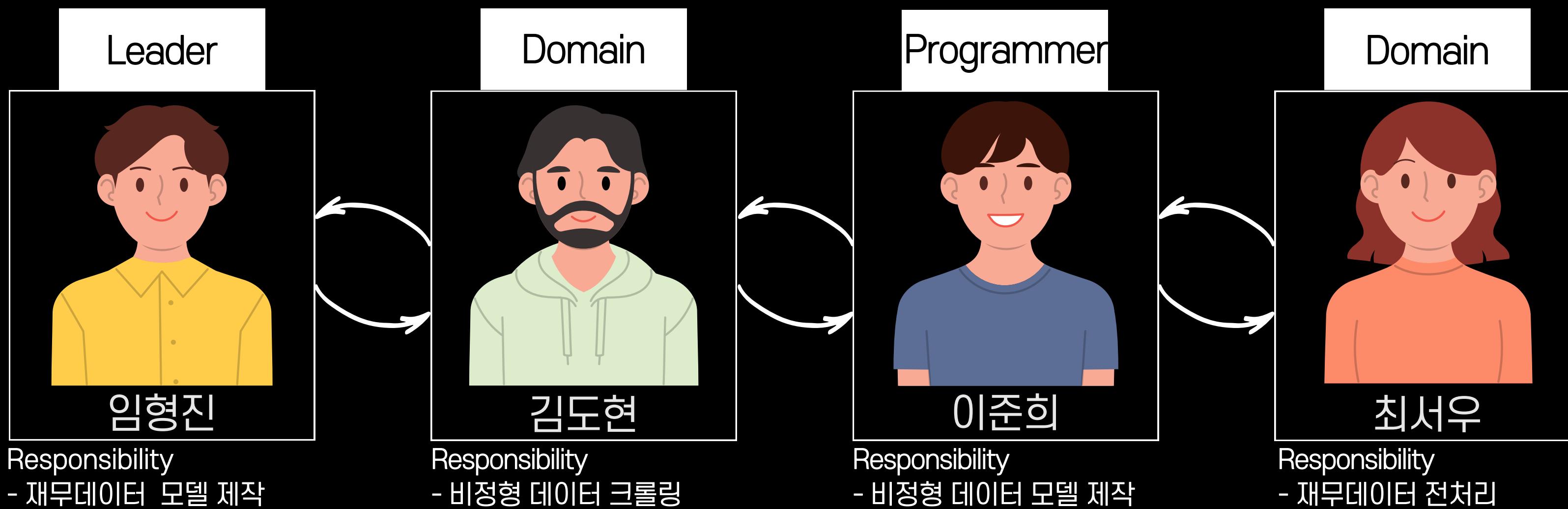
-  01. 프로젝트 개요
-  02. 프로젝트 결과 요약
-  03. 진행과정 - 1) 재무데이터
-  04. 진행과정 - 2) 비정형 데이터
-  05. 결론 도출 (의의 및 한계점)

03 Team Member

프로젝트 개요

컴.파.스 (Company Profiler Squad) :

4명이 한 팀을 이루어 **프로파일러**처럼 면밀한 분석으로
관리종목지정 기업을 찾아낸다는 뜻



04 Timetable

프로젝트 기간 : 9월 1일 - 10월 7일

프로젝트 개요



05 Motivation



재무 데이터를 활용한 부실예측의 한계

→ 적시성이 떨어지는 문제

→ 화폐단위로 측정할 수 있는 정보만 반영 가능

→ 기존 논문의 적시성 보완 방법
= 비정형데이터 중 뉴스, SNS



프로젝트의 차별점
"채용플랫폼" 데이터의 활용

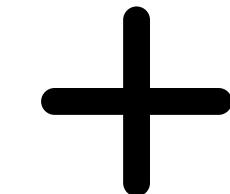
- 기존연구에는 없는 채용플랫폼 데이터를 활용한 부실예측모형으로 부실예측수준의 향상이 가능한지 확인
- 근로자들의 회사만족도 등의 기업내부의 반응이 기업부실과 연관이 있는지 확인하는 것이 목표

프로젝트 개요



비정형데이터

참고 논문:
머신러닝 기반 자연어 처리를 이용한
관리종목 예측에 관한 연구
(연세대 Learning Machine팀, 2021)



6개 모델 /
3개 데이터셋

참고 논문:
빅데이터와 인공지능 기법을
이용한 기업 부도예측 연구
(최정원 외 2명, 2017)

#관리종목 # 총화표본추출 #비정형데이터 #채용플랫폼

“창의적인 데이터의 활용과 다양한 모델의 비교”

재무데이터 모델

	Accuracy	Recall	Precision	F1-Score
Rogistic	0.9125	0.8500	0.9714	0.9067
SVM	0.9250	0.8625	0.9857	0.9200
Decision Tree	0.7625	0.7000	0.8000	0.7467
Stackin g Ensemble	0.9	0.8375	0.9571	0.8933
DNN	0.9438	0.9494	0.9375	0.9434
LSTM	0.6688	0.6667	0.6750	0.6708

VS

뉴스데이터 모델

재무+뉴스 모델2	Accuracy	Recall	Precision	F1-score
Logit	0.909	0.9667	0.8529	0.9063
SVM	0.8636	0.9629	0.7647	0.8525
Decision Tree	0.803	0.8182	0.7941	0.8059
Stacking Emsenble	0.8636	0.9032	0.8235	0.8615
DNN	0.8939	0.8824	0.9091	0.8955
LSTM	0.7576	0.9706	0.6875	0.8049

채용플랫폼 데이터 모델

재무+플랫폼 모델2	Accuracy	Recall	Precision	F1-score
Logit	0.9014	0.9608	0.9074	0.9333
SVM	0.8873	0.9792	0.8704	0.9216
Decision Tree	0.8732	1	0.8333	0.9091
Stacking Emsenble	0.8873	0.8966	0.9629	0.9286
DNN	0.9296	0.9444	0.9623	0.9533
LSTM	0.5634	0.5	0.8709	0.6353

부실기업의 정의

: Kosdaq 관리종목

관리종목이란? 상장법인이 갖추어야 할 최소한도의 유동성을 갖추지 못하였거나, 영업실적 악화 등의 사유로 부실이 심화된 종목으로 **상장폐지기준에 해당할 우려가 있는 종목**

관리종목지정 기업을 부실기업으로 정의한 이유:

관리종목으로 지정됨은 과거 사례와 연구들에 따르면 주식시장에서는 악재로 인식이 되며, 이는 타 기업들에 비하여 불안정한 구조를 가진 부실기업이라고 정의할 수 있음

손성규·오명전(2008) "관리종목 기업의 회계정보 효과"

김태혁·엄철준(1997) "관리대상종목의 수익률과 위험 속성에 관한 연구"

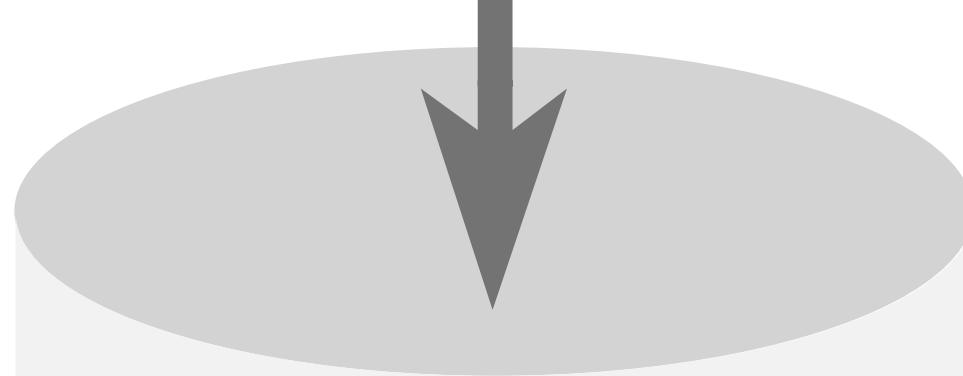
분석과정 체계도

진행과정

수집 및 전처리

2012~2020 코스닥시장 관리종목

- 데이터 출처: ts2000, 네이버, 채용플랫폼(잡플래닛)



- 후보변수 (X) :
 - Data set1 : 재무데이터 (40 개)
 - Data set2 : 재무데이터 (40 개) + 뉴스 데이터
 - Data set3 : 재무데이터 (40 개) + 채용 플랫폼
- 종속변수 (Y) : 관리종목 지정 여부

*재무 : 관리종목 지정일 직전 사업보고서

*비정형 : 관리종목 지정일 직전 6개월 데이터

변수 선정 및 모델링

총화표본추출

- 1) 기업규모
- 2) 업종구분

EDA
탐색적 분석

(1) 재무 데이터

검정 통계량 확인

유의한 변수 선정

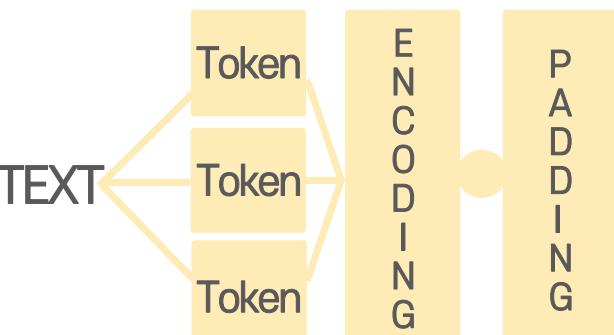
다중공선성 제거

Feature Selection

최종변수: 9개

Data Set 1, 2, 3

(2) 비정형 데이터



데이터 정제 과정
(TF-IDF), Word2Vec

Learning Algorithms

로지스틱 회귀분석

Decision Tree

SVM

LSTM DNN

Stacking Ensemble

예측 검정 및 평가

Train set

2012-2018

관리종목: 208개

Test set

2019-2020

관리종목: 80개

Resampling(총화표본추출)
정상기업 : 부실기업
1:1
2:1

평가 검정

정확도 (Accuracy)

Recall

Precision

F1 score

ROC AUC 그래프

오차행렬

최적의 부도 예측 모형 선정

09 코스닥 관리종목

진행과정

: 데이터 수집 기준

업종) ts2000 기준 제조업 대상 (금융 및 보험업 제외)

사업보고서 결산일) 12월 (K-IFRS 기준)

관리종목지정 기업 출처) KIND

01 수집 기간 ▶

* 한번 관리종목으로 지정되면 이후 데이터 삭제

관리종목지정 -----> X 관리종목 재지정



Trainset (2012-2018) - 208개

2019.04.01 ~ 2020.03.31

(상장폐지 기업 포함)

Testset (2019-2020) - 80개

2012.04.01 ~ 2019.03.31

(상장폐지 기업 포함)

02 제외기준 ▶ 국내상장된 외국기업 / SPAC주 제외

10 코스닥 관리종목 : 데이터 수집 기준

진행과정

<관리종목 지정요건 중 부실기업 선정기준>

구분	상세 요건
매출액	최근사업년도 30억원 미만
법인세차감전 계속사업손실	자기자본 50%초과(& 10억원이상) 계속사업손실 최근 3년간 2회 이상
장기영업손실	4년 연속 영업손실 발생
자본잠식 등	(A)사업연도(반기)말 자본잠식률 50%이상 / (B)사업연도(반기)말 자기자본 10억원 미만 등
시가총액	40억원 미만 30일간 지속
회생절차개시신청	회생절차개시 신청
파산신청	파산 신청
감사의견등 ⋮	반기보고서 부적정, 의견거절, 범위제한 한정 or 반기보고서 기한 경과 후 10일내 미제출 ⋮
정기보고서 미제출	분기, 반기, 사업보고서 미제출

03 관리종목 기준 ►

관리종목 지정요건 중 **부실과 관련있는**
지정요건으로 데이터 수집

부실기업과 관련없는 관리종목 지정 사유 제외
예시)

- 1) SPAC 상장예비심사청구서 미제출
- 2) 종류주식의 시가총액 요건 미달(5억원 미달 30일 계속)
- 3) 주식분산기준미달
- 4) 사외이사수 미달

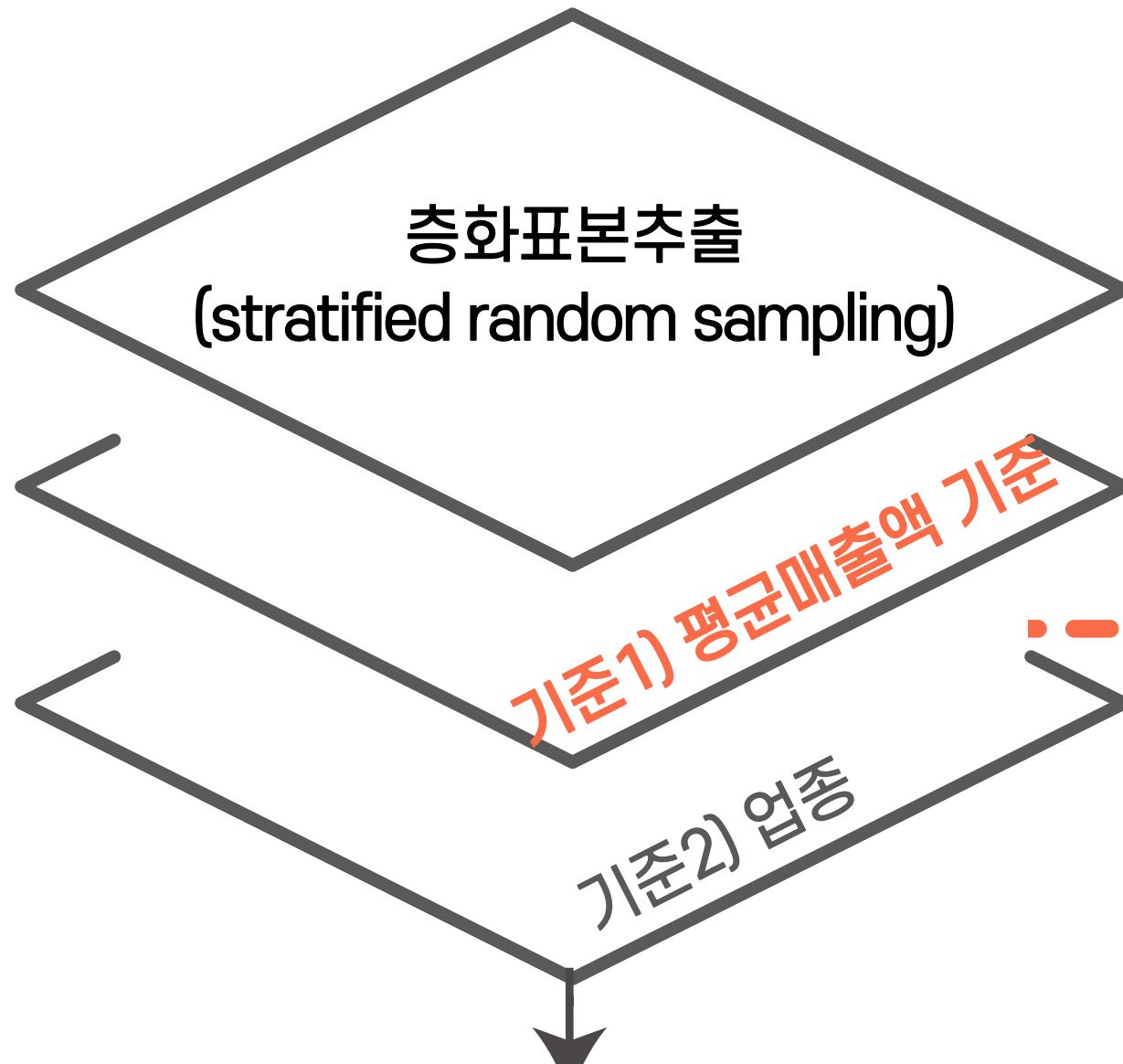
출처: KIND

11 층화표본추출 :

: 정상기업 vs 부실기업 기준 1) 평균매출액 기준

조사 대상인 모집단을 보다 동질적인 특성을 지니는 계층으로 나누고,
나누어진 층에서 단순 무작위 표본을 추출하는 방법

진행과정



〈 주된 업종별 평균 매출액 기준 (중소기업기본법 시행령 별표1,3) 〉

해당 기업의 주된 업종	분류번호	중소기업 (평균매출액)	소기업 (평균매출액)	
제조업 (6개업종)	의복, 의복액세서리 및 모피제품 제조업	C14	1,500억원 이하	120억원 이하
	가죽, 가방 및 신발 제조업	C15		
	펄프, 종이 및 종이제품 제조업	C17		
	1차 금속 제조업	C24		80억원 이하
	전기장비 제조업	C28		
	가구 제조업	C32		
농업, 임업 및 어업 과수	A		80억원 이하	
	B			
제조업 (12개업종)	식료품 제조업	C10	120억원 이하	
	담배 제조업	C12	80억원 이하	
	섬유제품 제조업(의복 제조업 제외)	C13		
	목재 및 나무제품 제조업 (가구 제조업 제외)	C16	120억원 이하	
	코크스, 연탄 및 석유정제제품 제조업	C19		
	화학물질 및 화학제품 제조업 (의약품 제조업 제외)	C20	80억원 이하	
	고무제품 및 플라스틱제품 제조업	C22		
	금속가공제품 제조업 (기계 및 가구 제조업 제외)	C25		
	전자부품, 컴퓨터, 영상, 음향 및 통신장비 제조업	C26	120억원 이하	
	그 밖의 기계 및 장비 제조업	C29		
	자동차 및 트레일러 제조업	C30		
	그 밖의 운송장비 제조업	C31	80억원 이하	
전기, 기아, 증기 및 기초설비 제조업	D		120억원 이하	
	E36		80억원 이하	
	F		50억원 이하	
	G			

1000억원 이하

1,000억원 이하

Train + Test set
전체 기업 수 ->

1000억원 이상 : 55개

1000억원 이하 : 233개

-> 총 288개

출처: 중소벤처기업부

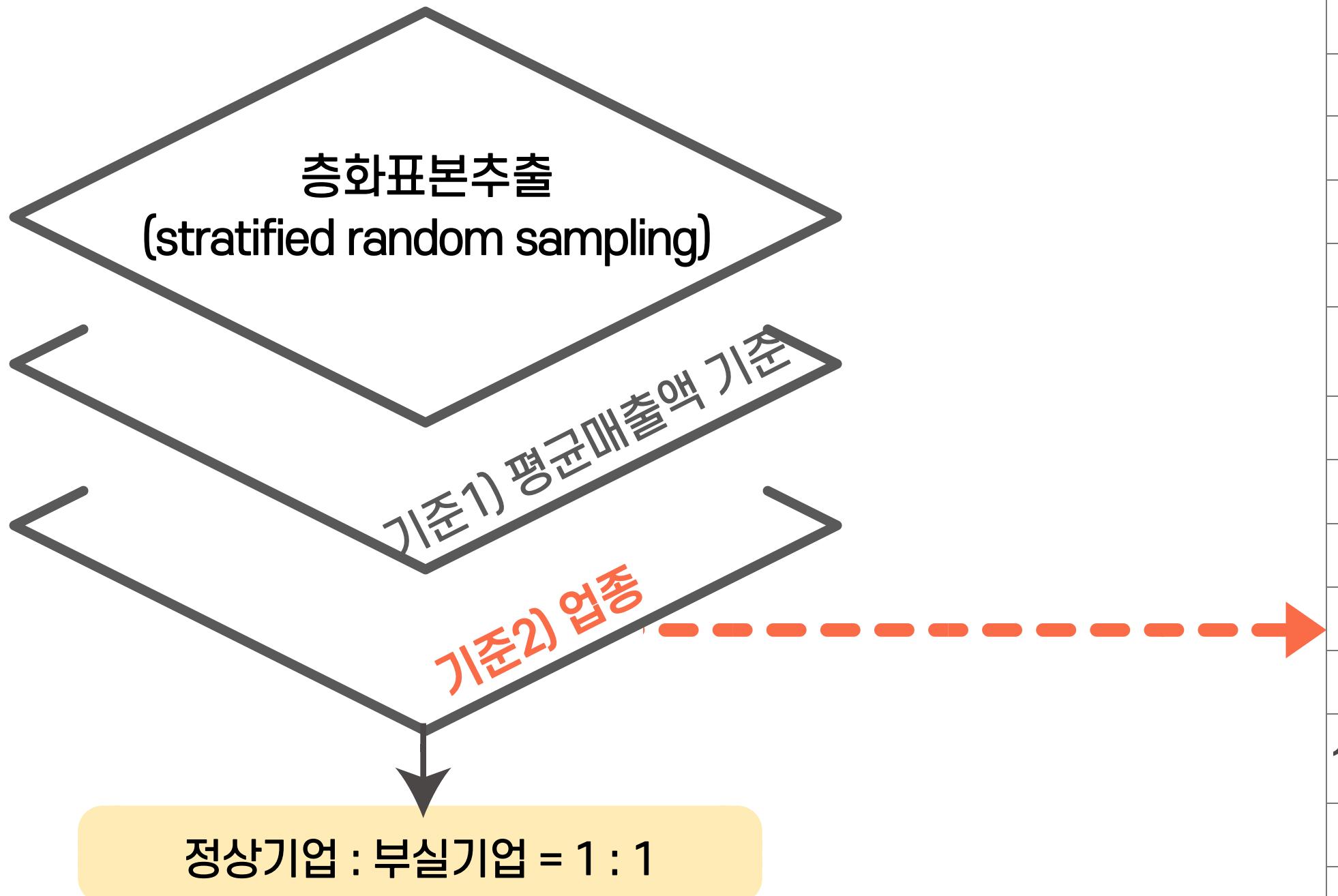
12

층화표본추출

: 조사 대상인 모집단을 보다 동질적인 특성을 지니는 계층으로 나누고,
나누어진 층에서 단순 무작위 표본을 추출하는 방법

진행과정

: 정상기업 vs 부실기업 기준 2) 업종 분류



통계청 한국표준산업분류
및 클라크의 산업분류 기반

한국표준산업분류 대분류	산업 분류	기업 수
농업, 임업 및 어업	제조업	1차 산업
광업		
제조업		
전기, 가스, 증기 및 공기조절 공급업		
건설업		
수도, 하수 및 폐기물 처리, 원료 재생업		
도매 및 소매업		
운수 및 창고업		
숙박 및 음식점업		
정보통신업		
전문, 과학 및 기술 서비스업	서비스업	86개
사업시설 관리, 사업 지원 및 임대 서비스업		
교육 서비스업		
예술, 스포츠 및 여가관련 서비스업		
Train + Test set 전체 기업 수		288개

Data Set 1 (재무데이터)

1) 데이터 수집 및 전처리 : 후보변수 선정

재무데이터 후보변수(40개) 출처(ts2000)

성장성(5)	수익성(10)	안정성(16)	활동성(6)	생산성(2)
총자본증가율	매출액순이익률 금융비용 대 총비용 총자본순이익률 자기자본순이익률 총자산영업이익률 순이익률 유보이익대총자산비율 매출총이익률 매출원가대매출액 판매관리비대매출액	유동부채비율 CASH FLOW 대 매출액 CASH FLOW 대 부채 CASH FLOW 대 총자본 이자보상배율 부채비율 유동비율 자기자본배율 당좌비율 순운전자본비율 현금비율 금융비용 대 부채 현금비율(CASHTA) 비유동자산비율 총자산부채비율(TLTA) 차입금의존도	자본금회전율 타인자본회전율 총자본회전율 매출채권회전율 유형자산회전율 재고자산회전율	설비투자효율 총자본투자효율
영업이익증가율				
순이익증가율				
매출액증가율				
유동자산증가율				

선행 연구:

회계정보와 시장정보를 이용한 부도예측모형의 평가 연구

재무비율을 이용한 부도예측에 관한 연구

감사의견, 감사법인 및 기업부실리스크의 예측

14 Data Set 1 (재무데이터)

진행과정

1) 데이터 수집 및 전처리 : 후보변수 수집

01.
데이터 출처

ts2000

02.
추출조건

- 연결 재무제표의 결측치 -> **개별 재무제표** 데이터로 대체
- 사업보고서 결산일 : **12월**
- 코스닥 상장 **K-IFRS**(한국채택국제회계기준) 공시 기업

03.
추출기간

관리종목지정 기업

Trainset - 회계년도 2010/12 - 2017/12

Testset - 회계년도 2017/12 - 2019/12

정상기업

Trainset - 회계년도 2016/12 - 2017/12

2018년 3월 31일 이전 상장사 까지 포함

Testset - 회계년도 2018/12 - 2019/12

2019년 1월 1일 이후 폐지사 포함

15 Data Set 1 (재무데이터)

진행과정

1) 데이터 수집 및 전처리 : 후보변수 계산

ts2000의 재무비율을 이용한 변수 - 27개

+

직접 계산한 파생변수 - 13개

1. 총자본증가율 / 영업이익증가율 / 순이익증가율 / 매출액증가율 / 유동자산증가율

-> 증가율의 경우 세가지 경우로 나누어 계산

1) 전기말이 양수인 경우

2) 전기말이 음수이면서 $\text{abs}(\text{전기말}) > \text{abs}(\text{당기말})$ 인 경우

3) 전기말이 음수이면서 $\text{abs}(\text{전기말}) < \text{abs}(\text{당기말})$ 인 경우

2. 재무제표 계정과목의 당기와 전기의 평균을 구하는 경우

총자산영업이익률 : 영업순익/(당기 자산과 전기 자산의 평균)

순이익률 : 당기순이익/(당기 자산과 전기 자산의 평균)

금융비용/부채 비율 : 이자비용/(당기 부채과 전기 부채의 평균)

3. 유보이익대총자산비율 (이익잉여금/총자산), 판매관리비대매출액 (판매관리비/매출액), 총자산의 로그값(lnTA), 현금비율(CASHTA) (현금성자산 / 총자산), 비유동자산비율 (비유동자산/자기자본), 총자산부채비율 (총부채/총자산)

16 Data Set 1 (재무데이터)

진행과정

2)-1 EDA (inf, 결측치 처리)

inf값 처리 - 분모가 0인 경우

- ★ 영업이익증가율 / 순이익증가율
당기년도 영업이익 / 순이익이
양수(+)인 경우 : → max,
음수(-)인 경우 : → min

★ 이자보상배율(이자비용)

999999999 → max

★ 재고자산회전율

inf → max

결측치 처리

- ★ 영업이익증가율, 순이익증가율
→ 전년도 데이터의 부재하는 경우 행 삭제
(회사명: 비지니스온커뮤니케이션)

제외 처리한 경우

- ★ SPAC 상장인 경우 제외
- ★ 인수목적합병된 경우 제외
- ★ 기타법인 제외) 같은 기업의 거래소코드가 2개 이상인 경우 하나 삭제
- ★ 순수지주회사 제외
예) 하림지주
- ★ 기술특례상장 제외
예) 아이진, 큐리언트,
- ★ 국내 상장된 외국기업 제외

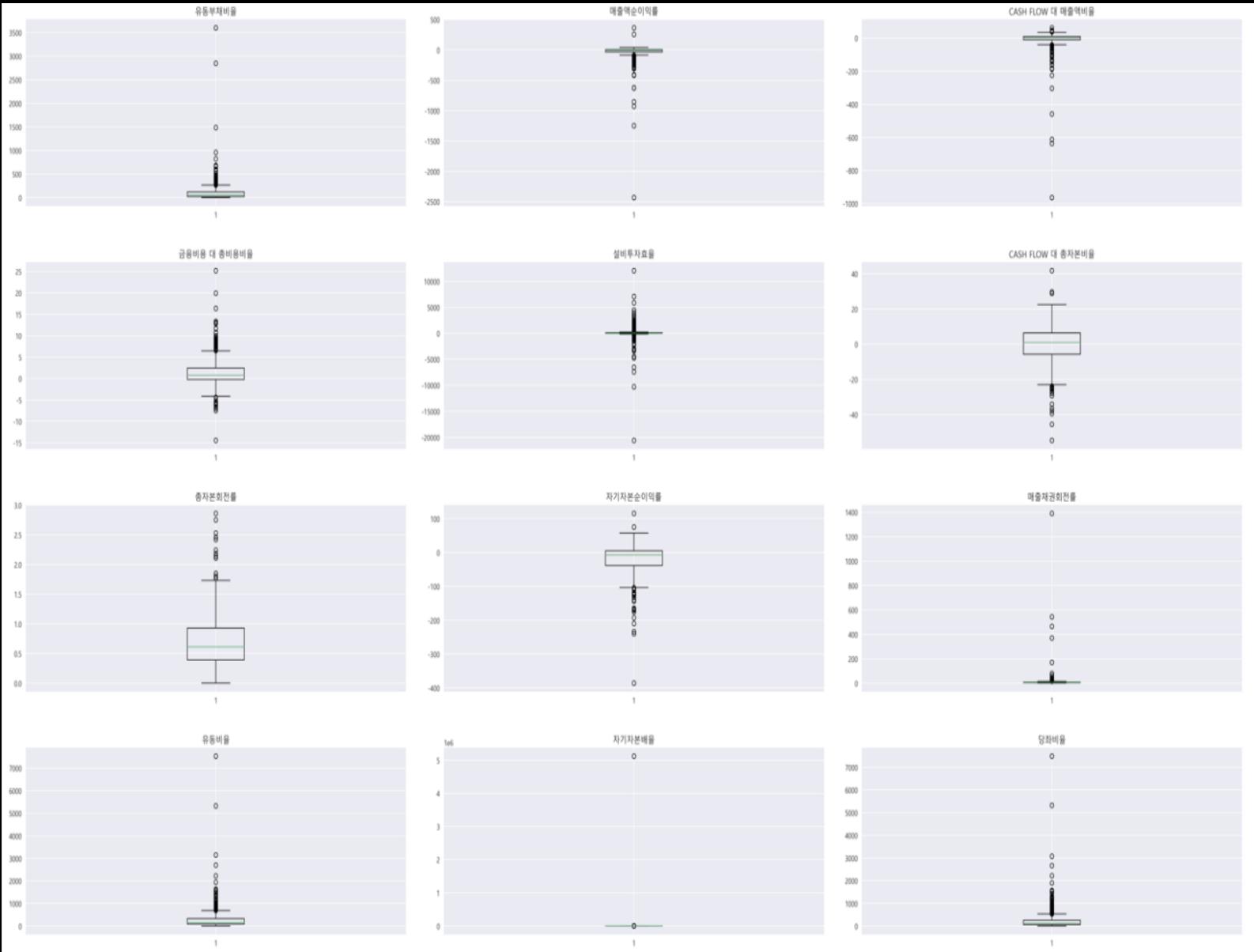
17 Data Set 1 (재무데이터) 2)-2 EDA (이상치 확인 : Boxplot)

진행과정

모든 값은 사업보고서를 확인했으나 값은 정확
따라서 winsorize를 통해 값 대체

정규성 그래프 꼬리와 Box plot을 토대로
양측, 단측 99% 백분위 winsorize 실행

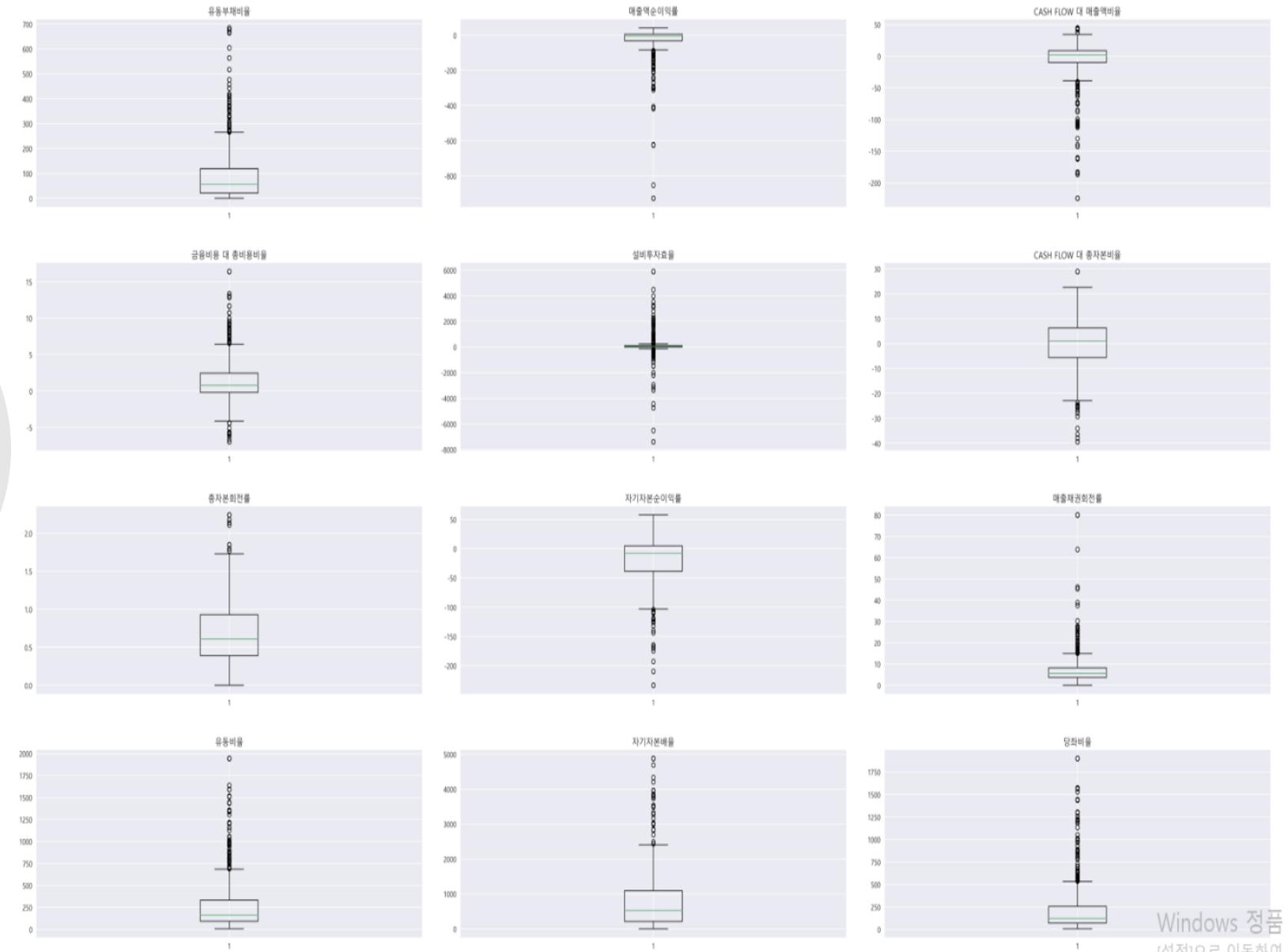
winsorize 전



winsorize



winsorize 후



Windows 정품

[설정]으로 이동하여

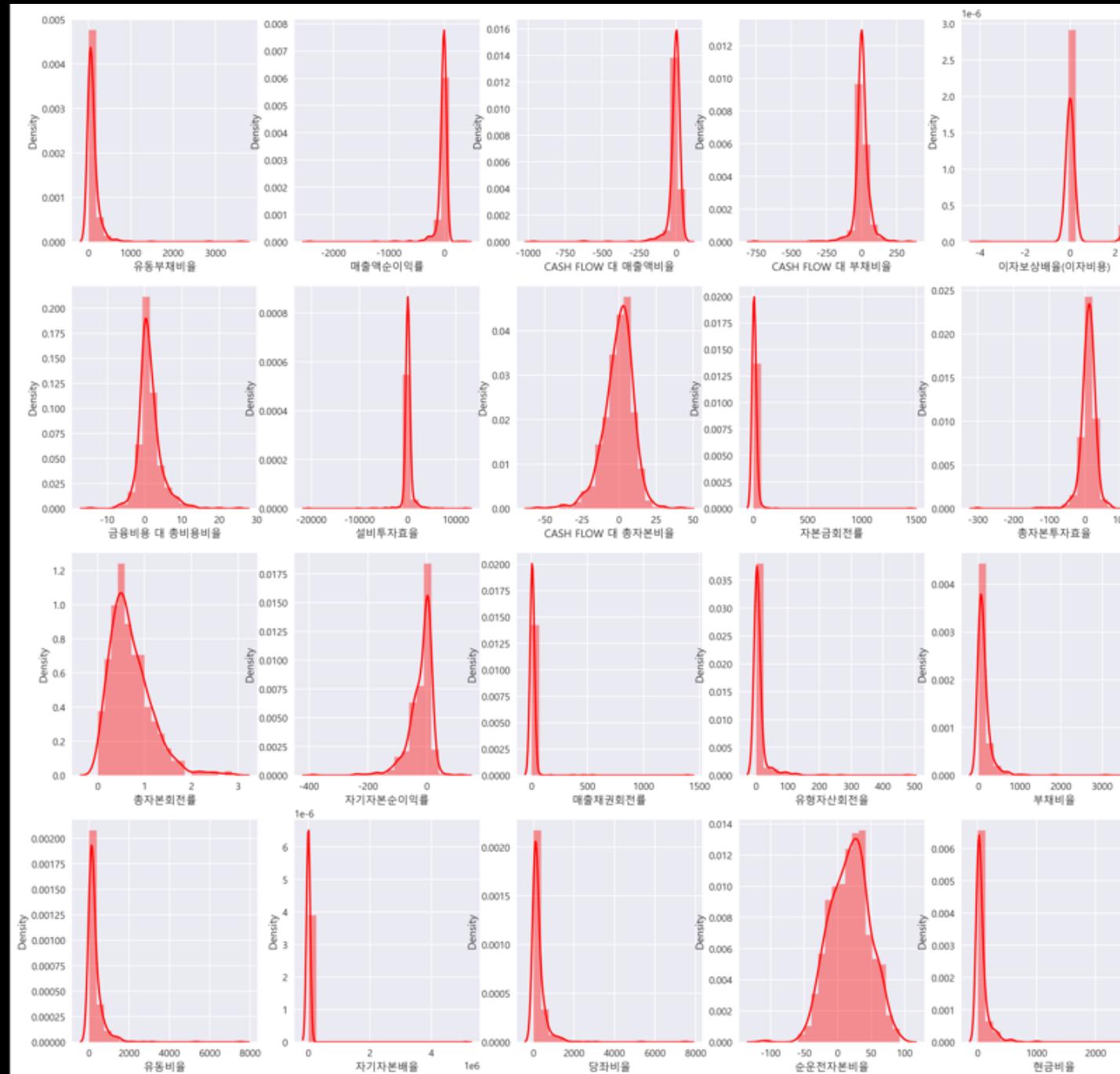
Winsorization : 양측 극단값을 더 작은 값으로 대체하는 과정

18 Data Set 1 (재무데이터)

2)-3 EDA (이상치 확인 : 정규성그래프)

진행과정

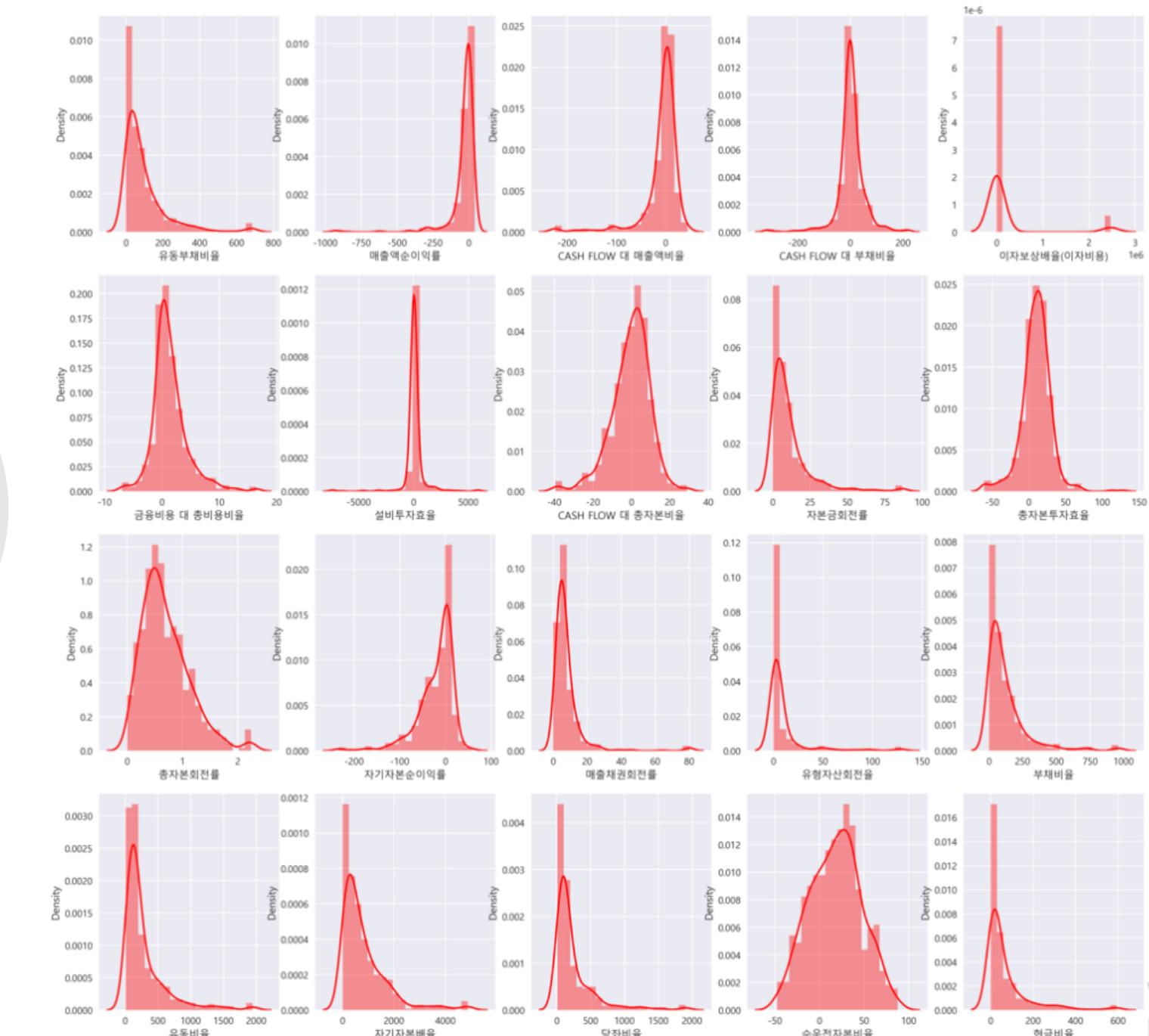
winsorize 전



winsorize



winsorize 후



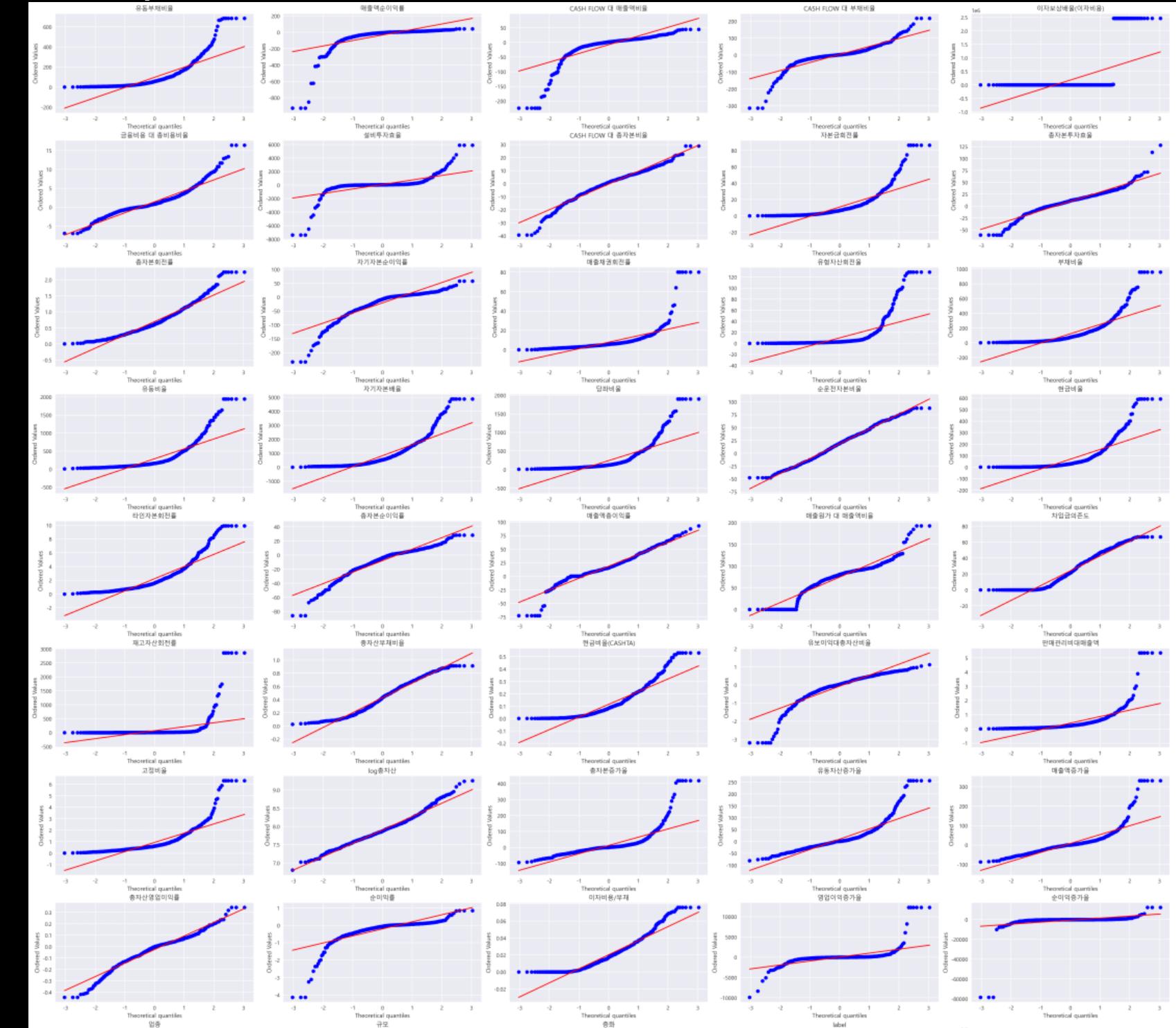
19 Data Set 1 (재무데이터)

2)-4 EDA (정규성 검정)

진행과정



Q-Q plot



20 Data Set 1 (재무데이터)

진행과정

2)-5 EDA (정규성 검정)

검정 결과

	Shapiro-Wilk Test	Anderson-Darling Test(1%)		K-S Test	
유동부채비율	7.998e-31	유동부채비율	> 1.085	유동부채비율	0
매출액순이익률	6.086e-39	매출액순이익률	> 1.085	매출액순이익률	3.107e-164
CASH FLOW 대 매출액비율	1.748e033	CASH FLOW 대 매출액비율	> 1.085	CASH FLOW 대 매출액비율	5.101e-114
이자보상배율 (이자비용)	3.046e-42	이자보상배율 (이자비용)	> 1.085	이자보상배율 (이자비용)	8.845e-118
⋮	⋮	⋮	⋮	⋮	⋮
금융비용 대 총비용비율	5.812e-19	금융비용 대 총비용비율	> 1.085	금융비용 대 총비용비율	9.849e-58

정규성을 만족하지 못함

정규성을 만족하지 못함

정규성을 만족하지 못함

21 Data Set 1 (재무데이터)

진행과정

2)-6 EDA (등분산성) Levene 검정

정상기업과 부실기업의 각각의 컬럼 : 등분산성, 이분산성 모두 존재

Output exceeds the [size limit](#). Open the full output data [in a text editor](#)

유동부채비율 5.112621680196903e-13 등분산성

매출액순이익률 0.0047896001207562 등분산성

CASH FLOW 대 매출액비율 0.017994929262914326 등분산성

CASH FLOW 대 부채비율 0.0017353242023570262 등분산성

이자보상배율(이자비용) 0.00038037111616936537 등분산성

금융비용 대 총비용비율 1.7116777099218153e-08 등분산성

설비투자효율 0.9507292645322216 이분산성

CASH FLOW 대 총자본비율 0.00862743586320241 등분산성

자본금회전률 0.0009617924282811028 등분산성

총자본투자효율 0.0045291949608444685 등분산성

총자본회전률 0.651524200270343 이분산성

자기자본순이익률 1.2714944837902155e-17 등분산성

매출채권회전률 0.3241298839288054 이분산성

유형자산회전율 0.21868536558848592 이분산성

부채비율 2.7231339840912076e-10 등분산성

유동비율 1.3464742731376127e-07 등분산성

자기자본배율 7.221470619046206e-08 등분산성

당좌비율 3.673880548889021e-07 등분산성

순운전자본비율 0.04818191677879478 등분산성

현금비율 1.8183414923324314e-06 등분산성

타인자본회전률 4.325947601600378e-10 등분산성

총자본순이익률 2.593954536554444e-08 등분산성

매출액총이익률 0.7193211301316189 이분산성

매출원가 대 매출액비율 0.2845967913292165 이분산성

차입금의존도 0.0015798622110110377 등분산성

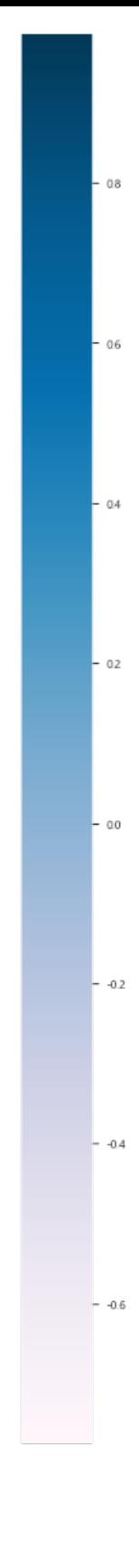
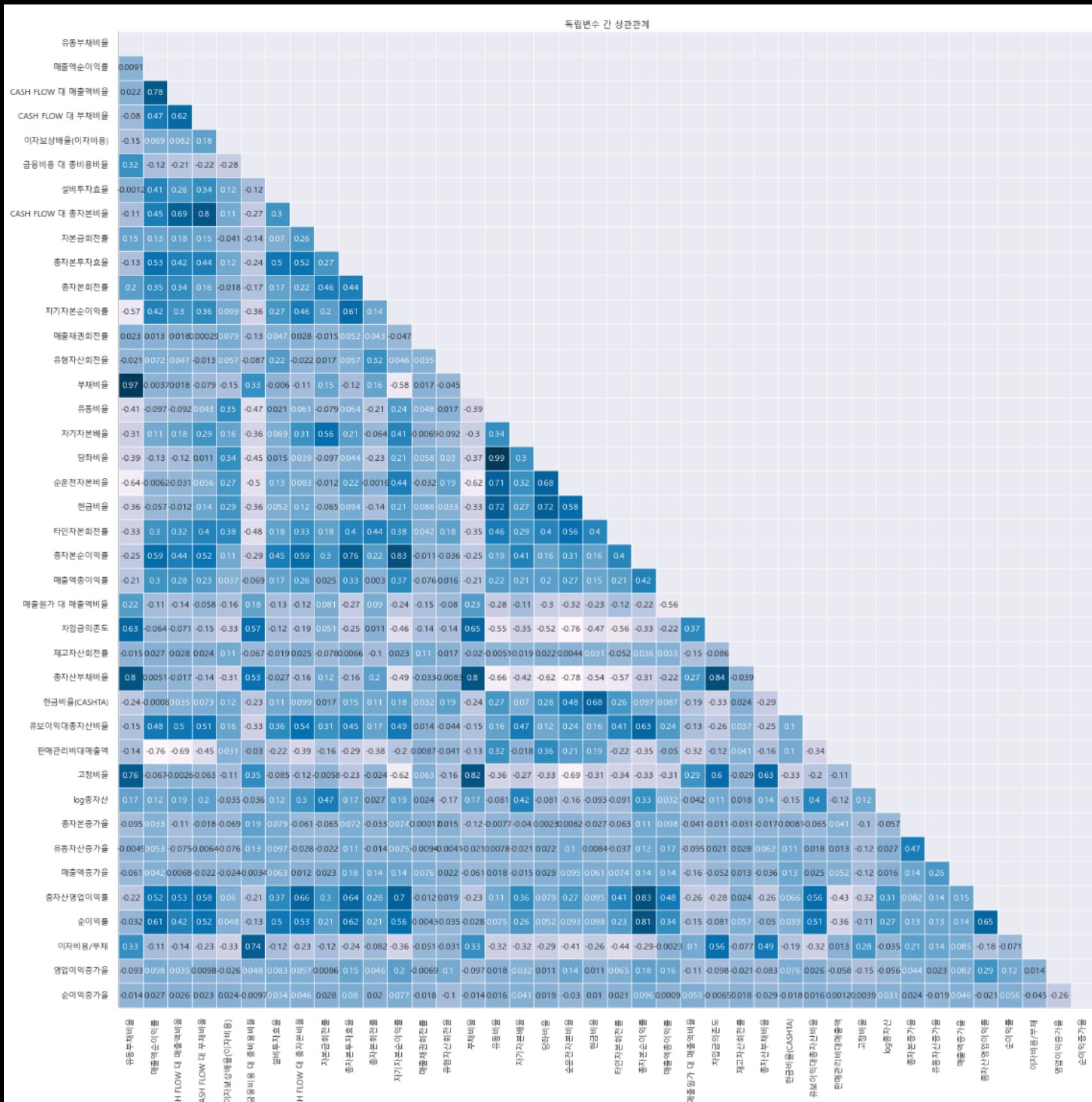
...

이자비용/부채 0.28748679575372044 이분산성

영업이익증가율 0.7628702232615598 이분산성

순이익증가율 0.7202338875217598 이분산성

업종 0.9490014215676592 이분산성



Data Set 1 (재무데이터)

2)-7 EDA (Heatmap)

Heatmap을 통해 변수 간 상관계수 확인

<상관계수가 높은 변수 예시>

당좌비율 <-> 유동비율 : 상관계수 0.99

부채비율 <-> 유동부채비율 : 상관계수 0.97
총자산부채비율 <-> 차입금의존도 : 상관계수 0.8

4

다중공선성 발생 여부 확인

23 Data Set 1 (재무데이터)

진행과정

2)-8 EDA (VIF 확인, 유의성 검정)

유의성 검정

	coef	std err	z	P> z	[0.025	0.975]
const	4.7144	0.946	4.95	0.000	2.861	6.568
유동부채비율	2.8142	2.055	1.39	0.171	-2.214	6.842
매출액순이익률	-2.4377	1.214	-2.08	0.045	-2.817	-0.058
CASH FLOW 대 매출액비율	-0.3164	0.631	-0.51	0.616	-1.554	0.921
CASH FLOW 대 부채비율	1.5708	0.584	2.60	0.007	0.426	2.715
이자보상배율(이자비용)	-0.3362	0.223	-1.59	0.131	-0.773	0.100
금융비용 대 총비용비율	1.9426	0.592	3.20	0.001	0.782	3.103
설비투자효율	0.1119	0.457	0.25	0.806	-0.783	1.007
CASH FLOW 대 총자본비율	-2.0514	0.803	-2.53	0.011	-3.626	-0.477
자본금회전률	-0.2954	0.688	-0.40	0.668	-1.643	1.052
총자본투자효율	-0.3596	0.468	-0.78	0.443	-1.277	0.558
총자본회전률	0.3960	0.723	0.58	0.584	-1.020	1.812
자기자본순이익률	0.1203	2.694	0.05	0.964	-1.160	5.401
매출채권회전률	-0.1349	0.257	-0.55	0.599	-0.638	0.368
유형자산회전율	0.0884	0.327	0.20	0.787	-0.553	0.730
부채비율	-3.1022	3.234	-0.99	0.337	-9.440	3.236
유동비율	-0.5411	1.392	-0.39	0.698	-3.270	2.188
자기자본배율	0.2669	0.504	0.50	0.596	-0.721	1.254
당좌비율	1.6754	1.191	1.46	0.160	-0.660	4.010
순운전자본비율	6.5983	1.156	5.77	0.000	-4.332	8.865
현금비율	-2.0880	0.764	-2.72	0.006	-3.586	-0.590
타인자본회전률	-1.0932	0.575	-1.93	0.057	-2.219	0.033
총자본순이익률	0.5214	2.452	0.23	0.832	-2.284	5.327

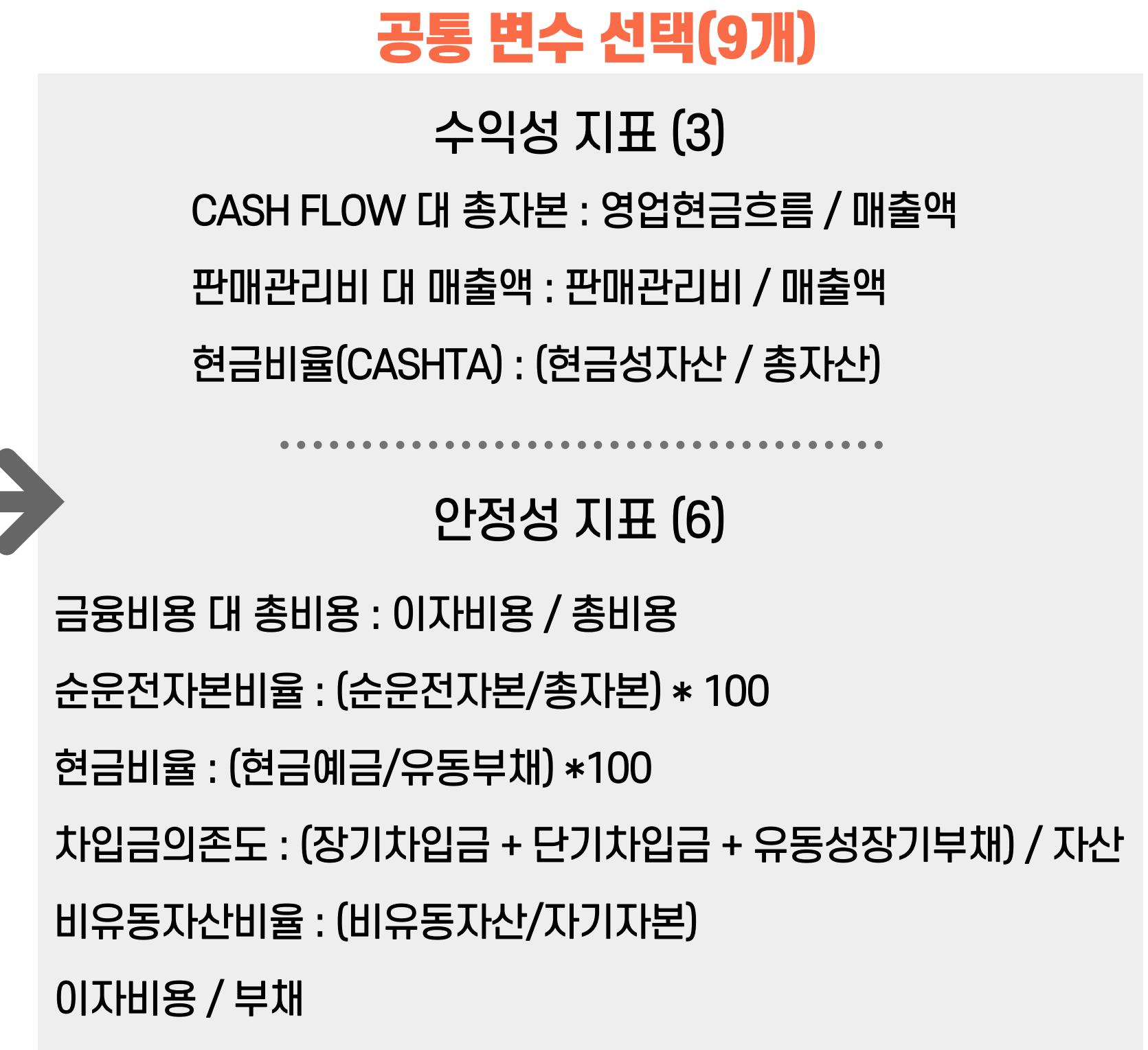
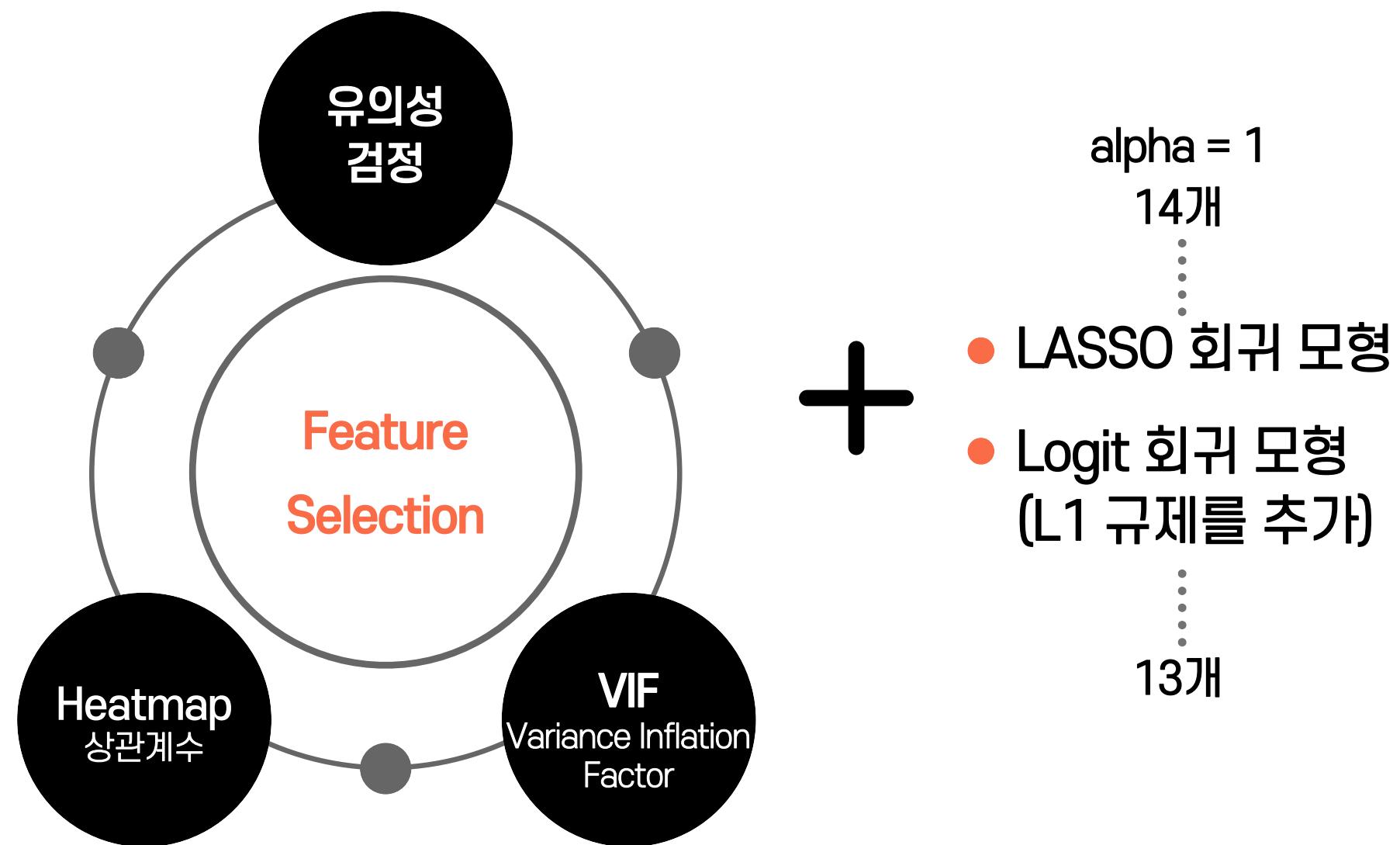
VIF 확인 -> 다중공선성 확인

VIF Factor	features	3	4.084697	CASH FLOW 대 부채비율	
15	61.930838	9	3.781868	총자본투자효율	
17	51.871939	27	3.776354	현금비율(CASHTA)	
14	34.780274	23	3.737256	매출원가 대 매출액비율	
0	24.980242	16	3.517959	자기자본배율	
21	18.852747	5	3.493134	금융비용 대 총비용비율	
26	14.532254	8	3.444711	자본금회전률	
11	9.240327	37	3.039712	이자비용/부채	
1	8.524671	28	2.499819	유보이익대총자산비율	
18	8.301548	22	2.147615	매출액총이익률	
29	7.673481	31	2.006315	log총자산	
2	6.946916	CASH FLOW 대 매출액비율	6	1.885740	설비투자효율
35	6.648214	총자산영업이익률	33	1.620286	유동자산증가율
30	6.131308	고정비율	32	1.542720	총자본증가율
19	5.983879	현금비율	13	1.503916	유형자산회전율
24	5.937986	차입금의존도	4	1.387676	이자보상배율(이자비용)
7	5.681622	CASH FLOW 대 총자본비율	38	1.326099	영업이익증가율
36	4.930667	순이익률	12	1.226235	매출채권회전률
20	4.848194	타인자본회전률	34	1.218370	매출액증가율
10	4.817975	총자본회전률	39	1.169649	순이익증가율
		25	1.133641	재고자산회전률	

24 Data Set 1 (재무데이터)

진행과정

3) Feature Selection



25 Data Set 1 (재무데이터)

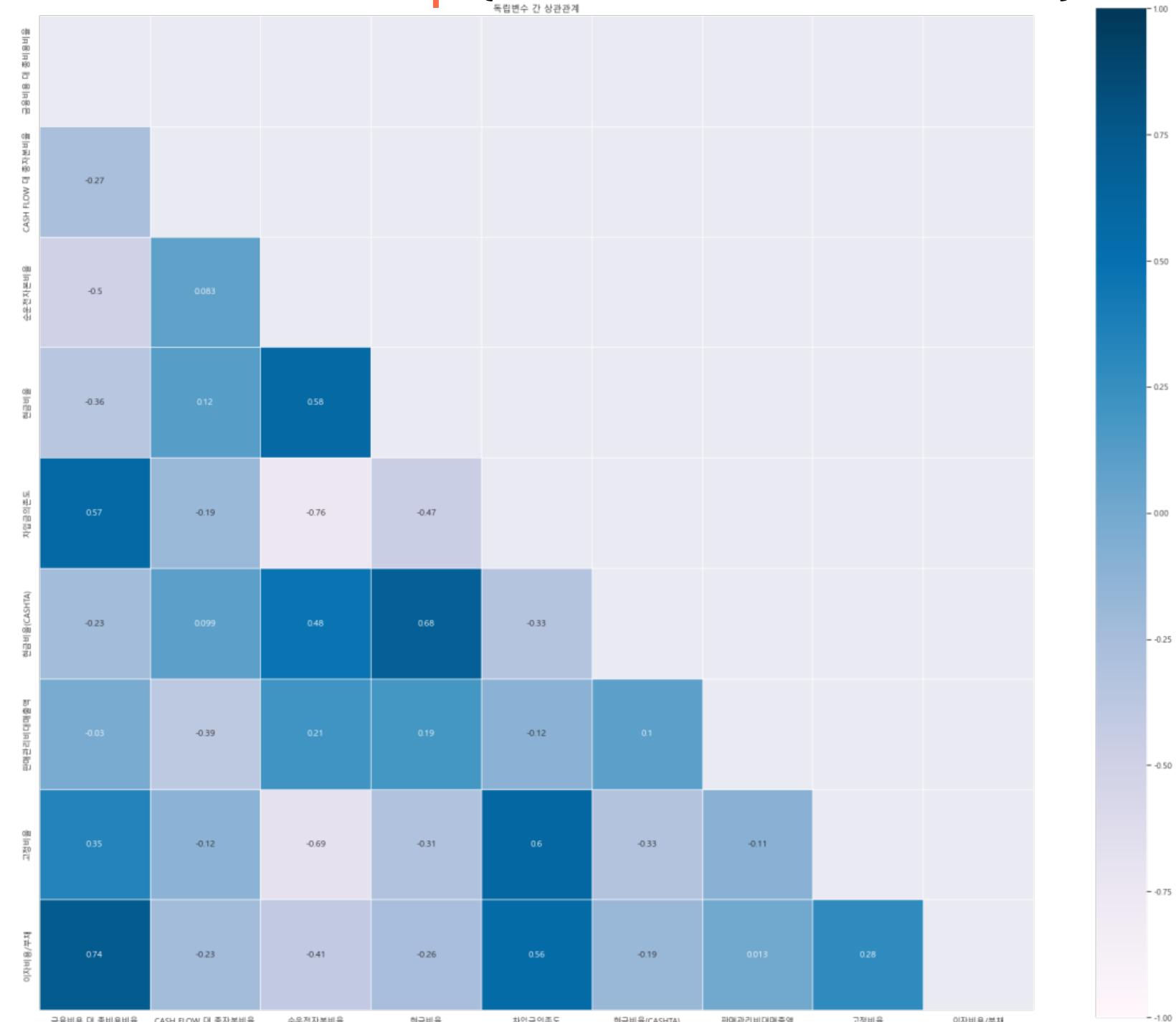
진행과정

3) Feature Selection

최종 변수 유의성 검증 (p-value가 모두 0.05 이하)

	coef	std err	t	P> z	[0.025	0.975]
const	6.8067	0.958	7.103	0.000	4.929	8.685
금융비용 대 총비용비율	0.9544	0.400	2.385	0.017	0.170	1.739
CASH FLOW 대 총자본비율	-1.2900	0.251	-5.135	0.000	-1.782	-0.798
순운전자본비율	7.5458	1.030	7.323	0.000	5.526	9.565
현금비율	-1.6392	0.392	-4.177	0.000	-2.408	-0.870
차입금의존도	1.4454	0.411	3.514	0.000	0.639	2.251
현금비율(CASHTA)	1.5879	0.333	4.771	0.000	0.936	2.240
판매관리비대매출액	-0.3741	0.167	-2.234	0.025	-0.702	-0.046
고정비율	24.7750	3.127	7.923	0.000	18.646	30.904
이자비용/부채	-1.0127	0.370	-2.736	0.006	-1.738	-0.287

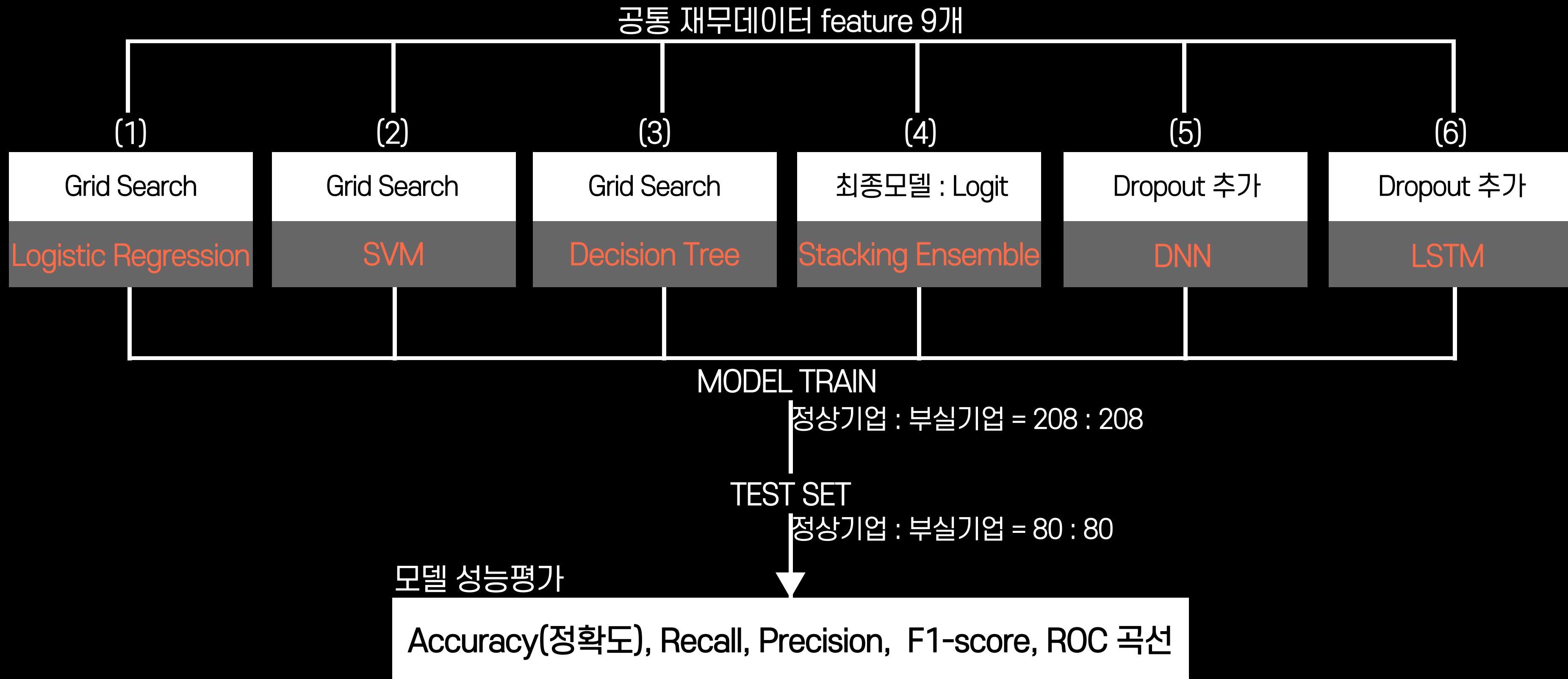
최종 변수 Heatmap (상관계수가 모두 0.8 이하)



26 Data Set 1 (재무데이터)

진행과정

4) Model



27 Data Set 1 (재무데이터)

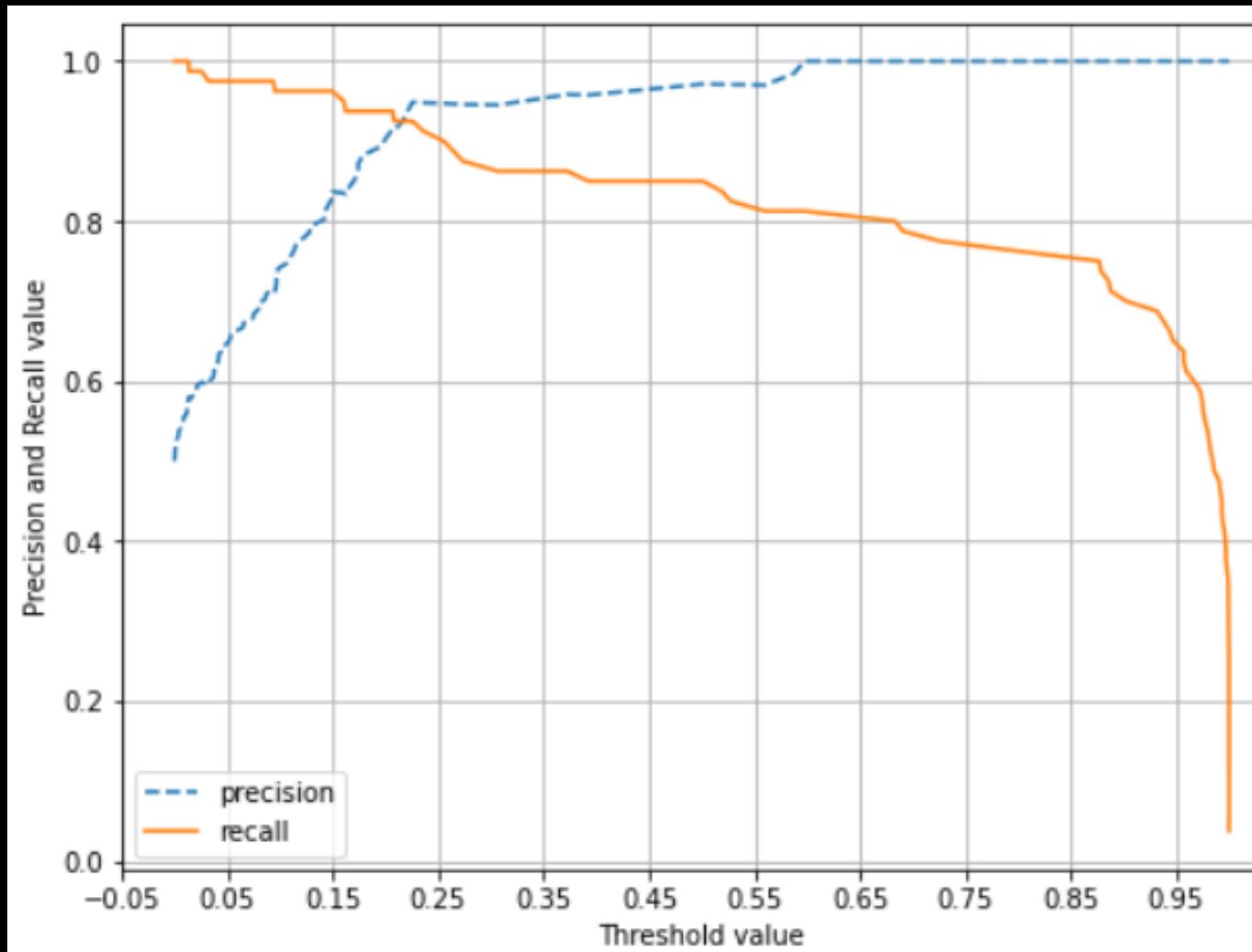
진행과정

4) Model Grid Search 활용 Logit

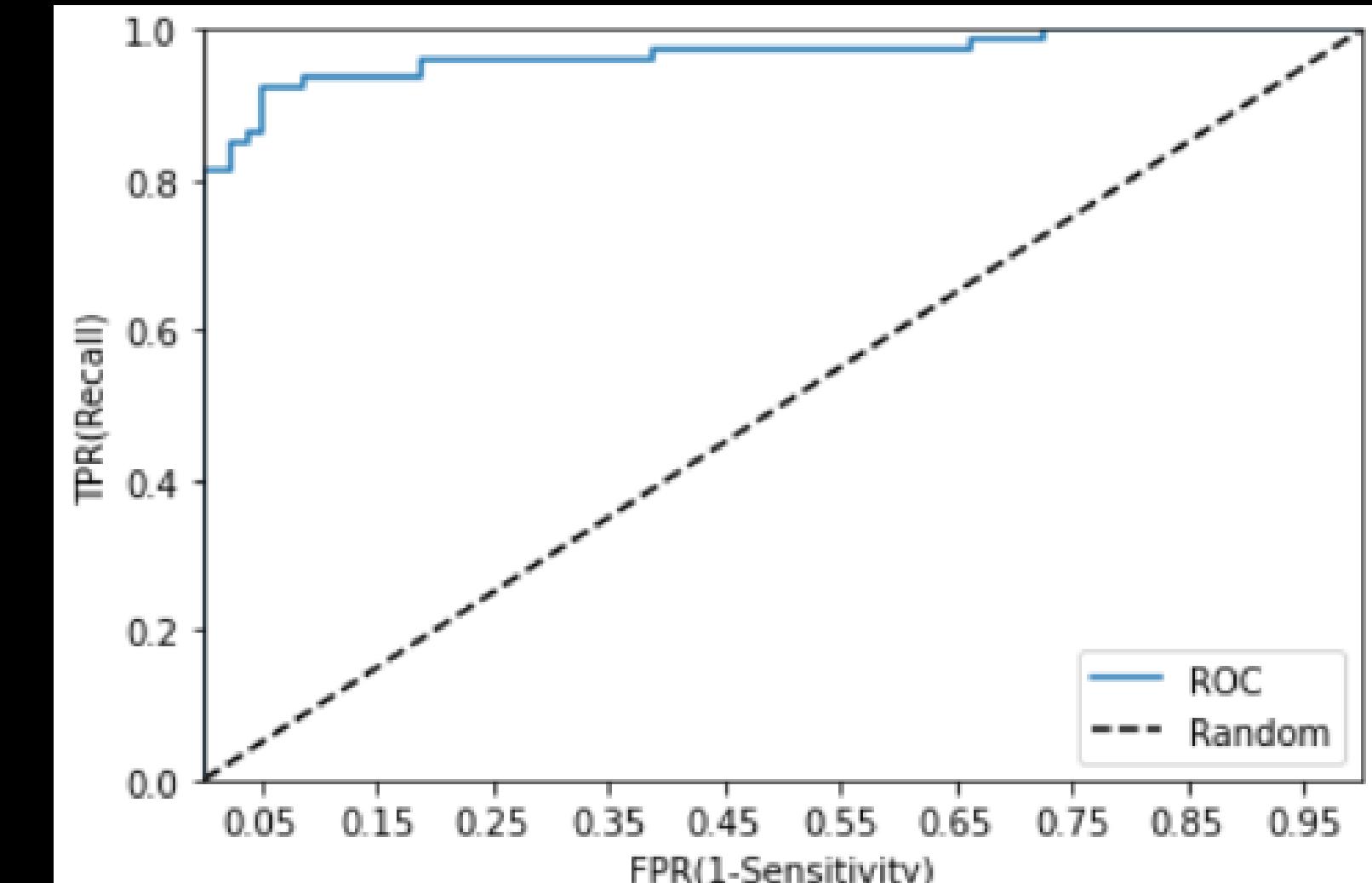
Hyper parameter : C = 10, penalty = 'l2'

정확도 : 0.9125
정밀도 : 0.9714285714285714
재현율 : 0.85
f1 score : 0.9066666666666667

임계값에 따른 precision , recall



ROC curve



28 Data Set 1 (재무데이터)

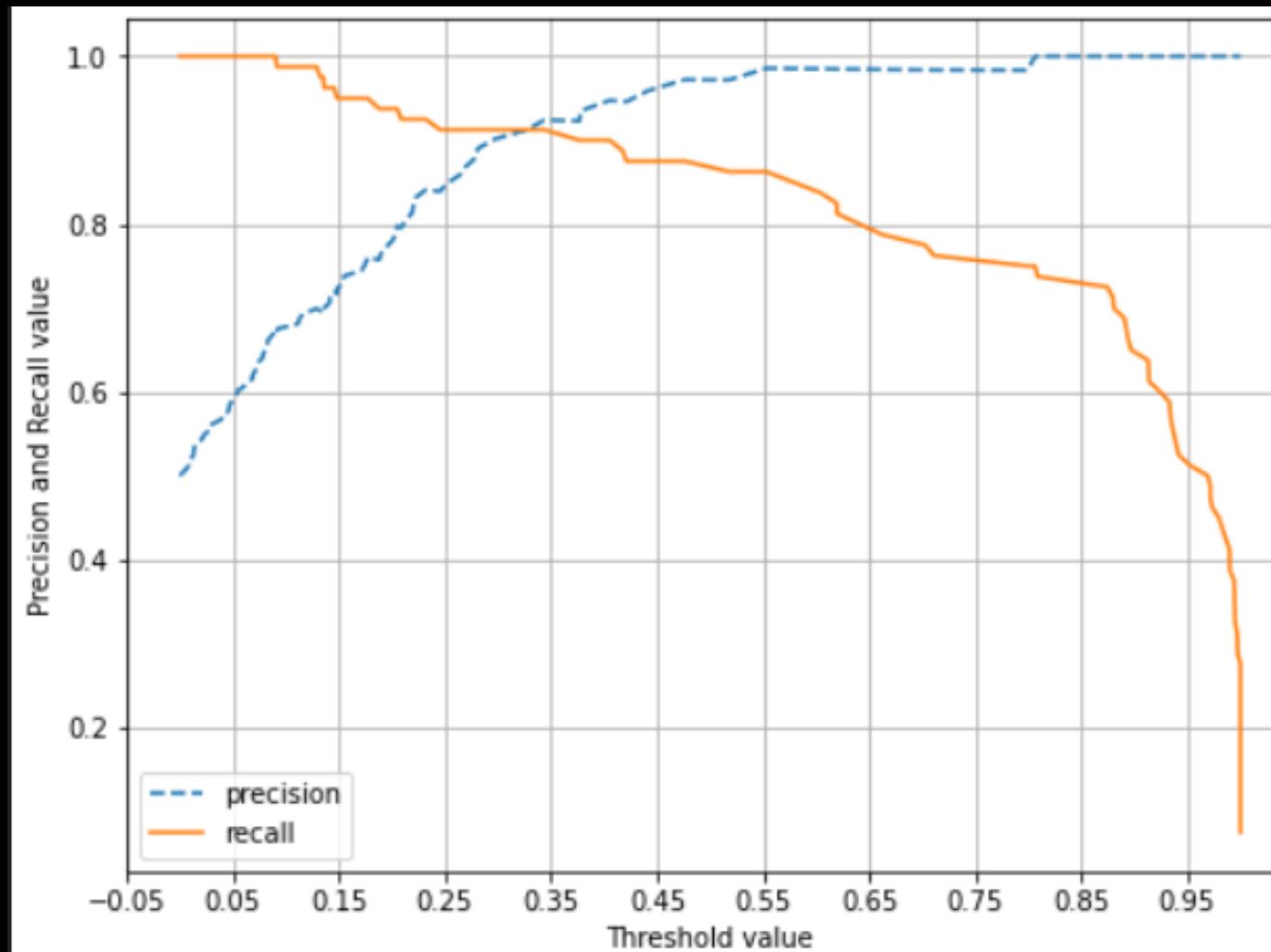
진행과정

4) Model Grid Search 활용 SVM

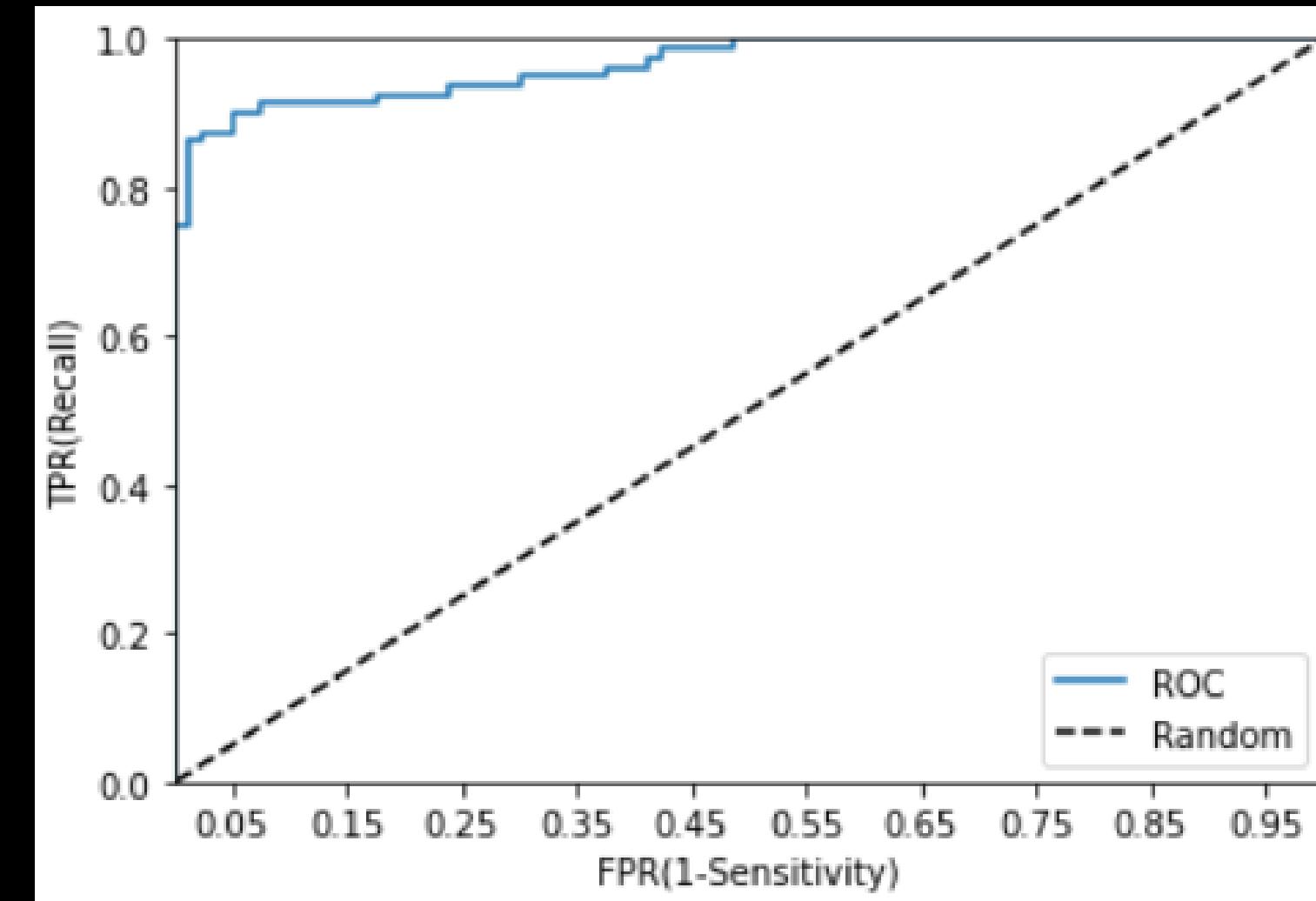
Hyper parameter : C = 1, gamma = 0.1

정확도 : 0.925
정밀도 : 0.9857142857142858
재현율 : 0.8625
f1 score : 0.9200000000000002

임계값에 따른 precision , recall



ROC curve



29 Data Set 1 (재무데이터)

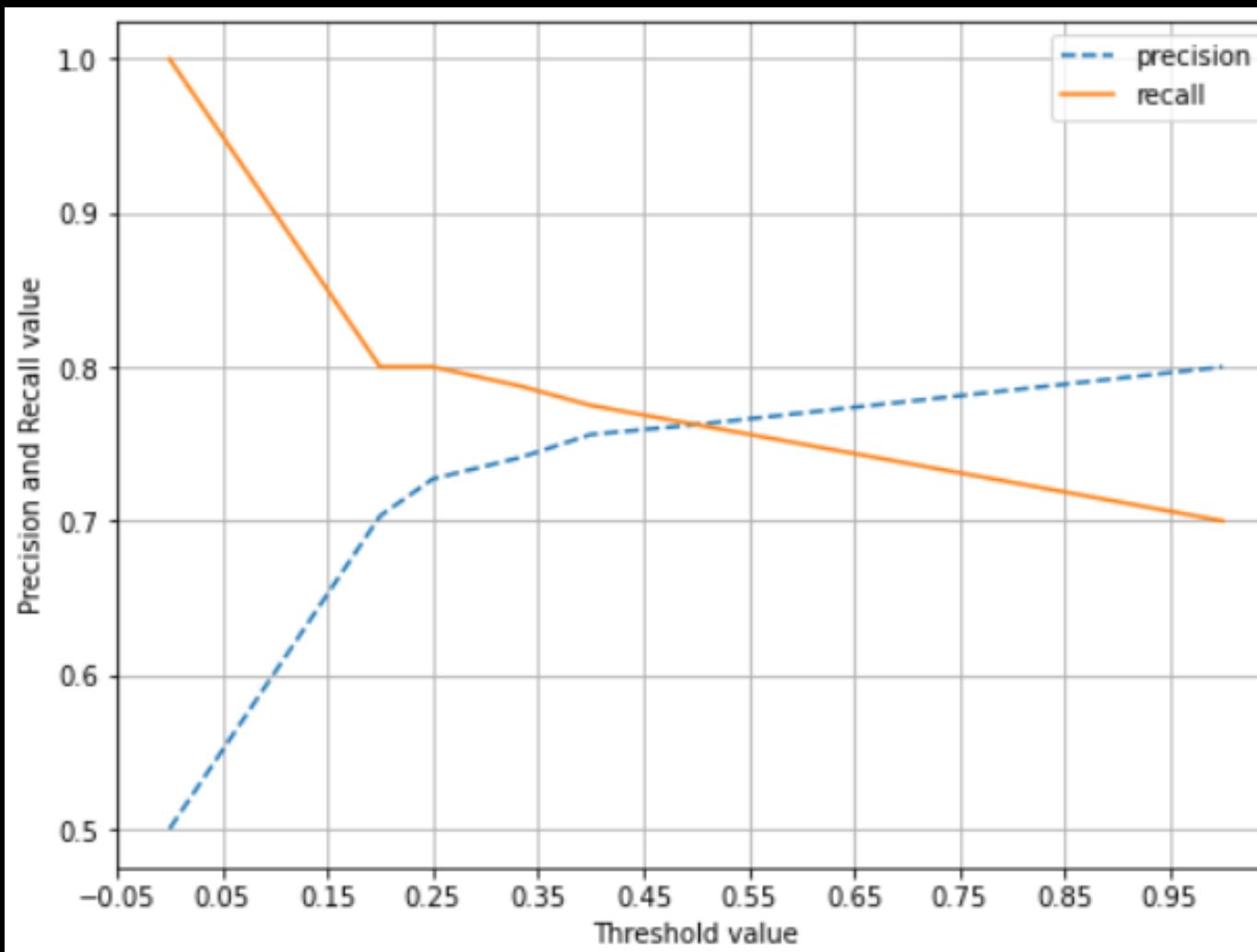
진행과정

4) Model Grid Search 활용 Decision Tree

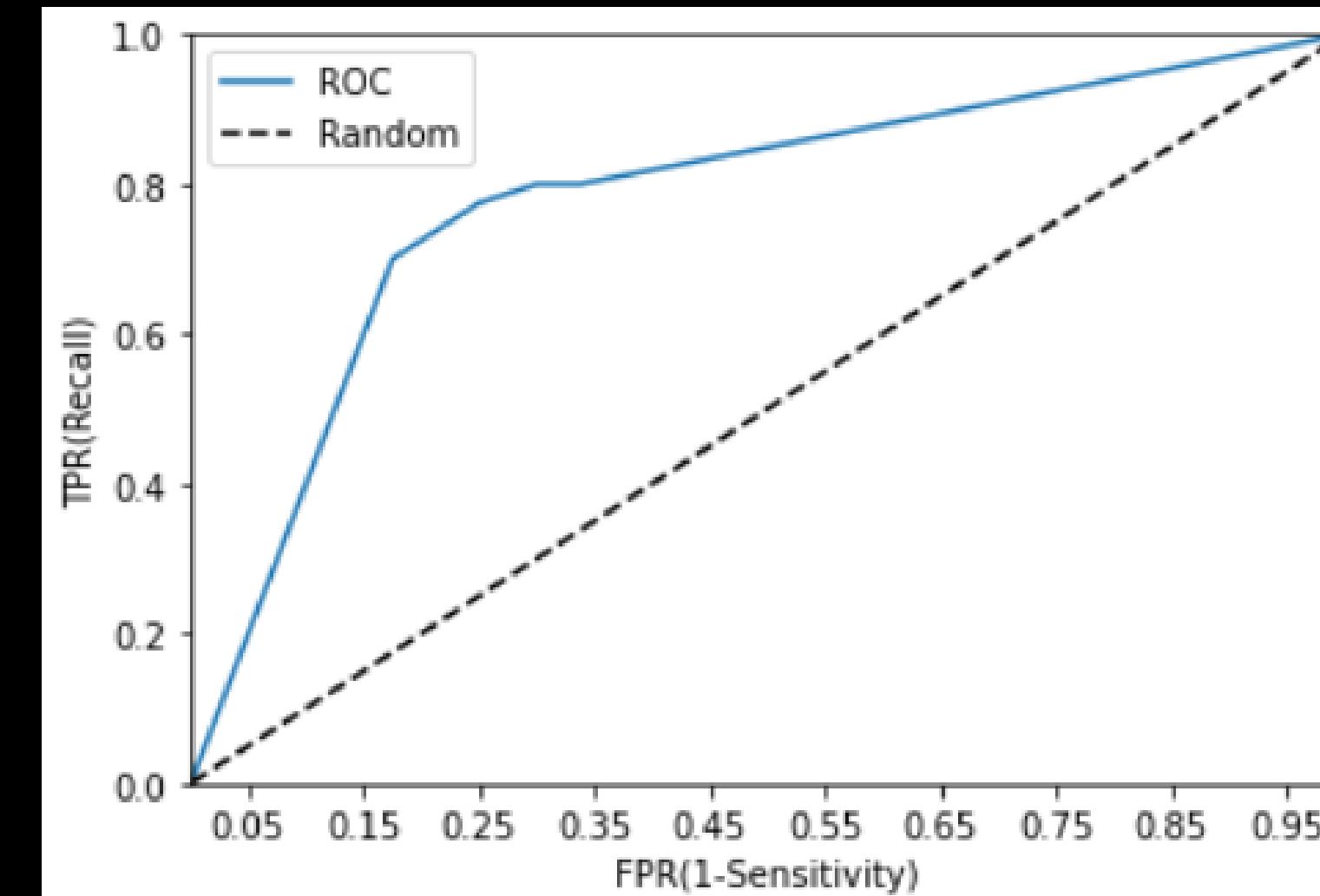
Hyper parameter : criterion = 'gini', 'max_depth' = 10,
max_features = 'auto', min_sample_split = 12

정확도 : 0.7625
정밀도 : 0.8
재현율 : 0.7
f1 score : 0.7466666666666666

임계값에 따른 precision , recall



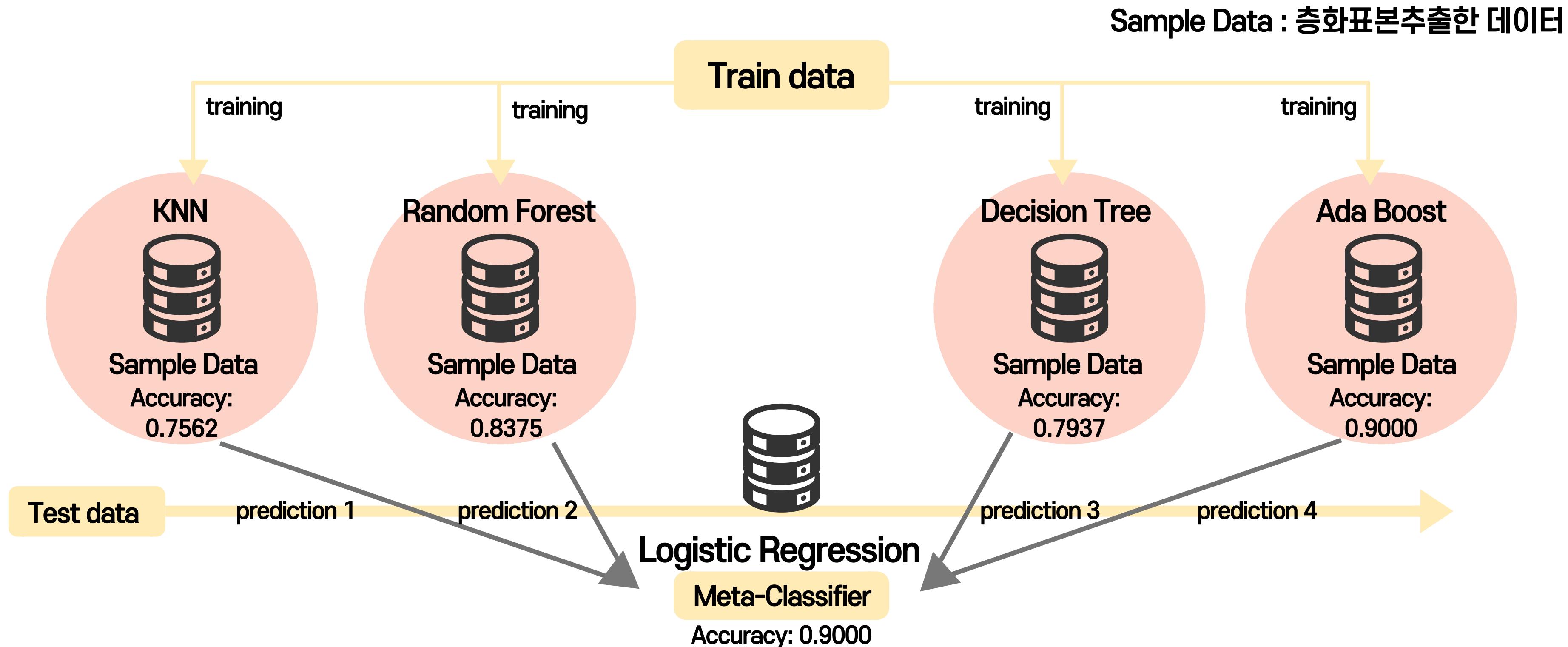
ROC curve



30 Data Set 1 (재무데이터)

진행과정

4) Model Stacking Ensemble



31 Data Set 1 (재무데이터)

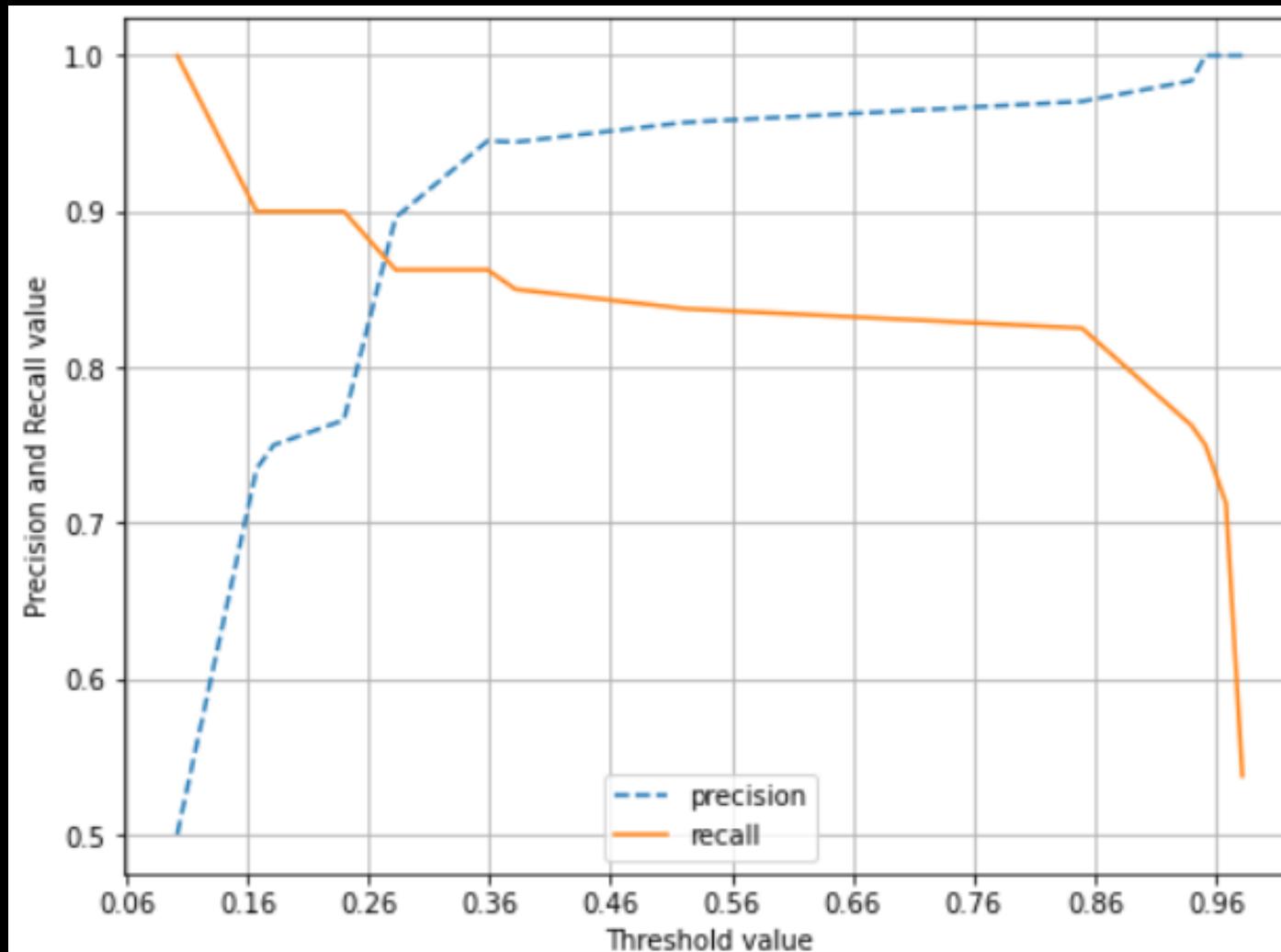
진행과정

4) Model Stacking Ensemble

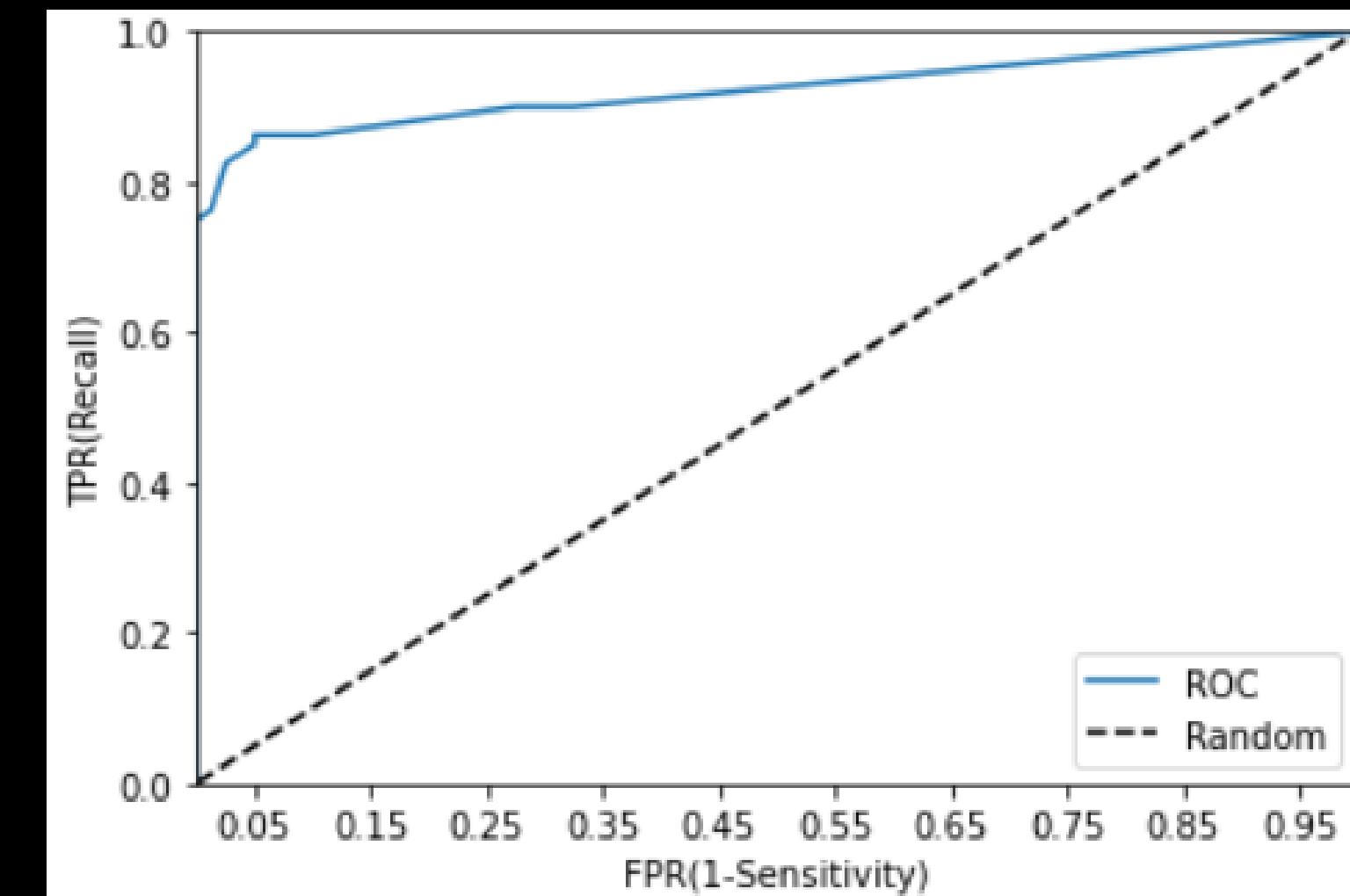
```
knn_clf = KNeighborsClassifier(n_neighbors=4)
rf_clf = RandomForestClassifier(n_estimators=100, random_state=0)
dt_clf = DecisionTreeClassifier(random_state=125)
ada_clf = AdaBoostClassifier(n_estimators=100)
```

정확도 : 0.9
정밀도 : 0.9571428571428572
재현율 : 0.8375
f1 score : 0.8933333333333334

임계값에 따른 precision , recall



ROC curve



32 Data Set 1 (재무데이터)

진행과정

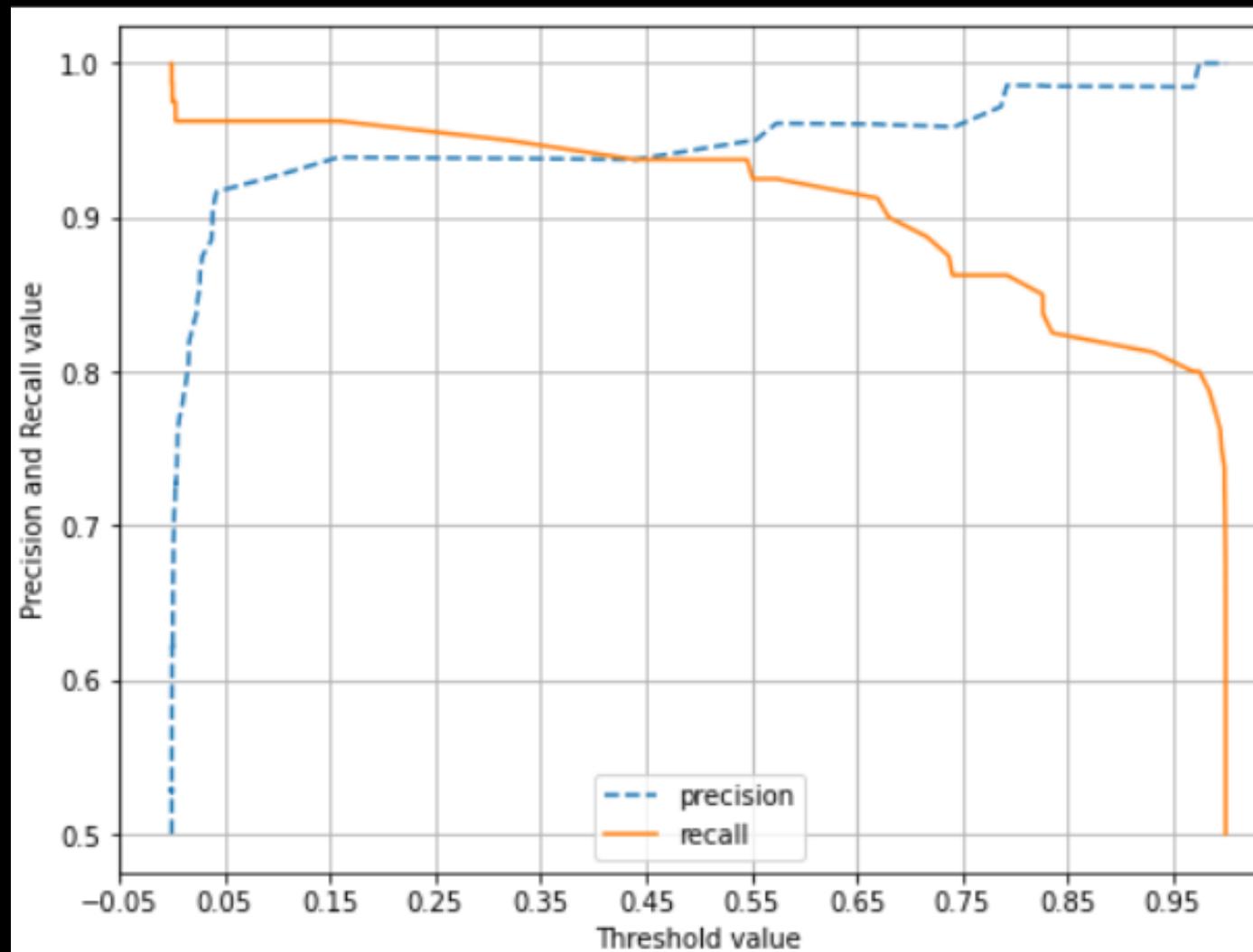
4) Model DNN 과적합 방지를 위해 Dropout 추가

Hyper parameter : dropout = 0.3, activation = 'relu',

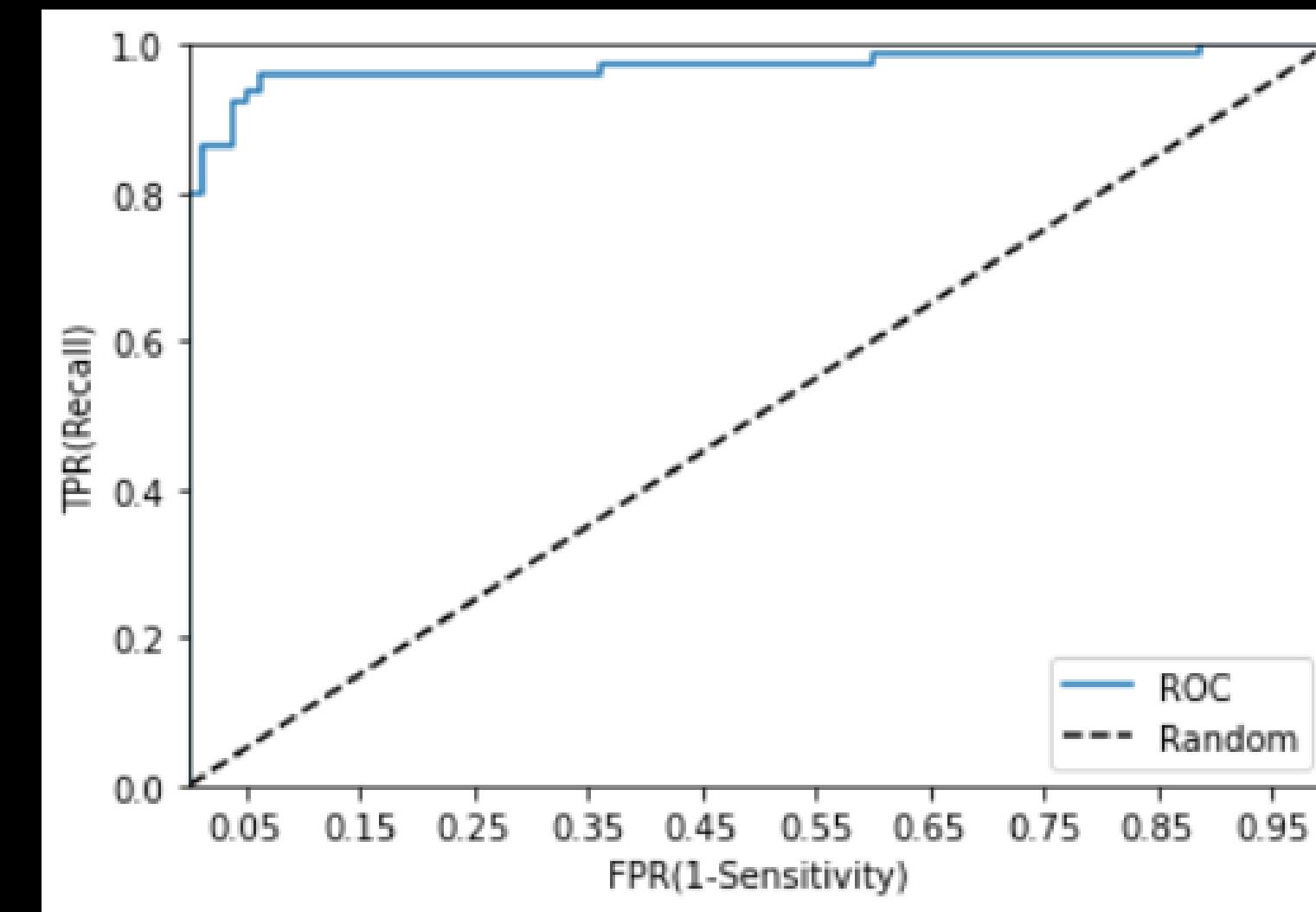
은닉층 수 = 2

정확도 : 0.94375
정밀도 : 0.9375
재현율 : 0.9493670886075949
f1 score : 0.9433962264150944

임계값에 따른 precision , recall



ROC curve



33 Data Set 1 (재무데이터)

진행과정

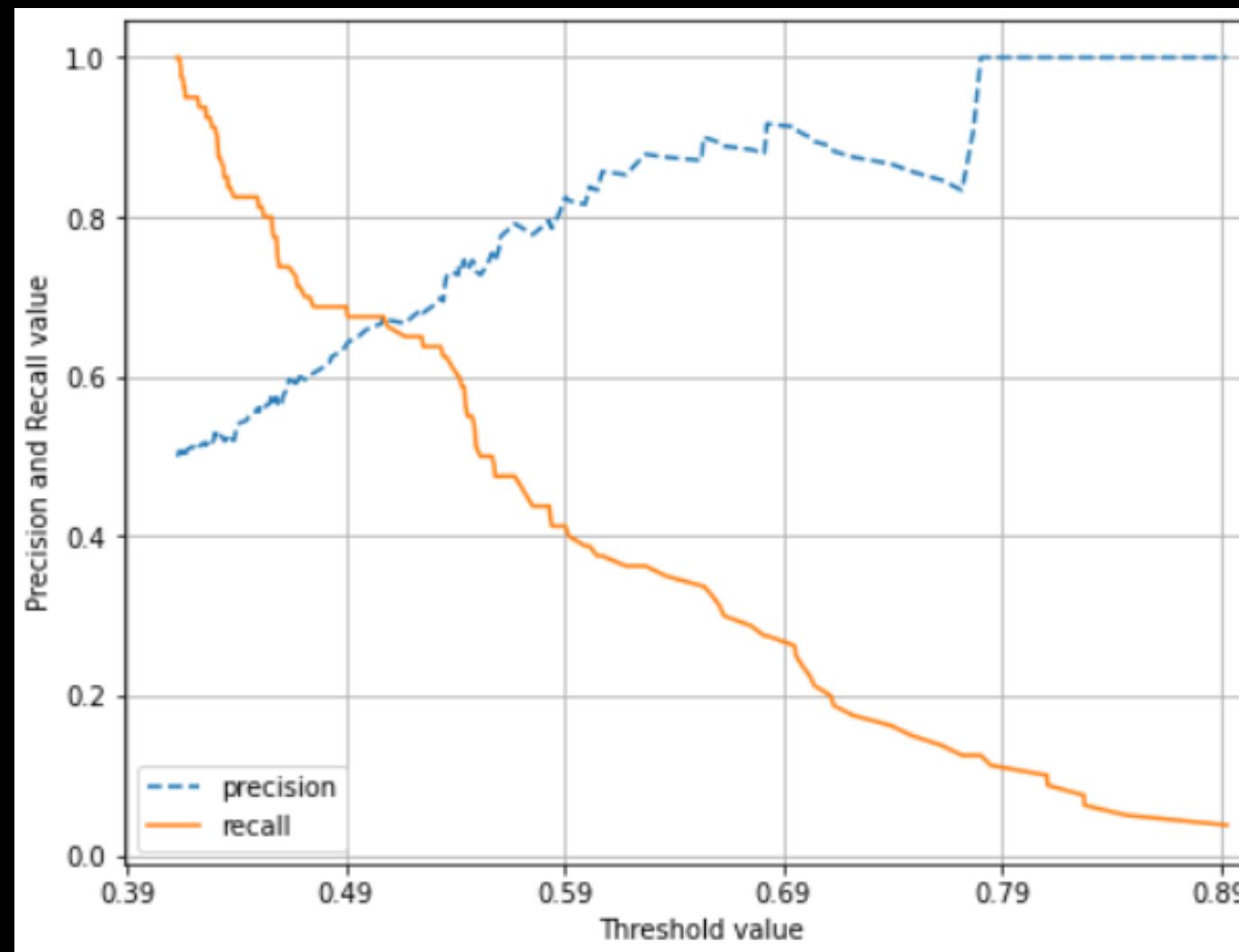
4) Model LSTM 과적합 방지를 위해 Dropout 추가

Hyper parameter : dropout = 0.3, activation = 'relu',

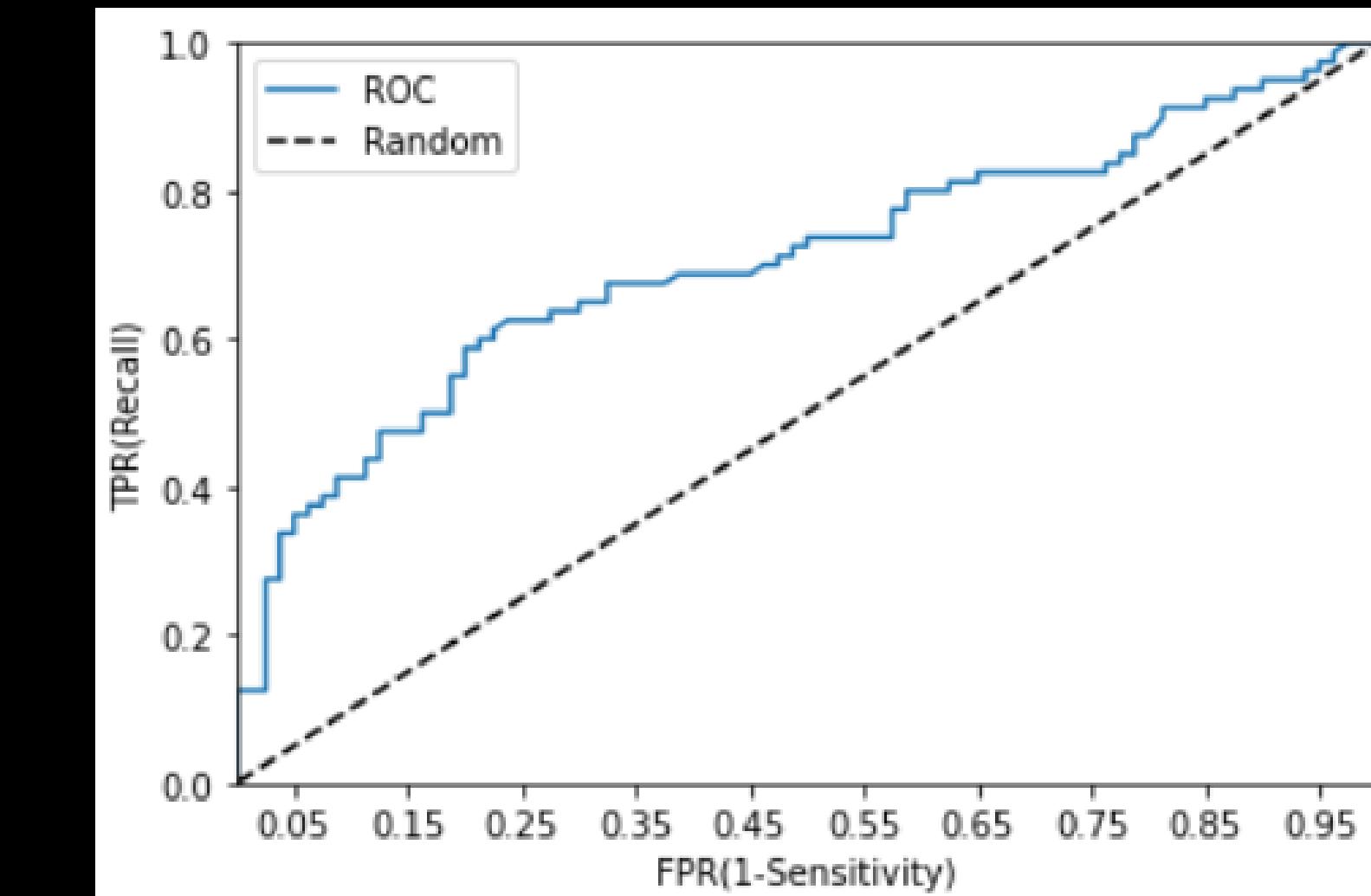
은닉층의 수 = 1

정확도 : 0.66875
정밀도 : 0.675
재현율 : 0.6666666666666666
f1 score : 0.670807453416149

임계값에 따른 precision , recall



ROC curve



34 Data Set 1 (재무데이터)

진행과정

4) Model 성능평가 비교 및 강건성 검증

증화추출표본 1:1 Sample

	Accuracy	Recall	Precision	F1-Score
Rogistic	0.9125	0.8500	0.9714	0.9067
SVM	0.9250	0.8625	0.9857	0.9200
Decision Tree	0.7625	0.7000	0.8000	0.7467
Stacking Ensemble	0.9	0.8375	0.9571	0.8933
DNN	0.9438	0.9494	0.9375	0.9434
LSTM	0.6688	0.6667	0.6750	0.6708

강건성을 위한 raw data (Scaling 0, 이상치 제거 X)

	Accuracy	Recall	Precision	F1-Score
Rogistic	0.4240	0.9858	0.1745	0.2965
SVM	0.4677	0.9787	0.1854	0.3117
Decision Tree	0.5895	0.9220	0.2207	0.3562
Stacking Ensemble	0.4672	0.9752	0.1848	0.3107
DNN	0.4930	0.9894	0.1942	0.3146
LSTM	0.5148	0.7943	0.1754	0.2874

VS

증화추출표본 2:1 Sample

	Accuracy	Recall	Precision	F1-Score
Rogistic	0.9417	0.8375	0.9853	0.9054
SVM	0.9458	0.8375	1.000	0.9117
Decision Tree	0.9375	0.9000	0.9114	0.9057
Stacking Ensemble	0.9708	0.9375	0.9740	0.9554
DNN	0.9625	0.9863	0.9000	0.9412
LSTM	0.7792	0.6667	0.6750	0.6708

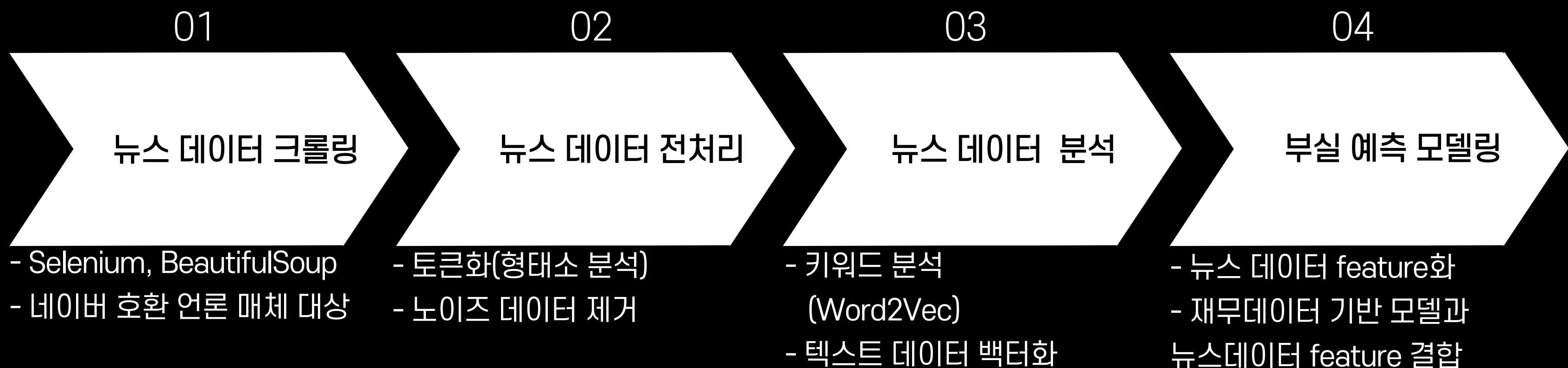
VS

강건성을 위한 raw data (Scaling 0, 이상치 제거 X)

	Accuracy	Recall	Precision	F1-Score
Rogistic	0.4572	0.9787	0.3056	0.4642
SVM	0.5044	0.9539	0.3970	0.5588
Decision Tree	0.4066	0.9504	0.3031	0.4609
Stacking Ensemble	0.4882	0.9929	0.3570	0.5236
DNN	0.4646	0.9965	0.3486	0.5165
LSTM	0.1249	0.9965	0.1961	0.3231

Data Set 2 (뉴스데이터)

1) 뉴스데이터 모델링 과정



36 Data Set 2 (뉴스데이터)

진행과정

2) 뉴스 데이터 크롤링

뉴스 콘텐츠 수집 대상 언론 매체

네이버 뉴스에 호환되는 모든
언론 매체 (네이버 뉴스 홈 >>
전체언론사)

뉴스 콘텐츠 수집 기간

2011~2020년 관리종목으로 지정된
코스닥 상장 기업에 대하여

관리 지정일 ~ 6개월 전

(정상기업의 경우 사업보고서
발표일 ~ 6개월 전)

크롤링 과정



뉴스 검색 기준

- 기사 제목에 기업명이 들어가도록 검색
- 다른 내용의 기사와 혼재하는 경우 네이버 상세검색 기능을 활용하여 “ ”로 묶어서 검색
- 상호변경이 된 기업의 경우 상호변경내역 수집 후, 과거상호명으로 기사 검색

뉴스 검색 제외 기준

- 일상에서 자주 쓰이는 단어가 기업 명인 경우 제외 (예: 지디, 연우, 퓨전, 리드)
- 증권, 경제 기사들과 직접적으로 관련있는 업체 제외 (예: 한국경제티브이, NICE신용평가)
- 기업당 기사의 개수가 10개 미만인 기업은 제외

37 Data Set 2 (뉴스데이터)

진행과정

3) 뉴스 데이터 전처리

토큰화(Tokenization)

Okt 라이브러리 → 명사 추출 (nouns 함수)

Null 값 제거

노이즈 데이터 제거 (정제)

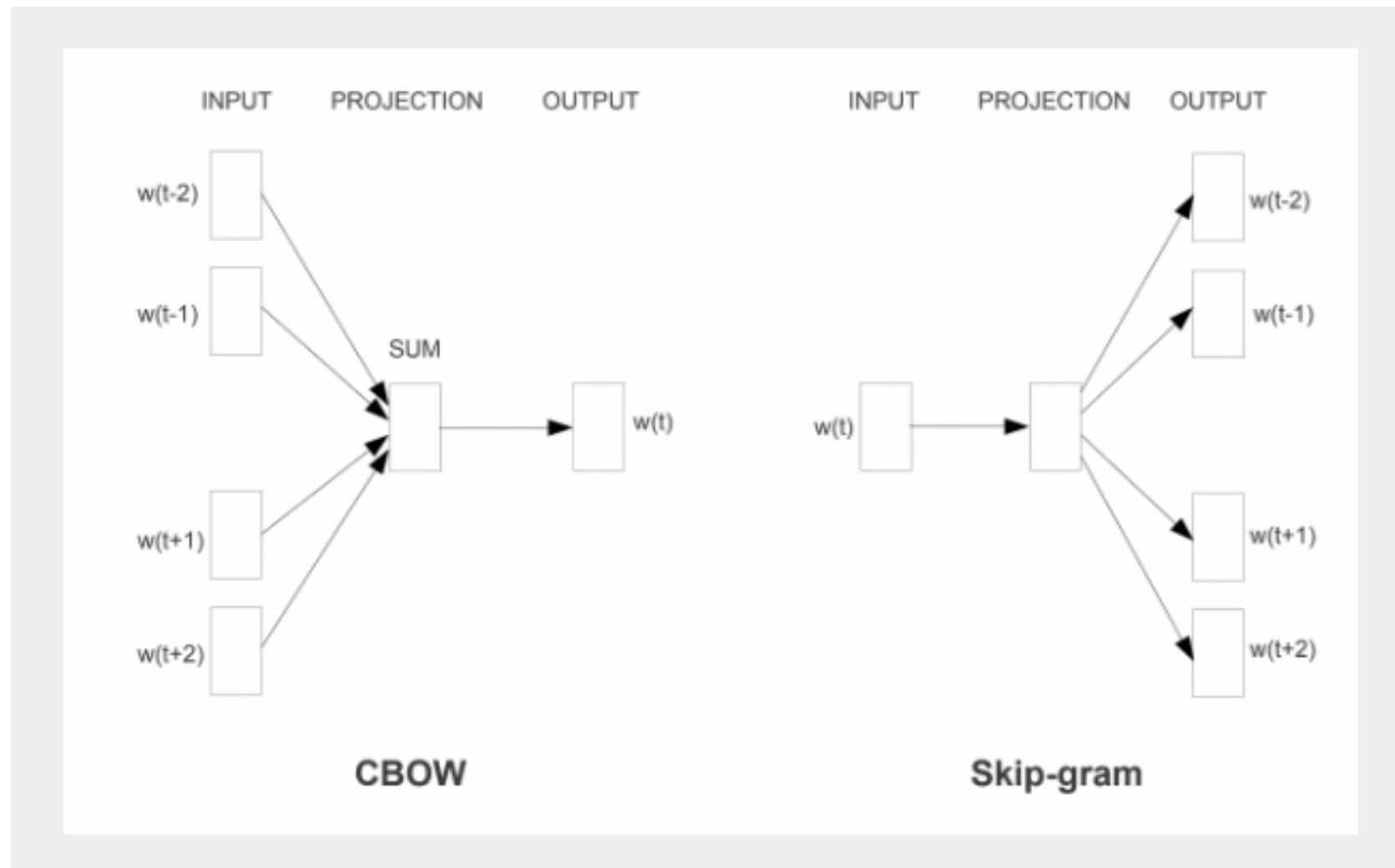
- 검색 빈도 수 10개 이하로 산출되는 키워드 제외
- 동사, 조사, 형용사 어미 형태 모두 제외
- 기업 활동과 상관없는 고유명사 제외 - 특정회사/제품/인물 이름
- 길이가 짧은 단어들을 제거 (1 글자 이하의 단어 제외)
- 불용어사전(korean_stopword.txt)을 정의하여 제거
- 특수문자 제거

38 Data Set 2 (뉴스데이터)

진행과정

4) 뉴스 데이터 분석 - 키워드 추출

Word2Vec : 단어 벡터 간 유의미한 유사도를 계산



Word2Vec를 활용하여
'관리', '관리종목'과 유사단어(키워드) 추출

단어	유사도	단어	유사도
정사유	0.8415	부실	0.5966
미달	0.8035	공적자금	0.5905
지정	0.6632	리스크	0.5416
우려	0.6473	워크아웃	
자본잠식	0.6338	자금난	

39 Data Set 2 (뉴스데이터)

진행과정

4) 뉴스 데이터 분석 - 데이터 벡터화 (TF-IDF)

TF-IDF → Feature 1

TF : 한 문서 내에 특정 단어가 나타나는 빈도수

IDF : $\log(\text{전체문서}/\text{특정 단어가 나온 문서})$

TF-IDF : 특정 단어가 한 문서에 몇 번 언급되며, 문서군에서는 얼마나 등장하였는지를 표현한 가중치



TF : Word2Vec 추출한 키워드 빈도수

IDF : $\log(1/\text{관리기사비율})$

TF-IDF =Word2Vec으로 구한 특정 단어들의 빈도수 * $\log(\text{부도기사비율})$

* IDF = inf or Nan 인 경우의미 없다고 판단하여 0

값으로 대체

40 Data Set 2 (뉴스데이터)

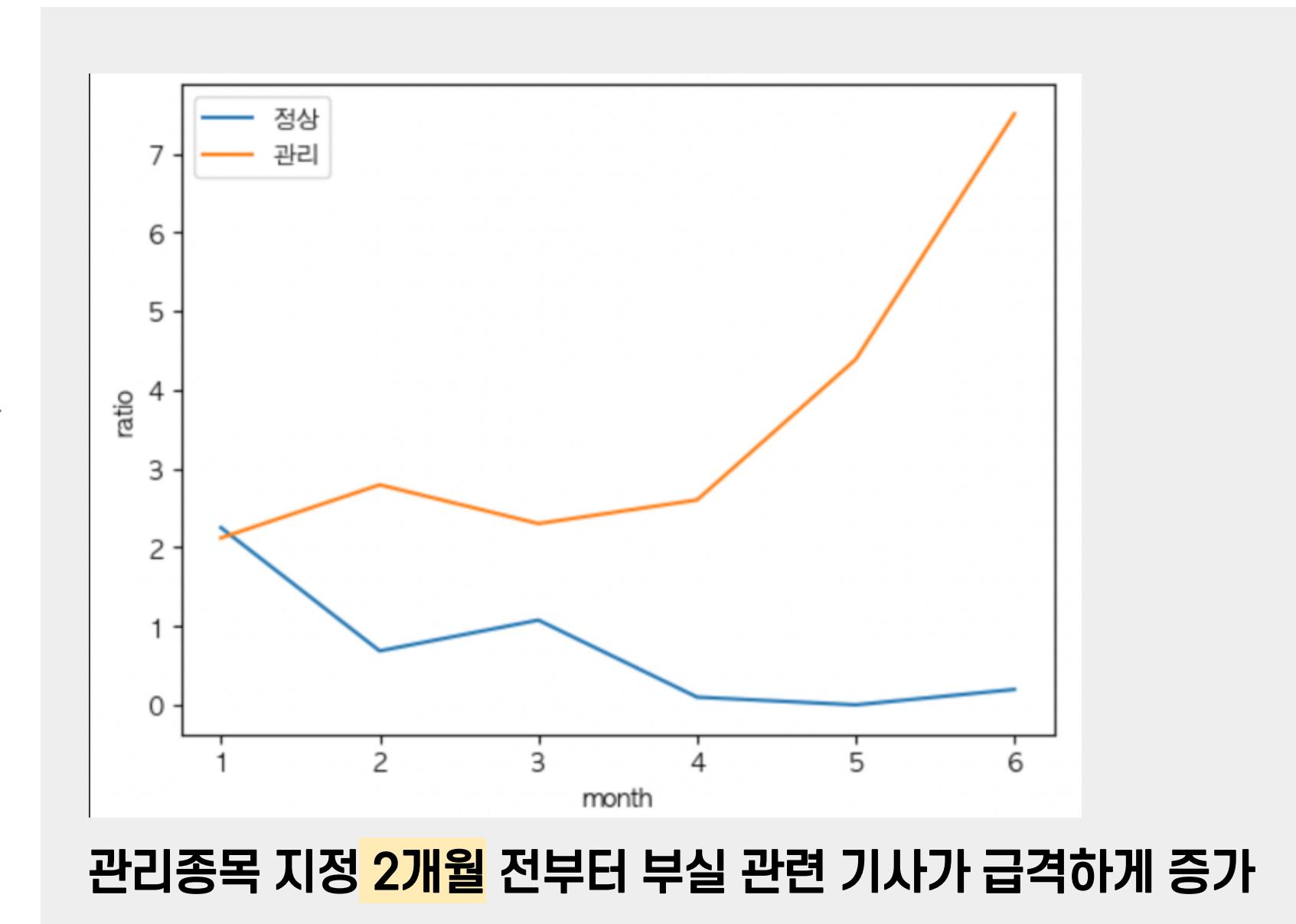
진행과정

4) 뉴스 데이터 분석 - 부실 관련 기사 비율

Feature 2

부실 관련 기사 비율

(부실 키워드 포함된 해당기업 기사수/해당기업 전체 기사 수)



41 Data Set 2 (뉴스데이터)

진행과정

5) 뉴스 데이터 feature화

Data set 2 부실기업 예측 모델 feature

종목명	관리기사비율	TF-IDF	Logit	...	LSTM
CMG제약	9.3750	3.084	1	...	0
국일신동	0	0	1	...	1
...
현진소재	0	0	0	...	1
협진	61.2903	1.2756	0	...	0

재무데이터 예측값 :

train : 훈련된 모델의 훈련데이터 예측 값 이용

test : 훈련된 모델의 평가데이터 예측 값 이용 y_test 예측

최종 변수 (8개)

1. 관리기사비율
2. TF-IDF
3. 재무데이터 Logit 모델 예측 결과
4. 재무데이터 SVM 모델 예측 결과
5. 재무데이터 Decision Tree 모델 예측 결과
6. 재무데이터 Stacking Ensemble 모델 예측 결과
7. 재무데이터 DNN 모델 예측 결과
8. 재무데이터 LSTM 모델 예측 결과

42 Data Set 2 (뉴스데이터)

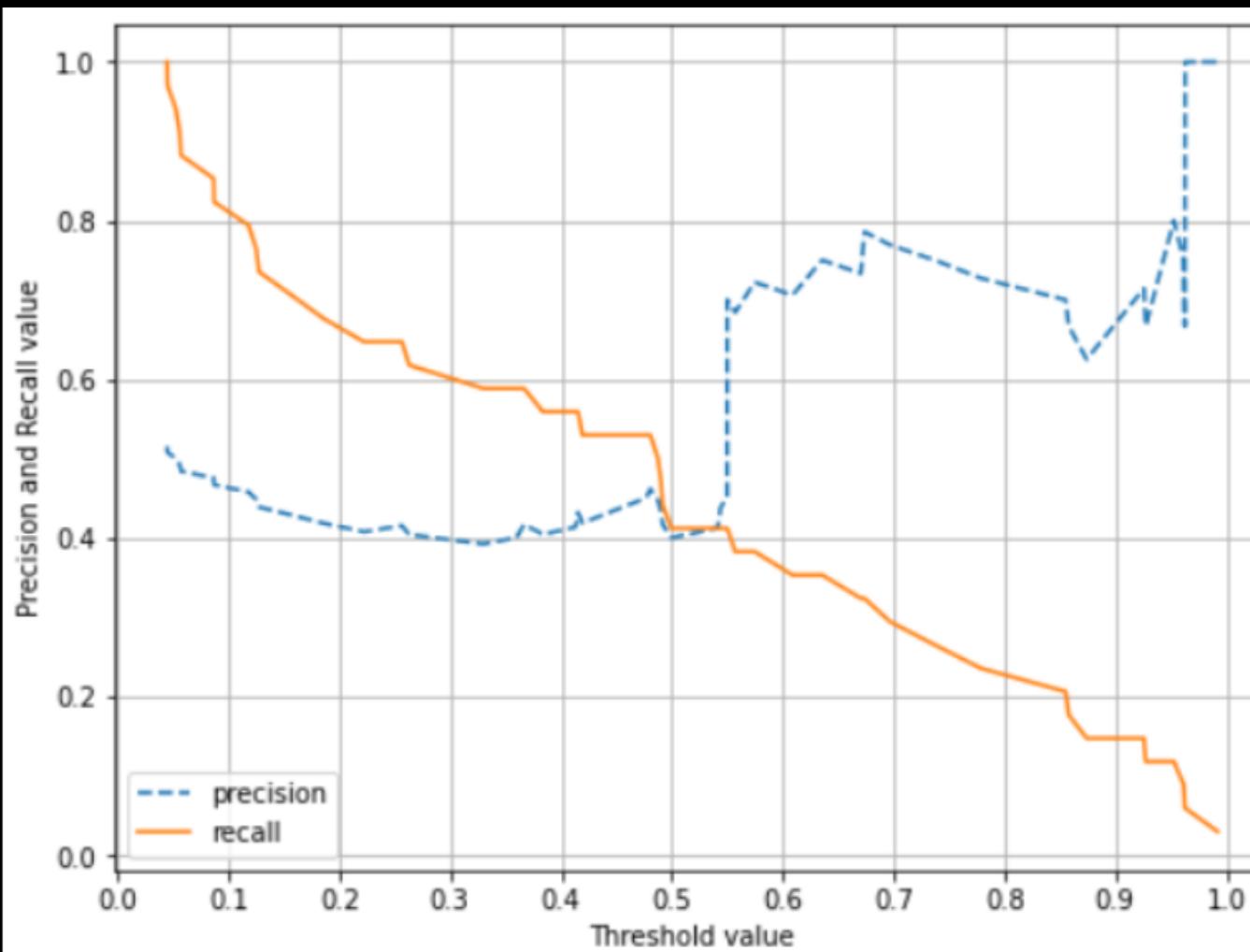
진행과정

6) Model Grid Search 활용 Logit

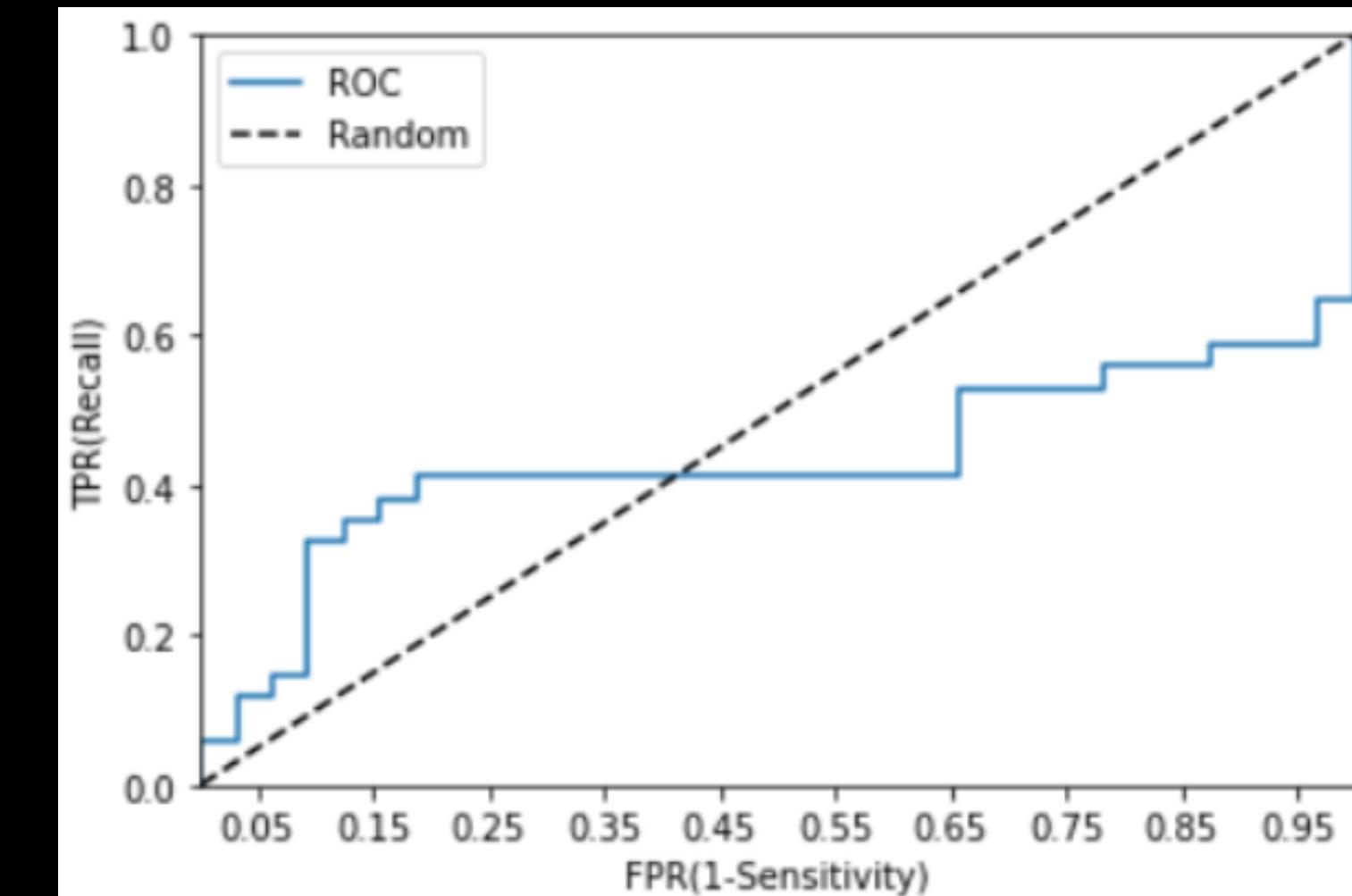
Hyper parameter : C = 1, penalty = 'l2'

정확도 : 0.3787878787878788
정밀도 : 0.40540540540540543
재현율 : 0.4411764705882353
f1 score : 0.4225352112676056

임계값에 따른 precision , recall



ROC curve



43 Data Set 2 (뉴스데이터)

진행과정

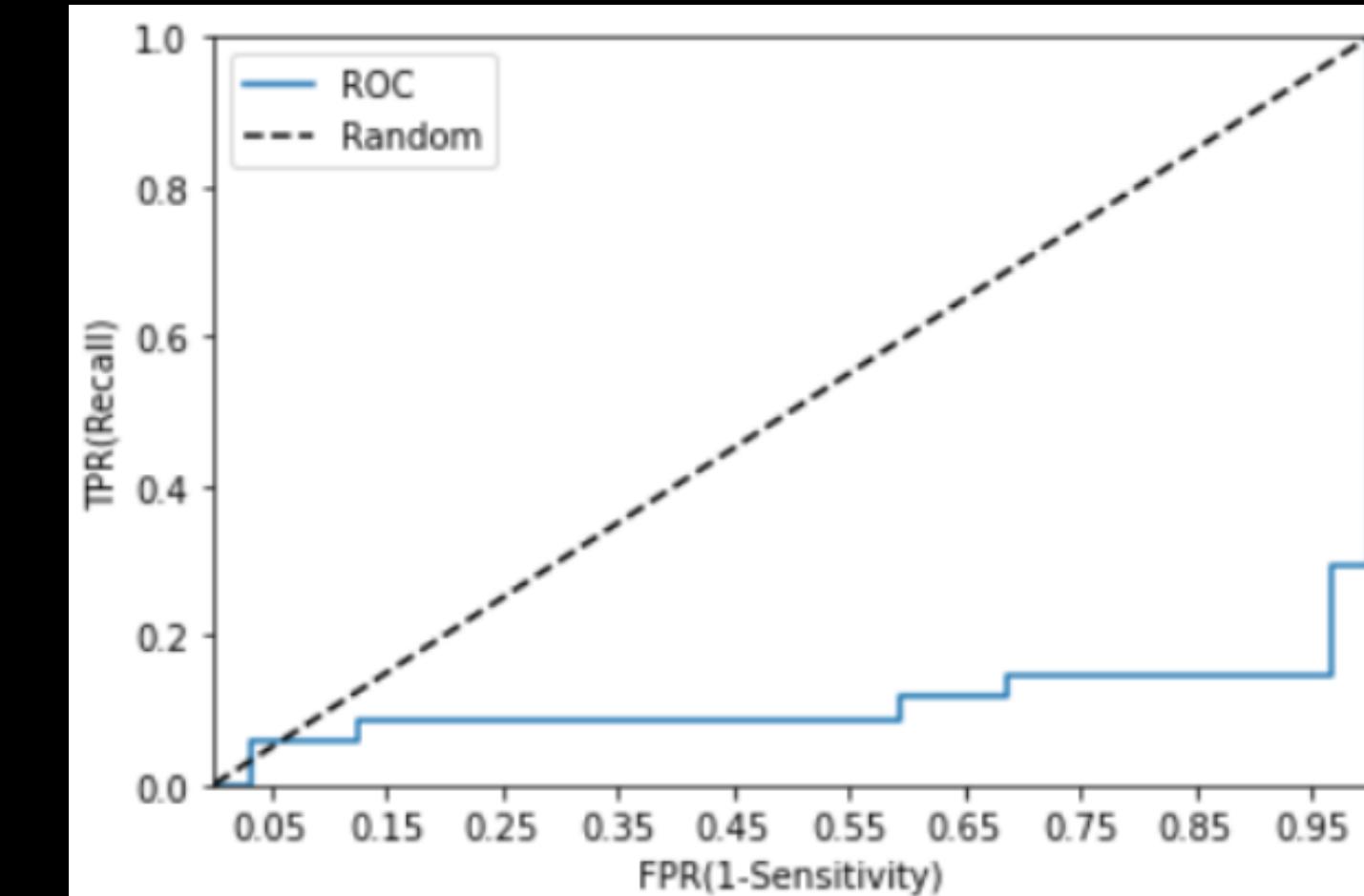
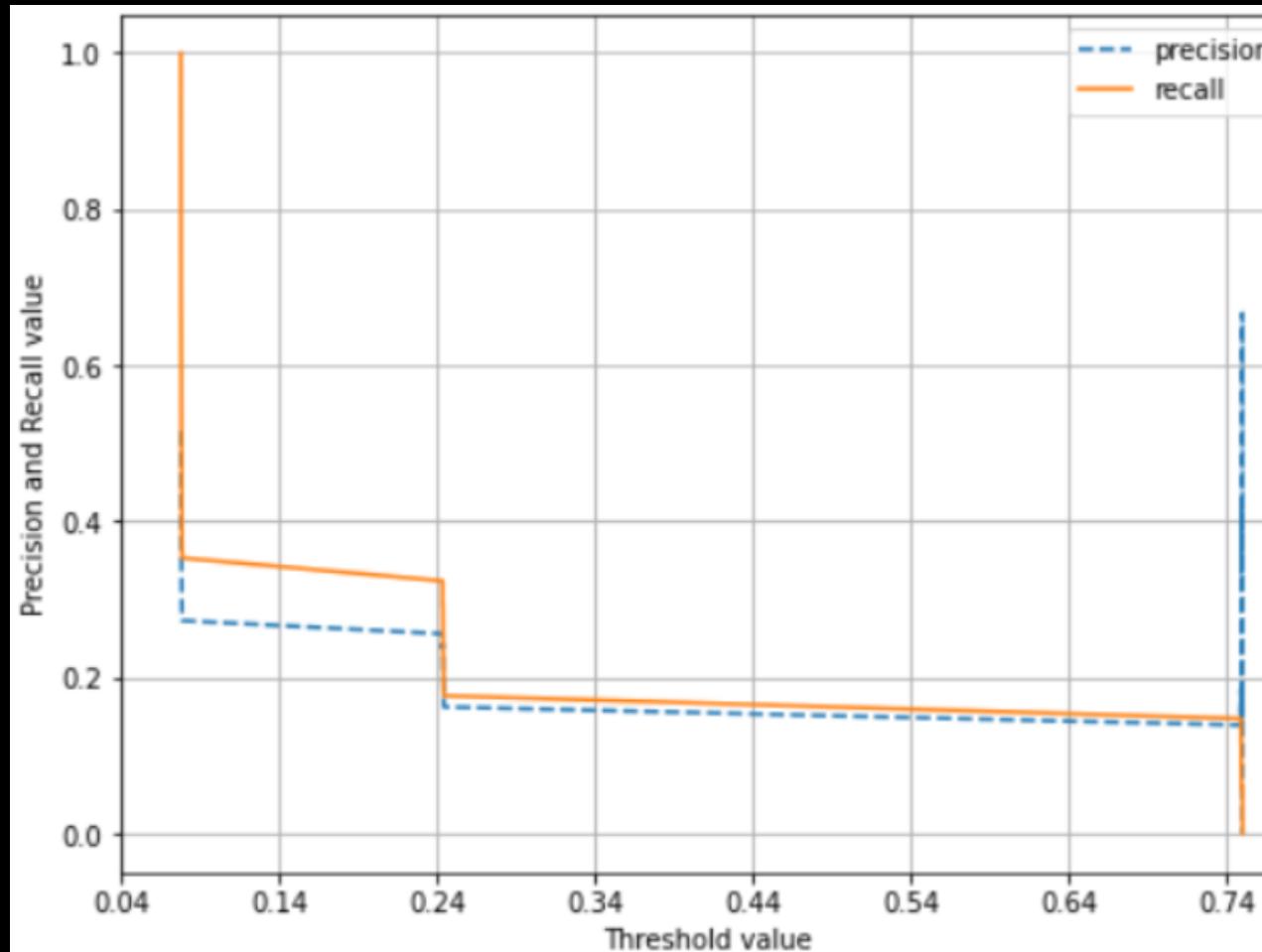
6) Model Grid Search 활용 SVM

Hyper parameter : C = 0.25, gamma = 0.1

정확도 : 0.1666666666666666
정밀도 : 0.24390243902439024
재현율 : 0.29411764705882354
f1 score : 0.2666666666666666

임계값에 따른 precision , recall

ROC curve



44 Data Set 2 (뉴스데이터)

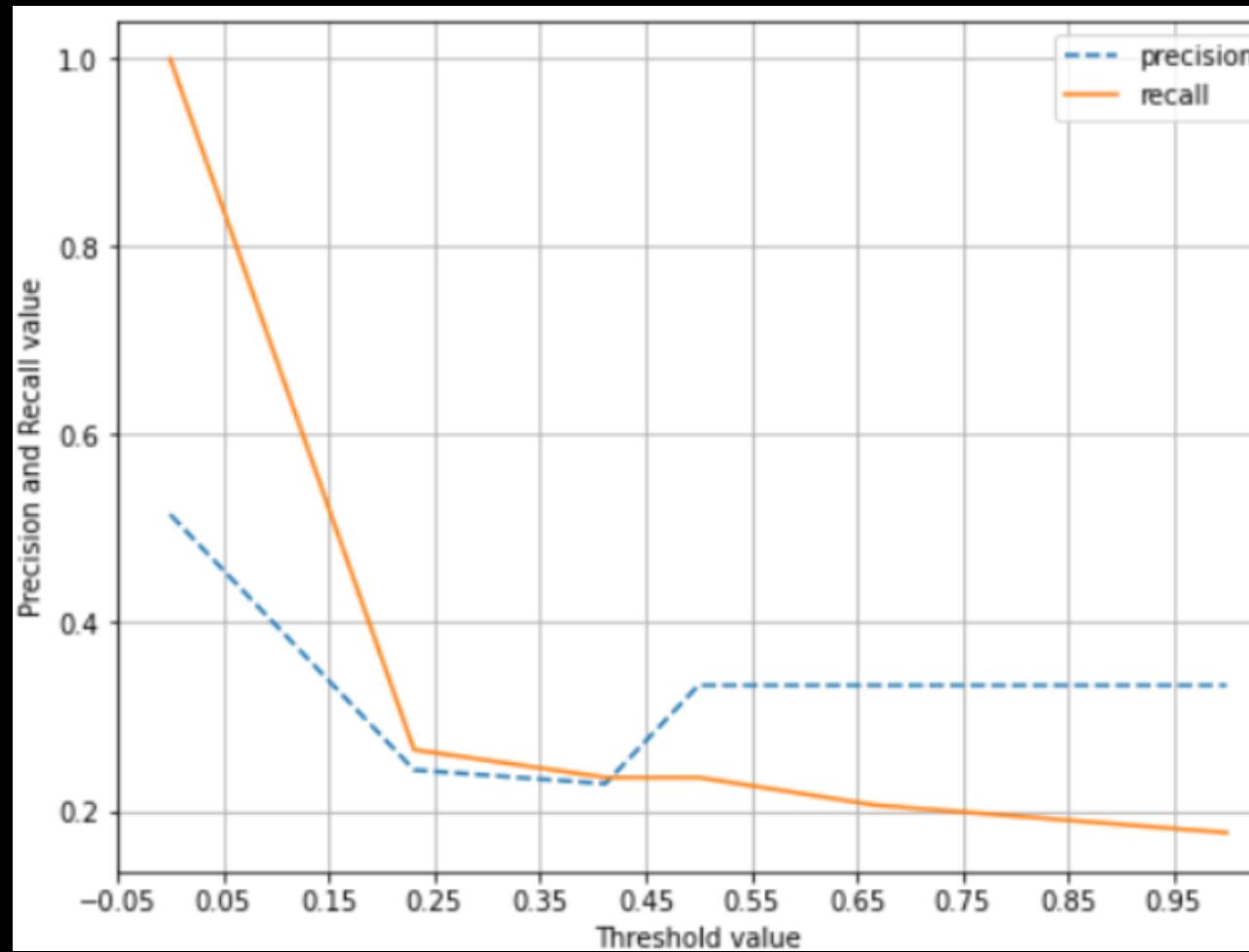
진행과정

6) Model Grid Search 활용 Decision Tree

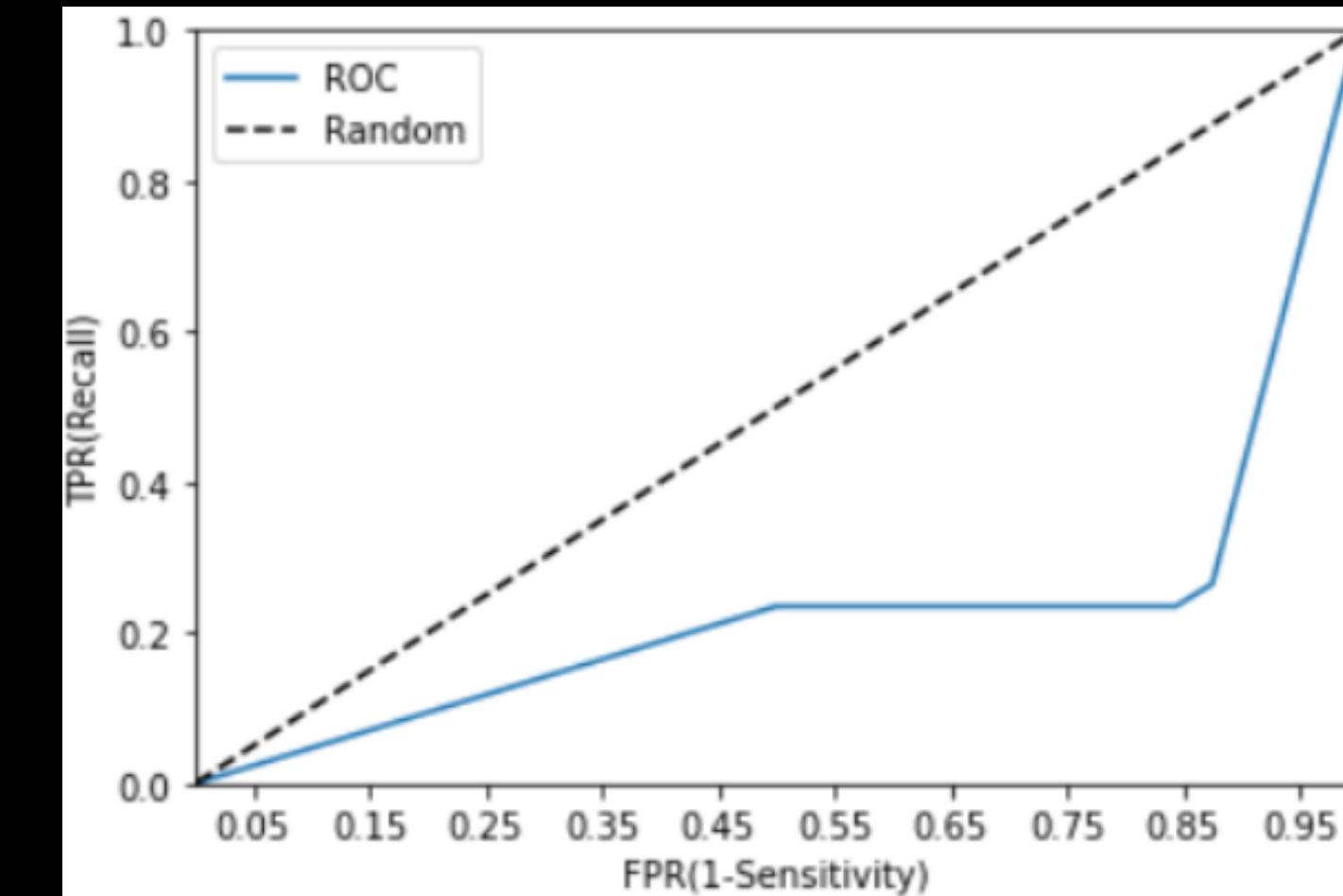
Hyper parameter : criterion = 'entropy', 'max_depth' = 10,
max_features = 'log2', min_sample_split = 6

정확도 : 0.4090909090909091
정밀도 : 0.3684210526315789
재현율 : 0.20588235294117646
f1 score : 0.2641509433962264

임계값에 따른 precision , recall



ROC curve



45 Data Set 2 (뉴스데이터)

진행과정

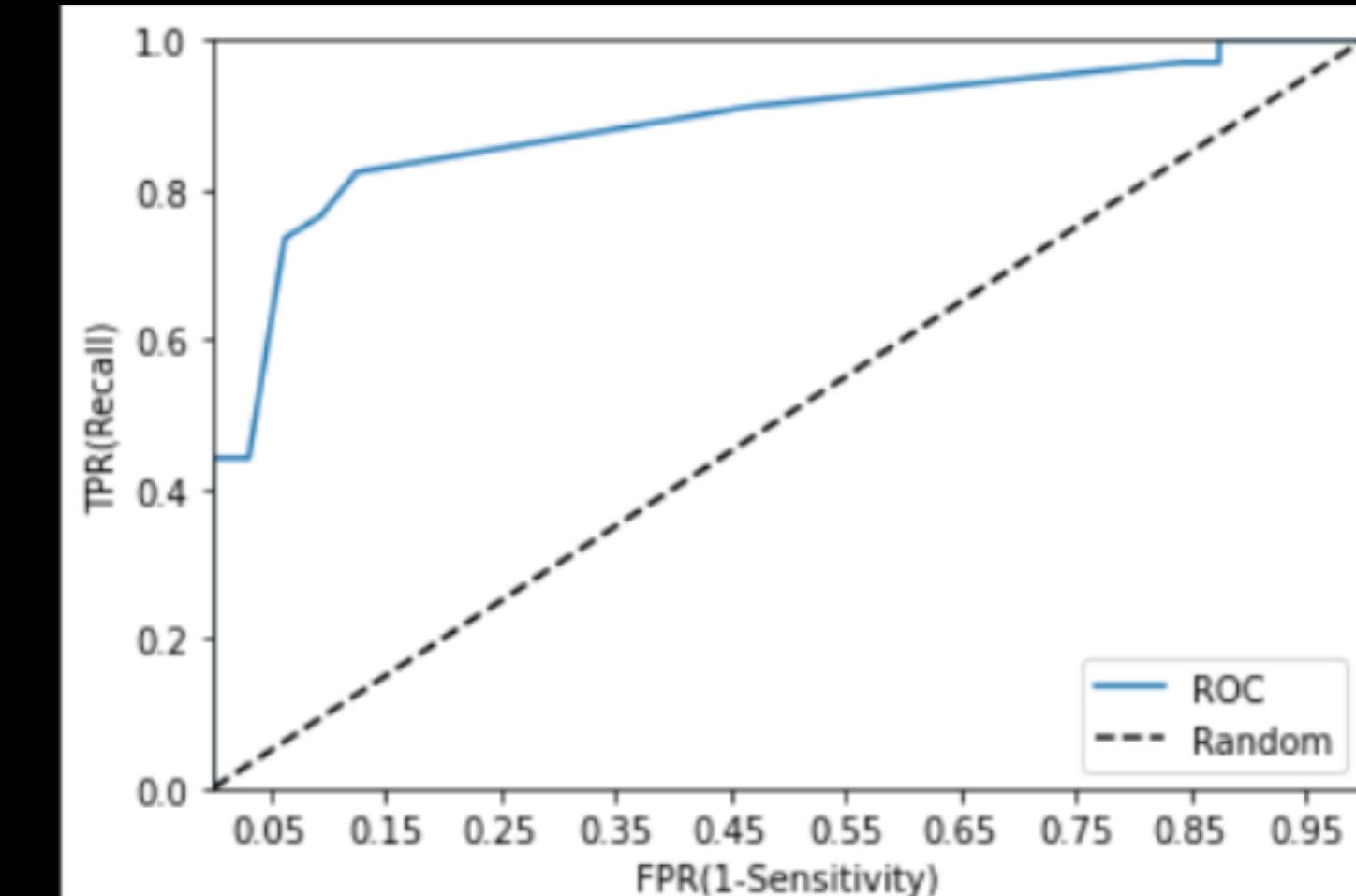
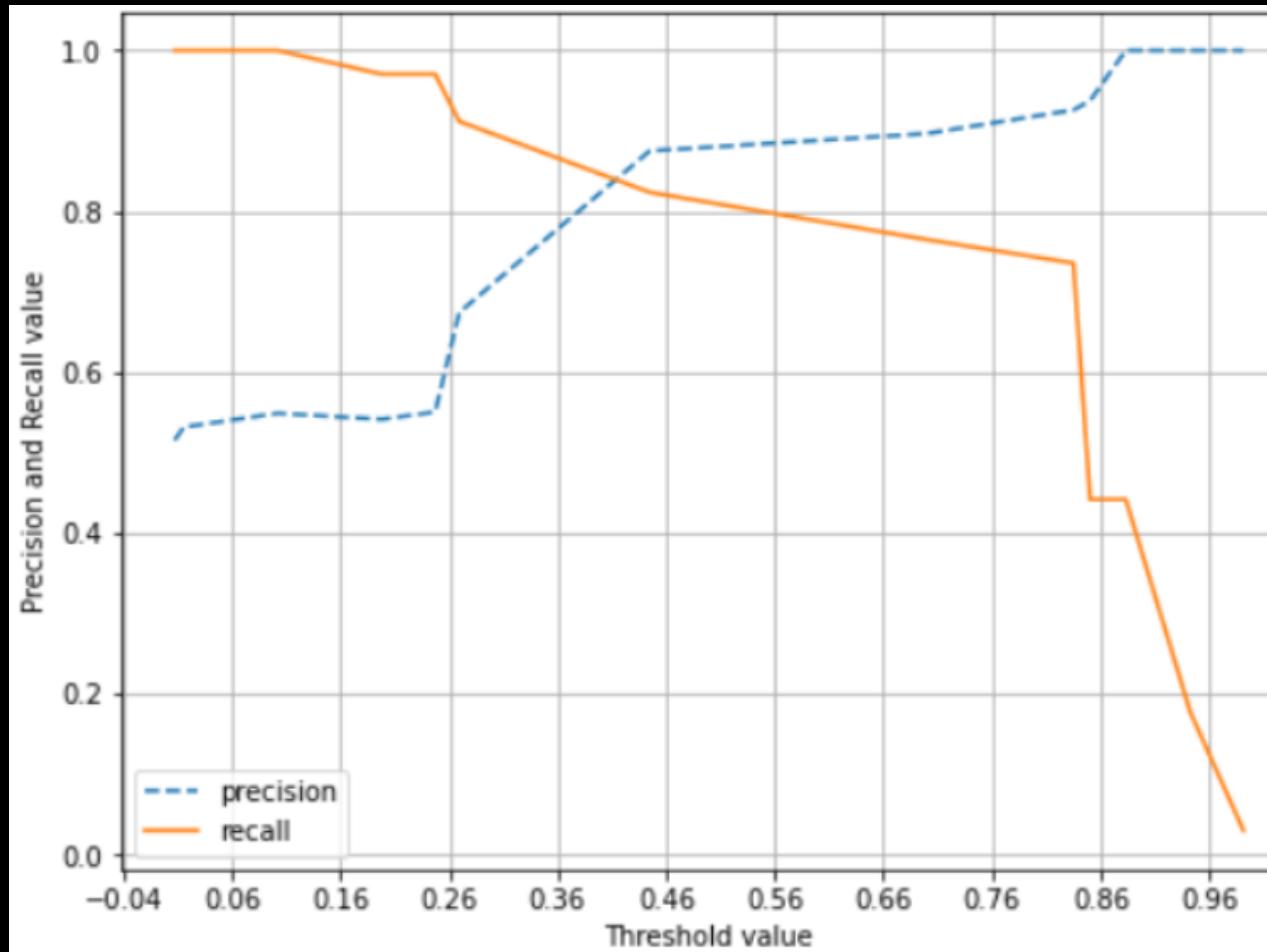
6) Model Stacking Ensemble

```
knn_clf = KNeighborsClassifier(n_neighbors=4)
rf_clf = RandomForestClassifier(n_estimators=100, random_state=0)
dt_clf = DecisionTreeClassifier(random_state=125)
ada_clf = AdaBoostClassifier(n_estimators=100)
```

정확도 : 0.8484848484848485
정밀도 : 0.9117647058823529
재현율 : 0.8157894736842105
f1 score : 0.8611111111111111

임계값에 따른 precision , recall

ROC curve



46 Data Set 2 (뉴스데이터)

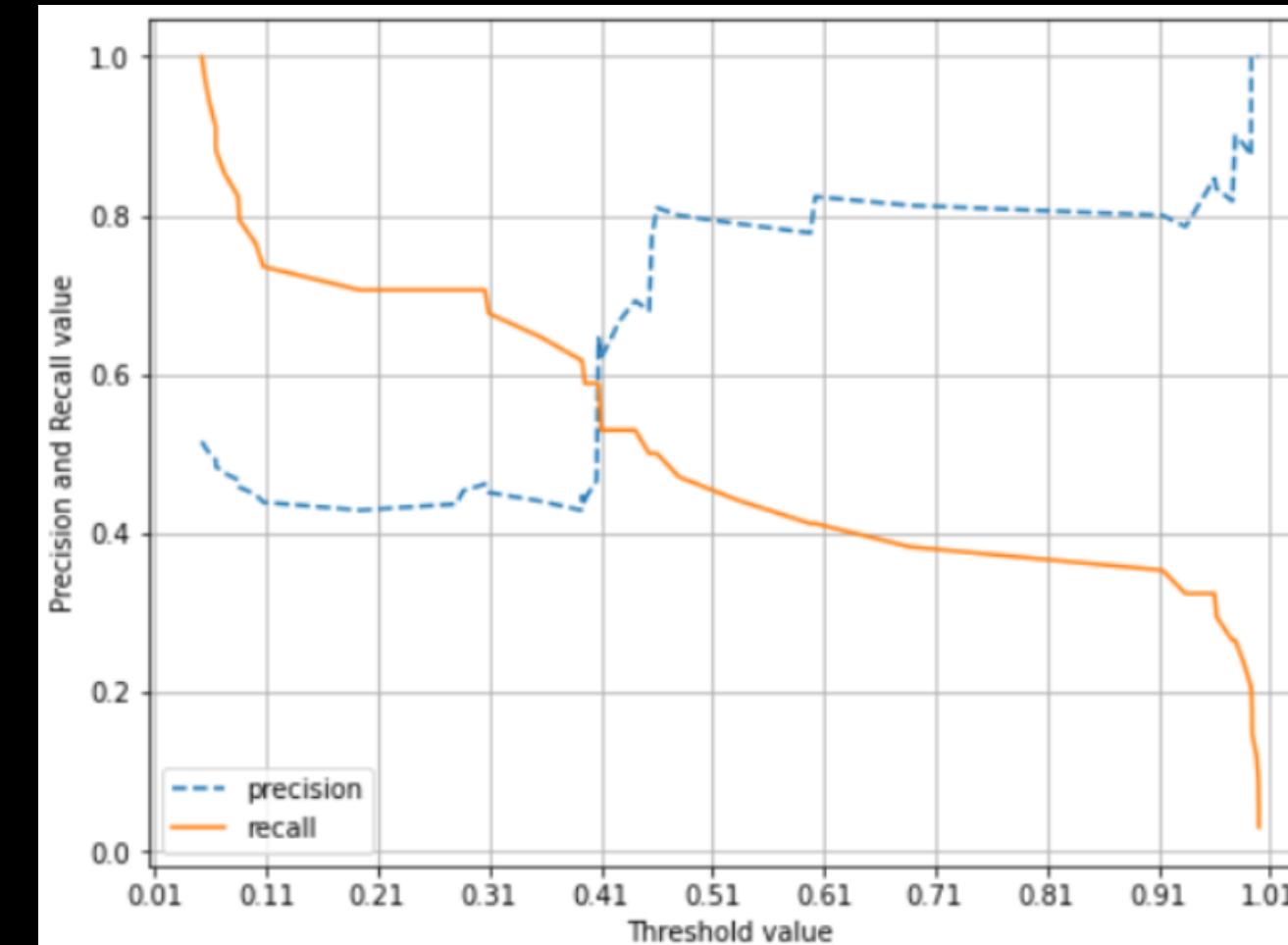
진행과정

6) Model DNN 과적합 방지를 위해 Dropout 추가

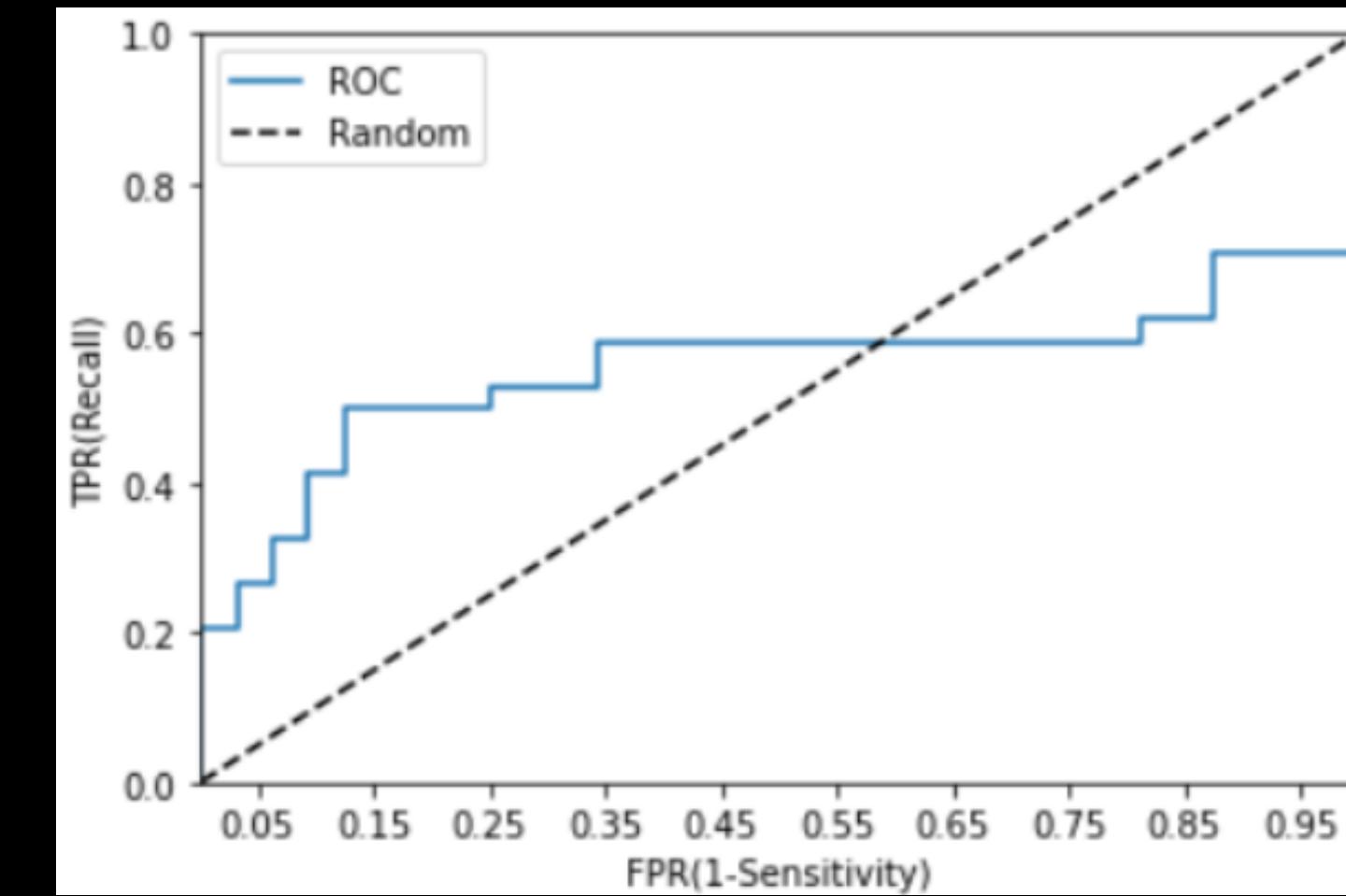
Hyper parameter : dropout = 0.3, activation = 'relu',
은닉층 수 = 2

정확도 : 0.6666666666666666
정밀도 : 0.5882352941176471
재현율 : 0.7142857142857143
f1 score : 0.6451612903225806

임계값에 따른 precision , recall



ROC curve



47 Data Set 2 (뉴스데이터)

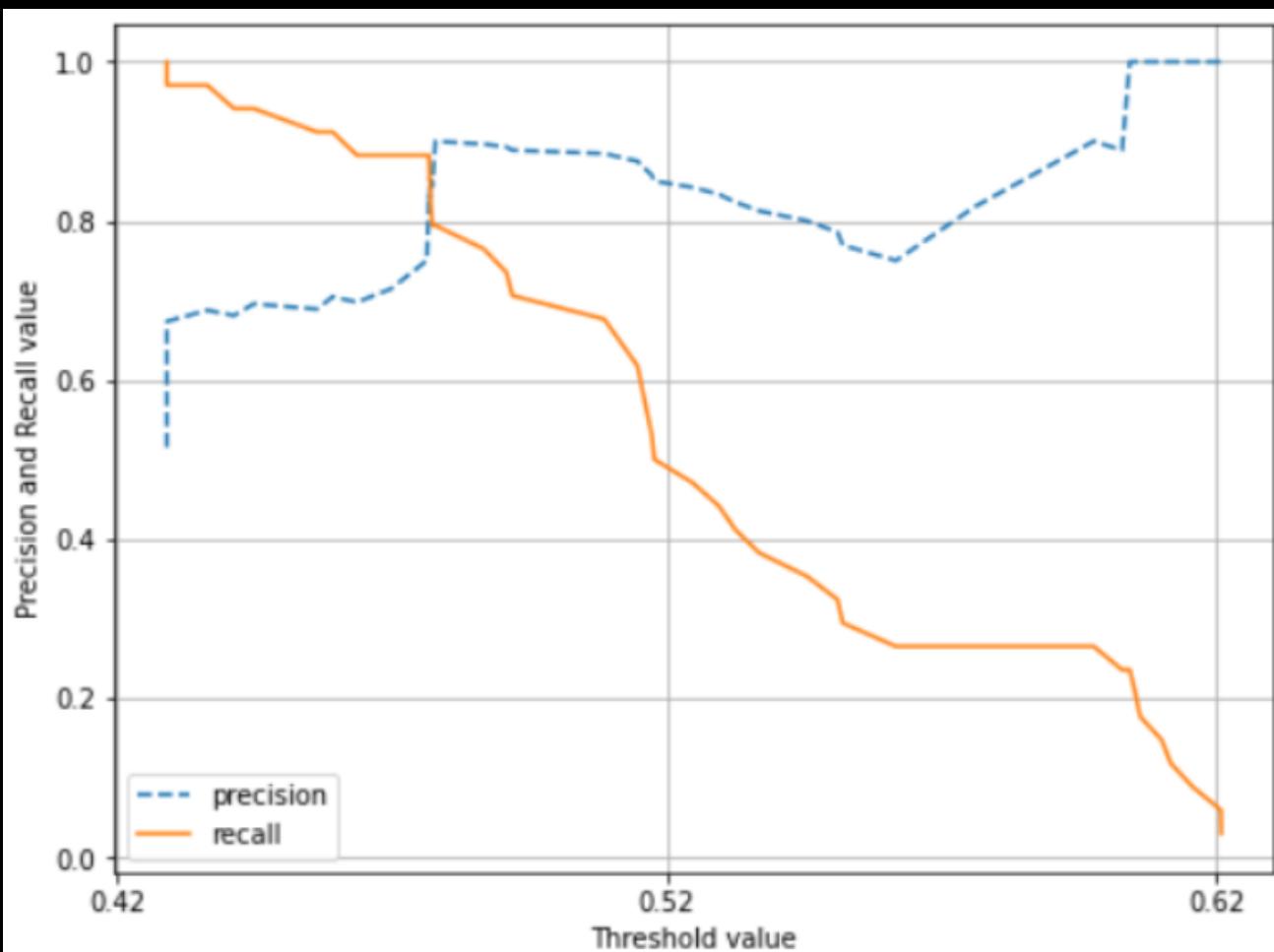
진행과정

6) Model LSTM 과적합 방지를 위해 Dropout 추가

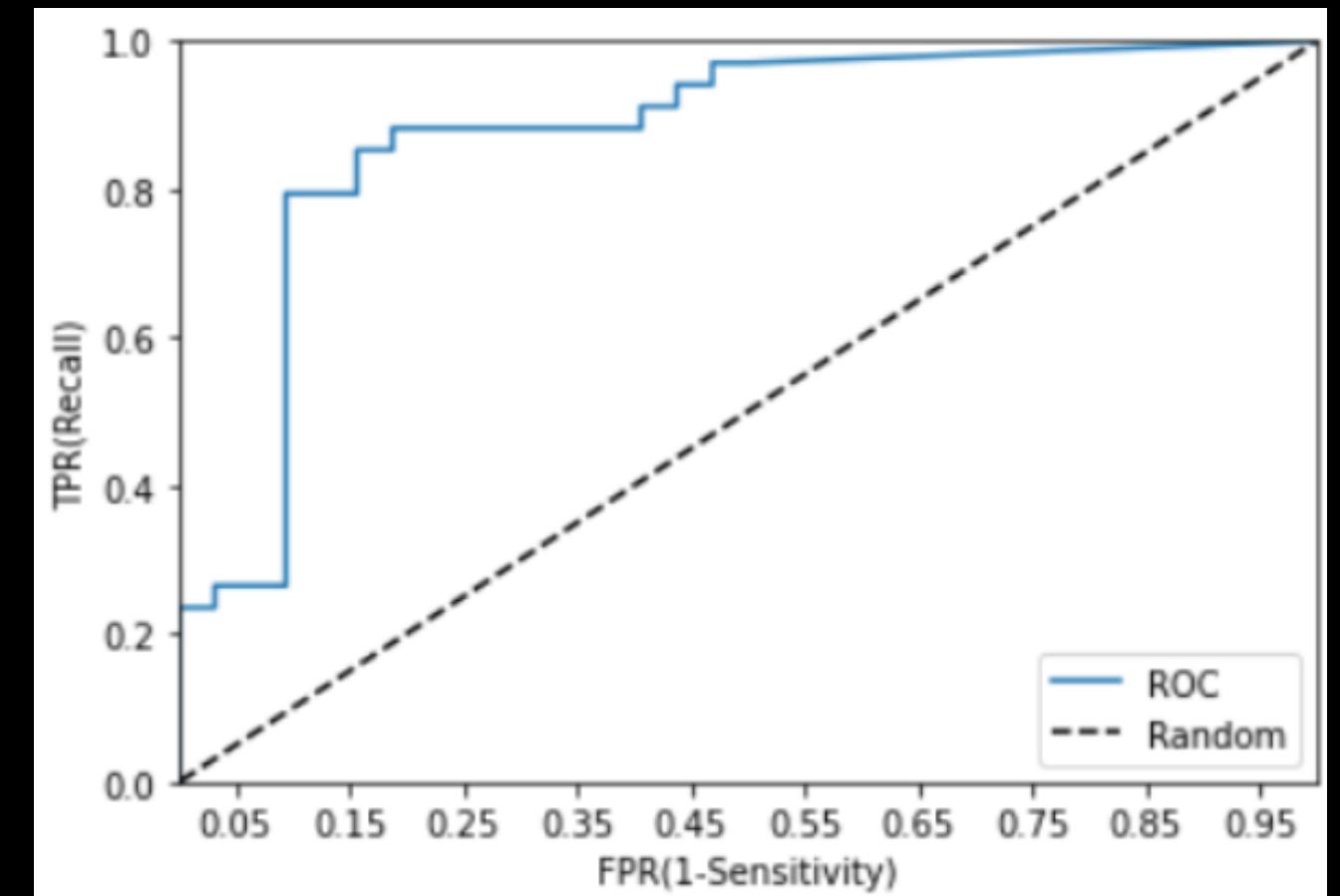
Hyper parameter : dropout = 0.3, activation = 'relu',
은닉층 수 = 2

정확도 : 0.7878787878787878
정밀도 : 0.6764705882352942
재현율 : 0.8846153846153846
f1 score : 0.7666666666666666

임계값에 따른 precision , recall



ROC curve



Data Set 3 (채용플랫폼 데이터)

1) 채용플랫폼 데이터 분석 과정

채용플랫폼 선정

잡플래닛

크롤링

Selenium,
BeautifulSoup

*잡플래닛으로 선정한 이유:
법적 제한이 없고
데이터의 양이 가장 많음

구글 api를 활용한 영문 번역

yi-yangkust 활용하여 정상기업과
관리종목 '경영진에게' 감성분석

경영악화된건 잖은 할인행사로 수익성이 떨어진 것과 자금관리를 허술하게 해서인듯 합니다.
늘 인력난에 시달린다. 인력충원이 급함. 인력이 부족해서 하는일은 많은데 급여는 너무 적음, 급여 및 복
동종업계와 비슷하게 금액적인 보상을 조금 더 신경써주시면 고맙겠습니다
경영진도 힘있어보이진 않음 꾸역꾸역 회사 생명 유지하고있긴하나 미래는 불투명

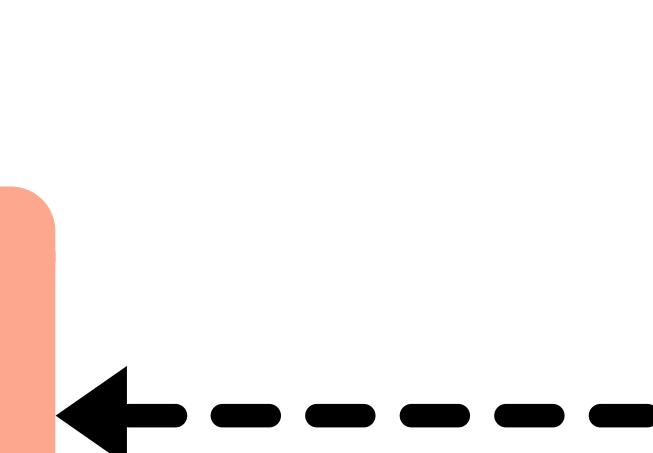
The deteri	Negative	1
I always su	Negative	0.99978
I would ap	Negative	0.977475
The mana	Negative	0.999976

feature화

0 : 부정

1 : 중립

2 : 긍정



49 Data Set 3 (채용플랫폼 데이터)

진행과정

2) feature화 과정

변수 1

- 0 : 부정
- 1 : 중립
- 2 : 긍정



원-핫 인코딩 * score



flatten + PCA
PCA : 최소 리뷰 개수 * 3

변수 2

- 0 : 긍정
- 1 : 중립
- 2 : 부정



수치화 된 감성변수 * score

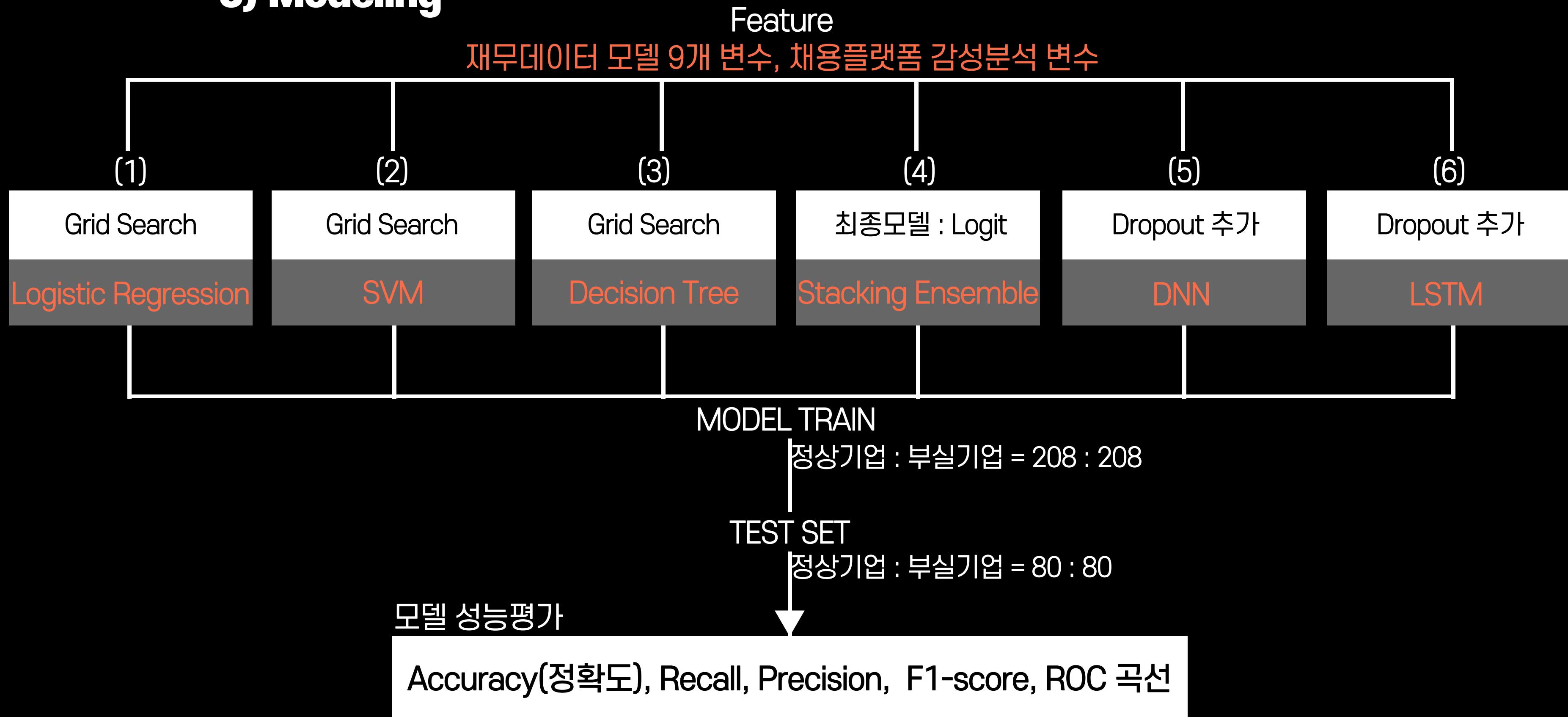


기업별 평균 도출

50 Data Set 3 (채용플랫폼 데이터)

진행과정

3) Modeling



51 Data Set 2, 3 (뉴스데이터 / 채용플랫폼)

결론 도출

4) Model 성능평가 비교

재무+뉴스 모델 1: 재무데이터 모델 예측값 + 뉴스데이터 변수

	Accuracy	Recall	Precision	F1-Score
Logistic	0.3788	0.4054	0.4412	0.4225
SVM	0.1667	0.2439	0.2941	0.2667
Decision Tree	0.4090	0.3684	0.2059	0.2642
Stacking Ensemble	0.8484	0.9117	0.8156	0.8611
DNN	0.6667	0.5882	0.7143	0.6452
LSTM	0.7879	0.6765	0.8846	0.7667

재무+뉴스 모델 2: 재무변수 9개 + 뉴스데이터 변수

	Accuracy	Recall	Precision	F1-Score
Logistic	0.9090	0.9667	0.8529	0.9063
SVM	0.8636	0.9629	0.7647	0.8525
Decision Tree	0.8030	0.8182	0.7941	0.8059
Stacking Ensemble	0.8636	0.9032	0.8235	0.8615
DNN	0.8939	0.8824	0.9091	0.8955
LSTM	0.7576	0.9706	0.6875	0.8049

VS

재무+플랫폼 모델1 : 재무변수 9개 + 감성분석 변수1

	Accuracy	Recall	Precision	F1-Score
Logistic	0.8592	1	0.8148	0.8979
SVM	0.8592	1	0.8148	0.8979
Decision Tree	0.6056	0.8421	0.5926	0.6957
Stacking Ensemble	0.8592	0.9231	0.8889	0.9057
DNN	0.7606	1	0.7606	0.8640
LSTM	0.7887	0.8704	0.8545	0.8624

VS

재무+플랫폼 모델 2: 재무변수 9개 + 감성분석 변수2

	Accuracy	Recall	Precision	F1-Score
Logistic	0.9014	0.9608	0.9074	0.9333
SVM	0.8873	0.9792	0.8704	0.9216
Decision Tree	0.8732	1	0.8333	0.9091
Stacking Ensemble	0.8873	0.8966	0.9629	0.9286
DNN	0.9296	0.9444	0.9623	0.9533
LSTM	0.5634	0.5	0.8709	0.6353

52 의의 및 한계점

결론 도출

의의



- 2:1 Sampling 재무데이터 모델에서 **데이터 불균형**이 강간성 검증의 과소적합을 발생시킨다는 것을 확인

- 적시성을 보완하기 위한 비정형 데이터 사용시 **LSTM 모형**이 뛰어남을 검증함

채용플랫폼을 중심으로

- 기존의 선행논문과는 차별화된 데이터를 활용
- **적시성** 보완
- 기업 내부의 상황을 잘 아는 **근로자의 평가**를 반영 가능

한계



- 비정형데이터의 데이터 수집 단계의 한계로 (ex. 일반적 단어, 광고 포함 등) 모델에서 과소적합이 발생하는 경우가 생김

- 온전히 비정형 데이터로만 이루어진 모델을 만드는데는 아직 예측에 한계가 있음을 발견

채용플랫폼을 중심으로

- 텍스트 그 자체로는 극성(긍정이나 부정)을 가지지 않는 표현이 존재하여 **감성분석의 한계** 직면
- **차원축소** 과정으로 인해 텍스트에 대한 논리적인 설명이 불가
- 감성사전을 직접 작성하지 못하여 모델에 최적화된 분석을 하지 못함

참고 논문

- "기업 리뷰 웹 사이트 텍스트 분석을 통한 직원 불만 표현 추출과 불만 원인 도출 및 해소 방안" 백혜연, 박용석 (2019)
- "빅데이터를 이용한 딥러닝 기반의 기업 부도예측 연구" 최정원, 오세경, 장재원 (2017)
- "텍스트마이닝 방법론을 활용한 기업 부도 예측 연구" 최정원, 한호선, 이미영, 안준모 (2015)
- "실제 사례 기반 비정형 데이터를 활용한 기업의 부실징후 예측에 관한 효용성 연구" 진 훈, 홍정표, 이강호, 주동원 (2018)
- "더블 앙상블 기법을 이용한 기술사업성평가 기반의 중소기업 부실 예측 연구" 이상훈, 유동희 (2022)
- "코스닥시장에서의 상장폐지위험과 이익조정" 손성규, 염지인 (2013)
- "비재무정보를 이용한 창업기업의 부실요인에 관한 실증연구" 남기정, 이동명, 진로 (2019)
- "회계정보와 시장정보를 이용한 부도예측모형의 평가 연구" 이인로, 김동철 (2015)
- "머신러닝 기반 KOSDAQ 시장의 관리종목 지정 예측 연구: 재무적 데이터를 중심으로" 윤양현, 김태경, 김수영 (2022)
- "관리대상종목의 수익률과 위험 속성에 관한 연구" 김태혁, 엄철준 (1997)
- "관리종목 기업의 회계정보 효과" 손성규, 오명전 (2008)
- "AUROC기반의 부도예측 앙상블 모형" 윤우섭, 김명종 (2021)
- "감사의견, 감사법인 및 기업부실리스크의 예측" 김경철, 김용덕 (2022)
- "기업부실예측모형에서 비재무적 정보의 유용성에 관한 연구" 민아영 (2021)

감사합니다
Thank you!

