머신러닝 기반 기업부도위험 예측모델 검증 및 정책적 제언: 스태킹 앙상블 모델을 통한 개선을 중심으로

저자: 엄하늘, 김재성, 최상옥

부도추격자

이정환, 이윤지, 박정호, 김영주

팀원소개

리더:이정환

시각화 마스터:김영주

도메인 마스터:박정호

통계 마스터:이윤지

CONTENTS

01 논문선정

02 데이터탐색

03 분석방법

04 실증결과 및 검증

05 기여효과 및 한계점

PART 1

논문선정

1) 선택논문요약
2) 선행연구
3) 동기
4) 논문선정이유

논문선정 1) 선택논문요약

목적	머신러닝 기반 부도위험 예측모형의 안정적인 예측력 확보를 위한 방안 제안 부도위험 예측 모형의 도입 기준과 정책 수립에도 활용
부도정의	기업의 부도확률을 부채변제 만기시점에서 가산가치가 부채가치보다 적을 확률로 계산
결론	스태킹 앙상블 모델의 부도확률 예측력이 가장 우수함

Keyword: 부도위험 예측, 스태킹 앙상블 모델, 머튼 모형, 랜덤 포레스트, 합성곱 신경망

논문선정 2) 선행연구

연구자	연도	제목	사용알고리즘 및 요약
Horrigan	1966	The Determination of Long-Term Credit Standing with Financial Ratios. Journal of Accounting	[다중회귀분석]독립변수: 재무비율, 채무변제 우선순위 종속변수: 신용상태
Altman	1968	Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy	[판별함수]운전자본/총자산, 이익잉여금/총자산, 영업이익/총자산, 자기자본 시장가치/총부채의 장부가치, 매출액/총자산 5개 재무비율
Ohlson	1980	Financial Ratios and the Probabilistic Prediction of Bankruptcy	[로짓분석]Altman 판별분석의 z-score 모형 개선
Zmijewski	1984	Methodological Issues Related to the Estimation of Financial Distress Prediction Models	[프로빗 모형]Altman 판별분석의 z-score 모형 개선
Jo, Ji., Ho	1998	A Study on Credit Rating System in Korea Bond Market	우량 제조업과 우량 기업으로 표본을 구분하고Altman 모형 개선한 kems1과 kems2 모형 제시
Wang, H., Q. Xu and L. Zhou	2015	Large Unbalanced Credit Scoring Using Lasso-Logistic Regression Ensemble	불균형 데이터 상에서 신용위험 예측 모델로 Lasso-logistic regression learning ensemble이 CART, LLR, RF보다 우수
Min, S. H.	2014	Bankruptcy prediction using an improved bagging ensemble	부도위험 예측에 있어 사례선택을 활용한 배깅(Bagging) 모형이 기존 SVM보다 예측력이 뛰어남
Cha, S., J	2018	Corporate Default Prediction Model Using Deep Learning Time Series Algorithm, RNN and LSTM	딥러닝 시계열 알고리즘인 RNN과 LSTM 기반의 부도예측모형이 다른 알고리즘에 비해 성능이 우수

논문선정 3) 동기

01 기존 연구는 회계적 재무정보 이외의 시장에서 평가하고 있는 부도위험을 충분히 반영하지 못함

02 희소한 부도사건을 오버 샘플링이나 언더 샘플링을 할 때 정보 왜곡이 발생

03 기존의 연구가 특정한 예측 모델에 기반하고 있기에 존재하는 편향을 제거하지 못함

논문선정 4) 논문선정이유

01 앙상블 기법 중 하나인 스태킹 앙상블에 대해 학습 할 기회를 얻기 위함

02 데이터 분석 결과에 대한 다양한 시각화

03 부도 발생을 이진분류 판별이 아닌 확률로 예측

PART 2

데이터탐색

1) 데이터
2) 데이터 탐색

*2194개 상장기업 7년치 연도별 기업데이터

- ☑ 기간 : 2012 - 2018
- ☑ 10545 rows * 160 columns
- ☑ 총 160개의 변수

데이터 탐색

1) 데이터 - 변수

Division		Num of Variables	Division		Num of Variables
Balance sheet	Total asset	1	Financial Ratio	Total assets ratio	14
	Current asset	6		total capital ratio	5
	Non-current assets	8		asset liability ratio	5
	Total liabilities	1		Cashflow ratio	4
	Current liabilities	9		Profit-related ratio	3
	non-current liabilities	9		Balance item	3
	capital assets	1		Sales profit item	12
	capital surplus	1		Return on Capital	7
	Comprehensive Income	1		Turnover Related Items	9
	Retained Earnings	1		Period related items	2
State-ment of comprehensive income	Sales	1		EPS	1
	Cost of Goods sold	1		BPS	1
	Gross profit or loss	1		SPS	1
	S&A expenses	8		CFPS	1
	Operating Income/Loss	1		EBITDAPS	1
	Financial income	4		PBR	1
	finance costs	3		PSR	1
	Other Income	1		Beta	1
	Other costs	1		WACC	1
	ILBIT	1		Total Cashflow	1
	Income Taxes	1		CAPEX	1
	Net Income or loss	1		EBITDA	1
	Other income	1	Year dummy variables	2012	1
	Total income	1		2013	1
State-ment of cash flows	Operating activities	6		2014	1
	Investing activities	1		2015	1
	Financing activities	1		2016	1
	Cash & cash equivalents	1		2017	1
	Cash at beginning	1		2018	1
	Cash at end	1	Other	Number of employees	1
Target	Default risk				1
Total	Total number of variables				160

1) 재무상태표항목 : 38개

2) 포괄 손익계산서 항목 : 26개

3) 현금흐름표 항목 : 11개

4) 재무비율지표 : 76개

5) 연도별 더미변수 : 7개

6) 직원 수 : 1개

* 연도별 외부사건에 대한 통제를 위해 더미 변수 추가

비재무데이터 대표 변수(3개)

	구분	개수	평균	표준편차	최솟값	최댓값
V1	Default probability(부도확률)	10545	0.34819	0.36766	0	1
V2	Market cap(시총)	10545	761	6550	3.3	329000
V31	The number of employees(직원수)	10545	876	3961	0	103011

재무상태표 대표 변수(7개)

	구분	개수	평균	표준편차	최솟값	최댓값
V3	current assets(유동자산)	10545	598	4070	0.9	175000
V4	Non-current assets(비유동자산)	10545	926	6240	0.1	166000
V5	Current liability(유동부채)	10545	458	2440	0.1	69100
V6	Non-current liability(비유동부채)	10545	348	2940	0	92300
V7	Capital stock(자본금)	10545	49.3	208	0	3660
V8	Capital surplus(자본잉여금)	10545	117	425	-574	7060
V10	Retained earnings(이익잉여금)	10545	530	5450	-3590	243000

손익계산서 대표 변수(14개)

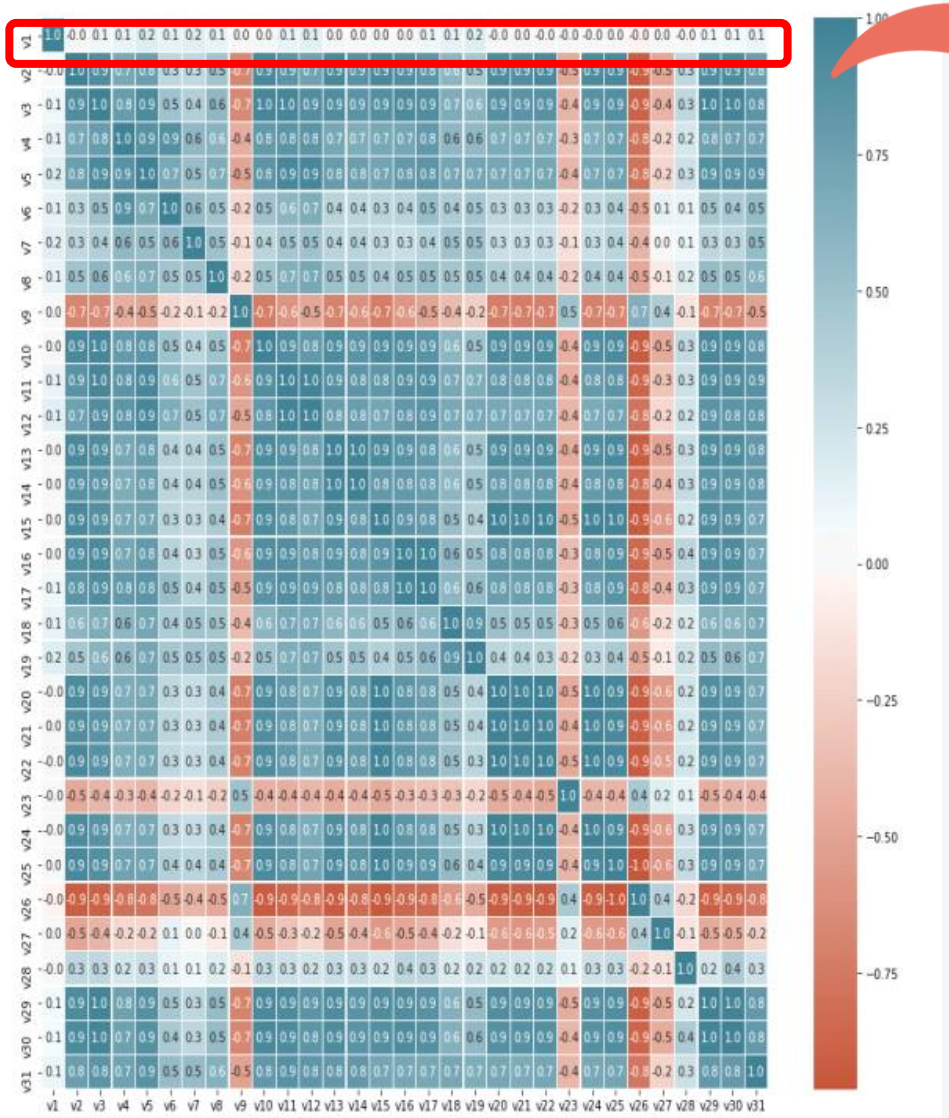
	구분	개수	평균	표준편차	최솟값	최댓값
V9	Accumulated Other Comprehensive income (기타포괄손익누계)	10,545	3.5	165	-7990	1940
V11	Sales(매출액)	10,545	1280	7720	0	244000
V12	Cost of sales(매출원가)	10,545	1000	5610	0	138000
V13	Gross profit(매출총이익)	10,545	286	2620	-1580	111000
V14	Selling, general and administrative expense(판관비)	10,545	201	1650	-273	56600
V15	Operating income(영업이익)	10,545	84.9	1090	-3270	58900
V16	Financial income(금융수익)	10,545	25.6	279	0	11400
V17	Financial cost(금융비용)	10,545	34.5	289	0	10700
V18	Other income(기타수익)	10,545	22.4	137	-0.6	4220
V19	Other expense(기타비용)	10,545	28.1	157	0	4420
V20	Continuing income and loss before income taxes(계속영업 세전이익)	10,545	78.5	1160	-4060	61200
V21	Tax expense(세금비용)	10,545	21.5	290	-985	16800
V22	Net income(당기 순이익)	10,545	59	878	-3080	44300
V23	Accumulated other comprehensive income (기타포괄손익누계)	10,545	-3.2	82.7	-5500	1990
V24	Total comprehensive income(총포괄손이익)	10,545	55.8	842	-3400	4330

현금흐름표 대표 변수(6개)

	구분	개수	평균	표준편차	최솟값	최댓값
V25	Cash flow from operating activities (영업활동 현금흐름)	10,545	115	1400	-3460	67000
V26	Cash flow from investing activities (투자활동 현금흐름)	10,545	-109	1170	-52200	4330
V27	Cash flow from Financing activities (재무활동 현금흐름)	10,545	0	356	-15100	8380
V28	Cash and cash equivalents (현금 및 현금성 자산)	10,545	5.4	168	-2510	9470
V29	Cash and cash equivalents at beginning of period (당기 초 현금 및 현금성 자산)	10,545	92.7	682	-5.4	32100
V30	Cash and cash equivalents at end of period (당기 말 현금 및 현금성 자산)	10,545	98.1	738	-5.5	32100

데이터 탐색

2) 데이터 탐색 -상관관계



〈Figure 6〉 Correlation analysis results for representative variables

상관관계	변수
0.1 (10개)	-유동자산 -비유동자산 -비유동부채 -자본잉여금 -매출액 -매출원가, -금융비용 -기타수익 -당기 초 현금 및 현금성 자산 -당기 말 현금 및 현금성 자산 -직원수
0.2 (3개)	유동부채, 자본금, 기타비용

부도확률이 재무지표가 아닌 주가 변동성과 같은 시장의 판단을 토대로 도출

PART 3

분석방법

1) 부도위험 기준 설정
2) 스태킹 앙상블 모델
3) 서브 예측 모델
4) 모델 비교
5) 하이퍼-파라미터

Merton 모형

1) **기하 브라운 운동** :회사 가치가 이 운동을 따른다고 가정

$$\frac{\mathrm{d}V}{V} = \mu\,\mathrm{d}t + \sigma_A\,\mathrm{d}W_t$$

V: 자산가치(주식가치 + 부채가치)
μ: 자산의 수익률
σ_A : 자산의 변동성
W_t : 표준 브라운 운동

2) **주식가치(KVM)** : **주식가치를 옵션가격결정모형으로 산정** 할 수 있다고 가정

: 기업의 부도확률을 부채변제 만기시점에서 **자산가치 < 부채가치**

$$E = VN(d_1) - e^{-rT}F(d_2)$$

$$d_1 = \frac{\ln(\frac{\mathrm{V}}{\mathrm{F}}) + (r + 0.5 \times \sigma_V^2)}{\sigma_v\sqrt{T}}$$

$$d_2 = d_1 - \sigma_v\sqrt{T}$$

F: 부채가치 = (유동부채 + 고정부채*0.5)
r: 무위험이자율 (국고채 3년 수익률)
σ_v : 자산가치 변동성
σ_E : 주식가치변동성

분석방법 1) 부도위험 기준 설정

3) 자산가치(v)와 자산가치의 변동성 (σ_v) 계산

: 시장 주식 가치(E)와 주식가치 변동성(σ_E)로부터 자산가치(V)와 자산가치변동성(σ_v)를 알아내는 것이 필요

$$\sigma_E = \frac{V}{E} N(d_1)\sigma_v$$

σ_E : 주식가치변동성

4) 부도거리

: 자산의 확률분포를 알 수 없기 때문에 부도거리를 통해 부도확률을 측정

$$DD = \frac{V - DP}{V \times \sigma_v}$$

DD : 부도거리
V :자산가치
DP : 부도점
σ_v : 자산가치변동성

5) KMV-머튼 모형의 내재부도확률

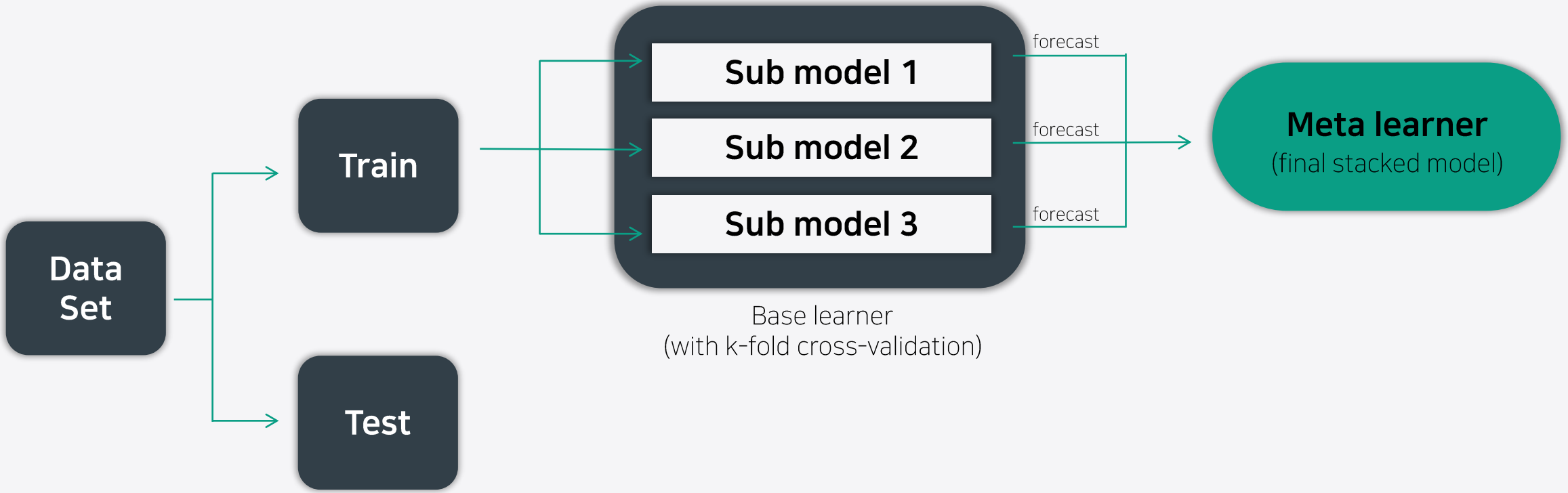
: 앞의 식들을 통해 자산가치(V)와 자산가치 변동성(σ_v) 을 추정
→ KMV-Merton 모형의 내재부도확률(π)

$$\pi = \mathrm{N}\left(-\left(\frac{\ln\left(\frac{\mathrm{V}}{\mathrm{F}}\right)+\left(r+0.5\times\sigma_V^2\right)}{\sigma_v\sqrt{T}}\right)\right)$$

스태킹 앙상블 모델

Model Stacking

서로 다른 모델을 조합해서 최고의 성능을 내는 모델 생성



분석방법 3) 서브 예측 모델

랜덤 포레스트

랜덤 노드 최적화와 배깅을 결합 한 방법과 같은 분류회귀트리(CART)를 사용해 상관관계가 없는 트리들로 포레스트를 구성

다층 퍼셉트론

다층 퍼셉트론은 투입층(input layer), 다수의 은닉층 (hidden layer), 출력층(output layer)으로 구성되어 있는 신경망 모형.

합성곱 신경망

이미지의 일정부분만을 인식하는 값을 필터를 통해서 새로운 값을 만들어주고 이 값들을 지속적으로 연결

분석방법 4) 모델 비교

Model comparison

Division	Model 1	Model 2	Model 3	Model 4
	Random Forest	MLP	CNN	Stacking Ensemble
Main logic	Decision tree	Neural network	Neural network	Combining predictions Using MLP
Input data for training	Training set	Training set	Training set	Base learner(RF, MLP,CNN) prediction results using segmented training set
Input data for testing	Test set	Test set	Test set	Base learner prediction results in test set
Training data dimension	(7382, 160)	(7382, 160)	(7382, 160)	(7382,4)
Test data dimension	(3163, 160)	(3163, 160)	(3163, 160)	(3163, 4)

· **Model 1 ~Model 3(Base learner) :** 스태킹 앙상블 모델의 예측력을 비교하기 위한 대조군

· **Model 4(Meta learner) :** 각 fold의 테스트에 집합에서 산출된 **Base learner들의 예측치를 토대로 훈련**

→ Base learner의 예측치를 결합시키는 알고리즘 MLP 사용

분석방법 5) 하이퍼-파라미터

Hyper-parameter setting results

Division	Model	Hyper Parameter	Value
Sub Model	Random forest	Number of trees	1000
		Loss function	MSE
		Random state	1
	MLP	Number of trees	100
		Number of hidden layers	2
		Epoch	300
		Loss function	MSE
		Optimizer	Adam
	CNN	Dimension	1
		Filter size	2
		Kerner size	2
		Pool size	2
		Epoch	300
		Loss function	MSE
		Optimizer	Adam
Stacking Ensemble Model	MLP	Number of nodes	20
		Number of hidden layers	2
		Epoch	300
		Loss function	MSE
		Optimizer	Adam

랜덤포레스트모델

- 생성할 트리 개수 1000개
- 오차함수 MSE사용

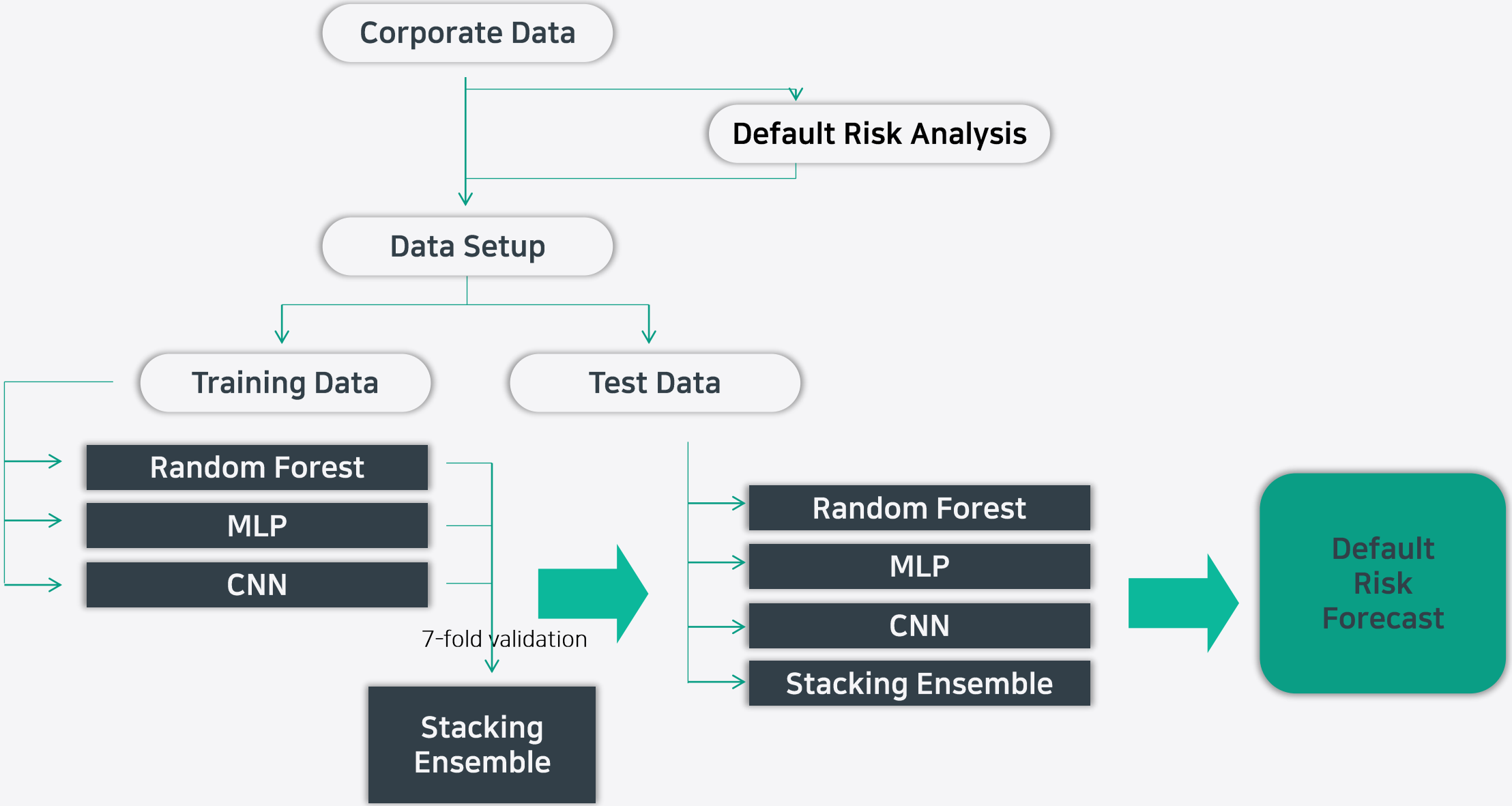
MLP모델

- 1개의 입력층과 2개의 은닉층
- 노드 수 100개
- 옵티마이져 Adam
- 손실함수 MSE
- 출력함수 시그모이드
- epoch 300

CNN모델

- 1차원 CNN모델 사용
- 필터 사이즈 2*2
- Pool 사이즈 2 사용
- epoch 300

분석방법 최종 분석 모델 로직



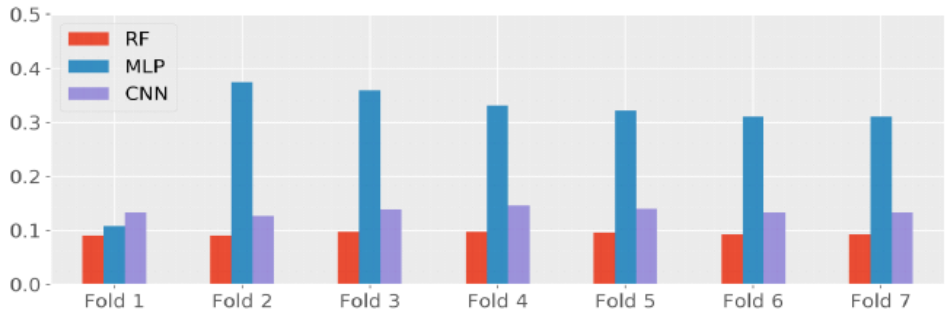
PART 4

실증결과 및 검증

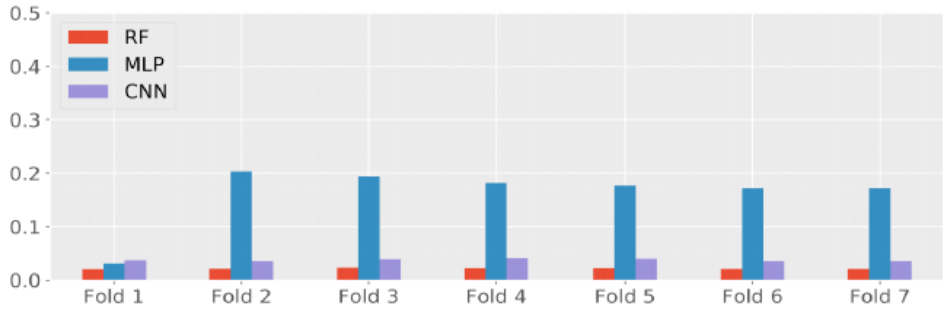
1) 서브모델 구현 결과
2) 부도위험 예측결과
3) 부도위험 예측결과 검정

실증 결과 및 검증

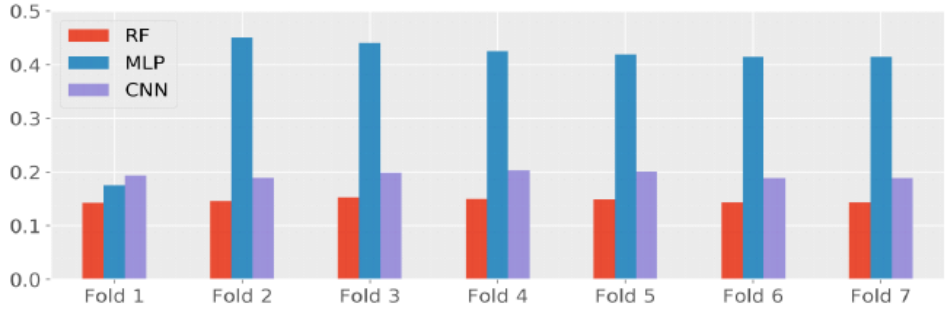
1) 서브모델 구현 결과



〈Figure 8〉 Prediction errors of meta-learner models (MAE)



〈Figure 9〉 Prediction errors of meta-learner models (MSE)



〈Figure 10〉 Prediction errors of meta-learner models (RMSE)

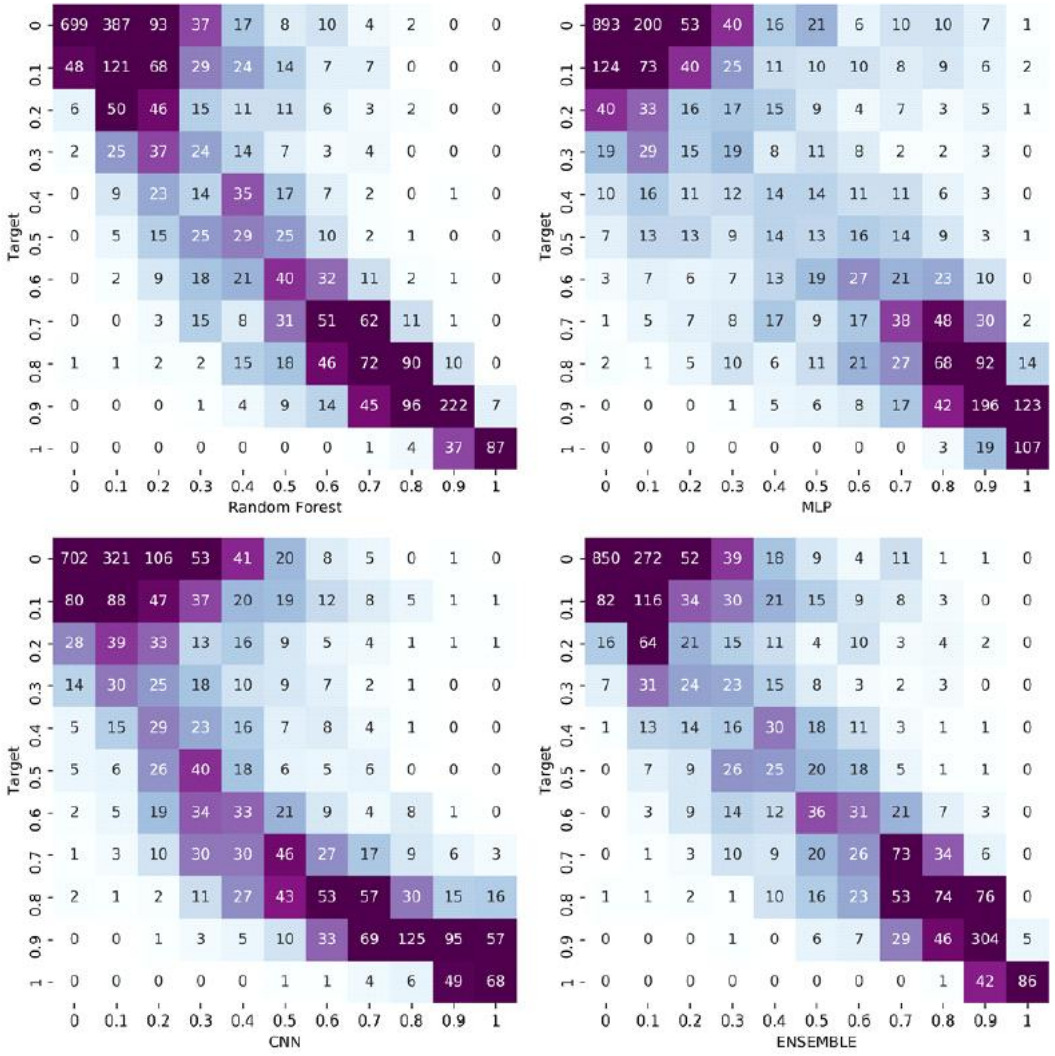
정확도

(막대 그래프 짧을수록 정확한 예측)

RF > CNN > MLP

실증 결과 및 검증

2) 부도위험 예측결과



〈Figure 11〉 Prediction comparison (Horizontal axis: prediction value, Vertical axis: target value)

예측력

(우하향 대각선의 정중앙에 가까울수록 정확한 예측)

가로축: 모델이 산출한 예측값
세로축: 실제 타겟값

RF, SE > MLP, CNN

실증 결과 및 검증

2) 부도위험 예측결과

Results of prediction error

Division	Model 1	Model 2	Model 3	Model 4
	Random Forest	MLP	CNN	Stacking Ensemble
MAE	0.094751	0.111714	0.134892	0.09079
MSE	0.022263	0.036053	0.037492	0.022175
RMSE	0.149209	0.189877	0.193628	0.148912

스태킹 앙상블 예측력

- **MAE 기준 :** RF 1.044배 | MLP 1.203배 | CNN 1.486배
- **MSE 기준 :** RF 1.004배 | MLP 1.626배 | CNN 1.691배
- **RMSE 기준 :** RF 1.002배 | MLP 1.275배 | CNN 1.300배

오차 결과

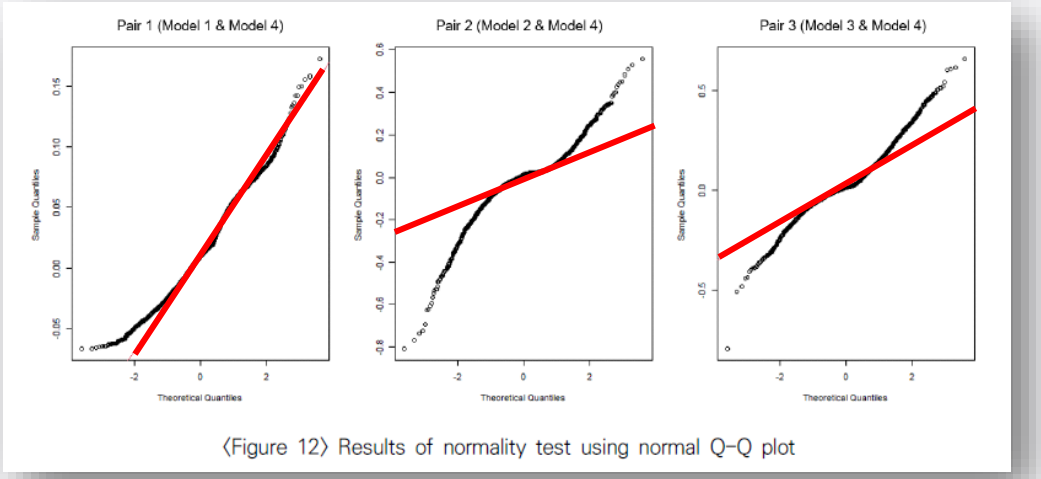
SE : 가장 낮은 오차

실증 결과 및 검증

3) 부도위험 예측결과 검정

Q-Q plot

모든 Pair가 붉은색 직선 바깥에 놓여져 있음



〈Figure 12〉 Results of normality test using normal Q-Q plot

샤피로-윌크 정규성 검정

모든 Pair가 정규성을 따르지 않음

Results of normality test (Shapiro-wilk normality test)

Division	Pair composition		Shapiro- wilk nomality test		
	Set 1	Set 2	W	p- value	Results
Pair 1	Stacking Ensemble	Random Forest	0.98303	< 2.2e-16	Non - nomal
Pair 2	Stacking Ensemble	MLP	0.89026	< 2.2e-16	Non - nomal
Pair 3	Stacking Ensemble	CNN	0.95428	< 2.2e-16	Non - nomal

윌콕슨 순위합 검정

정규성 검정 실패로 비모수적 방법인 윌콕슨 순위합 검정을 사용

Results of wilcoxon rank sum test with continuity correction

Division	Pair composition		Wilcoxon rank sum test		
	Set 1	Set 2	W	p- value	Results
Pair 1	Stacking Ensemble	Random Forest	4920409	0.2596	Non - difference
Pair 2	Stacking Ensemble	MLP	4687668	1.479e-05	difference
Pair 3	Stacking Ensemble	CNN	4753811	0.0006236	difference

- **Pair 1 :** SE & RF : 통계적으로 유의미한 차이 X
- **Pair 2 :** SE & MLP : 통계적으로 유의미한 차이
- **Pair 3 :** SE & CNN : 통계적으로 유의미한 차이

PART 5

기여효과 및 한계

1) 기여효과
2) 한계점
3) 정책적 제언

학문적 기여

01 메타 - 서브 구조 도입을 통한 과적합 문제 경감

02 예측을 위한 종속변수로 기존에 사용되는 변수(부도발생여부)대신
옵션가격결정모델에 기반한 **KMV-머튼모형의 부도거리를 계산한 부도확률** 을 사용

실무적 기여

01 주가 정보가 존재하지 않는 **비상장 기업에게도 적합한 부도위험** 을 평가할 수 있도록 함

02 전통적인 신용평가 모델을 예측모델에 반영할 수 있는 유연성을 제공

한계점

01 산출한 부도위험이 0~1 범위로 갖는 확률값으로 산출된다는 점

02 분석의 용이성을 위해 서브모델을 머신러닝 분야에서 대표적인 세 가지 모델로 한정한 점

한계점

01 feature selection 없이 160개의 변수를 모두 사용한 점

02 더미 변수의 개수와 범주의 개수와 동일하다는 점

03 비모수 검정을 사용하여 신뢰도가 낮다

04 외부데이터를 이용한 test가 없어 본 논문이 제시한 모델의 성능이 실제로 좋은 정확도를 보일지 불분명함

금융투자업규정

신용평가에 관한 과거의 통계자료 및 경험
미래의 시장환경 변화 등을 고려하여 평가방법의 적정성 검증 등을
포함한 평가방법을 마련할 것을 요구

스태킹 앙상블 모델

1. 부도확률에 대한 높은 예측력
2. 서브 모델을 금투업 규정에서 요구하는 적성성 기준에 부합되도록 세분화

머신러닝 기반 부도위험모델 도입에 제약요인으로
여겨졌던 문제들을 경감시켜 실무환경도입에 기여

Q & A