



나 일론마스크

IT기업 부실예측모형을 통한 증권 Portfolio 구성

홍영민 윤지혜 이정민 임상훈

CONTENTS



01 프로젝트 개요

02 데이터 확보

03 데이터 전처리

04 모델 및 성능비교

05 백테스팅 및 포트폴리오 구성

06 기대점 및 한계점

CHAPTER 01

프로젝트 개요

프로젝트 주제

“기업 부실 예측 모형을 통한 **투자 전략 개발**”

프로젝트

기업 부실 예측



기대 산출물

IT 기업 부실 예측 모형 개발

포트폴리오 종목 선정

투자전략 개발

팀원 및 역할 담당



홍영민
조장

- 프로젝트 계획 수립
- 모델링(RF, GAUSSIAN 外)



윤지혜
조원

- Logit 모델링 및 보조
- 보고자료 제작



이정민
조원

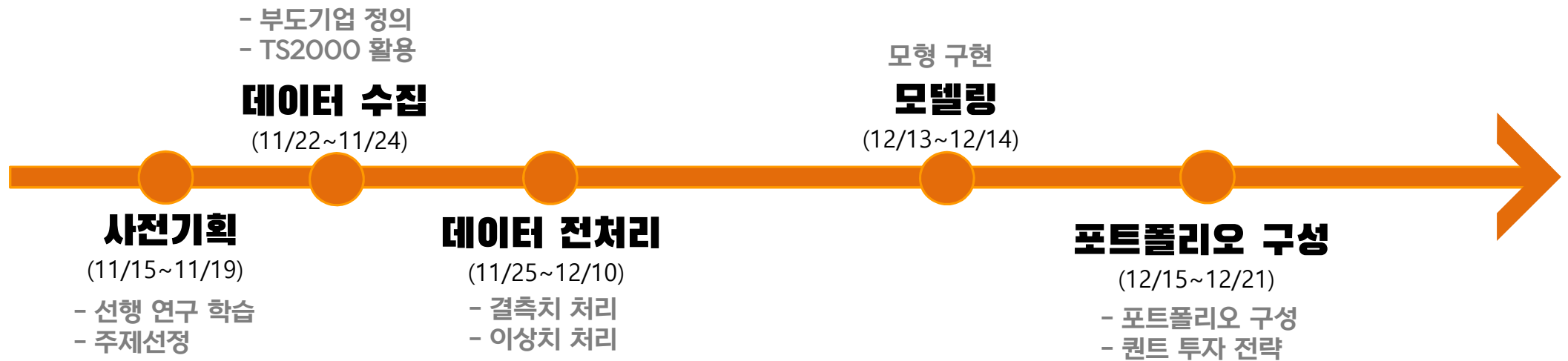
- SVM 모델링 및 보조
- 보고자료 제작



임상훈
조원

- ANN 모델링 및 보조
- 변수검정 (T-검정 등)

프로젝트 수행절차



사전조사

사전 논문 학습을 통해 변수, 모델을 선정하는데 있어 프로젝트 진행에 참고

Machine learning models and bankruptcy prediction(Flavio et al,2017)

- 변수 선정(성장성,수익성 범주 고려)
- 머신러닝 모형 참고

재무비율을 이용한 부도예측에 대한 연구(박종원,안성만)

- 산업별 분류 필요성
- 성능 평가 기준 참고

IT분야를 왜 선정하였는지?

변동성

E 이데일리

"FOMC 발표에 앞서 변동성 확대 조감 조경된 IT기업들이 주목" [이데일리 김겨레 기자] 한국투자증권은 8일 국내 정보기술(IT) ... 최근주가 조정으로 밸류에이션 부담이 완화된 기업에 집중할 시점"이라며 이같이...

2주 전

시사위크 2021.01.25.

"영원한 것은 없다"... '잘나가던' IT기업들이 몰락한 까닭

매순간 변화하는 소비자와 경제 트렌드를 읽지 못하고 과거의 영광에 취해 몰락해 버린 IT기업들의 사례는 '휴대폰 시장'에서 쉽게 찾아볼 수 있다. '노키아'와 '블랙...



FOMC 하루 앞 증시 변동성 ↑ ... "IT-성장주 조정 우려 과도"

증권가는 국내 증시도 이날 변동성 확대 국면에 진입할 것으로 전망했다. 14일(현지시간) 뉴욕증시에서... 그러면서 '국내 대형 성장 및 IT주들은 미국과 같은 쓸림...

[오늘의 투자전략] "FOMC 임박 국내증시 변동성 ↑, 중국 ... 이투데이 6일 전 [SEN투자전략]증시, 美 FOMC 경계감 속 변동성 확대..... 서울경제TV 6일 전

성장성

1	Apple	AAPL	\$2.784 T	\$169.75	-0.81%	
2	Microsoft	MSFT	\$2.401 T	\$319.91	-1.20%	
3	Alphabet (Google)	GOOG	\$1.884 T	\$2,648	-0.28%	
4	Amazon	AMZN	\$1.694 T	\$3,342	-1.73%	
5	Meta (Facebook)	FB	\$905.32 B	\$325.45	-2.50%	
6	Tesla	TSLA	\$903.77 B	\$899.94	-3.50%	
7	NVIDIA	NVDA	\$690.75 B	\$277.19	-0.29%	
8	Berkshire Hathaway	BRK.A	\$648.91 B	\$437,445	-1.32%	
9	UnitedHealth	UNH	\$455.60 B	\$483.73	-0.70%	
10	JPMorgan Chase	JPM	\$454.93 B	\$153.94	-1.80%	

CHAPTER 02

데이터 확보

데이터 선정

데이터베이스

한국 상장회사 협의회의 TS2000 활용

기간

2010.01.01 ~ 2020.12.31

대상

코스닥 / 코스피 상장기업



전체회사
640(개)

* IT 분야는 한국 소프트웨어 정책 연구소의 분류에 따름

IT분야 선정 기준

한국표준산업분류(KSIC)

- 제조업
 - 전자부품, 컴퓨터, 영상, 음향 및 통신장비 제조업 (IT 하드웨어 산업)
- 출판, 영상, 방송통신 및 정보서비스업
 - 출판업
 - 소프트웨어개발 및 공급업
 - 컴퓨터 프로그래밍, 시스템 통합 및 관리업
 - 컴퓨터프로그래밍, 시스템 통합 및 관리업
 - 정보서비스업
 - 자료처리,호스팅,포털 및 기타 인터넷 정보매개서비스업
 - 기타 정보 서비스업

부도정의기준

- ✓ 상장폐지 공시 사유 중 '재무적 위험성' 관련 사유가 있을 경우,
: '자본전액잠식', '감사의견부적정', '최종부도발생', '최근 5사업연도 연속 영업 손실 발생' 등
(43개 회사 해당)

*참고논문 : 빅데이터와 인공지능 기법을 이용한 기업 부도예측 연구 (최정원, 오세경, 장재원)

- ✓ 자본잠식상태가 3년간 지속될 경우, (46개 회사 해당)

*참고논문 : 기업도산예측을 위한 귀납적 학습지원 인공신경망 접근방법



총 60개의 회사를 **부도로 분류** (중복회사 존재/부도데이터:80)

변수 (Flavio et al, 2017 논문 변수 참고)

X1	$(\text{유동자산} - \text{유동부채}) / \text{총자산} = \text{운전자본비율}$
X2	$\text{이익잉여금} / \text{총자산} = \text{이익잉여금구성비율}$
X3	$\text{영업이익} / \text{총자산} = \text{총자산영업이익률}$
X4	$\text{시가총액} / \text{부채 총계} = \text{자본부채비율(레버리지비율)}$
X5	$\text{매출액} / \text{총자산} = \text{총자산회전율}$

OM	$\text{이자 및 세전 이익 (EBIT)} / \text{매출액} = \text{EBIT이익률}$
GA	$\text{당기말 총자산} - \text{전기말 총자산} / \text{전기말 총자산} = \text{총자산증가율}$
GS	$\text{t기 매출액} - \text{t-1기 매출액} / \text{t-1기 매출액} = \text{매출액증가율}$
GE	$\text{t기 종업원수} - \text{t-1기 종업원수} / \text{t-1기 종업원수} = \text{종업원수 증가율}$
CROE	$\text{t기 ROE} - \text{t-1기 ROE} = \text{ROE 변화량 (ROE [자기자본이익률] = 당기순이익/자본총액)}$
CPB	$\text{t기 PBR} - \text{t-1기 PBR} = \text{PBR 변화량 (PBR[주가순자산비율] = 주가 / 주당순자산)}$

유동성

안정성

수익성

안정성

활동성

수익성

성장성

CHAPTER 03

데이터 전처리

데이터 전처리 과정

결측치 처리

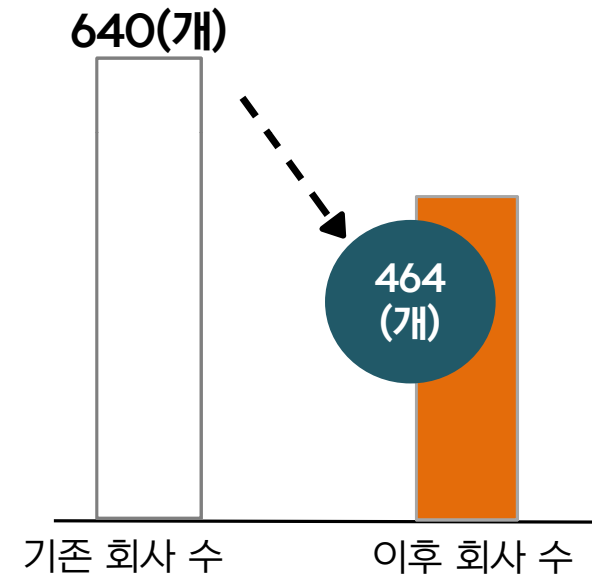
단순 탈락 또는 계산을 통해 결측값 처리

중복 데이터 제거

8개의 중복된 기업데이터 삭제

그 외

- 결산 월이 12월이 아닌 경우
- 1년 이하 기업일 경우



결측치 처리 방법

기준에 따른 단순 탈락 또는 계산을 통한 값 대체로 결측값 처리

name	0
code	0
year	0
casset	0
cdebt	0
asset	0
rearnings	0
debt	0
sales	0
employee	209
capital	0
equity	0
net income	0
stock_num	0
stock	250
roe	26
ebt	0
interest	0
facevalue	0
sector	0
dtype:	int64

단순 탈락으로 진행

상장 이전 데이터이므로 탈락 처리

Equity와 Net Income값으로 계산 처리

변수 별 박스플롯 극단치 확인

	x1	x2	x3	x4	x5	om	Ga	Gs	Ge	Croe	Cpb
총극단치수	21	146	242	253	87	359	267	205	27		
부도기업(개)	15	29	31	3	10	40	22	12	21	46	36
비율(%)	19%	36%	39%	4%	13%	50%	28%	15%	26%	58%	45%

부도기업 80개 중
극단치의 비율이 높음

➡ 부도기업의 극단치 비율이 높아 제거시 변별력 감소 우려하여 제거하지 않기로 결정

변수 유의성 검정(등분산성 검정)

```
x1 LeveneResult(statistic=59.548123398555894, pvalue=1.6211627209
x2 LeveneResult(statistic=175.17137718132412, pvalue=6.8937957399
x3 LeveneResult(statistic=317.48049297653256, pvalue=1.5634131093
x4 LeveneResult(statistic=6.114454911399377, pvalue=0.01346402946
x5 LeveneResult(statistic=9.160151522862428, pvalue=0.00249478412
om LeveneResult(statistic=255.27406938316417, pvalue=3.5510118813
ga LeveneResult(statistic=8.644754249328594, pvalue=0.0033055275
gs LeveneResult(statistic=4.644212059431207, pvalue=0.03123909216
ge LeveneResult(statistic=1.3299824412718, pvalue=0.2489017411268
```

x5, ga, gs, ge
등분산성 만족
(조건: p-value>0.01)

변수 유의성 검정(T-검정)

- T-검정의 조건 : 독립성, 정규성, 등분산성 만족

```
Variable x1 : The t-statistic and p-value assuming equal variances is 15.7
Variable x1 : The t-statistic and p-value not assuming equal variances is
Variable x2 : The t-statistic and p-value assuming equal variances is 21.6
Variable x2 : The t-statistic and p-value not assuming equal variances is
Variable x3 : The t-statistic and p-value assuming equal variances is 21.6
Variable x3 : The t-statistic and p-value not assuming equal variances is
Variable x4 : The t-statistic and p-value assuming equal variances is 3.02
Variable x4 : The t-statistic and p-value not assuming equal variances is
Variable x5 : The t-statistic and p-value assuming equal variances is -0.5
Variable x5 : The t-statistic and p-value not assuming equal variances is
Variable om : The t-statistic and p-value assuming equal variances is 18.5
Variable om : The t-statistic and p-value not assuming equal variances is
Variable ga : The t-statistic and p-value assuming equal variances is 6.19
Variable ga : The t-statistic and p-value not assuming equal variances is
Variable gs : The t-statistic and p-value assuming equal variances is 1.99
Variable gs : The t-statistic and p-value not assuming equal variances is
Variable ge : The t-statistic and p-value assuming equal variances is 2.55
Variable ge : The t-statistic and p-value not assuming equal variances is
```

등분산성 만족 여부에 따른
T값 도출



x5, croe, cpb 제외한 변수들의
유의성 확인 완료

최종변수

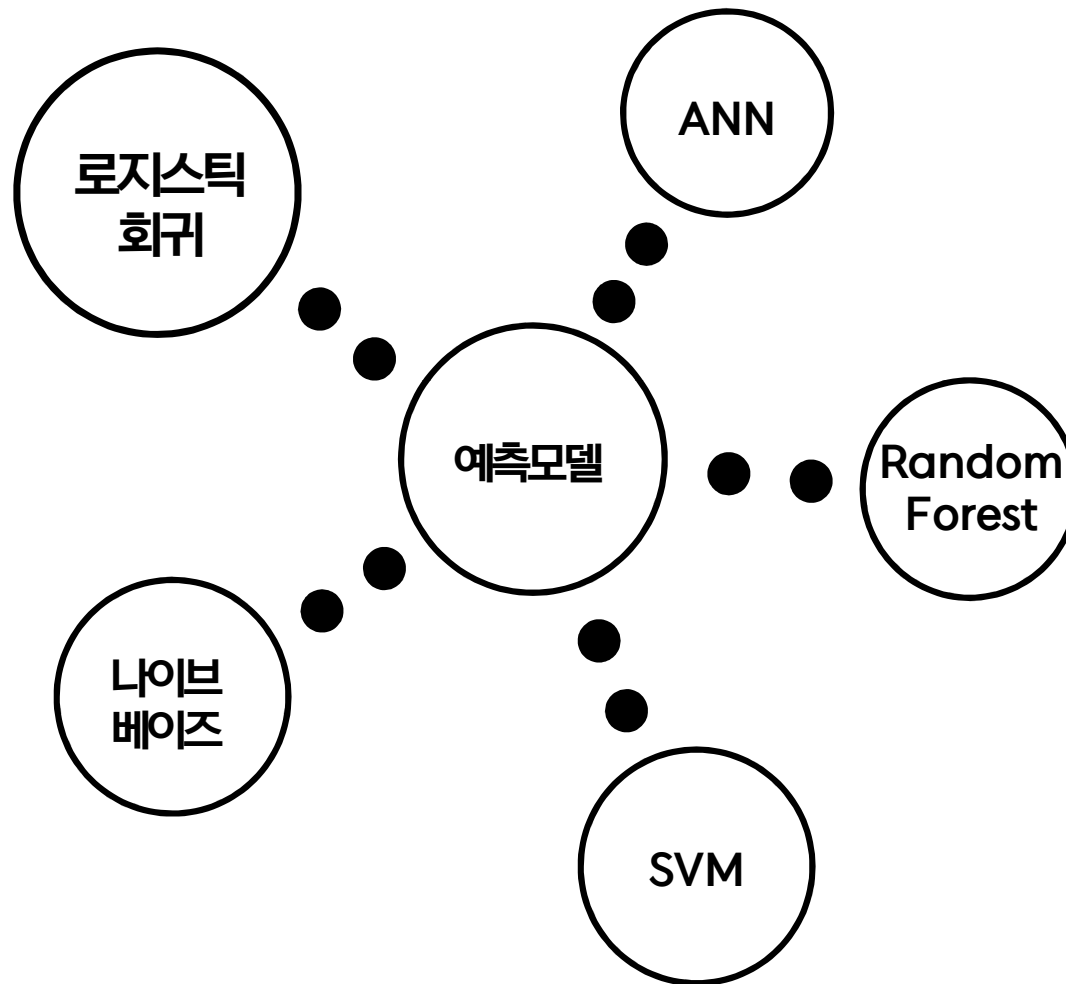
총 8개 변수 사용

X1	$(\text{유동자산} - \text{유동부채}) / \text{총자산} = \text{운전자본비율}$
X2	$\text{이익잉여금} / \text{총자산} = \text{이익잉여금구성비율}$
X3	$\text{영업이익} / \text{총자산} = \text{총자산영업이익률}$
X4	$\text{시가총액} / \text{부채 총계} = \text{자본부채비율(레버리지비율)}$
OM	$\text{이자 및 세전 이익 (EBIT)} / \text{매출액} = \text{EBIT이익률}$
Ga	$\text{당기말 총자산} - \text{전기말 총자산} / \text{전기말 총자산} = \text{총자산증가율}$
Gs	$\text{t기 매출액} - \text{t-1기 매출액} / \text{t-1기 매출액} = \text{매출액증가율}$
Ge	$\text{t기 종업원수} - \text{t-1기 종업원수} / \text{t-1기 종업원수} = \text{종업원수 증가율}$

CHAPTER 04

모델 및 성능비교

모델 선정(종합)



모델성능평가(기준)

	accuracy	Recall (파산기업)	f1-score	ROC_AUC
SVM	0.98	0.00	0.00	0.5
Logistic	0.98	0.31	0.36	0.65
ANN	0.97	0.31	0.33	0.65
Gaussian	0.94	0.69	0.35	0.82
Random Forest	0.98	0.31	0.44	0.65

모델성능평가(오버샘플링 적용/SMOTE)

	accuracy	Recall (파산기업)	f1-score	ROC_AUC
SVM	0.83	1.0	0.20	0.91
Logistic	0.84	1.0	0.21	0.92
ANN	0.93	0.62	0.27	0.77
Gaussian	0.90	0.69	0.24	0.79
Random Forest	0.94	0.62	0.30	0.78

모델성능평가(언더샘플링 적용/완전무작위)

	accuracy	Recall (파산기업)	f1-score	ROC_AUC
SVM	0.83	0.92	0.19	0.87
Logistic	0.83	1.0	0.21	0.92
ANN	0.72	1.0	0.14	0.87
Gaussian	0.90	0.77	0.24	0.83
Random Forest	0.73	1.0	0.14	0.86

모델성능비교

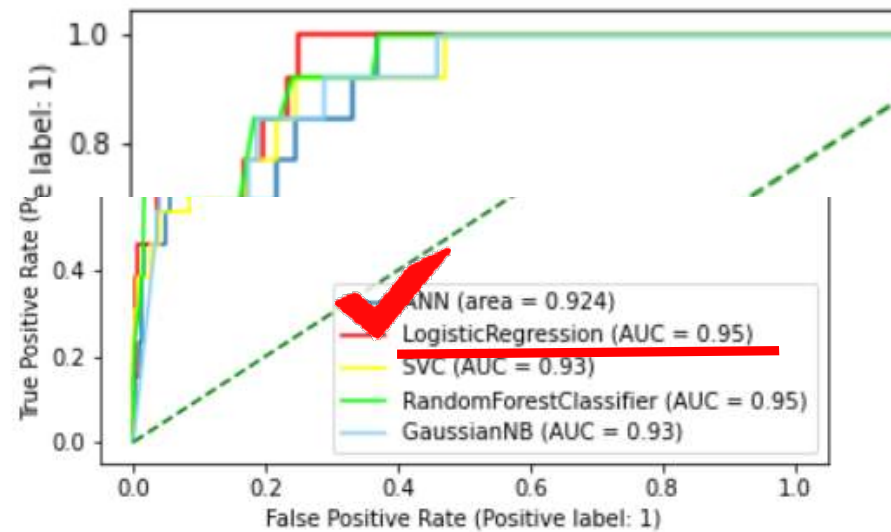
오버샘플링

	accuracy	Recall (파산기업)	f1-score	roc_auc
SVM	0.83	1.0	0.20	0.91
Logistic	0.84	1.0	0.21	0.92
ANN	0.93	0.62	0.27	0.77
Gaussian	0.90	0.69	0.24	0.79
RF	0.94	0.62	0.30	0.78

언더샘플링

	accuracy	Recall (파산기업)	f1-score	roc_auc
SVM	0.83	0.92	0.19	0.87
Logistic	0.83	1.0	0.21	0.92
ANN	0.72	1.0	0.14	0.87
Gaussian	0.90	0.77	0.24	0.83
RF	0.73	1.0	0.14	0.86

Roc_Auc Curve



- Logistic Regression 모델의 AUC 값이 가장 유의미한 결과를 나타냄

하이퍼 파라미터

ANN : Dense = 3, activation = relu, optimizer = adam, loss= binary crossentropy

SVM : var_smoothing = 1.0

Random Forest : bootstrapping / max_depth = 10 / max features = sqrt / min split
= 10 / min samples leaf = 2

SVM : c = 2,710 / gamma = 0.187


Logistic : c = 0.592 / solver = lbfgs

➡ 랜덤서치로 최상의 roc 값을 구하기 위해 피팅한 최적의 값

CHAPTER 05

백테스팅 및 포트폴리오 구성

포트폴리오 구성 방향성



투자 전략을 통한 그룹별
최적 포트폴리오 작성

Score 를 바탕으로
3개의 그룹화 및 백테스팅

기업 부도 예측
예측 확률 계산(Elon Score)

백테스팅 기준

기간

2010/01/01
~ 2020/12/31

투자 유니버스

- 기간 내 IT 산업군에 포함된 코스피 & 코스닥 상장 기업 대상
- 코스피 200 IT지수를 벤치마크로 비교

포트폴리오 구성 기준

산출된 부도 predicted probability를 기준으로, 3개의 등급으로 분류하여 접근

* 기준은 단발성 예시로, 추후 포트폴리오 방향에 따라 리밸런싱 가능

백테스팅 기준(1)

```
[ ] logistic_random.predict_proba(x_test)

array([[6.53339244e-01, 3.46660756e-01],
       [8.29853735e-01, 1.70146265e-01],
       [1.39975614e-04, 9.99860024e-01],
       ...,
       [7.45962933e-01, 2.54037067e-01],
       [8.44776773e-01, 1.55223227e-01]])
```

```
[ ] df_proba = pd.concat([header, proba], axis=1)
df_proba = df_proba.rename(columns={0:"safe",1:"risk"})
```

df_proba

	name	code	year	safe
0	(주)대유플러스	000300	2011	0.664591
1	(주)대유플러스	000300	2012	0.634537
2	(주)대유플러스	000300	2013	0.667358
3	(주)대유플러스	000300	2014	0.646174
4	(주)대유플러스	000300	2015	0.640270
...
2948	에스비아이핀테크솔루션즈(주)	950110	2017	0.698961
2949	에스비아이핀테크솔루션즈(주)	950110	2018	0.719348
2950	에스비아이핀테크솔루션즈(주)	950110	2019	0.685598

포트폴리오 구성 기준 :

Logistic regression에서 산출된 부도 predicted probability를 활용

해당 함수로, Safe, Risk의 확률을 계산

Safe - 값이 높을 수록 안전한 기업

Risk - 값이 높을 수록 부도확률이 높은 값

백테스팅 기준(2)

```
# 위험을 계산 완료 및 포트폴리오 기준 선정
def labeling(df,target):
    mean_rate = df.risk.mean()
    low_rate = df.risk.quantile(0.4)
    high_rate = df.risk.quantile(0.7)
    print(low_rate, high_rate,mean_rate)
    codes = df[(df.risk >= low_rate) & (df.risk <= high_rate) ].code.unique()
    another_codes = df[df.risk >= high_rate].code.unique()
    # 라벨링
```

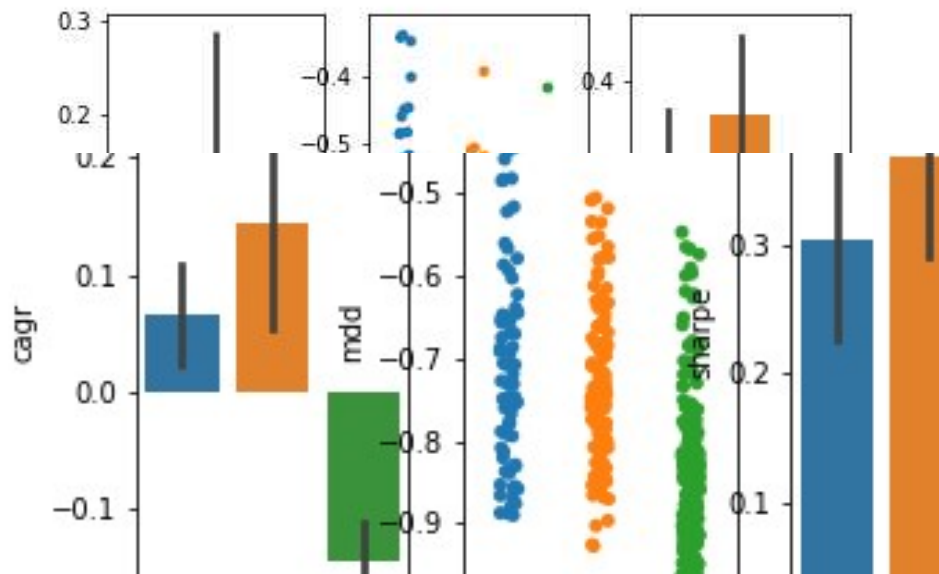
```
result2 = labeling(df_proba, result1)
result2
```

0.21283639148469366	0.3478259001011328	0.3088029		
	code	cagr	sharpe	mdd
0	000300	-0.089403	0.033072	-0.856828
1	000660	0.157213	0.531604	-0.578378
2	000990	0.206780	0.586872	-0.815447

세 그룹,
상위 33% (가장 위험) = portfolio [2]
중간 33% (보통) = portfolio [1]
하위 33% (안전) = portfolio [0]

=> 세 분류로 나누어서 접근

백테스팅 결과



라벨링

- [0] = 하위 33%
- [1] = 중간 33%
- [2] = 상위 33%

CAGR(Compound annual Growth Rate) :

연복리수익률

MDD(Maximum Draw Down) : 최대손실가능 수익률

Sharpe ratio : 해당펀드수익률-10년

미국국채)/(해당펀드수익률의 표준편차)

	CAGR(%)	MDD(%)	Sharpe
0그룹	6.57%	71%	0.31
1그룹	14.42%	73%	0.37
2그룹	-14.68%	88%	0.04
KOSPI200 IT	13.79%	37%	0.65

백테스팅 결과를 통한 투자 전략 수립

1. Buy & Hold 적용

여러 종목으로 구성된 포트폴리오/ETF 를
충분히 저평가되었을 때 매수 OR 주기적 매입
(적립식투자 방법)



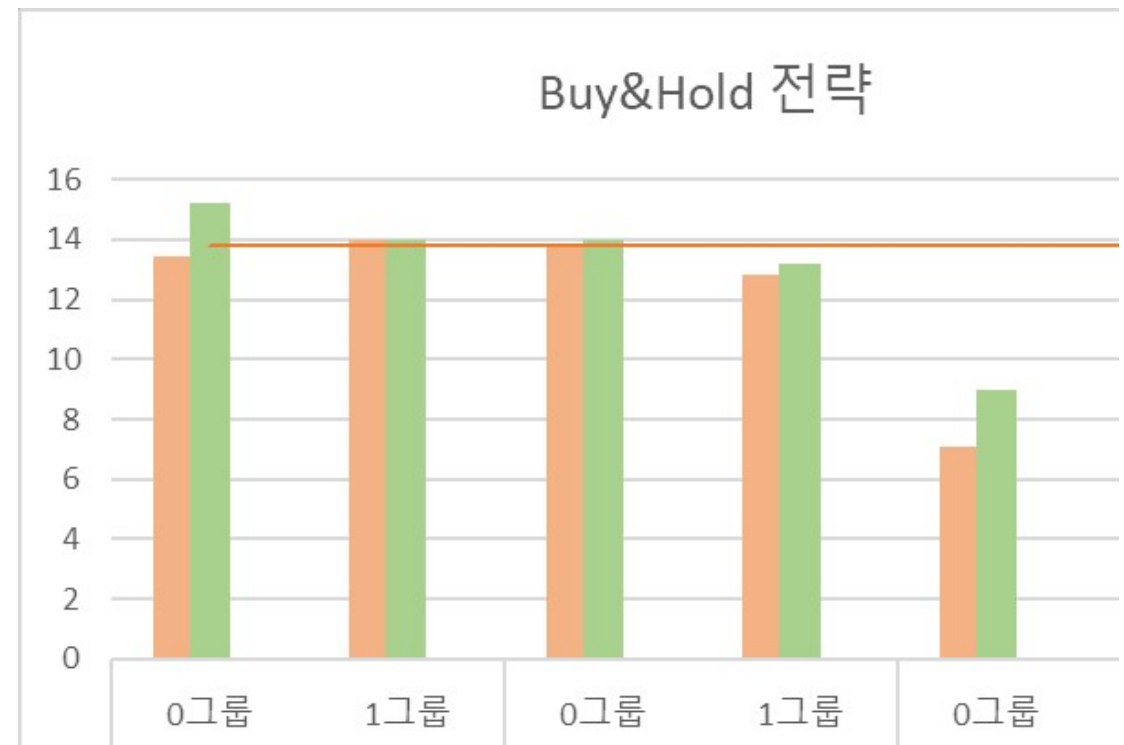
2. 이동 평균선 거래 적용(5일, 20일)

일정기간 동안의 주가를 산술 평균한 값인
주가이동평가를 차례로 연결해 만든 선



Buy & Hold

	그룹	CAGR	MDD	Sharpe
1년 단위 거래	0그룹	13.46	33.55	0.76
	1그룹	14.01	32.07	0.70
3년 단위 거래	0그룹	13.82	49.07	0.70
	1그룹	12.81	45.02	0.66
5년 단위 거래	0그룹	7.08	30.74	0.45
	1그룹	8.32	30.76	0.51



단순 이평선 거래

	그룹	CAGR	MDD	Sharpe
1년 단위 거래	0그룹	28.74	6.29	2.82
	1그룹	26.68	7.90	2.77
3년 단위 거래	0그룹	26.77	6.29	2.79
	1그룹	26.54	58.74	2.90



투자 결과

결과적으로, 0그룹과 1그룹 모두 5-20 이평선을 활용한 투자 전략이 가장 좋은 수익률과 sharpe값을 보여줌

BUT,

- 이 수치가 좀 신빙성 있는지에 대해 추후 보완 필요
- 0그룹과 1그룹간에 투자 결과 차이에 있어서 유의미한 차이가 보이지 않았다.

CHAPTER 06

기대효과 및 한계점

기대효과

- IT분야의 기업 부도 위험 예측 가능
- 실제 투자에 INDICATER 활용 가능
- 추후, IT 뿐만이 아닌 다른 분야로의 확장성 기대

한계점

- 자체적으로 새로운 변수 생성 X
- 특정 분야에 대한 접근 시도로 데이터 수 부족
- 수익률과 SHARPE 계산 값의 신빙성 부족
(계산 오류 가능성 높음)

Q & A
Thank you