
Trillion (1조)



TS2000 데이터를 이용한 기업 부도예측

정기호 이주희 신문혁 윤영주



CONTENTS

I. 프로젝트 개요

주제 설명 | 팀 구성 및 역할

II. 프로젝트 절차

부도 정의 | 모집단 선정 | 변수 설정
모델링 | 모델 평가 기준 | 순서도

III. 향후 계획

향후 계획

I. 프로젝트 개요

1. 주제 설명

“

”

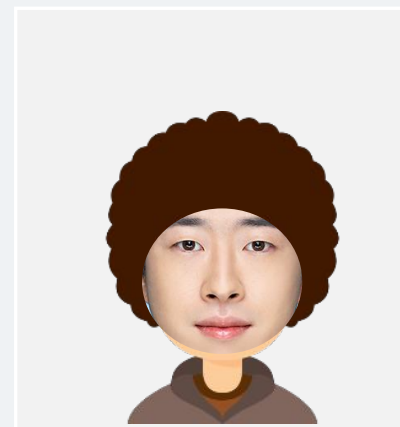
KOSPI, KOSDAQ 기업의 재무 데이터를 이용한 부도 예측

기간	2021/11/15~ 2021/12/23
내용	기업 재무제표 분석을 통한 부도 예측 with EDA, 전처리, 모델링, 시각화
목표	<ul style="list-style-type: none">→ TS2000 기업 데이터를 사용하여 부도기업에 대한 인사이트 도출→ 재무제표 구성 변수 이해 및 파생 변수 생성→ 정상, 부도 기업의 데이터 불균형에 대한 해결 연구→ 다양한 모델 적용과 하이퍼 파라미터를 통해 예측 성능 향상
활용	부도기업 예측 분류를 활용한 은행 여신 리스크 최소 방안 제시

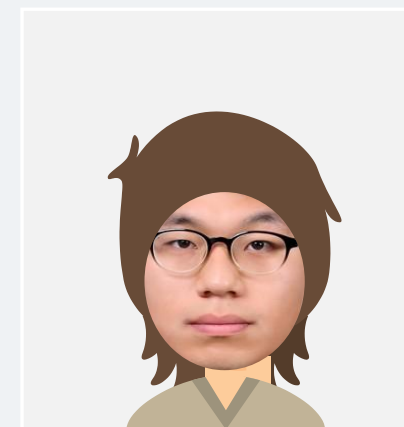
I. 프로젝트 개요

2. 팀 구성 및 역할

- TeamLeader
- 프로젝트 기획 및 총괄
- Modeling



정기호



이주희

- Pre-Processing
- 문헌 조사 및 자료 정리

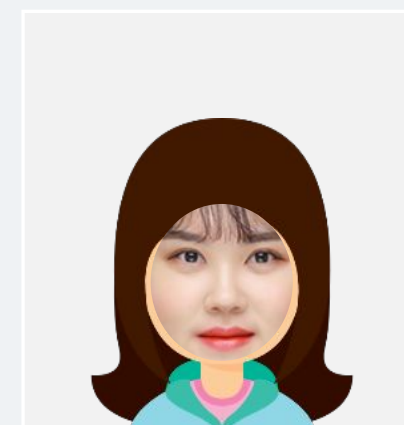
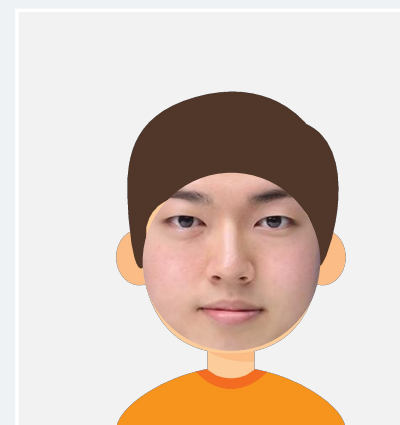


신문혁



윤영주

- Exploratory Data Analysis
- 시각화



- Pre-Processing
- Modeling

II. 프로젝트 절차

1. 부도 정의

부도 기준

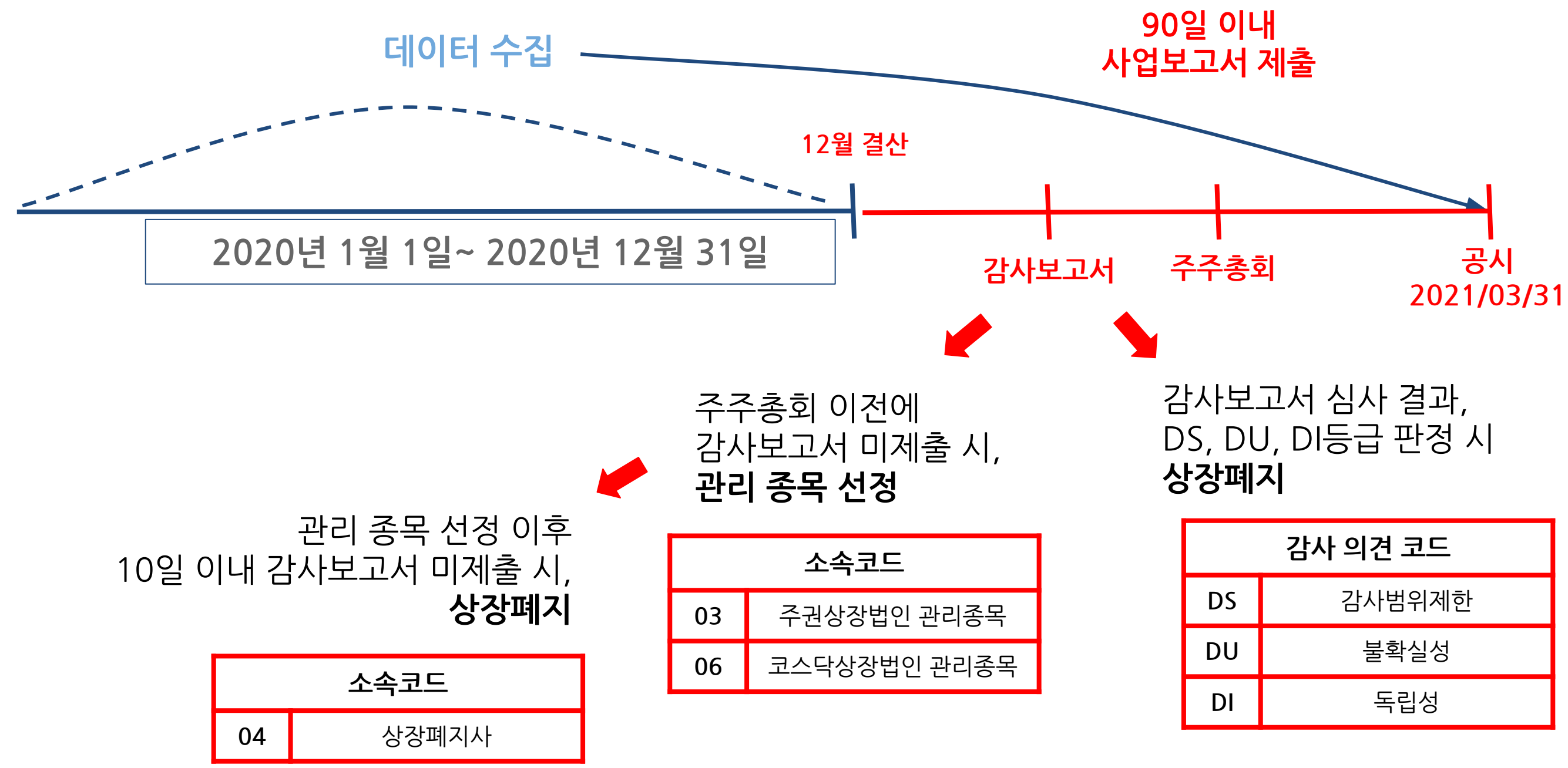
- 감사의견 코드가 ‘DS(감사의견제한)’, ‘DU(감사의견불확실) DI(독립성)’로 의견거절인 기업
- 소속코드가 04(상장폐지), 03(코스피관리종목), 06(코스닥관리종목)인 기업

감사의견 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

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

II. 프로젝트 절차

1. 부도 정의



II. 프로젝트 절차

1. 부도 정의

부도 기준

- 감사의견 코드가 ‘DS(감사의견제한)’, ‘DU(감사의견불확실) DI(독립성)’로 의견거절인 기업
- 소속코드가 04(상장폐지), 03(코스피관리종목), 06(코스닥관리종목)인 기업

	2001년	2002년	2003년	2004년	2005년
			↓ 부도	↓ 부도	↓ 부도
방법 1	정상	정상	부도	정상	부도
방법 2	정상	정상	부도	X	X
방법 3	정상	정상	부도	부도	부도

II. 프로젝트 절차

1. 부도 정의

TS2000 데이터 조건 식

The screenshot shows a window titled '조건실행' (Condition Execution) with two main panes. The left pane, '회사(3,007사)' (Company (3,007 companies)), lists various companies with their stock codes and names, such as [900010] (코스닥) 3노드디지털그룹 and [095570] (코스피) AJ네트웍스(주). The right pane, '항목(10개)' (Item (10 items)), lists financial and operational items like 산업코드 (Industry Code), 산업명 (Industry Name), 회계년월 (Fiscal Year/Month), and others. At the bottom, there are sections for '자료기간' (Data Period) set to 2000-2020 and '상장일 기준' (Listing Date Basis) set to '상장일 이전 자료' (Data before listing date). A 'Submit' button is at the bottom right.

- 회사 선택 조건:

2000년 01월 ~ 2020년 12월 기간 내의 KOSPI+KOSDAQ 상장기업

- 자료 선택 조건:

위 선정 기업의 10개의 재무 정보를 선정

산업코드, 산업명, 회계년월, 기수, 업종코드, 소속코드,
상장일, 상장폐지일, 감사의견코드, 감사의견)

II. 프로젝트 절차

1. 부도 정의

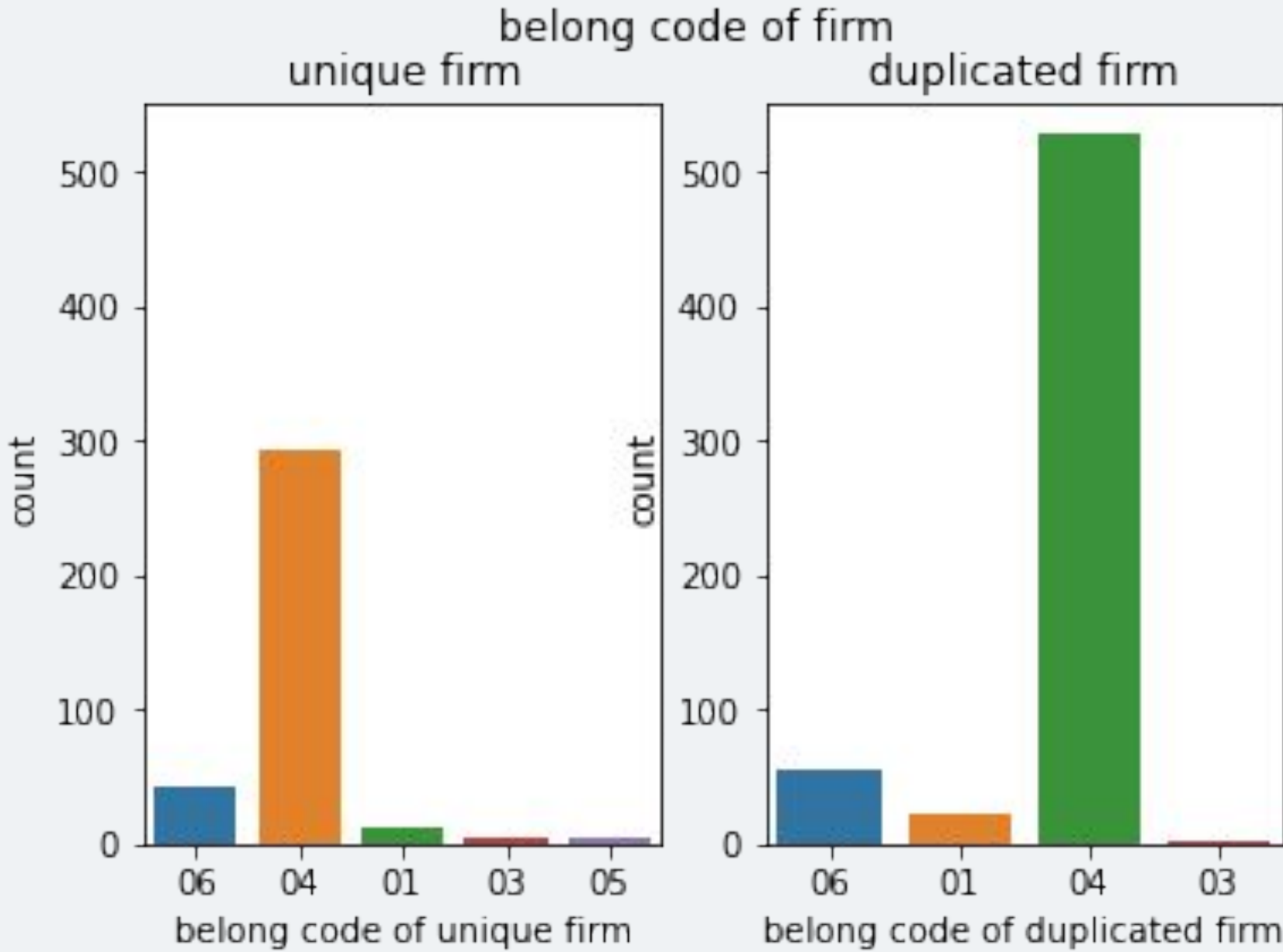
TS2000 데이터 조회 결과

	회사명	거래소코드	회계년도	산업코드	산업명	회계년월	기수	업종코드	소속코드	상장일	상장폐지일	감사의견코드	감사의견
0	(주)CMG제약	58820	2001/12	32102.0	의료용 물질 및 의약품 제조업	NaN	NaN	1.0	5.0	2001/08/31	NaN	UQ	NaN
1	(주)CMG제약	58820	2002/12	32102.0	의료용 물질 및 의약품 제조업	NaN	2	1.0	5.0	2001/08/31	NaN	UQ	NaN
2	(주)CMG제약	58820	2003/12	32102.0	의료용 물질 및 의약품 제조업	NaN	3	1.0	5.0	2001/08/31	NaN	UQ	NaN
3	(주)CMG제약	58820	2004/12	32102.0	의료용 물질 및 의약품 제조업	NaN	4	1.0	5.0	2001/08/31	NaN	UQ	NaN
4	(주)CMG제약	58820	2005/12	32102.0	의료용 물질 및 의약품 제조업	NaN	5	1.0	5.0	2001/08/31	NaN	UQ	NaN
...
135759	흥아해운(주)	3280	2015/12	85001.0	수상 운송업	2015/12	55.0	1.0	1.0	1976/06/29	NaN	UQ	NaN
135760	흥아해운(주)	3280	2020/12	85001.0	수상 운송업	2020/12	60.0	1.0	1.0	1976/06/29	NaN	UQ	NaN
135761	흥아해운(주)	3280	2015/12	85001.0	수상 운송업	2015/12	55.0	1.0	1.0	1976/06/29	NaN	UQ	NaN
135762	흥아해운(주)	3280	2015/12	85001.0	수상 운송업	2015/12	55.0	1.0	1.0	1976/06/29	NaN	UQ	NaN
135763	흥아해운(주)	3280	2019/12	85001.0	수상 운송업	2019/12	59.0	1.0	1.0	1976/06/29	NaN	DU	NaN

II. 프로젝트 절차

1. 부도 정의

부도 기준: 소속코드



Unique firm	
소속코드	count
01	11
03	5
04	293
05	4
06	42

Duplicated firm	
소속코드	count
01	21
03	2
04	529
05	0
06	55

II. 프로젝트 절차

2. 모집단 선정

“ 재무 비율을 이용한 부도 예측에 대한 연구 ”
(박종원, 안성만)

대상 : 외부 감사 기업

수집 데이터 기간 : 재무제표 2001/01/01 ~ 2006/12/31
부도기업 2005/01/01 ~ 2007/12/31

산업군 : 제조업, 건설업, 서비스업, 도소매업

부도 정의 : 은행 당좌거래 정지



“ TS2000 데이터를 이용한 기업 부도예측 ”
(Trillion)

대상 : KOSPI, KOSDAQ

수집 데이터 기간 : 1999/01/01 ~ 2020/12/31

산업군 : 제조업, 도소매업,

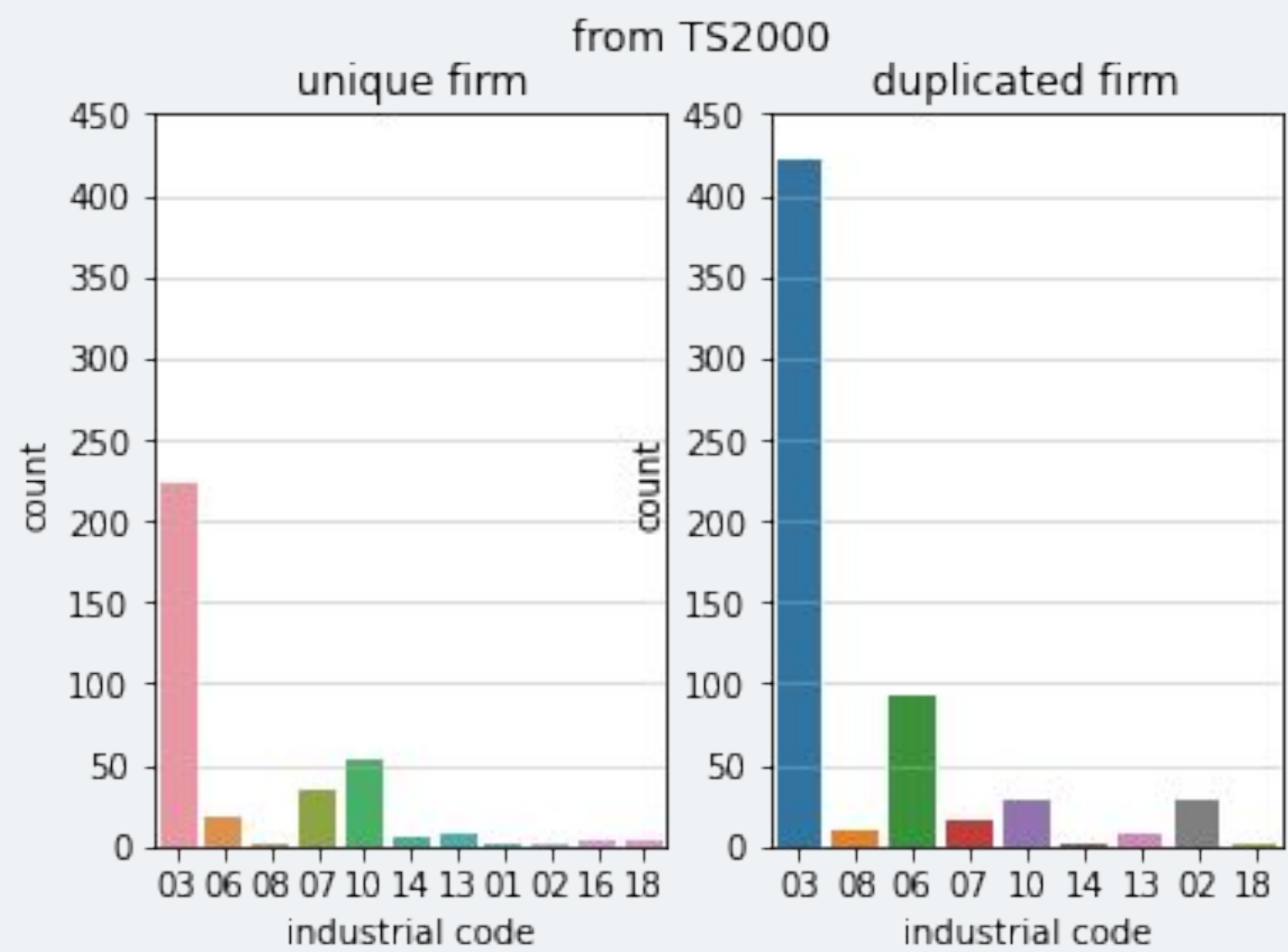
출판/영상/방송통신 및 정보서비스업

부도 정의 : 외부감사 의견 거절/ 상장폐지 및 관리종목 지정

II. 프로젝트 절차

2. 모집단 선정

산업 코드 선정



Unique firm	
소속코드	count
01	1
02	1
03	214
06	18
07	34
08	2
10	52
13	7
14	6
16	4
18	3

Duplicated firm	
소속코드	count
01	0
02	28
03	409
06	92
07	17
08	10
10	29
13	7
14	2
16	0
18	1

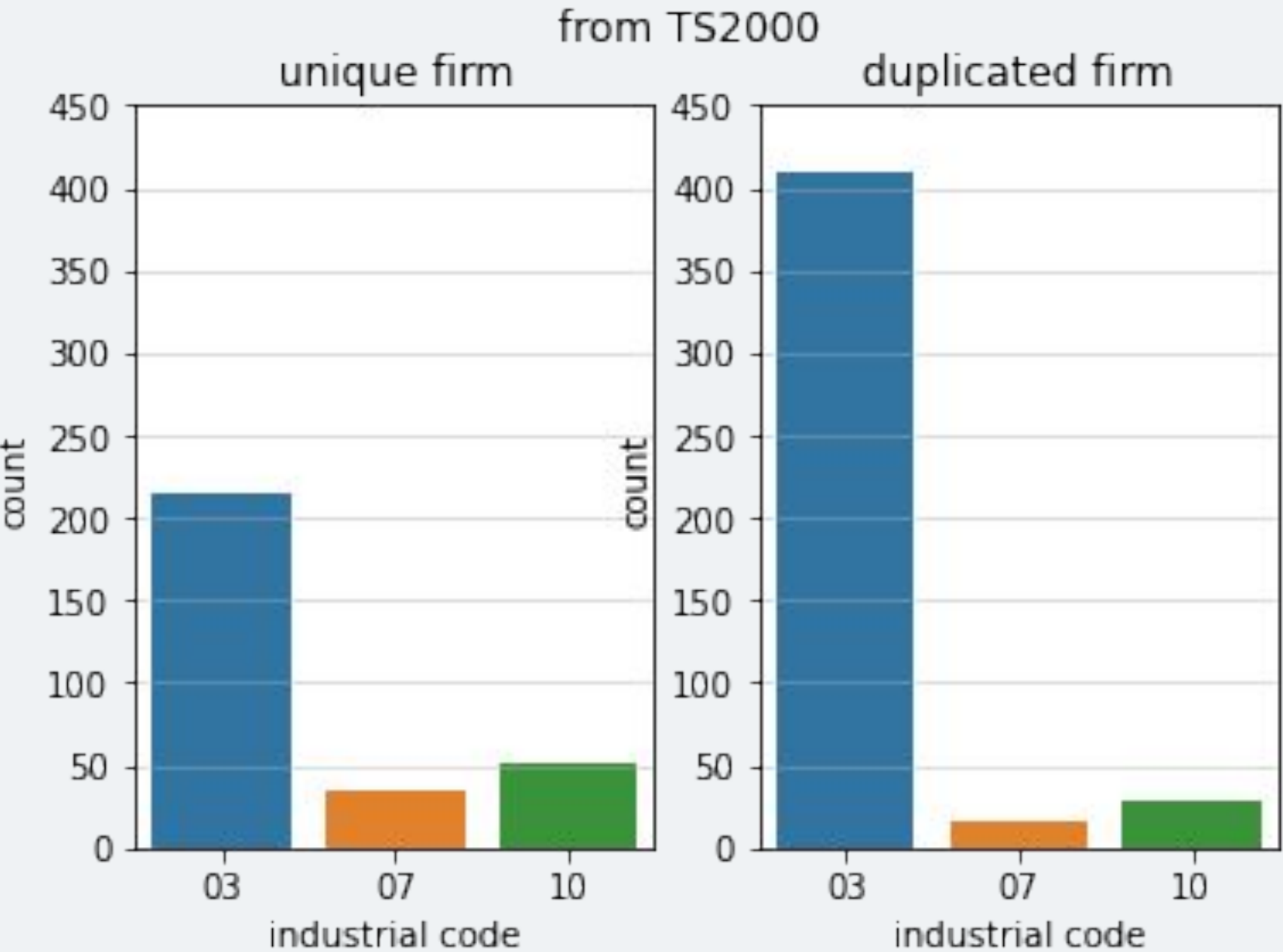
확인하면, 산업코드 03,07,10의 데이터가 상위 3개를 차지하고 있음을 확인할 수 있다.

06은 중복 기업에 꽤 많이 분포하고 있으며, 건설업이라는 산업의 특성상 (자기자본비율이 다름) model Generalize에도 좋지 않은 영향을 줄 것이라 생각되어 산업코드는 03, 07, 10을 선택한다.

II. 프로젝트 절차

2. 모집단 선정

산업 코드



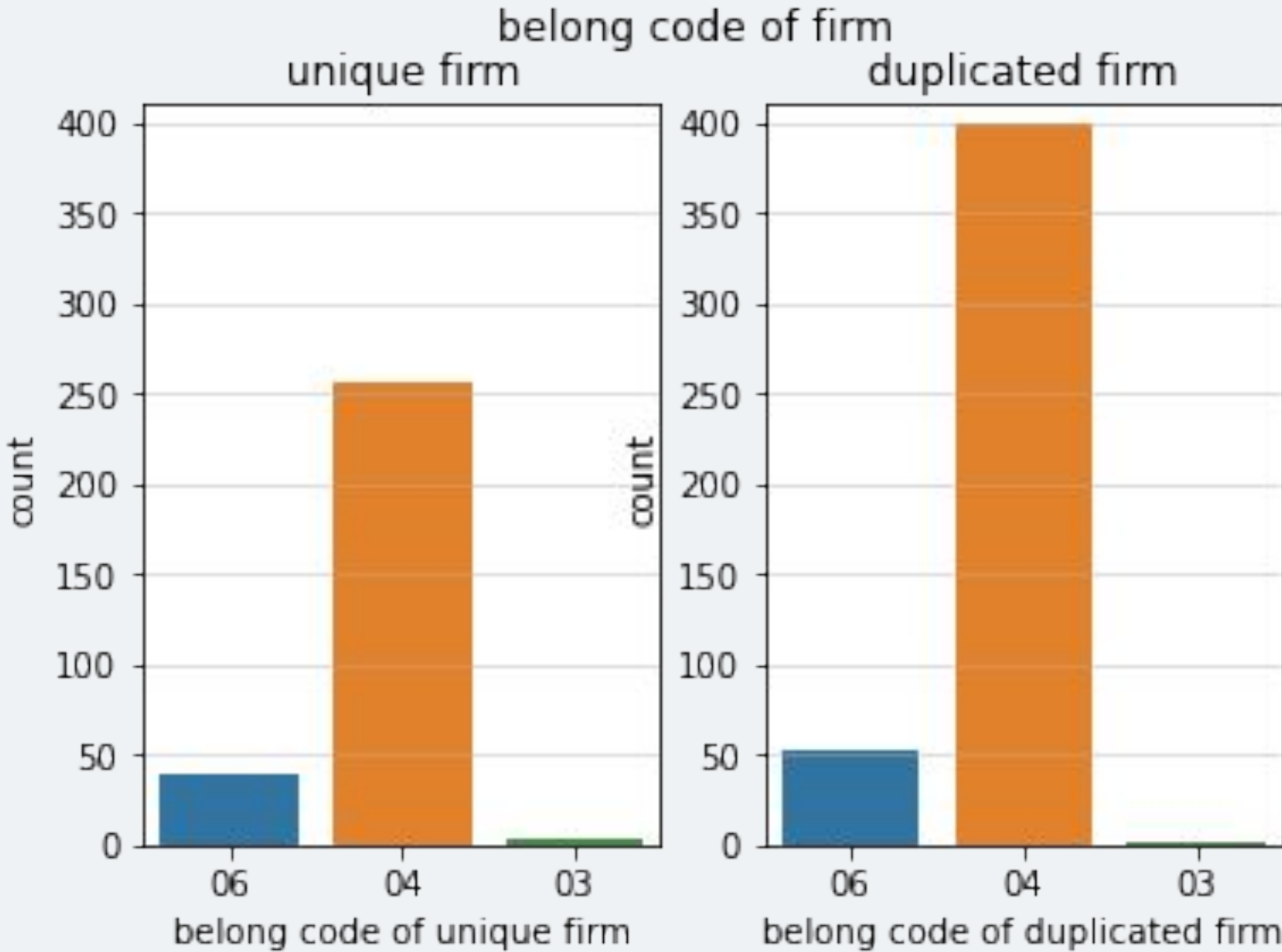
Unique firm	
산업코드	count
03	214
07	34
10	52

Duplicated firm	
산업코드	count
03	409
07	17
10	29

II. 프로젝트 절차

2. 모집단 선정

소속 코드



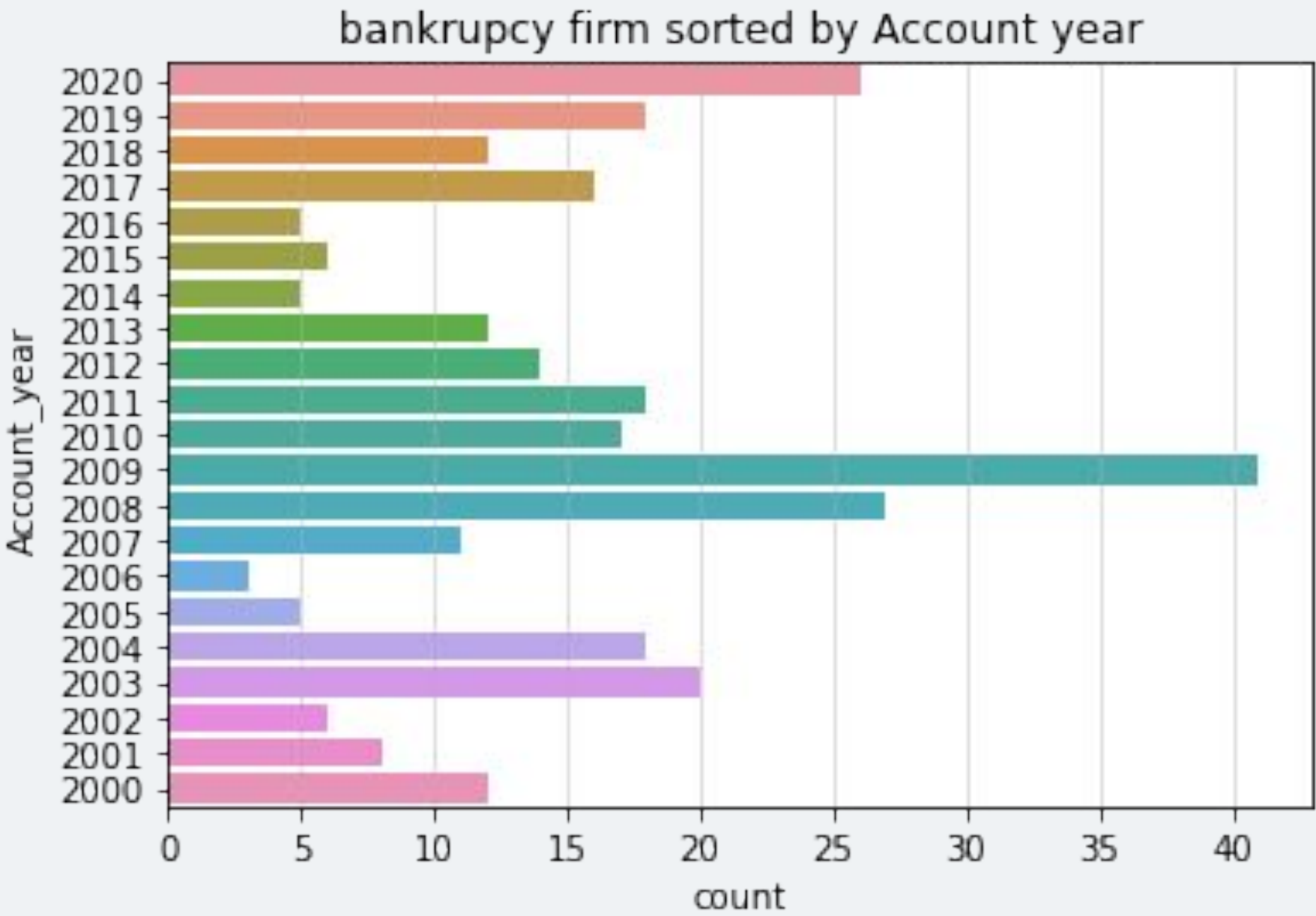
Unique firm	
소속코드	count
03	4
04	257
06	39

Duplicated firm	
소속코드	count
03	2
07	53
10	400

II. 프로젝트 절차

2. 모집단 선정

회계연도 별 부도기업 분포



	회계년도_new
2009	41
2008	27
2020	26
2003	20
2011	18
2004	18
2019	18
2010	17
2017	16
2012	14
2013	12
2018	12
2000	12
2007	11
2001	8
2015	6
2002	6
2014	5
2016	5
2005	5
2006	3

II. 프로젝트 절차

2. 모집단 선정

“ 재무 비율을 이용한 부도 예측에 대한 연구 ”
(박종원, 안성만)

대상 : 외부 감사 기업

수집 데이터 기간 : 재무제표 2001/01/01 ~ 2006/12/31
부도기업 2005/01/01 ~ 2007/12/31

산업군 : 제조업, 건설업, 서비스업, 도소매업

부도 정의 : 은행 당좌거래 정지



“ TS2000 데이터를 이용한 기업 부도예측 ”
(Trillion)

대상 : KOSPI, KOSDAQ

수집 데이터 기간 : 1999/01/01 ~ 2020/12/31

산업군 : 제조업, 도소매업,

출판/영상/방송통신 및 정보서비스업

부도 정의 : 외부감사 의견 거절/ 상장폐지 및 관리종목 지정

II. 프로젝트 절차

3. 변수 설정

Feature Selection

변수 유의성 검정 : 통계적 추정값의 신뢰도를 확인하기 위하여 통계적 이론에 근거하여 추론하는 통계검정

- ▶ T-검증

모집단의 분산과 표준편차를 알지 못할때, 모집단을 대표하는 표본으로부터 추정 통계치를 이용해 검정하는 방법

- ▶ 단변량 로짓

독립변수가 1개의 범주형 종속변수에 유의미한 영향을 미치는지 알아보기 위함.

변수 선택

- ▶ 단계적 추출법

여러 개의 후보 변수들이 있을 때, 다양한 변수 조합에 대해 통계 검정치를 비교하여 최적의 변수 조합을 찾는다

II. 프로젝트 절차

4. 모델링

비시계열 지도학습 알고리즘

- ▶ Logistic Regression

분석하고자 하는 대상이 두 집단 혹은 그 이상으로 나누어진 경우,
개별 관측치들이 어느 집단으로 분류될 수 있을지를 판단하는 분석 방법

- ▶ Random Forest

분류, 회귀 분석 등에 사용되는 앙상블 학습 방법의 일종으로,
훈련 과정에서 구성한 다수의 결정 트리로부터 부류(분류) 또는 평균 예측치(회귀 분석)를 도출하는 분석방법

시계열 지도학습 알고리즘

- ▶ RNN

시계열 데이터와 같이 시간의 흐름에 따라 변화하는 데이터를 학습하기 위한 인공신경망을 활용하는 분석방법

- ▶ LSTM

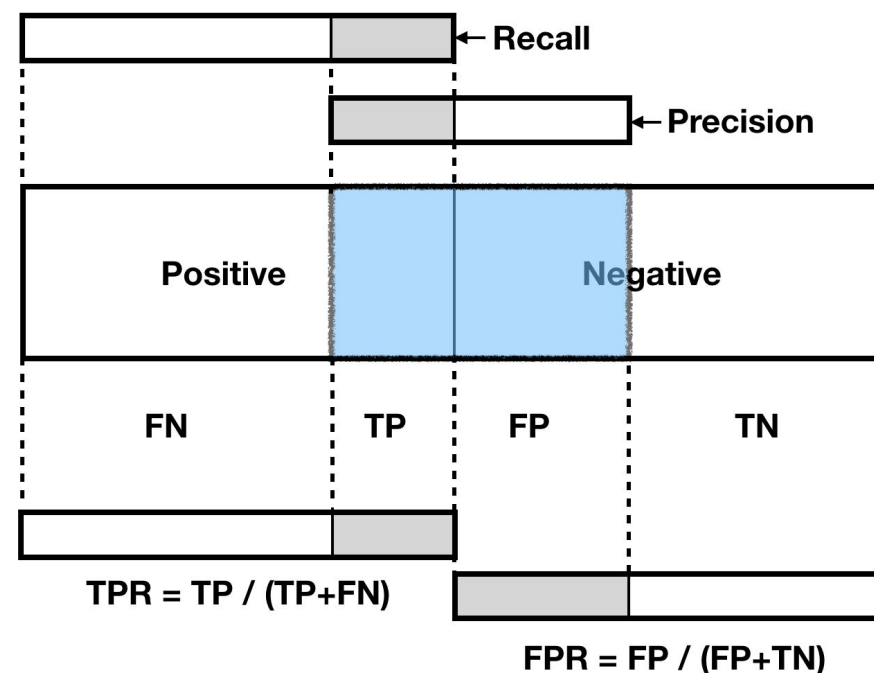
셀, 입력 게이트, 출력 게이트, 망각 게이트를 이용해 기존 순환 신경망(RNN)의 문제를
기울기 소멸 문제(vanishing Gradient Problem)를 방지하는 순환 신경망(RNN) 기법의 분석 방법

II. 프로젝트 절차

5. 모델 평가 기준

모델 평가기준

confusion matrix		Predicted	
		Yes	No
Actual	Yes	TP	FN
	No	FP	TN



▶ Precision(정확도) = $TP / (TP+FP)$

예측 True 중에 실제 True인 비율을 말함.
모형의 정확성을 나타냄.

Precision 값이 크다는 것은 실제 False를 True로 잘못 예측하는 오류가 낮다는 것.

▶ Recall(재현율) = $TP / (TP+ FN)$

실제 True 중에 True로 예측한 비율로 모형이 적중한 비율을 말함.
모형의 완전성을 나타내는 지표

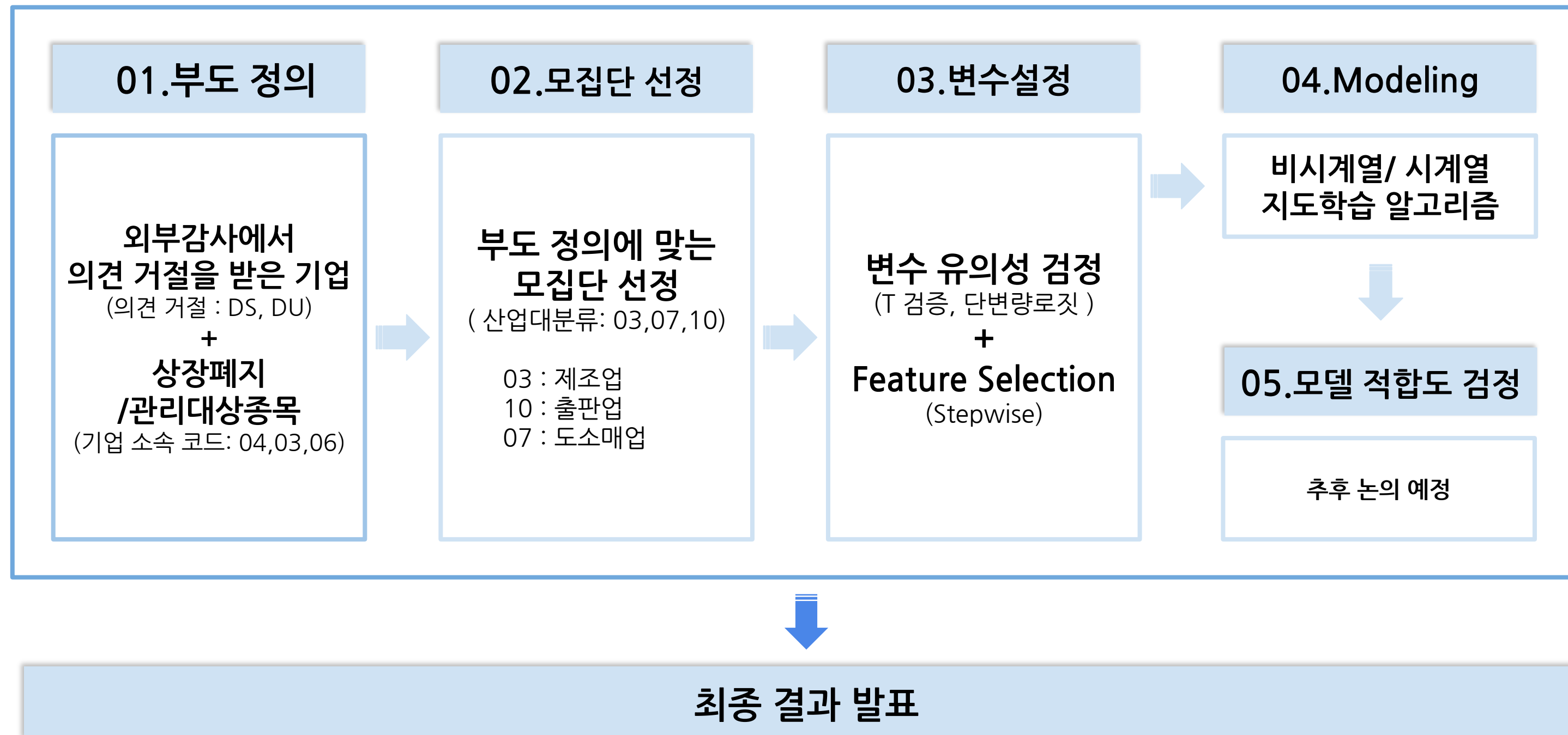
Recall 값이 크다는 것은 실제 True를 False로 잘못 예측하는 오류가 낮다는 것

▶ F1-score = $2 * (Precision * Recall) / (Precision + Recall)$

정확도와 재현율이 균등하게 반영될 수 있도록 정확도와 재현율의 조화평균을 계산한 값.
모형의 예측력을 종합적으로 평가하는 지표이다.

II. 프로젝트 절차

6. 순서도



III. 향후 계획

향후 계획

프로젝트 진행 일정					
1주	2주	3주	4주	5주	6주
선행 연구 논문 분석					
	데이터 수집				
	EDA				
	데이터 전처리				
		모델링			
			결과 분석		
				발표 준비	
					최종 정리

THANK YOU