뉴스기사 분석을 통한
기업 부도 예측

Team 빅데이터쉽G

Report Contents

뉴스 기사 분석을 통한
기업 부도 예측

01

contents 01
분석 기획

02

contents 02
데이터 준비

03

contents 03
데이터 분석

04

contents 04
결론 및 평가

Step 01

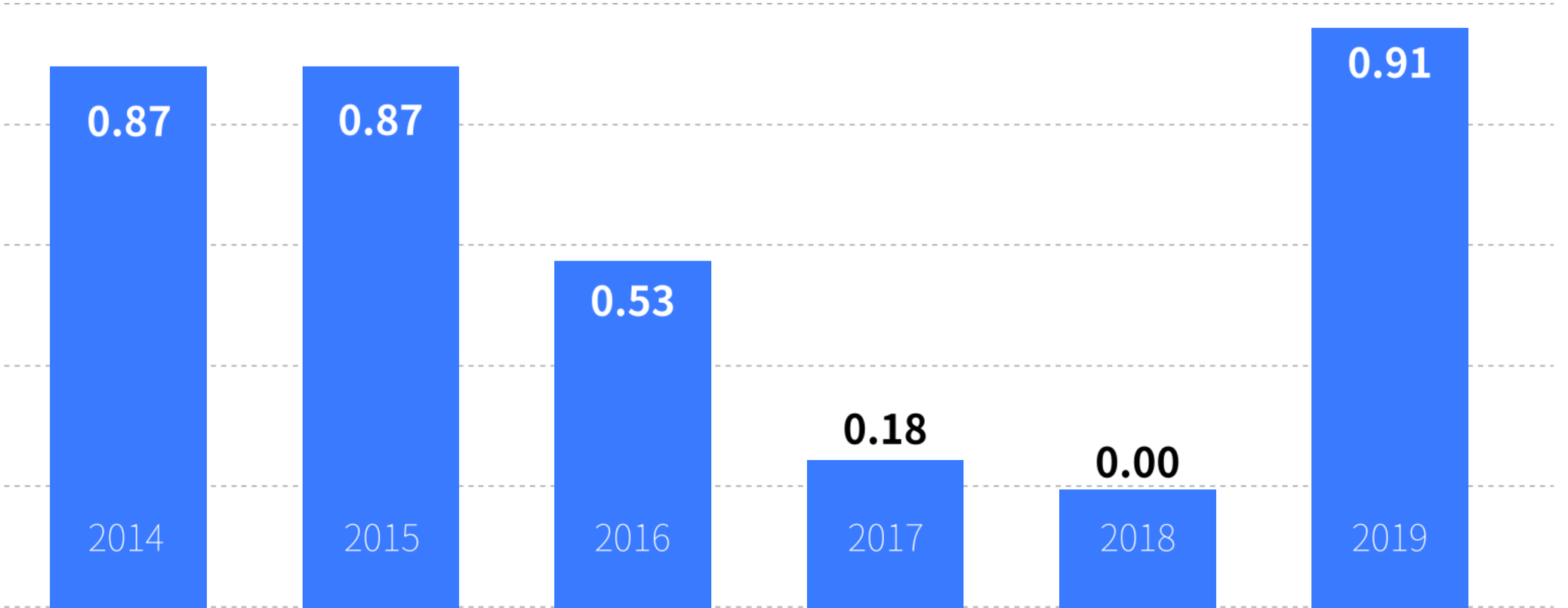
분석 기획 선정배경

| 2014년 이후 가장 높은 연간 부도율

지난해 국내 기업들 가운데 신용등급이 올라간 회사보다 내려간 회사가 더 많은 마이너스(-) 현상이 나타남

"코로나19 사태 등 최근 경제 상황에 비춰 볼 때 올해 신용등급 하락 및 부도율이 급속하게 심화될 것으로 예상"

금감원 관계자



Step 01

분석 기획 선정배경

| 코로나 19로 인해 큰 타격을 입은 국내외 경제

<딜링룸 백브리핑> "미국 기업 파산 속도, 2013년 이후 최고"

승인 2020.07.02 15:33 | 댓글 0

(서울=연합인포맥스) 국제경제부 = 미국 기업이 신종 코로나바이러스 감염증(코로나
빠른 속도로 파산 상태로 넘어가고 있다는 진단이 제기됐다.

코로나 강펀치에 '제조업 심장' 국가산단 녹다운…9월 '부도 쓰나미' 오나

5월 가동률 70.4%…IMF 직후보다 낮아
생산실적 19.8% 줄고 수출 30.8% 급락

김민혁 기자 | 2020-07-21 08:45:01 | 기업

파산 쓰나미 오나... 올해 상반기 법인 파산신청 522건 '역대 최대'

조선비즈 조은임 기자

입력 2020.07.21 17:04

올해 상반기 법원을 찾아 파산신청을 한 기업의 숫자가 사상 최대치를 기록했다. 신종 코로나 바이러스 감염증(코로나19)의 여파가 본격화되면서 기업들의 도산이 이어질 것이라는 우려가 현실화 되고 있다.

Step 01

분석 기획

문제인식

| 기업 부도 위험 예측의 필요성

기업의 부도 가능성 예측은 이해관계자들에게 예측 가능한 손실을 최소화할 수 있는 정보를 제공한다는 점에서 의의

경영자, 투자자 : 경영정책의 변화, 구조조정 등을 통해 손실 최소화
금융회사 : 신용평가 강화, 채권보전 조치 등을 통한 대손위험 최소화

| 기존 기업 부도 위험 예측의 불확실성

기업의 부도 위험은 재무정보로도 알 수 있지만 빠르게 파악하기 어렵다는 문제점
==> 텍스트 마이닝을 통한 빅데이터 모델을 구축

신용평가사가 예측하지 못한 기업의 부도를 AI가 예측에 성공한 사례

Step 01

분석 기획

주제선정

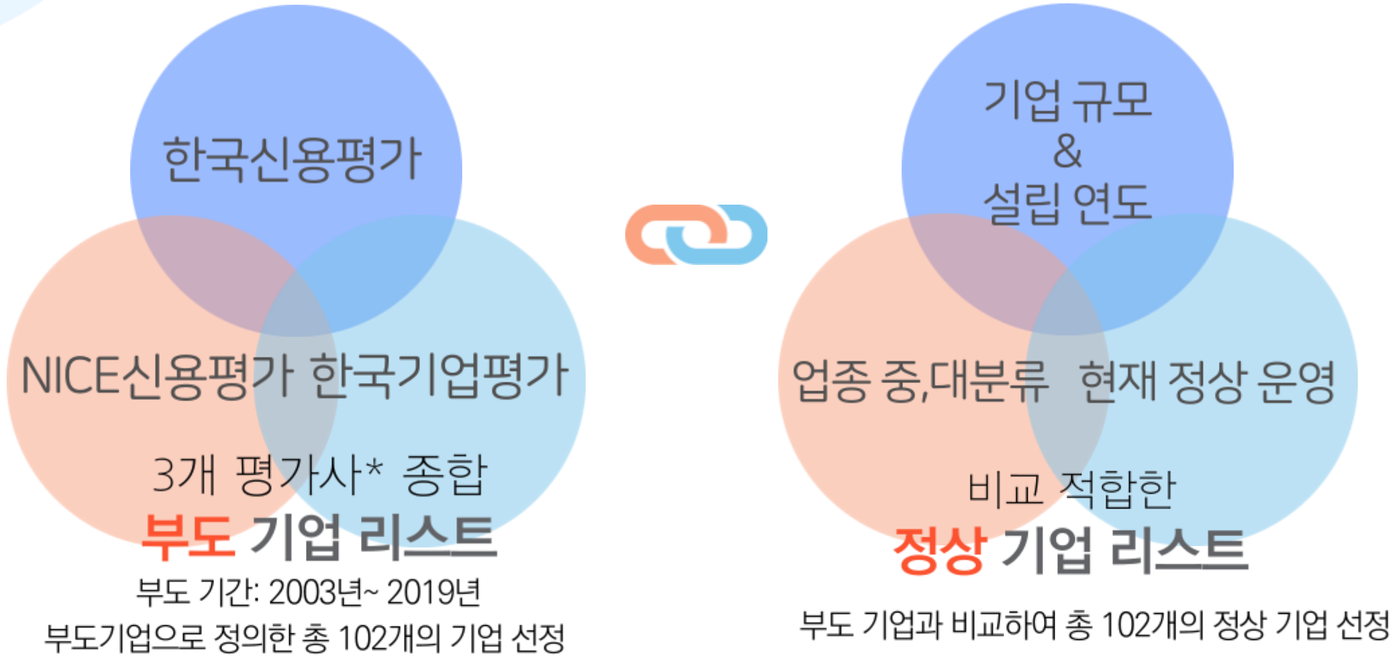
뉴스 기사 분석을 통한 기업 부도 예측

Step 02

데이터 분석 분석용 데이터 준비

부도 정의 및 기업 리스트 작성

유용한 결과를 얻기 위해 기업 부도(부실)의 명확한 정의 필요



"부도"는 원리금의 적기상환이 이루어지지 않거나
기업회생절차, 파산절차의 개시가 있는 경우를 포함
- 금융투자업규정 제8-19조의 9 제3항 2호-

부도는 등급의 정의 또는 평가방법론 등에 따라 다양하게 정의할 수 있으나, 당사는 '원리금의 상환 불능상태'를 부도로 정의하고 있으며 부도 시 'D등급'을 부여
- 한국신용평가사 -

원리금의 상환 불능상태에는 원리금의 적기상환이 이루어지지 않거나 기업회생절차·파산절차의 개시가 있는 경우 등 금융투자업규정 상의 부도

* 각 평가사에서 신용등급을 보유한 회사들만 측정을 하기 때문에 부도 리스트가 조금씩 다름.

Step 02

데이터 분석 분석용 데이터 준비

크롤링

웹상에 존재하는 정보들을 수집하는 작업
크롤링을 수행하는 프로그램: 크롤러

- 오픈 API 활용, 받은 데이터 중 필요한 데이터만 사용
- HTML 소스 가져와 원하는 정보 사용 - 정적 수집 방법
- 브라우저 조작으로 원하는 정보 사용- 동적 수집 방법

정적 수집
- 주소를 통한 접근으로 빠른 속도
- 수집 대상에 한계 존재

동적 수집
- 브라우저를 통한 연속적 접근으로 수집 대상에 한계 거의 없음
- 사용자와 상호작용하는 데이터 수집으로 매우 느린 속도

데이터 수집 진행
각 기업별로 수집기간 동안의 뉴스 수 다름 → 동적 크롤러(셀레니움) 활용이 유용
방대한 데이터 양과 데이터 수집에 오랜 시간이 걸리는 셀레니움의 단점 → 정적 크롤러 사용

Step 02

데이터 분석 분석용 데이터 준비

| 기업의 부실화는 단계별로 진행

손실 최소화하기 위해 조기 부실 징후 발견 및 대응 매우 중요

| 재무 정보 바탕의 부실 예측 모형

부실 1년 전에서 높은 예측력

예측 기간이 길수록 예측력 감소

예측의 정확성 → 부도 1년 전의 정보 사용

예측의 적시성 → 예측 기간 6개월 설정

기업이 부도된 시점에서 6개월 이전에 보도된
뉴스 기사의 제목, 본문 크롤링으로 수집
정상기업도 동일한 시점에 동일한 조건으로 진행

Step 02

데이터 분석

분석용 데이터 준비

1. 6개월로 수집 기간 설정

파이썬의 datetime 사용
부도일로부터 6개월 전 일자 : 변수 startday,
부도일: 변수 endday로 설정

+

2. 네이버 뉴스탭 접근

정적 크롤러(파이썬 request ,
BeautifulSoup 패키지) 활용,
기업명과 수집 기간을 쿼리로 사용

+

3. 정적 크롤러로 데이터 수집

동일한 html 구조로 구축된 사이트 확보
네이버 뉴스탭에서 검색되는 뉴스 중
'네이버 뉴스홈' 플랫폼 별도 제공 뉴스만 수집

> 뉴스 크롤링

기업당 최대 500개의 기사를 제한으로 수집

Step 03

데이터 분석

데이터 전처리

| 명사 추출

KoNLP(Korean Natural Language Processing)
: R 프로그램의 한글 자연어 분석 패키지

useNIADic : NIA 사전 활용, 약 100만개의 단어로 구성

→ 기사 제목과 내용에서 ' **명사** ' 추출

| 불용어 제거

특수문자•한자•영어 •숫자

→텍스트 분석에 적절 X, 공백으로 대체

전처리 전	전처리 후
'부활'한 쌍용차... 해고자들 복직요구	부활 쌍용차 해고 자 복직 요구
대한해운, 황당한 사기증자..."뒤통수 맞았다" 분통	대한 해운 황당 사기 증자 뒤통수 분통

Step 03

데이터 분석

텍스트 분석

01 동의어 처리

| 동일한 의미를 갖는 다른 표현들을 하나의 단어로 통일

동의어	처리대상
투자	주식투자, 설비투자, 기관투자, 투자유치, 시설투자, 집중투자, 민간투자, 투자비용, 창업투자, 신규투자, 투자펀드. 투자계획, 투자회수, 투자신탁, 투자확대
증권	유가증권, 증권정보, 증권거래소, 투자증권, 증권가
거래	매매거래, 신용거래, 주식거래
감사	감사보고서, 외부감사인, 감사인, 재감사, 회계감사, 외부감사, 재감사보고서, 국정감사
매각	지분매각, 자산매각, 매각대금, 매각설, 매각자
상장	상장사, 상장기업, 상장주식, 상장회사
회생	기업회생, 기사회생
성장	성장동력, 성장세, 급성장, 성장률, 고성장, 고속성장
부도	부도설, 부도위기, 부도처리, 부도금액

02 빈도•비중 분석

TFIDF 가중치 모델: 텍스트 마이닝을 위해 문서 내부에서 특정 단어의 중요도를 평가한 통계적 수치

$$TFIDF = TF \times \frac{1}{DF} = TF \times IDF$$

$TF = \frac{n}{N}$ n: 특정 단어 출현횟수, N: 문서 내 단어 개수

$IDF = \log\left(\frac{D}{d}\right)$ D: 문서 내 문장 개수, d: 특정단어를 포함하는 문장 개수

| ***TF***값: 지정한 문서 내에서 특정 단어의 빈도를 나타낸 값, 빈도가 높을수록 문서 내에서 중요도 상승

| ***DF***값: 특정한 단어가 일정한 범위의 문서들 간의 자주 사용되는 지수, 핵심 단어가 아닌 보편적 단어

| ***IDF***값: DF값의 역수

TF값 도출

DF값 도출

Step 03

데이터 분석

텍스트 분석

02 빈도•비중 분석

| 빈도: 뉴스기사 텍스트 내에서 해당 단어의 발생 빈도 수
| 비중: 전체 단어 발생 빈도 중 해당 단어의 비중

200개 ⇨ 46개

'공시', '감사', '상장폐지', '법정관리', '등급'
→ GAP > 0, 부도와 관련된 단어

'인수', '영업이익', '실적', '성장'
→ GAP < 0, 부도와 무관한 단어

'공시', '투자', '이사', '증권'
→ 부도기업과 정상기업의 빈도 • 비중 높음

46개 단어 빈도•비중 분석
정상기업과 비교

단어	부도기업		정상기업		GAP
	빈도	비중	빈도	비중	
공시	23321	1.222%	5402	0.645%	0.577%
투자	14377	0.753%	5139	0.614%	0.140%
이사	13318	0.698%	3952	0.472%	0.226%
증권	13018	0.682%	3870	0.462%	0.220%
감사	12294	0.644%	1541	0.184%	0.460%
상장폐지	8793	0.461%	1145	0.137%	0.324%
법정관리	8007	0.420%	357	0.043%	0.377%
인수	7656	0.401%	3632	0.434%	-0.033%
영업이익	7255	0.380%	4831	0.577%	-0.197%
등급	6268	0.328%	453	0.054%	0.274%
실적	3774	0.198%	2780	0.332%	-0.134%
성장	3447	0.181%	2540	0.303%	-0.123%
...	...	...	...	...	...
Total	1908368	100%	837407	100%	

Step 03

데이터 분석 텍스트 분석

03 연관성 분석

| 연관성 분석: 상품 혹은 서비스간의 관계를 살펴보고 이로부터 유용한 규칙 파악
구입항목의 집합에서 하나의 구입상품 또는 구입상품들 집합의 존재가 또 다른 구입상품의 존재를 암시하는 규칙을 발견하는 기법

| 연관규칙: 조건과 반응의 형태(if-then)로 구성
If A then B: 만일 A가 일어나면 B가 일어난다. *"아메리카노를 마시는 손님 중 17%가 스콘을 먹는다."*

| 연관성 측도

지지도

지지도 = $P(X \cap Y)$

생성된 연관규칙이 전체 항목에서 차지하는 비율
데이터베이스에 속한 전체 거래의 개수 중
그 연관규칙을 지지하는 거래의 개수 비율

신뢰도

신뢰도 = $P(X|Y) = P(X \cap Y)/P(X) =$ 지지도$/P(X)$

연관규칙의 강도
전제부 만족하는 거래가 결론부까지 만족하는 비율
X를 포함하는 거래 중 Y가 포함된 거래의 정도 의미

향상도

향상도 = $P(Y|X)/P(Y) = P(X \cap Y) /P(X)P(Y)$
= 신뢰도$/P(Y)$

두 거래품목 간의 연관성(독립성) 측정 지표
두 변수 완전한 독립 : Lift =1
양의 상관관계 : Lift 〉1
음의 상관관계 : Lift 〈 1

Step 03

데이터 분석

텍스트 분석

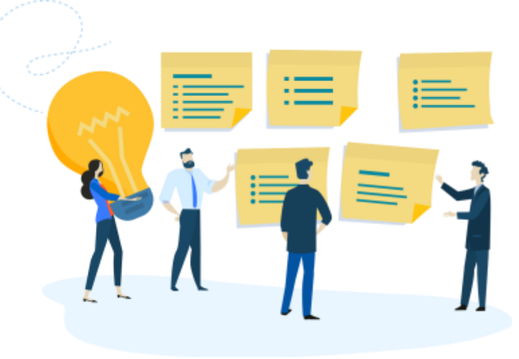
03 연관성 분석

| 향상도(Lift 지수) 활용

연관성 분석 진행

부도 기업 뉴스 기사에서 빈도 높은 46개의 단어 ⇄ 부도 발생 여부

단어	향상도	단어	향상도	단어	향상도
워크아웃	1.267018	법원	1.142941	처분	1.096879
자금난	1.254776	회생	1.142777	공시	1.092649
상장폐지	1.253094	손실	1.130004	자금	1.089513
회계법인	1.227669	구조조정	1.128239	발생	1.088638
퇴출	1.205449	M&A	1.125969	매각	1.087258
파산	1.199856	횡령	1.121429	증권	1.084934
법정관리	1.190867	우려	1.120915	전환	1.079542
한도액	1.166563	절차	1.120428	검찰	1.07927
등급	1.161309	영업이익	1.116643	주가	1.078412
채권단	1.154851	거래	1.114274	인수	1.070905
주주총회	1.153399	상장	1.111252	실적	1.061446
증자	1.152294	지분	1.111013	실적	1.061446
BW	1.146724	신용	1.106092	투자	1.057198
코스닥	1.145245	하락	1.101754	이사	1.045806
부도	1.14413	투자자	1.101754	감사	1.043767



단어가 뉴스 기사에 포함되지 않았을 때 기업의 부도 확률에 비해
단어가 포함된 경우 기업의 부도 확률 증가 비

'워크아웃', '자금난', '상장폐지'
→ 부도 직접 명시 단어 Lift 〉 1.2

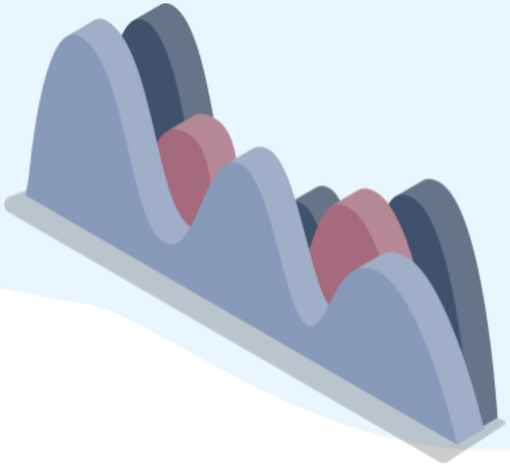
'회계법인', '거절', '퇴출'
→ 부정적인 단어

부도와 높은 연관성

Step 03

데이터 분석 04

모델링



데이터셋

23514(행) X 47(열)

독립 변수
부도 단어

종속 변수
부도 여부
부도(1) 정상(0)

예측모델링 실행

뉴스 기사의 단어 포함여부로 기업의 부도 여부 예측해 분류
대표적인 분류모델 의사결정나무와 랜덤 포레스트 활용

Decision tree
의사결정나무

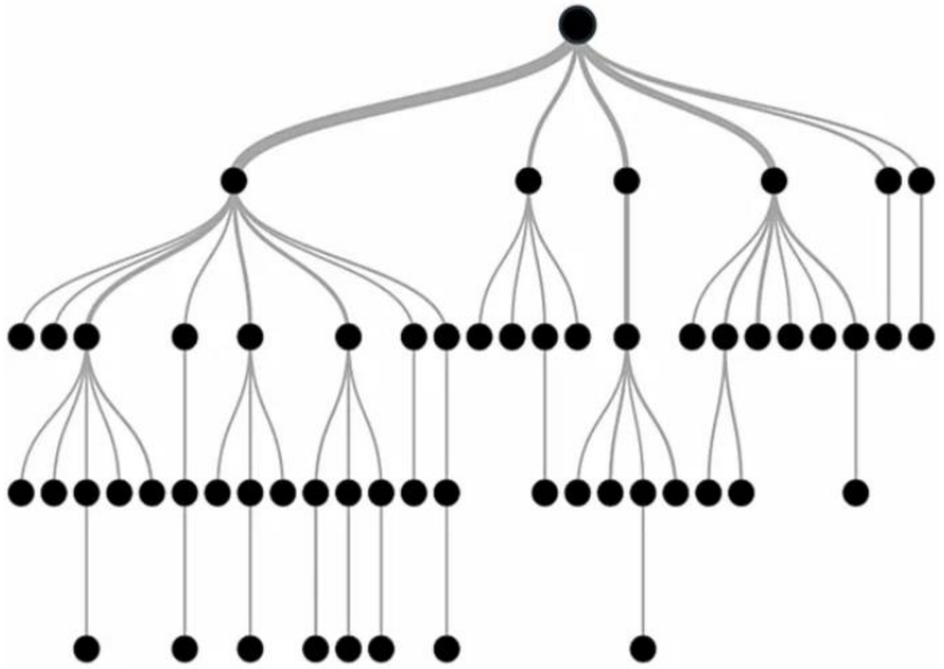
Random forest
랜덤포레스트

Step 03

데이터 분석 모델링

Decision tree
의사결정나무

의사결정규칙(decision rule) 도표화
→ 관심대상이 되는 집단 소집단으로 분류/예측 분석방법



| 의사결정나무 진행과정

01 분석의 목적과 자료구조에 따라 적절한 분리 기준, 정지규칙 지정하여 의사결정나무 생성
02 가지치기: 분류오류를 크게 할 위험이 높거나 부적절한 규칙 있는 가지 제거
03 타당성 평가: 이익도표, 위험도표, 검정용 자료에 의한 교차타당성 등을 이용해 평가
04 해석 및 예측: 의사결정나무 해석, 분류 및 예측모형을 설정

정지기준, 분리기준, 평가기준 등을 어떻게 지정하느냐에 따라서 서로 다른 의사결정나무 형성

장점

분류가 되는 의사결정 과정을 시각적으로 보여주어, 데이터 해석 용이
숫자형, 범주형 데이터 모두 사용 가능한 기법
데이터에 결측값 있어도 사용 가능
대규모의 데이터 분석 가능

Step 03

데이터 분석 모델링

Decision tree

의사결정나무

| 모형 평가

1) 빈도 분석 상위 7개의 단어

정분류율(accuracy) = 0.578

		Predicted	
		Positive	Negative
Observed	Positive	39	1931
	Negative	46	5039

2) 빈도 분석 단어 전체(46개)

정분류율(accuracy) = 0.804

		Predicted	
		Positive	Negative
Observed	Positive	863	409
	Negative	513	2918

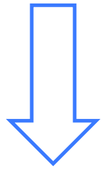
단어 수가 증가할수록 정확도 향상

Step 03

데이터 분석 모델링

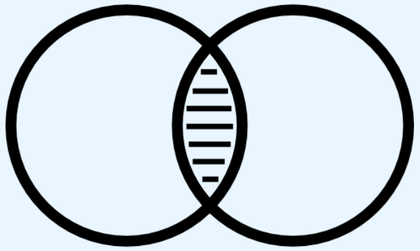
| 의사결정나무모델의 단점

01. 학습데이터의 미세한 변동에 최종결과가 크게 영향 받음
02. 나무의 노드 개수를 늘리면 과적합의 위험 발생
03. 가지치기로 과적합의 위험 해결 불가



| 앙상블(Ensemble) 기법: 여러 개의 분류모형의 결과를 종합해 분류의 정확도를 높이는 방식

앙상블 기법

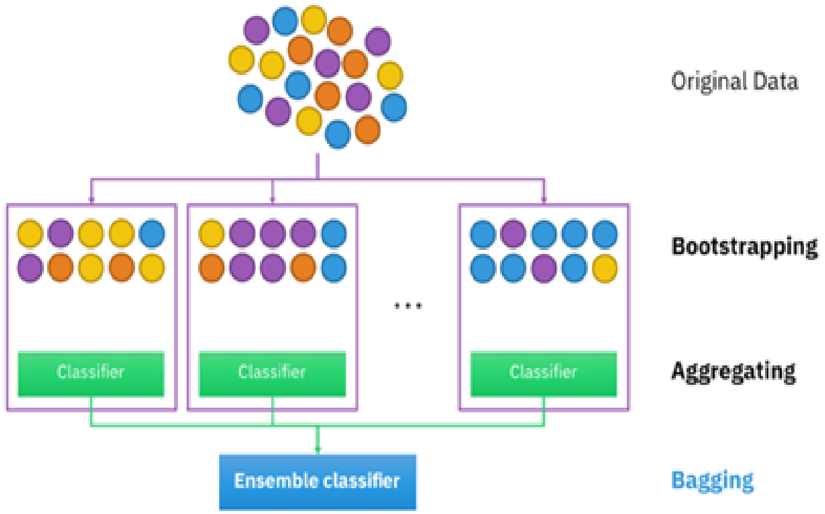


- 배깅(Bagging)
- 부스팅(Boosting)
- 랜덤 포레스트(Random forest)

Step 03

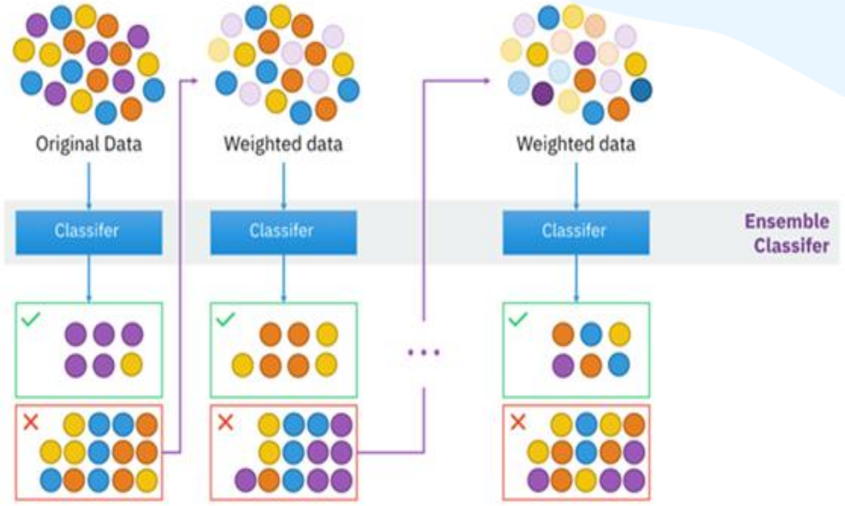
데이터 분석 모델링

| 배깅 (Bagging)



여러 개의 부트스트랩 자료를 생성한 후
각각의 예측모형을 생성, 예측된 결과를 결합해
최종 예측 결과를 도출하는 방식

| 부스팅 (Boosting)



예측력이 약한 예측모형을 결합해
예측력이 높은 예측모형을 만드는 기법

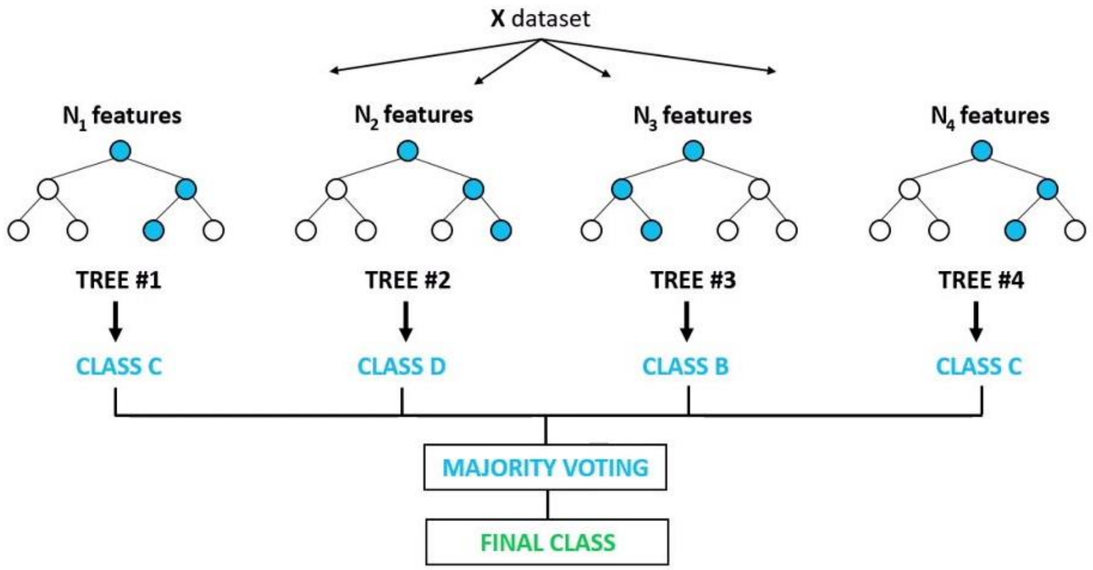
Step 03

데이터 분석 모델링

Random forest

랜덤 포레스트

| 다수의 의사결정나무모델에 의한 예측을 종합하는 앙상블 방식
| 배깅과 부스팅보다 더 많은 무작위성을 부여해 의사결정나무모델의 분산 ↓
| 결과 해석은 어렵지만 예측력 ↑

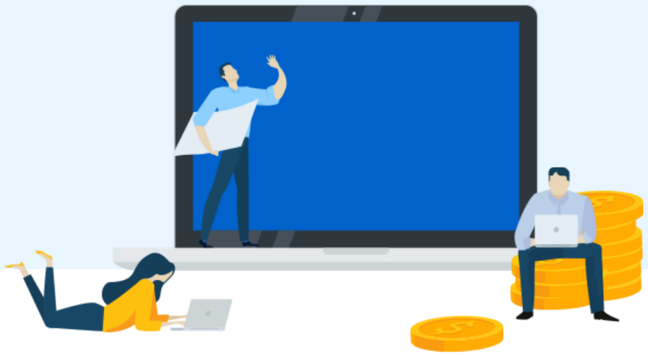


| 랜덤 포레스트 진행 과정

01. 부트스트랩 기법을 이용해 표본을 추출
02. 데이터를 생성하고 생성
03. train 데이터로 의사결정나무모델을 구축
04. 최종적으로 예측을 종합한 결과를 도출.

Step 03

데이터 분석 모델링



Random forest

랜덤 포레스트

| 모형 평가

1) 빈도 분석 상위 7개의 단어

정분류율(accuracy) = 0.726

		Predicted	
		Positive	Negative
Observed	Positive	25	1889
	Negative	14	5127

2) 빈도 분석 단어 전체(46개)

정분류율(accuracy) = 0.816

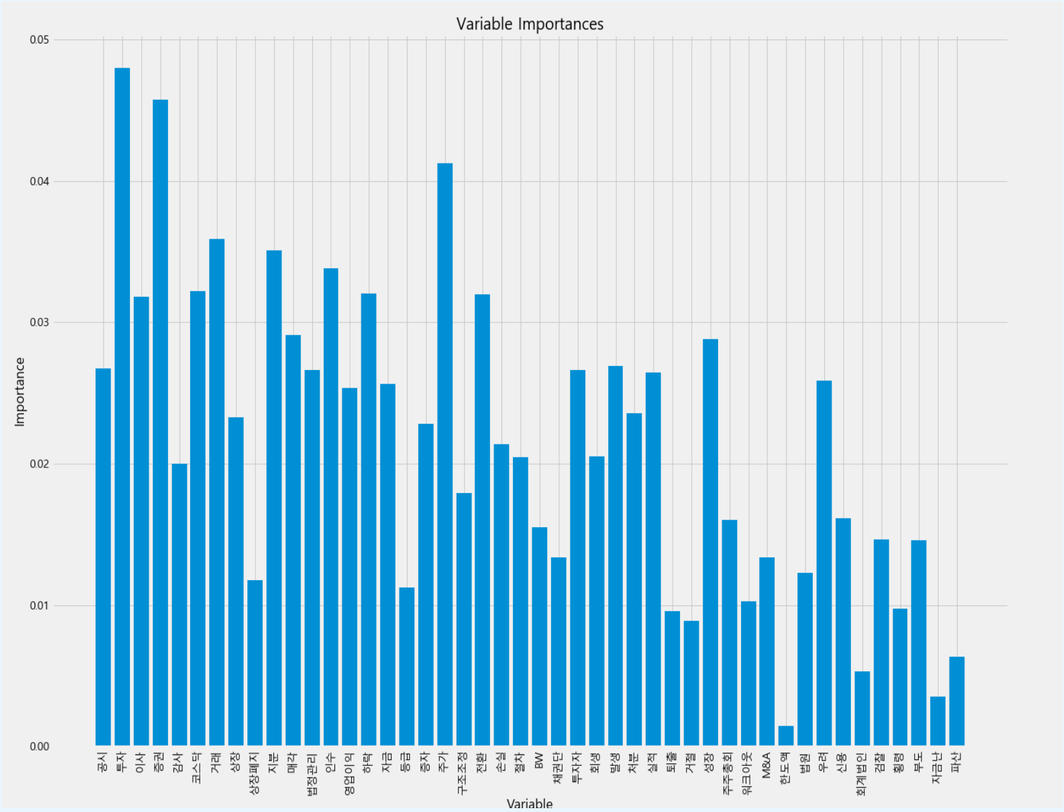
		Predicted	
		Positive	Negative
Observed	Positive	1092	865
	Negative	433	4665

46개 변수가 사용된 랜덤 포레스트 모델 최종 선정

Step 04

결론 및 평가

| 변수중요도



| 텍스트 및 모델

빈도•비중, 연관성 분석

- 부도기업: 부정적 단어의 수치 ↑

  ex) ‘상장폐지’, ‘자금난’

랜덤 포레스트 모델

- 단어 종류와 수가 많을수록 예측 정확도 ↑

| 성과

- 정상기업 뉴스 기사를 활용해 정상 운영이 가능한 기업의 예측도 가능
- 기존의 부도 예측 모형의 한계: 시장 정보를 활용, 비상장 기업 적용이 불가

  →뉴스 정보에서 비상장 기업 정보 수집 가능
- 뉴스 정보를 활용해 시장 정보 기반의 예측 대체 가능한 부도 예측 모형 제시

Step 04

결론 및 평가

| 한계점

- 크롤링 된 부도 기업과 정상 기업의 기사 수 차이
- 빈도 분석에서 적은 음의 GAP 단어 수
  → 신뢰도 부정적 영향
- 소규모 기업의 뉴스 기사 확보 어려움 (유명 기업에 편중되어 발행되는 기사)

| 향후 연구

- 비대칭성 해결: 부도 기업과 정상 기업의 기사 수를 맞춤으로써 비대칭성을 해결
- 산업군 별 부도 예측모형: 기업의 업종,산업의 속성에 따라 분류하여 진행
- 추가 정보 원천 결합: 뉴스 정보뿐만 아니라 웹 페이지, 공시자료 등 활용

Thank you

끝까지 들어주셔서 감사합니다.