

팜 경진대회 최종보고서

학생 팀별 작성용

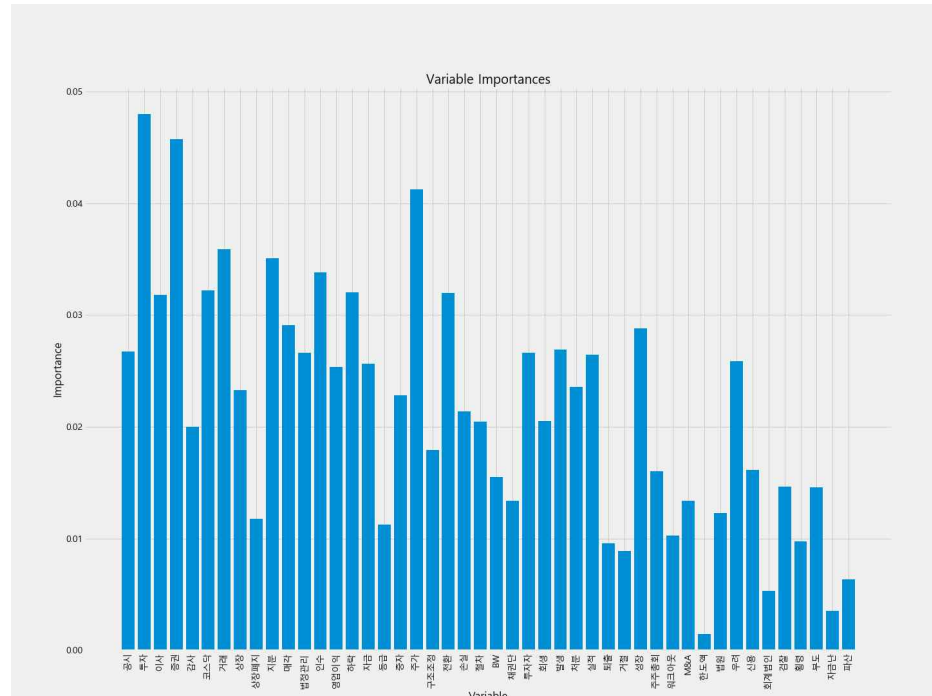
과제 수행원 현황						
수행 학기	<input type="checkbox"/> 2020년9월~2020년12월					
프로젝트명	뉴스 기사 텍스트 분석을 통한 기업 부도 예측 프로젝트					
팀명	빅데이터쉽G					
	학과	학번	성명	성별	연락처	E-mail
팀장	통계학과	2018*****	OOO			
팀원	통계학과	2018*****	OOO			
	통계학과	2018*****	이윤정			
	통계학과	2018*****	OO			
지도교수	교과목명					
	소속	<input checked="" type="checkbox"/> 컴퓨터공학전공 <input type="checkbox"/> 정보통신공학전공 <input type="checkbox"/> 멀티미디어공학전공 <input type="checkbox"/> 융합소프트웨어연계전공				
	성명	-				

과제 일반 현황					
작품(과제)명	뉴스 기사 텍스트 분석을 통한 기업 부도 예측				
보고서					
작품명 (프로젝트명)	텍스트 마이닝을 활용한 기업 부도 예측 모델 시스템				
# Key Words	텍스트 마이닝	기업 부도	뉴스 기사	신용평가	빅데이터
1.개발동기/ 목적/필요성및 개발 목표	1. 개발동기/목적/필요성 기업의 부도예측 모형은 금융 산업에서 중요한 과제로, 지속적으로 발전해 왔으며 여전히 중요한 연구 과제로 인식되고 있다. 특히 최근에는 코로나 사태로 인한 기업의 부도위험 증가로 그 관심은 더욱 커지고 있다.				

	<p>그렇다면 기업의 부도예측은 왜 중요한 과제로 인식되는 것일까? 그 이유는 빠르고 정확한 부도예측이 이해관계자들에게 예측 가능한 손실을 최소화할 수 있는 정보를 제공하기 때문이다. 경영자, 투자자에게는 구조조정, 경영정책 변화 등을 통해 손실을 최소화해주며, 금융회사의 경우 채권보전 조치, 신용평가 강화 등을 통한 대손 위험 최소화할 수 있다. 기업의 부도예측 모형은 이러한 중요성 때문에 그동안 많은 연구가 진행되어 정확도가 향상되었다.</p> <p>과거 연구의 대부분은 기업의 재무 정보와 시장 정보를 바탕으로 부도를 예측하였다. 재무 정보는 객관적이고 세부적인 기업에 대한 정보이다. 하지만 데이터 생성 주기가 연 혹은 분기 단위이기 때문에 상대적으로 길어 예측의 적시성이 떨어지는 한계가 있다. 시장 정보는 유가증권시장 참여자에 의하여 기업에 대한 정보가 가장 빠르게 반영되지만, 유가증권시장 상장 기업만 활용이 가능하고 비상장기업의 경우 적용이 불가능하다. 또, 주가에 영향을 주는 다른 요인에 대한 통제가 힘들다. 즉, 기업 경영환경, 거시경제 여건 등이 급속하게 변화하여 부도 기업에 대한 정확한 예측은 여전히 어려운 과제로, 재무 정보만으로도 예측할 수 있어도 빠르게 파악하기 어렵다. 위와 같은 단점을 가진 부도예측 모형을 보완하기 위해 본 조는 뉴스 기사를 활용하고자 한다. 데이터 분석과 인공지능 기법을 통해 새로운 데이터 원천인 뉴스 정보를 부도예측 과정에 활용할 수 있다. 더불어 정보량이 많은 이 뉴스 정보는 누구에게나 접근성이 뛰어나다. 빅데이터 연구 분야에 활용되는 텍스트 마이닝과 인공지능기법을 활용한다면 텍스트 정보를 측정 가능한 변수로 계량화할 수 있다. 뉴스는 기업에 대한 정보를 가장 빠르게 제공하는 원천 중 하나로, 기업 부실의 징후를 미리 알 수 있는 정보로서 충분한 가치가 있다고 판단했다.</p> <p>2. 개발목표</p> <p>2020년, 사회 전반에 걸쳐 큰 영향을 미치고 있는 코로나를 비롯해 기업의 외부적, 내부적 위험으로 인하여, 경영에 어려움을 가지고 있거나 부도가 난 기업들이 어떤 특징이 갖는지, 어떠한 징후가 나타났는지 텍스트 마이닝 방법론을 통하여 분석할 수 있다.</p> <p>현대 사회에서 한 기업의 부도는 그 기업과 연계된 다른 기업과 금융기관까지 영향을 주기 때문에 사회적으로 큰 손실이 발생한다. 이러한 사회적 손실을 줄이기 위해서 선행된 분석을 통하여 기업의 부도 가능성을 예측하고 미리 예방하여 이해관계자들의 피해를 최소화한다.</p>
<p>2.최종 결과물 소개</p>	<p>본 조의 프로젝트는 과거에 부도가 발생한 기업의 뉴스 기사 속 단어를 텍스트 마이닝을 통해 특징을 살펴보고, 현재 또는 미래에 발생할 수 있는 기업의 부도를 빠르고 정확하게 예측해보고자 하였다.</p> <p>부도기업의 뉴스 기사 속 단어 분석 결과 정상기업과 비교하여 ‘공시’, ‘감사’, ‘상장폐지’, ‘법정관리’와 같은 단어의 빈도가 극명하게 높은 것을 확인할 수 있었다. 단어의 연관성 분석에서는 ‘워크아웃’, ‘자금난’, ‘상장폐지’, ‘회계법인’, ‘거절’, ‘퇴출’의 단어가 부도여부와 큰 연관성을 갖는다는 결과가 나왔다. 주로 부정적인 단어의 빈도와</p>

연관성 수치가 부도기업에 대해 크게 나타났다.

랜덤 포레스트 모델로 기업의 부도를 예측해보았는데 예측결과는 정확도가 80%에 가까운 높은 수치를 얻을 수 있었다. 모델링 과정에서 뉴스 기사에 포함된 최대한의 많은 단어를 추출해 분석을 진행하면 더 높은 예측력과 정확도를 얻을 수 있다는 것을 알 수 있다.



<그림 1 랜덤포레스트의 변수 중요도>

비록 이번 프로젝트에서는 부도기업을 중심으로 진행했지만 정상기업 또한 뉴스 기사를 활용해 정상 운영이 가능한 기업의 예측도 충분히 가능하다.

기존의 부도 예측 모형은 시장 정보를 활용했기에, 시장 정보를 얻기 힘든 비상장 기업 등의 경우 적용이 불가능하다는 한계점이 있었다. 하지만 뉴스 정보의 경우 비상장 기업도 확보할 수 있고, 시장 정보 수준의 적시성 있는 데이터 수집이 가능하다. 따라서 뉴스 정보를 활용해 시장 정보 기반의 예측을 대체할 수 있는 부도 예측 모형을 제시하였다는 점에서 의미가 있다.

반면, 본 프로젝트 과정에서 몇 가지 한계점이 있었다. 우선, 크롤링된 기사들은 기업의 부도 시점부터 6개월 이전 기사들이다. 기존의 일반 기사에다가 부도일과 가까워지면서 부도 관련 기사의 양이 증가하여 결과적으로 부도 기업들의 기사 수는 많아지게 된다. 결국 부도 기업의 기사 수가 정상 기업의 기사 수보다 약 2배 가까이 차이나게 되었다. 결과적으로 부도 관련 단어의 비중이 큰 값을 가진다. 따라서 빈도 분석 시, 음의 GAP을 가지는 단어가 양의 값을 가지는 단어에 비해 그 수가 다소 적다. 이러한 요인들은 부도 예측 모형의 신뢰도에 부정적인 영향을 끼칠 수 있다는 한계점을 가지고 있다.

마지막으로 상대적으로 관심이 떨어지는 소규모 기업의 경우 양질의 뉴스 기사를 확보하는 것이 쉽지 않았다. 정상 기업의 경우, 6개월 동안 발행된 기사가 아예 없거나, 한 자릿수인 경우들이 있었다. 기업에 대한 뉴스 기사는 소수의 유명한 기업에 편중되어 발행되기 때문이다. 또한 분석 과정에서 형태소 단어 처리, 동의어 처리 등의 과정에서 데이터 전처리를 위한 분석자의 많은 수작업과 시간이 소요되었다. 이러한 한계점은 정보 수집 원천을 확대하고 정보 처리 기술 관련 연구가 지속적으로 진행된다면 많은 부분이 개선될 수 있을 것이라 기대된다.

향후 연구에서는 부도 기업과 정상 기업의 기사 수를 맞춤으로써 비대칭성을 해결하고, 업종별 부도 분석, 기업을 산업의 속성에 따라 분류하여 진행한다면 좀 더 정확하고 신뢰도 높은 산업군 별 부도 예측 모형을 만들 수 있을 것이다. 또한 본 연구에서는 뉴스 정보만 활용했지만, 웹 페이지, 공시자료 등 추가적인 정보 원천을 결합하여 사용한다면 보다 유의미한 결과를 얻을 수 있을 것이라 기대한다.

1. 데이터 준비

기업 부도예측 연구 과정에서 보다 유용한 결과를 얻기 위해서는 기업의 부도(부실)에 대한 명확히 정의 내려야 한다, 신용평가사에서 제공하는 부도기업 리스트의 경우, 각 평가사에서 신용등급을 보유한 회사들만 측정을 하기 때문에 조금씩 다를 수 있다. 본 조는 한국신용평가, NICE신용평가, 한국기업평가에서 정의한 부도를 따르며, 3개 평가사에서 제공한 부도 기업 목록을 활용하여 부도 기업리스트를 작성하였다. 기간은 2003년부터 2019년까지로, 부도기업으로 정의한 총 102개의 기업 선정하였다. 비교를 위해 부도기업의 규모, 형태, 중분류, 대분류, 설립연도와 일치하고 현재까지 정상적으로 운영 중인 102개의 기업을 선정하였다.

기업이 부도된 시점에서 6개월 이전에 보도된 뉴스 기사의 제목과 본문을 크롤링으로 수집하였다. 정상기업도 동일한 시점에 동일한 조건으로 진행하였다.

<표 1 수집한 뉴스 콘텐츠 기초통계량>

구분	기업 수	전체 기사 수	기업당 기사 수
부도기업	102	18628	182.63
정상기업	102	8759	85.87

기업 간 기사 수가 차이 나는 이유는 수집한 기사는 기업의 부도발생 6개월 이전의 기사이므로 부도기업과 관련된 기사가 정상기업의 기사보다 많다.

2 데이터 전처리

크롤링 한 직후의 데이터는 뉴스 기사를 그대로 가져온 것이기 때문에 정제되어 있지 않아 분석에 부적합하다. 따라서 텍스트 분석이 가능하도록 전처리를 시행해야 한다. 기업명과 기사 제목, 기사 본문, 3개의 변수로 구성된 부도 기업과 정상 기업의 데이터 셋 2개를 생성 후 전처리를 진행하였다.

2.1 명사추출

3.프로젝트 추진 내용

기사 제목과 내용에서 KoNLP 패키지를 활용해 대명사, 동사, 형용사, 부사, 조사를 제외한 명사만을 추출한다.

2.2 불용어 제거

분석에 영향을 줄 수 있는 특수문자/숫자/영어/한자 등의 불용어를 제거하였다.

<표 2 뉴스 기사 전처리 전후 비교>

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

3. 텍스트 분석

3.1 동의어 처리

뉴스 기사에서는 동일한 의미를 가진 단어가 다양한 형태로 표현될 수 있다. 효율적이고 정확한 텍스트 분석을 위해 동일한 의미를 갖는 표현들을 하나의 단어로 통일시켰다.

<표 3 동의어 처리대상>

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
상장폐지	상폐
법정관리	법정
감사	감사보고서, 외부감사인, 감사인, 재감사, 회계감사, 외부감사, 재감사 보고서, 국정감사
매각	지분매각, 자산매각, 매각대금, 매각설, 매각자
상장	상장사, 상장기업, 상장주식, 상장회사
지분	지분율, 보유지분, 출자지분
인수	인수합병, 인수의향서, 인수자, 인수자금, 인수대금, 인수설, 인수협상, 인수계약
하락	하락세, 하락폭, 주가하락
주가	종합주가지수
이사	대표이사, 사외이사, 이사회, 등기이사, 상무이사, 임시이사회, 이사진
절차	정리절차
자금	운영자금, 투자자금, 시설자금, 주식매입자금, 인수자금, 자금력, 여유자금
투자자	주식투자자, 일반투자자
채권단	주채권단인, 주채권단
손실	당기순손실, 순손실, 법인세비용차감전계속사업손실
전환	주식정환, 전환사채

영업이익	당기순이익, 순이익, 이익, 경상이익, 반사이익, 이익률
구조조정	기업구조조정, 분쟁조정, 조정
등급	신용등급
실적	영업실적, 경영실적, 실적호전, 수출실적
BW	국내BW행사, 해외BW행사, 국내BW, BW발행, BW행사
처분	가처분, 가처분신청, 주식처분
신용	신용등급, 신용도, 신용거래, 신용정보
회생	기업회생, 기사회생
우려	우려감
성장	성장동력, 성장세, 급성장, 성장률, 고성장, 고속성장
법원	서울중앙지방법원, 지방법원, 대법원
주주총회	정기주주총회, 주총
회계법인	상장법인
검찰	대검찰청
횡령	횡령혐의
증자	유상증자, 제자배정유상증자, 유상증자권리락, 유증,
M&A	흡수합병, 인수합병, 합병
자금난	자금지원
부도	부도설, 부도위기, 부도처리, 부도금액
파산	파산부

3.2 빈도 분석

KoNLP 패키지를 활용해 부도기업 데이터에서 총 440607개의 명사를 추출해 빈도 분석으로 상위 200개까지의 명사를 나열했다. 그중에서도 제품명, 사람 이름, 기업 이름, 지역과 같은 기업 활동과 관련 없는 단어를 제외하고 46개의 명사를 선택했다. 선택한 명사로 다시 빈도 분석을 실행하고 정상기업 데이터의 명사와의 빈도를 비교해보았다.

빈도: 뉴스 기사 텍스트 내에서 해당 단어의 발생 빈도 수 (전처리 이후 데이터 기준)

비중: 전체 단어 발생 빈도 중 해당 단어의 비중

<표 4 부도기업과 정상기업 간의 단어 빈도 비중 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%
코스닥	11313	0.593%	3034	0.362%	0.231%
거래	10088	0.529%	2753	0.329%	0.200%
상장	9422	0.494%	1634	0.195%	0.299%
상장폐지	8793	0.461%	1145	0.137%	0.324%
지분	8608	0.451%	3755	0.448%	0.003%
매각	8525	0.447%	2757	0.329%	0.117%
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%
하락	6998	0.367%	2688	0.321%	0.046%

자금	6678	0.350%	1290	0.154%	0.196%
등급	6268	0.328%	453	0.054%	0.274%
증자	6037	0.316%	1112	0.133%	0.184%
주가	5536	0.290%	2380	0.284%	0.006%
구조조정	5500	0.288%	1507	0.180%	0.108%
전환	5294	0.277%	1645	0.196%	0.081%
손실	5208	0.273%	1107	0.132%	0.141%
절차	5116	0.268%	582	0.070%	0.199%
BW	4960	0.260%	852	0.102%	0.158%
채권단	4857	0.255%	444	0.053%	0.201%
투자자	4716	0.247%	1035	0.124%	0.124%
회생	4450	0.233%	643	0.077%	0.156%
발생	4199	0.220%	987	0.118%	0.102%
처분	3862	0.202%	896	0.107%	0.095%
실적	3774	0.198%	2780	0.332%	0.134%
퇴출	3754	0.197%	317	0.038%	0.159%
거절	3493	0.183%	285	0.034%	0.149%
성장	3447	0.181%	2540	0.303%	0.123%
주주총회	3190	0.167%	812	0.097%	0.070%
워크아웃	2983	0.156%	210	0.025%	0.131%
M&A	2969	0.156%	1266	0.151%	0.004%
한도액	2812	0.147%	130	0.016%	0.132%
법원	2787	0.146%	422	0.050%	0.096%
우려	2627	0.138%	800	0.096%	0.042%
신용	2170	0.114%	359	0.043%	0.071%
회계법인	1990	0.104%	133	0.016%	0.088%
검찰	1868	0.098%	394	0.047%	0.051%
횡령	1543	0.081%	205	0.024%	0.056%
부도	1345	0.070%	76	0.009%	0.061%
자금난	947	0.050%	61	0.007%	0.042%
파산	571	0.030%	87	0.010%	0.020%

Total	1908368	100%	837407	100%	

GAP의 값이 양(+)의 값인 경우 부도기업에서 단어의 비중이 크게 나타나며 부도와 관련된 단어일 확률이 높아진다. 반면 음(-)의 값인 경우 정상기업에서 단어의 비중이 크게 나타나며 부도와 관련 없는 단어일 확률이 높다.

빈도 분석 결과, 부도기업에서는 ‘공시’, ‘감사’, ‘상장폐지’, ‘법정관리’, ‘등급’ 단어가 정상기업보다 빈도 비중이 확연한 차이로 높게 나타났다. ‘인수’, ‘영업이익’, ‘실적’, ‘성장’ 단어는 정상기업에서 부도기업보다 빈도 비중이 높게 나타났다. 부도기업은 부도와 관련된 부정적 단어가, 정상기업에서는 긍정적 단어가 큰 GAP 차이를 보여 준다.

‘공시’, ‘투자’, ‘이사’, ‘증권’과 같은 단어는 GAP의 차이가 있지만 빈도 비중이 부도기업과 정상기업에서 모두 높은 값을 갖는다.

3.3 연관성 분석

앞서 빈도 분석에서 살펴본 부도 기업 뉴스 기사에서 빈도가 높은 46개의 단어에 대해 연관성 분석 방법론 중 Lift 지수(향상도)를 측정해 부도 발생 여부와의 연관성을 분석해본다.

< 표 5 부도 여부와 부도단어 간의 연관성 분석 >

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

부도에 대해 직접적으로 나타내는 '워크아웃', '자금난', '상장폐지'와 같은 단어들은 1.2가 넘는 Lift값이 도출되었으며 그 뒤를 이어 '회계법인', '거절', '퇴출'과 같은 대부분 부정적인 단어가 부도와 높은 연관성을 갖는다는 결과를 얻었다.

4. 모델링

종속변수인 부도여부, 독립변수인 부도단어로 구성된 23514(행) x 47(열)의 데이터셋 형태로 예측모델링을 실행한다.

본 프로젝트에서는 뉴스 기사의 단어 포함 여부로 기업의 부도 여부를 예측해 분류하는 것이 핵심이다. 따라서 부도(1), 정상(0)을 알아보기 위해 대표적인 분류모델의 의사결정나무와 랜덤 포레스트를 활용한다.

4.1 의사결정나무(Decision tree)

의사결정나무 또는 나무 모형이라고 불린다. 의사결정 규칙을 나무의 구조로 나타내 전체 자료를 소집단으로 분류·예측하는 분석모델이다. 분류변수와 분류 기준값에 의해 상위노드에서 하위노드로 분류가 진행된다.

1) 최종적으로 선정한 46개의 단어 중 빈도 분석 상위 7개의 단어를 선정해 의사결정나무모델로 기업의 부도 여부를 예측해보았다.

		(오류분포표 1)	
		Predicted	
		positive	negative
Observed	positive	39	1931
	negative	46	5039

정확도 : 71.9%

2) 선정한 46개의 단어를 모두 포함한 의사결정나무 모델로 부도 여부를 예측해보았다.

		(오류분포표 2)	
		Predicted	
		positive	negative
Observed	positive	863	409
	negative	513	2918

정확도 : 80.4%

7개의 단어만 사용한 모델링 결과보다 높은 정확도를 도출했다. 단어 변수가 증가할수록 정확도가 향상된다는 것을 알 수 있다.

4.2 랜덤 포레스트

랜덤 포레스트는 다수의 의사결정나무모델에 의한 예측을 종합하는 방식이다. 여러 예측모형의 결과를 결합하기 때문에 분산을 낮추는데 효과적이며 의사결정나무모델보다 예측률이 뛰어나다.

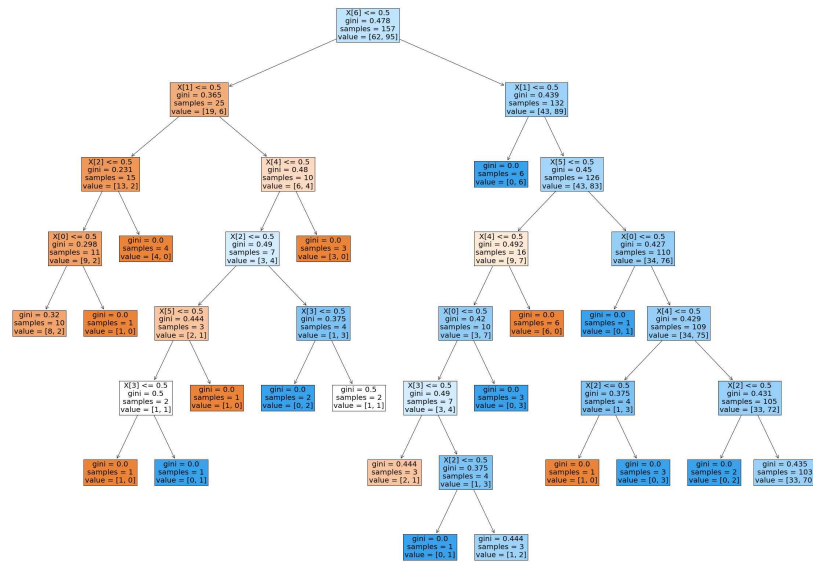
1) 의사결정나무모델에서와 동일하게 상위 7개의 단어를 뽑아서 예측해보았다.

		(오류분포표 3)	
		Predicted	
		positive	negative
Observed	positive	25	1889
	negative	14	5127

정확도 : 72.6%

상위 7개의 단어를 포함한 의사결정나무모델보다는 정확도가 향상된 것을 볼 수 있다. 하지만 46개의 단어를 사용한 의사결정나무모델보다는 현저하게 정확도가 낮아졌다.

2) 46개 단어를 모두 사용해 랜덤 포레스트 모델을 사용해보았다.



<그림2 랜덤포레스트에 사용된 DT>

(오류분포표 4)		Predicted	
		positive	negative
Observed	positive	1092	865
	negative	433	4665

정확도 : 81.6%

81.6%의 정확도로 분석을 위해 활용한 총 4개의 모델 중 가장 높은 정확도를 보였다. 본 조는 분석을 위해 46개의 단어 변수가 사용된 해당 랜덤 포레스트 모델을 최종 선정하였다.

1. 크롤링 기법

크롤링이란 웹상에 존재하는 정보들을 수집하는 작업을 말하며, 크롤링을 하는 프로그램은 크롤러라 부른다. 크롤링에는 여러 가지 방법이 있는데 크게 3가지를 뽑을 수 있다.

- 1) 오픈 API를 활용해 받은 데이터 중 필요한 데이터만 사용하는 방법
- 2) HTML 소스를 가져와 원하는 정보를 사용하는 방법 - 정적 수집 방법
- 3) 브라우저를 조작해 원하는 정보를 사용하는 방법 - 동적 수집 방법

크롤링에는 동적 수집과 정적 수집이 존재한다. 동적 수집은 브라우저를 통해 연속적으로 접근하여 수집 대상에 한계가 거의 없다는 장점이 있으나, 사용자와 상호작용하며 데이터를 수집하기 때문에 속도가 매우 느리다는 단점이 있다. 정적 수집은

동적 수집과 달리 주소를 통해 접근하므로 속도는 빠르지만, 수집 대상에 한계가 존재한다는 단점이 있다. 그러므로, 각 기업별로 수집기간 동안의 뉴스 개수가 다르기 때문에 동적 크롤러인 셀레니움을 활용하는 것이 유용하지만 방대한 데이터 양과 데이터 수집에 오랜 시간이 걸리는 셀레니움의 단점으로 인해 정적 크롤러를 사용하여 데이터 수집을 진행한다.

본 조에서 사용한 정적 수집의 원리를 알기 위해선 웹페이지의 구성을 알아야 한다. 웹페이지는 HTML 문서로 작성되어 있다. 그리고 이는 인터페이스를 참조할 수 있는 CSS 파일과 페이지 상호작용을 위한 JavaScript파일을 참조할 수 있다. 모든 웹페이지의 HTML 문서 확인이 가능하고 참조된 CSS 문서, 포함된 내용 모두 알 수 있다. 부도 기사를 크롤링하고 싶다면, 브라우저에 "부도"라고 검색된 페이지의 HTML 문서를 확인해본다면 기사제목을 어떤 HTML 태그로 사용했는지 알 수 있다. 이러한 방식으로 원하는 데이터를 추출하기 위해서는 HTML 문서를 확인한 뒤 원하는 데이터를 추출할 수 있도록 크롤러를 만들면 된다.

일반적으로 기업 부실은 부실징후가 일정 기간에 걸쳐 나타나는 경우가 많으므로 부실 정도의 측정이 어렵고 재무적, 경영적 그리고 법적인 처리 과정이 서로 다르며 국가별로 경제 환경이 다르므로 부실기업에 대한 정의나 개념을 한마디로 설명하기가 어렵다. ¹⁾ 그러나 일반적으로 부실 유형에 따라 경제적 부실, 기술적 지급 불능, 실질적 지급 불능 그리고 파산으로 구분하여 부실화 개념을 정의한다.

<그림 1> 기업 부실화 단계별 현상 및 원인

단계	구분	나타나는 현상	주요 원인
1단계	경제적 부실	<ul style="list-style-type: none"> 총수익(TR) < 총비용(TC) ROIC < WACC, EVA < 0 실현이익 < 기대이익 	<ul style="list-style-type: none"> 비효율적인 경영(비용 증가, 투자 실패) 제품 경쟁력 저하(매출 부진, 수익성 저하) 수익성 저하
2단계	기술적 지급 불능	<ul style="list-style-type: none"> 단기채무 > 보유현금 운전자금 부족 Payment Record 부실 	<ul style="list-style-type: none"> 재무구조 안정성 취약 방만한 자금관리 자금조달 능력 부족
3단계	실질적 지급 불능	<ul style="list-style-type: none"> 매출 < 차입금 영업CF 지속적(-) 자본잠식 등 순자산가치(-) 	<ul style="list-style-type: none"> 과다한 저수익 자산 보유(재고, 매출채권) 자금조달 곤란, 현금 흐름 악화 재무구조 악화(과다한 차입금)
4단계	파산단계	<ul style="list-style-type: none"> 유동성 부족 재무용통성 저하 회생가능성 상실 부도 	<ul style="list-style-type: none"> 파산적 지급 불능 상태 심화 유동성 악화 및 재무용통성 상실 총체적 신용상태 부실 계속기업가치 < 청산가치

<그림 3> 기업 부실화 단계별 현상 및 원인>

기업의 부실화는 단계별로 진행되는데, 손실을 최소화하기 위해서는 조기에 부실 징후를 발견하여 대응하는 것이 매우 중요하다.

기존의 재무 정보를 바탕으로한 부실 예측 모형에 따르면 부실 1년 전에는 예측력이 높게 나타나 기업 부도의 예측 모형으로 이용가능하다. Beaver의 연구의 경우 예

측 오류 비율값이 1년 전은 13%, 부실 5년은 22%로 기간 1년이 더 낮았고, Altman의 연구에 따르면 예측결과가 부실 1년 전에는 91.2~97%, 2년 전에는 72%, 3년 전에는 48%로 나타나 예측 기간이 길수록 예측능력은 떨어졌다. 본 조는 예측의 정확성을 위해 부도 1년 전의 정보를 사용하되 예측의 적시성을 위해 예측 기간을 6개월로 설정하고 데이터 수집을 진행하였다.

수집 기간 6개월에 대한 뉴스 기사를 크롤링하기 위해, 파이썬의 datetime을 사용하여 부도일로부터 6개월 전 일자를 변수 startday, 부도일을 변수 endday로 설정하는 작업이 선행되었다. 정적 크롤러인 파이썬의 request와 BeautifulSoup 패키지를 통해 기업명과 수집 기간을 쿼리로 사용하여 네이버 검색창 뉴스탭에 접근한다. 이때, 네이버 뉴스탭에서 검색되는 뉴스는 신문사 자체 사이트로 링크가 연결되어 신문사 별로 사이트의 html 구조가 다르므로 뉴스 전문을 크롤링하는 데 어려움이 있다. 정적 크롤러로 데이터 수집을 하기 위해서는 동일한 html 구조로 구축된 사이트가 필요하기 때문이다. 따라서, 네이버 뉴스탭에서 검색되는 뉴스 중 '네이버 뉴스홈' 플랫폼에서 별도로 제공되는 뉴스만 수집했고 해당 플랫폼에서 제공되지 않는 뉴스는 수집에서 제외했다. 기업별로 검색되는 총 기사 수의 수가 서로 상이하기 때문에 이에 최대 500개의 기사를 제한으로 수집했다.

2. 명사 추출

R 프로그램의 한글 자연어 분석 패키지인 KoNLP(Korean Natural Language Processing)에는 한국어를 분석할 수 있는 27개의 함수가 포함되어있다. KoNLP 패키지는 JAVA로 만들어진 패키지이기 때문에 rJAVA 패키지를 설치한 후에 사용할 수 있다.

KoNLP 패키지는 시스템 사전을 사용하는 useSysDic 함수와 NIA 사전을 사용하는 useNIADic, 세종 사전을 사용하는 useSejongDic을 지원한다. 본 프로젝트에서는 약 100만개의 단어로 구성된 NIA 사전을 활용한다.

크롤링 직후의 데이터는 3개의 변수, 기업명, 기사제목, 내용으로 구성되어있다. 기업명의 경우 고유명사이기 때문에 별도의 전처리는 하지 않는다. 데이터 셋에서 기업명을 분리하고 뉴스 기사 제목과 내용에 대해서만 전처리를 진행한다.

명사를 추출하는 함수 extractNoun으로 문장에서 동사, 조사, 형용사를 제외하고 명사만을 추출한다. 특수 문자와 한자, 영어, 숫자 또한 텍스트 분석에 적절하지 않기 때문에 gsub(pattern, replacement) 함수를 이용해 공백으로 대체한다. 그 결과 명사 단어로만 구성된 데이터를 얻을 수 있다.(<표2> 참고)

<표 6 부도기업 데이터 구성>

기업명	제목	내용
...

<표 7 정상기업 데이터 구성>

기업명	제목	내용
...

3. TF-IDF 빈도 분석

TF-IDF(Term Frequency-Inverse Document Frequency) 가중치 모델은 텍스트 마이닝을 위해 문서 내부에서 특정 단어의 중요도를 평가한 통계적 수치이다. TF-IDF 값이 큰 단어는 문서의 주제를 결정할 확률이 커지며 해당 수치를 활용하면 주요 키워드의 선정 기준을 설정할 수 있다.

TF 값은 지정한 문서 내에서 특정 단어의 빈도를 나타낸 값으로 빈도가 높을수록 문서 내에서의 중요도가 크다고 할 수 있다.

IDF 값은 DF(Document Frequency) 값의 역수를 뜻한다. 문서군에서 특정 단어가 자주 사용될 때는 핵심 단어가 아닌 보편적인 단어가 반복되는 것으로 볼 수 있다. 이것을 DF 값이라하며 보편적인 단어의 가중치 값을 제외시켜야 하기 때문에 역수인 IDF 값을 사용한다.

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

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

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

해당 개념을 활용해 추출한 단어들의 빈도 분석을 실행한다. 전처리 과정이 끝난 데이터에서 추출된 각 명사들의 빈도를 계산한 TF 값을 계산한다. TF 값은 전체문서가 아닌 하나의 문서에서의 값을 도출한다. 전체문서에서 보편적인 단어를 제외시키기 위해 IDF 값을 도출한다. 빈도가 높은 순서대로 나열한 후 상위 200개의 단어 중 기업과 관련 없는 단어를 수작업으로 제외한 다음 최종적으로 46개의 단어를 선정한다.

4. 연관성 분석

보통 기업의 데이터베이스에서 상품의 구매, 서비스 등 일련의 거래 또는 사건들 간의 규칙을 발견하여 IF - Then의 구조로 분석 결과의 연관성을 파악하는 데이터마이닝 방법론이 연관성 분석이다. 연관성 분석은 흔히 장바구니 분석 또는 서열분석이라고 불리기도 한다. 대규모 거래 데이터에 대해 작업을 할 수 있으며 이해하기 쉬운 규칙을 생성해준다는 장점이 있다. 하지만 작은 데이터셋에는 효율성이 떨어진다는 단점이 있다.

연관규칙은 조건과 반응의 형태(if-then)로 이루어져있다. (If A then B: 만일 A가 일어나면 B가 일어난다.) “아메리카노를 마시는 손님 중 17%가 스콘을 먹는다.”는 연관규칙의 예로 볼 수 있다.

연관성 규칙의 측도에는 지지도, 신뢰도, 향상도가 있다. 산업의 특성에 따라 측도값을 바탕으로 적합한 규칙을 선택해야한다.

1) 지지도(support)는 전체 거래 중 항목 A와 항목 B를 동시에 포함하는 거래의 비율이다.

$$\text{지지도} = P(A \cap B) = \frac{A \text{와 } B \text{가 동시에 포함된 거래 수}}{\text{전체 거래 수}} = \frac{A \cap B}{\text{전체}}$$

2) 신뢰도(confidence) 는 항목 A를 포함하는 거래 중에서 항목 A와 항목 B가 같이 포함될 확률로 연관성의 정도 파악이 가능하다.

$$\text{신뢰도} = P(A \cap B) = \frac{A \text{와 } B \text{가 동시에 포함된 거래 수}}{A \text{를 포함하는 거래 수}} = \frac{\text{지지도}}{P(A)}$$

3) 향상도(Lift)는 A가 구매되지 않았을 때 품목 B의 구매확률에 비해 A가 구매됐을 때 품목 B의 구매확률의 증가 비이다. 연관규칙 $A \rightarrow B$ 는 품목 A와 품목 B의 구매가 서로 관련이 없는 경우에 향상도가 1이 된다.

$$\text{향상도} = \frac{P(B|A)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)} = \frac{A \text{와 } B \text{가 동시에 포함된 거래 수}}{A \text{를 포함하는 거래 수} \times B \text{를 포함하는 거래 수}} = \frac{\text{신뢰도}}{P(B)}$$

연관성 분석 알고리즘으로는 1세대의 Apriori와 2세대의 FP-Growth이 있다.

Apriori는 최소 지지도 이상의 빈발항목집합을 찾은 후 그것들에 대해서만 연관규칙을 계산한다. (빈발항목집합: 최소 지지도보다 큰 지지도 값을 찾는 품목의 집합) 1994년에 발표된 알고리즘으로 구현과 이해가 쉽다는 장점이 있으나, 지지도가 낮은 후보 집합 생성 시 아이템의 개수가 많아지면 계산 복잡도가 증가한다는 문제점을 가지고 있다.

FP-Growth 알고리즘은 후보 빈발항목 집합을 생성하지 않고 FP-Tree를 만든 후 빈발항목 방식을 통해 사용해 Apriori 보다 더 빠르게 빈발항목집합을 추출할 수 있는 방법이다. Apriori의 약점을 보완하기 위해 고안된 알고리즘으로 데이터베이스 스캔 횟수가 작고 빠른 속도로 분석이 가능한 강점을 지니고 있다.

4. 예측모델

4.1 의사결정나무

의사결정나무는 분류함수를 의사결정 규칙으로 이루어진 나무 형태로 그리는 방법이다. 나무구조는 연속적으로 발생하는 의사결정 문제를 시각화해 의사결정이 이뤄지는 시점과 성과를 한눈에 볼 수 있다. 주어진 입력값에 대해 출력값을 예측하는 모형으로 분류나무와 회귀나무 모형이 있다. 목표변수가 이산형이면 분류나무를, 연속형이면 회귀나무를 사용한다. 의사결정 나무는 노드들로 구성되어 있는데, 상위 마디를 부모마디(parent node)라 하고, 하위 마디를 자식마디(child node)라 하며, 더 이상 분기되지 않는 마디를 최종노드(terminal node)라고 부른다
의사결정나무 형성과정은 크게 성장(growing), 가지치기(pruning), 타당성 평가, 해석 및 예측으로 이루어진다.

1) 성장단계 : 각 마디에서 적절한 최적의 분리규칙(splitting rule)을 찾아서 나무를 성장시키는 과정으로 적절한 정지규칙(stopping rule)을 만족하면 중단한다.

분리 규칙에서 분리기준은 이산형 목표변수의 경우 카이제곱 통계량 p 값, 지니지수, 엔트로피 지수가 있으며 연속형인 경우에는 분산분석에서 f 통계량, 분산의 감소량이

있다. 더 이상 분리가 일어나지 않고 현재의 끝마디가 되도록하는 정지규칙에서의 정지 기준은 의사결정나무의 깊이 지정 후, 끝마디의 레코드 수의 최소 개수를 지정함으로써 구한다.

2) 가지치기 단계 : 큰 오차가 생길 위험이 높거나 부적절한 추론규칙을 가지고 있는 가지 또는 불필요한 가지를 제거하는 단계이다.

3) 타당성 평가 단계 : 이익도표, 위험도표, 혹은 시험자료를 이용하여 의사결정나무를 평가하는 단계이다.

4) 해석 및 예측 단계 : 구축된 나무모형을 해석하고 예측모형을 설정한 후 예측에 적용하는 단계이다.

목표변수가 범주형 변수인 의사결정나무의 분류규칙을 선택하기 위해서는 카이제곱통계량, 지니 지수, 엔트로피 지수를 활용한다. 카이제곱통계량의 값은 그 값이 작을수록 자식노드 간의 이질성이 큼을 의미한다. 지니지수는 노드의 불순도를 나타내는 값으로, 값이 클수록 이질적이며 순수도가 낮다고 볼 수 있다. 엔트로피 지수는 무질서에 대한 측도로 값이 클수록 순수도가 낮다고 본다. 엔트로피 지수가 가장 작은 예측 변수와 이때의 최적분리 규칙에 의해 자식마디를 생성한다. 따라서 이 값들이 가장 작아지는 방향으로 가지분할을 진행해야한다.

의사결정나무의 주요 알고리즘에는 CART, C4.5와 C5.0, CHAID가 있다. CART는 앞서 설명한 방식이 가장 많이 활용되는 알고리즘으로 불순도 측도로 목표 변수가 범주형일 경우 지니지수를 이용, 연속형인 경우 분산을 이용한 이진분리를 사용한다. 개체 입력변수 뿐만 아니라 입력변수들의 선형결합들 중에서 최적의 분리를 찾을 수 있다. C4.5와 C5.0은 CART와 다르게 각 마디에서 다지분리가 가능하며 범주형 입력 변수에서는 범주 수만큼 분리가 일어난다. 불순도 측도로는 엔트로피지수를 사용한다. 마지막으로 CHAID는 입력변수가 반드시 범주형이어야하며 가지치기를 하지 않고 적당한 크기에서 나무모형의 성장을 중지한다. 불순도의 측도로는 카이제곱 통계를 사용한다.

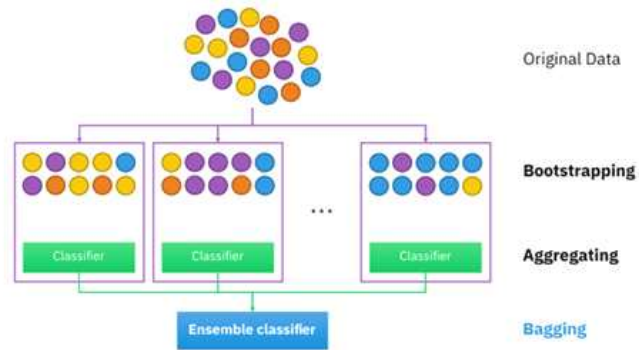
4.2 랜덤 포레스트

의사결정나무는 분류모델로 활용될 수 있지만 몇 가지의 단점을 갖고 있다. 학습데이터의 미세한 변동에도 최종결과가 크게 영향을 받으며 나무의 노드 개수를 늘리면 과적합이 발생할 위험이 있다. 가지치기를 시행해도 과적합의 위험이 남아있다. 의사결정나무모델의 단점을 해결하기위해 랜덤 포레스트 모델을 활용할 수 있다. 랜덤 포레스트 모델에 관해 자세하게 설명하기 전에 앙상블에 대한 개념을 설명한다.

앙상블(ensemble) 기법은 여러 개의 분류모형의 결과를 종합해 분류의 정확도를 높이는 방식이다. 앙상블 기법에는 대표적으로 배깅(bagging)과 부스팅(boosting), 랜덤 포레스트(random forest)가 있다.

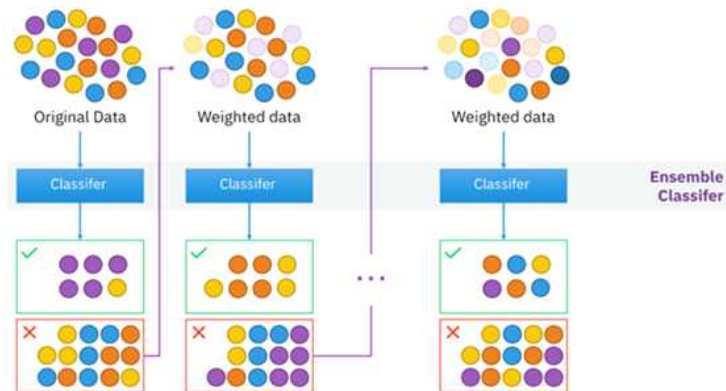
앙상블 기법은 개별 모형들의 평균을 취해 편의를 제거해 어느 쪽에도 치우치지 않는 결과를 얻을 수 있으며 두 번째로 모형을 종합함으로써 분산을 감소시킨다. 또, 과적합이 없는 각 모형으로부터 예측을 결합해 과적합의 가능성을 줄일 수 있다는

장점을 갖는다.



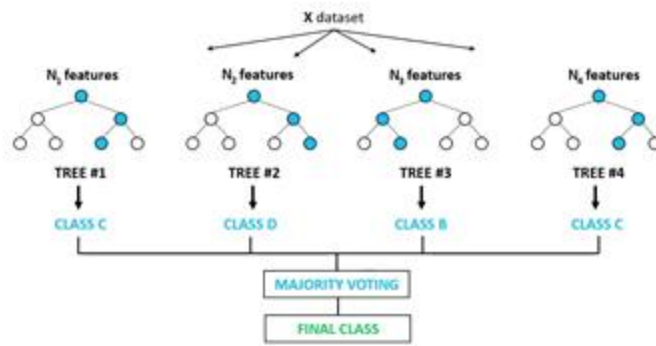
< 그림 4 배깅 기법 >

배깅은 여러 개의 부트스트랩 자료를 생성한 후 각각의 예측모형을 생성, 예측된 결과를 결합해 최종 예측 결과를 도출하는 방식이다. 반복추출을 하기때문에 동일한 데이터가 한 표본에 여러 번 추출될 수 있고 추출되지 않는 데이터가 존재할 수도 있다. 가지치기는 의사결정모델에서 가장 어렵고 중요한 부분이다. 하지만 배깅에서는 가지치기를 하지 않고 최대로 성장한 의사결정나무를 활용한다. 배깅은 주어진 예측모형의 평균예측모형을 구하고 분산을 줄여 예측력을 높일 수 있다는 장점이 있다.



< 그림 5 부스팅 기법 >

부스팅(boosting)은 예측력이 약한 예측모형을 결합해 예측력이 높은 예측모형을 만드는 기법이다. 배깅과 과정은 유사하지만 부트스트랩 표본을 구성하는 재표본 과정에서 각 자료에 같은 확률을 부여하지 않고 분류가 잘못된 데이터에 큰 가중치를 부여해 표본을 추출한다. 부스팅에서는 부트스트랩 표본을 추출해 분류기를 만들고 분류결과를 활용해 각 데이터가 추출될 확률을 조정해 다음 부트스트랩 표본을 추출하는 과정을 반복한다.



<그림 6 랜덤 포레스트 기법>

랜덤 포레스트(random forest)은 다수의 의사결정나무모델에 의한 예측을 종합하는 앙상블 방식이다. 배깅과 부스팅보다 더 많은 무작위성을 부여해 의사결정나무모델의 분산을 줄일 수 있다. 결과에 대한 해석은 어렵지만 예측력이 높다는 장점이 있다.

랜덤 포레스트는 다음과 같은 과정으로 진행된다.

- 1) 부트스트랩 기법을 이용해 표본을 추출한다.
- 2) 데이터를 생성하고 생성한다.
- 3) train 데이터로 의사결정나무모델을 구축한다.
- 4) 최종적으로 예측을 종합한 결과를 도출한다.


4.기대효과

코로나 사태로 혼란스러운 경제상황 속에서 빠르고 보다 정확하게 기업의 부도예측을 가능케 해 피해를 최소화할 수 있다. 재무상태와 시장정보를 통해 기업상태를 파악하는 것은 앞서 언급했듯이 많은 단점을 가지고 있다. 빠르게 정보를 파악하는 것이 불가능하며 비상장기업일 경우 데이터를 확보할 수 없다. 하지만 뉴스 기사 분석을 통해서 상장기업 뿐만 아니라 비상장기업의 정보를 얻을 수 있으며 기업의 부도 예측이 가능하다.

뉴스 기사를 다양하게 활용하는 모델의 연구를 기대할 수 있다. 본 조의 프로젝트를 통해 뉴스 기사는 단순 정보 전달만의 역할이 아닌 예측과 분석을 위한 데이터로 활용가능하다는 것을 알 수 있었다. 따라서 이후의 다양한 프로젝트에서 뉴스 기사 데이터가 활용되기를 기대할 수 있다.

5.주요성과

본 조의 조원들은 이전에도 다양한 프로젝트 경험이 있지만 이번 프로젝트에서는 여러가지 새로운 시도를 해보았다. 특히 데이터 수집과 텍스트 분석에 가장 많은 시간과 노력을 할애했다. 16년간의 인터넷 뉴스 기사를 수집하는 작업을 통해 크롤링에 대한 지식과 실습 경험을 얻을 수 있었다. 텍스트 마이닝이라는 분야는 조원들 모두 이번 프로젝트에서 처음 시도해보았다. 익숙하지 않아 실제 데이터에 적용하는 것에 어려움을 느꼈지만 다양한 자료를 활용하고 여러 번의 시도로 좋은 결과를 낼 수 있었다.

5.산학협력	<p>박지원 : 제안서 작성, 부도기업 데이터 수집, 데이터 전처리 진행, 텍스트 데이터 분석, 보고서 작성</p> <p>윤수연 : 제안서 작성, 정상기업 데이터 수집, 데이터 전처리 진행, 보고서 작성, 발표자료 제작</p> <p>이윤정 : 제안서 작성, 부도기업 데이터 수집, 데이터 전처리 진행, 텍스트 데이터 분석, 보고서 작성</p> <p>최솔 : 제안서 작성, 정상기업 데이터 수집, 데이터 전처리 진행, 보고서 작성, 발표자료 제작</p>
6.참고문헌	<p>최정원·한호선·이미영·안준모, 텍스트마이닝 방법론을 활용한 기업 부도 예측 연구, 2015</p> <p>최정원, 인공지능을 이용한 뉴스 정보 기반의 기업 부도예측 연구, 2019</p> <p>이성직, 김한준, TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출 기법, 2009</p> <p>심현우, 텍스트 마이닝을 활용한 개인정보유출 보고서의 군집 분석, 2019</p>
7.R&D성과	<p>해당사항 없음</p>
8.첨부	

1) 이기만. "부실 예방을 위한 기업 부실 조기진단 방법." 금융 728 (2014): 40-41.