



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

박사학위 논문

비정형 데이터를 활용한 머신러닝
기반 기업신용평가모형에 관한 연구

A Study on Machine
Learning-based Corporate Credit
Rating Model Using Unstructured
Data

2022년 12월

승실대학교 대학원

IT정책경영학과

김 용 환

박사학위 논문

비정형 데이터를 활용한 머신러닝
기반 기업신용평가모형에 관한 연구

A Study on Machine
Learning-based Corporate Credit
Rating Model Using Unstructured
Data

2022년 12월

승실대학교 대학원

IT정책경영학과

김 용 환

박사학위 논문

비정형 데이터를 활용한 머신러닝
기반 기업신용평가모형에 관한 연구

지도교수 김 광 용

이 논문을 박사학위 논문으로 제출함

2022년 12월

숭실대학교 대학원

IT정책경영학과

김 용 환

김 용 환 의 박 사 학 위 논 문 을 인 준 함

심 사 위 원 장 전 삼 현 인

심 사 위 원 최 정 일 인

심 사 위 원 박 중 우 인

심 사 위 원 지 용 득 인

심 사 위 원 김 광 용 인

2022년 12월

승실대학교 대학원

목 차

국문초록	viii
영문초록	xi
제 1 장 서론	1
1.1 연구의 배경	1
1.2 연구의 목적	2
1.3 연구방법	3
1.4 논문의 구성	4
제 2 장 이론적 배경 및 선행 연구	5
2.1 국내 기업신용평가모형 현황	5
2.2 기업 부도 예측 선행 연구	8
2.2.1 전통적인 기업 부도 예측 모형	10
2.2.2 최신 기업 부도 예측 모형	12
2.3 비정형 데이터 선행 연구	14
2.4 통계 모형 관련 이론	18
2.4.1 로지스틱 회귀 모형	18
2.5 머신러닝 모형 관련 이론	19
2.5.1 랜덤 포레스트 모형	19
2.5.2 그래디언트 부스팅 모형	21
2.5.3 심층 신경망 모형	22
2.6 비정형 데이터의 계량화 이론	24

2.6.1 의미 기반 방법	25
2.6.2 빈도 기반 방법	28
2.6.3 뉴스 기사 비율 변수	29
 제 3 장 기업신용평가 연구모형	30
3.1 개요	30
3.2 실험설계	30
3.2.1 개요	30
3.2.2 실험절차	31
3.3 데이터 정의	32
3.3.1 대상기업의 정의	32
3.3.2 부도의 정의	34
3.3.3 실적 관찰 기간 및 부도 예측 기간	36
3.3.4 재무 데이터의 정의	37
3.3.5 텍스트 데이터의 정의	38
3.4 데이터 수집과 전처리	39
3.4.1 재무 데이터	40
3.4.2 텍스트 데이터	44
3.5 연구데이터	46
3.5.1 연구데이터의 생성	46
3.5.2 연구데이터의 표본추출 및 분할	46
3.6 연구모형의 구성	48
3.6.1 로지스틱 회귀 모형	48
3.6.2 랜덤 포레스트 모형	49
3.6.3 그래디언트 부스팅 모형	49

3.6.4 심층 신경망 모형	50
3.7 연구모형의 성능 측정	50
3.7.1 연속형 모형의 평가	51
3.7.2 이진 분류 모형의 평가	54
제 4 장 실험 및 분석 결과	56
4.1 실험 환경 및 도구	56
4.2 개발 모집단	57
4.2.1 재무정보를 활용한 분석대상	57
4.2.2 뉴스 데이터와 결합	58
4.2.3 표본추출 및 데이터 분할	60
4.3 재무비율 평가항목	61
4.3.1 재무비율 전처리	61
4.3.2 단변량 분석 및 상관 분석을 통한 최종 변수 선별	64
4.4 뉴스 평가항목	66
4.4.1 뉴스 데이터 전처리	66
4.4.2 뉴스 데이터 키워드 선별	67
4.4.3 뉴스 평가항목 개발	73
4.5 연구모형 실험 결과	74
4.5.1 로지스틱 회귀 모형	74
4.5.2 랜덤 포레스트 모형	76
4.5.3 그래디언트 부스팅 모형	78
4.5.4 심층 신경망 모형	80
4.5.5 각 모형의 성능 비교	82

제 5 장 결론 및 향후 연구 방향	84
5.1 결론	84
5.2 연구의 의의	85
5.3 연구의 한계	86
5.4 향후 연구 방향	87
참고문헌	89
부 록	97

표 목 차

[표 2-1] 기업신용등급의 활용 현황	5
[표 2-2] 유형별 데이터의 특징 및 대표 데이터	15
[표 2-3] 자연어의 특성	16
[표 2-4] 랜덤 포레스트 모형 수행 단계	20
[표 2-5] CBOW와 Skip-gram의 단계별 학습 방법	26
[표 3-1] 업종별 대상기업 현황	34
[표 3-2] 부도의 정의	35
[표 3-3] 재무비율의 범주별 현황	40
[표 3-4] 단변량 분석 기준	43
[표 3-5] 연구데이터의 분할	47
[표 3-6] 분할 데이터의 역할	47
[표 3-7] 로지스틱 회귀 모형 변수 선택	48
[표 3-8] 혼동행렬	55
[표 4-1] 실험 환경 및 사용 도구	57
[표 4-2] 업종별 부도율 현황	58
[표 4-3] 연도별 부도율 현황	58
[표 4-4] 기업명의 글자 수 구성 현황	59
[표 4-5] 연도별 뉴스 데이터 현황	59
[표 4-6] 3글자 이상 기업의 뉴스 데이터 결합 현황	60
[표 4-7] 표본추출 및 데이터 분할 결과	60
[표 4-8] 변환 재무비율 변수 선택과정	64
[표 4-9] 최종 변환 재무비율의 기술 통계량	64
[표 4-10] 뉴스 데이터 전처리 예시	67

[표 4-11] 변별력 구간별 명사 현황	68
[표 4-12] 부도 관련 키워드	68
[표 4-13] 기업 성장 관련 키워드	70
[표 4-14] 주요 기업활동 영역별 키워드 분류 현황	71
[표 4-15] 키워드 분류 영역별 주요 뉴스	71
[표 4-16] Word2Vec을 활용한 키워드 추출 결과	73
[표 4-17] 뉴스 변수의 변별력	74
[표 4-18] 로지스틱 회귀 모형 구성	75
[표 4-19] 로지스틱 회귀 모형 성능 (검증 데이터)	76
[표 4-20] 로지스틱 회귀 모형 성능 (테스트 데이터)	76
[표 4-21] 랜덤 포레스트 모형 하이퍼 파라미터	77
[표 4-22] 랜덤 포레스트 모형 성능 (검증 데이터)	77
[표 4-23] 랜덤 포레스트 모형 성능 (테스트 데이터)	78
[표 4-24] 그래디언트 부스팅 모형 하이퍼 파라미터	79
[표 4-25] 그래디언트 부스팅 모형 성능 (검증 데이터)	79
[표 4-26] 그래디언트 부스팅 모형 성능 (테스트 데이터)	79
[표 4-27] 심층 신경망 모형 구조	80
[표 4-28] 심층 신경망 모형 하이퍼 파라미터	81
[표 4-29] 심층 신경망 모형 성능 (검증 데이터)	81
[표 4-30] 심층 신경망 모형 성능 (테스트 데이터)	81
[표 4-31] 각 모형별 성능 비교 (테스트 데이터)	82

그 립 목 차

[그림 2-1] 기업신용평가의 세부 구성	6
[그림 2-2] 신용등급 부여 및 검토 절차	8
[그림 2-3] 기업신용평가모형의 변화	9
[그림 2-4] 랜덤 포레스트 모형의 구조	20
[그림 2-5] 그래디언트 부스팅 모형의 구조	22
[그림 2-6] 심층 신경망 모형의 구조	24
[그림 2-7] Word2Vec의 구조	27
[그림 3-1] 실험 모형	31
[그림 3-2] 성능 비교를 위한 실험설계	32
[그림 3-3] 실적 관찰 기간 및 부도 예측 기간	37
[그림 3-4] 재무비율 전처리 절차	41
[그림 3-5] 상관 분석을 통한 최종 변수 선별 과정	44
[그림 3-6] 뉴스 데이터의 전처리 및 키워드 산출과정	45
[그림 3-7] ROC 곡선과 AUROC	52
[그림 3-8] K-S 통계량 Curve	53
[그림 4-1] 주요 재무비율의 원 재무비율과 변환 재무비율의 선형성 ...	62
[그림 4-2] 뉴스 변수와 부도의 선형 관계	74

국문초록

비정형 데이터를 활용한 머신러닝 기반 기업신용평가모형에 관한 연구

김 용 환

IT정책경영학과

승실대학교 대학원

기업의 부도를 사전에 예측하고 우량기업과 불량기업을 구분하는 것은 금융시장에 있어서 매우 중요한 문제이다. 기업신용평가모형의 부도 예측 성능을 높이기 위하여 다양한 시도 및 연구가 수행되어 왔다.

기업의 신용평가는 평가정보로써 재무제표를 활용하고 있으나, 기업의 경영위험 및 영업위험에 대한 다양한 분석이 필요하다. 본 연구는 기존 연구를 바탕으로 재무 데이터와 기업의 경영위험 및 영업위험을 반영하기 위해 비정형 데이터인 뉴스 데이터를 결합한 개선된 모형을 제시하고자 하였다.

재무 데이터는 변환 재무비율을 통해 재무비율의 비단조성을 해결하였으며, 변환 재무비율에 대해서 단변량 분석 및 상관 분석을 통해 유의한 39개 변수를 선별하였다.

비정형 데이터인 뉴스 데이터는 뉴스 제목을 활용하여 기업명이 존재하는 뉴스만을 발췌하였으며, 해당 뉴스의 내용을 통해서 기업의 영업위험과 영업성과를 예측할 수 있는 변수를 도출하였다. 형태소 분석을 통

하여 뉴스 제목에서 명사만을 발췌하였으며, 각 명사를 변수로 하여 뉴스 기사 비율 변수를 산출하여 단변량 분석을 실시하였다. 단변량 분석을 통해 부도 관련 키워드 35개와 기업 성장 관련 키워드 10개를 추출하였으며, 이를 통해 부도 관련 뉴스 기사 비율 변수와 기업 성장 관련 뉴스 기사 비율 변수를 도출하였다.

재무비율 변수와 뉴스 기사 비율 변수를 활용하여, 최적의 로지스틱 회귀(Logistic Regression) 모형, 랜덤 포레스트(Random Forest) 모형, 그래디언트 부스팅(Gradient Boosting) 모형 그리고 심층 신경망(Deep Neural Network) 모형을 도출하고 그 성능을 점검하였다.

실험 결과, 재무비율 변수들의 평균 정확도 비율(AR; Accuracy Ratio) 통계량은 약 49% 수준임에 반하여, 부도 관련 뉴스 기사 비율 변수의 AR 통계량은 4.6%, 기업 성장 관련 뉴스 기사 비율 변수의 AR 통계량은 21.5%로 뉴스 기사 비율 변수들의 변별력이 현저히 낮았으며, 이에 따라 전통적인 통계 모형인 로지스틱 회귀 모형에서는 뉴스 기사 비율 변수가 모형 변수에 선택되지 않았다.

그러나 머신러닝 모형에서는 재무 데이터만 활용했을 경우보다, 재무 데이터와 뉴스 데이터를 같이 활용할 경우 AR 통계량이 개선되는 결과를 확인하였다. 재무와 뉴스 데이터가 활용된 모형의 AR 통계량을 측정한 결과 랜덤 포레스트 모형은 78.97%로 재무 데이터만 활용했을 경우보다 1.45% 개선되었으며, 그래디언트 부스팅 모형은 78.05%로 재무 데이터만 활용했을 경우보다 4.41%로 개선되었으며, 심층 신경망 모형은 66.25%로 재무 데이터만 활용했을 경우보다 1.61% 개선된 결과를 보였다. 모형별로는 랜덤 포레스트 모형이 가장 우수한 성능을 보임을 확인하였다.

본 연구를 통해서 뉴스 데이터는 전통적인 통계 모형에서는 활용에 한

계가 있지만, 머신러닝 모형에서는 변별력의 개선 효과가 있었으며, 이를 통해 기업신용평가모형에 있어서 비정형 데이터가 의미 있는 평가정보로 활용될 수 있다는 가능성을 확인하였다.

ABSTRACT

A Study on Machine Learning-based Corporate Credit Rating Model Using Unstructured Data

Kim, Yong Hwan

Department of IT Policy and Management

Graduate School of Soongsil University

It is a very important issue in the financial market to predict the bankruptcy of a company in advance and to distinguish between a good company and a bad company. Various trials and studies have been conducted to improve the default predictive performance of the corporate credit rating model.

A company's credit evaluation uses financial statements as evaluation information, but various analyzes of the company's management and sales risk are required. Based on previous research, this study tried to present an improved model that combines news data which is unstructured data for reflecting management and sales risk, and financial data.

For financial data, non-monotonicity of financial ratios was resolved

through transformed financial ratios, and 39 significant variables were selected through univariate analysis and correlation analysis for transformed financial ratios.

For news data which is unstructured data, only news with a company name was extracted using the news title, and variables that could predict the business risk and performance of a company were derived from the contents of the news. Only nouns were extracted from news titles through morpheme analysis, and univariate analysis was performed by calculating news article ratio variables using each noun as a variable. Through univariate analysis, 35 keywords related to bankruptcy and 10 keywords related to corporate growth were extracted.

Using the financial ratio variable and the news article ratio variable, the optimal logistic regression model, random forest model, gradient boosting model and deep neural network model were derived and their performance was compared.

As a result of the experiment, the average accuracy ratio (AR) statistic of the financial ratio variables was about 49%, whereas the AR statistic of the ratio of news articles related to bankruptcy was 4.6% and the AR statistic of the ratio of news articles related to corporate growth was 21.5%. Therefore, the news article ratio variable was not selected as a model variable in the logistic regression model, which is a traditional statistical model.

However, in the machine learning model, AR statistics were improved when financial data and news data were used together,

compared to when only financial data was used. In the machine learning model, the AR statistics of the model using financial and news data were measured. As a result, the random forest model was 78.97%, which was 1.45% improved compared to the case where only financial data was used. The gradient boosting model's AR statistics was 78.05%, with was 4.41% improved compared to the case where only financial data was used. The deep neural network model's AR statistics was 66.25%, with was 1.61% improved compared to the case where only financial data was used. By model, it was confirmed that the random forest model showed the best performance.

Through this study, news data has limitations in utilization in traditional statistical models, but it has an effect of improving discrimination in machine learning models, and through this, it is possible to use unstructured data as meaningful evaluation information in corporate credit rating models confirmed.

제 1 장 서론

1.1 연구의 배경

금융기관들에게 있어서 보유한 자산에 대한 신용리스크를 관리하는 것은 매우 중요한 의사 결정 문제이다. 기업은 자산건정성의 관리를 위하여 부도에 관련된 많은 징후들을 정확하게 예측하기 위하여 노력하고 있으며, 이에 부도 예측에 대한 연구는 지속적으로 발전되어 왔다. 기업의 대내외적인 경영 환경도 계속 변화함에 따라 기업의 부도는 다양한 사전적인 흔적을 남기고 있으며, 이에 정교한 기업 부도 예측 방법에 대해 정책적인 규제와 함께 많은 방법론이 연구되고 있다.

금융감독원에서는 2004년 「신BIS 자기자본비율산출기준(안)」을 제시하였으며(금융감독원, 2004), 2005년 「신용리스크 내부등급법 기본 세부 지침(안)」을 통해 신BIS 자기자본비율산출기준(안)의 세부 기준을 보다 구체화하고 명확화하였다(금융감독원, 2005). 현재 운영되고 있는 자기자본비율 산출과 관련한 감독당국의 승인 체계 및 신용평가시스템의 설계 및 활용에 대한 세부적인 내용이 해당 지침안을 통해서 구체화되었다.

이에 기업의 신용평가모형은 감독기관에서 정한 바에 따라 체계적으로 발전하였으며, 현재는 재무정보를 통한 재무평가, 기업의 전망 및 영업환경에 대한 비재무평가, 대표자의 신용평가 등으로 구분되어 정교하게 운영되고 있다.

뉴스 기사 및 SNS는 텍스트 형식으로 저장되는 대표적인 비정형 데이터로써, 사회, 문화 및 경제 전반에 걸친 사회적인 특성을 잘 반영하고 있으며, 이를 통해 시대의 변화 및 기업의 경영현황, 경영전략 등에 대해 의미 있는 데이터를 추출할 수 있다(최요셉 & 최용석, 2014). 하지만 이러한 비정형 데이터는 기존의 정형 데이터 분석기법으로는 분석이 어려

운 텍스트라는 형식으로 저장되며, 언어의 특성상 데이터 간의 연결성을 고려해야 하는 어려움으로 인하여 정형화된 데이터인 기업의 재무 데이터 및 매출 데이터 등과 달리 분석에 상당한 제약이 있었다(장영재 외, 2020). 이렇게 비정형 데이터는 정형 데이터보다 더 의미있는 데이터를 추출할 수 있음에도 분석의 한계로 인하여 기업신용평가에서는 깊은 연구가 진행되지 못하였다.

이에 본 연구에서는 비정형 데이터를 기업신용평가에 활용하여 그 성과를 비교하고자 한다.

1.2 연구의 목적

금융감독원(2005)는 기업의 신용등급을 평가하기 위해서 양적인 정보와 질적인 정보를 축적하고 이를 반영해야 한다고 하였다. 양적인 정보는 자산, 부채, 매출액 규모 등 재무정보와 2개 이상 재무정보를 통해 생성되는 수익성 비율, 부채비율, 현금흐름 비율 등 재무비율을 포함한다. 질적인 정보는 경영진의 신뢰성 및 효율성, 산업의 전망 및 산업 내의 위치, 기업의 수익의 질을 포함한다.

기술의 발전에 따라 많은 양의 데이터가 생성 및 소비되고 있으며, 이를 빅데이터라고 한다. 대표적인 데이터로는 뉴스 데이터를 예로 들 수 있다. 뉴스 데이터는 정형화되어 있지 않지만 기업의 대내외 환경 등에 대한 많은 내용을 담고 있으며, 이 부분은 기업신용평가의 데이터 영역 중 질적인 정보의 영역에 해당된다. 기존 기업신용평가에 있어서 비정형 데이터인 뉴스 데이터는 그 활용에 대한 연구가 부족하였으며, 비정형 데이터를 활용함에 있어서 그 의미하는 바를 명확하게 구체화하고 있지 못하는 한계가 있었다.

이에 본 연구에서는 기업의 신용리스크를 측정하는 방법으로 재무 정

보와 대표적인 비정형 데이터인 뉴스 데이터를 활용하고, 기업신용평가에 주로 활용되는 통계 모형인 로지스틱 회귀(Logistic Regression) 모형 이외에, 다양한 머신러닝 알고리즘을 통해, 비정형 데이터인 뉴스 데이터의 가치를 실증분석하고, 다양한 머신러닝 알고리즘에서 뉴스 데이터가 어떠한 역할을 하는지에 대한 그 성과를 측정하고자 한다.

1.3 연구방법

전통적인 기업신용평가모형에서는 특정 회계기간의 성과를 나타내는 재무제표를 통해 미래의 부도 여부를 예측하고 있다. 기업의 재무제표는 수치적인 양적인 정보만을 포함하여, 기업의 전망 및 영업환경에 대한 질적인 정보는 포함하고 있지 못하다. 기업에 관련한 비정형 데이터인 뉴스 정보는 기업의 질적인 정보를 포함하고 있으며, 이를 통해 기업의 부도를 예측하는데 활용 할 수 있다는 가정하에서 본 연구를 시작하고자 한다. 기업에 대한 질적인 정보인 뉴스 정보는 기업의 전망, 경영자의 리스크 및 기업의 영업환경에 대한 다양한 정보를 담고 있으며, 이는 금융감독원(2008)이 제시하고 있는 신용등급산출 절차에서 비재무모형에서 활용하고자 하는 정보와 동일하다고 볼 수 있다.

각 기업과 관련한 뉴스 기사는 해당하는 기업이 처한 상황에 대한 우려 혹은 기대에 대하여 투자자들과 시장에 전하고자 하는 메시지로써, 특정한 키워드 또는 문장은 해당 기업의 미래의 성과를 내포하고 있다고 가정을 한다. 즉 재무제표에 표현된 재무적인 성과가 아닌 기사에 표현된 문장들과 그 문장으로 형성되는 맥락을 통해 기업의 숨겨져 있는 재무적 위험과 시장에서의 상황과 시장의 변화에 대응하는 기업의 대응 전략 및 시장의 평가가 내포되어 있을 것이라 볼 수 있다. 또한 동일한 내용에 대한 뉴스 기사일지라도, 그 기업이 보여주고 있는 성과의 다양성

속에서 해당 뉴스 기사가 차지하는 비중이 다르기 때문에 그 차이의 계량화를 통해 부도에 대한 가능성을 평가할 수 있다.

이에 재무제표는 NICE의 KIS-DATA에서 제공하는 108개 재무비율을 활용하며, 기업에 대한 시장의 부정적인 또는 긍정적인 평가는 뉴스 자체에 등장하는 단어의 통계를 활용하고자 한다.

또한, 다양한 알고리즘을 활용하여 재무 데이터만 활용했을 때와 재무 데이터와 뉴스 데이터를 같이 활용하였을 때의 성과를 측정하여 어떠한 차이가 있는지를 살펴보고자 한다.

1.4 논문의 구성

본 논문의 2장에서는 본 연구와 관련한 국내의 기업신용평가모형 현황 및 기업 부도 예측, 비정형 데이터의 텍스트 임베딩(Embedding) 관련 선행연구를 고찰한다. 또한 관련 이론을 살펴보는데 머신러닝 알고리즘을 살펴보고, 비정형 데이터를 계량화하기 위한 빈도 기반 텍스트 분석 기법과 관련한 이론들을 살펴본다. 3장에서는 본 연구의 실험 방법에 대해 설명하고 4장에서는 각 실험 방법에 따라 진행한 실험의 과정과 결과를 제시한다. 마지막으로 5장에서는 실험에 대한 결론과 한계 및 이후의 추가 연구 방향을 살펴보기로 한다.

제 2 장 이론적 배경 및 선행 연구

2.1 국내 기업신용평가모형 현황

기업신용평가는 기업의 채무 상환능력을 평가하여 금융회사의 여신 관리에 필요한 정보를 제공하기 위해 시작되었다. 기업신용평가는 기업이 채무상환을 불이행하여 발생할 수 있는 채권자의 손실을 최소화하기 위한 목적을 가지고 있으며, 평가결과는 채무불이행 가능성을 계량화하여 신용등급의 형태로 산출되어 제공된다(한국신용정보원, 2018).

기업신용평가는 금융회사와 개인, 기관투자자 등 다수의 이해관계자들이 관련되어 있으며, 이해관계자들 상호 간의 정보 비대칭을 해소해주는 역할을 하고 있다. 기업여신은 국가경제에 미치는 파급효과가 매우 큼으로, 기업신용평가모형에 대한 지속적인 수요가 존재한다. 기업신용평가는 다양한 목적으로 활용되며, 그 세부적인 사항은 [표 2-1]과 같다.

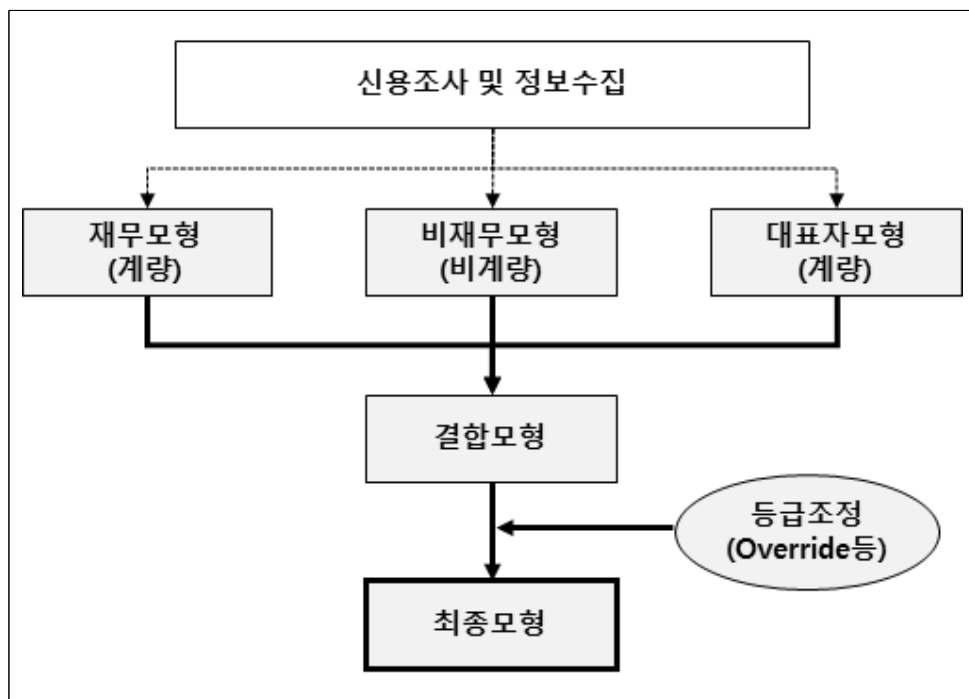
[표 2-1] 기업신용등급의 활용 현황

활용목적	세부내용
여신 의사결정	금융회사에서 여신실행 및 금리 등에 대한 기준으로 활용
위험가중자산 산정	금융회사의 BIS자기자본비율 (자기자본/위험가중자산)산출을 위한 위험가중자산 계산시 기업 신용등급에 따라 가중치가 차등 적용
대손충당금 산정	금융회사 여신자산에 대한 대손충당금 산정시 기업의 신용등급별로 대손충당금 적립 비율을 차등 적용
금융회사 영업 및 경영관리	영업전략 목표 고객 설정 및 거래처관리 등에 활용

출처: 한국신용정보원(2018)

기업신용평가는 Altman(1968)의 부도 예측 모형 연구를 시작으로 평가자 주관에 의한 오류 가능성을 최소화하고 평가과정을 표준화 및 객관화하기 위해 통계 방법론에 의한 신용평가모형 개발이 본격화되었으며, 바젤규약을 통해서 기업신용평가모형의 방법론은 체계화되었다.

현행 기업신용평가모형은 재무정보를 평가하는 재무모형, 재무 이외 영업환경 및 경영위험을 측정하는 비재무모형 그리고 대표자의 신용도를 측정하는 대표자모형을 하위모형으로 하며, 하위모형을 결합하여 최종 등급을 산출하게 된다. 재무모형과 대표자모형은 계량모형이며, 비재무모형은 비계량모형으로 심사자가 직접 신용평가를 통해 업체를 평가하게 되어 있다. 기업 신용평가모형의 세부 구성은 [그림 2-1]과 같다.



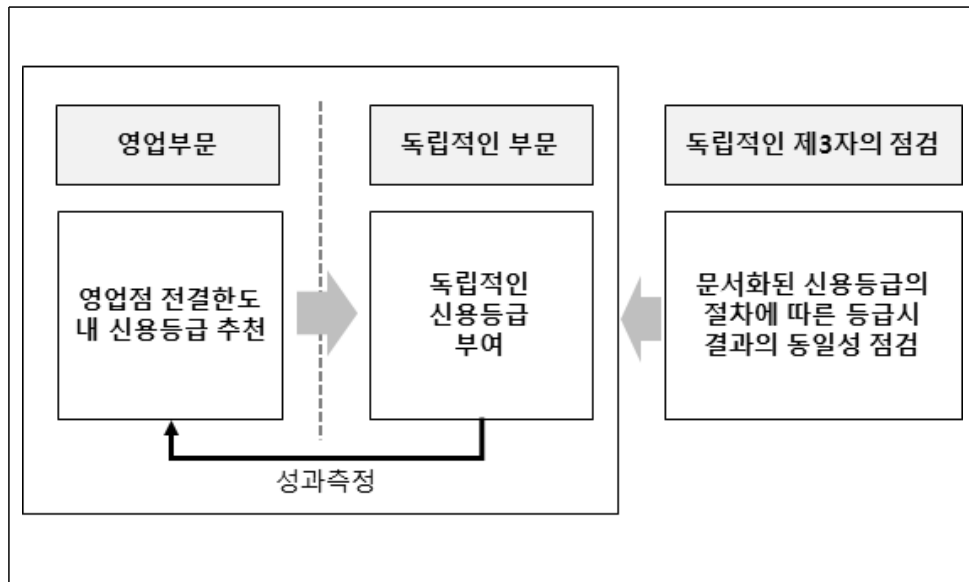
출처: 한국신용정보원(2018)

[그림 2-1] 기업신용평가의 세부 구성

금융감독원(2008)은 바젤 기준의 적용을 위한 세부지침에 대한 상세한 기준을 「바젤 II 下의 통합리스크관리 모범규준」을 통해 제시하였다. 바젤 II 下의 통합리스크관리 모범규준에는 기업 등 익스포저(Exposure), 소매 익스포저, 유동화 익스포저에 대한 신용리스크 관리체계, 운영리스크, 금리리스크 등 은행의 리스크 전반에 대한 세부내역이 기술되어 있으며, 이는 은행업 감독규정에 반영되어 운영되고 있다.

금융감독원(2008)에서는 기업신용평가에 활용되는 등급으로써 Master Scale Probability of Default(PD)를 제시하고 있다. Master Scale PD는 신용등급별 예상부도율의 수준을 사전에 정한 값이다. 차주들의 리스크 수준을 측정하여 동일한 리스크 수준의 차주들에 대해서 동일 등급으로 묶는 과정이 적용된다. 등급 계량화의 과정을 통해 Master Scale PD에 예상부도율이 할당되며, 이에 맞춰 신용등급이 산출되게 된다.

또한 신용평가 등급 부여의 무결성을 확보하기 위하여 신용등급 부여 업무의 독립성과 최소 1년 단위로 차주 및 여신에 대한 신용등급 재평가를 제시하고 있다. 신용등급의 독립성은 독립적인 부문에서 신용등급 부여가 진행될 것과 독립적인 제3자 점검의 2가지 사항을 요구하고 있다. 신용등급 부여 및 검토 절차는 [그림 2-2]와 같다.



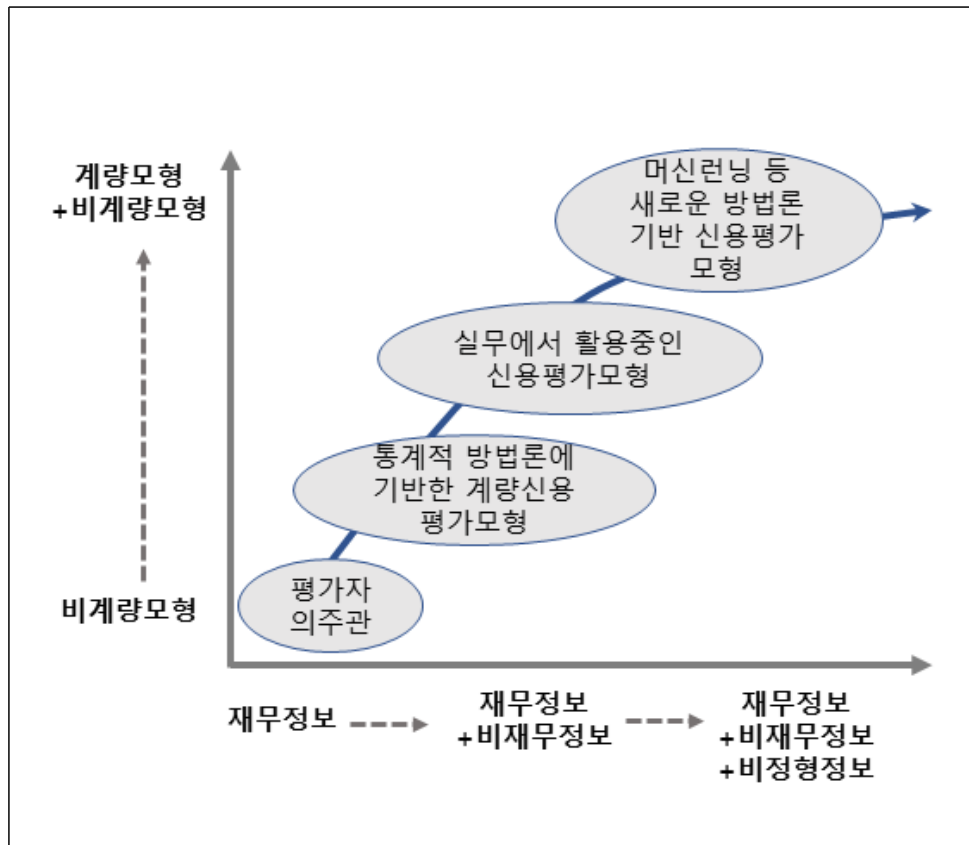
출처: 금융감독원(2008)

[그림 2-2] 신용등급 부여 및 검토 절차

2.2 기업 부도 예측 선행 연구

부도를 계량화하기 위한 기업신용평가의 활용 정보 측면에서는 재무 데이터만을 활용하였으나, 이후 재무 데이터와 주가 및 경기 상황 등의 비재무 데이터를 추가로 활용하였으며, 최근에는 뉴스 등의 비정형 데이터를 추가로 반영하고 있다.

모형적인 측면에서는 최초에는 경영위험, 산업위험 및 영업위험 등을 전문가의 지식과 경험을 통해 정성적으로 평가하는 전문가 판단모형인 비계량모형을 활용하였으나, 이후 다양한 통계 방법론 등을 활용한 계량모형을 비계량모형을 같이 혼합하여 활용하는 형태로 개선이 되었다. 이는 [그림 2-3]과 같다.



출처: 한국신용정보원(2018)

[그림 2-3] 기업신용평가모형의 변화

[그림 2-3]과 관련된 연구를 살펴보면 부도 예측에 이용된 정보의 원천으로 모형이 구분된다. 회계모형, 시장모형, 그리고 회계모형과 시장모형이 결합된 헤저드모형으로 분류되며, 최근에는 추가로 뉴스 기사 등 비정형 정보를 추가하려는 연구가 진행 중이다.

2.2.1 전통적인 기업 부도 예측 모형

부도 예측의 초기 연구는 Beaver(1966)가 통계 방법론을 사용한 이후로 과학적이고 통계적인 방법론에 대한 연구가 발전되면서 보다 정교한 예측을 위한 다양한 노력들이 진행되어왔다. 이러한 선행 연구들은 크게 3가지로 구분하여 살펴볼 수 있다.

첫번째는 초기 부도연구를 보완하고 개선하여 체계화를 했던 일련의 연구 흐름이다. Beaver(1966)는 단일변수를 활용하여 부도를 예측하는 단변량 분석을 통해서 부도를 예측하였다. Altman(1968)은 다양한 재무비율을 변수로 활용하여 부도를 예측하는 다변량 판별분석을 통해 부도 예측 모형을 연구하였다. 해당 연구에서는 운전자본비율, 이익잉여금 등 5개 재무비율 변수들이 활용되었으며, 이를 통해 부도기업을 구분하는 Z-Score 모형을 제시하였다. Altman et al.(1977)은 기존에 제시하였던 Z-Score 모형을 개선하여 ZETA 모형을 제안하였다. ZETA 모형에서는 선행연구인 Z-Score 모형에서 활용하였던 재무비율 변수들을 확대하여 7개의 재무비율 변수들을 활용하였으며, 재무비율 변수의 기준을 좀 더 명확하게 하였다. Ohlson(1980)은 앞서 Altman(1968)이 제시한 다변량판별분석을 활용한 Z-score 모형의 개선을 연구하였다. Ohlson(1980)은 Logit 모형을 통해 평가대상 기업의 부도를 예측하였으며, 기업의 규모를 비롯한 재무 구조, 성과 그리고 유동성의 4가지 요소를 연구에 활용하였다. 4가지 영역에 대해 9개의 회계정보를 활용하는 O-Score 모형을 제시하였으며, 정상기업 2,058개와 부도기업 105개에 대해서 O-Score가 높을수록 기업의 부도확률이 높은 것을 주장하였다.

두번째는 재무 정보와 시장 정보를 활용한 부도 예측에 대한 연구이다. Merton(1974)은 시장모형인 부도거리 모형을 제안하였다. 시장모형은 시장에서 거래되는 주가 등의 정보를 활용하여 부도를 예측하는 모형

으로, Merton(1974)는 만기시점에 부채금액과 기업의 자산가치 비교를 통해 부도 예측에 대한 연구를 하였다. Shumway(2001)는 기업의 부도 예측에 재무 정보와 시장 정보를 통합하여 모형의 예측력을 높이는 연구를 진행하였다. Shumway(2001)는 헤저드모형을 모형을 통해 기존의 재무정보와 주가정보를 결합하였으며, 각 정보를 개별적으로 활용했을 때보다 결합했을 때가 더 높은 예측력이 있음을 실증분석하였다. Campbell et al.(2008)은 시장총자산에 대한 정의 및 직전 10개월의 수익성 관련 변수의 가중평균 등을 통하여 새로운 변수를 개발하여 헤저드모형을 연구하였으며, 개선모형이 Shumway(2001)보다 예측력이 우수함을 실증분석하였다. 이인로 & 김동철(2015)는 회계모형, 시장모형, 그리고 헤저드모형에 대한 연구를 하였다. 회계모형에서는 판변분석모형과 로짓모형을 활용하였으며, 시장모형에서는 부도거리모형을 활용하였다. 그리고 헤저드모형은 기존 헤저드모형을 재추정하였다. 2001년부터 2013년까지 비금융 상장기업에 대한 부도 예측 연구를 하였으며, 이를 통해 헤저드모형이 회계모형 및 시장모형 대비 변별력이 우수함을 주장하였다.

세번째는 재무 정보를 바탕으로 기업의 특성을 반영한 부도 예측에 대한 연구이다. 박종원 & 안성만(2014)은 외감이상 기업을 대상으로 재무비율을 활용한 회계모형을 연구하였다. 2003년부터 2006년까지의 외감기업을 대상으로 다변량 로짓(Logit)분석을 활용하였으며, 재무비율 이외에 업종을 변수로 활용하여 건설업 더미, 제조업 더미의 변수를 모형에 활용하여 업종의 특성이 부도 예측에 유의함을 확인하였다. 권혁진(2017)은 K-IFRS도입에 따른 회계모형을 연구하였다. K-IFRS 도입에 따라 개별재무제표와 연결재무제표에서 각각 재무비율을 산출하였으며, 2010년부터 2014년까지 기업에 대해서 분석 결과를 바탕으로 연결재무제표를 활용한 회계모형을 제안하였다.

2.2.2 최신 기업 부도 예측 모형

최근 정보처리 기술의 발전에 따라, 데이터의 활용이 증가하였으며, 데이터 저장 및 처리에 대한 새로운 기술이 개발되었다. 이에 분석방법론적으로 많은 발전이 있었으며, 기업신용평가모형에도 새로운 연구들이 진행되었다. 최신 기업 부도 예측 연구는 재무 데이터를 중심으로 크게 3가지 영역으로 진행되었다.

첫 번째는 재무 데이터에 최신 알고리즘을 활용하여 그 예측력을 높이는 연구이다. 민성환(2014)은 부도 예측 모형으로 배깅(Bagging) 모형을 연구하였으며, 해당 모형은 사례 선택(Instance Selection)을 기반으로 하고 있다. 원 데이터에서 대표성있는 데이터를 선택하고 배깅을 통해 학습데이터에 영향을 주는 방법이 부도 예측에 유의함을 연구하였다. 총 8개 재무비율을 활용하여 서포트 벡터 머신 기반의 8개 모형을 제시하였으며, 이중 유전자 알고리즘을 이용한 사례 선택과 배깅을 연결한 모형인 Genetic Algorithm Instance Selection Bagging SVM (GAISBaggingSVM)의 성능이 가장 우수함을 확인하였다. 김성진 & 안현철(2016)은 기업신용등급 예측에 있어서 랜덤 포레스트(Random Forest) 모형을 활용하였다. 비교모형으로써 다중관별분석, 인공신경망, 서포트 벡터 머신을 활용하였으며, 이를 통해 랜덤 포레스트 모형의 성능이 우수함을 실증분석하였다. Le et al.(2019)는 재무제표를 활용한 기업의 부도 예측 모형을 연구하였으며, 분석 데이터로는 한국, 일본, 미국의 기업 파산 데이터를 활용하였다. 부도 예측을 위해 GPU 기반의 gDTC(Decision Tree Construction) Algorithm 구현을 사용하여 XGBS의 처리 시간을 가속화하는 gXGBS 알고리즘을 제안하였으며, 해당 알고리즘이 모형 변별력 지표인 Area Under ROC(AUROC) 및 기계학습의 처리 시간 측면에서 우수한 성능을 가지고 있음을 실증분석하였다.

두 번째는 재무 데이터에 비정형 데이터를 반영하여 부도 예측력을 높이는 연구이다. 비정형 데이터인 자연어를 처리하는 방법에 대해서 다양한 접근이 이루어졌다. Lu et al.(2013)은 금융 관련 뉴스를 추출하여, 텍스트 분석을 통해 말뭉치를 계량화하였으며, 이를 기존 재무 성과변수와 함께 활용하였다. 분석 기법은 로지스틱 회귀 모델을 활용하였으며, 재무 데이터와 금융 뉴스를 통합하여 재정적인 어려움을 겪는 기업을 선별하는 조기경보모형을 제안하였으며, 제안된 모형의 정확도 비율이 우수함을 주장하였다. 조남옥 & 신경식(2016)은 뉴스 데이터를 활용하여 감성 분석을 활용한 부도예측 모형을 제시하였다. 알고리즘은 로짓(logit) 분석, 인공신경망(Artificial Neural Network), 서포트 벡터 머신(Support Vector Machine)을 활용하였으며, 뉴스 데이터가 부도 예측의 유의함을 실증분석하였다. 김찬송(2018)은 뉴스를 유형별 분류, 수집 기간별 분류하여, 감성 분석 기반으로 부도 예측에 미치는 영향을 분석하였다. 뉴스 데이터는 Term Frequency - Inverse Document Frequency(TF-IDF) 분석을 통해 중요도가 높은 단어를 추출하였으며, 이를 뉴스의 상황에 맞춰 긍정과 부정으로 분류하여 빈도 기반의 뉴스 변수로 개발하였다. 빈도 기반의 뉴스 변수를 부도 예측 모형에 활용하여, 뉴스 데이터를 통한 감성 분석이 기업의 부도 예측에 유의한 성과를 보임을 실증분석하였으며, 부도 예측을 위한 뉴스 분류 방법을 제시하였다. 최정원(2019)는 뉴스 데이터, 시장 데이터, 재무 데이터를 활용하여 기업 부도 예측에 대한 연구를 하였다. 2010년부터 2016년까지 상장기업 전체에 대해서 뉴스 데이터를 수집하였으며, 워드투벡터(Word2Vec; Word to Vector)를 활용하여 부도연관 단어를 추출하였다. 추출된 부도연관 단어를 활용하여 뉴스의 빈도를 변수로 하는 새로운 변수를 개발하여 연구에 활용하였다. 알고리즘별로는 로지스틱 회귀 모형, 랜덤 포레스트 모형, 서포트 벡터

모형, 심층 신경망(Deep Neural Network) 모형을 활용하였으며, 각 알고리즘별로 뉴스 데이터, 시장 데이터, 재무 데이터를 결합한 모형의 성능 비교를 통해 뉴스 데이터가 기업의 부도 예측에 유의함을 제시하였다.

세 번째는 부도 데이터의 샘플링에 대한 연구이다. 김혜린(2020)은 부도 데이터의 특징을 정상 데이터 대비 발생 빈도가 낮은 불균형이라 정의하였으며, 적대적 생성 신경망 기반 오버샘플링 기법을 활용한 연구를 하였다. 적대적 생성 신경망을 비롯한 Random Over-Sampling(ROS) 및 Synthetic Minority Oversampling Technique(SMOTE)의 기법을 통한 오버샘플링된 데이터를 일반 선형 모형, 인공 신경망, 서포트 벡터 머신 모형을 적용하여 그 성능을 측정하였으며, 샘플링기법인 적대적 생성 신경망의 성능이 우수함을 실증분석하였다.

2.3 비정형 데이터 선행 연구

정형 데이터와 비정형 데이터를 구분하는 가장 큰 특징은 데이터 구조로, 정형 데이터는 DB 스키마(schema)의 표준 방식으로 구성이 되며, 비정형 데이터는 일반적인 파일(File) 시스템의 형태로 구성된다(조영임, 2013). 정형 데이터와 비정형 데이터의 특징을 가진 데이터를 반정형 데이터로 정의할 수 있으며, 각 유형별 데이터의 특징 및 대표적인 데이터는 [표 2-2]와 같다.

[표 2-2] 유형별 데이터의 특징 및 대표 데이터

데이터 유형	특징	대표 데이터
정형	고정된 필드에 저장된 데이터	관계형 데이터베이스, 스프레드시트 등
반정형	고정된 필드에 저장되어 있지 않지만 메타데이터나 스키마 등을 포함함	XML, HTML 등
비정형	고정된 필드에 저장되어 있지 않으며, 메타 데이터 및 스키마 등도 포함하고 있지 않음	텍스트 데이터, 음성 데이터, 이미지 데이터 등

출처: 이성훈 & 이동우(2013)

텍스트 데이터, 동영상 데이터, 음성 데이터 등 비정형 데이터를 수리적인 공간에 표현하여 정형 데이터로 나타내기 위한 다양한 연구가 진행되고 있다. 본 연구는 기업신용평가모형에 대한 연구이며, 기업에 대해 다양한 정보가 산출되는 대표적인 비정형 데이터인 텍스트 데이터에 관련된 선행 연구를 살펴보고자 한다.

텍스트 임베딩(Embedding)은 자연어인 텍스트를 수리적인 공간에 벡터로 표현하는 일련의 과정을 의미한다(이기창, 2019). 텍스트 임베딩은 최근 많은 관심을 받고 있으며, 다양한 분야에서 활용이 시도되고 있지만 많은 어려움이 있다. 임희석 & 고려대학교 자연어처리연구실(2019)은 텍스트 임베딩을 통해서 수치화가 되었더라도 하더라도 자연어라는 특성의 한계로 인하여 연구 및 응용에는 한계가 있다고 하였다. 텍스트 임베딩에는 자연어의 특성적인 한계로 인하여 대규모 컴퓨팅 자원이 필요하며, 텍스트 임베딩을 어렵게 하는 자연어의 특성은 언어의 중의성, 규칙의 예외, 언어의 유연성 및 확장성이 있다고 하였다. 텍스트 임베딩의 복잡성을 증가시키는 자연어의 특성은 아래 [표 2-3]과 같다.

[표 2-3] 자연어의 특성

구분	세부내용
언어의 중의성	단어 및 문장이 여러 가지 의미가 내포되어 있으며, 맥락에 따라서 의미하는 바가 달라짐
규칙의 예외	문장 안에서 단어와 형태소가 구성되는 방법을 규칙을 정하더라도 언어에 항상 예외가 존재함
언어의 유연성 및 확장성	유한한 단어와 소리를 조합하여 생성할 수 있는 문장이 무한함

이러한 자연어의 특성에 따른 한계에도 불구하고 많은 연구들이 진행되고 있으며, 다양한 분야에서 성과를 나타내고 있다.

Kraus & Feuerriegel(2017)은 공개된 재무 관련 뉴스를 통한 주가 예측에 대한 연구를 하였으며, 딥러닝 기법이 전통적인 기계학습보다 더 높은 정확도를 보여주고 있음을 실증분석하였다.

Huynh et al.(2017)은 온라인 금융 뉴스와 과거 주가 데이터를 활용하여 주가를 예측하는 모형을 연구하였다. Bidirectional Gated Recurrent Unit을 활용하였으며, 실험결과 S&P 500 지수 및 개별 주식의 예측율이 60%이상으로 우수한 성능을 보임을 실증분석하였다.

Liu et al.(2018)은 표현 학습을 위한 TransE모형과 Convolutional Neural Networks(CNN)을 통해 뉴스 기사의 특징을 추출하는 모형을 같이 활용하는 통합 모형을 제안하였으며, 추출된 정보를 활용하여 서포트 벡터 머신과 Long Short-Term Memory(LSTM)의 알고리즘을 통해 Apple의 주가 예측을 연구하였다. 해당 연구에서 제안된 통합 모형의 뉴스 분류의 정확도는 97.66%로 CNN을 통한 뉴스의 분류보다 더 높은 성능을 보임을 실증분석하였으며, 제안된 통합 모형으로 분류된 뉴스를 통한 주가 예측은 LSTM모형의 정확도가 55.44%로 서포트 벡터 머신보다 우수함을 실증분석하였다.

박대서 & 김화중(2018)는 통계 기반 키워드 추출 방식인 TF-IDF와 의미 기반 키워드 추출 방식인 Word2Vec을 결합하여 새로운 키워드 추출 방식을 제안하였다. TF-IDF 방법을 통해 뉴스 기사를 통계적 벡터로 변환하였으며, Word2Vec을 활용하여 뉴스 기사 내 특정단어와 그 외 단어들 간의 유사도 평균을 활용하여 벡터화하였다. 통계 벡터와 의미 벡터를 결합하여 키워드를 결합한 결합 벡터를 통해 키워드를 추출한 결과 키워드와 뉴스 기사의 일치도가 개선되었음을 실증분석하였다.

Mai et al.(2019)는 딥러닝을 통해 텍스트 데이터에서 특징을 추출하여 미국 상장 기업의 파산을 예측하는 모형을 연구하였다. 텍스트 데이터와 재무 데이터인 재무비율, 시장 기반 변수를 같이 활용한 연구모형이 파산 예측에 우수한 성능을 보임을 제시하였으며, 텍스트 데이터 분석에 있어서 평균 임베딩과 같이 단순한 모형이 신경망 모형보다 더 효과적임을 실증분석하였다.

조단비 외(2020)은 신문 뉴스 기사를 통해 정치적 성향의 편향성 분류에 대한 연구를 하였다. 정치적 성향과 밀접하게 관련이 있는 15개 키워드들에 대한 뉴스를 수집하였으며 형태소 분석기를 활용하여 임베딩을 하였다. 형태소 분석기는 Okt, Hannanum, Komoran을 활용하였으며, 이를 통해 문장 및 문서 임베딩으로 확장하였다. 학습데이터를 서포트 벡터 머신을 이용하여 학습한 결과 Okt 형태소 분석기를 활용하였을 때 테스트 데이터에서 높은 성능을 보임을 실증분석하였다.

2.4 통계 모형 관련 이론

2.4.1 로지스틱 회귀 모형

로지스틱 회귀(Logistic Regression) 모형은 종속 변수가 범주형일 때, 독립 변수의 선형 결합을 통해 종속 변수의 발생 가능성을 예측하는 통계 모형이다. 로지스틱 회귀 모형은 Cox(1958)에 의해 제안되었다. 이후 Cornfield et al.(1961) 및 Walker & Duncan(1967)를 통해 체계적으로 발전하였고, 통계학, 사회과학 등 다양한 분야에서 예측 또는 변수 간의 상관관계를 분석을 위한 목적으로 사용되고 있다.

로지스틱 회귀 모형은 독립변수와 종속변수의 선형 관계가 있음을 가정하고 있으며, 아래와 같은 선형수식으로 표현된다.

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n$$

수식에서 Y 는 종속변수, $X_1, X_2 \dots X_n$ 은 독립변수, $b_1, b_2 \dots b_n$ 은 회귀계수 그리고 b_0 은 상수항을 의미한다. 로지스틱 회귀 모형의 회귀계수는 최대 우도법을 통해 추정되며, 변수의 선택에는 전진 선택법, 후진 선택법, 단계적 선택법이 활용된다(홍세희, 2005).

로지스틱 회귀 모형은 다양한 분야에 활용되고 있으며, 특히 신용평가 분야에서 가장 중요한 모형으로 활용되고 있다. 로지스틱 회귀 모형은 박종원 & 안성만(2014), 이인로 & 김동철(2015) 및 서정구 & 김확열(2018) 등 기업 부도 예측 및 재무건정성 연구에 활용이 되었으며, 김중윤(2019)의 통신 빅데이터를 활용한 개인신용평가모형에 대한 연구에 활용되었다.

2.5 머신러닝 모형 관련 이론

2.5.1 랜덤 포레스트 모형

랜덤 포레스트(Random Forest) 모형은 Breiman(2001)에 의해 제안되었으며, 배깅(Bagging) 방식을 활용한 분류 모형이다. 배깅은 Bootstrap Aggregation의 줄임말이며 부트스트랩(Bootstrap) 샘플링 방식의 통한 예측 결과의 집합(Aggregation)을 의미한다.

랜덤 포레스트는 하나의 의사결정 나무를 확대하여, 여러 개의 의사결정 나무를 만들고, 각 의사결정 나무들을 연결한 모형이다. 전체 데이터를 부트스트랩 샘플링을 통해 여러 개의 데이터로 분할하고 개별적으로 학습하고 이를 결합하여 결과를 예측한다(Breiman, 2001).

한은정(2005)는 랜덤 포레스트 모형의 장점으로 무작위로 복원 추출된 데이터를 모형에 활용함에 따라 이상치 등의 영향이 상대적으로 적다고 하였으며, Raschka & Mirjalili(2021)는 랜덤 포레스트 모형이 여러 개의 의사결정 나무의 평균을 활용하기 때문에, 개별 의사결정 나무는 분산이 높을 수 있지만, 결합 모형인 랜덤 포레스트 모형은 견고한 모형으로써 과대적합의 위험이 낮다고 하였다.

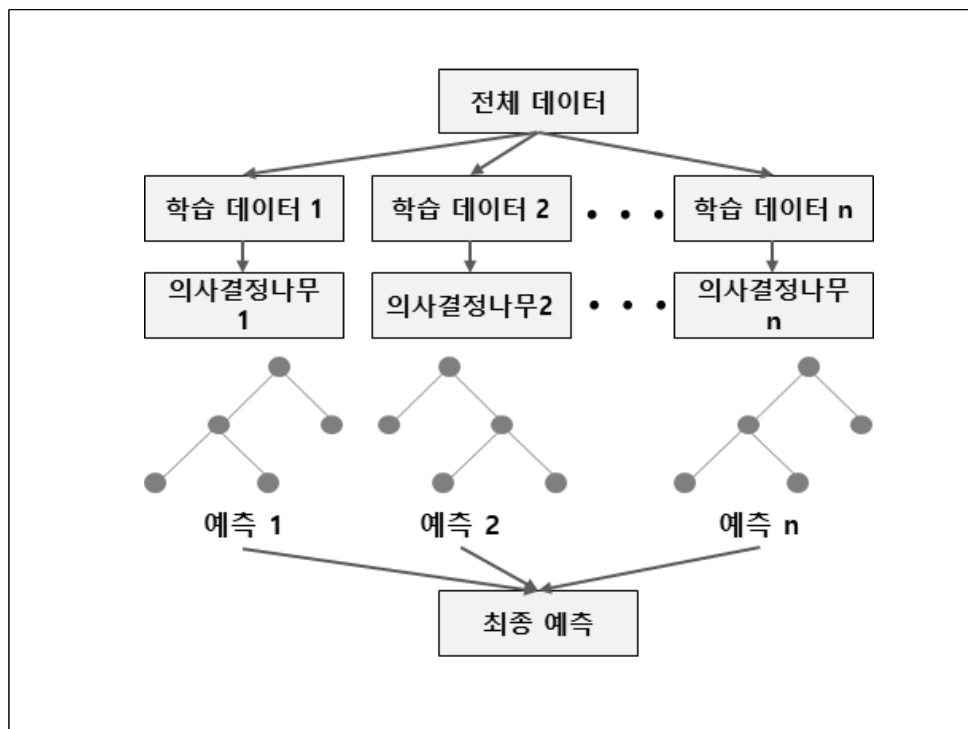
조용준(2018)은 랜덤 포레스트 모형의 제약사항으로 학습시간이 과다하게 소요되기 때문에 실시간 분석 등에 적용이 어려우며, 데이터의 관측치 및 변수가 적은 경우에는 모형 적합도를 높이는 데에 한계가 있다고 하였다.

랜덤 포레스트 모형은 4단계로 수행되며, [표 2-4]와 같으며, 모형의 구조는 [그림 2-4]와 같다.

[표 2-4] 랜덤 포레스트 모형 수행 단계

단계	세부내용
1단계	부스트트랩(Bootstrap) 샘플링 : 훈련 데이터에서 중복을 허용하는 랜덤한 n개의 샘플 선택
2단계	1단계를 통해 산출된 샘플을 통해 의사결정 나무 학습
3단계	1단계와 2단계를 반복 수행
4단계	각 의사결정 나무의 예측 결과들을 모아 다수결 투표를 진행하며, 이를 통해 최종 예측 결과를 할당

출처: Raschka & Mirjalili(2021)



출처: 김형수(2020)

[그림 2-4] 랜덤 포레스트 모형의 구조

랜덤 포레스트 모형은 다양한 분야에 활용되고 있다. 김성진 & 안현철

(2016), 조성빈(2020) 및 조경인 & 김영민(2021)은 기업부도 예측 연구에 랜덤 포레스트를 활용하였다. Zhang et al.(2018)는 랜덤 포레스트를 활용한 새로운 신용평가모형을 연구하였다. 권안나(2013)는 변수 선택에 대한 연구를 통해 랜덤 포레스트 모형이 다른 방법들에 비하여 잡음(Noise) 변수 제거에 효율적임을 실증분석하였다.

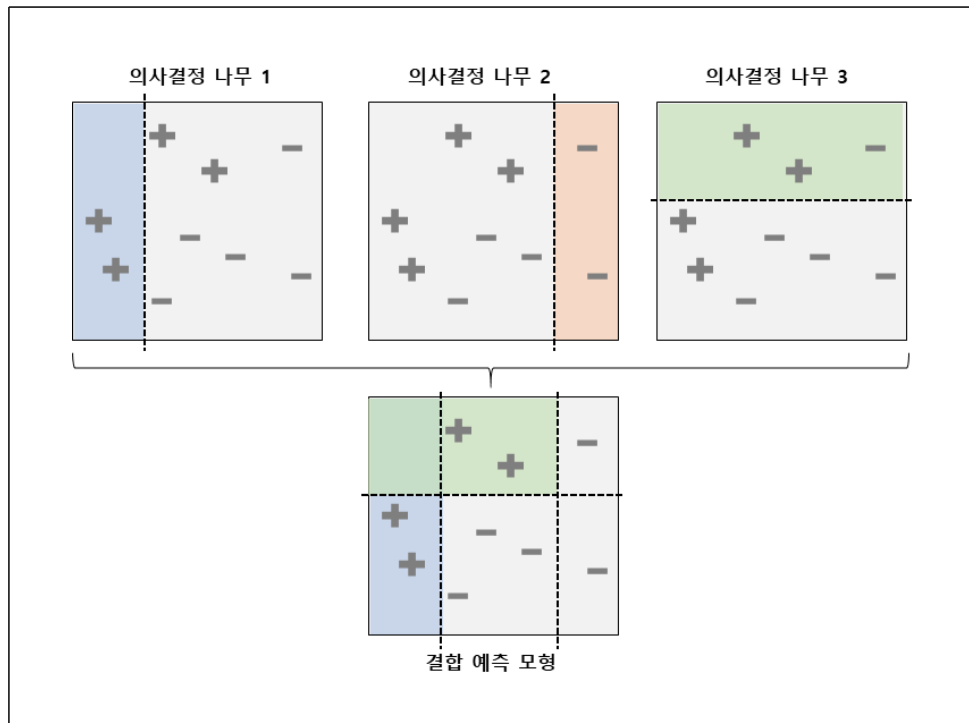
2.5.2 그래디언트 부스팅 모형

그래디언트 부스팅(Gradient Boosting)은 Breiman(1997)에 의해 제안되었으며, Friedman(2001)에 의하여 보다 체계적으로 발전되었다.

그래디언트 부스팅은 약한 학습기(weak learner)로 학습을 진행하고 이전 학습기의 모형 예측 오차를 보완하는 방법이다. 이전 모형의 오차는 손실함수로 표현하며, 경사하강법을 통해 손실함수를 최소화한다. 그래디언트 부스팅에서 약한 학습기로는 의사결정 나무를 활용한다(김형수, 2020). 그래디언트 부스팅의 작동원리는 첫번째 의사결정 나무가 도출되어, 실제 관측치와의 오차를 측정하게 되고, 두번째 의사결정 나무는 첫번째 모형의 오차를 종속변수로 하는 모형이 되며, 이러한 방법이 반복되어 예측치의 오차가 축소되게 된다(강성원 & 강희찬, 2020).

그래디언트 부스팅은 약한 학습기를 활용함에 따라 학습과 예측이 빠르다는 장점이 있으나, 의사결정 나무의 개수가 많아질 경우에는 과적합이 발생함에 따라 의사결정 나무의 개수 및 의사결정 나무의 최대 깊이 등 하이퍼 파라미터의 설정이 중요하다(김형수, 2020).

그래디언트 부스팅의 구조는 [그림 2-5]와 같다.



출처: 권철민(2021)

[그림 2-5] 그래디언트 부스팅 모형의 구조

Tian et al.(2020)는 그래디언트 부스팅 모형을 활용하여 신용위험평가 시스템에 대한 연구를 하였다. 그래디언트 부스팅 모형은 여러 가지 개선된 모형이 있으며, Chang et al.(2018)은 그래디언트 부스팅 계열의 방법인 eXtreme Gradient Boosting Tree를 활용하여 금융기관의 신용위험 평가모형을 연구하였으며, Taha & Malebary(2020)는 신용 카드 거래에서 사기를 탐지하기 위한 연구를 위하여 Light Gradient Boosting을 활용하였다.

2.5.3 심층 신경망 모형

인공 신경망(Artificial Neural Network)은 McCulloch & Pitts(1943)에

의하여 처음 소개가 되었다. 이후 인공 신경망은 심층 신경망(Deep Neural Network) 모형으로 발전하였다. 심층 신경망 모형은 인공 신경망 모형에서 여러 개의 은닉층을 추가하였으며, 더 많은 계층을 통해 스스로 학습을 진행하게 된다(Liu et al., 2017).

역전파 알고리즘(Back-Propagation)은 Werbos(1974)에 의하여 제안이 되었다. 역전파 알고리즘은 입력값에 대해서 순방향으로 출력의 예측값과 실측값의 오차를 계산하여, 이를 역방향으로 전파하여 오차를 줄이는 알고리즘으로 전파된 오차를 통해서 은닉층의 가중치를 변경하며, 이러한 과정을 전체 데이터에 대해서 분류가 정확히 이루어질 때까지 반복하게 된다(천인국, 2020). 역전파 알고리즘과 비선형 활성화함수가 심층 신경망 학습에 적용됨에 따라, 학습에 필요한 많은 계산량 및 선형적인 모형에 한정되었던 문제점이 개선되었다(이재성, 2016).

Hinton et al.(2006)은 심층 신경망 모형의 오차 역전파 오류 및 과잉 학습에 대한 대안으로 가중치 초기값들에 대한 초기화를 제안하였으며, 이를 통해 심층 신경망 학습에 대한 성과가 개선됨을 실증분석하였다(이재성, 2016).

심층 신경망은 우수한 성능을 가지고 있으며, 자연어 처리, 음성 및 이미지 분야 등 폭넓게 활용되고 있다.

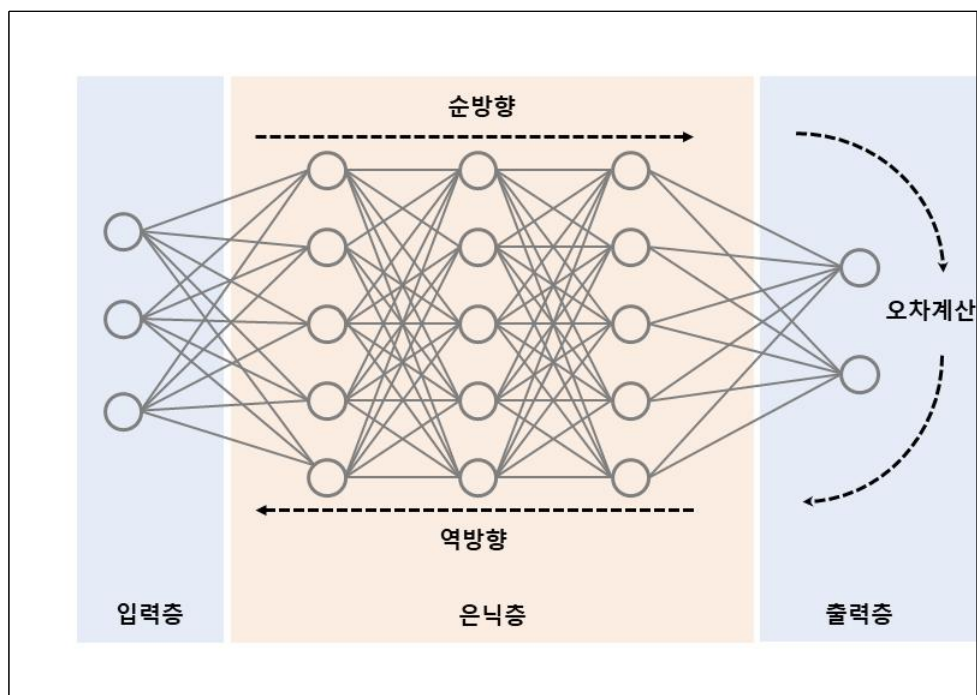
Salimov & 류재홍(2021)는 해상도가 낮은 영상 자료를 통해 얼굴의 표정 인식을 연구하였으며, 심층 합성곱 신경망 모형을 활용하였다. 조승현 외(2018)은 심층 신경망을 활용하여 이미지와 비디오의 압축 기술을 연구하였다.

지승은 & 김우일(2017)는 심층 신경망을 기반으로 하는 음성 인식 성능 지표를 연구하여 효과적인 음성 인식 평가 방안을 제시하였다.

홍동숙 외(2021)은 국내 제조업 개인사업자의 대출 및 연체 등 정보를

활용하여 심층 신경망을 활용한 연구모형을 제시하였다. Dharwadkar & Patil(2018)은 심층 신경망 모형을 활용하여 은행의 고객 유지 및 신용 위험 분석을 연구하였으며, 은행의 이익 제고를 위한 보다 효율적인 모형을 제안하였다.

심층 신경망의 구조는 [그림 2-6]과 같다.



출처: 천인국(2020)

[그림 2-6] 심층 신경망 모형의 구조

2.6 비정형 데이터의 계량화 이론

텍스트 정보는 비정형 데이터 중 가장 대표적인 정보이다. 텍스트 정보는 웹 페이지, 이메일 및 출판물 등 각종 문서를 통해서 확보할 수 있다. 최근 발전하고 있는 영상 인식 및 음성 인식 기술과의 결합을 통해 다양한 영역에서 정보의 확보가 가능하다. 빅데이터 분석 기법의 발전에

따라 텍스트 정보는 기존 연구모형에 활용되어 그 성능을 높이거나, 텍스트 분석 기법 자체만으로 충분한 하나의 연구모형으로 가치를 가질 수 있다.

인간의 언어인 자연어는 수학적으로 표현을 하거나 측정을 할 수 없는 데이터로써, 자연어 자체를 통계 모형 및 머신러닝 모형에 적용할 수 없다. 이에 자연어의 전처리를 통해 수치화하여 벡터의 공간에 표현하는 것을 임베딩이라고 한다. 임베딩의 개념은 기존에도 활용이 되어왔으나, Bengio et al.(2000)에 의해서 본격적으로 활용되었다(이기창, 2019).

임베딩의 가장 기본적인 방법은 단어를 표현하는 방법이다. 가장 대표적인 단어의 수치화 방법은 원-핫 인코딩(one-hot encoding)으로, 이 기법은 단어를 0과 1의 값만을 가지는 벡터로 표현하며, 1로 표현되는 단어는 인덱스의 역할을 하게 된다.

원-핫 인코딩은 단어의 의미와 특성이 고려되지 못하는 점과 분석하고자 하는 단어가 많아질수록 각 단어의 벡터의 크기가 증가하여 분석 효율이 저하되는 한계가 있다(전창욱 외, 2022).

본 연구를 위해서 텍스트 임베딩을 통해 이를 계량화하는 방법으로 의미에 기반한 방법과 빈도에 기반한 방법을 살펴보고자 한다.

2.6.1 의미 기반 방법

의미 기반 방법은 신경망 등의 모형을 활용하여, 문장 안에서 나올 수 있는 단어를 예측하고 이를 벡터로 만드는 기법이다(전창욱 외, 2022).

대표적인 예측 기반 방법에는 Word2Vec이 있다. Word2Vec은 사전학습 단계를 수행하지 않는 비지도 학습(Unsupervised Learning) 기반의 신경망 단어 임베딩(Embedding) 모델이다. Word2Vec은 단어와 단어 사이의 연관관계를 분석하여, 단어들 간의 관계를 계량적으로 산출하는 방

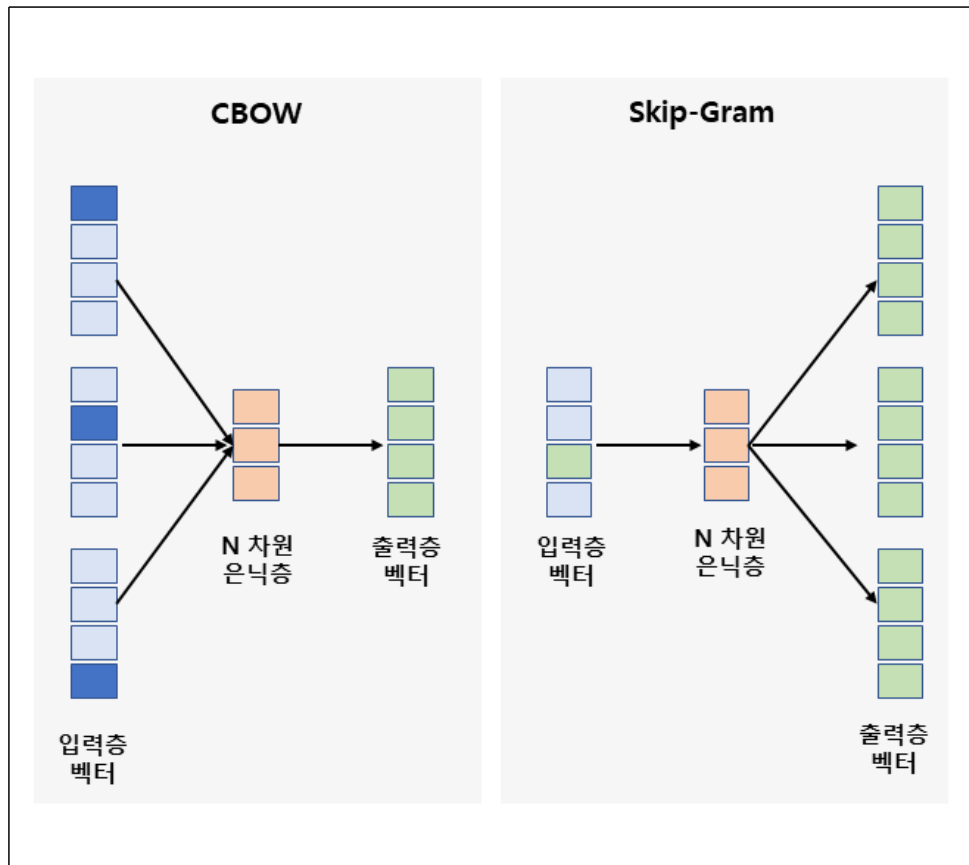
법이며, 단어의 순서를 고려하여, 앞의 단어와 뒤의 단어의 관계를 분석하여 그 거리를 벡터로 나타낸다(최정원 외, 2017).

Word2Vec은 Continuous Bag-Of-Word(CBOW) 그리고 Skip-gram의 2가지 모델로 구분된다. CBOW는 주변의 단어로부터 중심 단어가 나올 확률을 학습하고, Skip-gram은 중심 단어로부터 주변 단어가 나올 확률을 학습한다. CBOW와 Skip-gram의 단계별 학습 방법은 [표 2-5]와 같으며, 그 구조는 [그림 2-7]과 같다.

[표 2-5] CBOW와 Skip-gram의 단계별 학습 방법

단계	CBOW	Skip-gram
1단계	각 주변의 단어를 원-핫 벡터로 변환하고 이를 입력값으로 활용	하나의 단어를 원-핫 벡터로 변환하여 이를 입력값으로 사용
2단계	각 원-핫 벡터에 가중치 행렬을 곱하여 n 차원 벡터 생성	각 원-핫 벡터에 가중치 행렬을 곱하여 n 차원 벡터 생성
3단계	n차원 벡터의 값을 모두 합한 후에 개수로 나누어 평균 n 차원 벡터 생성	n 차원 벡터에 가중치 행렬을 곱하여 벡터 생성
4단계	n차원 벡터에 가중치 행렬을 곱하여 벡터 생성	만들어진 벡터와 예측하려는 주변 단어들의 원-핫 벡터와 비교하여 학습
5단계	만들어진 벡터와 예측하려는 단어의 원-핫 벡터와 비교하여 학습	-

출처: 전창욱 외(2022)



출처: 전창욱 외(2022)

[그림 2-7] Word2Vec의 구조

Word2Vec은 다양한 연구에 활용되고 있다. 최정원(2019)은 기업 관련 뉴스에서 부도 연관 단어를 Word2Vec을 통해 추출하여 기업 부도 예측 모형에 대한 연구를 하였으며, Jatnika et al.(2019)는 영문 Wikipedia를 활용하여 단어의 표현 기법에 따른 유사도에 대해 연구하였다.

2.6.2 빈도 기반 방법

단어를 표현하는 빈도 기반 방법은 문장 안에서 단어가 동시에 등장하는 빈도를 분석하는 방법이며, 이를 동시 출현(Co-occurrence)이라고 한다. 빈도 기반 방법은 기본적으로 동시 출현 횟수를 하나의 행렬로 나타내는 과정이다(전창욱 외, 2022).

대표적인 빈도 기반 방법으로는 TF(Term Frequency)가 있으며, 이는 이후 TF-IDF(Term Frequency-Inverse Document Frequency) 기법으로 개선되었다. TF-IDF는 단어의 빈도가 높다고 하더라도 단어만으로는 문서의 문맥을 예측하기 어렵다는 한계를 개선하기 위하여 제안된 방법이며, 수식은 아래와 같다.

$$TF-IDF(w) = TF(w) \times \log\left(\frac{N}{DF(w)}\right)$$

수식에서 TF(Term Frequency)는 단어가 특정 문서에서 발생 빈도, DF(Document Frequency)는 단어가 등장한 문서의 수, 그리고 IDF(Inverse Document Frequency)는 전체 문서의 수 N을 DF로 나눈 값에 log를 적용한 값이다.

TF는 빈도가 높은 단어가 문서 안에서 중요도가 높다는 것을 전제로 수식에서 활용되며, DF는 특정 단어가 등장하는 문서가 많을수록 해당 단어는 일반적인 단어라는 것을 가정하기 때문에, DF를 통해 생성되는 IDF는 그 값이 높을수록 특이한 단어라는 의미로 해석이 된다(이기창, 2019).

2.6.3 뉴스 기사 비율 변수

최정원(2019)은 기업에 부정적인 뉴스를 계량적인 변수로 산출하기 위하여 전체 기사 중 부도 관련 기사의 비중을 산출하여, 해당 비율이 높게 나타날 경우에는 부도의 징후가 있다고 판단하고 이를 모형의 평가항목으로 활용하였다.

최정원(2019)의 연구를 바탕으로 본 연구에 활용한 뉴스 기사 비율의 수식은 아래와 같으며, 부도 및 기업 성장 관련 뉴스 기사 비율 변수를 각각 설정하였다.

$$\text{부도 관련 뉴스 기사 비율} = \frac{\text{부도 관련 뉴스 기사 수}}{\text{총 뉴스 기사 수}}$$

$$\text{기업 성장 관련 뉴스 기사 비율} = \frac{\text{기업 성장 관련 뉴스 기사 수}}{\text{총 뉴스 기사 수}}$$

제 3 장 기업신용평가 연구모형

3.1 개요

본 연구는 기업의 부도 예측에 있어서 기존 재무정보 이외에 비정형 데이터인 뉴스 정보를 반영하여, 그 효용을 실증분석하고자 한다.

재무정보는 재무제표를 통해 산출되는 재무비율을 활용하였다. 기존 연구들은 선행연구들에서 활용된 재무비율을 기반으로 추가 정보를 반영하여 연구에 활용하였으나, 본 연구는 KIS-DATA를 통해서 가능한 많은 후보 재무비율 정보들을 확보하였으며, 극단치 처리, 재무비율의 선형성 확보를 위한 비율변환, 상관 분석을 통한 각 변수의 그룹화 및 대표 변수 추출 등의 절차를 통해 체계적으로 모형변수를 선별하여 연구에 활용하였다.

또한, 비정형 데이터인 뉴스 정보를 계량화함에 있어서 기존 연구들은 특정단어가 부도에 얼마나 연관이 있는지를 측정하고 이를 통해 부도와 양의 상관관계를 가지는 뉴스 변수를 산출하였으나, 본 연구는 뉴스에 등장한 단어를 하나의 변수로 보고 그 단어의 변별력 및 부도와의 방향성을 고려하여 부도와 양의 상관관계를 가지는 부도 관련 뉴스 변수와 음의 상관관계를 가지는 기업 성장 관련 뉴스 변수로 계량화하였으며, 이를 통해 다양한 측면에서 뉴스 정보의 유용성을 연구하고자 하였다.

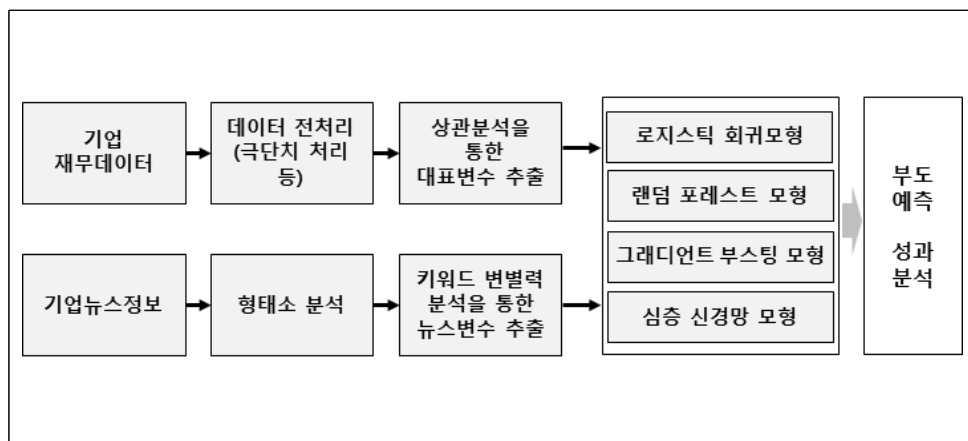
3.2 실험설계

3.2.1 개요

기업의 부도 예측 연구를 위하여 기업의 재무데이터와 뉴스 정보를 활용하였다. 기업의 재무데이터는 재무비율정보를 활용하였으며, 뉴스 정보

는 뉴스의 빈도분석을 통해 뉴스 변수로 계량화하여 분석하였다. 본 연구의 실험 모형은 [그림 3-1]에 나타내었다.

재무비율은 결측치 및 극단치 처리, 선형성을 위한 비율변환 등을 진행하여, 대표변수를 추출하였으며, 기업 뉴스 정보는 형태소 분석을 통해 키워드를 추출하고 해당 키워드의 변별력을 측정하여 뉴스 변수로 계량화하였다. 재무 및 뉴스 변수는 로지스틱 회귀 모형, 랜덤 포레스트 모형, 그래디언트 부스팅 모형 그리고 심층 신경망 모형에 활용하여 부도 예측의 성과를 측정하였다.



[그림 3-1] 실험 모형

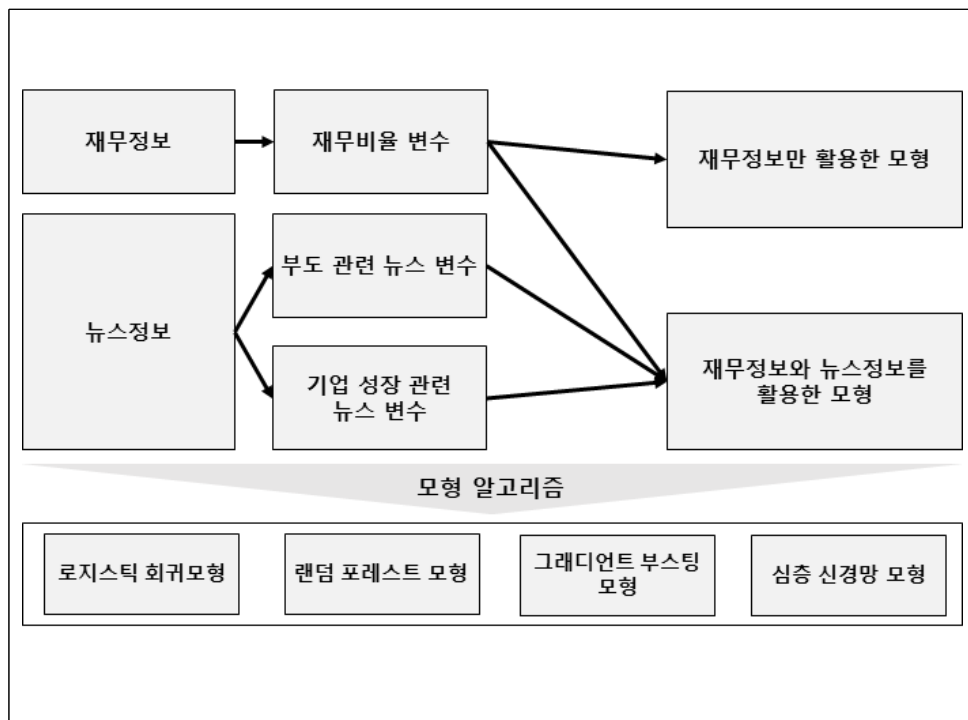
3.2.2 실험절차

본 연구를 위한 실험은 재무변수와 뉴스 변수를 선별하는 변수 선별단계와 연구모형에 재무변수만을 활용하였을 때와 재무변수와 뉴스 변수를 같이 활용하였을 때의 각 모형의 성능을 비교하는 모형 비교 단계로 구성된다.

모형의 성능 비교에는 전통적인 통계 모형인 로지스틱 회귀 모형과 머신러닝 모형인 랜덤 포레스트 모형, 그래디언트 부스팅 모형, 심층 신경

망 모델을 활용하였다.

모델의 성능은 우량과 불량을 비교하는 변별력 측정지표인 정확도 비율(AR; Accuracy Ratio) 등을 활용하였으며, 본 연구의 실험절차는 [그림 3-2]와 같다.



[그림 3-2] 성능 비교를 위한 실험설계

3.3 데이터 정의

3.3.1 대상기업의 정의

재무제표를 통해 기업의 경영실태 및 부실가능성을 예측하기 위해서는 입수된 재무제표가 신뢰성이 확보되어야 한다. 이시영(2021) 및 최정원(2019) 등은 재무제표의 신뢰성을 확보하기 위하여 KOSPI 상장기업을 대상으로 연구를 하였다. 선행연구들이 KOSPI 상장기업을 대상으로 연

구를 하였던 사유는 해당 기업들이 상대적으로 규모가 큰 기업들이기 때문에 자체적으로 회계의 신뢰성을 유지할 수 있는 프로세스를 가지고 있으며, 또한 정기적으로 외부 회계감사를 통해 회계 프로세스의 투명성을 감사받고 있기 때문이다.

본 연구에서는 NICE에서 제공하는 KIS-DATA를 통해 기업의 재무제표를 입수하였으며, 이 중 재무제표의 신뢰성을 확보하기 위하여 외부감사를 받는 외감이상 기업만을 연구대상으로 하였다.

KIS-DATA에서는 기업규모에 따른 기업규모구분 코드를 제공하고 있으며, 해당 코드는 기업의 규모를 대기업, 중소기업, 중견기업, 미해당으로 구분하고 있다. 대기업은 공정거래 위원회가 지정한 대기업집단 지정 결과에 포함되는 기업으로 공시대상 기업집단과 상호출자제한기업집단으로 구분이 된다. 대기업은 여러 기업집단의 재무제표를 통해 연결재무제표를 작성해야 되며, 연결재무제표에는 특정 기업의 성과뿐만 아니라 종속되는 기업의 성과도 포함됨으로 단일 기업의 재무제표와 그 성격이 다르다. 이에 외감기업 중 대기업은 본 연구에서는 제외하였다.

기업들의 업종은 한국표준산업분류에 따라 분류된다. 한국표준산업분류코드는 산업 관련 통계자료의 정확성 및 비교성을 확보하기 위하여 통계청에서 1963년 처음 제정된 이후 지속적으로 개정되어 왔으며, 현재는 제10차 개정이 2017년 7월부터 시행되고 있다(통계청, 2017). 표준산업분류코드는 전체 업종을 총 21개 대분류로 구분하고 있으며, 본 연구에서는 업종규모가 크고 업종간 재무제표의 일관성이 확보되는 제조업, 건설업, 부동산업을 분석대상으로 하였다.

기업의 영업활동 기간과 관련하여, 지속적인 영업을 영위하고 있는 기업과 신설기업은 같은 규모, 같은 업종의 기업이라고 할지라도 재무제표에서 산출되는 지표에서 차이가 존재한다. 이에 본 연구에서는 직전 회

계연도와 당해 회계연도, 2개년 동안 재무제표가 존재하는 기업만을 대상으로 하여, 업력이 상대적으로 짧은 신설기업 및 해당 기간 중 재무제표가 작성되지 않은 기업은 분석대상에서 제외하였다.

앞서 정의한 부분을 반영하여 본 연구에서는 2016년부터 2018년까지의 3개년의 NICE가 가진 전체 외감기업에 대해서 기업규모별, 업종별, 직전 2개년 재무제표 보유여부별 기준으로 연구대상을 확정하였으며, 그 현황은 [표 3-1]과 같다.

[표 3-1] 업종별 대상기업 현황

전체	제조업	건설업	부동산업
47,858	33,662	5,125	9,071

3.3.2 부도의 정의

본 연구는 재무정보 및 비정형 데이터인 뉴스 정보를 활용하여 기업의 신용도를 평가할 수 있는 모형을 연구하고 이를 통해 비정형 데이터의 가치를 실증분석함을 목적으로 한다. 신용도는 기업의 부도 가능성을 평가하여 얼마나 우량한지와 위험한지에 대한 상대적인 지표이며, 본 연구에 있어서 부도를 어떻게 정의할지와 부도를 인식하는 기간을 어떻게 정할지는 매우 중요한 부분이다.

금융감독원(2008)은 일관된 부도의 정의는 신용평가시스템의 가장 기본적인 사항이라고 하였다. 금융감독원(2008)의 부도 정의는 [표 3-2]와 같다.

[표 3-2] 부도의 정의

기준	정의
연체일수	은행에 부담하는 채무자의 채무가 90일 이상 연체한 경우
상환여력평가	보유 담보물의 처분과 같은 상환청구 조치를 취하지 않으면 채무자로부터 채무를 일부라도 상환받지 못할 것으로 판단되는 경우
손실발생	신용악화로 대손충당금을 설정 및 상각한 채권, 악화된 채무재조정으로 원금 및 이자 등 감소, 신용악화로 인한 상당한 경제적 손실이 발생된 채권의 매각

이인로 & 김동철(2015)는 부도의 요건을 상장폐지로 정의하였으며, 권혁진(2017)은 연체 90일 이상에 준하는 신용불량사건을, 최정원(2019)은 부도발생, 화의절차, 감사인의 의견 거절 등으로 발생한 상장폐지 및 워크아웃, 법정관리, 전액 자본잠식으로 인하여 주권 매매거래 정지를 부도의 요건으로 정의하였다. 송충석(2020)은 부실기업의 징후로 9개 요인을 제시하였으며, 전액 자본잠식을 중요한 부실의 징후로 보고 있다.

NICE는 기업고객의 연체정보를 해당 기업이 고객으로 있는 금융회사에만 제공하며, KIS-DATA를 통해서는 제공하고 있지 않다. 법정관리 및 화의정보는 금융회사에 제공되는 서비스 뿐만 아니라 KIS-DATA를 통해서도 제공된다. 이에 본 연구에서는 선행연구를 바탕으로 정리절차, 화의절차, 워크아웃, 회생절차 등 법정관리 및 화의정보의 개시 및 신청을 부도 요건으로 하였으며, 추가로 전액 자본잠식을 부도 요건에 반영하였다.

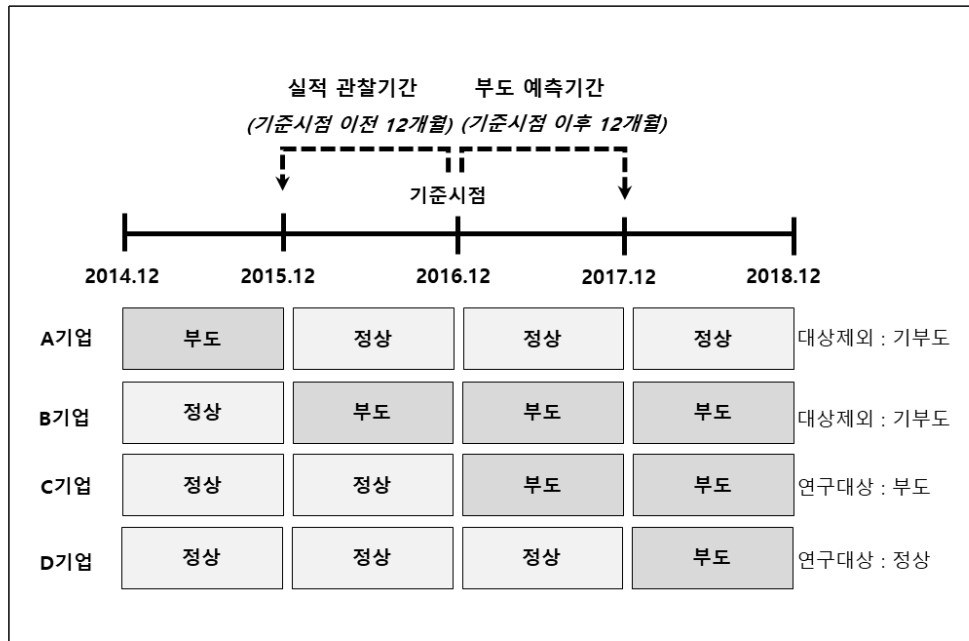
3.3.3 실적 관찰 기간 및 부도 예측 기간

기업 신용평가모형은 기업의 미래 부도가능성을 예측하는 통계 모형이다. 이에 본 연구의 목표변수는 앞서 정의한 부도이다. 이때 기준시점을 기준으로 일정 기간 동안 기업의 부도 발생 여부를 관찰하는 기간을 부도 예측 기간이라고 한다. 반대로 기준시점에 미래의 부도 발생을 예측하기 위해 필요한 과거 데이터의 수집 기간을 실적 관찰 기간이라고 한다. 본 연구는 기업의 부도 예측을 위하여 기업의 재무정보인 재무제표를 활용하고 있다. 재무제표는 특정 회계기간 동안의 기업 실적에 대한 데이터이며, 기업의 회계기간은 통상 12개월을 기준으로 하고 있다. 김종윤(2019)는 개인신용평가모형인 통신스코어를 연구하면서, 실적 관찰 기간을 12개월로 하였다. 기업의 통상적인 회계기간 및 선행연구를 고려하여, 본 연구의 실적 관찰 기간을 12개월로 하였으며, 이에 12개월간의 성과를 측정하는 결산재무제표를 연구에 활용하였다.

부도 예측 기간은 Campbell et al.(2008)은 36개월로 하였으나, 이인로 & 김동철(2015), 권혁진(2017) 및 최정원(2019) 등 최근 대부분 연구들은 부도 예측 기간을 1년으로 보고 있으며, 금융감독원(2005)은 부도확률의 계량화를 위해서 부도확률은 경기변동의 주기를 반영하여야 한다고 하였다. 경기변동을 검토하여 부도율이 높았던 시점과 낮았던 시점을 합리적으로 결합하여야 하며, 해당 기간 동안의 평균 부도 경험을 반영한 1년 부도율은 장기의 평균부도율을 의미한다고 하였다. 이에 본 연구에서는 선행연구를 바탕으로 부도 예측 기간을 1년으로 하였다.

기준시점에는 정상이었으나, 부도 예측 기간 동안에 부도사유가 발생한 기업을 부도로 분류하였으며, 그 외는 정상으로 분류하였다. 기준시점 이전에 부도 요건에 충족하는 사유가 발생한 기업은 기부도로 분류하여 연구대상에서 제외하였다. 또한 최초 부도로 분류된 이후 정상화가 되었

더라도 평가의 일관성을 위하여 이후 개발 대상에서는 제외하였다. 본 연구의 실적 관찰 기간 및 부도 예측 기간은 [그림 3-3]과 같다.



[그림 3-3] 실적 관찰 기간 및 부도 예측 기간

3.3.4 재무 데이터의 정의

기업의 재무제표는 그 기업의 재무상태를 보여주는 여러 가지 표로 구성이 되며, 이를 통해서 기업의 다양한 상태 및 활동을 살펴볼 수 있다. 재무제표는 재무상태표, 손익계산서, 현금흐름표의 중요한 세가지 자료로 구성된다.

재무상태표는 자본과 부채, 그리고 자산으로 구성되며, 이러한 항목을 통해서 해당 기업이 자본과 부채를 어떻게 조달했는지와 어떤 자산 등에 투자가 되었는지에 대한 정보를 가지고 있다.

손익계산서는 일정 기간에 걸쳐서 해당 기업의 수익과 비용을 보여주

고 있으며, 이러한 손익은 영업 관련 부분과 비영업 관련 부분으로 구분되어, 회사의 손익이 일시적인 부분인지와 영속적인 부분인지를 구분할 수 있게 한다.

현금흐름표는 일정 기간 동안에 발생한 실제 현금흐름을 보여주고 있으며, 영업활동, 투자활동, 재무활동에 따른 현금흐름으로 구분된다.

재무제표 항목은 매출액, 영업이익 등 개별 항목을 의미하며, 특정 두 개 항목을 결합하여 산출된 재무비율은 크게 안정성, 수익성, 활동성 등으로 분류된다. 이인로 & 김동철(2015), 권혁진(2017) 및 최정원(2019) 등 선행연구에서 재무비율을 연구에 활용하였다. KIS-DATA에서는 기업의 재무제표 및 재무비율 정보를 제공하고 있으며, 본 연구에서는 KIS-DATA에서 제공하는 재무비율을 기업의 부도를 예측하기 위한 연구모형의 독립변수로 활용하였다.

3.3.5 텍스트 데이터의 정의

기업의 부도 예측을 위하여 본 연구에서는 텍스트 데이터를 독립변수 중 하나로 활용하였다. 본 연구에서 활용된 텍스트 데이터는 기업의 뉴스 데이터이며, 해당 뉴스 데이터는 기업의 성과, 실적, 인사 등에 관련된 내용을 포함하고 있다. 뉴스 데이터는 신문, 방송 등 다양한 채널에서 생성되고 있으며, 공개된 웹사이트에서 이러한 뉴스들이 모두 게재되며, 이러한 뉴스들은 웹크롤링의 방식으로 수집되어 분석을 위한 데이터로 활용된다.

이시영(2021)은 기업의 성과 예측에 있어서 뉴스 본문과 제목에 대한 비교를 통해 뉴스 제목을 연구에 활용하였다. 뉴스 본문을 활용 시에는 단어의 양이 증가하여 학습에 유리한 부분이 있으나, 국내 뉴스의 경우 특정 기업에 대한 내용 및 경제 관련 뉴스를 전달하면서 직접적인 관련

은 없는 내용도 게시되는 경우가 있어서 연구모형의 성능 저하 요인으로 보았다. 이에 본 연구는 해당 기업에 대해서 명확하게 알리고자 하는 의도로 작성된 뉴스 정보만을 활용하고자 하며, 선행연구를 바탕으로 뉴스 제목을 연구에 활용하기로 하였다.

뉴스 정보를 독립변수로써 활용함에 있어서 김찬송(2018)은 TF-IDF를 활용하여 키워드를 추출하고 빈도 기반의 변수를 생성하여 감성 분석을 하였다. 최정원(2019)은 Word2Vec을 활용하여 부도연관 키워드를 추출하여, 이를 통해 빈도 기반의 변수를 생성하여 연구모형에 활용하였다.

뉴스 정보는 기업에 대해서 구독자들이 알고 싶어 하는 내용 또는 구독자들이 모르는 새로운 내용을 담고 있으며, 이러한 뉴스들은 부도와 양의 상관관계를 가지는 내용과 부도와 음의 상관관계를 가지는 내용을 모두 포함하고 있다. 이에 본 연구는 부도와 양의 상관관계를 가지는 키워드와 음의 상관관계를 가지는 키워드를 추출하여 이를 빈도 기반 변수로 생성하여 독립변수로 활용하였다.

3.4 데이터 수집과 전처리

본 연구에서는 기업의 재무데이터와 비정형 데이터인 텍스트 데이터를 활용하였다. 기업의 재무데이터는 재무비율정보를 활용하였으며, 텍스트 데이터는 뉴스 정보를 활용하였다. 기업의 재무데이터는 NICE KIS-DATA를 통해서 NICE가 보유한 외감이상 전체 기업의 데이터를 입수하였다. 입수된 기간은 2015년부터 2019년까지의 5개년이다. 텍스트 데이터인 뉴스 데이터는 네이버 포탈의 경제부문에 게재된 뉴스를 웹크롤링 방식으로, 2016년부터 2018년까지 3개년 데이터를 입수하였다.

3.4.1 재무 데이터

3.4.1.1 재무비율의 범주별 분류

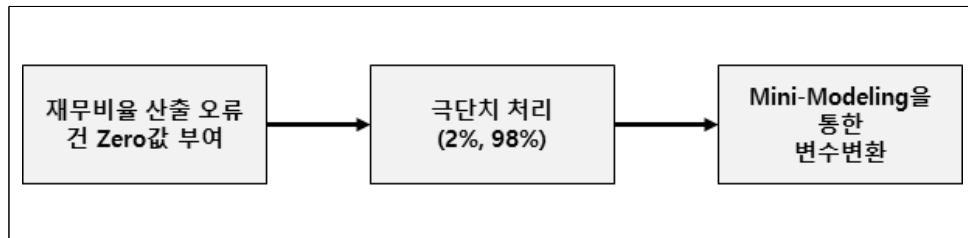
NICE KIS-DATA를 통해 108개 재무비율을 입수하였으며, 본 연구에서는 이를 연구모형의 재무정보의 후보 변수로 활용하였다. 선행연구에서는 기업의 재무비율을 산출하여 이를 여러 범주로 분류를 하였으며, 각 범주별로 대표변수를 선별하여 예측 모형 연구에 활용하였다. 이인로 & 김동철(2015)과 최정원(2019)은 6개 범주로 재무비율을 분류하였으며, 권혁진(2017)은 재무비율을 7개 범주로 분류하였다. 선행연구를 바탕으로 본 연구에서는 재무비율을 6개로 분류를 하였으며, 각 범주별 재무비율 현황 및 주요 재무비율 변수는 [표 3-3]과 같다.

[표 3-3] 재무비율의 범주별 현황

범주	재무비율 개수	주요 재무비율
건전성	26	유보액/총자산비율, 부채비율, 비유동자산비율 등
수익성	37	자기자본순이익율, EBITDA대금융비용, 매출액영업이익율 등
활동성	17	순운전자본회전율, 매입채무회전율, 자본금회전율 등
생산성	13	종업원1인당 순이익, 기계투자효율, 설비투자효율 등
성장성	10	종업원수증가율, 총자산증가율, 자기자본증가율 등
현금흐름	5	총C/F대부채비율, 총C/F대차입금비율, 순C/F대차입금비율 등
합계	108	

3.4.1.2 재무비율 전처리

데이터의 전처리는 모형개발에 있어서 필요한 부분이며, 전처리의 결과는 데이터 분석 전반에 걸쳐 매우 큰 영향을 미치고 있어 중요하게 다루어지고 있다. 본 연구에서 독립변수로 활용하고자 하는 재무비율은 아래와 같이 [그림 3-4]의 절차를 통해 극단치 처리 및 변수변환의 데이터 전처리 과정을 수행하였다.



[그림 3-4] 재무비율 전처리 절차

재무비율은 재무제표를 기반으로 산출되며, 재무제표의 값이 미산출된 경우에는 해당 재무제표 값을 사용하는 재무비율 값 또한 결측치로써 Null의 값이 부여된다. Null의 값이 부여된 관측치는 분석에 영향을 미치지 않는 값으로 대체가 필요하며, 본 연구에서는 결측치에 대해서는 Zero의 값으로 대체하였다.

극단치는 데이터의 분포가 다른 값들에 비하여 비정상적으로 떨어져 있는 관측치를 의미한다. 극단치 값은 통계분석에 영향을 미치고 그 결과를 왜곡할 수 있기 때문에 전처리를 한다. 본 연구에서는 극단치에 대해서 하위 2% 이하 관측치는 하위 2%의 값으로, 상위 98% 이상의 관측치는 98%에 해당하는 값으로 개별 관측치의 값을 대체하였다.

Falkenstein et al.(2000)은 기업의 부도율 예측 연구에 재무비율을 활용하였으며, 재무비율의 특징인 비선형성을 개선하는 방안을 제시하였다.

Falkenstein et al.(2000)은 재무비율은 50개의 서열화된 구간으로 구분하여, 각 구간별 실측 부도율을 비모수적인 방법으로 보정한 추정 부도율로 대체하여 독립변수로 활용하는 Mini-Modeling 방식을 제시하였다. 권혁진(2017)은 Mini-Modeling 방식을 활용하여 기업의 부도 예측 모형을 연구하였다. 30개의 서열화된 구간으로 각 재무비율을 구분하였으며, 비모수 방법인 Local Weighted Regression를 활용하여 각 구간의 평균 추세를 추정부도율로 변환하여 연구모형의 독립변수로 활용하였다. 현재 대부분은 국내 은행들은 기업신용평가모형 개발시 Mini-Modeling 방식을 통해 재무비율을 변환하여 활용하고 있다(윤동희, 2013).

이에 본 연구에서는 선행연구를 바탕으로 재무비율의 선형성 확보를 위하여 Mini-modeling 방식을 활용하였다. 연구를 위해 KIS-DATA를 통해 입수된 각 재무비율을 각 재무비율의 특성에 맞춰 50여개의 서열화된 등구간으로 구분하였다. 그리고 각 구간의 실측부도율을 Local Weighted Regression 변환을 통해 추정 재무비율로 변환하였으며, 변환된 재무비율을 연구의 독립변수로 활용하였다.

3.4.1.3 단변량 분석

단변량 분석은 변수의 특성을 분석하여 유의한 변수를 선별하는 중요한 과정이다. 최소운 & 안현철(2015)는 t-검증을 통해 164개 재무비율 중 유의한 재무비율을 찾아 연구에 활용하였다. 권혁진(2017)은 재무비율에 대해서 t-검증을 통해 정상기업과 부도기업 간의 평균의 차이가 있는지를 검토하였으며, 부도를 종속변수로 하는 단순 로지스틱 회귀 분석을 통해 추정된 회귀계수의 유의성을 점검하고, 각 변환 재무비율의 AR을 통해 변별력의 유의성을 점검하여 유의한 재무비율을 선별하였다. 김종운(2019)은 서열화에 대한 유의성 분석 및 변별력 분석을 통해

Kolmogorov-Sminrov(K-S) 통계량이 최소 10%를 충족하는 변수를 선별하였다.

본 연구는 선행연구를 바탕으로 변환 재무비율에 대해서 t-검증, 단순 로지스틱 회귀 분석을 통해 추정된 회귀계수의 유의성 검증, 변별력 분석을 진행하여 유의한 변수를 선별하였다. 본 연구에 활용된 단변량 분석 기준은 [표 3-4]와 같다.

[표 3-4] 단변량 분석 기준

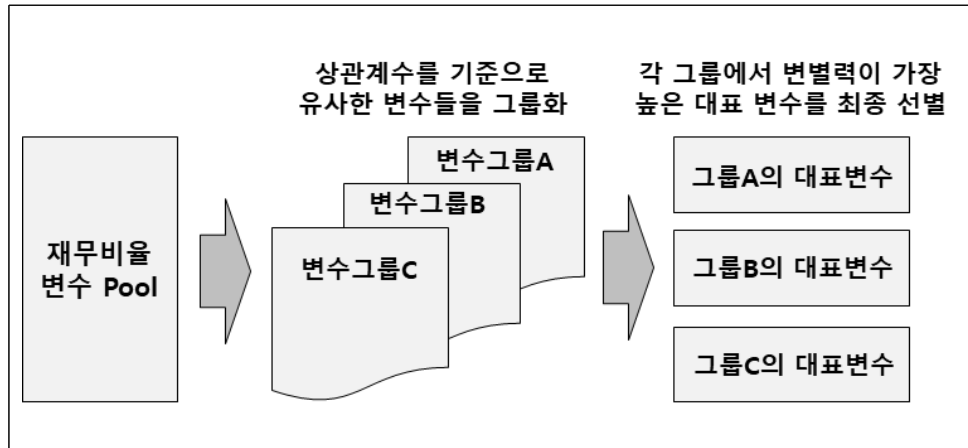
검증항목	세부 기준
t-검증	t-검증의 p-value가 0.05 미만
회귀계수의 유의성	단순 로지스틱 회귀 분석을 통해 추정된 회귀계수의 p-value 0.05 미만
변수의 변별력	AR 및 K-S 통계량이 15% 이상

3.4.1.4 상관 분석을 통한 최종 변수 선별

단변량 분석을 통과한 변수들에 대해서 상관관계 분석을 진행한다. 권혁진(2017)은 기업의 부도 예측 모형을 추정시 의미적으로나 구조적으로나 유사한 변수들이 모형에 포함될 경우 모형의 성능이 저하될 가능성이 있다고 하였으며, 이에 상관 분석을 통해서 변별력이 더 높은 변수를 선별하여 연구에 활용하였다.

본 연구에서는 모형의 안정성 및 간결성 확보를 위하여 단변량 분석을 통과한 변수들에 대해 상관 분석을 실시하였으며, 상관계수가 0.7 이상인 변수들을 하나의 그룹으로 분류하였다. 이후 각 그룹에서 변별력이 가장 높은 변수를 최종 변수로 선별하였다. 본 연구에서 상관 분석을 통해 최

중 변수를 선별하는 과정은 [그림 3-5]와 같다.



[그림 3-5] 상관 분석을 통한 최종 변수 선별 과정

3.4.2 텍스트 데이터

3.4.2.1 대상기업의 뉴스 데이터 정의

본 연구에서 텍스트 데이터는 뉴스 데이터를 활용하였다. 네이버 포털의 경제 섹션의 기사를 크롤링하여 수집된 뉴스 데이터는 뉴스 제목, 뉴스 본문, 뉴스 작성일시로 구성되며, 이 중 뉴스 제목과 뉴스 작성일시를 연구에 활용하였다. 관측치의 기업명이 뉴스 제목에 존재하는 경우 해당 기업의 뉴스로 분류하였으며, 실적 관찰 기간으로 정의된 직전 12개월 동안에 작성된 뉴스만을 기업의 부도 예측에 활용하였다.

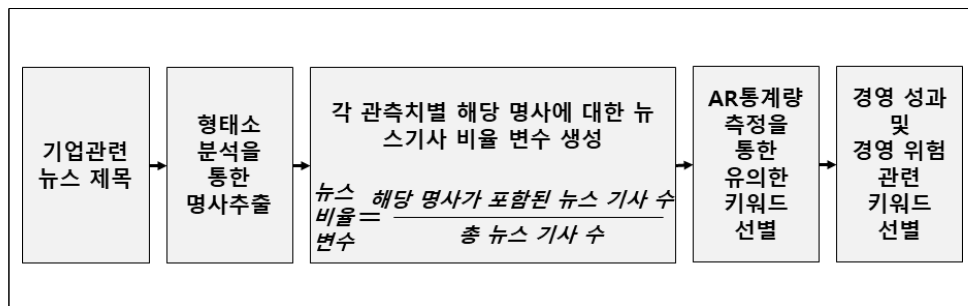
3.4.2.2 뉴스 데이터의 전처리 및 키워드 선별

뉴스 데이터를 기업의 부도 예측 연구에 활용하기 위해서는 전처리 과정을 통해 계량화가 필요하다. 최정원(2019)은 뉴스 데이터를 형태소 분석을 통하여 명사만을 추출하였으며, Word2Vec을 통해 부도 관련 키워드를 선별하였다. 김찬송(2018)은 뉴스 데이터를 형태소 분석을 통해서

명사를 추출하였으며, 명사의 빈도를 기반으로 감성 분석을 진행하여 기업의 긍정 및 부정적인 키워드를 선별하였다.

기업에 대한 뉴스의 키워드가 부도 예측에 활용되기 위해서는 변별력을 가지고 있어야 하며, 의미적으로도 설명력을 가지고 있어야 한다. 본 연구에서는 선행연구를 바탕으로 뉴스 제목을 형태소 분석하여 명사만을 추출하였다. 추출된 명사를 하나의 변수로 하여, 개별 관측치인 기업에 대해서 해당 명사의 뉴스 포함비율 변수를 생성하여 AR 통계량을 측정하였다.

뉴스에는 기업의 경영성과 및 위험을 의미하는 내용 이외에 마케팅, 광고목적의 내용, 인터뷰, 칼럼 등 다양한 내용이 포함되어 있다. 본 연구에서는 AR 통계량이 높은 키워드를 선별하고 이 중 기업의 경영성과 및 경영위험을 의미하는 키워드만을 최종 키워드로 선별하였다. 뉴스 데이터의 전처리 및 키워드 산출과정은 [그림 3-6]과 같다.



[그림 3-6] 뉴스 데이터의 전처리 및 키워드 산출과정

3.4.2.3 키워드를 활용한 뉴스 변수의 계량화

본 연구는 기업의 부도를 예측하기 위한 독립변수으로써 뉴스 데이터를 고려하였다. 형태소 분석을 통해 추출된 명사를 키워드로 하여, 각 키워드의 변별력을 측정하여 경영성과 및 경영위험에 관련된 키워드를 추출

하였으며, 추출된 키워드들에 대해 부도와 방향성을 측정하였다. 부도와 양의 상관관계를 가지는 키워드들은 하나로 묶어서 부도 관련 뉴스 기사 비율 변수를 생성하였으며, 부도와 음의 상관관계를 가지는 키워드들을 하나로 묶어서 기업 성장 관련 뉴스 기사 비율 변수를 생성하였다.

$$\text{부도 관련 뉴스 기사 비율} = \frac{\text{부도 관련 키워드들이 포함된 뉴스 기사 수}}{\text{총 뉴스 기사 수}}$$

$$\text{기업 성장 관련 뉴스 기사 비율} = \frac{\text{기업 성장 관련 키워드들이 포함된 뉴스 기사 수}}{\text{총 뉴스 기사 수}}$$

3.5 연구데이터

3.5.1 연구데이터의 생성

본 연구는 기업의 부도를 예측하기 위해 독립변수로써 기존 재무정보 외에 비정형 데이터인 뉴스 데이터를 고려하여 비정형 데이터의 가치를 실증분석하기 위한 것이다. 이에 기업의 재무 데이터와 뉴스 데이터가 모두 존재하는 기업만을 최종 연구대상으로 활용하였다.

3.5.2 연구데이터의 표본추출 및 분할

본 연구에서는 분석의 신속성과 분석 결과의 오차감소를 위하여 전체 연구데이터를 정상과 부도의 비율이 9 : 1 이 되도록 층화추출하여 표본을 구성하였다.

연구모형을 학습하기 위한 학습 데이터는 전체 표본 데이터의 70%로 구성하고, 학습된 모형의 성능을 점검하기 위한 테스트 데이터는 30%로 하여 데이터를 분할하였다.

머신러닝 모형을 활용시, 학습 데이터만을 활용하여 학습을 진행하였을 경우에 학습 데이터에 과적합되어 테스트 데이터에서는 모형의 성능이 크게 저하되는 경향을 보인다. 이에 과적합으로 모형의 성능 저하가 발생하지 않도록 학습 데이터는 다시 훈련 데이터와 검증 데이터로 분리하여 활용하였다. 훈련 데이터로 학습을 한 뒤 해당 모형을 검증 데이터를 통해 변별력을 점검하였다. 검증 데이터의 변별력이 최적화되도록 모형의 하이퍼 파라미터를 설정하여, 이를 최종모형으로 하였다. 본 연구에서는 학습 데이터 70%를 훈련 데이터 40%와 검증 데이터 30%로 분리하여 활용하였다. 본 연구에서 활용한 데이터의 분할은 [표 3-5]와 같으며, 각 데이터의 역할은 [표 3-6]과 같다.

[표 3-5] 연구데이터의 분할

전체 데이터 (100%)		
학습 데이터 (70%)		테스트 데이터 (30%)
훈련 데이터 (40%)	검증 데이터 (30%)	

[표 3-6] 분할 데이터의 역할

구분	역할
훈련 데이터	모형 추정
검증 데이터	추정한 모형의 성능이 적합한지를 평가하며 최적의 하이퍼 파라미터를 결정
테스트 데이터	최종적으로 선택된 모형의 성능 평가

3.6 연구모형의 구성

본 연구에서는 모형이 최적의 성능을 가질 수 있도록 하이퍼 파라미터를 단계적으로 선정하기 위한 실험을 진행하여 최적의 모형을 결정하였다. 연구모형은 재무 데이터만을 반영하였을 경우와 재무 데이터와 비정형 데이터인 뉴스 데이터를 반영하였을 경우로 나누어서 각각에 대하여 모형을 최적화하였다.

3.6.1 로지스틱 회귀 모형

로지스틱 회귀 모형은 종속변수가 0과 1의 값을 가지는 범주형 변수인 경우에 활용되는 전통적인 통계 모형으로 종속변수와 독립변수 간의 함수관계를 분석하는 방법이다(홍세희, 2005).

본 연구에서는 독립변수가 여러 개인 로지스틱 회귀 모형을 활용하였다. 로지스틱 회귀 모형에 활용되는 변수를 선택하는 방법에는 [표 3-7]과 같이 3가지 방법이 있으며 본 연구에서는 단계적 방법을 활용하여 변수를 선택하였다.

[표 3-7] 로지스틱 회귀 모형 변수 선택

구분	역할
전진 선택법	가장 유의한 변수부터 하나씩 추가를 하면서 모형의 성능을 비교
후진 제거법	모든 변수를 넣고 하나씩 제거해가면서 모형의 성능을 비교
단계적 선택법	변수를 하나부터 시작해서 추가하되 변수의 중요성이 낮은 변수는 제거하고 다른 변수를 추가하면서 모형의 성능을 비교

출처: 홍세희(2005)

3.6.2 랜덤 포레스트 모형

랜덤 포레스트 모형은 여러 개의 의사결정나무를 활용한 알고리즘으로, 동일한 알고리즘으로 여러 분류기를 만든 후 예측 결과를 보팅(voting)으로 결정하는 기법을 의미한다. 랜덤 포레스트 모형은 의사결정나무 모형의 쉽고 직관적인 장점을 그대로 가지고 있으며 앙상블 알고리즘 중 비교적 빠른 수행 속도를 가지고 있고 다양한 분야에서 좋은 성능을 나타낸다고 알려져 있다.

랜덤 포레스트 모형의 주요한 하이퍼 파라미터는 의사결정나무의 개수와 의사결정나무의 최대 깊이이다. 본 연구에서는 Random State는 다른 머신러닝 모형과 마찬가지로 동일한 결과를 얻기 위하여 난수값의 초기 Seed를 부여하였다.

최적의 하이퍼 파라미터를 찾기 위하여, 모든 하이퍼 파라미터는 초기값을 부여하고 의사결정나무의 개수를 1개에서 100개까지 순차적으로 시뮬레이션하여 최적의 의사결정나무의 개수를 선택하였으며, 이후 의사결정나무의 최대 깊이를 2개에서 20개까지 순차적으로 시뮬레이션하여 최적의 모형을 선택하였다.

3.6.3 그래디언트 부스팅 모형

그래디언트 부스팅 모형은 1997년 Breiman에 의하여 처음 소개가 되었으며, Jerome H Friedman에 의하여 발전되었다(Géron, 2020). 그래디언트 부스팅 모형은 약한 분류기를 결합하여 강한 분류기를 만드는 앙상블 모형으로 이전 모형의 예측 오류를 보완하는 형태로 학습을 반복 수행하게 되므로 수치 예측 및 분류 예측에 높은 성능을 보인다.

그래디언트 부스팅 모형의 주요한 하이퍼 파라미터는 의사결정나무의 개수, 의사결정나무의 최대 깊이와 학습률이다. 본 연구에서는 Random

State는 다른 머신러닝 모형과 마찬가지로 동일한 결과를 얻기 위하여 난수값의 초기 Seed를 부여하였다.

최적의 하이퍼 파라미터를 찾기 위하여, 모든 하이퍼 파라미터는 초기 값을 부여하고 의사결정나무의 개수를 1개에서 100개까지 순차적으로 시뮬레이션하여 최적의 의사결정나무의 개수를 선택하였으며, 이후 의사결정나무의 최대 깊이를 2개에서 20개까지 순차적으로 시뮬레이션하였으며, 마지막으로 학습률을 0.1에서 0.4까지 단계적으로 시뮬레이션하여 최적의 모형을 선택하였다.

3.6.4 심층 신경망 모형

심층 신경망 모형은 입력층과 출력층 사이에 여러 개의 은닉층(hidden layer)들로 이루어진 인공신경망으로 알고리즘을 통해 스스로 분류 레이블을 만들어 내며, 데이터를 구분하는 과정을 반복하는 학습과정을 통해 최적의 구분선을 도출한다. 심층 신경망의 학습은 오류역전파 알고리즘을 통해 진행되며, 이때 가중치들은 확률적 경사하강법을 통해 갱신된다.

본 연구에서는 모형의 특성에 맞춰 모형의 구조를 결정하였으며, 이후 학습횟수를 시뮬레이션하여 최적의 모형을 선택하였다.

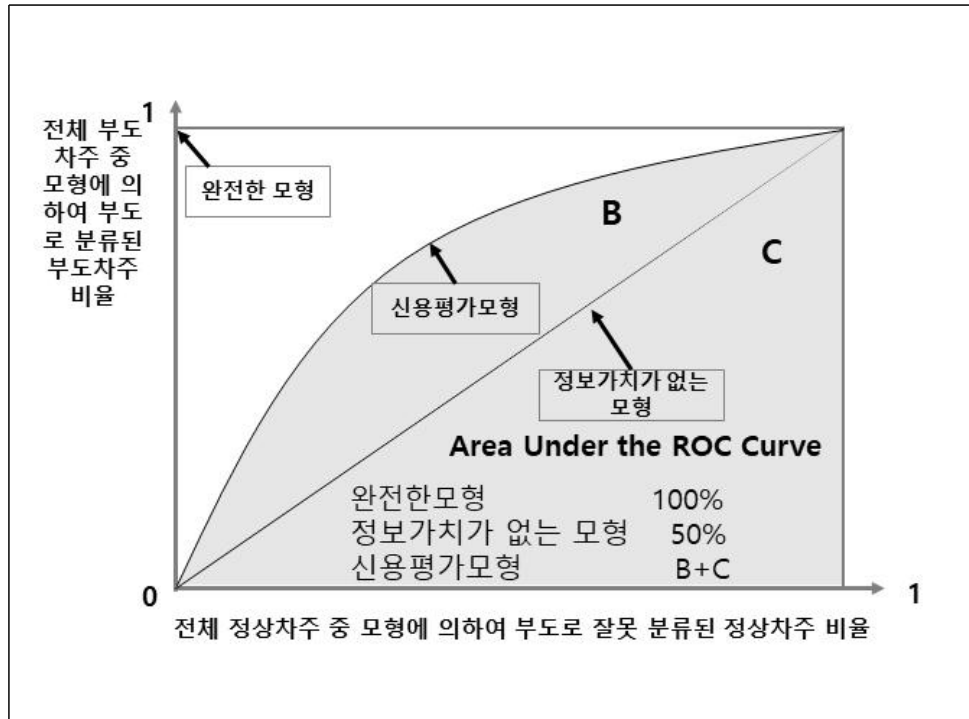
3.7 연구모형의 성능 측정

모형의 성능 평가는 연속형에 대한 평가와 이진 분류에 대한 평가로 구분된다. 본 연구에서는 연속형에 대한 평가지표로써 AR과 K-S을 활용하였다. 그리고 이진 분류에 대한 평가지표는 혼동행렬을 통해 산출되는 Accuracy, Precision, Recall 그리고 F1 Score을 활용하였다.

3.7.1 연속형 모형의 평가

신용평가모형은 산출된 각 등급에 대해서 우량과 불량을 측정하는 연속형 모형이다. 연속형 모형의 평가에는 상대적인 성과에 대한 측정지표인 AR과 K-S 통계량이 주로 활용된다. 금융감독원(2005)은 신용리스크 내부등급법 기본 세부 지침(안)에서 신용평가모형의 변별력 측정을 위한 지표로써 AR을 제시하였으며, 김종운(2019)은 신용평가모형의 변별력 측정지표로써 Area Under ROC(AUROC), K-S 통계량을 활용하였다.

AUROC는 ROC(Receiver Operator Characteristic) 곡선의 면적을 의미한다. ROC 곡선은 누적 우량고객 수를 X축으로, 누적 불량고객 수를 Y축으로 만들어지게 되며, 그 값이 클수록 등급에 따른 우량과 불량에 변별이 잘되고 있음을 의미한다. 신용평가모형은 ROC 곡선의 양극단에 위치하며, 그 수준에 따라 완전한 모형과 정보 가치가 없는 모형 사이에 존재하게 된다. 우량과 불량을 잘 구분하는 우수한 모형은 예측된 부도율이 높은 등급의 구간에서는 부도 차주의 구성비가 높게 되며, 정상 차주의 구성비가 낮게 된다. 따라서 변별력이 높은 모형일수록 ROC곡선의 면적인 AUROC가 커지게 된다. ROC 곡선과 AUROC는 [그림 3-7]과 같다. AR은 AUROC를 계량화한 지표로써, AUROC와 AR의 수식은 아래와 같다.



출처: 금융감독원(2005)

[그림 3-7] ROC 곡선과 AUROC

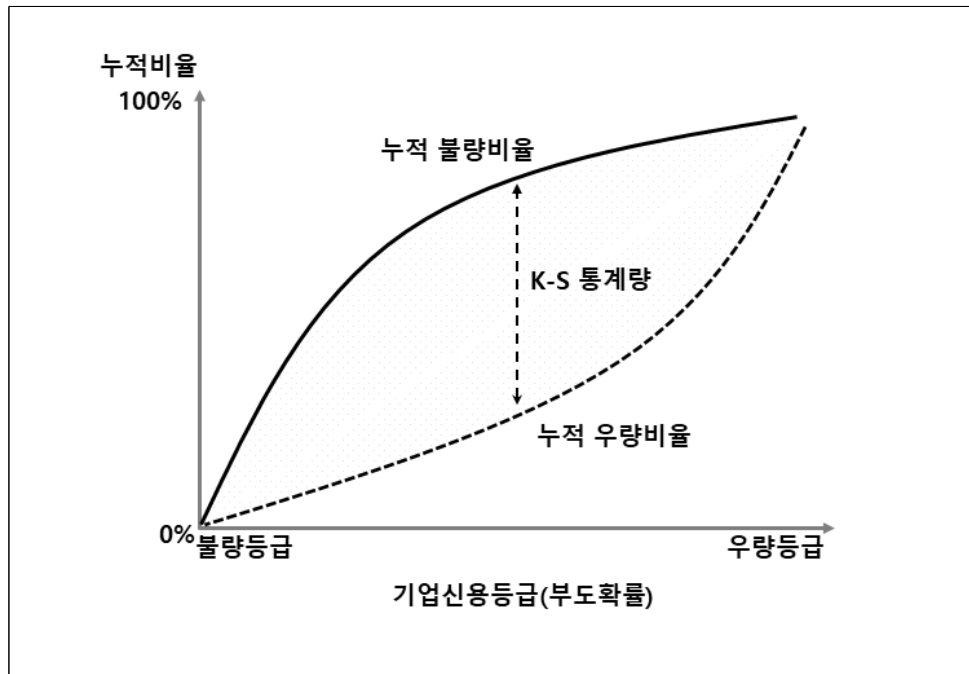
$$AUROC = \sum_{i=1}^n [1/2 Good_i\% \times Bad_i\% + (1 - cumGood_i) \times Bad_i\%]$$

$$AR = 2 \times AUROC - 1$$

위의 AUROC 수식에서 $Good_i\%$ 는 전체 우량고객 중 i 등급에 속한 우량고객 비중을, $Bad_i\%$ 는 전체 불량고객 중에 i 등급에 속한 불량고객 비중을 $cumGood_i$ 는 낮은 등급에서 i 등급까지 누적 계산된 $Good_i\%$ 를 의미한다. Hand(2009)는 AUROC의 값이 70% 이상인 경우, 우량과 불량을 잘 분류하고 있어 변별력이 우수하다고 판단하고 있으며 이를 AR로 환

산하면 40% 수준이다.

K-S 통계량은 모형의 변별력이 극대화되는 지점을 측정하여 평가하는 지표이며, 우량집단과 불량집단의 누적 분포 차이의 최대값으로 산출한다. 김종윤(2019)은 K-S 통계량이 40% 이상이면 변별력이 우수한 모형으로 판단하고 있으며, K-S 통계량은 [그림 3-8]과 같으며, 그 수식은 다음과 같다.



출처: 박종원 & 안성만(2014)

[그림 3-8] K-S 통계량 Curve

$$K-S = \text{MAX}[|cumGood_i\% - cumBad_i\%|]$$

3.7.2 이진 분류 모형의 평가

이진 분류에서는 일반적으로 두 개의 클래스인 Positive와 Negative에 대해 혼동행렬을 통해 예측 오류의 유형과 그 수준에 대해서 평가를 한다. 혼동행렬을 통해서 모형의 평가지표로써, Accuracy, Precision, Recall, F1 Score의 4개 지표가 산출된다. 각 지표에 대해서 살펴보면 아래와 같다.

Accuracy는 전체 예측한 클래스와 실제 클래스가 일치하는 비율을 의미한다.

Precision은 Positive로 예측한 클래스 중 실제 Positive인 클래스의 비율을 의미한다. 본 연구에서 낮은 Precision은 부도로 예측한 차주 중 우량고객의 비율이 높음을 의미하며, 부도 차주로 잘못 예측된 우량차주에 대해서 대출이 실행되지 못한 기회손실이 발생한다.

Recall은 실제 Positive인 클래스 중에 Positive로 예측된 클래스의 비율을 의미한다. 본 연구에서 낮은 Recall은 실제 부도 차주를 우량고객으로 예측한 비율이 높음을 의미하며, 우량차주로 잘못 예측된 부도 차주에 대해 대출이 실행될 경우 실질적인 금융 손실이 발생하게 된다.

F1 Score는 Precision과 Recall의 조화평균을 의미한다. 높은 F1 Score를 얻기 위해서는 Precision과 Recall이 모두 높아야 한다.

위에서 기술한 혼동행렬은 [표 3-8]과 같으며, 모형 성능 측정지표의 산출 수식은 다음과 같다.

[표 3-8] 혼동행렬

		Predict	
		Positive	Negative
Actual	Positive	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

제 4 장 실험 및 분석 결과

4.1 실험 환경 및 도구

본 연구의 실험은 개인 PC, 파이썬 및 SAS를 활용하여 수행하였다. 개인 PC의 운영체제는 MS Window 11을 사용하였다. 파이썬은 Google Colab의 파이썬 라이브러리를 활용하였으며, SAS는 SAS OnDemand for Academics를 활용하였다.

재무비율 등의 기업의 재무 데이터는 NICE KIS-DATA를 통해 입수하였으며, 뉴스 기사는 Google Colab을 사용하여 웹 크롤링을 통해 입수하여 활용하였다. NICE KIS-DATA를 통해 입수된 재무 데이터는 SAS를 활용하여 데이터 전처리를 수행하였다. 뉴스 기사는 Google Colab의 파이썬 라이브러리를 통해 전처리를 수행하였으며, 텍스트 분석을 위한 형태소 분석은 KoNLPy의 Mecab 형태소 분석기를 활용하였다.

분석모형 중 로지스틱 회귀 모형은 SAS를 활용하여 분석하였으며, 랜덤 포레스트, 그래디언트 부스팅 및 심층 신경망 모형은 Google Colab의 파이썬 라이브러리를 활용하였다.

실험 환경 및 사용 도구는 [표 4-1]과 같다.

[표 4-1] 실험 환경 및 사용 도구

구분		솔루션
개인 PC 환경		MS Window 11
데이터 수집	재무정보	NICE KIS-DATA
	뉴스정보	Google Colab 웹 크롤링
데이터 전처리	재무정보	SAS OnDemand for Academics
	뉴스정보	Google Colab KoNLPy의 Mecab 형태소 분석기
모델링	SAS	로지스틱 회귀 모형
	파이썬	랜덤 포레스트 모형 그래디언트 부스팅 모형 심층 신경망 모형

4.2 개발 모집단

4.2.1 재무정보를 활용한 분석대상

본 연구의 분석대상은 외감이상 기업 중 업종은 제조업, 건설업, 부동산업이며, 기간은 2016년부터 2018년까지 3개년이다. 3.3절 데이터의 정의에 따라 대기업, 신설기업 및 기부도를 제외하였으며, 기준시점을 결산 재무제표의 결산년월로 하여, 부도 예측 기간인 12개월 동안에 자본잠식, 법정관리 등의 사유가 발생한 경우에 부도로 분류를 하였다. 분석대상의 업종별 부도율 현황은 [표 4-2]와 같으며, 연도별 부도율 현황은 [표 4-3]과 같다.

[표 4-2] 업종별 부도율 현황

업종	전체 업체수	부도업체수	부도율
제조업	33,662	645	1.92%
건설업	5,125	136	2.65%
부동산업	9,071	572	6.31%
전체	47,858	1,353	2.83%

[표 4-3] 연도별 부도율 현황

기준연도	전체 업체수	부도업체수	부도율
2016년	15,115	340	2.25%
2017년	16,212	494	3.05%
2018년	16,531	519	3.14%
전체	47,858	1,353	2.83%

4.2.2 뉴스 데이터와 결합

본 연구는 재무정보와 비정형 데이터인 뉴스 정보의 결합을 통한 기업 신용평가모형에 대한 연구이며, 이를 통해 기업의 부도 예측에 있어 비정형 데이터인 뉴스 정보의 가치를 실증분석하고자 한다.

기업명과 뉴스 정보를 결합한 선행연구에서 이시영(2021) 등 선행연구에서는 정보검색에 어려움이 있는 2글자 이하의 기업명은 분석대상에서 제외하였다. 이에 본 연구에서는 선행연구를 바탕으로 재무정보를 통해 산출된 분석대상과 뉴스 정보를 결합할 때 기업의 이름을 지칭하는 고유 명사가 아닌 일반명사로 활용될 가능성이 높은 2글자 이하의 기업명을

가진 업체는 분석대상에서 제외하였으며, 3글자 이상의 기업명을 가진 업체 43,466건을 분석대상에 포함하였다. 기업명의 글자 수 구성 현황은 [표 4-4]와 같다.

[표 4-4] 기업명의 글자 수 구성 현황

구분	전체 업체 수	구성비
3글자 이상	43,466	91%
2글자 이하	4,392	9%
전체	47,858	100%

비정형 데이터인 뉴스 정보는 네이버 포탈의 경제 부분에 게재된 뉴스를 웹 크롤링 방식으로 2016년부터 2018년까지 총 3개년에 걸쳐 약 135만 건을 수집하였으며, 그 현황은 [표 4-5]와 같다.

3글자 이상의 기업명을 가진 업체를 대상으로 기업명이 기준시점을 기준으로 실적 관측 기간인 직전 12개월 동안 뉴스 제목에 존재하는 업체는 총 6,650건으로, 전체 분석대상 기업 중 약 15%만이 뉴스 제목에 기업명이 존재하였다.

이에 본 연구에서는 뉴스 데이터와 결합 가능한 6,650건의 업체를 분석대상으로 하였으며, 뉴스 데이터와 결합 현황은 [표 4-6]과 같다.

[표 4-5] 연도별 뉴스 데이터 현황

연도	2016년	2017년	2018년	합계
건수	438,040	464,340	450,526	1,352,906

[표 4-6] 3글자 이상 기업의 뉴스 데이터 결합 현황

구분	재무 정보 보유업체 수	뉴스 정보 보유업체 수	뉴스 정보 보유 비율
전체 업체	43,466	6,650	15.3%
부도 업체	1,232	118	9.6%
부도율	2.8%	1.8%	

4.2.3 표본추출 및 데이터 분할

연구를 위한 최종 분석대상은 뉴스 데이터와 결합되는 업체로 하였으며, 분석을 위해서 정상과 부도 건의 비율을 9 : 1이 되도록 표본추출을 하였다. 최종 개발 데이터 및 모형 성능 비교를 위한 데이터의 분할은 [표 4-7]과 같다.

[표 4-7] 표본추출 및 데이터 분할 결과

	비율	건수
훈련 데이터	40%	472
검증 데이터	30%	354
테스트 데이터	30%	354
전체	100%	1,180

4.3 재무비율 평가항목

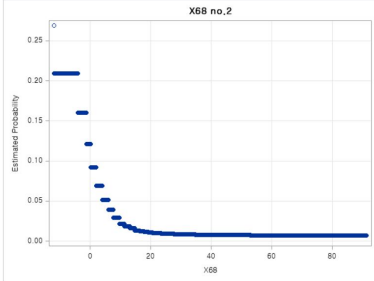
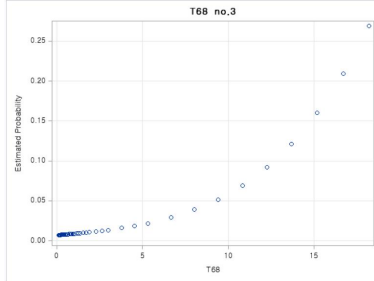
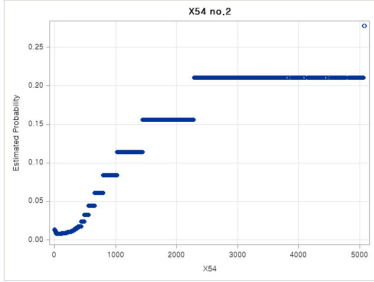
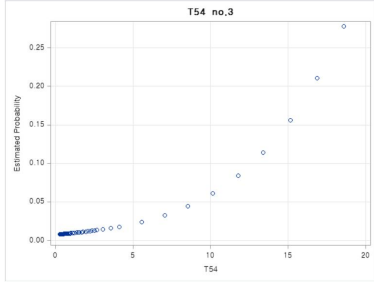
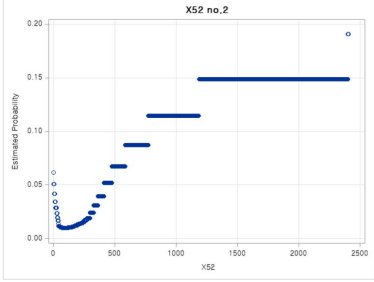
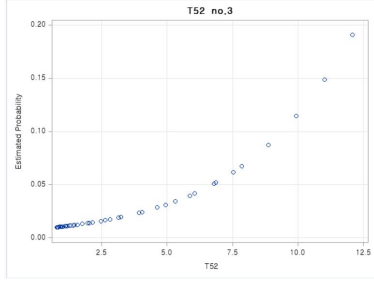
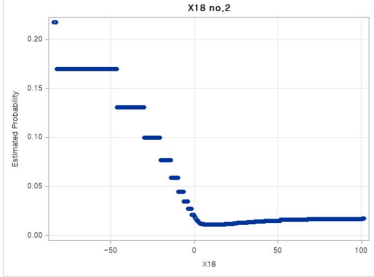
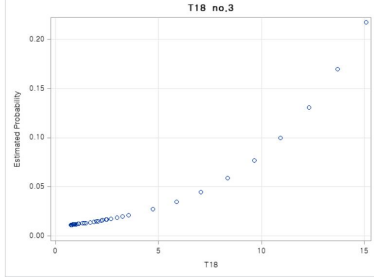
본 연구에 활용할 재무비율 평가항목의 산출을 위하여 전처리, 단변량 분석 및 상관 분석의 단계를 수행하였다. 해당 분석은 재무정보를 통해 확정된 분석대상인 47,858을 대상으로 진행하였다.

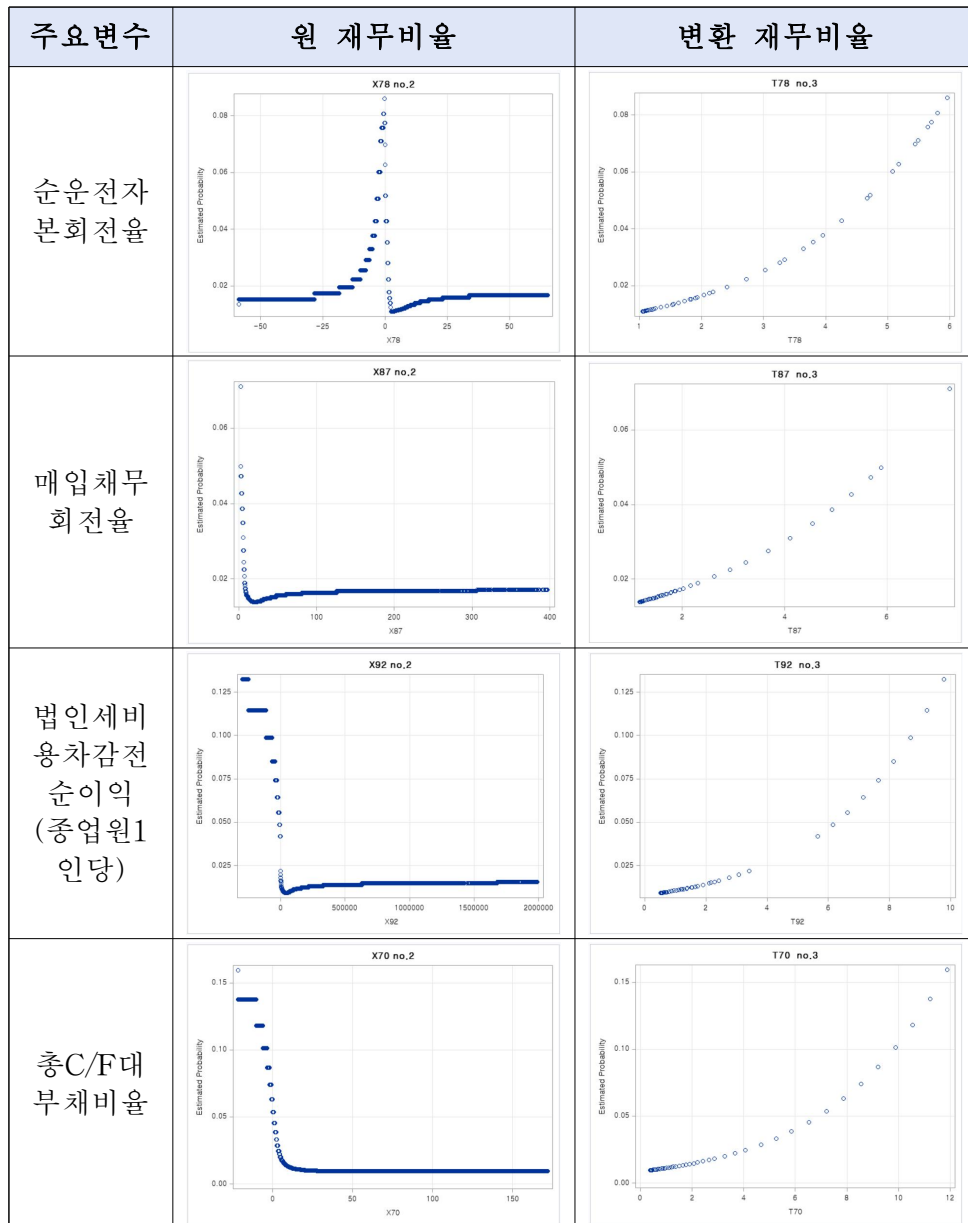
4.3.1 재무비율 전처리

재무비율은 결측치를 Zero로 처리하였으며, 상위 98% 이상 및 하위 2% 이하의 극단치에 대해서 98% 및 2%에 해당하는 재무비율의 값으로 대체하였다.

재무비율의 비선형성을 개선하기 위하여 Mini-Modeling을 통해 추정 부도율을 변환 재무비율로 활용하였다. Mini-Modeling은 비모수적인 방법인 Local Weighted Regression를 활용하였으며, SAS Institute(2018)는 Smoothing Parameter는 Default로 0.5의 값을 가진다고 하였으며, Cohen(1999)는 Smoothing Parameter를 결정할 때, AICC 통계량이 가장 작게 나오는 Smoothing Parameter가 최적의 Parameter라고 하였다.

이에 본 연구에서는 Smoothing Parameter로 0.4, 0.5 그리고 0.6의 3가지를 검토하였으며, AICC 통계량이 가장 작게 나오는 0.4를 최적의 Parameter로 선택하여 Mini-Modeling을 수행하였다. 원 재무비율과 변환 재무비율에 대해 각각에 대해서 단순 로지스틱 회귀 모형을 실행하여 추정된 부도확률에 대해 원 재무비율과 변환 재무비율의 선형성을 점검하였으며, [그림 4-1]과 같이 변환 재무비율이 선형성을 확보함을 확인하였다.

주요변수	원 재무비율	변환 재무비율
유보액/총 자산비율		
부채비율		
비유동자 산비율		
자기자본 순이익율		



[그림 4-1] 주요 재무비율의 원 재무비율과 변환 재무비율의 선형성

4.3.2 단변량 분석 및 상관 분석을 통한 최종 변수 선별

108개 재무비율을 Mini-Modeling을 통해 변환 재무비율로 변환하였다. 변환 재무비율에 대해서 단변량 분석 및 상관 분석을 통해서 39개 변수가 선별되었다. 변수의 선별과정은 [표 4-8]과 같으며, 최종 선정된 39개 변환 재무변수의 기술 통계량은 [표 4-9]와 같다.

[표 4-8] 변환 재무비율 변수 선택과정

수행 단계	변환 재무비율 수
1. 전체 변수	108
2. 단변량 통과 (t-검증, 회귀계수 유의성, 변별력)	72
3. 상관 분석을 통한 최종변수 선별	39

[표 4-9] 최종 변환 재무비율의 기술 통계량

범주	변수 코드	변수명	평균	표준 편차	AR	K-S
건전성	T68	유보액/총자산비율	3.05	4.76	74.42	61.10
건전성	T54	부채비율	3.12	4.58	70.75	58.53
건전성	T69	유보액/납입자본비율	3.06	4.17	62.51	48.55
건전성	T52	비유동자산비율	3.08	2.99	54.76	43.60
건전성	T56	비유동부채비율	2.90	2.81	50.02	43.98
건전성	T51	현금비율	2.89	2.43	46.67	36.37
건전성	T50	당좌비율	2.91	2.37	43.97	33.31
건전성	T60	매출채권/매입채무비율	2.94	2.14	39.00	33.33

범주	변수 코드	변수명	평균	표준 편차	AR	K-S
건전성	T67	사내유보율	3.11	3.74	37.74	41.99
건전성	T65	순운전자본/총자본 비율	2.94	1.81	33.89	25.16
건전성	T32	금융비용/총부채	2.93	1.46	29.14	22.93
건전성	T108	감가상각비(구성비)	2.91	1.14	24.64	23.37
수익성	T18	자기자본순이익율	2.92	3.58	58.51	50.66
수익성	T25	수지비율	3.02	3.43	52.93	42.19
수익성	T42	성환계수(세전이익)	3.12	3.15	52.04	40.95
수익성	T20	자본금순이익율	2.93	2.96	49.76	42.96
수익성	T47	EBITDA대금융비 용	3.14	2.58	47.13	39.35
수익성	T16	경영자본영업이익 율	2.95	2.91	47.09	39.59
수익성	T44	대출효율성계수(법 인세비용차감전순 이익)	3.20	2.73	44.22	32.93
수익성	T24	매출액영업이익율	2.83	2.50	42.34	39.23
수익성	T34	금융비용/총비용비 율	2.94	2.00	41.61	32.86
수익성	T46	EBITDA대매출액	3.04	2.47	40.79	38.11
수익성	T36	영업활동현금흐름 이자보상비율	2.92	1.85	37.58	30.71
수익성	T23	매출액총이익율	2.84	2.09	37.53	29.62
수익성	T30	조세공과/조세차감 전순이익비율	3.36	1.68	29.29	24.69
활동성	T78	순운전자본회전율	2.68	1.62	40.59	34.12
활동성	T87	매입채무회전율	3.01	2.24	38.62	32.86
활동성	T89	순영업자본회전율	2.89	2.06	35.31	31.18
활동성	T77	자본금회전율	2.87	1.96	34.36	25.96
활동성	T82	채고자산회전율	2.82	1.71	34.14	29.31
활동성	T79	경영자본회전율	2.82	1.84	33.78	26.60

범주	변수 코드	변수명	평균	표준 편차	AR	K-S
활동성	T86	매출채권회전율	3.08	2.42	32.59	26.34
활동성	T83	상(제)품회전율	2.79	1.29	27.20	24.26
활동성	T62	매입채무/재고자산 비율	2.95	1.46	24.02	21.14
생산성	T92	법인세비용차감전순 이익(종업원1인당)	3.06	2.73	50.83	41.86
현 금 흐 름	T70	총C/F대부채비율	2.92	3.31	59.20	46.73
현 금 흐 름	T71	총C/F대차입금비율	3.13	2.97	50.86	37.53
현 금 흐 름	T73	총C/F대매출액비율	2.87	2.84	46.80	40.95
현 금 흐 름	T74	순C/F대차입금비율	2.93	2.11	37.85	31.35

4.4 뉴스 평가항목

본 연구에는 뉴스 평가항목 개발을 3단계로 진행하였다. 첫 번째는 뉴스 데이터를 형태소 분석을 통해 전처리하였다. 두 번째는 형태소 분석을 통해 추출된 명사를 단일 변수화하여 변별력을 측정하였으며, 유의한 키워드를 추출하였다. 세 번째는 추출된 키워드들을 부도와 상관관계를 분석하여 양의 상관관계를 가지는 키워드들을 묶고, 음의 상관관계를 가지는 키워드들을 묶어서 뉴스 항목으로 개발하였다.

4.4.1 뉴스 데이터 전처리

뉴스 데이터는 뉴스 작성일, 뉴스 제목으로 데이터를 구성하였으며, 뉴스 제목은 KoNLPy의 Mecab 형태소 분석기를 통해서 명사만을 별도로 추출하였다. 형태소 분석기를 통해서 전처리된 뉴스 데이터는 [표 4-10]과 같다.

[표 4-10] 뉴스 데이터 전처리 예시

기준 년월일	뉴스 제목	형태소 분석 명사 추출
20160216	작년 영업이익 66억원…흑자전환	작년, 영업이익, 66억원, 흑자전환
20160404	거래소, ○○연구소에 조회 공시 요구	거래소, ○○연구소, 조회 공시, 요구
20160912	10억 규모 자사주 취득 신 탁계약 체결	10억, 규모, 자사주, 취득 신탁계약, 체결

4.4.2 뉴스 데이터 키워드 선별

본 연구에서는 빈도 기반 방법으로 키워드를 선별하였으며, Word2Vec을 활용하여 추가적인 검토를 하였다.

빈도 기반 방법은 3.4.2.2 뉴스 데이터의 전처리 및 키워드 선별에서 설계한 내용과 동일하게 수행하였다. 뉴스 제목은 형태소 분석을 통하여 명사만을 추출하였으며, 총 32,021개의 명사가 추출되었다. 본 연구는 뉴스 제목에서 기업의 경영성과 및 경영위험을 의미하는 키워드를 선별함을 목적으로 하고 있다. 이에 분석대상인 기업명과 동일한 명사 및 한글이 아닌 외국어를 제외하였다. 또한, 빈도수가 20개 미만인 명사는 변별력이 높더라도 일반화 가능성 및 분석 효율이 낮아 모수에서 제외하였다. 이에 변별력 분석에는 총 6,588개 명사를 활용하였다. 변별력 분석은 AR 통계량을 활용하였으며, 변별력 구간별 구성비는 [표 4-11]과 같다. 추출된 명사들의 84.8%는 변별력이 0% 수준으로 변별력이 거의 없었다. AR 통계량이 1.5 이상인 505개 단어에 대해서 해당 단어가 기업의 경영성과 및 경영위험을 의미하는지를 검토하였으며, 부도 관련 키워드 35개 및 기업 성과 관련 키워드 10개를 최종 선별하였으며, 그 결과는 [표 4-12] 및 [표 4-13]과 같다.

[표 4-11] 변별력 구간별 명사 현황

변별력 (AR)	단어 수	구성비
4 이상, 14이하	47	0.7%
3 이상	61	0.9%
2 이상	180	2.7%
1.5 이상	217	3.3%
1 이상	494	7.5%
0 이상	5,589	84.8%
전체	6,588	100%

[표 4-12] 부도 관련 키워드

NO	키워드	변별력(AR)	부도율
1	배정	6.20	5.37%
2	유상증자	5.60	4.00%
3	최대주주	4.97	2.81%
4	유치	4.89	4.41%
5	대표	4.68	2.11%
6	주식	4.38	2.39%
7	선임	3.74	2.02%
8	발행	3.52	2.77%
9	감사의견	3.41	3.73%
10	정상화	3.39	1.06%
11	혐의	3.25	3.80%
12	정지	3.15	2.92%
13	관리종목	3.08	1.16%
14	구속	2.91	9.90%

NO	키워드	변별력(AR)	부도율
15	지분	2.83	2.27%
16	손실	2.78	2.38%
17	거래소	2.59	2.23%
18	조치	2.33	1.93%
19	신규	2.30	1.81%
20	유증	2.29	3.88%
21	전환사채	2.17	3.24%
22	금지	2.16	2.22%
23	불성실	2.03	2.94%
24	매각	2.03	1.61%
25	상한가	1.99	3.52%
26	조사	1.98	0.98%
27	회생	1.97	2.73%
28	배임	1.86	5.26%
29	중속	1.83	3.85%
30	고소	1.79	3.77%
31	폐지	1.76	4.25%
32	폐점	1.65	77.78%
33	수사	1.54	7.75%
34	우려	1.52	0.95%
35	황령	1.51	3.41%

[표 4-13] 기업 성장 관련 키워드

NO	키워드	변별력(AR)	부도율
1	배당	13.95	0.76%
2	현금	11.71	1.30%
3	증가	6.30	1.13%
4	이익	6.17	1.01%
5	기대	3.74	0.74%
6	체결	3.72	0.97%
7	급등	3.08	0.70%
8	전망	2.42	0.13%
9	특허	2.07	0.76%
10	강세	1.74	0.82%

권성일(2016)은 기업활동에서 발견되는 부실징후를 경영자, 종업원, 재무활동, 구매 및 판매활동의 4가지로 구분하였으며, 세부항목으로 16개의 구체적인 징후에 대해서 정리를 하였다. 김권중(2020)은 기업의 도산원인을 기업의 내부요인과 외부요인으로 구분하여 10개 항목으로 정리를 하였다.

선행연구를 바탕으로 뉴스 데이터에서 선별된 키워드들을 부도 관련 5개 항목, 기업 성장 관련 3개 영역으로 구분하였다. 각 영역별 키워드 분류 현황은 [표 4-14]와 같으며, 각 영역별 주요 뉴스는 [표 4-15]와 같다.

[표 4-14] 주요 기업활동 영역별 키워드 분류 현황

구분	대분류	중분류	키워드 수	세부내역
기업 부도	경영자	경영자 교체	7	최대주주, 대표, 주식, 선임, 신규, 금지, 상한가
		경영자 배임	6	혐의, 구속, 배임, 고소, 수사, 횡령
	재무활동	경영자금 부족	9	배정, 유상증자, 유치, 발행, 지분, 유증, 전환사채, 매각, 종속
		부정적 외부감사	8	감사의견, 정상화, 정지, 관리종목, 거래소, 조치, 불성실, 조사
		경영악화	5	손실, 회생, 폐지, 폐점, 우려
기업 성장	재무활동	주주 현금배당	2	배당, 현금
		기업이익 증가	3	증가, 이익, 기대
	경영활동	경영성과 기대	5	체결, 급등, 전망, 특허, 강세

출처: 권성일(2016), 김권중(2020)

[표 4-15] 키워드 분류 영역별 주요 뉴스

구분	대분류	중분류	주요 뉴스
기업 부도	경영자	경영자 교체	“○○, 최대주주 ○○○로 변경” “○○, ○○○ 신입 대표이사 선임”
		경영자 배임	“○○, 대표이사 259억원 횡령·배임 혐의 고소” “○○, ○○○ 전 회장 경찰 출석... 횡령 혐의 부인”
	재무활동	경영자금 부족	“○○, 50억 규모 제3자배정 유상증자 결정” “○○, 외부자본 유치에 의한 매각 추진” “[특징주]○○, 최대주주 변경 소식에 상한가”

구분	대분류	중분류	주요 뉴스
기업 부도	재무 활동	부정적 외부감사	“○○ 반기 ‘감사의견 한정’ 받아” “거래소, ○○ 불성실공시법인지정” “거래소, ○○에 감사의견 비적정설 조회공시 요구”
		경영악화	“금감원, 손실 누장반영 ○○ 감리 나서” “○○, 작년 영업손실 52억원… 적자전환”
기업 성장	재무 활동	주주 현금배당	“○○, 주당 50원 현금배당 결정” “○○, 주당 250원 현금배당 결정”
		기업이익 증가	“○○, 11월 영업이익 33% 증가… 매출도 증가세” “○○, 1분기 영업이익 20억원… 40.7% 증가”
	경영 활동	경영성과 기대	“○○, 애플향 생산 수혜기업과 480억 규모 계약 체결” “○○, 80억 규모 레이저 제조장비 공급 계약 체결”

본 연구에서 부도의 요건으로 정한 키워드인 회생, 파산, 워크아웃, 화의, 법정관리, 자본잠식에 대해 Word2Vec을 통해 의미적으로 유사한 단어를 추출하였다. 추출결과 6개 키워드 중 회생과 자본잠식만 유사한 단어가 추출되었으며, 그 외는 추출이 되지 않았다. Word2Vec을 통한 부도 연관단어의 추출결과가 낮은 것은 부도 요건에 관련된 뉴스 기사는 대부분 해당 사건이 발생한 이후 뉴스에 기사화가 되며, 이는 기부도 요건에 해당되어 본 연구에서는 제외된 것으로 확인되었다. 이에 Word2Vec을 활용한 키워드 추출은 한계가 있어 추가로 연구에 활용하지 않았다. Word2Vec을 활용한 키워드 추출 결과는 [표 4-16]과 같다.

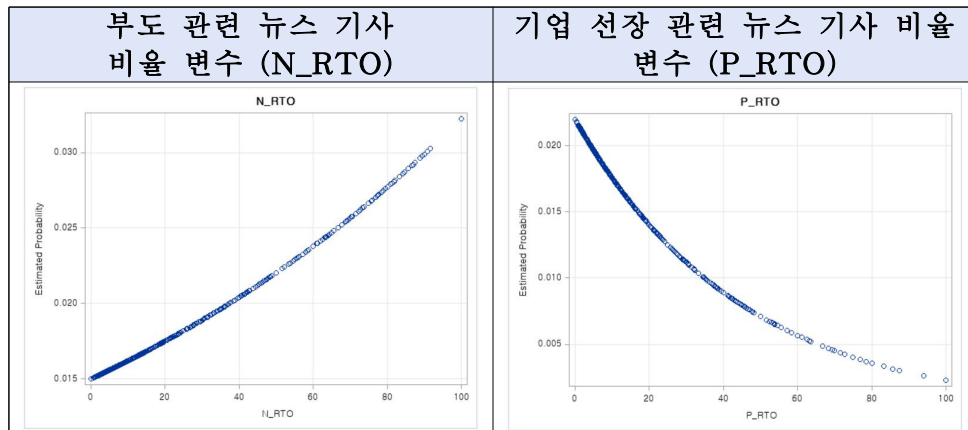
[표 4-16] Word2Vec을 활용한 키워드 추출 결과

부도 키워드	추출 키워드	유사도	변별력(AR)	부도와 상관관계
자본잠식	투자	0.98	3.40%	+
	유상증자	0.98	5.60%	+
회생	발행	0.97	3.52%	+
	결정	0.97	12.06%	-

4.4.3 뉴스 평가항목 개발

뉴스 데이터 분석을 통하여 선별된 부도 관련 키워드 35개와 기업 성장 관련 키워드 10개 통해서 최종 변수를 개발하였다. 3.4절에 정의된 바와 같이 해당 기업의 뉴스에 부도 관련 키워드들이 포함된 뉴스 수를 활용한 부도 관련 뉴스 기사 비율(N_RTO)와 기업 성장 관련 키워드들이 포함된 뉴스 수를 활용한 기업 성장 관련 뉴스 기사 비율(P_RTO)을 개발하였다.

개발된 각 변수에 대해 단순 로지스틱 회귀 모형을 실행하여 추정된 부도확률과 각 변수와의 선형 관계는 [그림 4-2]와 같다. 각 변수의 변별력을 측정하였으며, 변별력은 재무변수대비 낮은 수준으로 산출되었다. 산출결과는 [표 4-17]과 같다.



[그림 4-2] 뉴스 변수와 부도의 선형 관계

[표 4-17] 뉴스 변수의 변별력

구분	재무비율 변수 평균	부도 관련 뉴스 기사 비율 변수 (N_RTO)	기업 성장 관련 뉴스 기사 비율 변수 (P_RTO)
변별력(AR)	49.03	4.61	21.49

4.5 연구모형 실험 결과

본 연구에서는 재무 데이터인 39개의 재무비율 변수와 비정형 데이터인 2개의 뉴스 변수가 활용되었다. 비정형 데이터인 뉴스 데이터의 가치를 실증분석하기 위하여, 재무비율 변수만을 활용한 모형과 재무비율 변수와 뉴스 변수를 활용한 모형을 연구하여 각 모형의 변별력을 비교하였다.

4.5.1 로지스틱 회귀 모형

로지스틱 회귀 모형은 종속변수와 독립변수 간의 함수관계를 분석하는 방법으로, 독립변수가 여러 개인 로지스틱 회귀 모형을 활용하였다.

모형에 활용할 최적의 변수 선택은 모든 변수를 단계적으로 검토하는 단계적 방법을 활용하였다. 재무 변수를 활용한 모형에서는 3개 변수가 선택되었다. 재무 변수와 뉴스 변수를 같이 활용한 경우에는 뉴스 변수는 선택되지 않았으며, 재무 변수만 선택되었다. 4.4절에서 뉴스 변수의 변별력은 재무 변수 대비 낮은 수준을 보임을 확인하였다. 전통적인 통계 모형인 로지스틱 회귀 모형에서는 재무 변수대비 상대적으로 변별력이 낮은 비정형 데이터인 뉴스 변수의 활용가치가 낮음을 실증분석하였다.

로지스틱 회귀 모형의 구성은 [표 4-18]과 같으며, 검증 데이터 및 테스트 데이터의 모형의 성능은 [표 4-19] 및 [표 4-20]과 같이 정리하였다.

[표 4-18] 로지스틱 회귀 모형 구성

변수	재무모형		재무&뉴스모형	
	계수	Wald Chi-Square	계수	Wald Chi-Square
Intercept	-4.931	113.507	-4.931	113.507
T54	0.099	7.213	0.099	7.213
T68	0.126	10.045	0.126	10.045
T73	0.280	22.378	0.280	22.378

[표 4-19] 로지스틱 회귀 모형 성능 (검증 데이터)

구분	평가지표	재무모형	재무&뉴스모형
연속형	AR	75.51	75.51
	K-S	63.58	63.58
범주형	Accuracy	90.99	90.99
	Precision	61.11	61.11
	Recall	30.56	30.56
	F1 Score	40.74	40.74

[표 4-20] 로지스틱 회귀 모형 성능 (테스트 데이터)

구분	평가지표	재무모형	재무&뉴스모형
연속형	AR	76.52	76.52
	K-S	64.64	64.64
범주형	Accuracy	90.96	90.96
	Precision	58.82	58.82
	Recall	28.57	28.57
	F1 Score	38.46	38.46

4.5.2 랜덤 포레스트 모형

본 연구에서는 최적의 모형 도출을 위해서, 랜덤 포레스트 모형을 구성하는 하이퍼 파라미터들에 대해서 초기값을 부여를 하고, 의사결정 나무의 개수를 1개에서 100개까지 시뮬레이션하여 최적의 의사결정 나무의 개수를 선택하였다. 그 이후 의사결정 나무의 개수는 앞서 선택된 값으로 부여하고, 의사결정 나무의 최대 깊이를 2개에서 20개까지 순차적으

로 시뮬레이션하여 최적의 모형을 선택하였다.

각 모형별로 선정된 하이퍼 파라미터는 [표 4-21]과 같으며, 모형의 성능은 검증 데이터는 [표 4-22], 테스트 데이터는 [표 4-23]과 같이 정리하였다.

분석 결과, 검증 및 테스트 데이터에서 AR과 및 Precision은 재무&뉴스모형이 우수한 성능을 보이며, K-S 통계량에서는 재무모형이 우수한 성능을 보였다. 금융감독원(2005)에서는 모형 성능 검증지표로 AR을 제시하고 있음을 고려할 때, 재무&뉴스모형이 우량과 불량을 판단하는 변별력이 좀 더 우수하다고 판단된다.

[표 4-21] 랜덤 포레스트 모형 하이퍼 파라미터

변수	재무모형	재무&뉴스모형
의사결정 나무	41	47
최대 깊이	3	2

[표 4-22] 랜덤 포레스트 모형 성능 (검증 데이터)

구분	평가지표	재무모형	재무&뉴스모형
연속형	AR	76.92	77.26
	K-S	67.84	63.85
범주형	Accuracy	91.55	91.27
	Precision	68.75	69.23
	Recall	30.56	25.00
	F1 Score	42.31	36.73

[표 4-23] 랜덤 포레스트 모형 성능 (테스트 데이터)

구분	평가지표	재무모형	재무&뉴스모형
연속형	AR	77.52	78.97
	K-S	63.66	62.51
범주형	Accuracy	90.96	91.81
	Precision	61.54	80.00
	Recall	22.86	22.86
	F1 Score	33.34	35.56

4.5.3 그래디언트 부스팅 모형

그래디언트 부스팅은 약한 분류기를 결합하여 강한 분류기를 만드는 앙상블모형으로 이전 모형의 예측 오류를 보완하는 형태로 학습을 반복 수행하게 되므로 수치 예측 및 분류 예측에 높은 성능을 가진다.

최적의 모형 추정을 위하여, 하이퍼 파라미터들에 대해서 초기값을 부여하고 의사결정 나무의 개수를 1개에서 100개까지 시뮬레이션하여 최적의 의사결정 나무의 개수를 선택하였다. 그 이후 의사결정 나무의 개수는 앞서 선택된 값으로 부여하고, 의사결정 나무의 최대 깊이를 2개에서 20개까지 순차적으로 시뮬레이션하여 의사결정 나무의 최대 깊이를 선택하였다. 마지막으로 학습률은 0.2에서 0.5까지 시뮬레이션하여 최적의 모형을 선택하였다.

분석 결과, 검증 및 테스트 데이터에서 재무모형대비 재무&뉴스모형이 AR, K-S, Accuracy 및 Precision의 지표에서 우수한 성능을 보였다. 각 모형별로 선정된 하이퍼 파라미터는 [표 4-24]와 같으며, 모형의 성능은 [표 4-25] 및 [표 4-26]과 같이 정리하였다.

[표 4-24] 그레디언트 부스팅 모형 하이퍼 파라미터

변수	재무모형	재무&뉴스모형
의사결정나무 갯수	30	30
최대 깊이	2	2
학습률	0.4	0.4

[표 4-25] 그레디언트 부스팅 모형 성능 (검증 데이터)

구분	평가지표	재무모형	재무&뉴스모형
연속형	AR	69.80	70.55
	K-S	57.81	63.31
범주형	Accuracy	90.70	91.27
	Precision	56.00	59.26
	Recall	38.89	44.44
	F1 Score	45.90	50.79

[표 4-26] 그레디언트 부스팅 모형 성능 (테스트 데이터)

구분	평가지표	재무모형	재무&뉴스모형
연속형	AR	73.64	78.05
	K-S	61.26	65.68
범주형	Accuracy	89.83	90.11
	Precision	47.83	50.00
	Recall	31.43	28.57
	F1 Score	37.93	36.36

4.5.4 심층 신경망 모형

본 연구에서는 심층 신경망 모형은 1개 입력층과 3개의 은닉층 그리고 1개의 출력층으로 구성하였다. 입력층은 재무모형의 경우 39개 재무비율 변수가 입력되도록 하였으며, 재무&뉴스모형은 재무비율 변수 39개와 뉴스 비율 변수 2개를 포함한 41개 변수가 입력되도록 하였다. 은닉층은 총 3개로 구성하였으며, 재무모형은 60개, 39개, 10개의 노드로, 재무&뉴스모형은 60개, 41개, 10개의 노드로 구성하였다. 재무정보를 활용한 심층 신경망 모형의 구조 및 재무&뉴스정보를 활용한 심층 신경망 모형의 구조는 [표 4-27]과 같다. 학습횟수는 10회에서 400회까지 시뮬레이션하여 최적의 모형을 선택하였다.

분석 결과, 검증 및 테스트 데이터에서 재무모형대비 재무&뉴스모형이 AR 지표에서 우수한 성능을 보였다. 각 모형별로 선정된 하이퍼 파라미터는 [표 4-28]과 같으며, 모형의 성능은 [표 4-29] 및 [표 4-30]과 같이 정리하였다.

[표 4-27] 심층 신경망 모형 구조

구분	재무모형 노드수	재무&뉴스모형 노드수
입력층	39	41
은닉층1	60	60
은닉층2	39	41
은닉층3	10	10
출력층	1	1

[표 4-28] 심층 신경망 모형 하이퍼 파라미터

구분	재무모형	재무&뉴스모형
학습횟수(epochs)	43	59

[표 4-29] 심층 신경망 모형 성능 (검증 데이터)

구분	평가지표	재무모형	재무&뉴스모형
연속형	AR	62.21	64.09
	K-S	53.64	52.30
범주형	Accuracy	88.73	90.70
	Precision	41.67	60.00
	Recall	27.78	25.00
	F1 Score	33.34	35.29

[표 4-30] 심층 신경망 모형 성능 (테스트 데이터)

구분	평가지표	재무모형	재무&뉴스모형
연속형	AR	64.64	66.25
	K-S	56.07	59.35
범주형	Accuracy	91.81	89.83
	Precision	68.75	42.86
	Recall	31.43	8.57
	F1 Score	43.14	14.28

4.5.5 각 모형의 성능 비교

위의 연구에서 각 모형별로 검증 및 테스트 데이터에서의 성능을 비교하였다. 각 모형의 성능을 자세하게 비교하기 위하여, 테스트 데이터에서 재무모형 및 재무&뉴스모형에 대해 각 알고리즘의 성능을 살펴보았으며, 그 결과는 [표 4-31]과 같다.

[표 4-31] 각 모형별 성능 비교 (테스트 데이터)

구분	평가지표	재무모형				재무&뉴스모형			
		로지스틱 회귀	랜덤 포레스트	그래디언트 부스팅	심층 신경망	로지스틱 회귀	랜덤 포레스트	그래디언트 부스팅	심층 신경망
연속형	AR	76.52	77.52	73.64	64.64	76.52	78.97	78.05	66.25
	K-S	64.64	63.66	61.26	56.07	64.64	62.51	65.68	59.35
범주형	Accuracy	90.96	90.96	89.83	91.81	90.96	91.81	90.11	89.83
	Precision	58.82	61.54	47.83	68.75	58.82	80.00	50.00	42.86
	Recall	28.57	22.86	31.43	31.43	28.57	22.86	28.57	8.57
	F1 Score	38.46	33.34	37.93	43.14	38.46	35.56	36.36	14.28

재무모형에서는 랜덤 포레스트 모형이 AR이 가장 우수하였다. 로지스틱 회귀 모형은 K-S, 그래디언트 부스팅 모형에서는 Recall, 심층 신경망은 Accuracy, Precision, Recall 그리고 F1-Score가 우수하였다. 심층 신경망 모형이 범주형 분류모형에서는 우수한 성능을 보이고 있으나, 본

연구에서는 AR을 대표 검증 지표로 보고 있으며, 이에 랜덤 포레스트가 가장 우수한 성능을 가지고 있다고 판단된다.

재무&뉴스모형에서는 랜덤 포레스트가 AR, Accuracy 그리고 Precision이 가장 우수하였으며, 로지스틱 회귀 모형은 Recall과 F1-Score가 우수하였다. 그래디언트 부스팅은 K-S, Recall이 우수하였다. 재무&뉴스모형에서도 AR이 가장 높은 포레스트가 가장 우수한 성능을 가지고 있음을 확인하였다.

재무모형과 재무&뉴스모형 전체에 대해서 재무&뉴스모형의 랜덤 포레스트가 가장 성능이 좋았으며, 동일 알고리즘에서도 재무모형보다 재무&뉴스모형이 우수한 성능을 보이고 있다.

뉴스 변수가 변별력이 낮음으로 전통적인 통계 모형인 로지스틱 회귀 모형에서는 활용되지 않았지만, 머신러닝 모형인 랜덤 포레스트, 그래디언트 부스팅 모형 그리고 심층 신경망 모형에서는 뉴스 정보가 변별력을 높이는데 기여하고 있음을 확인하였으며, 기업 신용평가에 있어서 우량과 불량을 구분하는데 그 활용가치가 있음을 확인하였다.

제 5 장 결론 및 향후 연구 방향

5.1 결론

본 연구는 비정형 데이터인 뉴스 데이터를 계량화하여 기업신용평가의 새로운 평가정보 영역으로 활용하는 방법을 제시하였다.

재무정보인 재무비율은 Mini-Modeling 방법을 통해 변환하였으며, 단변량 및 상관관계 분석을 통해 유의한 39개 변수를 연구에 활용하였다.

뉴스 데이터는 뉴스 제목의 형태소 분석을 통해 명사를 추출하였으며, 각 명사에 대해서 단변량 분석을 통해 부도 관련 키워드 35개를 대상으로 하는 부도 관련 뉴스 기사 비율 변수, 기업 성장 관련 키워드 10개를 대상으로 하는 기업 성장 관련 뉴스 기사 비율 변수를 연구에 활용하였다.

실험 결과, 재무 데이터와 뉴스 데이터를 같이 활용한 랜덤 포레스트 모형의 AR 통계량이 78.97%, 정확도인 Accuracy는 91.81%로 높은 성과를 보였으며, 재무정보만 사용했을 경우보다 AR 통계량은 1.45%, Accuracy는 0.85%가 증가하여 뉴스 데이터를 통해 모형의 예측력이 개선되었음을 확인하였다. 본 연구의 실험 결과를 통한 결론은 다음과 같다.

첫째, 본 연구에서는 뉴스 데이터의 키워드 단변량 분석을 통해 기업 부도 관련 뉴스 키워드 35개와 기업 성장 관련 뉴스 키워드 10개가 도출되었다. 경영자의 배임, 교체 등 경영자 리스크 및 전환사채 발행 등의 자금 조달 키워드들은 부도와 깊은 상관관계를 가지고 있음을 확인하였으며, 현금배당, 기업의 이익증가 등의 키워드들은 부도와 음의 상관관계를 가지고 있음을 확인하였다.

둘째, 본 연구에서 뉴스 데이터를 통해 계량화된 뉴스 기사 비율 변수

들은 기존의 재무비율 변수 대비하여 변별력이 낮게 산출되었으며, 각 모형별로 활용되는 수준이 다름을 확인하였다. 전통적인 통계 모형인 로지스틱 회귀 모형에서는 뉴스 비율 변수들이 모형변수로써 선택되지 않았다. 머신러닝 모형인 랜덤 포레스트 모형, 그래디언트 모형 및 심층 신경망에서는 뉴스 비율 변수가 반영됨에 따라 변별력이 1.4%에서 4.4%까지 크게 증가하였다. 이를 통해서 머신러닝 모형에서는 뉴스 기사 비율 변수가 활용가치가 있음을 확인하였다.

세째, 본 연구에서는 랜덤 포레스트 모형이 가장 높은 변별력을 보였다. 재무 데이터만을 활용하였을 경우 및 재무 데이터와 뉴스 데이터를 같이 활용하였을 경우 모두 랜덤 포레스트 모형이 가장 높은 변별력을 보였다. 현행 기업신용평가모형에서는 전통적인 로지스틱 회귀 모형을 주로 활용하고 있으나, 랜덤 포레스트 모형과 같이 머신러닝 모형을 통해 부도 예측에 대한 정확도를 더 높일 수 있음을 확인하였다.

5.2 연구의 의의

본 연구가 가지는 의의는 크게 3가지이다.

첫째, 본 연구에서는 비정형 데이터인 뉴스 데이터의 활용가치를 실증 분석하였다. 비정형 데이터인 뉴스 데이터는 데이터 전처리 등의 어려움으로 재무정보와 달리 기업신용평가모형에서는 새로운 정보영역이었다. 본 연구에서는 뉴스 데이터를 형태소 분석하였으며, 단변량 분석을 통하여 유의한 키워드를 선별하였다. 선별된 키워드를 계량화하였으며, 머신러닝 모형인 랜덤 포레스트 모형, 그래디언트 부스팅 모형 그리고 심층 신경망 모형에서 재무 데이터를 활용하였을 때보다 재무 데이터와 뉴스 데이터를 같이 활용하였을 때 모형의 예측 정확도가 개선됨을 확인하였다.

둘째, 본 연구는 비정형 데이터인 뉴스 데이터의 계량화를 부도 관련 변수와 기업 성장 관련 변수로 구분하여 진행하였다. 기업신용평가에 활용되는 변수들은 부채와 같이 부도와 양의 상관관계를 가지는 변수들도 포함되며, 배당률과 같이 부도와 음의 상관관계를 가지는 변수들도 모두 포함된다. 선행연구들은 뉴스 정보에서 부도 관련 키워드만을 추출하여 기업의 부정적인 정보만을 활용하였으나, 본 연구는 부정적인 정보와 긍정적인 정보를 모두 활용하여 뉴스 데이터의 활용 범위를 확대하였다.

셋째, 본 연구에서는 연구모형으로 전통적인 통계 모형인 로지스틱 회귀 모형과 머신러닝 모형인 랜덤 포레스트, 그래디언트 부스팅 모형 그리고 심층 신경망 모형을 활용하였다. 실험 결과를 통해 로지스틱 회귀 모형보다 머신러닝 모형이 부도에 대한 예측 정확도가 개선되었음을 실증분석하였다. 또한 뉴스 데이터와 같이 변별력이 낮은 변수들은 전통적인 통계 모형인 로지스틱 회귀 모형에는 유의한 변수로 선택되지 않았으나, 머신러닝 모형에서는 뉴스 데이터가 반영되어 변별력이 개선되는 결과를 확인하였다. 이를 통해 머신러닝 모형이 전통적인 통계 모형보다 변별력이 우수하며, 또한 새로운 정보 영역에 대한 모형 활용 가능성이 높음을 확인하였다.

5.3 연구의 한계

비정형 데이터인 뉴스 데이터를 계량화하여 기업신용평가에 활용함에 있어 본 연구는 2가지 한계점을 가지고 있다.

첫째, 본 연구에서는 뉴스 데이터의 형태소 분석을 통해 명사를 추출하여, 부도 및 기업 성장 관련 키워드를 추출하여 연구에 활용하였다. 뉴스의 방향을 하나의 명사만으로 설명하기에는 한계가 있으며, 문맥에 대한 분석을 통해 그 방향성이 명확해진다. 예를 들면 "매출"이라는 키워

드는 "하락"과 "상승"이라는 단어와 결합할 때 좀 더 의미가 명확해진다. 이에 이러한 한계를 개선하기 위하여 Bidirectional Encoder Representations from Transformers(BERT), LSTM 등의 다양한 알고리즘을 활용하여 문맥에 기반한 뉴스 기사 계량화에 대한 연구가 필요하다.

둘째, 본 연구에서는 뉴스 제목을 활용함에 따라 보다 직관적인 모델링을 할 수 있었으나 뉴스 제목이 본문 대비 정보량이 적다는 단점이 있다. 경제 관련 뉴스의 본문에서는 경제 변화의 영향이 특정 산업군 및 기업에 미치는 영향 등 다양한 정보가 포함되어 있다. 비정형 데이터로써 뉴스 정보가 가지는 의미와 활용가치를 높이기 위해서는 뉴스 본문을 활용하여 뉴스 정보의 품질을 높이고 및 거시적인 지표 관련 뉴스를 반영하는 연구가 필요하다.

5.4 향후 연구 방향

본 연구는 뉴스 데이터에서는 부도 및 기업 성장 관련 키워드를 추출하였으며, 이를 재무 데이터와 결합하여 로지스틱 회귀 모형, 랜덤 포레스트 모형, 그래디언트 부스팅 모형 그리고 심층 신경망 모형의 성능 비교를 통해 뉴스 데이터의 가치를 실증분석하였다.

비정형 데이터의 계량화 측면에서, 본 연구에서는 단어의 빈도를 기반으로 하였으나, 단어를 포괄하는 개념인 문장이 가지는 의미를 계량화할 수 있는 다양한 분석 기법을 활용한다면 본 연구보다 뉴스 데이터의 품질을 개선할 수 있을 것으로 기대가 된다. 또한 비정형 데이터의 활용 대상 측면에서, 본 연구는 뉴스 데이터의 제목만을 활용하였으나, 다양한 정보를 포함하는 뉴스 본문을 활용한다면 뉴스 데이터의 변별력이 개선될 것으로 기대가 된다. 기업과 관련한 비정형 데이터에는 뉴스 데이터

이외에도 감사보고서 및 사업 관련 공시자료들이 있다. 뉴스 데이터를 포함한 다양한 비정형 데이터를 계량화하고, 다양한 최신 알고리즘을 활용한 연구가 진행된다면 좀 더 의미 있는 개선이 있을 것으로 기대된다.

기업의 신용평가영역은 신용평가모형을 통한 계량화된 평가 및 예측뿐만 아니라, 여신의 실행 등 다양한 의사결정을 위하여 기업의 재무적인 위험요소와 비재무적인 위험요소를 바탕으로 기업의 종합적인 위험 수준을 분석하고 제공함을 목적으로 한다. 다양한 비정형 데이터를 통해 품질 높은 정보를 추출하여 재무 데이터와 결합하여 기업의 신용도를 예측하는 연구와 더불어 이러한 연구의 목적이 다양한 참여자들이 합리적인 의사결정을 내릴 수 있도록 도와준다는 목적을 잊지 않는다면, 보다 깊이 있는 연구가 될 것으로 기대된다.

참고 문헌

[국내 문헌]

- Géron, A. (2020). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*(박해선, 역). 한빛미디어. (원본 출판 2019)
- Raschka, S., & Mirjalili, V. (2021). *머신 러닝 교과서 with 파이썬, 사이킷런, 텐서플로*(박해선, 역). 길벗. (원본 출판 2019)
- Salimov, S., & 류재홍. (2021). 저해상도 영상 자료를 사용하는 얼굴 표정 인식을 위한 소규모 심층 합성곱 신경망 모델 설계. *한국전자통신학회 논문지*, 16(1), 75-80.
- 강성원 & 강희찬. (2020). Gradient Boosting 모형을 이용한 중소기업 R&D 지원금 결정요인 분석. *한국전자거래학회지*, 25(4), 77-109.
- 권성일. (2016). *최신기업신용분석*(8판). 한국금융연수원.
- 권안나. (2013). *랜덤포레스트를 이용한 변수 선택*. 석사학위논문, 인하대학교 대학원.
- 권철민. (2021). *파이썬 머신러닝 완벽 가이드*. 위키북스.
- 권혁진. (2017). 연결재무제표와 별도재무제표를 활용한 부도예측모형의 비교 연구. *한국회계정보학회 학술대회발표집*, 2017(1), 109-170.
- 금융감독원. (2004). *신BIS 자기자본비율산출기준(안)*. <https://www.fss.or.kr/fss/bbs/B0000092/view.do?nttId=21059&menuNo=200121&cl1Cd=&sdate=&edate=&searchCnd=1&searchWrd=&pageIndex=15>.
- 금융감독원. (2005). *신용리스크 내부등급법 기본 세부지침(안)*. <https://www.fss.or.kr/fss/bbs/B0000092/view.do?nttId=21075&menuNo=200121&cl1Cd=&sdate=&edate=&searchCnd=1&searchWrd=&pageIndex=14>.
- 금융감독원. (2008). *바젤 II 下의 통합리스크관리 모범규준*. <https://www.fss.or.kr/fss/bbs/B0000092/view.do?nttId=21156&menuNo=200121&cl1Cd=&sdate=&edate=&searchCnd=1&searchWrd=&pageIndex=6>.

- 김권중. (2020). *K-IFRS 재무제표분석과 가치평가*(6판). 창민사.
- 김성진 & 안현철. (2016). 기업신용등급 예측을 위한 랜덤 포레스트의 응용. *산업혁신연구*, 32(1), 187-211.
- 김종윤. (2019). *개인신용평가모형(통신스코어) 개발 통신 빅데이터 활용*. 박사학위논문, 숭실대학교 대학원.
- 김찬송. (2018). *부도예측 모형에서 효과적인 감성분석을 위한 뉴스 분류 방법에 관한 연구*. 석사학위논문, 한양대학교 대학원.
- 김형수. (2020). *Step by Step 비즈니스 머신러닝 in 파이썬*. 프레딕스.
- 김혜린. (2020). *부도 데이터의 불균형 문제 해결을 위한 적대적 생성 신경망 (GAN) 기반 오버샘플링 기법*. 석사학위논문, 이화여자대학교 대학원.
- 민성환. (2014). 개선된 배깅 앙상블을 활용한 기업부도예측. *지능정보연구*, 20(4), 121-139.
- 박대서 & 김화중. (2018). TF-IDF 기반 키워드 추출에서의 의미적 요소 반영을 위한 결합벡터 제안. *한국정보기술학회논문지*, 16(2), 1-16.
- 박종원 & 안성만. (2014). 재무비율을 이용한 부도예측에 대한 연구: 한국의 외부감사대상기업을 대상으로. *경영학연구*, 43(3), 639-669.
- 서정구 & 김확열. (2018). 재무건전성 모형의 유용성에 관한 연구: 분식회계 기업을 대상으로. *윤리경영연구*, 18(1), 61-80.
- 송충석. (2020). *송충석의 재무제표 바로읽기*. 세학사.
- 윤동희. (2013년04월02일). 부산銀 신용평가모형 재구축...내부등급법 '성큼'. *The Bell*. <https://www.thebell.co.kr/free/content/ArticleView.asp?key=201304020100003090000177&lcode=00>.
- 이기창. (2019). *한국어 임베딩*. 에이콘출판.
- 이성훈 & 이동우. (2013). 빅데이터의 국내·외 활용 고찰 및 시사점. *디지털융복합연구*, 11(2), 229-233.
- 이시영. (2021). *뉴스 기사 텍스트 임베딩을 이용한 딥러닝 기반 기업성과*

- 예측 모델 연구. 박사학위논문, 숭실대학교 대학원.
- 이인로 & 김동철. (2015). 회계정보와 시장정보를 이용한 부도예측모형의 평가 연구. *채무연구*, 28(4), 625-665.
- 이재성. (2016). 심층 신경망의 발전 과정과 이해. *정보와 통신*, 33(10), 40-48.
- 임희석 & 고려대학교 자연어처리연구실. (2019). *자연어처리 바이블*. 휴먼싸이언스.
- 장영재, 손원, & 황희진. (2020). *비정형 데이터 분석*. 한국방송통신대학교출판문화원.
- 전창욱, 최태균, 조중현, & 신성진. (2022). *텐서플로2와 머신러닝으로 시작하는 자연어처리*(2판). 위키북스.
- 조경인 & 김영민. (2021). 통계적 학습을 이용한 다시점 기업부도 예측모형들의 비교. *한국데이터정보과학회지*, 32(3), 487-499.
- 조남옥 & 신경식. (2016). Bankruptcy prediction modeling using qualitative information based on big data analytics. *지능정보연구*, 22(2), 33-56.
- 조단비, 이현영, 박지훈, & 강승식. (2020). 형태소 임베딩과 SVM 을 이용한 뉴스 기사 정치적 편향성의 자동 분류. *한국정보처리학회 학술대회 논문집*, 27(1), 451-454.
- 조성빈. (2020). 기업도산예측에 대한 의사결정나무 앙상블 모델 평가. *한국경영공학회지*, 25(4), 63-71.
- 조승현, 김연희, 임웅, 김휘용, & 최진수. (2018). 딥 러닝 기반의 이미지와 비디오 압축 기술 분석. *방송공학회논문지*, 23(3), 383-394.
- 조영임. (2013). 빅데이터의 이해와 주요 이슈들. *한국지역정보화학회지*, 16(3), 43-65.
- 조용준. (2018). *빅데이터 SPSS 최신 분석기법*. 한나래.
- 지승은 & 김우일. (2017). 효과적인 음성 인식 평가를 위한 심층 신경망 기반의 음성 인식 성능 지표. *한국정보통신학회논문지*, 21(12), 2291-2297.

- 천인국. (2020). *인공지능 : 파이썬으로 배우는 머신러닝과 딥러닝*. 인피니티 북스.
- 최소운 & 안현철. (2015). 퍼지이론과 SVM 결합을 통한 기업부도예측 최적화. *디지털융복합연구*, 13(3), 155-165.
- 최요셉 & 최용석. (2014). 맵리듀스와 대응분석을 활용한 비정형 빅 데이터의 정형화와 시각적 해석. *응용통계연구*, 27(2), 169-183.
- 최정원, 오세경, & 장재원. (2017). 빅데이터와 인공지능 기법을 이용한 기업부도예측 연구. *한국재무학회 학술대회*, 2017(11), 396-435.
- 최정원. (2019). *인공지능을 이용한 뉴스 정보 기반의 기업 부도예측 연구*. 박사학위논문, 건국대학교 대학원.
- 통계청. (2017). *한국표준산업분류*. 경제서적.
- 한국신용정보원. (2018). *기업 신용평가모형의 현황과 변화 트렌드*. (CIS이슈리포트 2018-6호).https://www.kcredit.or.kr:1441/archive/cisReport.do?_csrf=e0bf4ecf-466a-4ecc-9967-c2fccafb4a0e&menuNo=420&hpBoardSn=CIS_REPORT&link=archive%2FcisReport.do.
- 한은정. (2005). *건강검진 자료에서 Random forests 를 이용한 백내장 발생 위험군 예측모형*. 석사학위논문, 연세대학교 대학원.
- 홍동숙, 백한중, & 신현준. (2021). 개인사업자 부도율 예측 모델에서 신용정보 특성 선택 방법. *한국시물레이션학회 논문지*, 30(1), 75-85.
- 홍세희. (2005). *이항 및 다항 로지스틱 회귀분석*. 교육과학사.

[국외 문헌]

- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4), 589-609.
- Altman, E. I., Haldeman, R. G., & Narayanan, P. (1977). ZETA analysis : A new model to identify bankruptcy risk of corporations. *The journal of banking & finance*, 1(1), 29-54.
- Beaver, W. H. (1966). Financial ratios as predictors of failure. *The journal of Accounting Research*, 4(3), 71-111.
- Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. *Advances in neural information processing systems*, 1(13), 932-938.
- Breiman, L. (1997). *Arcing the edge*. Technical Report 486, Statistics Department, University of California at Berkeley, https://scholar.google.com/scholar_lookup?title=Arcing%20the%20edge&publication_year=1997&author=Breiman%2CL.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *The journal of Finance*, 63(6), 2899-2939.
- Chang, Y. C., Chang, K. H., & Wu, G. J. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing*, 73, 914-920.
- Cohen, R. A. (1999). An introduction to PROC LOESS for local regression. *SUGI*, 24, 1584-1592.
- Cornfield, J., Gordon, T., & Smith, W. W. (1961). Quantal response curves for experimentally uncontrolled variables. *Bull Int Stat Inst*,

38(3), 97–115.

- Cox, D. R. (1958). The regression analysis of binary sequences. *The journal of the Royal Statistical Society: Series B (Methodological)*, 20(2), 215–232.
- Dharwadkar, N. V., & Patil, P. S. (2018). Customer retention and credit risk analysis using ANN, SVM and DNN. *International Journal of Society Systems Science*, 10(4), 316–332.
- Falkenstein, E. G., Boral, A., & Carty, L. V. (2000). RiskCalc for private companies: Moody's default model. *Moody's Risk Management Services Global Credit Research*.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 29(5), 1189–1232.
- Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine learning*, 77(1), 103–123.
- Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7), 1527–1554.
- Huynh, H. D., Dang, L. M., & Duong, D. (2017). A new model for stock price movements prediction using deep neural network. *In Proceedings of the Eighth International Symposium on Information and Communication Technology*, 57–62.
- Jatnika, D., Bijaksana, M. A., & Suryani, A. A. (2019). Word2vec model analysis for semantic similarities in english words. *Procedia Computer Science*, 157, 160–167.
- Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, 104, 38–48.

- Le, T., Vo, B., Fujita, H., Nguyen, N. T., & Baik, S. W. (2019). A fast and accurate approach for bankruptcy forecasting using squared logistics loss with GPU-based extreme gradient boosting. *Information Sciences*, 494, 294–310.
- Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, 234, 11–26.
- Liu, Y., Zeng, Q., Yang, H., & Carrio, A. (2018). Stock price movement prediction from financial news with deep learning and knowledge graph embedding. *In Pacific rim knowledge acquisition workshop*, 2018, 102–113.
- Lu, Y. C., Shen, C. H., & Wei, Y. C. (2013). Revisiting early warning signals of corporate credit default using linguistic analysis. *Pacific-Basin Finance Journal*, 24, 1–21.
- Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European journal of operational research*, 274(2), 743–758.
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4), 115–133.
- Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The journal of finance*, 29(2), 449–470.
- Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *The journal of accounting research*, 18(1), 109–131.
- SAS Institute. (2018). *SAS/STAT 15.1 User's Guide*. https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/statug/titlepage.htm.

- Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The journal of business*, 74(1), 101-124.
- Taha, A. A., & Malebary, S. J. (2020). An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. *IEEE Access*, 8, 25579-25587.
- Tian, Z., Xiao, J., Feng, H., & Wei, Y. (2020). Credit risk assessment based on gradient boosting decision tree. *Procedia Computer Science*, 174, 150-160.
- Walker, S. H., & Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1-2), 167-179.
- Werbos, P. (1974). *Beyond regression: new tools for prediction and analysis in the behavioral sciences*. Ph. D. dissertation, Harvard University.
- Zhang, X., Yang, Y., & Zhou, Z. (2018). A novel credit scoring model based on optimized random forest. *In 2018 IEEE 8th annual computing and communication workshop and conference*, 60-65.

부 록

부록. 전체 변환 재무비율의 기술 통계량

N o	범주	변수 코드	변수명	평균	표준 편차	AR	K-S
1	건전성	T68	유보액/총자산비율	3.05	4.76	74.42	61.10
2	건전성	T48	자기자본비율	3.10	5.03	74.24	59.69
3	건전성	T54	부채비율	3.12	4.58	70.75	58.53
4	건전성	T55	유동부채비율	3.13	4.50	67.80	56.47
5	건전성	T58	차입금/자기자본	3.14	4.39	67.03	51.03
6	건전성	T69	유보액/납입자본비율	3.06	4.17	62.51	48.55
7	건전성	T57	차입금의존도	3.00	3.59	57.61	44.66
8	건전성	T52	비유동자산비율	3.08	2.99	54.76	43.60
9	건전성	T56	비유동부채비율	2.90	2.81	50.02	43.98
10	건전성	T53	비유동자산장기적합율	3.06	2.55	48.46	39.19
11	건전성	T51	현금비율	2.89	2.43	46.67	36.37
12	건전성	T50	당좌비율	2.91	2.37	43.97	33.31
13	건전성	T35	금융비용/매출액비율	3.05	2.00	41.36	35.79
14	건전성	T60	매출채권/매입채무비율	2.94	2.14	39.00	33.33
15	건전성	T67	사내유보율	3.11	3.74	37.74	41.99
16	건전성	T59	차입금/매출액비율	2.95	1.72	37.63	30.90
17	건전성	T65	순운전자본/총자본비율	2.94	1.81	33.89	25.16
18	건전성	T63	재고자산/순운전자본비율	3.00	1.48	32.14	27.61
19	건전성	T49	유동비율	2.88	1.56	31.46	27.17
20	건전성	T64	비유동부채/순운전자본비율	2.92	1.35	31.22	27.63
21	건전성	T32	금융비용/총부채	2.93	1.46	29.14	22.93
22	건전성	T107	조세공과(구성비)	2.93	1.09	24.80	23.08
23	건전성	T108	감가상각비(구성비)	2.91	1.14	24.64	23.37

N o	범주	변수 코드	변수명	평균	표준 편차	AR	K-S
24	건전성	T105	금융비용(구성비)	2.73	0.97	24.61	20.49
25	건전성	T104	인건비(구성비)	2.91	1.08	23.41	22.52
26	건전성	T106	임차료(구성비)	2.91	0.92	19.81	18.81
27	수익성	T66	적립금비율	3.04	4.06	61.46	48.38
28	수익성	T18	자기자본순이익율	2.92	3.58	58.51	50.66
29	수익성	T17	자기자본법인세비용 차감전순이익율	3.01	3.61	57.98	49.61
30	수익성	T13	총자본순이익율	2.97	3.26	53.03	43.53
31	수익성	T25	수지비율	3.02	3.43	52.93	42.19
32	수익성	T12	총자본법인세비용 차감전순이익율	2.96	3.26	52.82	43.40
33	수익성	T42	성환계수(세전이익)	3.12	3.15	52.04	40.95
34	수익성	T38	(구)경상이익이자보 상비율	2.99	2.79	51.83	41.23
35	수익성	T39	법인세차감전순이 익이자보상비율	2.99	2.79	51.83	41.23
36	수익성	T43	부채상환계수	3.10	3.10	51.61	40.45
37	수익성	T22	매출액순이익율	2.77	2.75	50.11	43.81
38	수익성	T37	영업이익이자보상 비율	2.96	2.64	50.01	39.16
39	수익성	T19	자본금법인세비용 차감전순이익율	2.93	2.97	49.84	42.43
40	수익성	T21	매출액법인세비용 차감전순이익율	2.81	2.77	49.79	43.97
41	수익성	T20	자본금순이익율	2.93	2.96	49.76	42.96
42	수익성	T14	기업법인세비용차 감전순이익율	2.97	3.01	47.57	40.24
43	수익성	T47	EBITDA대금융비 율	3.14	2.58	47.13	39.35
44	수익성	T16	경영자본영업이익 율	2.95	2.91	47.09	39.59
45	수익성	T15	기업순이익율	2.97	3.01	47.00	40.54
46	수익성	T11	총자본영업이익율	2.96	2.96	46.84	39.54

N o	범주	변수 코드	변수명	평균	표준 편차	AR	K-S
47	수익성	T44	대출효율성계수(법 인세비용차감전순 이익)	3.20	2.73	44.22	32.93
48	수익성	T45	EBIT대매출액(세 전이익)	2.82	2.63	43.55	40.35
49	수익성	T24	매출액영업이익율	2.83	2.50	42.34	39.23
50	수익성	T34	금융비용/총비용비 율	2.94	2.00	41.61	32.86
51	수익성	T46	EBITDA대매출액	3.04	2.47	40.79	38.11
52	수익성	T36	영업활동현금흐름 이자보상비율	2.92	1.85	37.58	30.71
53	수익성	T23	매출액총이익율	2.84	2.09	37.53	29.62
54	수익성	T26	매출원가율	3.09	2.14	36.34	30.48
55	수익성	T30	조세공과/조세차감 전순이익비율	3.36	1.68	29.29	24.69
56	수익성	T27	감가상각율	2.92	1.67	28.44	23.45
57	수익성	T33	차입금평균이자율	2.89	1.18	25.93	21.60
58	수익성	T103	법인세차감전순이 익(구성비)	3.44	1.48	25.47	25.64
59	수익성	T41	배당성향	2.89	1.13	17.77	17.78
60	수익성	T28	감가상각비/총비용 비율	2.86	0.85	17.65	16.40
61	수익성	T31	조세공과/총비용비 율	2.83	0.78	16.24	14.66
62	수익성	T40	배당율	2.89	1.04	15.85	15.85
63	수익성	T29	인건비/총비용비율	2.84	0.75	15.75	13.79
64	활동성	T76	자기자본회전율	2.95	2.51	48.65	37.73
65	활동성	T78	순운전자본회전율	2.68	1.62	40.59	34.12
66	활동성	T87	매입채무회전율	3.01	2.24	38.62	32.86
67	활동성	T89	순영업자본회전율	2.89	2.06	35.31	31.18
68	활동성	T77	자본금회전율	2.87	1.96	34.36	25.96
69	활동성	T82	재고자산회전율1	2.82	1.71	34.14	29.31

N o	범주	변수 코드	변수명	평균	표준 편차	AR	K-S
70	활동성	T79	경영자본회전율	2.82	1.84	33.78	26.60
71	활동성	T75	총자본회전율	2.77	1.67	32.83	24.77
72	활동성	T86	매출채권회전율	3.08	2.42	32.59	26.34
73	활동성	T88	재고자산회전율2	2.87	1.68	32.17	26.77
74	활동성	T81	유형자산회전율	2.87	1.64	31.22	23.90
75	활동성	T84	원재료회전율	2.87	1.52	30.14	25.75
76	활동성	T80	비유동자산회전율	2.82	1.35	29.31	22.61
77	활동성	T83	상(제)품회전율	2.79	1.29	27.20	24.26
78	활동성	T61	매출채권/상, 제품비율	2.81	1.27	26.37	23.65
79	활동성	T62	매입채무/재고자산비율	2.95	1.46	24.02	21.14
80	활동성	T85	재공품회전율	2.76	0.82	19.42	14.67
81	생산성	T92	법인세비용차감전 순이익(종업원1인당)	3.06	2.73	50.83	41.86
82	생산성	T93	순이익(종업원1인당)	2.94	2.64	50.73	41.94
83	생산성	T91	매출액(종업원1인당)	3.11	2.00	28.84	24.72
84	생산성	T95	노동장비율(종업원1인당)	2.98	1.70	27.33	21.16
85	생산성	T98	총자본투자효율	2.80	1.26	26.91	23.89
86	생산성	T96	기계장비율(종업원1인당)	2.93	1.32	26.68	24.03
87	생산성	T97	자본집약도(종업원1인당)	3.06	1.53	24.36	22.41
88	생산성	T99	설비투자효율	2.87	1.12	23.68	22.16
89	생산성	T102	노동소득분배율	2.91	1.08	23.41	22.52
90	생산성	T100	기계투자효율	2.85	1.19	23.04	22.70
91	생산성	T101	부가가치율	2.85	1.10	22.31	20.42

No	범주	변수 코드	변수명	평균	표준 편차	AR	K-S
92	생산성	T90	부가가치(종업원1인당)	2.85	1.04	20.24	20.04
93	생산성	T94	인건비(종업원1인당)	2.84	0.84	16.03	14.85
94	성장성	T5	자기자본증가율	2.91	3.07	51.06	43.70
95	성장성	T8	법인세비용차감전 순이익증가율	2.37	1.35	47.34	46.49
96	성장성	T9	순이익증가율	2.34	1.33	47.23	46.49
97	성장성	T7	영업이익증가율	2.42	1.16	42.74	41.32
98	성장성	T1	총자산증가율	2.99	1.89	35.93	28.21
99	성장성	T6	매출액증가율	2.84	1.38	34.78	32.59
100	성장성	T3	유동자산증가율	2.97	1.64	30.94	24.23
101	성장성	T10	종업원수증가율	2.59	0.74	23.09	21.92
102	성장성	T2	유형자산증가율	2.80	0.83	22.03	23.69
103	성장성	T4	채고자산증가율	2.83	0.63	18.53	16.48
104	현금흐름	T70	총C/F대부채비율	2.92	3.31	59.20	46.73
105	현금흐름	T72	총C/F대총자본비율	2.98	3.31	52.93	42.70
106	현금흐름	T71	총C/F대차입금비율	3.13	2.97	50.86	37.53
107	현금흐름	T73	총C/F대매출액비율	2.87	2.84	46.80	40.95
108	현금흐름	T74	순C/F대차입금비율	2.93	2.11	37.85	31.35