

딥러닝 시계열 알고리즘 적용한 기업부도예측모형 유용성 검증*

차성재

(주)에이젠글로벌, 연구원
(sungjae.cha@aizen.co)

강정석

(주)에이젠글로벌, 대표이사
(js.kang@aizen.co)

본 연구는 경제적으로 국내에 큰 영향을 주었던 글로벌 금융위기를 기반으로 총 10년의 연간 기업데이터를 이용한다. 먼저 시대 변화 흐름에 일관성있는 부도 모형을 구축하는 것을 목표로 금융위기 이전(2000~2006년)의 데이터를 학습한다. 이후 매개 변수 튜닝을 통해 금융위기 기간이 포함(2007~2008년)된 유효성 검증 데이터가 학습데이터의 결과와 비슷한 양상을 보이고, 우수한 예측력을 가지도록 조정한다. 이후 학습 및 유효성 검증 데이터를 통합(2000~2008년)하여 유효성 검증 때와 같은 매개변수를 적용하여 모형을 재구축하고, 결과적으로 최종 학습된 모형을 기반으로 시험 데이터(2009년) 결과를 바탕으로 딥러닝 시계열 알고리즘 기반의 기업부도 예측 모형이 유용함을 검증한다.

부도에 대한 정의는 Lee(2015) 연구와 동일하게 기업의 상장폐지 사유들 중 실적이 부진했던 경우를 부도로 선정한다. 독립변수의 경우, 기존 선행연구에서 이용되었던 재무비율 변수를 비롯한 기타 재무정보를 포함한다. 이후 최적의 변수군을 선별하는 방식으로 다변량 판별분석, 로짓 모형, 그리고 Lasso 회귀분석 모형을 이용한다. 기업부도예측 모형 방법론으로는 Altman(1968)이 제시했던 다중판별분석 모형, Ohlson(1980)이 제시한 로짓 모형, 그리고 비시계열 기계학습 기반 부도예측모형과 딥러닝 시계열 알고리즘을 이용한다.

기업 데이터의 경우, ‘비선형적인 변수들’, 변수들의 ‘다중 공선성 문제’, 그리고 ‘데이터 수 부족’이란 한계점이 존재한다. 이에 로짓 모형은 ‘비선형성’을, Lasso 회귀분석 모형은 ‘다중 공선성 문제’를 해결하고, 가변적인 데이터 생성 방식을 이용하는 딥러닝 시계열 알고리즘을 접목함으로써 데이터 수가 부족한 점을 보완하여 연구를 진행한다.

현 정부를 비롯한 해외 정부에서는 4차 산업혁명을 통해 국가 및 사회의 시스템, 일상생활 전반을 아우르기 위해 힘쓰고 있다. 즉, 현재는 다양한 산업에 이르러 빅데이터를 이용한 딥러닝 연구가 활발히 진행되고 있지만, 금융 산업을 위한 연구분야는 아직도 미비하다. 따라서 이 연구는 기업 부도에 관하여 딥러닝 시계열 알고리즘 분석을 진행한 초기 논문으로서, 금융 데이터와 딥러닝 시계열 알고리즘을 접목한 연구를 시작하는 비 전문가에게 비교분석 자료로 쓰이기를 바란다.

주제어 : 최적 변수 선별, Lasso 회귀분석, 딥러닝 시계열 알고리즘, 기업부도, RNN, LSTM

논문접수일 : 2018년 10월 24일 논문수정일 : 2018년 11월 16일 게재확정일 : 2018년 11월 26일

원고유형 : 일반논문(급행) 교신저자 : 강정석

* 이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.2017-0-01013, 딥러닝 LSTM을 활용한 개인데이터의 시계열 상태변화 패턴을 예측하는 모형개발)

1. 서론

1.1 연구배경

기업의 부도는 해당 부도기업의 경영자, 종업원, 채권자, 투자자를 비롯한 이해관계자들 이외에도 지역경제, 국가경제까지 파급효과를 미친다. 아시아 외환위기 발생 전에는 중소기업만을 대상으로 분석을 진행하였고, 다양한 방식의 부도 모형 개발이 아닌, 계량분석 모형 위주로 부도예측모형의 예측력을 높이고자 하였다. 그 결과, 소위 재벌기업이라 하는 대기업까지도 부도로 이어지게 되었다. 이외에도 과거 기업 부도에 대한 연구는 특정 변수를 중심으로 분석이 진행되어 왔다. 그리고 정부 또한 글로벌 금융위기 발생 직후 기업 구조조정을 진행함에 있어 ‘부채비율’과 같은 주요 변수만을 중심으로 구조조정을 진행하였다. 글로벌 금융위기에서 Swedberg(2009)가 제공한 ‘리만 브라더스 사례’처럼 다양한 이해관계가 한 순간에 쉽게 무너지지 않기 위해서는 기업부도 예측 모형에 대한 다각적인 연구가 필수적이다.

기업부도 예측에 사용되는 주요 변수들은 시간에 따라 변화된다. 이는 Deakins(1972)의 연구를 통해 기업의 실패에 주요한 영향을 미치는 요소가 변화됨을 Beaver(1967,1968)와 Altman(1968)의 분석 방법을 통해 확인하였다. 이후 Grice(2001)의 연구에서 또한 예측 변수의 중요도가 Zmijewski(1984) 및 Ohlson(1980) 모델을 통해 시기에 따라 변하고 있음이 다시 재확인되었다. 즉, 과거 정적 모형으로 진행되어오던 연구들은 시간에 흐름에 따라 변화되고 있는 부분을 고려하지 않고 편향되어 있는 경우가 대부분이다. 따라서 일관성있는 기업부도 예측모형

구축을 위해서는 동적 변화를 반영한 딥러닝 시계열 알고리즘을 통해 시간에 따른 편향을 보완하는 것이 필요하다.

한국 정부는 2017년 4차산업혁명의 근간이 되는 인공지능·ICT 등 핵심기술 확보하고, 신산업 및 새로운 서비스를 육성하기 위해 대통령 직속 산하 4차산업혁명위원회를 설립했다. ‘Data Industry Promotion Strategy - I-KOREA 4.0 Data Field Plan, I-DATA+’(2018)이란 보고서에 따르면, ‘데이터’는 4차 산업혁명을 견인하는 핵심 동인이며, 빅데이터를 통해 사회문제 해결능력을 강화하는 것을 핵심과제로 정하였다. 또한 2018년 3월 13일 4차산업혁명위원회의 첫 회의에서 문재인 대통령은 인공지능, 사물인터넷, 빅데이터를 위한 투자를 확대하여 혁신생태계를 조성할 것임을 밝히면서 금융 분야에서의 빅데이터 분석에 관한 연구의 필요성이 높아지고 있다.

해외 주요 국가들을 살펴보면 ‘The Fourth Industrial Revolution in Major Countries and Growth Strategy of Korea: U.S., Germany and Japan Cases’(2018)라는 Korea Institute for International Economic Policy의 Kim et al.(2017)의 보고서를 통해 4차 산업혁명을 앞장서서 준비하고 있음을 확인할 수 있었다. 독일의 경우 2013년부터 플랫폼 인더스트리 4.0이라는 산관학 협력기구를 설립하여 비교경쟁우위를 갖는 제조업에 IT를 접목하는데 힘쓰고 있다. 미국은 2014년 Industrial Internet Consortium을 필두로 기술혁신, 창업생태계 조성, 제조업 혁신을 위한 각종 지원책을 시행하고 있다. 그리고 일본 또한 2016년 인공지능기술전략회의를 설치하고, 4차 산업혁명을 이끌 미래성장동력 연구에 한발 앞서서 진행 중이다.

미래선도기술인 빅데이터 기술은 분석을 비롯

하여 인공지능을 넘어 초연결지능화의 방향으로 향하고 있다. 아직 시계열 알고리즘을 통한 기업부도 예측모형 연구는 초기단계임에도 불구하고, 기업부도 예측모형 구축시, 딥러닝 모형이 과거 회귀분석 모형을 이용할 때에 비해 시간을 더욱 단축된다. 또한 예측력 면에서 더욱 효과적이다. 따라서, 다각적인 분석이 가능한 딥러닝 기반의 시계열 알고리즘을 통해 기업부도 예측모형의 효과성을 높이는 것이 필요해졌다.

1.2 연구내용

본 연구의 내용은 2가지 단계로 구성되어 딥러닝 시계열 알고리즘의 유용성을 검증한다. 데이터는 기존 논문들에 적용되었고, 국내 금융감독기관에서 기업구조조정을 위해 연구를 하였던 주요 변수들이 포함된 약 196개 재무변수이다. 먼저 3가지 최적 변수 선정 알고리즘에 의해 변수 분석을 진행하고, 분석에 용이한 최적 변수들을 추출한다. 이후 3가지 변수군을 바탕으로 기존 score모형, 비시계열 기계학습 알고리즘, 그리고 최종적으로 딥러닝 시계열 알고리즘까지 총 9개의 모형을 만들고 딥러닝 시계열 알고리즘 기반 부도 예측 모형과 다른 부도 예측 모형들의 결과를 비교하였다.

최적 변수를 선정하는데 적용된 알고리즘은 다음과 같다.

- 다중판별분석: 설명변수들이 정규분포를 따르며, 집단간 분산(공분산)이 동일하다는 가정을 바탕으로 하는 분석방법이다. 다중공선성 문제 감소 및 종속변수와의 상관성을 고려한다.
- 로지스틱 회귀분석: 로짓모형 기반 비선형 회귀분석으로서, 비선형 변수의 특징을 보

이는 경우 적용되는 방법이다.

- Lasso 회귀분석: 선형 회귀분석에서 다중공선성 문제를 대표적으로 보완해주는 방법으로 머신러닝 혹은 최근 금융산업에서 차원 축소 및 최적 변수 선정에 주로 이용되는 방법이다.

최적 변수 선정 이후 가변적인 데이터 생성법을 이용해 RNN과 LSTM 시계열 모형을 구축하였다. 시계열 분석이 진행되지 않았던 기존의 모형들에 비해 표본내 데이터(Insample)로 학습 및 검증시 예측력이 우수함을 비교를 통해 확인한다. 표본외 데이터(Out-of-sample)에 적용한 결과 또한 RNN 및 LSTM 모형으로 기업부도모형을 구축하는 것이 기존 모형들보다 안정성 있는 예측력을 보여줌을 확인한다.

1.3 문헌연구

과거 선행연구에 따르면 Altman(1968)은 다변량 판별분석을 통해 Z-score 모형을 제안하였다. 이는 1970년대 부도예측 및 신용평가모형으로 주로 이용되어왔다. 이후 다변량 판별분석에서는 설명변수들이 정규분포를 따라야 한다는 가정에서 일관성을 가지지 못하는 문제점이 발생했다. Ohlson(1980)은 종속변수가 기업부도와 같이 0과 1로 표현되는 이항변수로 표현될 때 이용 가능한 로지스틱 회귀분석(로짓모형)을 제안하여 정규분포성 문제를 해결하려 했다. 선택 확률은 로지스틱 함수를 따른다는 가정이 필요하며, 이는 실무에서 주로 이용되어지고 있다.

인공지능(딥러닝) 기법은 비교적 최신의 방법론으로서 금융 및 재무 분야에서는 전통적인 방법론에 의한 예측 방법론에 비하여 연구의 양과 질 모두 부족하지만 최근 연구가 매우 급속하게

증가하는 추세이다. Lee(1993)은 인공지능경망 기반 부도예측모형이 높은 예측력을 가짐을 확인하였다. Lee et al.(1995)은 기존의 재무정보만 활용한 부도예측의 한계가 있음을 지적, 비재무정보를 활용한 인공지능경망 기반의 부도예측 모형을 제시하였다. Kim(2009, 2010, 2012)은 GM-Boost(Geometric Mean-based Boosting) 및 MGM-Boost(Multiclass GM-Boost) 알고리즘을 통해 기업부실 예측 데이터의 불균형 정도에 관계없이 견고한 학습성능을 보임을 확인하였다. Kim et al.(2010)은 SVM 기법으로 부도예측을 수행, 정보가 상대적으로 부족한 중소기업의 경우 기존의 방법론보다 인공지능 기법의 예측 성능이 더 우수함을 실증하였다. Bac(2010)은 Voting 알고리즘과 인공지능경망을 통합한 알고리즘 기법이 부도 예측에 우수함을 실증하였다. Park(2014)은 외부감사대상기업을 대상으로 다변량 로짓분석을 적용하면 유용함을 확인하였다. Kim et al.(2014)은 랜덤 포레스트를 활용한 기업채권평가 모형의 예측력이 우수함을 확인하였다. Wang(2015)은 Lasso-logistic regression learning ensemble 기법으로 신용위험을 평가하는 것이 효과적임을 실증하였다. Kim et al.(2016)은 금융기관의 기업신용등급 예측 과정에 랜덤 포레스트 방법을 적용한 바 있다. Yeh et al.(2015)은 Deep Belief Networks(DBN)이 SVM보다 부도예측 성능이 우수함을 실증하였다. Min(2014, 2016)은 Bagging Ensemble 기법과 K-Nearest Neighbors (KNN) 알고리즘이 부도예측에 우수함을 확인하였다. Jo et al.(2015, 2016)은 하이브리드 인공지능경망 기법을 적용하여 부도 유형을 예측하는 연구를 진행하였다. Addal(2016)은 인공지능경망, K-Nearest Neighbors(KNN) 등의 방법론을 이용하여 기업부도예측 모형이 우수한 예측력을 보임을 실증하

였다.

변수 선별방식에서 이용된 Lasso 회귀분석은 선형회귀분석에서 다중공선성을 감안한 알고리즘으로 알려져 있으며 차원 축소에 대표적으로 이용되는 알고리즘이다. Kapinos & Mitnik(2015)는 연구를 통해 Lasso 회귀분석에 의해 은행 성과지표를 거시경제지표로 이용하는 것이 효과적임을 밝혔으며, Perdeiy(2009)의 연구에서는 전통적이지 않은 재무지표들을 공변량으로 사용하여 파산을 예측하는데 이용하였다

2. 연구 설계

2.1 연구 방법론

데이터는 총 10년간의 기업금융 데이터를 세 기간으로 분류하여 연구를 진행하였다. 표본내 데이터는 2000~2006년(7년) 데이터가 학습 데이터로, 그리고 2007~2008년(2년) 데이터가 검증 데이터로 이용하고, 표본 외 데이터로는 2009년(1년) 데이터를 테스트 데이터로 이용하였다. 이는 표본내 데이터에서 학습데이터가 금융위기 이전, 검증 데이터가 금융위기 데이터를 포함하고, 금융위기 이후를 테스트 데이터로 연구에 이용함으로써 급격한 위기에 따른 변화 속에서 일관된 모형을 찾고자 하였다. 연구 순서는 표본내 데이터를 기반으로 세 종류의 변수 선정 알고리즘에 의해 최적의 변수를 도출한 후 기존 연구에서 이용되었던 모형(다중판별분석, 로짓 모형, 인공 신경망 모형 등)과 RNN 및 LSTM 부도예측모형의 비교 분석을 통해 효과를 검증했다.

일반적으로 기업부도예측모형 분석 시에는 정상데이터의 비율이 현저히 높아 정확도가 95%

를 넘는 경우가 대부분이다. 따라서 Accuracy로 예측력을 판단하는 것은 무리가 있다고 판단하였다. 따라서 Precision, Recall, F1 Score, ROC AUC, PR AUC를 분석평가지표로 이용한다. 또한 모형의 Threshold 기준은 부도데이터의 비율이 연간별 2%를 전후로 나타나므로 Threshold 기준을 [0.025, 0.05, 0.075, ..., 0.3]와 같이 12가지로 세분화 하여 분석 결과를 확인한다. Parameter Tuning은 Training Set과 Validation Set을 통해 각 모델마다 ROC AUC 및 PR AUC가 안정적으로 유지될 수 있도록 진행하였다. 이후 학습 및 유효성 검증 모두에서 F1 Score 값이 높을 때를 확인하여, Threshold의 범위를 모형에 포함하여 추천한다. 모델 선정 기준으로는 과 최적화가 되지 않은, 그리고 시대가 변화되었음에도 불구하고 안정성이 높은, 즉 Consistent하고 Robust한 모델 선정을 목표로 연구를 진행하였다.

뒤에서 제시될 모형 검증 지표인 ROC AUC, PR AUC는 Threshold 범위가 변화될 때 그 값이 넓이로서 표현되어 값이 클수록 모형의 변별력이 더 높아진다. 따라서 본 논문에서는 ROC AUC가 표본내 데이터에서 높은 값으로 유지되며, 비슷한 ROC AUC를 가지는 우수한 모형 중에서도 상대적으로 PR AUC가 높은 모형을 최우수 모델로 선정하고자 하였다. 즉, 다양한 변수 선정 모델에 의해 선별된 변수 및 여러 알고리즘과의 조합을 크게 두 가지 평가 지표를 바탕으로 최적 모델을 선정한다. 이를 통해서 선별된 모형을 바탕으로 In-sample data 전체를 가지고 재학습을 진행하여 최종적으로 Testing Set에 최종 모형을 적용한 결과값을 분석하였다.

연구에 이용한 SW는 Python이며, Scikit-learn, Keras, Tensorflow, sklearn 기반의 library를 이용하였다. OS는 리눅스를 이용하였으며, 이용한 컴

퓨터의 사양은 프로세서 2.7 GHz Inter Core i7, RAM 메모리 16GB, 그래픽은 Radeon Pro 455 2048 MB이다.

2.1.1 변수 분석 방법론

Kim(2000)의 연구는 기업부도예측모형 생성 시 현금흐름변수의 중요성을 검증하였다. 또한 Park(2014)의 연구는 외부감사대상기업의 경우 안정성 및 유동성과 관련된 재무비율 변수가 성장성 혹은 수익성 요인보다는 기업부실에 더욱 중요한 영향을 미치고 있음을 확인하였다. Hong(2003)의 연구는 유전자 알고리즘을 이용한 최적화를 통하여 입력 변수군을 도출하는 방법을 기존 다변량 관별분석과 로짓분석, 전문가 선정 변수 등의 변수군을 통해 인공지능망 모형을 구축한 결과 유전자 알고리즘에 의한 변수 선정이 유용함을 확인하였다. 이처럼 변수 선정 방식에 따라 기업부도예측모형의 효과는 달라지게 된다. 따라서 본 연구는 딥러닝 시계열 알고리즘에 대한 효과성을 검증하는 것으로서 대표적인 3가지 변수 선정 방식에서 모두 시계열 알고리즘이 여타 기존 연구들에서 제공되었던 모형들과 비교했을 때 우수함을 보이고자 하였다.

• 다중판별분석 변수 선정

ANOVA 분석은 4가지 기본 가정이 필요하다. 무작위 추출, 각 집단내 오차의 독립성, 각 집단내 오차는 정규분포를 따름, 모든 집단의 분산은 동일[등분산]하다는 가정이 필요하다. 정규성은 Shapiro(1965)의 정규성 검정으로 진행 가능하다. 또한 F-test를 통해 등분산성 여부를 확인하였다. 이후 등분산 가정이 만족될 때는 ANOVA분석인 Student's t-test를 진행하고, 등분산 가정이 만족하지 않을 시에는 비모수 통계 검

정인 Welch(1958)'s t-test를 진행한다.

이후 이를 기반으로 선정된 독립변수들을 기반으로 상관성 분석을 진행한다. 이는 다중공선성을 해결하는데 이용된다. 다중판별분석을 통해서 이에 맞는 변수를 다시 선택하고, 최종 선택된 변수들을 기반으로 분석을 진행한다.

• 로짓 모형 변수 선별

로지스틱 분석에서 독립변수를 줄여가며 모델의 성능을 향상시키는 방식인 Backward stepwise selection 방법에 의해 변수를 선택하였다. 로지스틱은 비선형모델로서 다중공선성문제를 감안하지 않은 모형이다. 로지스틱 분석의 경우 정상기업, 부도 기업 데이터의 수가 각각 설명변수수의 최소 10배 이상이 되어야 다중 로지스틱 회귀분석을 이용하는 것이 적절하다. 따라서 본 연구에서는 117개의 부도기업 데이터가 존재하므로 11개 변수를 선정하였다.

• Lasso 회귀분석 통한 변수 선정

Tibshirani(1996)에 의해 차원 축소 방법으로서, 최근 기계학습 및 인공지능 모형 분석시 독립변

수들이 종속변수에 영향을 미치는 다중공선성을 고려하여 최적의 변수를 도출하는 변수 선정 모델이다. 방법론은 L1-norm(Manhattan) Distance: $|X-Y|$ 을 기반으로 Parameter λ 를 적용하여 선형 회귀분석을 진행한다. 이후 변수들의 계수가 0이 되는 변수가 발생하게 되어 이 변수는 제외 가능해진다. L2-norm(Euclidean Distance: $\sqrt{X^2 + Y^2}$)을 이용하는 Ridge 회귀분석 또한 다중공선성을 고려한 회귀분석 방법으로 쓰인다. 하지만 독립변수의 계수를 0으로 만들지 못하고 0에 가깝게 추정하여 효과적인 변수 선정에는 어렵다는 단점이 있다. 따라서 λ 를 조정 후 Lasso 회귀분석을 진행하여 본 연구에는 총 11개의 최적 변수를 선정하였다.

2.1.2 다중판별분석 및 로짓모형

Altman(1968)의 Z-score 모형에 해당하는 다중판별분석 및 Ohlson(1980)의 O-score 모형에 해당하는 로짓모형은 기존 산업 및 연구에서 주로 쓰이는 대표적인 모형이다. 재무분야에서 대표적으로 기업부도모형 구축시 쓰이는 기존 모형

<Table 1> Original Models

Algorithm	Description	Detailed Description
Multiple Discriminant Analysis	Algorithm that minimizes the classification information between classes and reduces the dimension. It collapses dimensions in a way that maximizes 'variance between classes' over 'variance within classes'	Calculates variance between classes($V_B = \sum_{i=1}^c N_i (\mu_i - \mu)(\mu_i - \mu)^T$), variance within class($V_{W_i} = \sum_{x \in \omega_i} N_i (x - \mu_i)(x - \mu_i)^T$). Conducts an Eigen value analysis on $V_W^{-1}V_B$, select the Eigen vector corresponding to q largest Eigen values, and find a transformation matrix W that use it as q columns. Converts data to be trained and tested as ($y = W_i^T x$). Decides the class from average comparison of converted test data and training data.
Logit Model	A special case of a general linear model. However, when applied to binomial data, the result of the dependent variable y is limited to the range [0, 1]. And distribution of the conditional probability ($P(y x)$) as the dependent variable follows the binomial distribution instead of the normal distribution.	Success Probability: Suppose that $p = \Pr(Y = 1 x)$ Odds ratio: Define that $ODDs = \frac{p}{1-p}$ Logit transform: $\ln\left(\frac{p}{1-p}\right) = \alpha + \beta x + e$ Logit model: $\text{logit}(p) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_i$

인 만큼 딥러닝 시계열 알고리즘과의 비교분석을 진행하고자 한다(Table 1).

2.1.3 비시계열 지도학습 알고리즘

기계학습 알고리즘의 역할에는 분류와 군집이 있다. 분류란 이미 정의된 몇 개의 클래스로 구분하는 것을 말하며, 군집이란 주어진 데이터의

특성을 분석하여 특성에 따라 그룹으로 나누는 것을 말한다. 학습 방법의 경우는 크게 지도학습, 준지도학습, 비지도학습 3가지로 나뉜다. 먼저 지도학습이란 종속변수에 해당하는 변수들의 값이 모두 주어진 채로 학습에 이용가능하며, 준지도학습의 경우는 그 종속변수에 해당하는 값이 일부 존재하지 않아 값이 주어진 데이터만을 이

〈Table 2〉 Non-Time-Series Supervised Algorithms

Algorithm	Description
Decision Tree	<ul style="list-style-type: none"> • A target variable classification algorithm with a finite number of values based on finite depth tree branches. • It is useful for confirming the explanatory power of the independent variables. • However, in Decision Tree algorithm, when an error occurs in the middle of learning, the tree is generated with the error in the next step. • There is a disadvantage that the model greatly fluctuates according to the learning data.
Random Forest	<ul style="list-style-type: none"> • A modeling technique used for classification and regression analysis • A model that ensembles a number of Decision Tree models through bubbling and arbitrary node optimization. • Generates N training data sets through bootstrapping method, and then learn Decision Tree for each set and form an ensemble model through the same method as average and majority voting.
K-Nearest Neighbors(KNN)	<ul style="list-style-type: none"> • Nonparametric algorithms. • Learning by adding more weights so that nearby neighbors contribute more to the average than farther neighbors. • The larger K is, the less the Bias is, but the boundary between the items is also reduced. Therefore, it is important to set the appropriate K value. In addition to Euclidean distance as well as various analysis criteria, classification accuracy can be improved.
Support Vector Machine(SVM)	<ul style="list-style-type: none"> • Non-probabilistic linear classification model. • An algorithm that finds boundaries with the largest margins to be classified and represented in bounded spaces. • Since there are many overlapping data in the classifier, supplementing the data overlap modeling through the concept of soft margin.
Multi-Layer Perceptron(MLP)	<ul style="list-style-type: none"> • An algorithm to place one or more hidden layers at the midpoint of the input and output layers of existing artificial neural networks. • Cost function is applied by backpropagation using average square error between output and target, and commonly stochastic gradient descent method. <div style="text-align: center;"> <p>The diagram illustrates a Multi-Layer Perceptron (MLP) model. It consists of three layers: an Input Layer at the bottom with four nodes labeled $input_1$, $input_2$, $input_3$, and $input_4$; a Hidden Layer in the middle with five nodes; and an Output Layer at the top with one node labeled $Output$. Dotted lines represent the connections between nodes in adjacent layers, showing a fully connected structure between the input and hidden layers, and between the hidden and output layers.</p> </div>

〈Figure 1〉 MLP Model Diagram

용해 먼저 학습을 진행하고, 이후 종속변수에 해당하는 값이 부족한 부분을 채움으로서 지도학습을 진행하게 된다. 비지도학습은 지도학습과 준지도학습이 분류라는 것을 진행하는 것과는 다르게 독립변수만을 바탕으로 학습을 진행하여 군집을 하는 경우를 말한다.

기업부도에서는 부도를 예측을 함에 있어 부도 여부가 0과 1인 경우가 주어진 데이터를 바탕으로 지도학습을 진행하고 이를 통해 새로운 데이터를 0과 1로 분류하는 모델을 형성하고자 한다.

MLP(Multi-Layer Perceptron)와 같은 딥러닝 알고리즘은 복잡한 비선형 문제를 비지도 방식 학습으로 해결하는 데 효과적이다. 즉, 데이터가 많을 때 다양한 변수들을 이용해서 분석을 진행할 때, 비선형적인 체계를 가진 데이터에서 분석이 용이하다. 따라서, 기존의 연구 진행방식에서 비선형적인 연구를 진행할 때, 시간비용 및 인적 비용이 높았던 점을 특정 변수를 선택하지 않고도 쉽고 빠르게 분석이 가능하다, Training 시간이 오래걸리는 단점이 있지만 이는 GPU 및 CPU

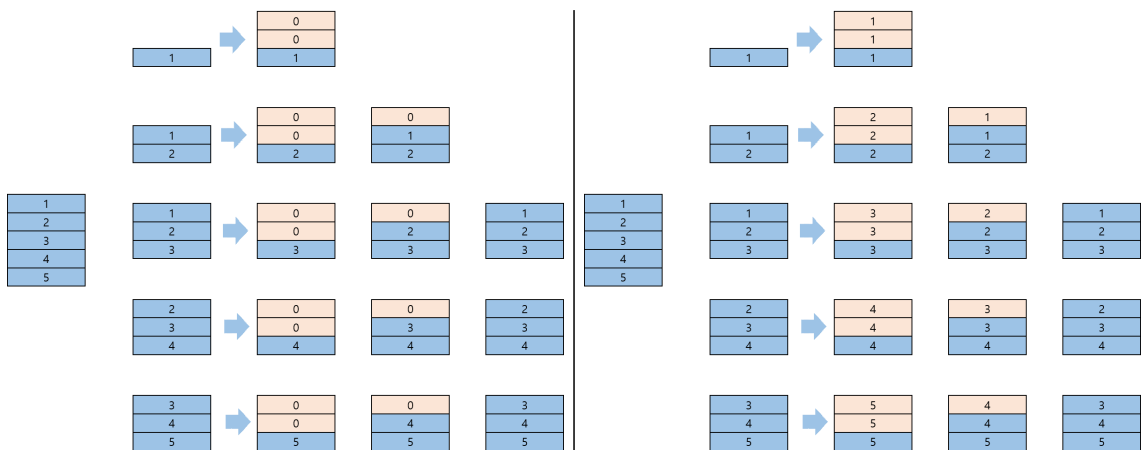
성능이 좋은 컴퓨터를 이용하여 속도를 개선할 수 있다(Table 2).

2.1.4 비시계열 지도학습 알고리즘

지도학습에서 표현된 딥러닝 MLP 알고리즘은 인공 신경망 모형 중 정적모형에 해당한다. 하지만 기업 데이터가 시계열 데이터라는 점, 또한 정적 모형은 시간이 변함에 따라 기업의 변화를 분석하는데 한계가 존재한다는 단점이 존재해 기업부도예측모형 개발에 있어서 동적 부도 예측모형의 개발은 필수적이다.

또한 시계열 알고리즘 분석의 강점은 가변적으로 데이터 추가 생성이 가능하다는 점이다. 아래의 그림은 5기간의 데이터가 존재할 때, 3기간의 연속 시계열 데이터를 이용하여 분석하고자 할 때 가변적으로 데이터를 생성하는 방식을 도식으로 나타내었다. 앞의 기간 데이터가 존재하지 않아도, Padding이란 개념을 통해서 빈 기간의 데이터에 0이나 혹은 일반적으로 직전값 혹은 직후값을 채워 분석에 이용가능하다.

따라서 시계열 알고리즘인 RNN(Recurrent



(Figure 2) Data Generation Method for Time-Series Deep Learning Algorithms

Neural Network)와 LSTM(Long Short Term Memory)을 이용해서 시간이 변함에 따라 변화를 반영가능한 동적 모형을 개발하고자 한다.

- RNN(Recurrent Neural Network)

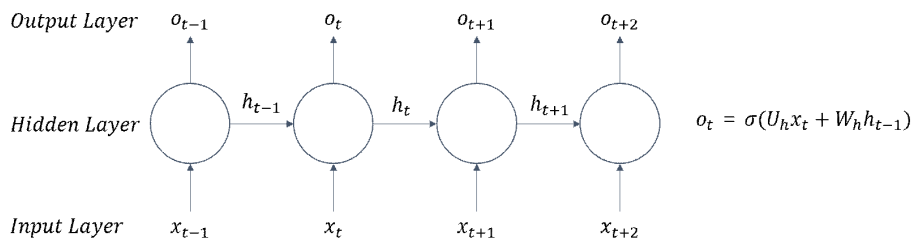
RNN이란 인공신경망 알고리즘 중 이전 기록의 데이터들까지도 다시 현시점의 데이터 학습에 이용하는 알고리즘으로서 이전 기록의 Hidden Layer에 남아있는 기록을 현시점의 데이터와 함께 Forward-propagation화하여 아래와 같은 관계를 Back-Propagation을 통해서 학습하는 알고리즘이다(Figure 3).

RNN은 이전의 정보를 활용한 분석을 진행한

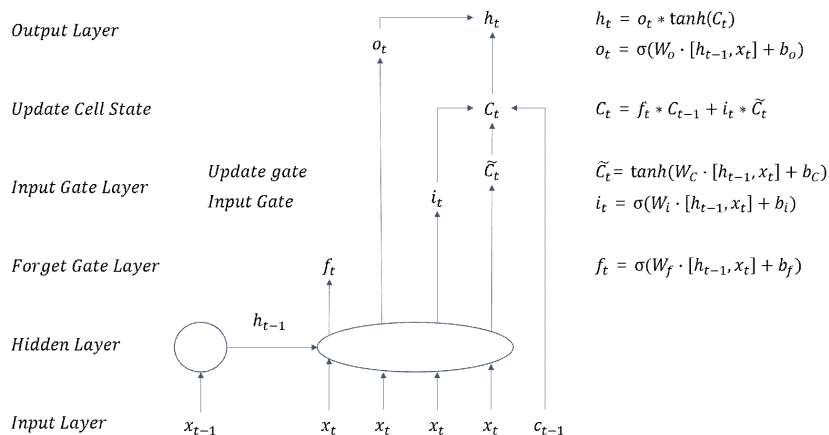
다. 하지만, 시점간 간격이 멀어질수록 Back-Propagation Through Time방법을 이용함에 있어 gradient값들이 0에 점점 가까워진다. 따라서 점점 Gradient가 줄어드는 Vanishing Gradient현상이 발생하게 되는 단점이 존재한다. 따라서 이를 보완하기 위해 생성된 알고리즘이 LSTM(Long Short Term Memory)이다.

- LSTM(Long Short Term Memory)

LSTM이란 Cell State라 불리는 Layer층을 하나 더 넣어서 Weight함수를 기억할 것인지 아닌지(추가 혹은 삭제하는 방식을 통해) 결정한다. 즉, Cell State 기능을 기반으로 RNN에서 발생했



<Figure 3> RNN Model Diagram



<Figure 4> LSTM Model Diagram

던 Gradient Vanishing 문제를 해결한 모형이다. Process 진행방식은 <Figure 4>와 같다.

2.1.5 모형 평가 기준

기업부도예측모형을 평가하고자 할 때, 정상/부도 데이터 편향성으로 인해 모든 알고리즘에서 정확도가 매우 높게 나온다. 따라서 모형을 평가하고자 할 때, 정확도를 포함하여 다양한 지표를 통해서 평가하고자 한다. 이용하는 평가지표 다음과 같다.

임계값(Threshold) 0~1 사이의 확률값을 기준으로 정하여 정상/부도로 예측한 결과 값과 실제 값을 비교하는 대표적인 지표로 쓰이는 4가지 평가지표는 <Table 3>과 같다.

위의 4가지 지표를 도표로 나타내면 <Table 4>와 같은 혼동행렬이 생성된다.

True는 정답을 맞춘 경우, False는 오답의 경우

를 말한다. 또한 Positive는 예측시 정상기업으로 예측할 확률이며, Negative는 예측에 의해 부도기업으로 예측할 확률을 말한다.

Precision, Recall, F1 Score(Precision과 Recall의 조화평균), Accuracy(정확도), ROC AUC(Receiver Operating Characteristics Area Under Curve), PR AUC(Precision-Recall Area Under Curve)이다. ROC AUC란 수신자 조작 특성으로서 X축의 값은 부도기업일 때, 정상기업으로 예측할 확률, Y축의 값은 정상기업일 때, 정상기업으로 예측했을 확률로 표현한 그래프의 아래 적분 값(넓이)을 의미한다(Table 5).

평가 기준은 주로 ROC AUC, PR AUC를 이용한다. F1 Score는 Threshold별 F1 Score가 높은 구간을 추천함으로써 Threshold를 추천하는 지표로 활용된다.

2.2 데이터 수집 및 정제

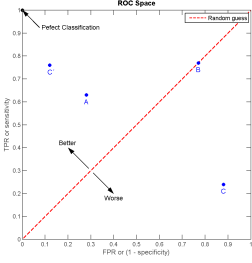
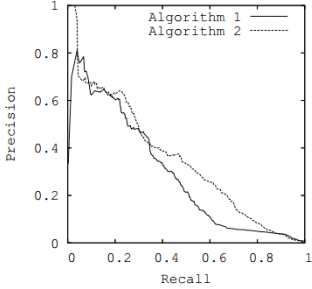
<Table 3> Evaluation Index

Evaluation Index	Description
True Positive	Predicted: normal, Actual: normal (True)
True Negative	Predicted: default, Actual: default (True)
False Positive	Predicted: normal, Actual: default (False)
False Negative	Predicted: default, Actual: normal (False)

<Table 4> Confusion Matrix

		Prediction Result	
		Normal	Default
Actual Result	Normal	True Positive [TP]	False Negative [FN] (Type II error)
	Default	False Positive [FP] (Type I error)	True Negative [TN]

〈Table 5〉 Advanced Evaluation Index

Evaluation Index	Description	Formula
Precision	Ratio which is predicted as default among the actual default data.	$\frac{TP}{TP + FP}$
Recall	Ratio which is actual default among predicted as default	$\frac{TP}{TP + FN}$
F1 Score	Harmonic mean of Precision and Recall	$\frac{2 \times Precision \times Recall}{Precision + Recall}$
Accuracy	Probability that the prediction is true.	$\frac{TP + TN}{TP + TN + FP + FN}$
ROC AUC	 <p>Area under curve that FPR, TPR are expressed as X axis, Y axis.</p>	<p>X axis: $FPR = \frac{FP}{FP + TN}$</p> <p>Y axis: $TPR = \frac{TP}{TP + FN}$</p>
PR AUC	 <p>Area under curve that Recall, Precision are expressed as X axis, Y axis.</p>	<p>X axis: Recall</p> <p>Y axis: Precision</p>

㈜에프엔가이드에서 제공하는 DataGuide 터미널을 통해서 연구에 사용된 자료는 상업적으로 구독해야하는 데이터로서 데이터를 이용하는데 제약이 있을 수 있다. 데이터는 KOSPI, KOSDAQ

기업들의 2000년부터 2009년까지 연단위 데이터이다. 연단위로 주어지는 재무상태표, 포괄손익계산서, 현금흐름표, 안정성, 성장성, 수익성, 활동성 지표와 관련된 변수를 모두 가져와서 데이

터 전처리(결측값 처리 및 변수별 데이터 표준화)를 진행한 후 분석에 이용하였다.

• 기업의 부도 정의

기존 참고 문헌들에 의하면, 기업 부도 정의는 금융감독원, 신BIS협약, 은행연합회의 불량정보 기준 등이 대표적이다. 신BIS협약의 정의는 잠재적 부실가능성을 포함하여 가장 포괄적으로 부도를 정의하며, 은행연합회의 불량정보기준은 국내 대부분의 금융기관에서 사용하는 기준으로 신BIS협약에서 포괄적으로 정의하고 있는 부도에 해당하는 사유를 구체적으로 제시하고 있다. 하지만 본 연구에서 이용할 부도의 정의는 Lee (2015)의 연구와 같이 실적부진의 사유로 상장폐지가 일어난 경우를 부도로 정의하고자 하였다. 즉, 부도발생, 화의절차개시신청, 회사정리절차개시신청, 감사인의 의견 거절, 은행거래 정지 등이 부도에 해당되며, 부도와 관련이 없는 사유인 신규/변경 상장, 특수 목적에 의한 상장 폐지,

자진등록취소 등은 부도에서 제외하였다. 위의 기준에 따라 부도 발생 직전 년도에 부도가 발생하는지 여부를 정하고 분석을 진행하고자 하였다. 따라서 부도 기준 직전 년도 데이터에 부도 여부 변수값을 ‘1’, 부도 이전 정상일 때는 ‘0’, 부도 이후의 데이터는 제거하였다(Table 6).

2000년부터 2009년까지 데이터에 의하면 총 1,743개 기업, 그리고 254개 부도기업이 존재한다. 이후 금융권 기업들을 제외 및 결측값 전처리를 진행하였다. 그 결과 총 890개의 기업데이터와 134개의 부도기업데이터가 연구에 이용되었다.

10년간 총 8,018개의 데이터 내 부도데이터는 134개에 해당되며, 가변적인 데이터 생성 이전과 이후의 데이터 분포를 비교하면 <Table 7>과 같다.

3. 결과

<Table 6> Original data distribution (Left) and data distribution after preprocess (Right)

Year	Num of Default	Num of Company	Default Ratio
2000	28	1676	1.67%
2001	36	1682	2.14%
2002	21	1654	1.27%
2003	38	1635	2.32%
2004	30	1594	1.88%
2005	4	1557	0.26%
2006	9	1554	0.58%
2007	11	1554	0.71%
2008	42	1539	2.73%
2009	35	1486	2.36%
All	254	1743	14.57%

Year	Num of Default	Num of Company	Default Ratio
2000	11	864	1.27%
2001	22	860	2.56%
2002	17	839	2.03%
2003	14	825	1.70%
2004	16	807	1.98%
2005	4	788	0.51%
2006	8	783	1.02%
2007	5	774	0.65%
2008	20	756	2.65%
2009	17	722	2.35%
All	134	890	15.06%

<Table 7> Distribution Before/After Time-Series Data Generation

	2000-2006 [Train Set]		2007-2008 [Validation Set]	
	All	Default	All	Default
Before Time-series Data Generation	5,766	92	1,530	25
After Time-series Data Generation (3 year period)	14,649	227	1,530	25
	2000-2008 [Train + Validation Set]		2009 [Test Set]	
	All	Default	All	Default
Before Time-series Data Generation	7,296	117	722	17
After Time-series Data Generation (3 year period)	19,239	302	722	17

3.1 변수 선정 결과

- 다변량 판별분석 변수 선정

표본 데이터에서 부도기업 데이터의 수는 117(30이상)개로 정규성 검정을 진행하지 않고 정규성 검정 없이 정규 분포를 따른다고 가정한다. 이후, 유의수준 5%에 의해 부도기업과 정상기업 두 분포가 동일한 분산을 가진다는 귀무가설을 기반으로 F-test를 진행한 결과, 귀무가설이 기각된 경우에는 부도/정상 데이터의 분산이 동일하다는 가정이 유의미하지 않다고 판단하여 이분산(Heteroscedasticity of variance)으로 T-test (Welch's t-test)를 진행하였고, 귀무가설이 기각되지 않은 경우에는 귀무가설이 유의하다고 판단하여 등분산(Homogeneity of variance)을 가정하여 T-Test(Student's t-test)를 진행하였다. 그 결과, 총 29개의 변수들이 <Table 8>과 같이 선정

되었다.

표본 내 데이터는 총 7,296개 중에서 정상기업 데이터는 7,179개, 부도기업데이터는 총 117개이므로 상대적으로 분산이 같아지는 경우가 존재하기 어렵다는 점을 감안한다면 일반적으로 부도기업 데이터와의 분포 비교에서 이분산으로 표현되는 것은 타당하다. 결과적으로, 이분산을 가정한 Welch's t-test를 통해서 진행된 변수 선정 결과는 5%의 유의수준에서 모두 유의미한 차이를 보인다는 것을 확인할 수 있었다.

이후 변수들의 상관계수를 직접 계산하여 상관성 높은 변수군들을 생성한다. 이후 상관성 높은 변수군들 내에서 T-test의 p-value가 가장 낮은 변수를 선별한 결과는 <Table 9>와 같다.

〈Table 8〉 Feature Selection Result from applying F-test and T-test

category	section	feature	F-test		T-test
			p-value	variance	p-value
Financial Statements	Balance Sheet (1,000 won)	Accumulations	5.07×10^{-05}	hetero	1.30×10^{-04}
		Retained Earnings	8.37×10^{-04}	hetero	7.06×10^{-04}
		Net assets of controlling shareholders (before capital stock reduction)	9.43×10^{-03}	hetero	5.90×10^{-03}
		Owners of Parent Equity	6.58×10^{-03}	hetero	9.48×10^{-04}
		Total Equity	7.37×10^{-03}	hetero	2.33×10^{-05}
	Comprehensive Income Statement (1,000 won)	Earnings before tax	1.82×10^{-02}	hetero	4.15×10^{-02}
		(Total Comprehensive Income Attributable to) Owners of Parent Equity	3.59×10^{-04}	hetero	3.59×10^{-04}
		Total Comprehensive Income	7.41×10^{-03}	hetero	3.22×10^{-02}
	Cash Flow Statement (1,000 won)	Cash Flow	4.41×10^{-04}	hetero	1.13×10^{-02}
	Financial Ratio	Stability (%)	Intangible Asset Ratio	7.47×10^{-04}	hetero
Equity Capital Ratio			2.34×10^{-39}	hetero	1.48×10^{-20}
Borrowings and Bonds Payable Ratio			2.04×10^{-49}	hetero	9.92×10^{-05}
Borrowed Capital Ratio			2.04×10^{-40}	hetero	9.63×10^{-06}
Cash Flow/ Total Debt			3.08×10^{-06}	hetero	2.23×10^{-08}
Cash Flow/ Total Equity			3.63×10^{-20}	hetero	1.65×10^{-02}
Cash Flow/ Total Asset			1.73×10^{-29}	hetero	8.07×10^{-09}
Growth (yearly) (%)		Total Asset Growth Rate	3.51×10^{-07}	hetero	2.88×10^{-14}
Profitability (%)		Operating Revenue/ Operating Expense	5.42×10^{-12}	hetero	6.71×10^{-13}
		Profit Margin Ratio	2.44×10^{-04}	hetero	2.45×10^{-04}
		ROA(Current Net Income)	1.96×10^{-22}	hetero	4.18×10^{-09}
		ROA(Earnings before tax)	3.86×10^{-23}	hetero	1.79×10^{-09}
		ROA(Operating Profit)	2.40×10^{-67}	hetero	2.06×10^{-10}
		ROA(Total Comprehensive Income)	2.60×10^{-21}	hetero	4.51×10^{-09}
		ROE(Current Net Income)	1.10×10^{-33}	hetero	5.58×10^{-04}
		ROE(Earnings before tax)	1.37×10^{-48}	hetero	2.89×10^{-05}
		ROE(Operating Profit)	1.86×10^{-35}	hetero	2.52×10^{-03}
Activity (times)		ROE(Net profit of controlling shareholders)	3.86×10^{-34}	hetero	5.21×10^{-04}
		Total Debt Turnover	1.10×10^{-14}	hetero	5.68×10^{-27}
		Total Asset Turnover	1.17×10^{-03}	hetero	1.99×10^{-05}

〈Table 9〉 Feature Selection by Correlation Analysis

category	section	feature	F-test		T-test
			p-value	Variance	p-value
Financial Statements	Balance Sheet (1,000 won)	Total Equity	7.37×10^{-03}	Hetero	2.33×10^{-05}
	Comprehensive Income Statement (1,000 won)	(Total Comprehensive Income Attributable to) Owners of Parent Equity	3.59×10^{-04}	Hetero	3.59×10^{-04}
Financial Ratio	Stability (%)	Intangible Asset Ratio	7.47×10^{-04}	Hetero	4.84×10^{-02}
		Equity Capital Ratio	2.34×10^{-39}	Hetero	1.48×10^{-20}
		Borrowed Capital Ratio	2.04×10^{-40}	Hetero	9.63×10^{-06}
		Cash Flow/ Total Debt	3.08×10^{-06}	Hetero	2.23×10^{-08}
	Growth (yearly) (%)	Total Asset Growth Rate	3.51×10^{-07}	Hetero	2.88×10^{-14}
	Profitability (%)	Operating Revenue/ Operating Expense	5.42×10^{-12}	Hetero	6.71×10^{-13}
		Profit Margin Ratio	2.44×10^{-04}	Hetero	2.45×10^{-04}
		ROA(Earnings before tax)	3.86×10^{-23}	Hetero	1.79×10^{-09}
		ROA(Operating Profit)	2.40×10^{-67}	Hetero	2.06×10^{-10}
		ROE(Earnings before tax)	1.37×10^{-48}	Hetero	2.89×10^{-05}
		ROE(Operating Profit)	1.86×10^{-35}	Hetero	2.52×10^{-03}
Activity (times)	Total Debt Turnover	1.10×10^{-14}	Hetero	5.68×10^{-27}	

• 로짓 모형에 의한 변수 선정

로짓 모형 변수 선정 방법론에 의해서, 115개의 부도 기업 데이터 학습 개수를 맞추면 총 11개 이하의 변수를 이용하는 것이 전형적이라고 하였다. 따라서 Backward stepwise selection 방법에 의해 변수를 하나씩 제거하면서 유의미한 변수 선정을 했을 때 최종으로 남아 있었던 순서대로 1-11번을 매겨서 기록하였다(Table 10).

• Lasso 회귀분석에 의한 변수 선정

Lasso 회귀분석을 통해 λ 를 0.004로 조정하여 Logistic Regression에서 뽑았던 변수 개수와 동일한 개수의 변수를 선정하였다. 변수별 계수의 절대값이 높으면 영향력이 높고 예시는 ROA(영업이익), ROE(세전계속사업이익), 총부채회전을 등이 선정되었다(Table 11).

<Table 10> Feature Selection Result from applying Logit Model

category	section	feature	rank
Financial Statements	Balance Sheet (1,000 won)	Current tax liabilities	6
		Retained Earnings	11
		Owners of Parent Equity	2
	Income Statement(1,000won)	Payroll cost	9
	Cash Flow Statement(1,000 won)	Investments Earnings in Associates	5
Financial Ratio	Stability (%)	Equity Capital Ratio	10
	Growth (yearly) (%)	Total Debt Growth Rate	4
		Total Asset Growth Rate	1
	Profitability (%)	ROA(Earnings before tax)	7
		ROA(Total Comprehensive Income)	8
Activity (time)	Total Debt Turnover	3	

<Table 11> Feature Selection Result from applying Lasso Regression

category	section	feature	rank
Financial Statement	Balance Sheet (1,000 won)	Non-Current Provisions for Employee Benefits	11
	Cash Flow Statement (1,000 won)	Interest Expenses	3
Financial Ratio	Stability (%)	Equity Capital Ratio	4
		Borrowings and Bonds Payable Ratio	8
		Borrowed Capital Ratio	6
		Cash Flow/ Total Asset	9
	Growth (yearly) (%)	Total Asset Growth Rate	10
	Profitability (%)	Selling Administrative Expense Rate	7
		ROA(Operating Profit)	1
		ROE(Earnings before tax)	2
		ROE(Operating Profit)	5
Activity (times)	Total Debt Turnover	3	

<Table 12>는 3가지 변수 선별 모형에 의해 선정된 최종 변수 목록이다.

최종적으로 선택된 변수 목록을 살펴보면, 자기자본비율, 타인자본비율, 총자산증가율, ROA

(세전계속사업이익), ROA(영업이익), ROE(세전계속사업이익), 총부채회전율이 공통적으로 선별되었다. 이외에도 지배주주와 관련된 변수인 ‘지배주주총포괄이익’ 및 ‘지배주주지분’, 총포

〈Table 12〉 Feature Selection Summary from 3 models

category	section	MDA	Logit	Lasso
Financial Statement	Balance Sheet	Total Capital	Current tax liabilities	Non-Current Provisions for Employee Benefits
			Retained Earnings	
			Owners of Parent Equity	
	Income Statement	(Total Comprehensive Income Attributable to Owners of Parent Equity)	Payroll cost	Interest Expenses
Cash Flow Statement		Investments Earnings in Associates		
Financial Ratio	Stability	Intangible Asset Ratio	Equity Capital Ratio	Equity Capital Ratio
		Equity Capital Ratio		Borrowings and Bonds Payable Ratio
		Borrowed Capital Ratio		Borrowed Capital Ratio
		Cash Flow/ Total Debt		Cash Flow/ Total Asset
	Growth	Total Asset Growth Rate (yearly)	Total Debt Growth Rate (yearly)	Total Asset Growth Rate (yearly)
			Total Asset Growth Rate (yearly)	
	Profitability	Operating Revenue/ Operating Expense	ROA (Earnings before tax)	Selling Administrative Expense Rate
		Profit Margin Ratio	ROA(Total Comprehensive Income)	ROA(Operating Profit)
		ROA(Earnings before tax)		ROE(Earnings before tax)
		ROA(Operating Profit)		
		ROE(Earnings before tax)		
		ROE(Operating Profit)		
	activity	Total Debt Turnover	Total Debt Turnover	Total Debt Turnover

2 Times 3 Times

팔이익과 관련된, ‘지배주주총포팔이익’과 ‘ROA (총포팔이익)’이 상관관계가 있을 것으로 표현되는 요소이며, 부채와 관련된 변수는 ‘현금흐름/총부채’, ‘총부채회전율’, ‘당기법인세부채’, ‘총부채증가율’, ‘비유동종업원급여충당부채’가 존재했다.

Kim(2011) 등의 연구에 따르면, 기업부실화의 원인분석을 통해서도 자기자본을 많이 보유할수

록 기업 부도확률이 줄어든다는 결과를 확인하였으며, 다른 변수들의 영향력에 대해서도 조사하였다. 하지만, 기업 부도에 영향력을 미치는 재무변수들의 순차적인 중요성을 파악하지 못하는 문제로 인하여 재무변수의 부도확률 설명력 순서에 관한 연구의 필요성을 인식하였다. Ahn(2014) 저자의 재무비율을 이용한 부도예측에 대한 연구에서는 117개 변수내 선정된 17개

변수에 따르면, 자기자본비율, 차입금의존도, 총 부채회전율이 본 연구에서 선정된 변수들과 공통적으로 선정되었다.

본 연구에서는 다변량 판별분석을 기반으로 가장 p-value가 낮은 변수 5개를 차례로 선정하였고, 로짓 모형 및 Lasso 회귀분석은 이와 관련하여 backward stepwise selection 방식을 기반으

로 우선순위를 정할 수 있었다. 이에 각 변수선정모형별 상위 5위에 해당하는 8개 변수인 총부채회전율, 자기자본비율, 영업수익/영업비용, ROA(영업이익), 총자산증가율, 지배주주지분, 총부채증가율, ROE(세전계속사업이익)에 대해 정리하면 <Table 13>과 같다.

결과적으로 변수 선정하는 방식은 각각 달랐

<Table 13> Train/validation result with features selected by Multiple Discriminant Analysis

Evaluation Index	Description	Formula
Total Debt Turnover	To pay off the debt, the total debt turnover rate should be high. In other words, the higher the total debt turnover rate, the less the default probability.	Revenue / [(Total debts of the beginning at this year+ Total debts at the end of this year)/2]
Equity Capital Ratio	The higher the ratio of equity capital in operating a business, the more likely the probability of default is reduced.	(Equity Capital / Total Capital) × 100
Operating Revenue/ Operating Expense	The higher the cost-to-revenue ratio, the better the company is operating. That is, the larger the value of the operating profit / operating cost, the less the probability of default.	Operating Revenue/ Operating Expense
ROA(Operating Profit)	The higher the earning based on assets including capital and debt, the less the probability of default.	Net income / Total assets
Total Asset Growth Rate	Which is the ratio of assets to assets at the end of the year. This means that there is a payoff depending on whether the capital is increased or not and the result is that the probability of default decreases when capital increases and the probability of default increases when debt increases.	(Total assets at the end of this year / Total assets at the end of last year) × 100-100
Owners of Parent Equity	The amount of the economic entity's stockholders' equity attributable to the parent excludes the amount of stockholders' equity which is allocable to that ownership interest in subsidiary equity which is not attributable to the parent (noncontrolling interest, minority interest). The higher the controlling shareholder's equity, the less the probability of default	-
Total Debt Growth Rate (yearly)	Which is the ratio of debt to debt at the end of the year. As a general rule, the greater the liability increases, the greater the probability of default.	(Total liabilities at the end of this year / Total liabilities at the end of last year) × 100-100
ROE(Earnings before tax)	A measure of the profitability of a business in relation to the equity, also known as net assets or assets minus liabilities. ROE is a measure of how well a company uses investments to generate earnings growth. The higher the return on equity, the lower the probability of default.	Net income / Total equity

으나, 일정부분 공통적인 부분들이 발견되었다. 그러므로 각각의 변수 선정 모델들이 서로 상관관계가 있음을 감안하여 각각의 알고리즘별 비교 분석을 진행해보았다.

3.2 모형별 분석 결과 비교

- 다변량 판별 분석 선정 변수 적용 (Train/Validation Set)

<Table 14>의 결과는 각 알고리즘마다 Train 및 Validation Set에서 ROC AUC 기준으로 값이 높은 경우와 그 차가 많이 나지 않는 경우를 Parameter Tuning 후 선정된 모형의 분석 결과이다. 2000~2006년의 분석 결과를 살펴보면, SVM

을 제외한 모든 모형은 학습 결과 ROC AUC가 0.85 이상이었다. 학습은 유의하게 잘 되고 있음을 확인할 수 있으며, 학습에서 ROC AUC가 0.9를 넘는 모형은 Random Forest와 KNN 그리고 LSTM이었다. 2007~2008년 데이터 분석 결과에 따르면, ROC AUC가 오히려 더 커진 경우가 존재하였다. 이는 Under-fitting된 경우로도 볼 수 있으나, 부도 발생의 위험이 높았던 금융위기 시점의 데이터가 포함된 유효성 검증 데이터이므로 모형의 ROC AUC값이 오히려 더 커지는 경향을 나타내고 있었다. 결과적으로 2007~2008년의 경우 ROC AUC가 0.9를 초과하는 경우는 DT, KNN, SVM을 제외한 모두였다.

본 연구에서는 금융위기 이전과 이후의 모형

<Table 14> Train/validation result with features selected by Multiple Discriminant Analysis

Algorithm	Training Result			Validation Result		
	Accuracy	ROC AUC	PR AUC	Accuracy	ROC AUC	PR AUC
LDA	0.976	0.851	0.182	0.976	0.919	0.258
LR	0.984	0.861	0.266	0.980	0.912	0.255
DT (Max_depth=6)	0.991	0.897	0.626	0.982	0.888	0.282
RF (Max_depth=3, n_estimator=500)	0.987	0.919	0.455	0.982	0.915	0.279
KNN (K=30)	0.984	0.950	0.277	0.983	0.869	0.190
SVM (C=0.001)	0.987	0.735	0.287	0.982	0.682	0.062
MLP (Neuron=300, Dropout=50%, EarlyStop=10)	0.985	0.870	0.275	0.980	0.935	0.278
RNN (Neuron=300, Dropout=50%, EarlyStop=10)	0.986	0.899	0.345	0.978	0.914	0.231
LSTM (Neuron=250, Dropout=50%, EarlyStop=10)	0.981	0.909	0.413	0.980	0.901	0.246

이 일관되고, 꾸준히 성과가 좋은 모형을 선정하고자 하였기에, 금융위기 이전과 이후의 ROC AUC값이 크게 벗어나지 않는 모형을 우수 모형으로 선정하여 확인을 진행하였다. 따라서, 위 결과값을 기준으로 학습과 유효성 검증을 통해 살펴본 결과로는 DT와 LSTM의 ROC AUC 결과값의 차이가 0.01도 채 나지 않았으며, 오히려 그 평균은 LSTM이 높은 것을 확인할 수 있었다. 또한 Tree계열 알고리즘인 DT와 RF 모형은 PR AUC에서는 상당히 Over-Fitting이 되어있었다.

ROC AUC가 상대적으로 다른 모형에 비해서 높은 LSTM과 RF 모형을 비교해보면 상대적으로 양상불 모델을 적용한 RF 모형이 우수함을 확인할 수 있었다. 하지만, 두 모형 모두 앞서 금융위기의 Risky한 데이터 기반으로 유효성 검증

을 적용하였음에도 PR AUC가 감소한 것으로 볼 때, Over-fitting되어있는 것으로 보여진다. 오히려 LDA, LR, MLP, RNN과 같이 ROC AUC 혹은 PR AUC가 점차 성장하는 모형으로서 금융위기의 Risky한 부분을 감안하였을 때, Over-fitting한 부분이 다소 적은 모형으로 보여진다.

- 다변량 판별 분석 선정 변수 테스트 결과 (Train + Validation Set/Test Set)

학습 및 유효성 검증을 통해 선정된 모형들로 다시 Train Data와 Validation Data를 합산하여 모형을 학습 후 Test Set에서 시험을 진행하였다. 학습 모형 및 검증 모형이 서로 많이 상이한 SVM 모델의 경우에는 결과 분석에서 제외하였다(Table 15).

〈Table 15〉 Test result with features selected by Multiple Discriminant Analysis

Algorithm	Train+Validation Result			Test Result		
	Accuracy	ROC AUC	PR AUC	Accuracy	ROC AUC	PR AUC
LDA	0.975	0.868	0.192	0.979	0.804	0.306
LR	0.983	0.882	0.510	0.982	0.702	0.373
DT (Max_depth=6)	0.989	0.897	0.587	0.975	0.645	0.101
RF (Max_depth=3, n_estimator=500)	0.986	0.922	0.423	0.975	0.861	0.381
KNN (K=30)	0.984	0.956	0.270	0.977	0.832	0.242
MLP (Neuron=300, Dropout=50%, EarlyStop=10)	0.984	0.892	0.295	0.981	0.756	0.396
RNN (Neuron=300, Dropout=50%, EarlyStop=10)	0.987	0.926	0.457	0.978	0.875	0.330
LSTM (Neuron=250, Dropout=50%, EarlyStop=10)	0.987	0.926	0.439	0.977	0.844	0.159

Test 결과, ROC AUC 기준, 상대적으로 대다수의 모형들이 예측력이 크게 떨어짐을 확인할 수 있었다. 표본내 데이터 학습의 경우 92이상인 경우는 KNN, RNN, LSTM, RF 순서대로 총 4개의 모형이 있으나, 표본외검증인 2009년 Test 결과는 RNN, RF, LSTM, KNN순으로 높았다. 상대적으로 KNN은 학습은 잘했지만 예측력 면에서 RNN, RF, LSTM에 비해서 Over-fitting 되었음을 확인할 수 있었다.

PR AUC의 경우는 위에서 효과적이었던 RNN, RF, LSTM 3 가지 모형을 비교해보면, 상대적으로 RF, RNN 모형이 테스트 결과 0.33이상으로 높았다.

• 로짓 모형 선정 변수 적용(Train/Validation Set)
다변량 판별분석 분석 결과와 동일하게 로짓 모형 변수 선정 기반으로 9가지 부도예측모형의 안정성을 분석한 결과, ROC AUC 기준에서는 DT 및 KNN이 상대적으로 Over-fitting 되어있으며, RF, RNN, LSTM이 높은 값과 안정성을 두루 갖추고 있었다(Table 16).

유효성 검증 데이터 내에서 PR AUC의 경우, RF는 상대적으로 금융위기 시점을 포함하였음에도 크게 줄어든 것으로 보아 Over-fitting되어 있으며, LSTM 모형의 경우가 0.302로 가장 그 결과값이 높았다.

<Table 16> Train/validation result with features selected by Logit model

Algorithm	Training Result			Validation Result		
	Accuracy	ROC AUC	PR AUC	Accuracy	ROC AUC	PR AUC
LDA	0.982	0.822	0.169	0.976	0.881	0.167
LR	0.984	0.858	0.268	0.982	0.884	0.255
DT (Max_depth=6)	0.988	0.912	0.595	0.981	0.836	0.170
RF (Max_depth=3, n_estimator=500)	0.987	0.916	0.433	0.982	0.930	0.270
KNN (K=30)	0.983	0.958	0.253	0.984	0.877	0.246
SVM (C=100)	0.984	0.850	0.484	0.984	0.479	0.053
MLP (Neuron=300, Dropout=50%, EarlyStop=10)	0.985	0.855	0.251	0.982	0.889	0.247
RNN (Neuron=300, Dropout=50%, EarlyStop=10)	0.985	0.893	0.268	0.985	0.899	0.270
LSTM (Neuron=250, Dropout=50%, EarlyStop=10)	0.985	0.910	0.355	0.983	0.928	0.302

• 로짓 모형 선정 변수 테스트 결과

(Train + Validation Set/Test Set)

학습 및 유효성 검증을 통해 선정된 모형들로 다시 Train Data와 Validation Data를 합산하여 모형 학습 후 Test Set에서 시험을 진행하였다. 검증 모형의 ROC AUC가 약 0.48인 점으로 보아 효과적이지 않은 것으로 판단하여 SVM 모형의 경우에는 시험 결과 분석에서 제외하였다(Table 17).

시험 분석 결과, ROC AUC의 경우는 LDA, LR, MLP의 모형은 테스트 검증에서 오히려 그 값이 크게 떨어져 Over-fitting된 상태로 보여진다. 앞서 유효성 검증을 통해서 Over-fitting 된 것으로 확인되었던 DT모형의 경우는 오히려 효과가 높게 나왔다. 결과적으로, 앞서 선정된 RF,

RNN, LSTM 모형은 약간의 Over-fitting된 경향을 보이지만 상대적으로 RF와 RNN의 결과값이 덜 Over-fitting되었다.

테스트 결과에서 PR AUC의 경우는 RF와 RNN의 경우가 전체 모델 중에서 높은 편에 속했는데, RNN이 그 중에서도 0.5123으로 가장 높은 값을 보여주었다.

• Lasso 회귀분석 모형 선정 변수 적용

(Train/Validation Set)

학습데이터와 유효성 검증데이터의 ROC AUC 결과값을 확인해보면, 앞서 두 변수 선정 모형의 결과와 동일하게 LDA, LR, MLP가 금융 위기 시점이 포함되지 않은 모형을 학습함으로써 상대적으로 Under-fitting 된 형태를 보이고 있

<Table 17> Test result with features selected by Logit model

Algorithm	Train+Validation Result			Test Result		
	Accuracy	ROC AUC	PR AUC	Accuracy	ROC AUC	PR AUC
LDA	0.981	0.839	0.190	0.978	0.693	0.327
LR	0.983	0.873	0.250	0.978	0.684	0.340
DT (Max_depth=6)	0.989	0.919	0.618	0.971	0.897	0.245
RF (Max_depth=3, n_estimator=500)	0.986	0.922	0.410	0.981	0.899	0.390
KNN (K=30)	0.982	0.959	0.253	0.979	0.881	0.414
MLP (Neuron=300, Dropout=50%, EarlyStop=10)	0.983	0.871	0.250	0.982	0.714	0.426
RNN (Neuron=300, Dropout=50%, EarlyStop=10)	0.984	0.906	0.259	0.981	0.862	0.512
LSTM (Neuron=250, Dropout=50%, EarlyStop=10)	0.985	0.921	0.375	0.978	0.830	0.233

〈Table 18〉 Train/validation result with features selected by Lasso Regression

Algorithm	Training Result			Validation Result		
	Accuracy	ROC AUC	PR AUC	Accuracy	ROC AUC	PR AUC
LDA	0.975	0.851	0.188	0.973	0.927	0.199
LR	0.983	0.862	0.237	0.980	0.902	0.203
DT (Max_depth=6)	0.991	0.933	0.601	0.978	0.846	0.211
RF (Max_depth=3, n_estimator=200)	0.987	0.914	0.458	0.983	0.924	0.262
KNN (K=30)	0.984	0.949	0.236	0.984	0.892	0.245
SVM (C=50)	0.984	0.910	0.711	0.984	0.513	0.137
MLP (N=200, Dropout=50%, EarlyStop=10)	0.985	0.875	0.275	0.981	0.940	0.265
RNN (N=200, Dropout=50%, EarlyStop=10)	0.987	0.906	0.417	0.982	0.914	0.182
LSTM (N=200, Dropout=50%, EarlyStop=10)	0.986	0.914	0.443	0.980	0.915	0.247

으며, DT와 KNN이 다소 Over-fitting이 심하였다. 과거와 동일하게 SVM은 유효성 검증시 0.51의 값을 가지게 되어 예측력이 무의미하였다 (Table 18).

PR AUC의 결과값을 살펴보면 RF, KNN, MLP, 그리고 LSTM이 각각 0.262, 0.245, 0.265, 0.247로서 우수했다. 상대적으로 학습 및 유효성 검증 결과는 이번 모형에서는 LSTM이 RNN에 비해서 상대적으로 더 효과적이었다.

• Lasso 회귀분석 모형 선정 변수 테스트 결과 (Train + Validation Set/Test Set)

ROC AUC의 경우는 표본내 데이터 학습기준으로 RF, KNN, RNN, LSTM이 0.9를 상회했다. KNN의 경우 이전 시험결과에서와 동일한 방향으로 Over-fitting 정도가 상대적으로 심했다. RF 모형은 가장 ROC AUC의 분포가 안정적이었으며, RNN과 LSTM은 그 중간정도로서 안정적인 편에 속했다(Table 19).

PR AUC의 결과는 RF, KNN, RNN, LSTM을 비교해보면 상대적으로 RF와 LSTM이 우수했으

(Table 19) Test result with features selected by Lasso Regression

Algorithm	Train+Validation Result			Test Result		
	Accuracy	ROC AUC	PR AUC	Accuracy	ROC AUC	PR AUC
LDA	0.974	0.867	0.195	0.979	0.836	0.309
LR	0.983	0.882	0.216	0.981	0.702	0.330
DT (Max_depth=6)	0.989	0.885	0.537	0.974	0.836	0.324
RF (Max_depth=3, n_estimator=500)	0.986	0.922	0.413	0.977	0.917	0.382
KNN (K=30)	0.984	0.954	0.264	0.975	0.844	0.375
MLP (N=200, Dropout=50%, EarlyStop=10)	0.984	0.895	0.281	0.979	0.773	0.401
RNN (N=200, Dropout=50%, EarlyStop=10)	0.985	0.916	0.394	0.978	0.869	0.314
LSTM (N=200, Dropout=50%, EarlyStop=10)	0.986	0.919	0.381	0.979	0.873	0.418

며, 시험 결과에서는 LSTM이 0.418로서 전체 모형들 중 가장 우수했다.

4. 결론

4.1 결론

본 연구는 3가지 변수 선정 모형(다중판별분석 모형, 로짓 모형, Lasso 회귀분석)을 통해서 최적 변수군을 3개 생성하였다. 이를 모형별로 Parameter Tuning진행을 함에 있어서는 경제적으로 안정화된 시기뿐만 아니라 글로벌 금융위기와 같은 큰 위험 속에서 일관되고 안정적인 예측

력을 가지도록 선정하였다. 결과적으로 ROC AUC, PR AUC를 기반으로 모형들을 비교분석한 결과, 과거 공통적으로 뽑히는 재무비율 변수 이외에도 각종 재무제표 변수들 또한 주요 변수로서 유의미하게 뽑히게 되었고, 이를 기반으로 딥러닝 시계열 알고리즘 기반의 부도예측모형이 유용함을 확인하였다.

최종 테스트 결과, 표본내 검증 데이터에서 우수 모형으로 뽑혔던 RF, RNN, LSTM 모형을 바탕으로 비교해보면, 다중판별분석에 의한 선정 변수 기준, ROC AUC 기준 및 PR AUC 기준 RF, RNN 모형이 효과적이었다. 로짓 모형 선정 변수 기준으로는 RF와 RNN 알고리즘이 우수 모형으로 선정, PR AUC의 결과는 RF가 0.390인데 비

〈Table 20〉 Test result based on ROC AUC

Rank	1	2	3	4	5	6	7	8
MDA	RNN	RF	LSTM	KNN	LDA	MLP	LR	DT
Logit	RF	DT	KNN	RNN	LSTM	MLP	LDA	LR
Lasso	RF	LSTM	RNN	KNN	DT	LDA	MLP	LR

ROC AUC High Ranks

〈Table 21〉 Test Result based on PR AUC

Rank	1	2	3	4	5	6	7	8
MDA	MLP	RF	LR	RNN	LDA	KNN	LSTM	DT
Logit	RNN	MLP	KNN	RF	LR	LDA	DT	LSTM
Lasso	LSTM	MLP	RF	KNN	LR	DT	RNN	LDA

PR AUC High Rank

해 RNN이 0.512으로 매우 높았다. Lasso 변수 선정 모형으로는 RF와 LSTM이 우수모형으로 선정되었고, PR AUC는 각각 0.382, 0.412가 나오게 되었다. 결과적으로 ROC AUC는 모두 비슷하게 90전후로 높았으며, 따라서 PR AUC의 순서를 비교해보면 가장 우수한 모형이 비선형 데이터를 고려한 로짓 모형에서의 RNN모형이었으며, 다음으로 선형 데이터 내 다중공선성을 고려하는 Lasso 회귀분석 선별 변수 기반 LSTM모형이었다.

Min(2014) 저자의 연구에 따르면, 불균형 데이터로 이루어진 10억에서 70억 사이인 국내 비외국 기업의 데이터들을 Undersampling기법을 통해서 1,832개로 줄여 1,832개 중 부도기업 수를 916개로 두고 연구를 진행하였다. SVM 앙상블 모형을 통해서 정확도는 75.66%, ROC AUC의 경우는 0.8129였다.

Kwon et al.(2017) 저자의 연구에 의하면

2010~2016년의 유가증권시장, 코스닥시장, 코넥스 시장에 상장된 비 금융기업을 대상으로 Undersampling을 진행하여 약 450개 기업데이터를 대상으로 연구를 진행하였는데, Altman(1968)의 추천 변수군을 이용한 결과는 SVM의 Accuracy가 0.780, ROC AUC가 0.879였으며, MDA는 Accuracy가 0.755, ROC AUC가 0.818이었다. 이에 비해 RNN의 경우는 Accuracy가 0.811, ROC AUC가 0.889로서 가장 높은 결과값을 확인하였다. Kim et al.(2015)의 추천 변수군을 이용한 결과는 SVM의 Accuracy가 0.771, ROC AUC가 0.818였으며, MDA는 Accuracy가 0.732, ROC AUC가 0.782이었다. 이번 추천 변수군 또한 RNN의 경우 Accuracy가 0.828, ROC AUC가 0.891로서 가장 높은 결과값을 확인하였다.

변수 선정 방식, 이용 연구 데이터 기간, 데이터 분석 대상, 샘플링 등이 본 연구와 동일하지

는 않지만, 다양한 변수군을 기반으로 진행된 점은 동일하며 위 결과값 대비 정확도와 ROC AUC 지표의 값은 매우 높은 결과가 나오게 되었다.

Lasso 변수 선정방식을 적용하여 기존 방식에 다중공선성이 제거된 변수군을 추가로 이용하였다. 일반적으로 기계학습시 데이터가 많아야 학습이 유의미한 강점을 가지므로 불균형 데이터 본연 그대로를 사용하였다. 그리고 최종적으로 시계열 알고리즘으로 분석시 가변적으로 데이터를 생성해서 학습에 이용할 수 있었다는 점을 기반으로 데이터가 부족할 수 있는 문제를 해결하였다. 결론적으로는 Kwon et al.(2017)의 연구를 통해 RNN 효과성을 입증하였던 연구를 발판삼아 RNN 그리고 LSTM이 모두 다른 알고리즘에 비해 그 성능이 우수함을 확인하였다.

RNN과 LSTM이 기계학습 알고리즘에서 앙상블 모형으로 유명한 Random Forest에 상응하는 효과 혹은 그 이상을 나타내는 경우가 있음은 앙상블 기반의 알고리즘은 랜덤하게 선정된 다양한 경우에서 알고리즘을 통해 효과를 분석하고 여러 우수한 모형들을 묶어 모델링을 구축하기 때문에 기본적으로 모형의 예측력이 매우 우수하다. 하지만 딥러닝 시계열 알고리즘을 이용한 단일 모형 임에도 충분히 효과적인 결과를 확인할 수 있었다. 따라서 추후 연구로는 딥러닝 시계열 알고리즘 또한 Random Forest와 같은 앙상블 기법을 적용하여 Robust한 모형을 구상해야 할 필요성을 느끼게 되었다. 따라서 추후 이를 보완한 연구를 통해 현재보다 더욱 개선된 성능을 보이는 안정적인 딥러닝 시계열 알고리즘 기반 부도예측모형을 제시할 수 있을 것이다.

4.2 시사점

기존 연구들과는 달리 본 논문에서는 다중공선성을 제거해주는 대표적인 모형인 Lasso 회귀 분석을 변수 선정 방식에 새로 적용하였다. 또한 불균형 데이터를 undersampling 기법을 이용하지 않고 연구를 진행한 케이스로서 데이터를 축소하지 않았고, 따라서 결과적으로 딥러닝 시계열 알고리즘에 이용가능한 가변적인 학습데이터 생성 방식까지 추가 적용한 분석을 적용하였다. 그 결과 기존 모형 대비 유의미한 차이를 보이는 결과를 확인할 수 있었다. 본 연구 결과에서는 분량이 많아 다양한 실험데이터를 담지 않았지만, 가변적인 학습데이터 생성방법 중 데이터가 빈 공간을 직후값으로 채워주는 방식이 아닌 다른 경우는 딥러닝 시계열 알고리즘의 우수성을 발견하기가 어려울 정도로 예측력이 떨어짐을 확인했다. 따라서 딥러닝 시계열 알고리즘 기반의 기업부도 예측모형 구축시 가변적인 데이터 생성 방식은 매우 중요한 요소이다.

추가적으로 기업 부도 모형의 데이터의 원천 데이터를 살펴보고 이를 데이터 분석에 이용하려면 데이터 전처리에 가장 큰 힘을 써야한다는 사실을 알게 되었다. ‘데이터 전처리’는 원천 데이터를 바탕으로 분석에 이용가능한 데이터로 제작하는 과정을 말한다. 연구를 위해서 데이터를 각종 제약조건을 바탕으로 전처리를 진행하였는데 사소한 기준 하나 차이로 남아있어야 할 데이터가 사라져 있거나 어느 순간 변질되어 있음을 확인하게 되었다. 이는 기업들이 모두 공통적인 데이터로 표현되는 부분도 있으나 그렇지 않은 부분들 또한 있었다. 기업별로 시계열 분석이 진행 가능하도록 기존 데이터의 형식을 수정해서 분석을 진행해야했고, 최종적으로 딥러닝

시계열 알고리즘을 분석을 위해서는 기존 모형들과는 다르게 가변적인 데이터 생성을 한 후 연구를 진행할 수 있도록 데이터 전처리를 추가로 진행해야했다.

본 연구에 쓰인 기술적인 부분은 Python언어를 기반으로 scikit-learn, tensorflow, keras 등 대표적인 머신러닝 툴을 이용하였다. 하드웨어적인 측면에서는 또한 회사 내부의 서버를 통해서 연구를 진행하였다. 다행히 기업 부도 분석을 진행하는데 있어 기술적 어려움이나 하드웨어 적인 어려움은 없었다.

따라서 후속연구를 진행하는 사람들에 있어서, Lasso 회귀분석 기반의 변수 선정 방식 및 시계열 딥러닝 알고리즘 이용시 가변적인 데이터 생성법을 잘 이용하길 바란다. 또한 데이터 전처리시 분석에 문제가 발생하지 않도록 다양한 경우를 잘 고려하여, 딥러닝 시계열 알고리즘을 통한 연구가 금융산업에서 자주 진행될 수 있기를 바란다.

참고문헌(References)

- Addal, S., “Financial forecasting using machine learning”, African Institute for Mathematical Science, (2016), 1~32.
- Ahn, S. M., and J. W. Park, “Corporate Bankruptcy Prediction Using Financial Ratios: Focused on the Korean Manufacturing Companies Audited by External Auditors”, Korean Management Review, Vol.43, No.3, (2014), 639~669.
- Altman, E. I., “Financial Ratios, Discriminant Analysis and the Predication of Corporate Bankruptcy”, Journal of Finance, Vol.23, No.4, (1968), 589~609.
- Bae, J. K., “An Integrated Approach to Predict Corporate Bankruptcy with Voting Algorithms and Neural Networks”, Korean Business Review, Vol.3, No.2, (2010), 79~101.
- Beaver, W. H., “Financial ratios as predictors of bankruptcy”, Journal of Accounting Research, Supplement, (1966), 71~102.
- Deakin, E. B., “A Discriminant Analysis of Predictors of Business Failure”, Journal of Accounting Research, Vol.10, No.1, (1972), 167~179.
- Grice, J. S. and M. T. Dugan, “The Limitations of Bankruptcy Prediction Models: Some Cautions for the Researcher”, Review of Quantitative Finance and Accounting, Vol.17, No.2, (2001), 151~166.
- Hong, S. H. and K. S. Shin, “Using GA based Input Selection Method for Artificial Neural Network Modeling; Application to Bankruptcy Prediction”. Journal of Intelligence and Information Systems, Vol.9, No.1, (2003), 227~249.
- Jo, N. O., H. J. Kim and K. S. Shin. “Bankruptcy Type Prediction Using A Hybrid Artificial Neural Networks Model.” Journal of Intelligence and Information Systems, Vol.21, No.3, (2015), 79~99.
- Jo, N. O. and K. S. Shin. “Bankruptcy Prediction Modeling Using Qualitative Information Based on Big Data Analytics”, Journal of Intelligence and Information Systems, Vol.22, No.2, (2016), 33~56.
- Kapinos, P., and O.A. Mitnik, “A Top-Down Approach to Stress-Testing Banks”, Journal of Financial Services Research, Vol.49, No.2,

- (2016), 229~264.
- Kim, G. P., H. K. Lee, J. H. Kim and H. J. Kwon, "The Fourth Industrial Revolution in Major Countries and Growth Strategy of Korea: U.S., Germany and Japan Cases", Korea Institute for International Economic Policy, Policy Analysis, (2017).
- Kim, J. B. and J. S. Lee, "Usability of Cash Flow Data in Predicting Bankruptcy Using Artificial Intelligence Techniques: The Case of Small and Medium Sized Firms", Korean Journal of Business Administration, No.26, (2000), 229~250.
- Kim, M. J., "Ensemble Learning for Solving Data Imbalance in Bankruptcy Prediction", Journal of Intelligence and Information Systems, Vol.15, No.3. (2009), 1~15.
- Kim, M. J., H. B. Kim and D. K. Kang, "Optimizing SVM Ensembles Using Genetic Algorithms in Bankruptcy Prediction", Journal of information and communication convergence engineering, Vol.8, No.4, (2010), 370~376.
- Kim, M. J., "Ensemble Learning with Support Vector Machines for Bond Rating", Journal of Intelligence and Information Systems, Vol.18, No.2, (2012), 29~45.
- Kim, S. B., P. Ji and K. J. Jo, "The Analysis on the Causes of Corporate Bankruptcy with the Bankruptcy Prediction Model", Journal of Market Economy, Vol.40, No.1, (2011), 85~106.
- Kim, S. J. and H. C. Ahn, "Estimation Model applied Random Forest for Corporate Bond Ratings", Journal of Intelligence and Information Systems, Spring Conference, (2014), 371~376.
- Kim, Y. D., C. H. Jun and H. S. Lee, "A new classification method using penalized partial Least squares", Journal of the Korean Data and Information Science Society, Vol.22, No.5, (2011), 931~940.
- Kim, Y. T. and M. H. Kim, "An Artificial Neural Network Model for Business Failure Prediction", Korean Journal of Accounting Research, Vol.6, No.1, (2001), 275~294.
- Kwon, H. K., D. K. Lee and M. S. Shin, "Dynamic forecasts of bankruptcy with Recurrent Neural Network model", Journal of Intelligence and Information Systems, Vol.23, No.3, (2017), 139~153.
- Lee, I. R. and D. C. Kim, "Evaluation of Bankruptcy Prediction Model Using Accounting Information and Market Information", Journal of Korean Finance Association, Vol.28, No.4(2015), 626~666.
- Lee, J. S. and J. H. Han, "Test of Non-Financial Information in Bankruptcy Prediction using Artificial Neural Network - The Case of Small and Medium - Sized Firms -)", Journal of Intelligence and Information Systems, Vol.1, No.1, (1995), 123~134.
- Lee, K. C., "Comparative Study on the Bankruptcy Prediction Power of Statistical Model and AI Models : MDA , Inductive Learning , Neural Network)", Journal of the Korean Operations Research and Management Science Society, Vol.18, No.2, (1993), 57~81.
- Min, S. H., "Bankruptcy prediction using an improved bagging ensemble", Journal of Intelligence and Information Systems, Vol.20, No.4, (2014), 121~139.
- Min, S. H., "Simultaneous optimization of KNN ensemble model for bankruptcy prediction",

- Journal of Intelligence and Information Systems, Vol.22, No.1, (2016), 139~157.
- No, G. M. and W. G. Han, “ICT Policy Direction After 100-days Moon Jae-in government launched.”, National Information Society Agency, Hot Issue Report, (2017).
- Ohlson, J. A., “Financial Ratios and the Probabilistic Prediction of Bankruptcy”, Journal of Accounting Research, (1980), 109~131.
- Park, J. Y., Y. W. Kim and M. Y. Lee, “A Prediction Model of Small Business Bankruptcy”, Journal of Korean Logos Management, Conference, (2007), 202~204.
- Presidential Committee on the Fourth Industrial Revolution, “Data Industry Promotion Strategy - I-KOREA 4.0 Data Field Plan, I-DATA+”, (2017).
- Shapiro, S. S. and M. B. Wilk, “An analysis of variance test for normality (complete samples)”, Biometrika, Vol.52, (1965), 591~611.
- Swedberg, R., “The Structure of Confidence and the Collapse of Lehman Brothers”, Research in the Sociology of Organizations, (2009).
- Tibshirani, R., “Regression Shrinkage and Selection via the Lasso”, Journal of the Royal Statistical Society, Series B (Methodological), Vol.58, No.1, (1996), 267~288.
- Wang, H., Q. Xu and L. Zhou, “Large Unbalanced Credit Scoring Using Lasso-Logistic Regression Ensemble”, PLoS One, San Francisco, Vol.10, No.2, (2015).
- Welch, B. L., “‘Student’ and Small Sample Theory”, Journal of the American Statistical Association, Vol.53, No.284, (1958), 777~788.
- Yeh, S., C. Wang and M. Tsai, “Corporate default prediction via deep learning”, Wireless and Optical Communication Conference, Vol.24, 1~8.
- Zmijewski, M. E., “Methodological issues related to the estimation of financial distress prediction models”, Studies on Current Econometric Issues in Accounting Research, Vol.22, (1984), 59~82.

Abstract

Corporate Default Prediction Model Using Deep Learning Time Series Algorithm, RNN and LSTM

Sungjae Cha* · Jungseok Kang**

In addition to stakeholders including managers, employees, creditors, and investors of bankrupt companies, corporate defaults have a ripple effect on the local and national economy. Before the Asian financial crisis, the Korean government only analyzed SMEs and tried to improve the forecasting power of a default prediction model, rather than developing various corporate default models. As a result, even large corporations called 'chaebol enterprises' become bankrupt. Even after that, the analysis of past corporate defaults has been focused on specific variables, and when the government restructured immediately after the global financial crisis, they only focused on certain main variables such as 'debt ratio'. A multifaceted study of corporate default prediction models is essential to ensure diverse interests, to avoid situations like the 'Lehman Brothers Case' of the global financial crisis, to avoid total collapse in a single moment.

The key variables used in corporate defaults vary over time. This is confirmed by Beaver (1967, 1968) and Altman's (1968) analysis that Deakins'(1972) study shows that the major factors affecting corporate failure have changed. In Grice's (2001) study, the importance of predictive variables was also found through Zmijewski's (1984) and Ohlson's (1980) models. However, the studies that have been carried out in the past use static models. Most of them do not consider the changes that occur in the course of time. Therefore, in order to construct consistent prediction models, it is necessary to compensate the time-dependent bias by means of a time series analysis algorithm reflecting dynamic change.

Based on the global financial crisis, which has had a significant impact on Korea, this study is conducted using 10 years of annual corporate data from 2000 to 2009. Data are divided into training data, validation data, and test data respectively, and are divided into 7, 2, and 1 years respectively. In order to construct a consistent bankruptcy model in the flow of time change, we first train a time series deep learning algorithm model using the data before the financial crisis (2000~2006). The parameter tuning of

* research engineer, AIZEN GLOBAL

** Corresponding Author: Jungseok Kang

CEO, AIZEN GLOBAL

AIZEN GLOBAL, 30, Eunhaeng-ro, Yeongdeungpo-gu, Seoul, 07242, Korea

Tel: +82-10-3579-5738, E-mail: js.kang@aizen.co

the existing model and the deep learning time series algorithm is conducted with validation data including the financial crisis period (2007~2008). As a result, we construct a model that shows similar pattern to the results of the learning data and shows excellent prediction power. After that, each bankruptcy prediction model is restructured by integrating the learning data and validation data again (2000 ~ 2008), applying the optimal parameters as in the previous validation. Finally, each corporate default prediction model is evaluated and compared using test data (2009) based on the trained models over nine years. Then, the usefulness of the corporate default prediction model based on the deep learning time series algorithm is proved. In addition, by adding the Lasso regression analysis to the existing methods (multiple discriminant analysis, logit model) which select the variables, it is proved that the deep learning time series algorithm model based on the three bundles of variables is useful for robust corporate default prediction.

The definition of bankruptcy used is the same as that of Lee (2015). Independent variables include financial information such as financial ratios used in previous studies. Multivariate discriminant analysis, logit model, and Lasso regression model are used to select the optimal variable group. The influence of the Multivariate discriminant analysis model proposed by Altman (1968), the Logit model proposed by Ohlson (1980), the non-time series machine learning algorithms, and the deep learning time series algorithms are compared.

In the case of corporate data, there are limitations of 'nonlinear variables', 'multi-collinearity' of variables, and 'lack of data'. While the logit model is nonlinear, the Lasso regression model solves the multi-collinearity problem, and the deep learning time series algorithm using the variable data generation method complements the lack of data.

Big Data Technology, a leading technology in the future, is moving from simple human analysis, to automated AI analysis, and finally towards future intertwined AI applications. Although the study of the corporate default prediction model using the time series algorithm is still in its early stages, deep learning algorithm is much faster than regression analysis at corporate default prediction modeling. Also, it is more effective on prediction power. Through the Fourth Industrial Revolution, the current government and other overseas governments are working hard to integrate the system in everyday life of their nation and society. Yet the field of deep learning time series research for the financial industry is still insufficient. This is an initial study on deep learning time series algorithm analysis of corporate defaults. Therefore it is hoped that it will be used as a comparative analysis data for non-specialists who start a study combining financial data and deep learning time series algorithm.

Key Words : Optimal Feature Selection, Lasso Regression, Deep Learning Time Series Algorithm, Corporate Bankruptcy, RNN, LSTM

Received : October 24, 2018 Revised : November 16, 2018 Accepted : November 26, 2018

Publication Type : Regular Paper(Fast-track) Corresponding Author : Jungseok Kang

저 자 소개



차성재

성균관대학교 수학과, 그리고 한국과학기술원(KAIST) 금융전문대학원 금융공학부 석사 과정 재학생이며, 현재 (주)에이젠글로벌에서 연구원으로 재직 중이다. 주요 관심분야는 인공지능 응용, 빅데이터 분석/비즈니스 애널리틱스, 데이터마이닝 등이다.



강정석

서울대학교 언어학/경제학 학사, 석사 및 미국 시카고대학 MBA (Chicago Booth School of Business MBA) 를 졸업하였다. 현재 인공지능 금융 회사인 (주)에이젠글로벌 대표이사를 역임중이다. 또한 LG CNS, Citi Group, 국회재경위 경력을 겸한 금융/IT 전문인력으로 현재 한국정보화진흥원 AI 전문위원 및 디지털금융연구센터 공식 어드바이저로 활동중이다. 주요 관심 분야는 금융, 머신러닝, 핀테크, AI application 등이다.