논문 22-47-01-13 The Journal of Korean Institute of Communications and Information Sciences '22-01 Vol.47 No.01
https://doi.org/10.7840/kics.2022.47.1.98

오차 앙상블모형을 활용한 기업 신용평가모형의 비교 연구

김 용 환◆, 김 도 형*, 허 재 혁**, 김 광 용°

A Study on Corporate Credit Rating Model Using Ensemble Model Based on Error

Yong-hwan Kim◆, Do-hyung Kim*, Jae-hyuk Heo**, Gwang-yong Gim°

요 약

본 연구에서는 기업 신용평가모형의 개선을 연구하기 위하여 로지스틱 회귀, 랜덤 포레스트, 그래디언트 부스팅 3개 단일모형과 로지스틱 회귀과 그래디언트 부스팅의 앙상블모형을 최적화하고, 각 모형의 변별력을 비교하였다. 전체 33,317 기업 중 1,295 기업이 표본으로 활용하였으며, 108개 재무비율을 검토하여, 31개의 유의한 변수를 선별하여 각 모형별로 최적화된 모형을 도출하였다. 실증분석 결과, 오차 앙상블모형은 랜덤 포레스트와 그래디언트 부스팅과 비교하여 일부 지표에서만 우수한 성능을 가지는 한계가 있었으나, 로지스틱 회귀보다 전반적인 성능이 개선되었으며, 로지스틱 회귀가 가지는 해석력을 강화시킨다는 관점에서 신용평가모형으로써의 활용 가능성을 확인할 수 있었다.

Key Words : Corporate credit rating model, Logistic Regression, Random Forest, Gradient Boosting

ABSTRACT

In order to study the improvement of the corporate credit rating model, the three single models of logistic regression, random forest, and gradient boosting and the error ensemble model combined logistic regression and gradient boosting were optimized and the discriminative power of each model was compared in this study. 1,295 companies out of 33,317 companies were used as samples. Each models are optimized 31 significant variables selected by reviewing 108 financial ratios. As a result of empirical analysis, error ensemble model had a limitation that it had excellent performance only in some indicators compared to random forest and gradient boosting. But error ensemble model's overall performance was improved compared to logistic regression, and from the viewpoint of strengthening the interpretive power of logistic regression, its applicability as a credit rating model could be confirmed through this study.

Ⅰ. 서 론

기업 신용평가모형은 바젤 내부등급법에 따라 그동안 은행업권을 중심으로 체계적으로 발전해왔다. 기업 신용평가모형은 기업에 대한 신용도를 측정하며, 대출심사 및 내부등급법에 따른 위험가중자산, BIS비율 등에 활용이 되고 있다. 기업 신용평가모형은 신용 리스크의 가장 중요한 요소이며, 다양한 업무와 연관이

◆ First Author : Graduate School of IT Policy and Management, Soongsil University, berlnui@daum.net, 정회원
° Corresponding Author : Soongsil University, gygim@ssu.ac.kr, 정회원
* Graduate School of IT Policy and Management, Soongsil University, dhykim1@naver.com
** Graduate School of IT Policy and Management, Soongsil University, picro@daum.net
논문번호 : 202108-202-0-SE, Received August 14, 2021; Revised September 23, 2021; Accepted October 26, 2021

있어, 부도의 예측력이 우수하고, 인과관계의 설명이 용이한 로지스틱 회귀모형이 주로 활용되고 있다.

알고리즘의 발전에 따라 다양한 머신러닝 모형이 여러 분야에 활용이 되고 있다. 본 연구에서는 표본 추출된 1,295개 기업에 대해 로지스틱 회귀(Logistic Regression)와 대표적인 머신러닝(machine learning) 모형인 랜덤 포레스트(Random forest), 그래디언트 부스팅(Gradient Boosting)을 활용하여 기업의 부도를 예측하였다. 각 모형의 변별력 및 차이점을 비교하였으며, 로지스틱 회귀와 그래디언트 부스팅을 결합한 오차 앙상블(Ensemble)모형을 도출하여 예측 성능의 향상과 그 한계점에 대해서 살펴보았다.

Ⅱ. 연구배경 및 연구모형

2.1 기업 신용평가모형 현황

기업 신용평가모형은 기업의 채무상환능력을 평가하는 것을 목적으로 하며, 신용등급으로 산출되어 표 1과 같이 금융회사의 다양한 의사결정에 활용이 되고 있다[1]. 기업 신용평가모형은 금융회사의 다양한 업무에 활용되고 있어 각 이해관계자들이 신뢰할 수 있는 충분한 정보가 제공될 수 있어야 한다.

바젤Ⅱ 내부등급법에 따라 금융회사들은 대출 자산 등에 대해 은행 내부의 신용평가모형을 활용하여 위험가중자산을 산출하고 있으며, 바젤Ⅱ에서는 '은행은 내부검증 절차를 통해 내부등급과 리스크 측정 시스템의 성과를 지속적으로 의미있게 평가 하고 있음을 감독기관에 반드시 보여 줄 수 있어야 한다'라고 정의하여 각 금융기관에 체계적이고 투명한 신용평가관리 체계를 요구하고 있다[2].

바젤Ⅱ를 위하여 금융감독원에서는 2005년 "신용리스크 내부등급법 기본 세부지침(안)"을 마련하여 신용평가시스템의 구축 및 운영, 적합성검증에 대한 구체적인 지침을 제시하고 있으며, 은행업감독규정에는 신용평가모형을 비롯한 리스크 측정요소의 변경 등에 대한 세부적인 규정이 반영되어 있다.

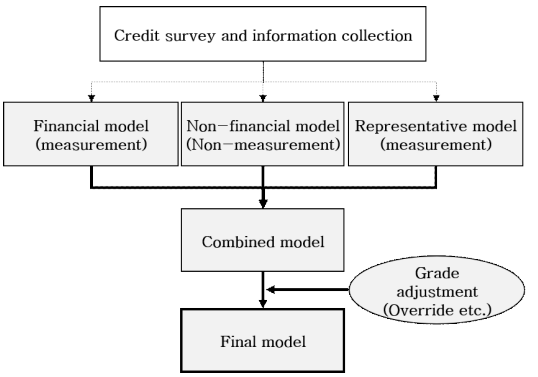
기업 신용평가의 기준 및 등급부여, 성과측정 등에 대한 세부적인 사항은 2008년 "바젤 II 下의 통합리스크관리 모범규준"에 구체적으로 기술이 되어 있다. 기업 신용평가모형에서는 재무정보 및 재무비율 등 양적지표와 이익 및 현금흐름의 질, 산업전망, 경영진 신뢰성 등 질적지표가 활용되고 있다[3].

기업 신용평가모형은 재무제표를 활용한 재무모형, 심사자의 신용조사를 통한 비재무모형, 대표자의 신용정보를 활용한 대표자모형으로 구성이 되며, 각 모형의 결합을 통해 최종 신용등급이 산출되게 된다[1]. 그 구성은 그림 1과 같다.

바젤 내부등급법에 따라 신용평가모형 관리에 대한 중요성이 강조되었으며, 이를 통해 기업 신용평가모형은 체계적으로 발전해왔다. 그 동안 기업 신용평가모형은 로지스틱 회귀모형을 활용하여 장기적인 경기변동에도 안정적인 성능을 가질 수 있도록 모형의 정교화에 중점을 두었으며, 최근에는 우리은행의 빅데이터 및 머신러닝 기법을 반영한 기업진단시스템 Big Eye와 같이 머신러닝 기법을 기업 신용평가모형에 반영하고자 하는 노력이 진행 중이다[1].

표 1. 기업 신용평가모형의 활용 영역
Table 1. Areas of corporate credit rating model's application

application area	details
Loan screening	Decisions on loan execution and interest rates
Calculation of risk-weighted assets	Calculations of risk-weighted assets and BIS equity capital ratio based on credit rating grade
Calculation of allowance for bad debt	Calculation the loan loss provisions based on credit rating grade



출처 : 최성민(2018), "기업 신용평가모형의 현황과 변화 트렌드", 한국신용정보원, p10 내용수정 및 재구성

그림 1. 기업 신용평가모형의 구성
Fig. 1. Composition of corporate credit rating model

2.2 선행연구

Altman(1968)은 미국의 제조업 기업을 대상으로 다변량 판별분석을 활용하여 부도기업을 예측하는 Z-score모형을 제시하였다. Altman(1968)은 22개 재무비율을 검토하였으며, 운전자본비율, 이익잉여금,

영업이익, 부채비율, 총자산회전율의 5개 재무비율을 모형에 활용하였다.

Ohlson(1980)은 Logit분석을 이용하여 부도가능성을 예측하는 O-score모형을 제시하였으며, 105개 부도기업과 2,058개 정상기업을 표본으로 연구하였다. Ohlson(1980)은 기업 규모, 재무 구조, 성과 및 유동성이 부도에 밀접한 관련이 있는 요인임을 실증분석하였다.

안성만과 박종원(2014)은 다중 로지스틱 회귀분석을 이용하여, 52,052개 외감기업에 대한 부도예측모형을 연구하였다. 금융비용대부채비율, 자기자본비율, 차입금의존도, 건설업더미 등 17개 변수가 부도예측에 유의한 결과를 보였으며, 부도예측모형의 구축시에 산업별 특성이 반영되어야 함을 실증분석하였다.

김성진과 안현철(2017)은 랜덤 포레스트를 활용하여 1,295개 제조업 기업에 대한 부도예측 모형을 연구하였다. 김성진과 안현철(2017)은 랜덤 포레스트모형과 인공신경망, SVM모형을 비교하였으며, 이를 통해 랜덤 포레스트모형의 변별력이 좀 더 유의함을 실증분석하였다.

권혁진(2017)은 다중 로지스틱 회귀분석을 통해 연결재무제표와 개별재무제표를 통해 부도예측모형의 연구를 하였으며, 부도예측에 있어서 연결재무제표의 중요성을 확인하였다.

2.3 연구모형

기업 신용평가모형은 바젤 내부등급법에 따라 금융기관 내부의 많은 의사결정에 활용이 되며, 감독기관 등과 밀접한 관련이 있으므로 예측력이 우수하고 모형 결과의 설명이 용이한 로지스틱 회귀모형이 주로 활용이 되고 있다.

랜덤 포레스트는 2001년 Breman에 의하여 개발된 분류기법으로, 기존 단일 의사결정나무를 여러 개로 확장한 형태의 머신러닝 기법이다. 여러 개의 의사결정나무를 종합하여, 예측을 수행하게 되므로 정확도와 안정성이 높아지게 된다[4].

그래디언트 부스팅은 2017년 Breman에 의하여 처음 소개가 되었으며, Jerome H Friedman에 의하여 발전이 되었다[5]. 그래디언트 부스팅은 약한 분류기를 결합하여 강한 분류기를 만드는 앙상블모형으로 이전 모형의 예측오류를 보완하는 형태로 학습을 반복 수행하게 되므로 수치예측 및 분류예측에 높은 성능을 보이고 있다[6].

각 모형의 장점 및 단점은 표 2와 같다.

본 연구에서는 선행연구를 바탕으로 전통적인 로지스틱 회귀와 머신러닝 기법인 랜덤 포레스트, 그래디언트 부스팅을 활용한 신용평가모형을 도출하고 각 모형 간의 차이점을 비교하고자 한다. 또한 로지스틱 회귀와 그래디언트 부스팅의 결합을 통해, 로지스틱 회귀의 장점인 높은 해석력과 안정적인 예측력, 그리고 그래디언트 부스팅의 장점인 높은 예측력을 결합한 오차 앙상블모형의 성능에 대해 살펴보고자 한다.

표 2. 각 모형의 장점 및 단점
Table 2. Each model's Advantages and disadvantages

Category	Advantage	disadvantage
Logistic Regression	• Easy interpretation throu regression formula • Stable predictive power	• Assumptions about linearity and correlations should be considered • Low predict power about high varianced population
Random Forest	• Excellent generalization and predictive power throu the combination of decision trees	• Difficulty of Interpretation and overfitting due to complex trees
Gradient Boosting	• High predictive power through prediction of residuals	• Difficulty of Interpretation and overfitting

Ⅲ. 연구의 설계

본 연구에서는 기업 신용평가모형의 하부모형 중 계량적인 접근을 통한 모형의 비교를 위하여 재무평가영역을 대상으로 하였으며, NICE의 KIS-DATA를 활용하였다.

KIS-DATA에서는 업체개요, 재무제표, 재무비율, 법정관리 및 화의정보 등을 제공하고 있으며, 본 연구를 위해서 2015년부터 2020년까지 외감이상 기업 전체에 대한 데이터를 입수하였다.

3.1 개발 모집단의 정의

재무비율을 통한 기업의 신용평가시, 분석대상 기업의 재무비율을 동질적인 경제환경하에서 활동을 하는 동종기업들과 비교를 해야 하며[7], 통계청에서는 산업활동의 유사성을 바탕으로 표준산업분류 코드를 통해 전체 업종을 21개 대분류로 분류하고 있다[8].

이에 본 연구에서는 평가모형의 일관성을 위하여

표준산업분류 대분류를 준용하여 제조업으로 분류된 기업을 대상으로 하였다.

추가로 재무제표의 연속성을 위하여 당해연도 및 직전년도 결산재무제표를 보유한 기업으로 개발 모집단을 정의하여, 2016년~2019년 결산재무제표를 보유한 제조업 기업 33,317건을 개발 모집단으로 하였다.

3.2 부도기업의 정의

본 연구는 재무비율을 통한 부도예측 모형의 비교를 목적으로 하며, 이에 부도의 정의는 연구에 있어서 매우 중요한 부분이다.

Campbell et al.(2008)은 부도의 관측기간을 36개월로 제시를 하고 있으나, 이인로&김동철(2015) 및 권혁진(2017) 등 최근 대부분 연구들은 부도의 관측기간을 1년으로 보고 있으며, 이에 본 연구도 부도의 관측기간을 1년으로 하였다.

부도의 정의는 바젤협약에 따라 90일 이상 연체를 하거나, 채무자가 은행에 채무상환을 지연 또는 회피하기 위해 파산 또는 유사한 조치를 취하는 경우로 정의를 하고 있다[3]. 연체 일수에 대한 정보는 NICE에서 제공이 되지 않고 있다. 이에 본 연구에서는 파산, 워크아웃 등 법정관리 정보를 부도로 평가하였다. 추가로 전액자본 잠식은 부실의 중요한 요소로 판단을 하고 있으며[9], 전액 자본잠식은 부실기업 상장폐지의 실질적인 부도발생 요건으로 인식을 하고 있다[10]. 이에 본 연구에서는 선행연구를 바탕으로 법정관리 정보 및 전액자본 잠식을 부도의 요건으로 하였으며, 그 외 기업은 정상기업으로 분류하였다.

관측시점에 부도요건에 해당하는 기업은 기부도로 분류하여 개발 대상에서 제외하였으며, 최초 부도가 색인된 이후에 정상화가 되었더라도 이후 개발 대상에서는 제외하였다.

3.3 표본 데이터

표본 데이터는 샘플링을 통해 정상기업수와 부도기업수를 5:5로 표본 데이터를 정의하였으며, 모형의 개발 및 검증을 위하여 표본 데이터를 7 : 3으로 개발 데이터와 검증 데이터를 분리하였다.

표본 추출은 결산연도를 기준으로 층화 단순 임의추출(Stratifed random sampling)하여 균형이 잡힌 표본 데이터가 추출될 수 있도록 하였다.

개발 데이터의 부도율은 표 3과 같으며, 표본 데이터의 현황은 표 4와 같다.

표본 데이터는 업체명, NICE에서 업체별로 부여하는 업체코드, 결산년월, 부도여부 및 108개 재무비율을 변수로 가지고 있으며, 부도여부는 부도기업은 1, 정상기업은 0의 값을 부여하였다. 분석에 활용한 재무비율은 표 5와 같다.

표 3. 개발 데이터 부도율 현황
Table 3. Default rate of Development population

	Total number of companies	Number of default companies	Default rate
Total	73,028	1,891	2.6%
Manufacturing industry	33,317	647	1.9%

표 4. 표본 데이터 분포
Table 4. Sampling data Distribution

	Normal Companies	Default Companies	Total
Development	453	453	906
Validation	195	194	389
Total	648	647	1,295

표 5. 재무비율 변수 현황
Table 5. Financial ratio variables

Variable	Variable Name
X1	Growth rate of total assets
X2	Growth rate of property, plant and equipment
X3	Growth rate of current assets
X4	Growth rate of inventories
X5	Growth rate of stockholders' equity
X6	Growth rate of sales
X7	Growth rate of operating profit
X8	Growth rate of net income before tax
X9	Growth rate of net income
X10	Increase in number of employees
X11	Operating profit / total capital
X12	Net income before tax / total capital
X13	Net income / total capital
X14	Net income / total capital
X15	Corporate net profit margin
X16	Operating return on operating capital
X17	Return on Equity Ratio before tax
X18	Return on Equity Ratio
X19	Return on Capital Ratio before tax
X20	Return on Capital Ratio
X21	Net Sales before tax margin on sales

X22	Net Sales margin on sales
X23	Gross profit margin on sales
X24	Operating profit margin on sales
X25	Resin ratio
X26	Cost of sales ratio
X27	Depreciation expenses ratio
X28	Depreciation expenses/total expense ratio
X29	personnel expense/total expense ratio
X30	Taxation/Net Income Before Tax Ratio
X31	Taxation/total expense ratio
X32	Financial expense/total debt ratio
X33	Average interest rate on borrowings
X34	Financial expense/total expense ratio
X35	Financial expense/sales ratio
X36	Cash flow and compensation ratio for business activities
X37	Operating profit interest compensation ratio
X38	Current interest and compensation ratio
X39	Net profit and compensation ratio before corporate tax deduction
X40	Dividend ratio
X41	Dividend payout ratio
X42	Debt Service Coverage Ratio
X43	Debt Service coefficient
X44	Loan efficiency coefficient
X45	EBITDA/Sales(pre-tax profit)
X46	EBITDA/Sales
X47	EBITDA/financial expenses
X48	Shareholders' Equity ratio
X49	Current ratio
X50	Quick ratio
X51	Cash ratio
X52	Non-current assets ratio
X53	Non-current asset long-term suitability ratio
X54	Debt Ratio
X55	Current Liabilities Ratio
X56	Non-current debt ratio
X57	Reliance on borrowings
X58	borrowings / stockholders' equity
X59	borrowings / Sales ratio
X60	Trade receivable / trade payable ratio
X61	Trade receivable / Goods ratio
X62	Trade payable / Inventories ratio
X63	Inventories / Net working capital
X64	Non-current liabilities / Net working capital
X65	Net working capital / Capital
X66	Reserve ratio
X67	Retention ratio
X68	Retention / total Assets
X69	Reserved amount/Paid-in capital ratio
X70	Total C/F to Debt Ratio
X71	Total C/F to Borrowing ratio
X72	Total C/F to Capital ratio
X73	Total C/F to sales Capital ratio
X74	Net C/F to Borrowing ratio
X75	Total Capital Turnover
X76	Equity Turnover Ratio
X77	Capital Turnover Ratio
X78	Net Working Capital Turnover Ratio
X79	Management Capital Turnover Ratio
X80	Non-current assets Turnover Ratio
X81	Tangible assets Turnover Ratio
X82	Inventory Turnover ratio
X83	Goods Turnover Ratio
X84	Raw materials Turnover Ratio
X85	Work in process Turnover Ratio
X86	Accounts receivable Turnover Ratio
X87	Payable Turnover Ratio
X88	adjusted Inventory Turnover
X89	Net operating capital turnover
X90	Value added per employee
X91	Sales per employee
X92	Net income before tax per employee
X93	Net income per employee
X94	personnel expense per employee
X95	Labor equipment ratio per employee
X96	Mechanical equipment ratio per employee
X97	Capital intensity per employee
X98	Total capital investment efficiency
X99	Gross value-added to property, plant and equipment
X100	Gross value-added to machinery & equipment
X101	Value-added ratio
X102	Labor income share
X103	Composition ratio of net income before tax
X104	Composition ratio of personnel expense
X105	Composition ratio of Financial expenses

X106	Composition ratio of Rent fee
X107	Composition ratio of Tax Service
X108	Composition ratio of depreciation expenses

3.4 연구방법

본 연구는 앞서 정의한 개발 모집단을 대상으로 NICE에서 입수한 108개 재무비율을 변수 변환, 단변량, 상관관계 분석을 하였으며, 이를 통해 최종 변수를 선별하였다, 예측모형의 성능에 대한 측정은 연속형에 대한 평가지표인 AR(Accuracy Ratio)과 K-S(Kolmogorove-Simirnov)통계량과 이진 분류에 대한 평가지표인 Accuracy, Recall, Precision, F1 Score를 활용하였다.

3.4.1 변수 변환 및 최종 후보 변수 선별

(1) 극단치의 처리

극단치는 T검증 등 통계분석에 유의하지 않은 영향을 미칠 수가 있다[11]. 본 연구에서는 각 원 재무비율은 하위 2%, 상위 98%의 해당하는 값으로 극단치를 조정하였다.

(2) 재무비율의 범주 및 변환

재무비율을 통한 기업의 부도를 예측한 선행연구에서는 재무비율을 각 범주로 분류하여 예측모형에 활용하였다. 이인로(2015), 최정원(2019)는 재무비율을 6개 범주로 분류를 하였으며, 권혁진(2017)은 7개로 분류하였다. 선행연구를 바탕으로 본 연구에서는 재무비율을 표 6와 같이 6개 범주로 분류하였다.

원 재무비율은 비선형성 및 비단조성의 특성으로 인하여 로지스틱 회귀모형에 활용하기 위해서는 변환이 필요하다[12]. Eric Falkenstein et al.(2000)은 재무비율은 50개의 서열화된 구간으로 구분하여, 각 구간별 실측부도율을 비모수적인 방법으로 보정한 추정부도율을 기업 신용평가에 설명변수로 활용하는 Mini-Modeling방식을 제시하였으며, 권혁진(2017)은 Mini-Modeling을 활용하여 재무비율은 30개의 서열화된 구간으로 구분하고, 비모수적 방법인 Loess(Local Weighted Regression)를 활용하여 평균추세를 조정한 추정부도율을 변환 재무비율로 활용하였다. 대부분의 은행에서도 기업 신용평가모형 개발시 Mini-Modeling방식을 통해 재무비율을 변환하여 활용하고 있다[13].

본 연구에서는 선행연구를 바탕으로 각 재무비율을 50개의 서열화된 등구간으로 구분하고, 각 구간의 실측부도율을 Loess변환을 통해 추정 재무비율로 변환한 변환 재무비율을 연구에 활용하였다. 그림 2와 같

표 6. 각 범주별 재무비율 현황
Table 6. Financial ratios by each categories

Category	No of variables	Major Financial Ratios
Soundness	26	Shareholders' Equity ratio, Quick ratio, Non-current assets ratio, Trade receivable / trade payable ratio, Retention ratio, Reserved amount/Paid-in capital ratio, etc.
Profitability	17	Average interest rate on borrowings, Dividend payout ratio, Debt Service Coverage Ratio, EBITDA/Sales, EBITDA/financial expenses, etc.
Activity	13	Total Capital Turnover, Inventory Turnover ratio, Payable Turnover Ratio, etc.
Productivity	10	Value added per employee, Net income before tax per employee, Gross value-added to property, plant and equipment, etc.
Growth potential	37	Growth rate of total assets, Growth rate of stockholders' equity, Growth rate of sales, Increase in number of employees, etc.
Cash flow	5	Total C/F to Debt Ratio, Total C/F to Borrowing ratio, Net C/F to Borrowing ratio, etc.
Total	108	

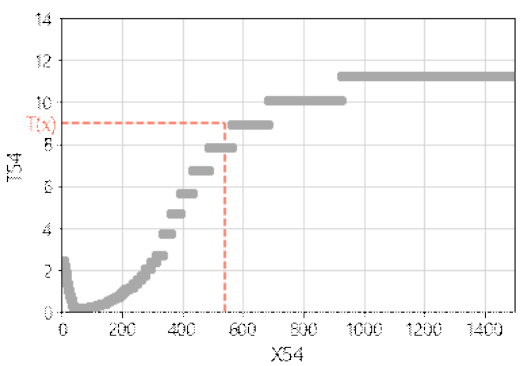


그림 2. 원 재무비율의 변환
Fig. 2. Conversion of the original financial ratio

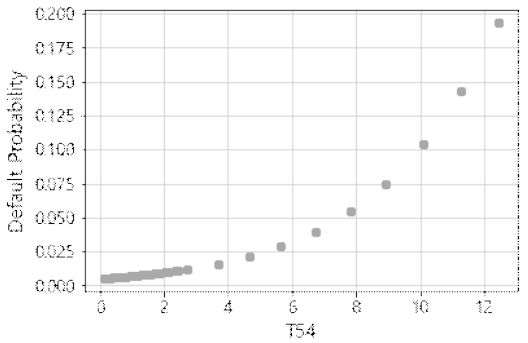


그림 3. 변환 재무비율의 선형성
Fig. 3. Linearity of the conversion financial ratio

은 과정을 통해 원 재무비율(X)이 변환 재무비율(T)로 변환이 되며, 연구에 활용된 변환 재무비율 변수들이 부도에 대해서 선형성을 가지고 있음을 그림 3과 같이 확인하였다.

(3) 단변량 분석

단변량 분석을 통해 각 변환 재무비율이 정상기업과 우량기업을 구분하는데 유의한지 검증하였다. 권혁진(2017)은 T-Test를 통해 정상기업 및 부도기업간의 평균의 차이가 있는지를 검토하였으며, 부도를 종속변수로 하는 단순 로지스틱 회귀분석을 통해 추정된 회귀계수가 유의성을 가지는지 검토하였다. 또한 각 변환 재무비율의 AR을 통해서 변별력의 유의성을 검토하였다. 김종윤(2019)는 단변량분석에서 K-S통계량을 활용하였다.

본 연구에서는 선행연구를 바탕으로 변환 재무비율에 대해서 T-Test, 단순 로지스틱 회귀분석, AR 및 K-S통계량을 활용하여 변수의 변별력을 검토하였다.

(4) 상관분석

부도 예측모형에 있어서 유사한 변수들이 모형에 포함되는 경우 모형의 안정성이 저하될 위험이 존재한다[12]. 본 연구에서는 개발 모형의 안정성 및 모형의 간결성 확보를 위하여 단변량 분석을 통과한 변수들에 대해서 각 범주별로 상관분석을 진행하였다. 추가로 각 범주별 상관분석을 통과한 변수들에 대해서 전체 변수 간의 상관관계를 검토하여 최종 후보 변수들을 선별하였으며, 최종 변수 선별에 대한 기준은 표 7와 같다.

표 7. 최종 변수 선별 기준
Table 7. Final Variable Selection Criteria

Step	Major Financial Ratios
1.Outlier adjustment	Adjust the outliers value to the values corresponding to the bottom 2% and top 98%
2.Univariate analysis	1.Satisfy T-TEST 2.Satisfy the significance of the regression coefficient of the simple logistic regression model 3.AR 20 or more & K-S 15 or more
3.Correlation analysis	Grouping variables that have high correlation(over 0.7) and select the variable with the highest discriminative power for each group

3.4.2 연구모형의 도출

본 연구에서는 3개 단일모형과 1개 앙상블모형을 도출하였다.

단일모형은 다중 로지스틱 회귀, 랜덤 포레트스, 그래디언트 부스팅을 활용하였다. 다중 로지스틱 회귀모형은 단계 선택방식으로 최종 변수를 선별하여 최종 모형을 도출하였으며, 랜덤 포레스트 및 그래디언트 부스팅모형은 의사결정나무의 개수 및 학습률 등 파라미터를 최적화한 모형을 도출하였다.

앙상블모형에서는 다중 로지스틱 회귀모형을 기반으로 그래디언트 부스팅모형을 결합한 새로운 모형을 연구하였다. 다중 로지스틱 회귀모형에서 예측된 부도확률값과 실측부도값과의 차이를 그래디언트 부스팅을 통해 예측하는 모형을 도출하였으며, 두 모형을 결합하여 부도예측의 성능을 높이고자 하였다.

3.4.3 모형의 적합성 평가

모형의 적합성 평가는 연속형과 이진 분류에 대한 2가지로 구분이 되며, 본 연구에서는 연속형에 대한 평가지표인 AR과 K-S, 그리고 이진 분류에 대한 평가지표인 Accuracy, Precision, Recall, F1 Score를 통해 모형의 적합성을 평가하였다.

(1) 연속형 모형의 평가

신용등급과 같은 연속형 모형에 대한 상대적인 서열화에 대한 성과측정치로 활용되는 지표에는 AR과 K-S통계량이 활용된다[3,14].

AR은 AUROC(Area Under ROC)를 다른 방식으로 나타내는 지표이며, AUROC는 ROC(Receiver

Operator Characteristic)곡선 아래의 면적을 의미하며, 값이 클수록 등급에 따른 우량과 불량의 변별이 잘 이루어지고 있음을 의미하며[3], 식(1), 식(2)와 같이 계산된다.

$$AR = 2 \times AUROC - 1 \tag{1}$$

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

$Good_i\%$는 전체 우량고객 중 i등급에 속한 우량고객 비중을, $Bad_i\%$는 전체 불량고객 중에 i등급에 속한 불량고객 비중을 $cum\, Good_i$는 낮은 등급에서 i등급까지 누적 계산된 $Good_i\%$를 의미한다.

K-S통계량은 모형의 변별력이 극대화되는 지점을 측정하여 평가하는 지표이며, 우량집단과 불량집단의 누적분포 차이의 최대값으로 산출하며[14], 식(3)과 같이 계산된다.

$$KS = MAX[|cum\, Good_i\% - cum\, Bad_i\%|] \tag{3}$$

(2) 이진 분류 모형의 평가

이진 분류에서는 일반적으로 두 개의 클래스인 Positive와 Negative에 대해서 표 8의 혼동행렬을 통해 예측 오류의 유형과 그 수준에 대해서 평가를 할 수 있다.

Accuracy는 전체 예측한 클래스와 실제 클래스가 일치하는 비율을 의미하며 식(4)와 같다.

Precision은 Positive로 예측한 클래스 중 실제 Positive인 클래스의 비율을 의미하며, 하며 식(5)과 같다. 본 연구에서 낮은 Precision은 부도로 예측한 차주 중 우량고객의 비율이 높음을 의미하며, 부도차주로 잘못 예측된 우량차주에 대해서 대출이 실행되지 못한 기회손실이 발생된다.

Recall은 실제 Positive인 클래스 중에 Positive로 예측된 클래스의 비율을 의미하며 식(6)와 같다. 본 연구에서 낮은 Recall은 실제 부도차주를 우량고객으로 예측한 비율이 높음을 의미하며, 우량차주로 잘못 예측된 부도차주에 대해 대출실행시 실질적인 금융손실이 발생하게 된다.

F1 Score는 Precision과 Recall의 조화평균을 의미한다. 높은 F1 Score를 얻기 위해서는 Precison과 Recall이 모두 높아야 하며, 식(7)과 같이 산출이 된다.

표 8. 혼동행렬
Table 8. Confusion Matrix

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

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\mathrm{Precisionl} = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1\, Score = 2 \times \frac{\mathrm{Precision} \times Recall}{\mathrm{Precision} + Recall} \tag{7}$$

Ⅳ. 실증 분석 결과

4.1 기술통계량

108개 변환 재무비율을 통해 31개 변수가 도출이 되었으며, 그 결과는 표 9과 같다. 최종 후보 변수들의 평균, 표준편차, 중위수에 대한 통계량을 살펴보면, 각 변수의 평균은 1.89~2.61사이에 분포하고 있으며, 표준편차도 0.63~3.59이내에 분포하고 있어 Mini-Modeling을 통해 각 변수의 크기가 조정되었음을 알 수 있다.

T-test 및 단순 로지스틱 회귀분석의 회귀계수의 유의수준이 0.05에서 유의하며, 각 변수별 변별력은 비유동자산비율, 총C/F(Cash flow)대부채비율, 유보액/납입자본비율, 부채비율, 자기자본비율이 AR이 60% 이상으로 높게 산출이 되었다.

각 변수들간의 상관계수도 0.7이하로 최종 후보변수들의 선택이 유의하였다고 판단된다.

각 범주별로 살펴보면, 표 10와 같이 건전성 및 수익성 관련된 변수가 총 17개로 54%를 차지하고 있으며, 변별력은 건전성과 현금흐름이 AR이 50%이상으로 높게 나타났다.

표 9. 최종 후보변수 통계량
Table 9. Statistics of the final candidate variable

Category	Variable	Variable Name	Mean	Standard deviation	Med-ian	AR	K-S	T-test p-value	Simple Logistic Regessioin p-value
Soundness	T35	Financial expense/sales ratio	1.99	2.14	0.87	55.96	44.69	<.0001	<.0001
	T48	Shareholders' Equity ratio	2.09	3.59	0.38	77.81	62.18	<.0001	<.0001
	T50	Quick ratio	2.00	2.17	1.13	55.89	43.45	<.0001	<.0001
	T52	Non-current assets ratio	2.10	2.68	0.93	63.57	53.79	<.0001	<.0001
	T54	Debt Ratio	2.12	3.06	0.82	71.51	60.11	<.0001	<.0001
	T56	Non-current debt ratio	2.05	2.11	1.26	52.59	44.06	<.0001	<.0001
	T60	Trade receivable / trade payable ratio	1.97	1.43	1.18	35.53	29.6	<.0001	<.0001
	T67	Retention ratio	2.17	2.89	0.69	45.14	45.91	<.0001	<.0001
	T69	Reserved amount/Paid-in capital ratio	2.09	3.04	0.75	68.12	52.1	<.0001	<.0001
Profitability	T18	Return on Equity Ratio	2.06	2.60	0.82	57.06	49.86	<.0001	<.0001
	T23	Gross profit margin on sales	2.02	1.65	1.37	37.48	27.38	<.0001	<.0001
	T30	Taxation/Net Income Before Tax Ratio	2.61	1.49	3.91	33.08	23.45	<.0001	<.0001
	T33	Average interest rate on borrowings	1.92	1.53	0.96	41.87	36.24	<.0001	<.0001
	T41	Dividend payout ratio	1.97	0.86	2.42	21.42	21.39	<.0001	<.0001
	T42	Debt Service Coverage Ratio	2.17	2.46	1.04	56.61	42.82	<.0001	<.0001
	T46	EBITDA/Sales	2.28	2.22	1.07	45.21	38.39	<.0001	<.0001
	T47	EBITDA/financial expenses	2.29	2.15	1.40	53.32	43.69	<.0001	<.0001
Activity	T75	Total Capital Turnover	1.94	0.77	1.61	20.71	16.66	<.0001	<.0001
	T82	Inventory Turnover ratio	1.92	0.98	1.38	27.18	22.68	<.0001	<.0001
	T87	Payable Turnover Ratio	1.96	1.44	1.30	35.98	28.42	<.0001	<.0001
Productivity	T90	Value added per employee	1.90	0.63	2.27	20.04	16.38	<.0001	<.0001
	T92	Net income before tax per employee	2.18	2.53	1.21	60.46	43.8	<.0001	<.0001
	T99	Gross value-added to property, plant and equipment	1.89	0.68	2.42	23.66	19.32	<.0001	<.0001
	T100	Gross value-added to machinery and equipment	1.90	0.64	2.39	21.92	18.7	<.0001	<.0001
Growth potenti-al	T1	Growth rate of total assets	2.01	1.37	1.36	31.91	26.72	<.0001	<.0001
	T5	Growth rate of stockholders' equity	2.06	2.14	1.26	50.63	43.46	<.0001	<.0001
	T6	Growth rate of sales	1.96	1.05	1.64	27.73	20.79	<.0001	<.0001
	T10	Increase in number of employees	1.94	0.95	1.77	23.89	16.32	<.0001	<.0001
Cash flow	T70	Total C/F to Debt Ratio	2.02	2.66	0.84	66.84	52.76	<.0001	<.0001
	T71	Total C/F to Borrowing ratio	2.28	2.40	1.31	56.18	42.57	<.0001	<.0001
	T74	Net C/F to Borrowing ratio	1.99	1.22	2.10	33.55	25.76	<.0001	<.0001

표 10. 각 범주별 변수 선택 결과
Table 10. Variable selection result for each category

Category	No of Variab-les	Univaria-te analysis	Correlati-on analysis	discrimination power	
				AR	K-S
Soundness	26	20	9	58.46	48.43
Profitability	17	32	8	43.26	35.40
Activity	13	7	3	27.96	22.59
Productivity	10	6	4	31.52	24.55
Growth potential	37	4	4	33.54	26.82
Cash flow	5	5	3	52.19	40.37
Total	108	74	31		

4.2 다중 로지스틱 회귀모형 결과

총 9개 변수가 단계적 선택법에 의하여 설명변수로 선정이 되었다. 추정결과는 표 11과 같으며 다중 로지스틱 회귀함수 식은 식(8)과 같다.

차입금 평균이자율 및 배당성향은 수익감소에 영향을 주게 됨으로 강한 (+)의 영향을 미치고 있음을 보여준다. 매출액증가율, 자기자본비율, 매출채권/매입채무비율, 순CF대차입금비율, 재고자산 회전율은 10%수준의 (+)의 영향을 미치고 있다. 이와 반대로 사내유보율 및 종업원1인당 순이익은 상대적으로 약한(+)의 영향을 미치고 있음을 보여주고 있다.

기업의 수익감소와 관련된 변수들이 부도확률에 강한 영향을 미치고 있으며, 기업의 수익증대 및 생산성에 관련된 변수들은 상대적으로 약한 영향을 미치고 있음을 판단할 때, 본 연구모형은 타당성을 확보한 것으로 해석된다.

$$\begin{aligned} Y = & -7.1368 + 0.2512\,T6 + 0.4246\,T33 \\ & + 0.7087\,T41 + 0.2405\,T48 \\ & + 0.2706\,T60 + 0.0743\,T67 \\ & + 0.2463\,T74 + 0.2748\,T82 \\ & + 0.1539\,T92 \end{aligned} \quad (8)$$

검증통계량은 회귀계수 p-value는 유의수준 0.05에서 유의한 수준을 보이며, 표 12과 같이 검증 데이터에 대해 연속형 모형 검증지표인 AR은 84.94, K-S는 68.64이며, 이진 분류 검증지표도 80이상의 높은 성능을 보여주고 있음으로 구축된 모형은 상당히 높은 예측력을 보유한 것으로 판단된다.

표 11. 다중 로지스틱 회귀모형 추정 결과
Table 11. Multiple Logistic Regression Estimation Results

Varialbe	Category	Variable Name	Regression coefficient	P-value
Intercept				<.0001
T6	Growth potential	Growth rate of sales	-7.1368	0.0046
T33	Profitability	Average interest rate on borrowings	0.2512	<.0001
T41	Profitability	Dividend payout ratio	0.4246	0.0172
T48	Soundness	Shareholders' Equity ratio	0.7087	<.0001
T60	Soundness	Trade receivable / trade payable ratio	0.2405	<.0001
T67	Soundness	Retention ratio	0.2706	0.0248
T74	Cash flow	Net C/F to Borrowing ratio	0.0743	0.0072
T82	Activity	Inventory Turnover ratio	0.2463	0.0054
T92	Productivity	Net income before tax per employee	0.2748	<.0001

표 12. 로지스틱 회귀모형 성능
Table 12. Logistic Regression Model Performance

Category	Verification Statistics	Development	Validation
Continuous model	AR	86.62	84.94
	K-S	73.29	68.64
Binary classification	Accuracy	85.54	83.80
	Precision	85.46	84.66
	Recall	85.65	82.47
	F1 Score	85.56	83.55

4.3 랜덤 포레스트 모형 결과

랜덤 포레스트모형은 의사결정나무의 개수와 의사결정나무에 활용될 최대변수의 개수를 적절하게 지정해야 한다. 최적의 모형을 도출하기 위하여 의사결정나무의 개수를 탐색하였으며, 이후 최적의 최대변수 개수를 탐색하는 과정으로 진행을 하였다.

의사결정나무의 개수는 그림 4와 같이 1~100개까지 탐색하여, 90개가 최적의 의사결정나무 개수로 선

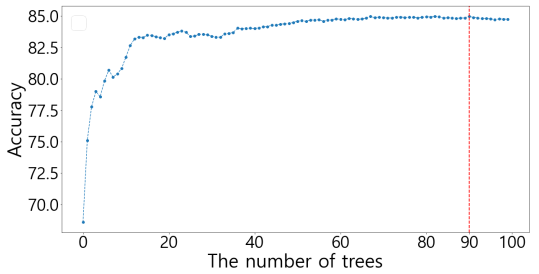


그림 4. 최적의 의사결정나무 개수 탐색 결과
Fig. 4. Search result for the optimal number of decision trees

택이 되었다. 최대 변수의 제곱값이 파이썬 랜덤 포레스트모형의 기본값이며[4], 이에 본 연구는 2~8개까지 탐색을 하였으며, 8개가 최적값으로 선택이 되었다. 최종모형의 성능은 표 13과 같다.

표 13. 랜덤 포레스트 모형 성능
Table 13. Random Forest Model Performance

Category	Verification Statistics	Development	Validation
Continuous model	AR	99.80	85.13
	K-S	96.47	69.68
Binary classification	Accuracy	97.79	84.32
	Precision	96.96	83.42
	Recall	98.68	85.57
	F1 Score	97.81	84.48

4.4 그래디언트 부스팅 모형 결과

그래디언트 부스팅모형은 의사결정나무의 개수와 의사결정나무에 활용될 최대 변수의 개수, 학습률을 적절하게 지정해야 한다. 의사결정나무의 수는 1~10, 최대 변수의 개수는 1~30, 학습률은 0.1~3.0까지 순차적으로 검토하여, 의사결정나무의 수는 2개, 최대 활용변수의 수는 10개, 학습률은 0.2로하는 예측모형을 도출하였으며, 모형의 결과는 표 14와 같다.

표 14. 그래디언트 부스팅 모형 성능
Table 14. Gradient Boosting Model performance

Category	Performance metrics	Development	Validation
Continuous model	AR	88.34	85.31
	K-S	75.28	69.15
Binary classification	Accuracy	87.20	83.80
	Precision	85.62	81.95
	Recall	89.40	86.60
	F1 Score	87.47	84.21

4.5 오차 앙상블모형에 대한 연구

로지스틱 회귀는 안정적인 예측력과 높은 해석력을 가지는 반면에 예측력에 있어서는 다른 머신러닝 모형대비 낮은 단점을 개선하고자, 로지스틱 회귀과 그래디언트 부스팅을 결합하는 프로세스를 통한 오차 앙상블모형을 제시하고자 한다.

연구모형은 앞서 적합된 다중 로지스틱 회귀모형의 예측된 부도확률값과 실측부도값의 오차를 그래디언트 부스팅 모형으로 예측하여, 두 모형을 결합하는 프로세스로 산출이 되며, 그림 5와 같다.

다중 로지스틱 회귀모형의 예측된 부도확률값과 실측부도값의 오차를 예측하는 그래디언트 부스팅 모형은 의사결정나무의 수는 1~10, 최대 변수의 개수는 1~30, 학습률은 0.1~3.0까지 순차적으로 검토하여 의사결정나무의 수는 3개, 최대 활용변수의 수는 2개, 학습률은 1로 하는 모형을 도출하였으며, 최종 오차 앙상블모형의 성능은 표 15와 같다.

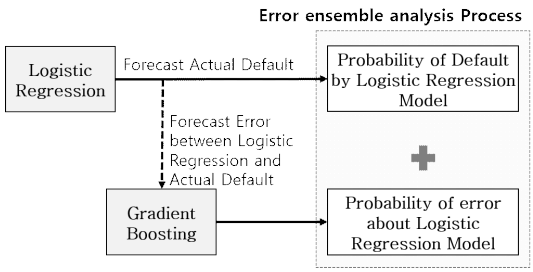


그림 5. 오차 앙상블모형 분석 프로세스
Fig. 5. Error Ensemble model analysis process

표 15. 오차 앙상블모형 성능
Table 15. Error Ensemble model performance

Category	Performance metrics	Development	Validation
Continuous model	AR	88.72	84.97
	K-S	77.26	71.21
Binary classification	Accuracy	87.53	85.09
	Precision	87.95	85.05
	Recall	86.98	85.05
	F1 Score	87.46	85.05

4.6 각 모형의 성능 비교

각 모형에 대한 성능은 표 16과 같으며, 개발과 검증에 대해서 각 모형별 성능을 측정하기 위하여 활용한 6개 지표에 대해서 우수한 성능을 가지는 지표의 개수로 각 모형을 평가한 결과는 표 17과 같다.

표 16. 각 모형의 성능 비교
Table 16. Comparison of each Model's performance

Development					
Category	Performance metrics	Logistic Regression	Random Forest	Gradient Boosting	Error Ensemble Model
Continuous	AR	86.62	99.80	88.34	88.72
	K-S	73.29	96.47	75.28	77.26
Binary classification	Accuracy	85.54	97.79	87.20	87.53
	Precision	85.46	96.96	85.62	87.95
	Recall	85.65	98.68	89.40	86.98
	F1 Score	85.56	97.81	87.47	87.46
Validation					
Category	Performance metrics	Logistic Regression	Random Forest	Gradient Boosting	Error Ensemble Model
Continuous	AR	84.94	85.13	85.31	84.97
	K-S	68.64	69.68	69.15	71.21
Binary classification	Accuracy	83.80	84.32	83.80	85.09
	Precision	84.66	83.42	81.95	85.05
	Recall	82.47	85.57	86.60	85.05
	F1 Score	83.55	84.48	84.21	85.05

표 17. 각 모형의 성능 평가
Table 17. Evaluation ranking of each model's performance

Development		
Rank	Model	Reason
1	Random Forest	6 metrics are high
2	Ensemble Model	4 metrics are slightly high (AR, K-S and etc.)
3	Gradient Boosting	2 metrics are slightly high (Recall, F1 Score)
4	Logistic Regression	
Validation		
Rank	Model	Reason
1	Ensemble Model	4 metrics are high (K-S, F1 Score and etc.)
2	Gradient Boosting	2 metrics are high (AR, Recall)
3	Random Forest	5 metrics are slightly high (AR, K-S and etc.)
4	Logistic Regression	

개발 대상에서는 랜덤 포레스트가 모든 지표에서 가장 우수한 성능을 보이고 있다. 검증 대상으로는 오차 앙상블모형이 4개 지표(K-S, Accuray, Precision, F1 Score)에서 우수한 성능을 보이고 있으며, 그래디언트 부스팅 모형이 2개 지표(AR, Recall)에서 우수한 성능을 보이고 있다.

의사결정나무에 참여한 변수들에 대해 평균적인 불순도의 감소로 변수의 중요도를 측정할 수 있다[15]. 랜덤 포레스트 및 그래디언트 부스팅 단일모형과 오차 앙상블모형에 활용된 그래디언트 부스팅모형에 대해서는 불순도의 감소에 따른 각 변수의 중요도를 측정하였으며, 로지스틱 회귀모형에 대해서는 모형에 활용된 전체 변수의 회귀계수의 합에 대해서 개별 변수의 회귀계수가 차지하는 비중으로 중요도를 계산하였다. 그 결과는 표 18과 같다.

표 18. 각 모형별 변수의 중요도 비교
Table 18. Comparison of importance of variables for each model

Variable	Random Forest	Gradient Boosting	Logistic Regression	Error forecast Gradient Boosting
T41	0.2		26.8	
T33	3.7	3.1	16.1	24.7
T82	1.8		10.4	
T60	2.4	1.7	10.2	8.7
T6	1.8		9.5	
T74	1.6	1.0	9.3	
T48	14.0	67.0	9.1	13.7
T92	2.8	0.3	5.8	
T67	3.3	2.9	2.8	
T99	1.0			25.2
T52	7.4			10.5
T10	1.2			10.0
T50	2.6			7.2
T47	3.1	5.6		
T70	8.9	4.1		
T35	3.1	4.0		
T18	4.2	3.9		
T69	6.5	3.6		
T42	2.6	1.7		
T87	2.1	1.2		
T54	9.6			
T5	3.3			
T56	2.7			
T71	1.7			
T46	1.5			
T1	1.4			
T75	1.4			
T23	1.2			
T30	1.1			
T100	1.0			
T90	0.9			

랜덤 포레스트 단일모형에서는 31개 변수가 모두 활용이 되었으며, 이중 자기자본비율(T48)의 중요도가 14%로 가장 높았다. 그래디언트 부스팅 단일모형에서는 13개 변수가 활용되었으며, 자기자본비율(T48)의 중요도가 67%로 높았다. 로지스틱 회귀모형에서는 9개 변수가 활용되었으며, 배당성향(T41)의 중요도가 26.8%로 높았다. 오차 앙상블모형에 활용된 그래디언트 부스팅모형에서는 7개 변수가 활용되었으며, 설비투자효율(T99)의 중요도가 25.2%로 높았다.

각 모형별로 알고리즘이 다르며, 그에 따라 모형에 활용된 변수 및 각 변수의 중요도가 차이가 있음을 알 수 있다.

오차 앙상블모형은 로지스틱 회귀모형과 비교하여 성능이 개선되었으며, 로지스틱 회귀모형의 오차가 그래디언트 부스팅모형을 통해 감소가 되었음을 알 수 있다.

오차 예측에 활용된 그래디언트 부스팅모형은 변수는 7개로, 이 중 로지스틱 회귀모형과 중복되는 변수영역의 중요도는 47%이며, 중복되지 않는 변수영역의 중요도는 53%이다. 이를 통해 로지스틱 회귀모형이 가지는 오차가 모형 변수에서 발생된 영역과 모형 변수로 설명되지 않는 영역에 걸쳐 고르게 보완이 되었음을 알 수 있다.

4.7 오차 앙상블모형의 한계

로지스틱 회귀모형과 오차 앙상블모형을 개발 대상의 분포를 비교하면 그림 6과 같으며, 로지스틱 회귀모형을 중심으로 앙상블모형이 일정한 범위내에서 변동되었음을 볼 수가 있다.

오차 앙상블모형은 부도 예측에 있어서 로지스틱 회귀모형보다 더 높은 예측력을 가지고 있으며, 로지스틱 회귀모형을 기반으로 함에 따라 랜덤 포레스트 모형 및 그래디언트 부트팅모형과 비교하여 회귀식으로 해석이 가능한 영역이 존재하는 장점이 있다.

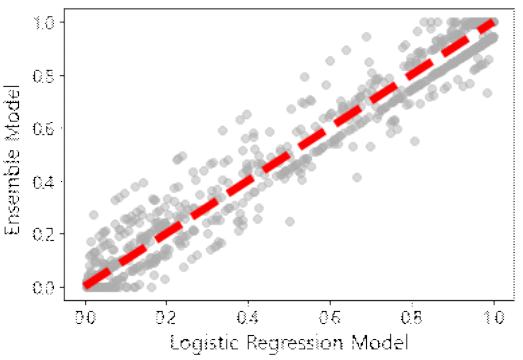


그림 6. 오차 앙상블모형과 로지스틱 회귀모형의 분포
Fig. 6. Distribution of the error ensemble model and logistic regression model

다만, 모형의 성능만을 비교해볼 때 개발 대상에서는 랜덤 포레스트모형이 더 성능이 우수하였으며, 검증 대상에서는 오차 앙상블모형이 K-S등 일부 지표에서만 우수한 성능을 보인 점을 고려할 때, 부도 예측에 있어서 랜덤 포레스트모형 및 그래디언트 부스팅 모형대비 오차 앙상블모형의 성능 개선은 한계가 존재함을 알 수 있었다.

V. 결 론

본 연구에서는 1,295개 제조업 기업에 대해서 다중 로지스틱 회귀모형, 랜덤 포레스트, 그래디언트 부스팅모형 그리고 로지스틱 회귀와 그래디언트 부스팅의 오차 앙상블모형에 대해서 비교를 하였다.

활용변수는 108개 재무비율을 사용을 하였으며, Mini-Modeling을 통해 변환된 재무비율에 대해 단변량, 상관분석을 하여 부도에 유의한 31개 변수를 선별하였다.

로지스틱 회귀모형에서는 총 9개 변수가 모형변수로 선택이 되었으며, 랜덤 포레스트모형, 그래디언트 부스팅모형, 오차 앙상블모형에 대해서 최적의 파라미터로 적합하였다.

각 모형의 변별력을 비교한 결과 개발 대상에서는 랜덤 포레스트가 우수한 성능을 보이고 있으며, 검증 대상에서는 오차 앙상블모형이 K-S등 일부 지표에서 우수한 성능을 가지고 있었다.

오차 앙상블모형은 로지스틱 회귀모형대비 모형의 성능이 개선되었으며, 로지스틱 회귀모형으로 해석이 가능한 영역이 존재하는 부분을 검토할 때, 오차 앙상블모형이 기업 신용평가모형으로써 장점이 있다고 판단이 된다.

다만, 랜덤 포레스트모형 및 그래디언트 부스팅모형대비 더 높은 예측력을 가지고 있지 못하는 점은 금번 연구의 한계라고 할 수가 있다.

앞서 각 모형별로 부도 고객을 예측함에 있어서 활용변수 및 활용변수의 중요도에 있어서 많은 차이가 있음을 확인하였다. 다양한 머신러닝 모형과 각 모형에서 부도 예측에 있어서 중요하게 고려하는 변수를 반영하여 변별력을 개선하고, 해석 가능한 영역을 높이는 연구가 진행된다면 보다 완성된 연구가 될 것으로 기대된다.

References

[1] S. M. Choi, “Current status and change trend of corporate credit rating model,” *CIS Issue Report*, vol. 2018, no. 6, pp. 1-12, 2018.

[2] Donald van Deventer and Kenji lmai, *Credit Risk Models & the Basel Acccords*, John Wiley & Sons Inc press, 2003.

[3] Financial Supervisory Service, *Best Practices for Integrated Risk Management in Basel Ⅱ* (2008), Retrieved Oct. 1, 2021, from https://www.fss.or.kr.

[4] S. J. Kim and H. C. Ahn, “Application of random forests to corporate credit rating prediction,” *J. Ind. Innovation Res.*, vol. 32, no. 1, pp. 187-211, May 2016.

[5] G. Aurelien, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, Hanbit Media Inc Press, 2019.

[6] H. S. Kim, *Step by Step Business Machine Learning in Python*, PREDICS Press, 2020.

[7] K. J. Kim, *Finance Statement Analysis & Valuation*, ChangMinSa Press, 2020.

[8] Statistics Korea, *Korean Standard Industrial Classification*, Economic Book Press, 2017.

[9] C. S. Song, *Properly Read financial statements*, Sehaksa Press, 2020.

[10] J. W. Choi, “Forecasting corporate default using artificial intelligence based on news information,” Ph.D. dissertation, Dept. of Business Administration Graduate School of Konkuk University, 2019.

[11] W. R. Oh and E. G. Kim, *Writing a Thesis using SAS*, Tamjin Press, 2020.

[12] H. J. Kwon, “Comparative study of default prediction model using consolidated and separate financial statements,” *J. Korean Accounting Inf. Assoc.*, vol. 2017, no. 1, pp. 109-170, May 2017.

[13] D. H. Yoon, *Reconstruction of Busan's credit rating model...Internal grading method*(2013), Retrieved Oct. 1, 2021, from https://www.thebell.co.kr.

[14] J. Y. Kim, “Development of a personal credit scoring model(TELCO Score) telecommunications,” Ph.D. dissertation, Dept. of IT Policy & Management of Soongsil Univerisity, 2019.

[15] Sebastian Raschka and Vahid Mirjalili, *Python Machine Learning*, Gilbut press, 2021.

[16] E. Altman, “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy,” *J. Finance*, vol. 23, no. 4, pp. 589-609, 1968.

[17] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, pp. 5-32, 2001.

[18] J. Y. Campbell, J. Hilscher, and J. Szilagyi, “In search of distress risk,” *J. Finance*, vol. 63, no. 6, pp. 2899-2939, 2008.

[19] E. G. Falkenstein, A. Boral, and L. Carty, “RiskCalc for private companies: Moody's default model,” *Rating Methodology*, pp. 1-88, 2000.

[20] Financial committee, *Banking Supervision Regulations*(2021), Retrieved Oct. 1, 2021, from https://www.law.go.kr.

[21] Financial Supervisory Service, *Basic Detailed Guidelines for Credit Risk Internal Rating Act*(2005), Retrieved Oct. 1, 2021, from https://www.fss.or.kr.

[22] N. K. Kim, *WooriBank, Introduction of Enterprise diagnosis system using Big data*(2018), Retrieved Oct. 1, 2021, from http://it.chosun.com/site/data/html_dir/2018/03/27/2018032785000.html.

[23] I. R. Lee and D. C. Kim, “An evaluation of bankruptcy prediction models using accounting and market information in Korea,” *Asian Rev. Financial Res.*, vol. 28, no. 4, pp. 625-665, Nov. 2015.

[24] J. W. Park and S. M. Ahn, “Corporate bankruptcy prediction using financial ratios: focused on the korean manufacturing companies audited by external,” *J. Korean Manag. Rev.*, vol. 43, no. 3, pp. 639-669, 2014.

김 용 환 (Yong-hwan Kim)

2002년 8월 : 고려대학교 통계학과 학사

2017년 8월 : 숭실대학교 IT정책경영학과 석사

2020년 3월~현재 : 숭실대학교 IT정책경영학과 박사과정, 농협은행 카드신용관리부 근무 중

<관심분야> 신용평가모형, 머신러닝, 빅데이터

[ORCID:0000-0002-9127-3420]

김 도 형 (Do-hyung Kim)

1995년 2월 : 숭실대학교 전자계산학과 졸업

1997년 2월 : 숭실대학교 컴퓨터학과 석사

1996년 12월~현재 : LG-CNS, 인터파크, 오픈타이드코리아, CJ올리브네트웍스, 숭실대학교 IT정책경영학과 박사과정

<관심분야> 인공지능, 빅데이터 분석 및 전략, 디지털 비즈니스

[ORCID:0000-0001-9083-6132]

허 재 혁 (Jae-hyuk Heo)

2020년 3월~현재 : 숭실대학교 IT정책경영학과 박사과정 타임소프트 상무 재직 중

[ORCID:0000-0002-5922-0281]

김 광 용 (Gwang-yong Gim)

1984년 : 고려대학교 공학사 졸업

1991년 : 조지아 주립대학 보험수리학 석사

1995년 : 미국 조지아 주립대학 박사

1999년~현재 : 숭실대학교 경영학부 교수

<관심분야> 데이터사이언스, 디지털트랜스포메이션, 인공지능, 빅데이터, 블록체인, 클라우드, IOT, 전자정부, 핀테크, 비즈니스 모델링(디자인싱킹,TRIZ, 캔버스모델 등) 등