Asian Review of Financial Research 34(4): November 2021
재무연구 제34권 제4호 (2021년 11월), pp.199~234
https://doi.org/10.37197/ARFR.2021.34.4.6
ISSN 1229-0351

기계학습 기반 기업신용정보 분석을 통한 채무불이행 예측

송민찬 NICE디앤비 과장
류두진* 성균관대학교 경제학과 교수

요약

본 연구는 신용정보 표본DB 원격분석시스템에서 제공하는 기업신용정보를 분석하여 기업의 채무불이행을 예측한다. 사업자 구분에 따라 분석대상을 나누고 표본DB에서 제공하는 기업신용정보의 활용에 따른 실험을 구성한다. 또한, 다양한 기계학습 기법을 모수 추정 방식에 따라 모수적 방법론, 비모수적 방법론, 준모수적 방법론으로 구분하여 예측성과를 비교한다. 표본DB 원시데이터를 활용한 분석보다 대출 및 연체 종류에 따라 가공한 자료를 활용하는 경우 각 기계학습 모형별 성능개선이 관측되었으나, 기업 차주의 특성정보와 기술신용평가 정보의 활용은 모형별 성능개선에 기여하지 못하였다. 모든 세그먼트에서 준모수적 방법론에 해당하는 심층신경망 모형에 대해 성능이 가장 우수한 것으로 확인되었으며, 트리계열이 아닌 비모수적 방법론의 경우 재현율이 낮게 관측되어 채무불이행 예측 문제에 적합하지 않았다. 기존 실무에서 사용되는 모수적 방법론을 활용한 경우보다 준모수적 방법론을 활용할 경우 분류 성능이 향상됨을 확인하였다. 본 연구는 실제 기업신용정보에 대해 구성된 표본DB를 활용하여 기업부실 예측을 시도한 최초의 연구이며, 기업신용정보를 활용하는 여신금융기관과 신용정보사의 자료 활용 및 모형 구축에 대한 방향성을 제시한다.

주요단어 기계학습, 부실위험, 여신거래정보, 채무불이행 예측, 표본DB원격분석시스템

투고일 2021년 05월 07일
수정일 2021년 07월 19일
게재확정일 2021년 08월 27일

* 교신저자: sharpjin@skku.edu, 전화: +82-2-760-0429
본 논문은 송민찬의 데이터사이언스융합학과 학위논문을 확장한 연구로, 한국신용정보원 표본DB 원격분석시스템(Credb)의 분석용 데이터 및 분석환경을 활용함. 한국경제학회 주관 경제학 공동학술대회와 재무금융 공동학술대회에서 발표되었으며, 연구의 발전을 위해 유익한 조언을 해주신 송교직 편집위원장님, 김장현 교수님(성균관대), 김형준 교수님(영남대), 송수영 교수님(중앙대), 이지형 교수님(성균관대), 최형석 교수님(이화여대), 허정규 교수님(전남대), Marty Janowiecki(PWC)께 감사드립니다.

Asian Review of Financial Research 34(4): November 2021
재무연구 제34권 제4호 (2021년 11월), pp.199~234
https://doi.org/10.37197/ARFR.2021.34.4.6
ISSN 1229-0351

Predicting Loan Delinquency by Analyzing Sample DB with Machine Learning

Minchan Song Manager, NICE Dun&Bradstreet Co., Ltd.
Doojin Ryu* Professor, Department of Economics, Sungkyunkwan University

Received 07 May. 2021
Revised 19 Jul. 2021
Accepted 27 Aug. 2021

Abstract

This paper investigates the ability to predict corporate default rates using loan-sample data from the Korea Credit Information Service's financial big data open system (CreDB). The corporate loan from financial institution increases financial institution's credit exposure. Because measurement of the impact on the credit risk in the financial institution is used in determining the pricing model and structure of loan products, it is an essential factor for the financial institution that affects its profit structure. In terms of risk management, predicting delinquency using loan data is necessary for 5,000 Korean financial institutions. In several studies, bankruptcy forecasting was conducted on listed companies that disclosed financial and stock price information. However, this study increases the practical utility by extending the analysis target to individual entrepreneurs and small and medium-sized enterprises(SMEs). In addition, this study presents representative big data analysis results by utilizing loan, delinquency, and technology credit information of approximately 1.1 million corporations, which is 20% of almost 5.6 million domestic sole proprietors and non-listed corporations. For loan data, it includes ten monthly loan type codes and eleven overdue reason codes. Prediction targets are separated by individual and corporate entrepreneurs. Also, analyses are divided by use of the processed dataset. For efficient analysis, the data dimension was reduced by changing the table structure through nested iterative operations while expanding the variable composition from a table consisting of N rows to one column. To reflect the

* Corresponding Author: sharpjin@skku.edu, Tel: +82-2-760-0429

characteristics of the data as much as possible, exploratory data analysis and feature-engineering were performed to process the data. Also, classification models are classified by four groups using a parametric method that nine models train for classification. Group 1 consists of Logistic Regression and Linear discriminant analysis based on the parametric method, group 2 consists of several algorithms that calculate the distance for model learning. In addition, group 3 consists of tree-based algorithms, which are also non-parametric methods. Group 4 consists of the semi-parametric method, which is deep neural network. However, out of the total 438,697 corporations, 810 defaulted, accounting for only 0.2% of the forecast, so the target distribution is severely imbalanced. For this reason, before model fitting, under sampling of imbalanced data was performed. The bias of the sampled training and validation data is minimized by performing. K-fold cross validation as much as the level of K=5. Finally, the analysis result suggests a significant effect on classification performance when the processed data is used. However, this study suggests no significant effect on performance when loan owner's characteristics are included. Moreover, tech-credit rating (TCB) information gives any meaningful effect regarding the type of corporation. Also, classification with Deep Neural Network (DNN), which is based on the Semi-parametric method, makes the best performance of binary classification. Non-parametric and Non-tree based models are not appropriate methods for analyzing loan data. In the case of the DNN based on the semi-parametric methodology, the highest classification performance was confirmed for all analyses and entrepreneurs' classifications performed in this study. The neural network used in this study consists of 14 hidden layers. According to the neural network baseline design, the sigmoid function was applied to the activation function's initial value, the relu function was applied to the hidden layer, and optimization was performed through the Adam optimizer. In particular, the analysis of credit transaction information based on credit information of all financial institutions in Korea was conducted, and there is a possibility for alleviating information asymmetry of individual credit institutions regarding risk management targets. In addition, in the case of parametric methodologies used in classical studies and most used in practice, the average classification performance for major segments was inferior to that of semi-parametric methodologies. Furthermore, the difference between these performances is up to 16 percent. This paper suggests the direction of using loan-sample data. It is foundational research for financial institutions that are using loan data for credit risk management. It is necessary to expand research focusing on semi-parametric methodologies about corporate credit information analysis.

Keywords Corporate Loan Data, Distress Risk, Machine Learning, Predicting Loan Delinquency, Sample DB Remote Analysis System

Ⅰ. 서 론

본 융합연구에서는 재무금융 분야의 기존 연구들과는 달리, 재무제표 및 재무지표가 아닌 기업의 여신거래정보 및 기술신용정보를 활용하여 기업의 채무불이행 발생에 대한 기계학습 기반의 예측모형을 수립하고 분석자료와 모형 활용의 유의성을 검증하고자 한다.

자금조달(financing)은 기업 활동에서 가장 중요한 영역의 하나이다. 실제로, 외부회계감사법인(이하, 외감) 이상 규모의 국내 제조업의 경우 2분위수 기준 37% 수준의 차입금의존도를 보이는 것이 일반적이며, 개인사업자를 포함한 중소기업은 2018년과 2019년 2개년 동안 38% 수준의 차입금의존도를 보여 22% 수준의 차입금의존도를 보이는 대기업 및 중견기업 대비 높은 레버리지와 금융비용 부담을 통해 사업을 영위하는 실정이다(한국은행, 2020). 다만, 해당 차입금은 여신금융기관의 재무제표상 부채로 계상되었다가 원금의 상환이 완료됨과 동시에 이자수익과 함께 자본으로 전환되는 계정인 이유로, 연체 및 부도에 따른 차입금 회수에 대한 채무불이행 위험관리를 포함한 부실채권 관리는 여신금융기관의 리스크(risk)와 직결된다(이민준, 이정환, 2019). 즉, 은행, 캐피탈 등 여신금융기관에서 발생하는 대출은 동 여신기관의 신용위험을 증가시키는 원인이 되는 것으로 볼 수 있다. 또한, 여신금융기관의 신용위험관리는 가치평가 및 상품설계에도 활용이 되는 이유로 수익구조 설계에 영향을 미치는 중요한 요소이기도 하다. 이 때문에 대부분 여신기관은 원금 및 이자 상환에 대한 차주의 채무불이행 발생에 대비하여 신용평가 모형을 직접 구축하여 운용하거나, 신용평가기관의 컨설팅을 통해 통계모형을 구축하여 신용위험 관리체계를 운영하고 있다.

이에 따라 기업부실과 기업부도 예측을 위한 다양한 연구가 진행되었으나, 주식거래자료나 기업의 재무자료를 활용한 연구가 대부분이다. 반면, 여신금융기관으로부터 발생하는 대출 및 연체를 반영하는 실제 여신거래정보나 차주의 특성을 대변하는 비재무 정보를 활용한 연구는 미미한 실정이다.[1] 또한, 외감법에 따라 외부감사 대상이 아닌 중소기업과 소상공인에 대해서는 분석 가능한 수준의 재무자료 확보가 제한되어 분석대상 자체도 한정되었다. 한편, 2019년 5월 금융분야에서도 금융위원회를 중심으로 금융정보 개방화가 추진되며 한국신용정보원의 표본DB 원격분석시스템이 구축되었다.[2] 표본DB 원격분석시스템은 국내 약 5,000개에 해당하는 국내

1) 비재무정보는 K-IFRS기준 표준재무제표에서 제시하는 재무계정 147개 항목에 해당하지 않는 모든 정보를 의미한다. 기업의 규모 및 업력 뿐 아니라 대표자의 신용평점과 대출잔액 등은 모두 비재무정보에 해당한다.

2) 한국신용정보원은 신용정보법 제25조의2(종합신용정보집중기관의 업무)에 의거 일반신용정보 집중 · 관리를 담당하는 기관으로, 신용정보법 시행령 제21조 제3항에 따라 [별표2]에서 규정하는 일반신용정보를 금융기관, 공공기관 및 신용조회사 이용자에게 제공하고 있다(https://www.kcredit.or.kr/).

전체 여신금융기관으로부터 발생한 기업의 대출 및 연체에 대한 발생 및 해제정보뿐 아니라 기술신용평가(TCB) 정보가 집중하는 한국신용정보원의 원본 DB를 바탕으로 비식별화(de-identification) 및 층화추출(stratified sampling)된 표본을 학술연구 목적 이용자에게 제공하는 시스템이다.[3)]

표본DB 원격분석시스템을 활용할 경우, 여신금융기관의 주요 관심 대상인 여신거래 데이터를 활용한 분석이 가능할 뿐 아니라 제도적 뒷받침을 받으며 추진 중인 기술신용평가 정보 활용의 유의성에 대해서도 검증 가능하다는 장점이 있다. 또한, 재무자료 확보가 제한되는 중소기업과 개인사업자에까지 연구 대상을 확대할 수 있다. 무엇보다도, 대출과 연체로 구성된 기업신용정보는 차입금의존도와 금융비용부담률 등 기업 재무상태 진단 시 안정성 지표에 직접 영향을 미치는 자료로써 분석의 가치가 있다. 이에, 본 논문에서는 표본DB 원격분석시스템에서 제공하는 대출 및 연체정보와 함께 해당 정보의 대상이 되는 차주특성정보 및 기술신용정보를 활용하여 기업의 채무불이행을 예측하고자 한다.[4)]

최근 재무금융 분야에서는 예측모형의 성능을 향상시키거나 변수선택의 유의성을 확보하기 위해 인공지능 기법을 적용한 연구가 진행되고 있다. 기계학습의 지도학습 기법을 활용하여, 기존 계량 및 통계분석의 한계로 실증분석이 제한되었던 연구 분야에 대해서도 유의한 결과가 계속하여 제시되고 있다. 다만, 기계학습 지도학습의 경우 분석 주제에 따라 데이터의 특성 및 모형에 따른 학습결과의 차이가 존재하는 가운데, 다수의 기존 연구에서는 데이터의 적절한 활용 방향성을 찾기보다는 단순히 학습모형의 성능평가 결과만을 제시한다. 채무불이행 예측에 관한 연구의 경우에도, 기계학습 및 딥러닝(deep learning) 기반의 다양한 알고리즘을 활용한 연구가 진행되었으나 자료의 특성에 대한 분석을 반영한 데이터 활용에 관한 연구는 미미한 실정이다. 이에, 본 연구에서는 채무불이행 예측을 위한 표본DB의 활용에 따라 기계학습 알고리즘 유형별 성능 차이를 분석하여 기계학습 활용에 대한 방향성을 제시하고자 한다.

본 연구에서 기계학습 기법을 기반으로 기업신용정보 표본DB의 분석 방향성을 살펴보기 위해 수행한 내용은 세 가지로 설명할 수 있다. 첫째, 분석정보의 활용에 따라 실험을 구성하였다. 표본DB 원격분석시스템에서 제공하는 원시데이터에 대한 활용과 대출코드 및 연체사유를 기준으로 구성한

3) 층화추출은 표본에 구성된 틀에 대하여 관심변수와의 관계를 기준으로 보조변수에 의한 층을 구성하는 방식으로 계층별 특성을 고려할 때 일반적으로 사용하는 표본설계 방법론이다(이상은, 2019).

4) 금융위원회는 '14년 06월 부로 기술신용평가제도를 시행하였으며, 기술보증기금과 3개 CB(credit bureau) 회사가 기술신용평가 수행기관으로 선정되었다(금융위원회 보도자료, 2014.06.20.).

가공데이터의 활용에 대한 유의성을 확인하고자 하였으며, 차주의 특성정보와 그 확장의 성격을 지닌 기술신용평가 정보의 반영 등 데이터 활용에 따라 네 가지 실험을 수행하였다. 둘째, 데이터 특성을 반영한 탐색적 데이터 분석(EDA; Exploratory Data Analysis)와 데이터 처리(data processing)를 수행하였다. 채무불이행 예측의 경우 소수 비정상 클래스에 대한 예측을 수행하는 문제에 해당한다. 따라서 소수 클래스가 보이는 특성을 최대한 반영한 학습 수행을 위해 재표본추출(resampling) 외에도 원-핫 인코딩(one-hot encoding), 특성 공학(feature engineering), 스케일링(scaling) 등 다양한 기법을 적용하여 모형에 적합 시 데이터의 특성을 최대한 반영하고자 하였다. 셋째, 기계학습 베이스라인으로 활용되는 9가지 지도학습 알고리즘을 모수 추정 방식에 따라 구분하였으며, 비모수적 방식의 방법론 중 트리(tree)를 활용한 알고리즘을 별도로 구분하여 총 네 가지 그룹의 알고리즘에 대한 실험 성능을 추출하였다. 모수적(parametric) 기법과 비모수적(non-parametric) 기법, 트리계열(tree-based)의 알고리즘, 준모수적(semi-parametric) 기법 등 각 알고리즘이 소속된 그룹 간 비교를 수행할 경우, 데이터 활용에 더욱 적합한 알고리즘의 유형을 파악할 수 있다는 장점이 있다.

본 연구는 여신거래정보 및 기술신용정보를 제공하는 기업신용정보 표본DB 활용에 관한 연구로, 재무자료 입수가 가능한 제한된 대상에 대해서만 분석한 기존 연구와는 차별성이 있다. 또한, 모형의 변형과 상관관계 도출 등 설명력을 제시하거나 성능평가 결과를 제시하는 대부분의 연구와 달리 자료 활용과 처리에 초점을 맞추어 기업 채무불이행 예측의 방향성을 제시한다는 기여가 있다. 또한, 한국신용정보원에서 제공하는 기업신용정보 및 기술신용정보를 활용하여 실제 위험관리 시스템을 구축·운영하고 있는 은행, 카드, 캐피탈 등 여신금융기관에 기계학습 기반 신용정보 활용 방향성을 제시하는 의의가 있다.

본 연구의 구성은 다음과 같다. 제 Ⅰ장에서 연구의 배경과 의의를 살펴본 후 제 Ⅱ장에서는 기업부실에 관한 기존 연구와 인공지능 기법을 적용한 최근 연구에 대해 살펴보았다. 제 Ⅲ장에서는 본 연구에 사용된 자료에 대한 특징과 한계에 대하여 설명하였다. 제 Ⅳ장에서는 연구의 자료에 대한 탐색적 데이터 분석과 함께 실험에 대한 구성 및 데이터 가공 프로세스를 설명함과 동시에 기계학습을 활용한 분석방법론과 성능평가 기준을 정의한다. 제 Ⅴ장에서 각 모형을 기준으로 학습이 진행된 결과를 분석한다. 제 Ⅵ장은 본 연구의 결론을 제시한다.

Ⅱ. 선행연구

기업의 대출과 연체 등 여신거래정보를 활용하여 채무불이행 예측을 수행한 연구는 드문 편이나, 기업 채무불이행과 높은 관련성을 보이는 기업의 부실 및 부도예측에 대해서는 위험관리 분야의 전통적인 연구주제로 이미 많은 연구가 진행되었다. 특히, 고전적인 연구의 경우 판별분석과 이진(binary) 모형, 그리고 위험 모형 등 3가지 방법론을 중심으로 전개되었는데 모두 통계적 방법론에 기반하였다는 공통점이 있다. Beaver(1966)와 Altman(1968)은 판별분석을 바탕으로 신용점수를 구성하여 부도 위험을 단계적으로 제시하는 축약형 모형을 구성하였다. 한편, Ohlson (1980)과 Zmijewski(1984)는 이진 모형을 활용하여 분석대상 기간 다음 시점의 부도율을 계산하는 방식을 제시하였으며, 로짓(Logistic) 회귀분석과 프로빗(Probit) 회귀분석을 활용한 분석결과를 제시했다. 나아가, Shumway(2001)은 Ohlson(1980)과 Zmijewski(1984)가 각각 제시한 이진 모형을 확장하여 분석대상 기간 이후 여러 시점에 대한 부도율을 예측할 수 있는 다기간 모형을 통해 시계열 변수 특성을 반영한 위험 모형을 제시했다. 고전 연구에서 제시한 방법론을 확장하는 방향의 다양한 연구가 수행되었으며(Altman, 1993; Campbell, Hilscher, and Szilagyi, 2008; Chava and Jarrow, 2004; Shumway, 2001), 미국 기업에 대한 분석결과를 토대로 국내기업에 대한 부도예측 방법론 적용 방향을 제시하는 연구도 진행되었다(이인로, 김동철, 2015). 다만, 대부분의 연구에서 재무 및 주가정보가 연구자료로 활용되어 자료 확보가 가능한 대기업 및 상장기업만을 대상으로 연구가 진행되었다는 한계가 있다.

한편, Samuel(1959)이 기계학습의 개념을 제시한 이래로 Hinton(2006)이 딥러닝의 개념을 제시하면서 금융 및 재무분야에서도 기계학습 및 인공신경망(artificial neural network) 기법을 활용한 다양한 연구가 수행되었다. 특히, 기계학습 기법의 경우 복잡한 구조에 대한 분석과 비선형성을 지니는 자료에 대한 분석이 용이하다는 측면에서, 고전 연구에서 활용된 통계적 기법과 계량경제학 기반으로 모형을 확장하는 방법 대비 부실 예측에 있어 높은 성능을 보이는 것으로 확인되었다(김형준, 류두진, 조훈, 2019). Odom and Sharda(1990)는 신경망과 판별분석을 비교하였으며, Tam and Kiang(1992)은 ID3(Iterative Dichotomiser 3)와 로짓분석을 더하여 정확성과 적응력 측면에서의 기계학습 모형을 평가한다. 이후, Zhang, Hu, Patuwo, and Indro(1999)는 인공신경망을 활용해 로짓분석 대비 우수한 예측 성능을 제시하였다.

국내에서도 기업부실 예측을 위해 기계학습 기법을 적용한 연구가 활발히 진행되었다. 기존

연구가 주가 및 재무정보를 활용한 기업부실과의 상관관계 추정에 국한되었다면, 기계학습 기법을 적용한 최근 연구들은 새로운 방법론을 활용하여 성능개선 결과를 제시하는 방향과 다양한 자료를 분석에 활용하는 방향으로 진행되었다.

실제로, 권혁건, 이동규, 신민수(2017)는 인공신경망을 기반으로 순환신경망(RNN: Recurrent Neural Network)을 활용한 기업 부도예측모형을 구성하여 모형 성능이 개선된 결과를 보였다. 동 연구에 사용된 순환신경망은 자료의 순서를 반영한 학습을 지원하는 알고리즘으로, 기존 모형과 다르게 시계열 재무데이터의 동적 특성을 반영한 모형을 구성한 결과 다변량판별분석, 로짓분석, 인공신경망 분석 대비 개선된 예측력을 확인할 수 있었다. 또한, 차성재, 강정석(2018)은 순환신경망과 LSTM(Long Short-Term Memory)을 활용해 딥러닝 기반 시계열 학습을 통한 기업부도예측이 고전 연구에서 활용된 통계적 분석 방법론과 그 확장 모형 대비 성능이 우수함을 실증하였다. 다만, 기계학습 기반의 학습에서는 단일 모형 기반의 예측이 진행됨에 따른 편향(bias)이 존재하였는데, 엄하늘, 김재성, 최상옥(2020)은 스태킹 앙상블 기법을 적용한 기계학습 기반 예측모형의 성능개선 가능성을 제시하였다. 한편, 기계학습 기법의 발전으로 비정형 정보에 대한 분석이 가능해지며 새로운 자료를 활용한 연구가 진행되고 있다. 송서하, 김준홍, 김형석, 박재선, 강필성(2019)은 은행에 대한 민원데이터와 뉴스기사 등 텍스트에 대한 자연어처리 기법을 적용하여 조기경보 모형을 개발하였다. 이주희, 동학림(2018)은 상권정보를 활용하여 소상공인 부도예측 모형을 구성한 결과 의사결정나무(Decision Tree)와 인공신경망 기반의 모형이 판별분석 대비 분류성능이 우수함을 보였다.

다만, 이러한 기존 연구는 상장 또는 외감이상 규모를 가진 기업을 대상으로 분석한다. 개인사업자 및 중소기업에 대한 부실예측에 관한 연구도 일부 진행되었으나, 기계학습 기법을 활용한 연구는 드물다. 또한, 기업부실 예측 분야에서 새로운 자료를 활용하는 방향으로도 다양한 연구가 진행되었으나 자료의 활용보다는 모형의 성능추출 중심의 연구가 주를 이룬다. 현재까지 기업 여신거래에서 파생된 정보를 활용한 연구와 해당 자료의 특성을 중심으로 기계학습 기법 적용의 방향성을 제시하는 연구는 드물다.

본 연구가 기존 연구와의 차별성을 갖는 부분은 다음의 세 가지로 요약할 수 있다. 첫째, 재무자료를 중심으로 진행된 기존 연구와 달리 여신거래 정보를 활용하여 개인사업자를 포함한 전체 기업으로 분석의 대상을 확대하였다. 둘째, 기존 연구에서 새로운 자료의 활용이 유의성에 집중하였다면, 본 연구에서는 자료의 특성을 반영한 가공 변수의 활용까지도 고려하였다. 셋째, 기존 연구에서

기계학습 기법을 활용한 모형 성능개선에 중점을 두었다면, 본 연구는 기계학습 베이스라인 알고리즘을 특성에 따라 그룹화함으로써 모형 활용의 타당성을 검토하였다. 즉, 본 연구는 새롭게 활용되기 시작한 여신거래정보에 대한 분석에 있어 정보의 활용뿐 아니라 기계학습 기법 적용의 방향성을 제시하고 있다는 점에서 기여가 있다.

Ⅲ. 연구 설계

1. 연구 자료

본 연구에서는 2019년 5월에 출범한 한국신용정보원의 표본DB 원격분석시스템에서 제공하는 기업신용정보 표본DB를 활용한다. 표본DB는 신용정보원의 기업신용공여 제공 대상 기업으로 등록된 국내 개인사업자 및 법인 약 560만 개의 20% 수준에 해당하는 약 110만개 차주의 대출, 연체, 기술신용 정보로 구성되었으며 층화추출 기법이 적용되었다.[5] 또한, 층화추출을 위한 층화변수는 차주 특성정보 중 개인/법인여부와 지역으로 한정되었으나, 업종과 대출 및 연체 사유코드에 대해서도 표본 검증이 진행되었다. 본 DB에서는 16년 6월부터 20년 3월까지 총 46개월에 대한 월별 정보를 제공하고 있다. 또한, 대출의 경우 원화 대출금에 대한 자료가 구성되어 있고 연체의 경우 신용도 판단정보 기준 3개월 이상 연체 건을 기준으로 정보가 구성되어 있다.

2. 자료 분석 및 처리

2.1. 자료의 구성

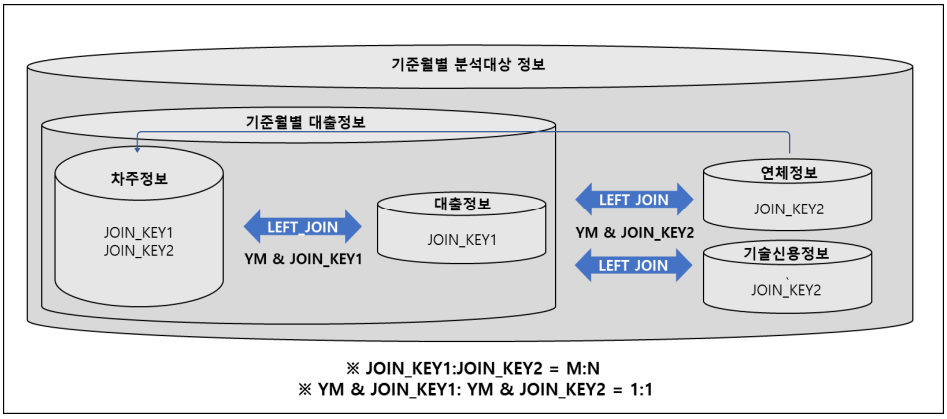
표본DB 원격분석시스템에서는 다음의 네 개의 테이블에 해당하는 정보를 제공한다. 첫째, 차주정보 테이블(table)을 통해 기준년월, 차주일련번호1, 차주일련번호2, 개인/법인코드, 업종코드, 지역코드, 등록년월 등 차주에 대한 식별값과 특성을 설명하는 총 7개의 칼럼(column)을 제공한다. 해당 테이블에는 2019년 6월과 2020년 3월을 기준으로 여신거래 이력을 보유한 차주정보가 존재한다.

5) 표본DB 원격분석시스템에서 제공하는 표본의 구성은 모집단 정의, 층화변수 정의, 표본크기 및 배분방법 결정, 표본추출, 추출결과 검증, 비식별 조치 등 총 6단계로 진행되었다. 특히, 표본추출에는 모집단과 표본에 관한 특성을 바탕으로 틀을 구성하는 방식의 층화추출 기법이 사용되었다.

또한, 차주일련번호1과 차주일련번호2는 각각 대출정보와 연체정보로 연결하기 위한 식별정보로 사용된다. 둘째, 대출정보 테이블은 차주 기업의 대출관련 정보를 기준월별 월말 시점의 스냅샷(snapshot) 형태로 제공한다. 여기서 기준년월은 대출정보가 등록된 기준년월을 의미하며 앞서 제시한 차주일련번호1를 통해 차주정보 테이블과의 연결이 가능하다. 한편, 대출이 발생한 업권을 7개로 구분하고 대출 계정에 따른 유형을 10개로 구분하여 대출업권코드와 대출 종류코드가 구성되어 있다. 셋째, 연체정보 테이블은 차주 기업의 연체 및 부도 관련 등록 및 해제 정보로 구성되어 있다. 넷째, 기술신용정보 테이블을 통해 기술신용평가를 진행한 차주에 대한 사업경쟁력과 기술경쟁력의 평가 등급정보를 제공하며, 해당 평가가 진행된 평가일자를 제공한다. 본 연구에서는 데이터의 특성에 따라 차주 식별값인 JOIN_KEY와 기준년월을 의미하는 YM 두 가지 항목을 활용하여 〈그림 1〉과 같이 차주별 기준월별 분석대상 정보를 구성하였다.

〈그림 1〉 분석대상 정보 구성도

표본DB 원격분석시스템에서 제공하는 차주정보 테이블에서는 JOIN_KEY1와 JOIN_KEY2와 같이 두 가지 차주일련번호를 식별정보로 제시한다. 본 DB에서 제공하는 총 네 개의 테이블은 그림과 같이 좌측에 제시된 차주정보 테이블을 기준으로 연결(LEFT JOIN)이 가능하다. JOIN_KEY1와과 기준년월을 의미하는 YM을 활용하여 차주정보와 대출정보간의 1:1연결이 가능하며, JOIN_KEY2와 YM을 기준년월을 의미하는 활용하여 차주정보와 연체정보 및 기술신용정보와의 1:1연결이 가능하다.



한편, 표본DB 원격분석시스템에서 제공하는 대출 및 연체 정보는 월 단위 스냅샷 형태를 보인다. 다만, 해당 스냅샷에서 연체 및 대출에 대한 식별번호는 제공하지 않기 때문에 개별 대출 및 연체 건에 대한 변동을 확인하는 것이 불가하다. 이를테면, 전월 대비 대출잔액의 증가가 확인되는 경우에도 대출잔액의 증가만이 있었는지 대출잔액 증가분이 감소분보다 많아 결과적으로 대출잔액이

증가한 것인지는 알 수 없다. 또한, 정보가 등록된 일자와 실제로 정보의 등록사유가 발생한 일자와의 차이가 존재하여 등록 기준년월 당월 실제 발생한 데이터를 확인하는 것이 제한적이다. 또한, 차주정보와 연체 및 대출을 연결하는 조인 키(join key)가 상이하여 대출과 연체에 대한 직접적인 연결이 불가한 바, 연체의 기준이 되는 대출을 확인할 수 없고 차주와 기준년월을 모두 공유하는 대출 및 연체정보에 대한 연결만이 가능하다. 이에 따라, 기준월별 말일을 기준으로 집계된 월별 데이터로 일별 데이터에 따른 시차(lag)가 반영되지 못한다. 또한, 연체의 경우 금액에 대한 정보가 제공되지 않고 천원 단위 이자 연체에서부터 원화 대출금 전체에 대한 연체까지 모든 경우를 포함하고 있어 채무불이행 대상 금액 파악이 제한되는 등, 리스크관리 시점의 위험에 노출(risk exposure)된 정도를 적절히 반영하지 못하는 한계가 있다.

2.2. 변수의 구성

본 연구에서는 자료의 특성을 반영한 분석을 수행하기 위해 앞서 확인된 자료를 재구성하였다. 분석대상 기간은 2019년 1월부터 2020년 1월까지 총 13개월로 특정하였다.[6] 또한, 2020년 1월 기준 대출정보가 등록된 차주 전체를 대상으로, 2020년 1월 당월 연체 발생여부에 따라 채무불이행과 정상을 구분한다. 연체 등록년월이 당월이면서 해제일자가 없는 미해제 연체사유코드를 보유한 경우를 당월 신규연체가 발생한 채무불이행으로 간주하였다.[7] 정상차주는 채무불이행 차주를 제외한 대출정보 보유 차주 전체로 정의하였다. 사업자 구분에 따라 정상 및 채무불이행 차주의 현황을 살펴본 결과 〈표 1〉과 같은 불균형 상태를 확인할 수 있었다.

〈표 1〉 개인/법인 구분에 따른 채무불이행 차주 구성

표본DB 원격분석시스템상 최신 차주정보등록년월인 '20년 1월을 기준으로 집계되었으며, 채무불이행 차주 비중은 정상 차주와 채무불이행 차주의 합계 대비 채무불이행 차주 수를 의미한다.

구분	개인	법인	합계
정상 차주	381,639	56,248	437,887
채무불이행 차주	477	333	810
합계	382,116	56,581	438,697
채무불이행 차주 비중	0.125%	0.589%	0.185%

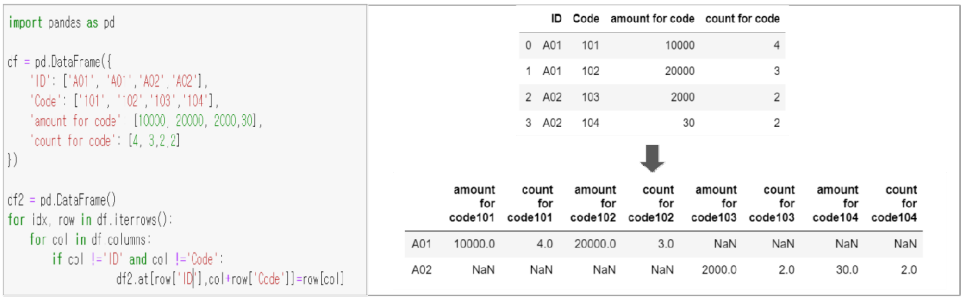
6) 원격분석시스템에서 제공하는 DBMS 성능 제약에 따라 분석기능 기간은 최대 1년으로 제한되며, 분석환경의 폐쇄성에 따라 거시경제 변수에 대한 결합은 불가하다.

7) 스냅샷 데이터 형태에 따라 과거 발생 연체도 정보 제공시점에 따라 당월에 등록되는 경우가 있으나, 해제시점이 당월 또는 과거일 경우 이미 조회시점 기준 연체가 해제된 것으로 볼 수 있다.

채무불이행 예측을 위한 독립변수를 구성하기 위해 표본DB 원격분석시스템에서 제공하는 각 테이블에서 출력한 원시데이터를 활용하였다.[8] 다만, 원시데이터만을 활용할 경우 각 차주가 보유한 여신 총액을 확인하는 것은 가능하지만, 대출 종류별 금액과 건수의 영향도에 따른 분석은 제한된다. 이에, 본 연구에서는 〈그림 2〉와 같은 중첩반복(overlapping iteration) 연산을 통해 대출과 채무불이행의 유형을 반영한 가공데이터를 생성하여 〈표 2〉와 같은 분석자료를 구성하였다.

〈그림 2〉 대출종류코드 및 연체종류코드 활용을 위한 데이터 구조 변경 연산

대출종류코드와 연체사유코드를 활용하기 위해 중첩반복 연산을 수행하였다. 본 연산을 통해 동일 차주(ID)의 특정 기준년월 정보에 대한 N개의 열(Row)로 구성된 테이블을 1개의 열을 가진 테이블로 구조를 변경하여 변수 구성을 확대하면서도 자료의 차원은 축소하였다.



```
import pandas as pd

df = pd.DataFrame({
    'ID': ['A01', 'A0','A02' 'A02'],
    'Code': ['101', '02','103','104'],
    'amount for code'  [10000, 20000, 2000,30],
    'count for code': [4, 3,2 2]
})

df2 = pd.DataFrame()
for idx, row in df.iterrows():
    for col in df.columns:
        if col !='ID' and col !='Code':
                df2.at[row['ID'],col+row['Code']]=row[col]
```

	ID	Code	amount for code	count for code
0	A01	101	10000	4
1	A01	102	20000	3
2	A02	103	2000	2
3	A02	104	30	2

	amount for code101	count for code101	amount for code102	count for code102	amount for code103	count for code103	amount for code104	count for code104
A01	10000.0	4.0	20000.0	3.0	NaN	NaN	NaN	NaN
A02	NaN	NaN	NaN	NaN	2000.0	2.0	30.0	2.0

〈표 2〉 신용정보 표본DB를 활용한 분석자료 구성

표본DB 원격분석시스템에서 제공하는 테이블을 쿼리(query)를 활용해 출력한 정보를 원시데이터로 구분하였으며, 원시데이터에 대해 중첩반복연산이 수행된 자료를 가공데이터로 구분하였다.

구분	대상 항목	내용
신용정보 원시데이터	대출정보	예측 기준이 되는 시점으로부터 최근 1개월부터 6개월까지를 관측시점으로, 각 시점별 차주의 대출기관수, 대출잔액
신용정보 가공데이터	대출정보	예측 기준이 되는 시점으로부터 최근 1개월, 3개월, 6개월 각 시점별 대출종류코드별 대출보유 건수 및 대출보유금액
	연체정보	예측 기준이 되는 시점으로부터 최근 3개월, 6개월, 12개월 각 시점별 연체등록 사유코드별 미해제 연체 발생여부
기술신용정보 가공데이터	기술신용정보	사업경쟁력등급(사업 경쟁력 수준에 따라 4단계로 구분) 및 기술경쟁력등급(기술 경쟁력 수준에 따라 4단계로 구분)

8) 본 자료에서 제시되는 기업의 고유 식별정보인 사업자번호는 가명처리 되었으며 재식별이 불가능하여, 채무불이행 예측에 관련한 기업 재무 및 개요 등의 추가 반영은 제한된다.

2.3. 탐색적 데이터 분석

연구 진행에 앞서 자료의 특성을 바탕으로 분석대상을 구분하고 자료 분포를 확인하기 위해 대출과 연체의 관점에서 탐색적 데이터 분석을 실시한다.[9] 첫째, 채무불이행 발생 시의 사업자 구분에 따른 부도율의 차이를 살펴본 결과, 분석기간 동안 법인의 경우 채무불이행 발생 시 20% 수준의 부도율을 보이나, 개인사업자의 부도율은 3%에 미치지 못하는 수준으로 확인되어 채무불이행 사건의 발생이 부도에 미치는 영향은 법인이 개인과 비교해 월등히 높게 나타나는 것이 확인되었다. 둘째, 차주의 업종과 소재 지역에 따른 대출기관 수와 대출잔액 현황을 확인한 결과, 중공업 및 철강업을 기반으로 대규모 산업단지가 분포한 울산지역에서 차주의 평균 대출잔액이 가장 높은 것으로 확인되었으며, 차주 수가 가장 많이 관측된 지역은 경기 및 서울지역, 총 대출잔액이 가장 많이 집계된 지역은 서울, 경기, 경남 순으로 확인되었다. 셋째, 기술신용평가 결과로 제시되는 등급에 따른 연체 및 정상차주의 분포를 통해서는, 기술신용평가를 통해 평정된 두 등급체계에서는 공통으로 등급이 우량한 기업일수록 채무불이행 차주에 대한 구성비가 감소하였다. 또한, 각 등급 모두 A 등급 구간에서 채무불이행이 발생한 사업자가 없었으며, D등급 구간에서는 채무불이행이 발생한 사업자의 비중이 가장 높게 나타났다.

2.4. 변수의 처리

기준월별 차주에 대한 연체정보는 단일 연체사유코드에 대한 발생년월 및 해재년월을 제시하고 있어, 차주마다 연체사유별 연체 보유여부에 따른 데이터 구성이 상이하다. 이에 따라, 차주 식별정보(join key)와 기준년월을 기준으로 사유코드별 연체 발생여부를 원핫인코딩을 통해 발생(1)과 미발생(0)으로 구분하여 종속변수를 구성하였다. 한편, 채무불이행 차주의 특성을 기반으로 분석대상 데이터에 대한 특성공학(feature engineering)[10] 방법론을 적용하여 독립변수를 구성하였다. 첫째, 결측 데이터에 대한 변수 제거를 진행하였다. 채무불이행 차주에 대해서만 대출 및 연체 변수에 따른 결측치를 확인한 결과 51개 변수에 대해서 보유 건이 전혀 없는 것으로 확인되었으며, 그 외 25개 변수에 대해서도 95% 수준의 차주에 대한 결측이 확인되었다. 다만, 소수 클래스에 대한 데이터 소실 방지를 위해 〈표 3〉과 같이 완전 결측을 보인 변수는 분석대상에서

9) 탐색적 데이터 분석(EDA: Exploratory Data Analysis)은 다양한 관점으로 분석대상 자료를 이해하고 관찰하여 데이터의 특징을 추출하는 과정을 의미한다(Tukey, 1977)

10) 특성공학(feature engineering)은 기계학습 알고리즘의 작동을 위해 도메인 지식을 활용해 특성(feature)을 구분하는 기법이다. 독립변수 중심의 통계 및 계량경제 분석과 달리 종속변수 기준의 그룹화를 통한 데이터 특징 추출도 가능하다(Zheng and Casari, 2018).

제외하였다.

둘째, 차주 특성정보에 대한 그룹화를 진행하였다. 산업분류 대분류를 기준으로 채무불이행 차주의 업종 구분은 3개 업종에 대한 구성이 각각 15% 이상으로 확인되어, 해당 3개 산업군에 해당하는 경우 연체 발생 가능성이 큰 요주의 업종으로 구분하고, 해당 3개 산업군에 해당하지 않는 경우 일반 업종으로 구분하였다. 또한, 채무불이행 차주의 지역분포를 확인한 결과 수도권 소재 기업이 전체 대비 38.27%로 확인되어, 채무불이행 차주의 지역 특성을 수도권 및 지방으로 구분하였다.

〈표 3〉 연체차주의 대출 가공데이터에 대한 결측률 확인

당월 채무불이행 발생 차주로 구분된 대상에서 관측되지 않는 비중을 결측률로 정의한 후 결측률 100%에 해당하는 항목에 대한 집계 결과를 항목 수에 제시하였다. 항목 상세에서 제공하는 각 대출종류코드 및 연체사유코드에 대한 정의는 [별표1]에 수록하였다.

구분	결측률	항목 수	항목 상세
대출잔액(원본)	100%	9개	대출종류코드 103, 111, 113 , 109 에 대한 각 기준월별 발생금액
대출 건수(원본)	100%	9개	대출종류코드 101, 103, 109, 111, 113, 115, 117, 129 에 대한 각 기준월별 발생건수
연체 여부(가공)	100%	14개	연체사유코드 0403, 0402, 0299, 0103, 0401, 0404 에 대한 각 기준월별 발생여부

분석자료 중 대출금액과 같은 수치형 변수의 경우 각 변수의 평균을 0, 분산을 1을 갖도록 정규분포로 변환하는 Standard-Scaler를 적용하였다. 이상치에 대한 영향도를 반영해야 하므로 Standard-Scaler의 사용은 Min-Max Scaler 또는 Robust Scaler보다 본 데이터 분류학습에 더욱 적합하다. 또한, 범주형 변수에 대해서는 종속변수 처리와 같은 방식의 인코딩을 통해 더미에 대한 전처리를 수행하였다.

3. 실험의 구성

본 연구에서는 채무불이행 예측에 대한 분석방법론의 적절성과 함께 기업신용정보 표본DB활용의 유용성을 검증하기 위해 표본DB에서 제공하는 대출, 연체, 차주, 기술신용 정보의 반영 여하에 따른 분류모형 성능개선 여부를 검증할 수 있도록 〈표 4 〉 및 〈그림 3〉과 같이 연구 과정을 설계하였다. 실험1에서 표본DB에서 제공하는 대출 정보 원본을 그대로 활용한 분류학습을 실행했다면, 실험2를 통해서는 원본 DB의 가공을 통해 대출종류코드에 따른 대출잔액 및 대출 건수가 채무불이행 예측에 유의하였는지를 살펴본다. 또한, 실험3에서는 개별 차주의 특성에 대한 영향도를 확인하기 위해

차주 특성정보를 반영한 결과를 제시하고, 실험4를 통해 기술신용평가 정보를 반영하여 차주의 사업경쟁력과 기술경쟁력에 대한 평가자료 활용의 유의성을 살펴보고자 한다.

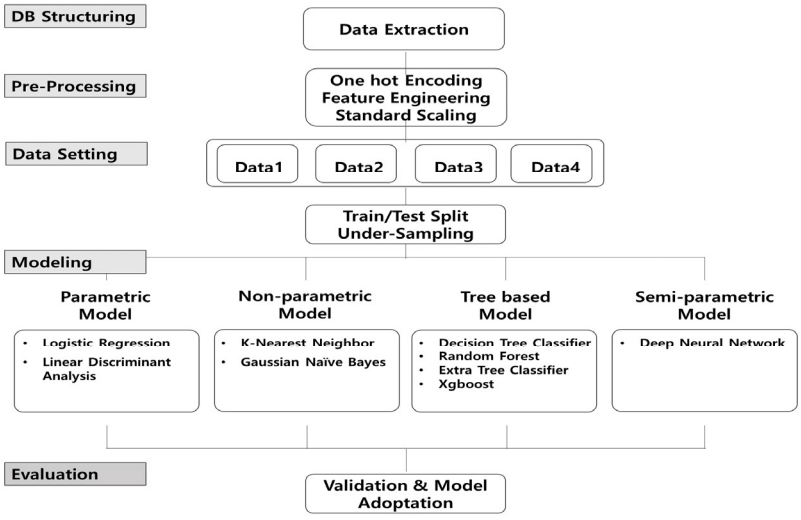
〈표 4〉 신용정보 표본DB 활용에 따른 실험 구성

본 연구에서는 구분의 정보에 따라 4개의 실험을 구성하였다. 동일 테이블에서 제공되는 데이터의 경우라도 원본과 가공본을 구분하였으며, 구성 항목 수를 통해 제시하는 각 항목의 상세 정보에 대해서는 [별표3]을 통해 제시한다.

구분	실험1	실험2	실험3	실험4	내용	구성 항목 수
대출잔액 (원본)	○	○	○	○	전체 대출에 대한 최근1개월부터 6개월까지 6개 시점의 월별 대출잔액	6개
대출 건수 (원본)	○	○	○	○	전체 대출에 대한 최근1개월부터 6개월까지 6개 시점의 월별 대출 건수	6개
연체 여부 (가공)	○	○	○	○	연체사유코드11개에 대한 최근1개월, 6개월, 12개월 시점의 대출 건수 정보	33개
대출잔액 (가공)		○	○	○	대출종류코드10개에 대한 최근1개월, 3개월, 6개월 시점의 대출잔액	30개
대출 건수 (가공)		○	○	○	대출종류코드10개에 대한 최근1개월, 3개월, 6개월 시점의 대출 건수	30개
차주 특성 (가공)			○	○	개인/법인구분, 지역구분, 업종구분	3개
기술 신용 (가공)				○	차주의 최근 기술신용평가 결과에 따른 사업경쟁력등급 및 기술경쟁력등급	2개

〈그림 3〉 분석 흐름도

본 연구의 흐름은 표본DB의 출력(data extraction)을 통한 분석자료 구성에서 시작한다. 자료의 분석을 바탕으로 데이터 처리를 수행하였으며, 구성된 실험에 사용되는 변수에 따라 데이터셋(dataset)을 4개로 분리한 이후 훈련 및 검증에 활용되는 데이터를 구성(data setting)하였다. 총 9가지 기계학습 기법을 활용해 분석모형을 구성하고, 학습결과를 통해 성능추출을 시도하였다.



또한, 본 연구에서는 모수 추정 방식에 따라 기계학습 알고리즘을 구분하여 각 세그먼트에 대한 실험을 수행하여 대출정보 분석방법론 선택의 방향성에 대해서도 살펴보고자 한다. 이에, 과세등록 구분에 따른 차주의 유형에 따라 전체 차주와 개인사업자, 그리고 법인으로 세그먼트를 구성하고, 각 방법론을 대표하는 총 9가지 기계학습 방법론에 따라 〈표 4〉와 같이 구성된 4개의 실험을 수행하여 〈그림 3〉의 절차에 따라 총 144회의 분류학습을 수행하였다.

Ⅳ. 실증 분석

1. 분석방법론

1.1. 언더샘플링을 통한 불균형 처리

본 연구에서 다루는 데이터는 예측 대상 레이블인 채무불이행 차주의 정보가 정상 차주의 정보 대비 매우 부족한 불균형 상태를 이루고 있다. 전체 차주 438,697개 사업자 중 채무불이행이 발생한 810개 사로 예측 대상에 대한 비중이 0.2%에 미치지 못하는 바, 데이터에 대한 언더샘플링(under-sampling)을 통한 훈련 및 검증 세트 재구성을 진행하였다. SMOTE(synthetic minority over-sampling technique)알고리즘과 부트스트래핑(bootstrapping)등을 활용한 오버샘플링(over-sampling)기반 재표본추출도 가능하나, 본 연구에서는 다수 클래스인 정상 차주 데이터에 대한 소실을 감안하더라도 소수 클래스에 대한 왜곡 없이 채무불이행 차주가 지닌 여신거래 특성에 대한 반영도를 최대화하기 위해 언더샘플링 기법을 최종 채택하였다.[11] 그 결과, 정상 차주의 데이터는 채무불이행 차주의 데이터와 동일하게 전체, 개인, 법인 등 세 가지 차주 구분에 따라 각각 810건, 333건, 477건으로 조정되었다. 또한, K=5수준의 교차검증(K-fold Cross Validation)을 수행하여 샘플링된 훈련 및 검증 자료의 편향을 최소화하였다.

1.2. 지도학습 방법론

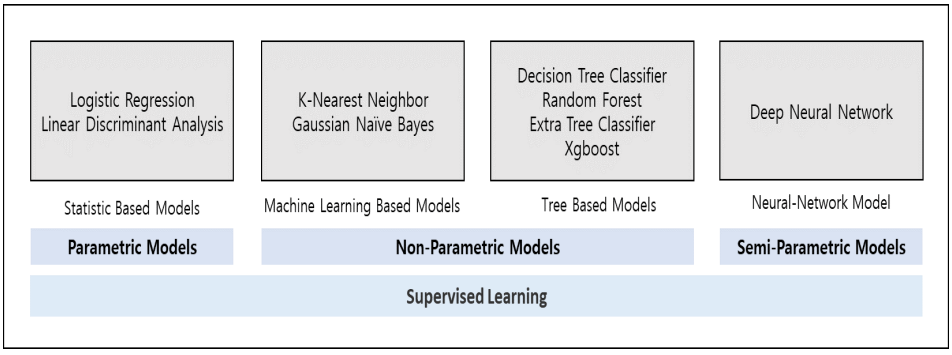
본 연구에서는 훈련데이터에 레이블을 포함하여 함께 훈련시키는 방식인 기계학습 지도학습

[11] 언더샘플링(Under-Sampling)은 데이터 소실 최소화가 필요한 경우 활용되는 기법이며 무작위 추출을 통해 재표본 데이터가 구성된다(Chawla, 2010).

알고리즘을 활용한 분석을 수행하였다. 지도학습 알고리즘은 모수를 추정하는 과정이 포함되어 있는지에 따라 모수적 방법론과 비모수적 방법론으로 구분된다. 모수적 방법론의 경우 오차항의 분포가 정규분포를 가정하는 방식으로 대표적으로 회귀 및 판별분석 등 통계 방법론에 기초한 알고리즘들이 이에 해당된다. 반면, 비모수적 방법론의 경우 확률분포와 관계없이 K-최근접이웃(K-Nearest Neighbors; 이하, KNN) 알고리즘과 같이 거리 척도에 기반하거나 트리계열 방법론과 같이 앙상블 기법을 활용하여 예측 및 분류 문제를 해결한다. 한편, 비모수적 방법론에 해당하나 앙상블 기법에 근간을 두고 있는 트리계열의 방법론에 대해서는 베이스라인 알고리즘 자체에 대한 확장이 활발히 진행되어 분류학습 분야에서 성능과 속도가 모두 개선되는 방향으로 발전되어 별도의 방법론으로 구분하였다. 마지막으로, 모수가 존재하나 확률분포를 가정하지 않는 준모수적 방법론을 활용하는 방법으로 심층신경망(Deep Neural Network; 이하, DNN) 알고리즘이 있다. 이에, 본 연구에서는 〈그림 4〉와 같이 기계학습 지도학습에서 각 방법론의 특징을 대표하는 알고리즘을 활용하여 성능을 추출하는 방식으로 분류학습을 수행하였으며, 각 방법론에 대한 하이퍼파라미터 최적화를 위해 그리드 서치(Grid Search)기법을 적용하였다.[12)]

〈그림 4〉 분석 방법론에 따른 지도학습 알고리즘

지도학습의 경우 모수 추정의 방식과 분석기법에 따라 4가지 부류로 알고리즘을 구분할 수 있다. 본 그림에서는 각 방법론의 특징을 대표하는 총 9개의 알고리즘을 제시하고 있다.



1.2.1 모수적 방법론

모수적 방법론에서 대표적으로 활용되는 알고리즘으로는 로지스틱 회귀(Logistic Regression;

12) 그리드 서치(Grid Search)는 기계학습 모델의 최적화를 위해 가능한 조합을 모두 시도하여 최고 성능을 도출하는 기법으로, 각 방법론에 활용된 주요 하이퍼파라미터는 [별표3]과 같다.

이하, LR)와 선형판별분석(Linear Discriminant Analysis; 이하, LDA)이 있다. LR의 경우, 종속변수가 0 또는 1인 이진 변수에 대하여 반응함수 값을 1로 수렴시키는 방식의 분류학습을 수행하는 모형이다. 분류분석에 활용되는 회귀분석 기반 방법론으로, 종속변수에 대하여 로그 기반 변수 변환의 과정을 거쳐 회귀식의 계수를 추정한다. 기계학습 분야에서 비교 대상인 베이스라인으로 가장 많이 활용되는 모델 중 하나로 과적합의 가능성이 낮고 오차를 최소화하여 관계의 선형성을 확인하는 것에 우수한 성능을 보인다(Olivia, 2001). 본 분석의 주제가 이진 분류에 해당하는 점을 비추어 볼 때 분석방법론으로 적합하다. 한편, LDA는 클래스 라벨(class label) 정보를 이용하여 각 클래스에 대한 분류를 수행하는 지도학습 기반의 알고리즘이다(Martinez, Kak, 2001). LDA는 2개 이상의 범주형 변수가 있을 경우 LR대비 분석 결과에 대한 안정성이 높은 분석방식으로, 종속변수가 주어졌을 때의 독립변수의 분포를 통하여 간접적인 방식으로 결과에 대한 추정을 진행한다. 분석대상 자료 중 차주에 대한 특성과 신용등급 정보가 범주형 변수로 주어진바, LR 대비 안정적인 분류성능을 보일 수 있는 분석방법론이다.

1.2.2 비모수적 방법론

비모수적 방법론에는 거리 계산을 기반으로 학습을 수행하는 KNN과 사전확률 계산을 통해 학습을 수행하는 나이브베이즈(Naive Bayes; 이하, NB)기법이 해당한다. KNN의 경우, 입력 및 출력으로 구성된 학습데이터에 대하여 새로운 입력 데이터가 주어질 경우, 데이터간 유클리드 거리를 기반으로 가장 가까운 거리를 갖는 값을 기준으로 분류를 수행하는 기법이다(이건명, 2019). 작은 데이터셋에 보다 적합한 형태의 알고리즘이나 기계학습 베이스라인 알고리즘의 하나로 이진분류 문제 해결을 위해 활용 가능한 분석방법론이다. 비지도학습에 해당하는 클러스터링(clustering)과 거리기반 분석을 수행한다는 면에서의 유사성이 있으나, 타겟이 되는 종속변수에 대한 클래스를 사전에 훈련에 포함하는 차이가 있어 지도학습에 해당한다.

한편, NB는 사전확률 정보와 관측을 통해 측정한 우도를 곱하여 산출한 사후 확률을 토대로 분류학습을 수행하는 알고리즘이다. 분류 대상 객체의 각 속성인 독립변수가 상호 독립의 관계에 있다는 전제하에 각 속성을 대상 카테고리로 분류하는데, 각 속성 변수가 대상 카테고리에 속할 확률을 산출한 후, 해당 확률이 가장 큰 값을 보이는 카테고리로 분류하게 된다(Alpaydin, 2014). 확률론적 방법 중에서는 유연하면서도 계산과정은 다소 복잡한 편이나, 확률에 기반하여 본 연구 주제에 대한 분석을 실시한다는 측면에서 방법론으로 활용에 의미가 있다(유석종, 2019).

1.2.3 트리계열 방법론

트리계열 방법론은 비모수적 방법론에 해당하지만 의사결정나무에 기반을 둔 알고리즘으로 결정트리(Decision Tree Classifier; 이하, DT), 랜덤포레스트(Random Forest; 이하, RF), 엑스트라트리(Extra Tree Classifier; 이하, ET), XGboost(Extreme Gradient Boosting; 이하, XGB)가 있다. DT는 지니계수 또는 엔트로피를 기반으로 불순도를 측정하여, 변수에 대한 이진분류 수행을 계속적으로 진행하여 트리를 구성하는 방식으로 분류학습을 수행한다. DT를 활용할 경우, 나무 구조의 도표를 통해 분류 규칙을 확인할 수 있어 설명력이 높고, 연속형 변수와 범주형 변수를 동시에 활용 가능하여 대출과 채무불이행뿐 아니라 차주 특성정보를 자료로 활용하는 본 분석에 적용이 적합하다(진서훈, 최종후 2005).

이에 대한 확장으로 제시되는 RF의 경우, 의사결정나무의 과적합 문제에 대한 개선안에서 출발한 알고리즘으로, 다수의 의사결정나무를 생성하여 정보이득(information gain)이 높은 특성(feature)을 선택하는 방식으로 배깅(bagging)을 수행하는 앙상블(ensemble) 학습을 통해 분류를 수행한다. 학습 프로세스에서 다양한 종류의 대출 및 연체 유형 각각에 대해 무작위로 변수를 선정하여 구성한 세트를 트리로 구성하여 분류학습을 수행한다는 점에서 앙상블 학습의 효과를 관찰할 수 있고, 의사결정나무 기반 학습에 대한 성능개선을 확인할 수 있는 기계학습기반 분류 분석의 베이스라인 방법론으로 볼 수 있다(박호연, 김경재, 2018).

한편, ET 역시 의사결정나무를 기반에 둔 분석기법으로, RF와 유사하나 DT를 의사결정나무가 아닌 ExtraTreeClassifier를 사용한다는 점에서 차이가 있다. ET는 학습 과정에서 각 변수의 선택뿐 아니라 선택된 변수의 특성을 무작위로 분할하는 방식으로 무작위성을 증가시키는데, 대량의 데이터를 분석하는 경우에 의사결정나무 기법 대비 빠른 연산속도를 보인다(Sharma, Giri, Granmo, and Goodwin, 2019).

마지막으로 XGB의 경우, 실제값과 예측값의 차이인 잔차를 감소시키기 위해 순차적으로 모델을 개선하며 학습하는 Gradient Boosting 알고리즘에 기반하면서도 단일 트리를 활용하는 의사결정나무기법 대비 학습 손실을 줄이면서 속도와 성능을 모두 개선시킨 알고리즘이다(Chatzis, Siakoulis, Petropoulos, Stavroulakis, and Vlachogiannakis, 2018). 트리계열 앙상블 학습 중 빠른 성능과 과적합 방지 기능을 통해 분류와 회귀 문제에 있어 높은 성능을 보여, 본 분석에서도 트리계열 다른 방법론 대비 유의한 성능을 관측할 수 있을지 확인할 필요가 있다.

1.2.4 준모수적 방법론

입력층(input layer)과 출력층(output layer)사이 은닉층(hidden layer)를 구성하여 학습을 수행하는 방식에 대하여 인공신경망으로 정의하며, 은닉층의 개수가 2개 이상인 경우 심층신경망으로 분류된다(Bayraci and Susuz, 2019). 신경망 기법을 활용할 경우, 학습 과정에서 활성화 함수(activation function)을 통해 비선형관계에 대한 출력값과 이에 대한 가중평균합계를 계산하며 연산을 수행하는 형태가 실제 신경망을 구성하는 뉴런의 연산구조와 동일하다는 특징이 있다(Nielsen, 2015). 신경망 학습은 데이터 본연의 특성에 대한 별도의 로짓처리 및 가중치 부여에 대한 영향도가 다른 기계학습 기법 대비 상대적으로 낮은 편이며, 변수 간의 비선형관계를 찾아내는 것에 우수한 성능을 보인다. 다만, 복잡한 모형 구조로 인하여 과대 적합의 경향이 있고 의사결정나무 또는 회귀분석 대비 해석이 제한된다는 단점이 있으나(Ripley, 1996), 드롭아웃(Drop-out) 및 배치 정규화(Batch-Normalization) 등 전체적인 학습 과정에 대한 안정화 기법이 고안되어 회귀와 분류뿐 아니라 강화학습에도 적용되고 있다.

1.3. 성능평가

일반적인 분류모형의 성능평가는 혼동행렬(confusion matrix)을 기준으로 진행된다. 분류에 대한 시행과 실제 결과에 대한 모든 경우의 수를 제공하는 방식이기 때문이다. 이에, 본 연구에서는 〈표 5〉와 같이 채무불이행 예측 혼동행렬을 구성하였다. 본 행렬을 활용할 경우 양성을 음성으로 예측한 위음성률(false negative)와 실제 음성을 양성으로 예측한 위양성률(false positive)를 통해 제1종 오류(type1 error)와 제2종 오류(type2 error)를 확인할 수 있다. 다만, 국내 금융기관의 보수적인 여신리스크관리 관행을 반영할 경우, 채무불이행이 발생하지 않을 것으로 예측했으나 실제로 채무불이행이 발생하는 위음성률에 주목할 필요가 있다.

〈표 5〉 채무불이행 예측 혼동행렬

이진 분류기의 혼동행렬은 실제값에 대한 예측값의 조합을 제공하며, 이를 활용하여 진양성률(True Positive), 위양성률,(False Positive), 위음성률(False Negative), 진음성률(True Negative)를 구할 수 있다. 각 성과지표는 0에서 1사이의 값을 갖는다.

		실제	
	구분	채무불이행	정상
예측	채무불이행	True Positive	False Positive
	정상	False Negative	True Negative

한편, 혼동행렬에서 제시한 True Positive(TP), True Negative(TN), False Negative(FN), False Positive(FP)를 토대로 아래 수식과 같이 성능평가 수행을 위한 평가지표를 계산할 수 있다. 정확도(Accuracy)는 혼동행렬에서 제시하는 모든 경우 중 부도와 정상을 바르게 예측한 경우를 의미하며, 정밀도(Precision)는 채무불이행으로 예측한 경우 중 실제 채무불이행 차주로 확인된 비율이다. 또한, 재현율(Recall)은 실제 채무불이행 차주에 대하여 채무불이행 차주로 옳게 예측한 비율로 정밀도와 상반된 의미를 갖는다. 한편, TP와 FP의 비율을 통해 ROC(Receiver Operating Characteristic) 곡선 이하의 면적을 의미하는 AUC(Area Under the Curve)를 계산할 수 있다.[13)]

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

2. 결과 분석

신용정보 표본DB 활용에 따라 실험을 구성하고, 분석 방법론에 따라 그룹(Group)을 구성하여 차주 세그먼트별 성능평가를 수행하였다. 〈표 6〉은 대출 거래 총 건수와 총 잔액 등 대출정보 원본만을 활용해 데이터 분석을 수행한 결과를 제시한다. 이 경우 대출 종류별 잔액 및 건수 등에 대해서는 고려하지 않는다. AUC를 활용한 분류성능 평가 결과, 모든 차주 세그먼트에 대해 준모수적 방법론에 해당하는 DNN의 분류성능이 가장 우수한 것으로 나타났다. 또한, KNN알고리즘의 경우 DNN에 이어 우수한 성능을 보였으나, 정확도가 낮게 확인되는 점을 감안할 경우 모수적 방법론 기반의 LR과 트리계열 방법론에서 더욱 안정적인 분류학습을 수행할 수 있는 것으로 확인된다. 한편, 개인 세그먼트에 대해서는 전체와 법인 대비 전체 분석방법론에 대한 평균 분류성능이 낮은 것으로 확인되었다. 이는 개인의 경우 정상과 채무불이행에 대한 불균형 분포가 세 가지 세그먼트 중 가장 심하게 나타나는 가운데, 본 실험에서 대출종류에 대한 구분이 없는 원본 자료 활용에 따라 학습대상 특성자료 자체가 부족한 것에 기인하는 결과로 추정된다.

13) 일반적인 상황에서 이진분류기의 AUC(Area Under the Curve)는 0.50에서 1사이의 값을 갖는다.

〈표 6〉 실험1: 대출정보 원본을 활용한 기계학습 모형별 채무불이행 예측 성능

Seg1, Seg2, Seg3은 각각 차주 구분에 따라 전체, 개인, 법인을 의미한다. Group1에서 Group4까지는 각 알고리즘의 계열에 따라 구분되어, Group1은 모수적 방법론, Group2는 비모수적 방법론, Group3은 트리계열 방법론, Group4는 준모수적 방법론을 의미한다. 각 세그먼트에 대한 성능평가 지표로 제시된 Acc는 정확도(Accuracy), Pre는 정밀도(Precision), Rec는 재현율(Recall)을 의미한다. Mean은 그룹별 성능평가 결과의 산술평균이며, Average는 전체 성능평가 결과의 산술평균이다.

Model		Seg1				Seg2				Seg3			
		Acc	Pre	Rec	Auc	Acc	Pre	Rec	Auc	Acc	Pre	Rec	Auc
Group 1	LR	0.811	0.005	0.476	0.644	0.488	0.002	0.802	0.645	0.744	0.009	0.385	0.565
	LDA	0.802	0.004	0.458	0.631	0.558	0.002	0.659	0.609	0.718	0.010	0.477	0.598
	Mean	0.807	0.005	0.467	0.637	0.523	0.002	0.731	0.627	0.731	0.009	0.431	0.582
Group 2	KNN	0.603	0.003	0.708	0.655	0.712	0.002	0.473	0.593	0.535	0.008	0.677	0.606
	NB	0.970	0.004	0.065	0.518	0.968	0.002	0.055	0.512	0.978	0.016	0.046	0.515
	Mean	0.786	0.004	0.387	0.587	0.840	0.002	0.264	0.552	0.757	0.012	0.362	0.560
Group 3	DT	0.839	0.005	0.429	0.634	0.444	0.002	0.835	0.640	0.450	0.008	0.754	0.601
	RF	0.821	0.005	0.464	0.643	0.465	0.002	0.802	0.634	0.614	0.009	0.631	0.622
	ET	0.778	0.004	0.500	0.639	0.436	0.002	0.824	0.630	0.709	0.007	0.354	0.533
	XGb	0.818	0.005	0.464	0.642	0.541	0.002	0.681	0.611	0.609	0.010	0.677	0.643
	Mean	0.814	0.005	0.464	0.640	0.472	0.002	0.786	0.629	0.596	0.009	0.604	0.600
Group 4	DNN	0.828	0.005	0.484	0.724	0.456	0.002	0.786	0.655	0.676	0.011	0.672	0.732
	Mean	0.828	0.005	0.484	0.724	0.456	0.002	0.786	0.655	0.676	0.011	0.672	0.732
Average		0.808	0.005	0.450	0.637	0.563	0.002	0.658	0.614	0.670	0.010	0.519	0.602

한편, AUC 기준의 분류성능 측정과 함께 재현율에 대해 살펴볼 필요가 있다. 재현율은 정밀도와 함께 혼동행렬을 통해 산출할 수 있는 성능평가 지표이나, 두 지표는 서로 상충관계(trade-off)에 있다. 재현율은 채무불이행 차주를 정상으로 예측하는 오류를 줄이는 것에 중점을 두고 있다는 점에서 본 연구와 같이 리스크관리 분야의 문제를 다루는 경우 정밀도보다 보수적인 의미를 갖는 지표라고 볼 수 있다. 실험1에서는 모든 세그먼트에서 정밀도 대비 재현율이 높게 나타나 채무불이행 차주를 분류가 리스크관리 목적에 적합한 방향으로 학습이 수행된 것을 확인할 수 있었다.

〈표 7〉은 대출 종류별 정보를 반영한 가공데이터를 활용한 실험2의 모형별 분류성능이 원본데이터만 활용한 실험1 대비 각 모형이 보이는 성능이 전반적으로 개선되는 결과를 제시하고 있다. 이를 통해 대출 종류별 잔액과 건수를 반영한 가공 자료의 활용에 대한 유의성을 확인할 수 있었다. 모든 차주 세그먼트에서 DNN 모형의 성능이 가장 높게 나타났으며, 전체 및 개인 차주에 대해서는 ET모형이, 법인 차주에 대해서는 RF모형이 다음으로 높은 분류성능을 보였다. 한편, LDA와 DT 등 일부 모형에서 분류성능이 다소 감소하였으나, 전반적인 방법론별 분류성능이

〈표 7〉 실험2: 대출 가공정보를 활용한 기계학습 모형별 채무불이행 예측 성능

Seg1, Seg2, Seg3은 각각 차주 구분에 따라 전체, 개인, 법인을 의미한다. Group1에서 Group4까지는 각 알고리즘의 계열에 따라 구분되어, Group1은 모수적 방법론, Group2는 비모수적 방법론, Group3은 트리계열 방법론, Group4는 준모수적 방법론을 의미한다. 각 세그먼트에 대한 성능평가 지표로 제시된 Acc는 정확도(Accuracy), Pre는 정밀도(Precision), Rec는 재현율(Recall)을 의미한다. Mean은 그룹별 성능평가 결과의 산술평균이며, Average는 전체 성능평가 결과의 산술평균이다.

Model		Seg1				Seg2				Seg3			
		Acc	Pre	Rec	Auc	Acc	Pre	Rec	Auc	Acc	Pre	Rec	Auc
Group 1	LR	0.823	0.005	0.494	0.659	0.531	0.002	0.769	0.650	0.798	0.014	0.492	0.646
	LDA	0.811	0.005	0.446	0.629	0.461	0.001	0.670	0.566	0.734	0.009	0.431	0.583
	Mean	0.817	0.005	0.470	0.644	0.496	0.002	0.720	0.608	0.766	0.012	0.462	0.615
Group 2	KNN	0.810	0.005	0.536	0.673	0.879	0.004	0.374	0.626	0.809	0.014	0.462	0.636
	NB	0.995	0.025	0.036	0.517	0.989	0.014	0.121	0.556	0.987	0.075	0.108	0.550
	Mean	0.903	0.015	0.286	0.595	0.934	0.009	0.247	0.591	0.898	0.045	0.285	0.593
Group 3	DT	0.881	0.005	0.321	0.602	0.487	0.002	0.747	0.617	0.556	0.009	0.708	0.631
	RF	0.831	0.006	0.518	0.675	0.478	0.002	0.846	0.662	0.635	0.012	0.754	0.694
	ET	0.822	0.006	0.542	0.682	0.383	0.002	0.945	0.664	0.938	0.034	0.354	0.648
	XGb	0.834	0.006	0.518	0.676	0.512	0.002	0.769	0.641	0.660	0.012	0.692	0.676
	Mean	0.842	0.006	0.475	0.659	0.465	0.002	0.827	0.646	0.697	0.017	0.627	0.662
Group 4	DNN	0.890	0.008	0.466	0.760	0.492	0.002	0.802	0.725	0.884	0.023	0.455	0.750
	Mean	0.890	0.008	0.466	0.760	0.492	0.002	0.802	0.725	0.884	0.023	0.455	0.750
Average		0.855	0.008	0.431	0.652	0.579	0.003	0.672	0.634	0.778	0.022	0.495	0.646

개선되었을 뿐 아니라 트리계열 및 DNN 모형의 성능 증가 폭이 더 크게 나타나 평균 분류성능이 개선된 것으로 확인되었다. 특히, 준모수적 방법론을 활용한 DNN의 경우 법인 세그먼트에서 직전 실험 대비 7% 수준의 성능개선이 관측되었는데, 개인사업자에 대한 대출 종류별 잔액 및 건수 정보가 비선형성을 가진 신경망 학습에 유의한 영향을 미친 것으로 해석된다. 한편, 실험1에서와 마찬가지로 정밀도 대비 재현율이 높게 관측되어 채무불이행 차주에 대한 분류가 정상적으로 수행된 것이 확인되었다. 다만, KNN과 NB 등 트리계열이 아닌 비모수적 방법론의 경우 평균 재현율이 네 가지 방법론 그룹 중 가장 낮게 나타난 것으로 확인되었다. 그중에서도 NB의 경우 세그먼트에 따라 재현율이 10%에서 30%까지도 나타나 타 모형 대비 현저한 차이를 보였는데, 베이지안 확률 기반의 학습이 분류 수행을 방해하는 것으로 해석된다.

〈표 8〉은 기업 채무불이행 예측에 있어 차주 특성정보를 활용한 실험3의 결과를 제시한다. 준모수적 방법론과 트리계열 방법론에서 역시 안정적인 분류성능이 도출된 가운데, 차주특성으로 주어진 지역 및 업종이라는 더미 변수의 활용은 성능개선에 유의한 결과를 제시하지 못하는 것으로 확인된다.

〈표 8〉 실험3: 차주 특성정보를 활용한 기계학습 모형별 채무불이행 예측 성능

Seg1, Seg2, Seg3은 각각 차주 구분에 따라 전체, 개인, 법인을 의미한다. Group1에서 Group4까지는 각 알고리즘의 계열에 따라 구분되어, Group1은 모수적 방법론, Group2는 비모수적 방법론, Group3은 트리계열 방법론, Group4는 준모수적 방법론을 의미한다. 각 세그먼트에 대한 성능평가 지표로 제시된 Acc는 정확도(Accuracy), Pre는 정밀도(Precision), Rec는 재현율(Recall)을 의미한다. Mean은 그룹별 성능평가 결과의 산술평균이며, Average는 전체 성능평가 결과의 산술평균이다.

Model		Seg1				Seg2				Seg3			
		Acc	Pre	Rec	Auc	Acc	Pre	Rec	Auc	Acc	Pre	Rec	Auc
Group 1	LR	0.753	0.005	0.601	0.677	0.578	0.002	0.780	0.679	0.657	0.010	0.585	0.621
	LDA	0.790	0.004	0.470	0.630	0.454	0.002	0.725	0.590	0.588	0.006	0.446	0.518
	Mean	0.772	0.004	0.536	0.654	0.516	0.002	0.753	0.634	0.623	0.008	0.515	0.569
Group 2	KNN	0.791	0.005	0.536	0.663	0.849	0.003	0.396	0.622	0.671	0.011	0.615	0.643
	NB	0.996	0.026	0.036	0.517	0.989	0.012	0.099	0.545	0.987	0.075	0.108	0.550
	Mean	0.893	0.015	0.286	0.590	0.919	0.008	0.247	0.583	0.829	0.043	0.362	0.597
Group 3	DT	0.824	0.005	0.494	0.659	0.484	0.002	0.791	0.637	0.924	0.025	0.323	0.625
	RF	0.825	0.006	0.536	0.681	0.483	0.002	0.857	0.670	0.638	0.012	0.754	0.695
	ET	0.840	0.006	0.512	0.676	0.420	0.002	0.912	0.666	0.907	0.029	0.462	0.686
	XGb	0.794	0.005	0.560	0.677	0.530	0.002	0.769	0.649	0.676	0.012	0.677	0.677
	Mean	0.821	0.006	0.525	0.673	0.479	0.002	0.832	0.656	0.786	0.019	0.554	0.671
Group 4	DNN	0.698	0.004	0.680	0.751	0.586	0.002	0.713	0.722	0.806	0.016	0.500	0.742
	Mean	0.698	0.004	0.680	0.751	0.586	0.002	0.713	0.722	0.806	0.016	0.500	0.742
Average		0.812	0.007	0.492	0.659	0.597	0.003	0.671	0.642	0.762	0.022	0.497	0.640

가공정보를 활용하지 않는 실험1에 비해서는 모든 차주 구분에서 개선된 성능이 관측되었으나, 실험2 대비 대부분의 모형에 대한 성능변화가 미미한 수준이다. 트리계열의 경우 전반적으로 소폭의 성능 개선결과가 나타났으나 신경망의 경우 1% 이내의 미미한 변화가 나타났다. 한편, 실험3을 통해 확인된 재현율은 앞선 실험에서와 같이 트리계열이 아닌 비모수적 방법론을 활용할 경우 가장 낮게 나타나는 것을 확인할 수 있다.

〈표 9〉에서는 앞서 살펴본 채무불이행 차주와 기술신용평가정보의 상관관계를 토대로 기술경쟁력과 사업경쟁력 반영에 따른 모형별 분류성능 개선을 파악하는 실험4의 결과를 제시한다. 가공정보를 활용하지 않는 실험1에 비해서는 역시 모든 차주 구분에서 개선된 성능이 관측되었으나, 가공정보를 바탕으로 모형을 구성한 실험2와 실험3에 비교할 경우 분류성능에 유의한 영향을 미치지 못하는 것으로 확인되었다. 다만, 법인 차주 세그먼트에서 DNN 등 일부 모형에 대해서는 2% 내외의 성능개선이 나타났다.

실험4의 경우에도 방법론별 성능에 대해서는 준모수적 방법론 기반 DNN의 성능이 모든

〈표 9〉 실험4: 기술신용정보를 활용한 기계학습 모형별 채무불이행 예측 성능

Seg1, Seg2, Seg3은 각각 차주 구분에 따라 전체, 개인, 법인을 의미한다. Group1에서 Group4까지는 각 알고리즘의 계열에 따라 구분되어, Group1은 모수적 방법론, Group2는 비모수적 방법론, Group3은 트리계열 방법론, Group4는 준모수적 방법론을 의미한다. 각 세그먼트에 대한 성능평가 지표로 제시된 Acc는 정확도(Accuracy), Pre는 정밀도(Precision), Rec는 재현율(Recall)을 의미한다. Mean은 그룹별 성능평가 결과의 산술평균이며, Average는 전체 성능평가 결과의 산술평균이다.

Model		Seg1				Seg2				Seg3			
		Acc	Pre	Rec	Auc	Acc	Pre	Rec	Auc	Acc	Pre	Rec	Auc
Group 1	LR	0.760	0.005	0.613	0.687	0.610	0.002	0.725	0.667	0.674	0.010	0.585	0.629
	LDA	0.786	0.004	0.411	0.599	0.454	0.001	0.659	0.557	0.531	0.007	0.554	0.542
	Mean	0.773	0.004	0.512	0.643	0.532	0.002	0.692	0.612	0.602	0.009	0.569	0.586
Group 2	KNN	0.788	0.005	0.536	0.662	0.833	0.003	0.374	0.603	0.784	0.014	0.523	0.654
	NB	0.996	0.026	0.036	0.517	0.993	0.032	0.154	0.574	0.987	0.075	0.108	0.550
	Mean	0.892	0.015	0.286	0.589	0.913	0.017	0.264	0.589	0.886	0.045	0.315	0.602
Group 3	DT	0.882	0.007	0.423	0.653	0.283	0.002	0.967	0.624	0.861	0.014	0.338	0.601
	RF	0.830	0.006	0.506	0.668	0.475	0.002	0.835	0.655	0.658	0.012	0.723	0.690
	ET	0.822	0.006	0.554	0.688	0.439	0.002	0.857	0.648	0.934	0.038	0.431	0.684
	XGb	0.826	0.006	0.530	0.678	0.564	0.002	0.714	0.639	0.699	0.013	0.692	0.696
	Mean	0.840	0.006	0.503	0.672	0.440	0.002	0.843	0.642	0.788	0.019	0.546	0.668
Group 4	DNN	0.818	0.005	0.535	0.756	0.463	0.002	0.831	0.727	0.880	0.024	0.458	0.761
	Mean	0.818	0.005	0.535	0.756	0.463	0.002	0.831	0.727	0.880	0.024	0.458	0.761
Average		0.834	0.008	0.460	0.656	0.568	0.005	0.680	0.633	0.779	0.023	0.490	0.645

세그먼트에서 가장 우수한 것으로 확인되었으며, 트리계열에 해당하는 분류모형에서 다음으로 우수한 성능이 관측되었다. 전체 차주에 대해서는 모수적 방법론 중 LR모형에 대해서도 70%에 가까운 분류성능이 관측되었는데, 이는 기술신용평가정보의 자료 형태가 선형관계를 나타내기 적합한 것에 따른 결과로 해석된다.

〈표 10〉과 같이 준모수적 방법론 기반의 DNN의 경우 본 연구에서 수행한 모든 실험 및 차주 구분에 대하여 가장 높은 분류성능을 보였다. 본 연구에서 사용된 신경망은 〈그림 5〉와 같이 14개의 은닉층으로 구성되었으며, 신경망 베이스라인 설계에 따라 입력층의 활성화 함수에는 시그모이드(sigmoid) 함수를, 은닉층의 경우 렐루(relu) 함수를 적용하였고, Adam Optimizer를 통해 최적화를 수행하였다. 또한, 학습손실(loss)의 경우 채무불이행 예측값과 실제값이 같을 때 손실함수의 값을 1로 출력하는 이진 교차엔트로피(binary cross entropy)를 활용하였으며, 출력층에서 0.5 수준의 드롭아웃을 적용하여 계산비용을 절감하면서도 과적합과 가중치 소멸을 방지하였다. DNN 모형은 AUC 기준의 분류성능뿐 아니라 재현율에 대해서도 안정적인 성능을 보여 채무불이행 차주 분류에 적합한 모형임을 확인할 수 있었다.

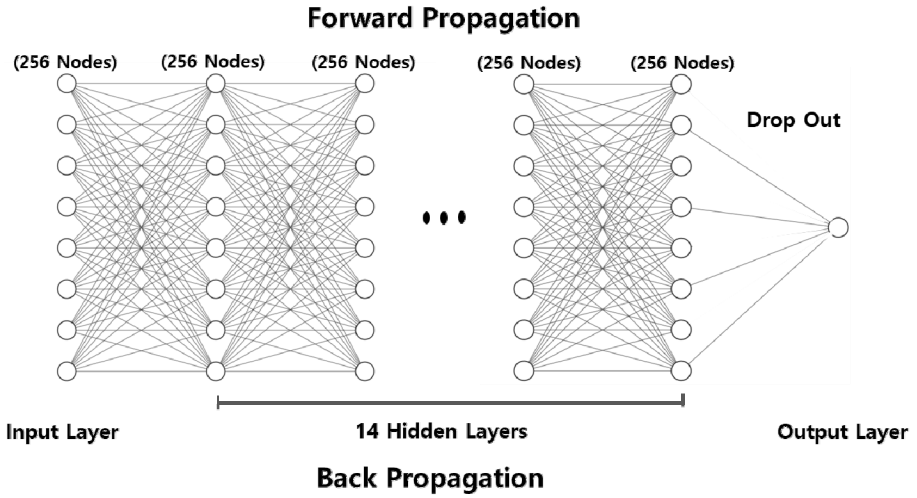
〈표 10〉 차주 세그먼트에 따른 방법론별 채무불이행 예측 평균 성능

Seg1, Seg2, Seg3은 각각 차주 구분에 따라 전체, 개인, 법인을 의미하며, Group1은 모수적 방법론, Group2는 비모수적 방법론, Group3은 트리계열 방법론, Group4는 준모수적 방법론을 의미한다. 각 세그먼트별 성능평가 지표로 제시된 Acc는 정확도(Accuracy), Pre는 정밀도(Precision), Rec는 재현율(Recall)을 의미하며, 각 수치들은 Group에서 실행한 전체 실험결과의 산술평균을 반영한다.

Model	Seg1				Seg2				Seg3			
	Acc	Pre	Rec	Auc	Acc	Pre	Rec	Auc	Acc	Pre	Rec	Auc
Group1	0.792	0.005	0.496	0.644	0.517	0.002	0.724	0.620	0.681	0.009	0.494	0.588
Group2	0.869	0.012	0.311	0.590	0.902	0.009	0.255	0.579	0.842	0.036	0.331	0.588
Group3	0.829	0.006	0.492	0.661	0.464	0.002	0.822	0.643	0.717	0.016	0.583	0.650
Group4	0.809	0.006	0.541	0.748	0.499	0.002	0.783	0.707	0.812	0.019	0.521	0.746

〈그림 5〉 심층신경망 모형 구조도

준모수적 방법론 기반의 심층신경망은 총 16개의 층(Layer)로 구성되었으며, Input Layer, Hidden Layer, Output Layer는 각각 입력층, 은닉층, 출력층을 의미한다. 출력층을 제외한 각 층의 노드는 256개이다. 출력층에서는 0.5 수준의 드롭아웃을 적용하였다.



물론, 이는 AUC 측면에서 라쏘(lasso) 변수선택을 통해 상장사를 대상으로 90%를 상회하는 부도예측을 수행한 권누리, 김영민, 최광신(2019)의 연구결과 대비 낮은 성과를 나타낸 것으로 해석할 수 있다. 다만, 본 연구에서는 채무불이행 발생 기업이 전체의 0.2%에 해당하는 심한 불균형 상태에 해당한다는 측면에서, 전체 1,254개의 분석대상 기업 중 부도기업이 608개인 48.4%로 균형에 가까운 분포를 나타내는 권누리, 김영민, 최광민(2019)의 연구결과와 동일선상에서의 비교는

제한적이다. 한편, 본 연구와 같이 여신거래정보를 기반으로 기업이 아닌 일반 개인에 대한 채무불이행 예측을 시도하여 70% 수준의 재현율과 5% 미만의 정밀도를 관측한 조남용, 안동욱, 박훈, 여신영, 심창운(2019)의 연구와 유사한 연구결과를 나타내는 수치이다.

한편, 금융사를 제외한 미국 내 재무정보와 주가정보를 모두 활용하여 시계열 부도예측을 수행한 Kim , Cho, and Ryu(2021)의 연구에서의 성능평가 결과와는 유사한 수준의 예측력을 확인할 수 있다. 해당 연구에서는 0.6% 이하의 정밀도와 21%에서 47% 수준의 재현율이 관측되었으며, 앙상블 모형에서는 가장 높은 73.1% 수준의 AUC가 관측되어 본 연구에서의 성능평가 결과와 매우 유사한 지표별 성능평가 결과 분포를 확인할 수 있었다.

한편, 비모수적 방법론 중 트리계열에 해당하는 분류모형의 경우 DNN 대비 최대 10% 수준까지 성능차이가 나타나지만 AUC와 재현율을 종합적으로 고려할 경우 타 방법론 대비 안정적인 분류를 수행한다. 하지만, 트리계열이 아닌 비모수적 방법론의 경우 재현율이 낮게 관측될 뿐 아니라 재현율의 변동폭도 실험에 따라 크게 나타난 것으로 확인되어 보수적인 리스크관리 목적의 분류모형 구성에는 적합하지 않은 방법론으로 확인된다. 또한, 실제 여신금융기관 및 신용조회사의 부도율 예측 모형 수립에 가장 활발히 활용되는 모수적 방법론은 준모수적 방법론 및 트리계열 방법론 대비 열위한 분류성능을 보이는 것이 확인되었다. 최근 국내 부동산 시장에 관해 부실예측을 위해 조기경보 모형을 구축한 연구(박대현, 김정환, 류두진, 2021)에서도 기계학습 방법론 중 모수 추정에 기반한 모형에서 상대적으로 열위한 성과를 나타내었다는 것으로 확인되어 부실화 예측에 있어 모수적 방법론의 활용은 제한적이라는 본 논문의 연구결과를 활용한 해석이 가능하다.

Ⅴ. 결 론

본 연구에서는 국내 기업신용정보 표본DB를 활용한 자료 분석과 함께 기계학습 기반의 채무불이행 예측을 위한 분류모형 수립 방향성에 대해 분석하였다. 자료 분석의 경우, 모형 구축에 앞서 국내기업 전체를 모집단으로 하는 대표성 높은 표본DB를 활용하여 차주 구분에 따른 다각도의 분석을 수행했을 뿐 아니라 부실예측 문제에 활용을 위해 채무불이행 차주를 중심으로 자료의 활용 방향을 모색하였다는 것에 의의가 있다. 또한, 실제로 한국신용정보원으로부터 기업신용정보를 입수하여 활용하는 국내 여신금융기관에서는 동 기관에서 발생한 정보만을 활용할 수 있는 상황이나, 본 연구는 전체

여신금융기관의 정보를 바탕으로 구성된 여신거래정보 분석결과를 제시하고 있어 리스크관리 대상에 대한 개별 여신금융기관의 정보 비대칭을 완화할 여지가 있다. 무엇보다 실제 한국신용정보원의 기업신용정보를 입수하여 활용하는 국내 여신금융기관에 여신정보 활용 및 모형 구축의 방향성을 제시한다는 점에서 실무적 함의를 갖는다.

본 연구에서는 기계학습 방법론과 학습대상 차주 세그먼트에 따른 총 144회의 분류학습 실험 구성을 통해 채무불이행 예측 자료의 활용 및 방법론 적용의 방향성에 대해서도 살펴보았다. 분류학습 결과, 첫째, 대출 계정에 따른 대출잔액 및 대출 건수 정보 등 가공 변수의 활용이 데이터 원본만을 활용한 경우보다 분류 분석에 있어 최대 7% 수준의 유의한 성능개선 효과를 보이는 것을 확인했다. 둘째, 특성공학을 기반으로 전처리된 기업의 소재 지역과 업종 등 차주 특성정보와 기술신용평가 정보의 활용은 채무불이행 예측모형의 성능개선에 유의한 영향을 미치지 못하는 것으로 확인되었다. 셋째, 준모수적 방법론을 활용한 신경망 방식의 학습이 여신거래정보를 활용한 기업 채무불이행 예측에 가장 적합하며, 재현율에 대한 안정적인 관측이 제한적인 트리계열 외 비모수적 방법론의 활용은 채무불이행 예측에 적합하지 않은 것으로 확인된다. 또한, 고전 연구에서 활용되었을 뿐 아니라 실무적으로 가장 많이 사용되는 모수적 방법론의 경우에도 주요 세그먼트에 대한 평균 분류성능은 준모수적 방법론의 활용 대비 열위할 뿐 아니라 최대 16% 수준까지 성능 차이가 나타난 것으로 확인되어 기업신용정보 분석에 대한 새로운 방향의 접근이 필요함을 시사한다.

다만, 분류학습의 성능 향상을 위한 재표본추출, 표준정규화, 특성공학 등 다양한 방식으로 자료의 특성을 반영하기 위한 데이터 처리뿐 아니라 신경망을 제외한 8개 모형에 대해서는 0.1%이내 성능변동 관측 시까지 그리드 서치 기반의 최적화와 21개 하이퍼파라미터에 대한 튜닝(tuning)이 수행되었음에도, 심한 데이터 불균형에 따라 심층신경망을 제외한 각 기계학습 기법에서는 주요 성능평가 지표인 AUC가 재무자료 활용 기업부도 예측에 관한 선행연구 대비 상대적으로 낮은 수준으로 확인되었다. 금번 연구는 실험수행과 모형 구축이 인터넷 통신이 제한되는 폐쇄된 원격분석 환경을 통해서 수행되었다. 또한, 분석환경의 폐쇄성으로 인해 딥러닝 분석을 위해 요구되는 의존성을 모두 확보하는 것에 제약이 있었으며, 분석환경 메모리의 한계로 전체 기간에 대한 시계열 데이터 구성 및 학습이 제한되어 분석대상 기간 내 특정 시점을 기준으로 앞선 시점에서 발생한 대출 및 채무불이행에 대한 변수를 구성하는 방식으로 학습과 검증을 진행하였다. 추후, 원격분석시스템 분석환경을 개선하여 상대적으로 성능이 우수한 준모수적 방법론 중심으로 자료와 모형을 확장하여 분석할 필요가 있다.

추가로, 분석대상 데이터 중 업종 및 지역으로 구성된 차주 특성정보의 경우 차주 기업의 속성을 충분히 반영하지 못하는 항목으로 학습 효과에 개선을 가져오지 못하는 것으로 확인되었다. 또한, 기술신용평가 정보 역시 등급 외 정보 활용에 대한 제약에 따라 상관분석 결과와 달리 자료 활용에 따른 유의한 예측력 개선을 발견하지는 못했다. 추후, 표본DB 원격분석시스템에서 실제 국내 개인사업자 및 법인에 대한 설명력을 높이는 방향으로 기업차주의 특성정보를 추가하여 분석자료를 구성할 수 있다면, 채무불이행 예측에 있어서 보다 유의미한 실험이 가능할 것으로 예상한다. 또한, 기술기업의 가치평가뿐 아니라 기술금융 관련 정책적 진단을 제시하는 방향의 연구도 후속하여 진행될 수 있을 것이다.

References

권누리, 김영민, 최광신, “거시경제 변수를 고려한 한국기업부도 모형 구축 방법 연구,” 한국데이터정보과학회지, 제30권 제5호 (2019), pp. 1037-1050.

(Translated in English) Gwon, N. R., Y. M. Kim, and K. S. Choi, “On modeling Korean corporate bankruptcy using macroeconomic variables,” *Journal of the Korean Data and Information Science Society*, Vol. 30, No. 5 (2019), pp. 1037-1050.

권혁건, 이동규, 신민수, “RNN(Recurrent Neural Network)을 이용한 기업부도예측모형에서 회계정보의 동적 변화 연구,” 한국지능정보시스템학회지, 제23권 제3호 (2017), pp. 139-153.

(Translated in English) 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), pp. 139-153.

김형준, 류두진, 조훈, “기업부도예측과 기계학습,” 금융공학연구, 제18권 제3호 (2019), pp. 131-152.

(Translated in English) Kim, H. J., D. J. Ryu, and H. Cho, “Corporate default predictions and machine learning,” *Korean Journal of Financial Engineering*, Vol. 18, No. 3 (2019), pp. 131-152.

박대현, 김정환, 류두진, “기계학습 기반 주택시장의 조기경보체계,” 부동산분석, 제7권 제1호 (2021), pp. 29-45.

(Translated in English) Park, D. H., J. H. Kim, and D. J. Ryu, “Early warning system of housing

market using machine learning," *Journal of Real Estate Analysis*, Vol. 7, No. 1 (2021), pp. 29-45.

박호연, 김경재, "시뮬레이티드 어니일링 기반의 랜덤 포레스트를 이용한 기업부도예측," 지능정보연구, 제24권 제4호 (2018), pp. 155-170.

(Translated in English) Park, H. Y., and K. J. Kim, "Predicting corporate bankruptcy using simulated annealing-based random forests," *Journal of Intelligence and Information Systems*, Vol. 24, No. 4 (2018), pp. 155-170.

송서하, 김준홍, 김형석, 박재선, 강필성, "금융 데이터 및 텍스트 데이터를 활용한 금융 기업 조기 경보 모형 개발: 부실은행 예측을 중심으로," 대한산업공학회지, 제45권 제3호 (2019), pp. 248-259.

(Translated in English) Song, S. H., J. H. Kim, H. S. Kim, J. S. Park, and P. S. Kang, "Development of early warning model for financial firms using financial and text data: A case study on insolvent bank prediction," *Journal of the Korean Institute of Industrial Engineers*, Vol. 45, No. 3 (2019), pp. 248-259.

엄하늘, 김재성, 최상옥, "머신러닝 기반 기업부도위험 예측모델 검증 및 정책적 제언: 스태킹 앙상블 모델을 통한 개선을 중심으로," 지능정보연구, 제26권 제2호 (2020), pp. 105-129.

(Translated in English) Eom, H. N., J. S. Kim, and S. O. Choi, "Machine learning-based corporate default risk prediction model verification and policy recommendation: Focusing on improvement through stacking ensemble model," *Journal of Intelligence and Information Systems*, Vol. 26, No. 2 (2020), pp. 105-129.

유석종, "나이브 베이즈 분류를 활용한 부동산 추천 기법 연구," 한국정보기술학회논문지, 제17권 제10호 (2019), pp. 115-120.

(Translated in English) Yu, S. J., "A study on recommendation method for real estate using naive bayes classification," *Journal of KIIT*, Vol. 17, No. 10 (2019), pp. 115-120.

이건명, 인공지능 (제2판), 생능출판, 2019.

(Translated in English) Lee, K. M., *Artificial intelligence* (2nd Ed.), Life and Power Press, 2019.

이민준, 이정환, "대출 포트폴리오와 대출금 변수를 중심으로 본 국내은행 신용위험의 결정요인," 산업경제연구, 제32권 제1호 (2019), pp. 49-76.

(Translated in English) Lee, M. J., and J. H. Lee, “Determinants of credit risks of Korean banks: A comprehensive study,” *Journal of the Korean Data And Information Science Society*, Vol. 32, No. 1 (2019), pp. 49-76.

이상은, “빅데이터분석에 정보적 표본설계의 적용,” 통계연구, 제24권 제3호 (2019), pp. 33-49.

(Translated in English) Lee, S. E., “Application of informative sampling on big data,” *Journal of the Korean Official Statistics*, Vol. 24, No. 3 (2019), pp. 33-49.

이주희, 동학림, “소상공인의 자금공급 확대를 위한 빅데이터 활용 방안연구,” 벤처창업연구, 제13권 제3호 (2018), pp. 125-140.

(Translated in English) Lee, J. H., and H. L. Dong, “Research on the application methods of big data within SME financing: Big data from trading-area,” *Asia-Pacific Journal of Business Venturing and Entrepreneurship*, Vol. 13, No. 3 (2018), pp. 125-140.

조남용, 안동욱, 박훈, 여신영, 심창운, “금융 빅데이터 개방시스템(CreDB)을 활용한 채무불이행 예측 모형에 관한 연구,” 한국IT서비스학회 학술대회 논문집 (2019), pp. 186-189.

(Translated in English) Jo, N. Y., D. W. An, H. Park, S. Y. Yeo, and C. W. Sim, “A machine learning based loan delinquency prediction using korean Financial Bigdata Open System(CreDB),” *Korea Society of IT Service* (2019), pp. 186-189.

진서훈, 최종후, 데이터 마이닝의 현장, 자유아카데미, 2005.

(Translated in English) Jin, S. H., and J. H. Choi, *Data mining*, Freeca, 2005.

차성재, 강정석, “딥러닝 시계열 알고리즘 적용한 기업부도예측모형 유용성 검증,” 지능정보연구, 제24권 제4호 (2018), pp. 1-32.

(Translated in English) Cha, S. J., and J. S. Kang, “Corporate default prediction model using deep learning time series algorithm, RNN and LSTM,” *Journal of Intelligence and information Systems*, Vol. 24, No. 4 (2018), pp. 1-32.

한국은행, 2019년 기업경영분석, 2020.

(Translated in English) Bank of Korea, *Financial statement analysis for 2019*, 2020.

Alpaydin, E., *Introduction to machine learning* (4th Ed.), The MIT Press, 2020.

Altman, E. I., “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy,” *Journal of Finance,* Vol. 23, No. 4 (1968), pp. 589-609.

Altman, E. I., *Corporate financial distress and bankruptcy: A complete guide to predicting and avoiding distress and profiting from bankruptcy* (2nd Ed.), Wiley, 1993.

Bayraci, S., and O. Susuz, "A Deep Neural Network (DNN) based classification model in application to loan default prediction," *Theoretical and Applied Economics*, Vol. 26, No. 4 (2019), pp. 75-84.

Beaver, W. H., "Financial ratios as predictors of failure," *Journal of Accounting Research*, Vol. 4, No. 3 (1966), pp. 71-111.

Campbell, J. Y., J. Hilscher, and J. Szilagyi, "In search of distress risk," *Journal of Finance*, Vol. 63, No. 6 (2008), pp. 2899-2939.

Chatzis, S. P., V. Siakoulis, A. Petropoulos, E. Stavroulakis, and N. Vlachogiannakis, "Forecasting stock market crisis events using deep and statistical machine learning techniques," *Expert Systems with Applications*, Vol. 112 (2018), pp. 353-371.

Chava, S., and R. A. Jarrow, "Bankruptcy prediction with industry effects," *Review of Finance*, Vol. 8 (2004), pp. 537-569

Chawla, N. V., *Data mining for imbalanced datasets: An overview*, Springer, Boston, MA, 2010.

Hinton, G. E., S. Osindero, and Y. W. Teh, "A fast learning algorithm for Deep Belief Nets," *Neural Computation*, Vol. 18, No. 7 (2006), pp. 1527-1554.

Kim, H. J., H. Cho, and D. J. Ryu, "Corporate bankruptcy prediction using machine learning methodologies with a focus on sequential data," *Computational Economics* (2021), pp. 1-19.

Martinez, A. M., and A. C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 2 (2001), pp. 228-233.

Nielsen, M., *Neural networks and deep learning*, Determination Press, 2015.

Odom, M. D., and R. Sharda, "A neural network model for bankruptcy prediction," *IJCNN International Joint Conference* (1990), pp. 163-168.

Ohlson, J. A., "Financial ratios and the probabilistic prediction of bankruptcy," *Journal of Accounting Research*, Vol. 18, No. 1 (1980), pp. 109-131.

Olivia, P. R., *Data mining cookbook: Modeling data for marketing, risk, and customer relationship management* (1st Ed.), Wiley, 2001.

Ripley, B. D., *Pattern recognition and neural networks* (1st Ed.), Cambridge University Press, 1996.

Samuel, A. L., "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, Vol. 3, No. 3 (1959), pp. 210-229.

Sharma, J., C. Giri, O. C. Granmo, and M. Goodwin, "Multi-layer intrusion detection system with ExtraTrees feature selection, extreme learning machine ensemble, and softmax aggregation," *EURASIP Journal on Information Security* (2019), pp. 1-16.

Shumway, T., "Forecasting bankruptcy more accurately: A simple hazard model," *Journal of Business*, Vol. 74, No. 1 (2001), pp. 101-124.

Tam, K. Y., and M. Y. Kiang, "Managerial applications of neural networks: The case of bank failure predictions," *Management Science*, Vol. 38, No. 7 (1992), pp. 926-947.

Tukey, J. W., *Exploratory data analysis* (1st Ed.), Addison-Wesley Publishing Company, 1977.

Zhang, G., M. Y. Hu, B. E. Patuwo, and D. C. Indro, "Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis," *European Journal of Operational Research*, Vol. 116, No. 1 (1999), pp. 16-32.

Zheng, A., and A. Casari, *Feature engineering for machine learning: Principles and techniques for data scientists* (2nd Ed.), O'Reilly Media, Inc., 2018.

Zmijewski, M. E., "Methodological issues related to the estimation of financial distress prediction models," *Journal of Accounting Research*, Vol. 22 (1984), pp. 59-82.

Appendix

[별표 1] 대출 종류코드 및 연체사유코드

기업신용정보 표본DB원격분석시스템에서 제공되는 대출 및 연체 종류 구분코드를 수록하였다. 대출종류코드 10개와 연체사유코드 11개로 구분되며, 연체사유코드는 4개의 부도사유코드를 포함한다.

1) 대출종류코드

대출종류코드	대출종류명
101	상업어음할인
103	당좌대출
105	일반운전자금
107	일반시설자금
109	무역금융
111	신탁계정대출
113	기업어음할인
115	전자방식외상매출채권담보대출
117	구매자금대출
129	기타

2) 연체사유코드

구분	연체사유코드	연체사유	
연체	0101	대출금연체	대출원금 · 이자연체(3개월 이상)
	0103		장기카드대출(카드론)연체(3개월 이상)
	0104		신용카드대금연체(3개월 이상)
	0199		기타(대출금연체 중 0101, 0103, 0104 사유 외)
	0201	지급보증 대지급	지급보증대지급금(3개월 이상)
	0202		신용보증대지급금
	0299		기타(대위변제 · 대지급정보 중 0201, 0202 사유 외) **0203,0204,0205,0207
부도	0404	부도사실	부도사실 어음 · 수표부도사실
	0401	최종부도	가계수표최종부도
	0402		당좌수표최종부도
	0403		약속어음최종부도

[별표 2] 실험별 변수의 구성

본 실험에 활용된 변수에 대하여 대출 및 연체코드와 기준년월(yyyymm 또는 yymm)을 반영한 변수명을 제시한다. 변수 구분의 i는 독립변수이며, d는 종속변수를 의미한다.

No.	변수구분	구분	변수명	실험1	실험2	실험3	실험4	완전결측여부
1	d	채무불이행 여부	DLQ_CD	O	O	O	O	
2	i	차주 속성	I_C_FLAG	O	O	O	O	
3	i	대출 건수(원본)	LN_TOTAL_CNT_201912	O	O	O	O	
4	i	대출 건수(원본)	LN_TOTAL_CNT_201911	O	O	O	O	
5	i	대출 건수(원본)	LN_TOTAL_CNT_201910	O	O	O	O	
6	i	대출 건수(원본)	LN_TOTAL_CNT_201909	O	O	O	O	
7	i	대출 건수(원본)	LN_TOTAL_CNT_201908	O	O	O	O	
8	i	대출 건수(원본)	LN_TOTAL_CNT_201907	O	O	O	O	
9	i	대출 잔액(원본)	LN_TOTAL_AMT_201912	O	O	O	O	
10	i	대출 잔액(원본)	LN_TOTAL_AMT_201911	O	O	O	O	
11	i	대출 잔액(원본)	LN_TOTAL_AMT_201910	O	O	O	O	
12	i	대출 잔액(원본)	LN_TOTAL_AMT_201909	O	O	O	O	
13	i	대출 잔액(원본)	LN_TOTAL_AMT_201908	O	O	O	O	
14	i	대출 잔액(원본)	LN_TOTAL_AMT_201907					O
15	i	대출 건수(가공)	LN_CNT129_1912		O	O	O	
16	i	대출 건수(가공)	LN_CNT129_1910		O	O	O	
17	i	대출 건수(가공)	LN_CNT129_1907		O	O	O	
18	i	대출 건수(가공)	LN_CNT117_1912		O	O	O	
19	i	대출 건수(가공)	LN_CNT117_1910					O
20	i	대출 건수(가공)	LN_CNT117_1907		O	O	O	
21	i	대출 건수(가공)	LN_CNT115_1912		O	O	O	
22	i	대출 건수(가공)	LN_CNT115_1910		O	O	O	
23	i	대출 건수(가공)	LN_CNT115_1907		O	O	O	
24	i	대출 건수(가공)	LN_CNT113_1912					O
25	i	대출 건수(가공)	LN_CNT113_1910					O
26	i	대출 건수(가공)	LN_CNT113_1907					O
27	i	대출 건수(가공)	LN_CNT111_1912					O
28	i	대출 건수(가공)	LN_CNT111_1910					O
29	i	대출 건수(가공)	LN_CNT111_1907					O
30	i	대출 건수(가공)	LN_CNT109_1912		O	O	O	
31	i	대출 건수(가공)	LN_CNT109_1910		O	O	O	
32	i	대출 건수(가공)	LN_CNT109_1907		O	O	O	
33	i	대출 건수(가공)	LN_CNT107_1912		O	O	O	
34	i	대출 건수(가공)	LN_CNT107_1910		O	O	O	
35	i	대출 건수(가공)	LN_CNT107_1907		O	O	O	
36	i	대출 건수(가공)	LN_CNT105_1912		O	O	O	
37	i	대출 건수(가공)	LN_CNT105_1910		O	O	O	
38	i	대출 건수(가공)	LN_CNT105_1907		O	O	O	
39	i	대출 건수(가공)	LN_CNT103_1912					O
40	i	대출 건수(가공)	LN_CNT103_1910					O
41	i	대출 건수(가공)	LN_CNT103_1907					O
42	i	대출 건수(가공)	LN_CNT101_1912		O	O	O	
43	i	대출 건수(가공)	LN_CNT101_1910		O	O	O	
44	i	대출 건수(가공)	LN_CNT101_1907		O	O	O	
45	i	대출 잔액(가공)	LN_AMT129_1912		O	O	O	
46	i	대출 잔액(가공)	LN_AMT129_1910		O	O	O	
47	i	대출 잔액(가공)	LN_AMT129_1907		O	O	O	
48	i	대출 잔액(가공)	LN_AMT117_1912		O	O	O	
49	i	대출 잔액(가공)	LN_AMT117_1910		O	O	O	
50	i	대출 잔액(가공)	LN_AMT117_1907		O	O	O	
51	i	대출 잔액(가공)	LN_AMT115_1912		O	O	O	
52	i	대출 잔액(가공)	LN_AMT115_1910		O	O	O	
53	i	대출 잔액(가공)	LN_AMT115_1907		O	O	O	
54	i	대출 잔액(가공)	LN_AMT113_1912					O

[별표 3] 기계학습 모형별 주요 하이퍼파라미터(Hyper Parameter)

본 실험에 활용된 모형을 구성하는 주요 하이퍼파라미터를 제시한다. Model은 각 기계학습 모형을 의미하며, Param은 각 모형에서 제시하는 하이퍼파라미터의 명칭이다. Type, default를 통해 각 자료의 유형과 초기설정 값을 제시한다.

Model	Param	Type	Default
Logistic Regression	fit_intercept	bool	TRUE
	max_iter	int	100
Linear Discriminant Analysis	solver	{'svd', 'lsqr', 'eigen'}	'svd'
	n_components	int	None
	store_covariance	bool	FALSE
K-Neighbors Classifier	n_neighbors	int	5
	metric	str or callable	'minkowski'
Gaussian NB	var_smoothing	float	1e-9
Decision Tree Classifier	max_depth	int	None
	max_features	int, float or {"auto", "sqrt", "log2"}	None
	random_state	int, RandomState instance or None	None
Random Forest Classifier	max_depth	int	None
	random_state	int, RandomState instance or None	None
Extra Trees Classifier	max_depth	int	None
	random_state	int, RandomState instance or None	None
XGB Classifier	max_depth	int	reg:linear
	booster	–	gbtree
	alpha	int, default=0	0
Deep Neural Network	activation	{'identity', 'logistic', 'tanh', 'relu'}	'relu'
	solver	{'lbfgs', 'sgd', 'adam'}	'adam'
	hidden_layer_sizes	tuple, length = n_layers - 2	(100,)