

재무정보 및 동태적 채권거래 정보를 활용한 부도위험 예측 연구

정정윤¹, 유헌창²

¹고려대학교 SW·AI 융합대학원 빅데이터융합학과 석사과정

²고려대학교 컴퓨터학과 교수

god129@naver.com, yuhc@korea.ac.kr

Predicting Bankruptcy Risk in the Construction Industry Using Financial Data and Dynamic Trade Credit Transactions

Jungyoun Jung¹, Heonchang Yu²

¹Dept. of Bigdata Convergence, Graduate School of SW·AI Convergence, Korea University

²Dept. of Computer Science & Engineering, Korea University

요 약

본 연구는 건설업종을 대상으로 기업 부도위험을 조기에 탐지하기 위해, 전통적 재무제표 정보와 함께 금융결제원의 채권거래 데이터를 활용하는 새로운 접근을 제안한다. 재무제표는 기업의 건전성을 평가하는 대표적 수단이지만, 보고 주기가 길어 적시성이 부족하다는 한계가 있다. 이에 반해 채권거래 데이터는 기업 간 지급·결제 활동을 반영하여 유동성과 거래 안정성을 보다 민감하게 포착할 수 있다. 실험 결과, 채권거래 기반 파생변수를 포함한 모형은 ROC 만으로는 미세한 차이가 드러나지 않았으나 AUPRC 지표에서 일관된 개선을 보여 회귀 사건인 부도 식별력 향상에 기여하였다. 이러한 결과는 건설업 부도위험 관리에서 재무제표와 거래정보를 통합적으로 고려할 필요성을 시사하며, 금융기관과 정책당국의 선제적 리스크 관리에도 실질적인 활용 가능성을 제공한다.

1. 서론

기업 부도 예측은 금융 산업뿐만 아니라 국가 경제 전반에서 신용 위험을 사전에 방지하고 관리하기 위한 핵심 과제로 꾸준히 연구되어 왔다. 최근 들어 고금리·고물가·고환율이라는 복합 경제 위기 속에서 기업들의 재무 건전성이 약화되면서, 특히 부동산 프로젝트 파이낸싱(PF) 부실을 기점으로 건설업계에 연쇄적인 부도가 발생하는 사례가 주목받고 있다. 대형 건설사의 부도 위험은 단일 기업 차원을 넘어 다수의 하청업체, 협력사, 나아가 해당 프로젝트에 자금을 공급한 금융기관까지 직·간접적으로 영향을 미치며, 하나의 부도가 산업 생태계 전체의 리스크로 확산되는 파급 효과를 야기한다. 이러한 이유로 다양한 산업 중에서도 건설업종에 대한 부도위험 예측 연구의 필요성이 그 어느 때보다 높아졌다.

전통적인 부도 예측 연구는 주로 재무제표(회계) 정보와 주식시장 데이터를 기반으로 수행되었다[1]. 재무제표는 기업의 재무 현황을 객관적이고 표준화된

방식으로 제공한다는 장점이 있으나, 데이터 생성 주기가 연간 단위로 길어 적시성이 떨어지는 근본적인 한계가 존재한다.

본 연구는 이러한 한계에 대응하여, 일별 발생하는 기업의 매입/매출 채권거래 데이터를 활용하여 건설업종 기업의 부도위험을 예측하는 새로운 방법론을 제시한다. 금융결제원의 채권거래 데이터는 기업 간 실제 지급·결제 네트워크를 반영하기 때문에, 기업의 단기 유동성과 거래 안정성을 시의성 있게 파악할 수 있다는 점에서 기존 재무제표 정보를 보완할 수 있을 것이다. 특히 건설업종은 원도급사-하청업체-자재공급사 등 복잡한 공급망 구조를 가지고 있으며, 어느 한 기업의 부도가 거래 네트워크 전반으로 확산될 가능성이 크다는 점에서, 거래 데이터의 활용은 실무적 가치가 높다.

따라서 채권거래 데이터를 전통적인 재무제표와 결합하여 예측 성능을 향상시키는 동시에, 데이터 불균형 문제를 완화하여 기업의 위험 요인을 식별하고 개

선 방향을 도출하고자 한다.

2. 관련 연구

기업 부도 예측 연구는 1960년대부터 본격적으로 진행되었으며, 초기에는 전통적인 통계 기법들이 주로 활용되었다. 1980년대 후반부터는 인공신경망(ANN), 의사결정나무(Decision Tree), 서포트 벡터 머신(SVM)과 같은 비모수적 머신러닝 기법을 활용한 연구가 진행되기 시작했다[2]. 특히 인공신경망은 기업 부도 예측 연구에서 가장 높은 예측력을 보이며 활발하게 활용되었다. Odom and Sharda(1990)는 인공신경망이 기존 판별 분석보다 높은 예측 성능을 보여주었으며, Tam and Kiang(1992), Wilson and Sharda(1994), Jo and Han(1996) 등의 연구에서도 인공신경망의 우수성이 입증되었다. 그러나 인공신경망은 과적합 문제와 연구자가 임의로 설계해야 하는 하이퍼파라미터가 많아 모델 설계에 많은 시간이 소요된다는 문제가 제기되었다.

최근에는 랜덤 포레스트(Random Forest)와 XGBoost와 같은 앙상블 기법이 뛰어난 예측 성능을 보이며 주목받고 있다. 랜덤 포레스트는 다수의 의사결정나무 결과를 결합하여 예측 오차를 낮추며, 의사결정나무 수가 증가해도 과적합되지 않는다는 장점이 있다[3]. XGBoost는 학습 손실을 최소화하면서 과적합을 방지하기 위해 나무의 복잡도를 통제하는 방식으로 최적화된 모델을 생성하며, 빠른 처리 속도와 뛰어난 성능으로 다양한 데이터 분석 경진 대회에서 우승 알고리즘으로 활용되어 왔다. 실제로 빅데이터를 이용한 딥러닝 기반의 기업 부도예측 연구에서는 랜덤 포레스트와, SVM과 DNN(Deep Neural Network)도 높은 예측력을 나타냈다[4].

3. 부도위험 예측 방안

3.1 데이터 셋

본 연구에서 금융결제원 채권거래 데이터의 활용 효과를 검증하기 위해 실험 데이터셋을 설정하고 비교 분석을 수행한다(그림 1). 외부감사 대상 기업 중 건설업종에 해당하는 약 5,000여 개 기업을 분석 대상으로 하고 기업의 공시자료 기준 향후 1년간의 부도를 측정하도록 설계하였다.

구체적으로, 2022년 공시자료가 존재하는 기업의 경우 2023년 5월부터 2024년 5월까지의 기간 동안 실제 부도 발생 여부를 관측하여 종속변수(Target)로 설정하였다. 동일한 방식으로, 2023년 공시자료가 존재하는 기업의 경우에는 2024년 5월부터 2025년 5월까지의 부도 여부를 예측 대상으로 한다. 이러한 설계는 특정 시점의 재무제표와 거래 데이터를 기반

으로, 이후 1년간의 단기적인 부도 발생을 예측하는 구조를 갖는다.

특히 본 연구는 동일한 기업이라 하더라도 연도별로 구분하여 독립적인 관측치로 처리한다. 예를 들어, A기업의 2022년 재무제표와 거래 데이터는 2023년 공시 후 부도 여부를 예측하는 데 사용되고, 2023년 재무제표와 거래 데이터는 다시 2024년 공시 후 부도 여부를 예측하는 데 활용된다. 즉, 같은 기업이라도 시점별로 다른 표본으로 간주한다. 이러한 접근 방식은 일반적인 부도 예측 연구에서 장기간의 관측 기간을 설정하는 방식과 달리, 단기적인 예측 정확도를 높이기 위한 설계 전략에 해당한다.

실제 건설업은 프로젝트 단위의 매출 구조와 원도급-하청-자재공급사로 이어지는 다층적 거래망을 특징으로 한다. 따라서 경기 변동, 원자재 가격 상승, 금융시장 불안정 등 외부 요인에 따라 단기간 내에도 급격한 유동성 악화와 연쇄적 부도가 발생할 수 있다. 이와 같이 건설업종 산업적 특수성을 반영할 때, 1년 단위의 단기 예측이 실무적으로 적합하다고 판단하였다.

SET A 건설업 전체							
기업명_년도	부도 여부	총자산 금액 (백만원)	부채 비중	...	최근1년 채권거래 건수	최근1년도 채권거래 금액 (백만원)	전년도 대비 거래 증감률
A기업_2022	Y	501	201		0	0	0
B기업_2022	N	70,000	140		0	0	0
C기업_2022	N	410	70		0	0	0
D기업_2023	N	10	30		50	24	0
E기업_2023	Y	511,257	400		10	350	0.07
F기업_2023	N	649,821	151		4000	174,501	1.25

SET B
금결제원 채권거래 보유기업

(그림 1) 실험 데이터 셋

3.2 변수 구성

부도위험 예측을 위한 독립변수는 크게 세 가지 범주로 구성되며 <표 1>과 같다.

첫째, 재무 변수는 기업의 재무제표에서 도출된 8개 주요 항목으로, 부채비율, 총자산, 총부채, 자본금 등을 포함한다. 위 변수는 당해년도(t) 값뿐 아니라 전년도(t-1) 값까지 산출하여 기업의 재무 구조 변화를 시계열적으로 반영하였다. 예를 들어, 부채비율의 변화나 자본금 증감은 단일 시점의 수치만으로는 파악하기 어려운 추세를 드러낼 수 있으며, 이는 단기 부도위험의 중요한 신호로 작용할 수 있다. 이러한 시계열적 설계는 단순한 절대적 재무 상태뿐 아니라, 변화의 방향성과 속도를 고려하는 데 초점을 둔다.

둘째, 금융결제원 채권거래 변수는 당해년도(t)에 기록된 채권거래 금액과 건수로 구성된다. 채권거래

데이터는 기업 간 실질적인 지급·결제 흐름을 반영하기 때문에, 단순히 재무제표에서 드러나는 자산·부채 규모와는 다른 차원의 정보를 제공한다. 특히 거래 금액의 수준은 기업의 거래 규모와 신뢰도를, 거래 건수의 빈도는 거래망의 활발함을 간접적으로 나타낼 수 있다.

셋째, 파생 변수는 재무 변수와 채권거래 변수를 조합하여 생성된다. 예를 들어 자산 대비 채권거래 금액 비율, 차입금/매출채권 등 부채비율과 거래규모 간의 상호작용 항은 기업의 재무적 건전성과 거래 행태를 동시에 반영하는 복합 지표로 활용될 수 있다. 이러한 파생 변수는 서로 다른 데이터 출처에서 얻어진 정보를 결합함으로써, 개별 변수로는 드러나지 않는 상호작용 효과를 모델에 반영한다. 이는 단일 데이터셋의 한계를 보완하고, 기존 재무제표 정보를 보완할 수 있는 중요한 대안 정보원으로서 예측력을 제고하기 위한 전략적 접근이다.

<표 1> 데이터셋

범주	주요변수
재무변수	총자산, 자본금액, 영업이익, 부채비율 등
금융결제원 변수	연간 금융결제원 거래금액, 연간 금융결제원 거래건수, 연간 채권거래금액 변동성 등
파생변수	총자산 대비 금융결제원 채권거래금액 비중 금융결제원 매출 채권 대비 차입금 비중 등

4. 실험 결과

4.1 실험 설계 및 평가지표

본 연구에서는 관련 연구 검토를 통해 기업 부도 예측 분야에서 XGB, Random Forest, LGBM, Logistic 이 효과적인 것을 확인함에 따라 네 가지 모델을 사용하여 부도 예측 실험을 수행하고자 한다. 극단적 불균형(부도율 < 1%)을 고려하여, 각 알고리즘에는 학습·평가는 동일 폴드를 공유하는 Stratified k-fold 교차검증으로 수행하였다. 모델의 성능 평가는 부도와 같은 희귀 타겟에서 일반적으로 사용되는 AUC 만으로는 모델 차이를 과소평가할 수 있으므로 데이터 불균형 문제에 강건한 AUPRC(Area Under the Precision-Recall Curve)를 핵심 지표로 함께 사용하였다[5]. 이는 부도와 같이 드물게 발생하는 사건을 예측할 때, 모델이 소수 클래스(부도 기업)를 얼마나 잘 탐지하는지를 정밀도(Precision)와 재현율(Recall)의 조화를 통해 평가하므로 보다 실질적인 예측력을 나타낸다.

먼저 금융결제원 채권거래 변동성이 부도에 유의미한지 검토하기 위하여, 금융결제원 채권거래를 보유한 기업(SET B)을 대상으로 단변량 분석을 진행하였다. t-test 결과 부도군과 정상군 두 집단의 평균은 통계적

으로 유의하게 달랐다($p = 0.0008$). 이는 유의수준 5% 하에서 해당 변수가 부도 발생 여부와 연관된 차이를 보인다는 점을 시사한다.

이후 재무변수만으로 구성된 실험에서는 AUPRC 0.0896 의 성능을 보였으며, 금융결제원 채권거래 변수, 이와 조합한 파생변수를 활용한 실험에서는 AUPRC 0.0884 의 성능을 기록하였다. 표면적으로는 재무변수 단독 모형이 더 높은 ROC 를 나타냈으나, 주목할 점은 금융결제원 파생변수만으로 구성된 모형 또한 극히 낮은 부도율(약 0.8%)을 고려할 때 baseline AUPRC(0.008 수준)를 10 배 이상 상회했다는 것이다. 이는 금융결제원 데이터가 독립적 변수로서는 제한적일 수 있으나, 그 안에 부도와 관련된 잠재적 신호가 내재되어 있음을 의미한다.

다음 건설업 전체 기업(SET A)을 대상으로, 금융결제원 채권거래 정보의 추가적 설명력을 검증하기 위해 비교군을 다음 두 가지로 정의하였다.

- 재무변수 모형 : 재무지표 16 개만 투입
- 확장 모형 : 재무지표 16 개 + 금융결제원 원시변수 및 도메인 지식을 반영한 파생변수를 결합한 모형

4.2 분석 결과

<표 2>를 보면 네 알고리즘 모두에서 확장 모형은 ROC 에서는 혼합적인 변화를 보였으나, AUPRC 에서 일관된 개선을 확인하였다. ROC 변화가 제한적이었던 것은 다수 클래스 영역에서의 구분력에는 영향이 크지 않지만, 타겟(부도) 순위 상위권의 질적 개선이 발생했기 때문으로 해석된다. 특히 알고리즘별로 XGB 는 ROC 와 AUPRC 에서 모두 성능향상을 보였으며, Random Forest (RF)는 재무변수만 사용한 경우 AUPRC 가 0.0451 에 그쳤으나, 금융결제원 변수와 파생변수를 추가한 경우 0.062 로 가장 큰 폭으로 향상되었다. 이는 절대적으로는 +0.0169 포인트 개선이며, 상대적으로는 약 37% 증가한 수치이다. 부도율이 0.08% 수준(무작위 분류기의 baseline AUPRC ≈ 0.0008)에 불과한 상황을 고려하면, 금융결제원 파생변수를 추가함으로써 baseline 대비 약 77 배 수준으로 성능이 향상된 것이다.

이러한 결과는 금융결제원 데이터가 타겟 클래스 탐지가 핵심인 AUPRC 지표에서는 실질적인 개선 효과를 제공했음을 의미한다. 따라서 부도율이 극히 낮은 환경에서 부도기업을 상위 위험군으로 선별하는 데 있어 보완적 신호로 기능한다고 해석할 수 있다.

<표 2> 재무변수 및 금융결제원 변수 추가시 성능비교

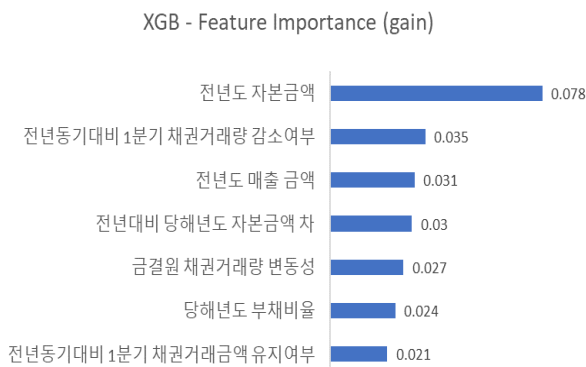
알고리즘	재무변수 모형		확장 모형	
	ROC	AUPRC	ROC	AUPRC

XGB	0.7581	0.0402	0.7666	0.0425
RF	0.7617	0.0451	0.766	0.062
LGBM	0.7515	0.0325	0.7509	0.0372
LOGIT	0.6962	0.0297	0.6752	0.0414

전통적 재무정보는 공시 주기의 장기성(연·분기)에 기인하여 예측의 적시성이 제한되며, 거시·산업 요인에 대한 통제가 구조적으로 어렵다는 한계를 가진다. 이러한 제약을 보완하기 위하여 본 연구는 금융결제원 기반의 채권거래 변수를 결합하였다. 초기 기준모형(로지스틱 회귀 및 일부 트리 계열의 기본 설정)에서는 금결원 변수 도입이 예측 성능을 제한적으로만 개선하였고, 변수중요도 또한 상대적으로 낮게 관찰되었다.

그러나 알고리즘 및 중요도 산정 방식의 이질성을 고려하여 재평가한 결과, (그림 2)와 같이 금결원 변수 다수가 상위권에 배치되는 경향을 확인하였다. 특히 분기별 거래건수·거래금액 수준 및 증감률, 채권거래 변동성 등 동태적 거래 신호는 여러 모형에서 유의미한 기여를 보였다. 이는 금결원 변수가 기업의 단기 유동성 압력과 현금흐름 민감도를 포착함으로써 재무정보로 설명되지 않는 잔여 위험을 보완하는 국소적·상호작용적 신호로 작동한 것으로 해석된다.

아울러 금결원 자료는 표본 커버리지와 관측기간의 제약을 내포하고 있어, 변수중요도는 데이터 범위, 알고리즘 선택, 불균형 데이터 평가척도(AUPRC 등)에 민감하게 변동될 수 있으나 조기경보 지표로서 실질적 효용을 갖는다. 재무제표만으로는 포착이 어려운 미세한 신용상태 변화를 선행적으로 드러내어, 공시, 시장정보가 빈약한 비상장 중소기업 집단에서 특히 높은 실무적 가치를 제공할 것으로 기대된다. 종합하면, 전통 재무정보에 금결원 변수를 결합하는 접근은 예측력의 가시적 제고 가능성을 보여줄 뿐 아니라, 정책금융기관의 부실 선제관리 및 자금배분 효율성 제고에 기여할 수 있는 실증적 근거를 제공한다.



(그림 2) 변수 중요도 (feature importance)

5. 결론 및 향후 연구계획

본 연구는 금융결제원 채권거래 데이터와 재무제표 데이터를 통합적으로 활용하여 기업의 동태적 자금흐름과 재무 건전성을 동시에 고려하는 새로운 접근 방식을 모색하였다. 향후 연구에서는 본 연구의 성과와 한계점을 바탕으로 기업 부도 예측 모델의 정확도와 실용성을 더욱 제고하기 위해 두 가지 방향으로 연구를 확장하고자 한다.

첫째, 다양한 비재무 데이터 및 비정형 데이터를 활용하여 부도 예측력을 강화하고자 한다. 뉴스 텍스트 정보, 증권/투자 관련 게시판, 기업 홈페이지 등으로 확대하여 예측력을 개선할 수 있을 것으로 기대된다.

둘째, 채권거래 데이터를 바탕으로 기업 간 거래관계를 모델링하여 연쇄 부도 위험을 측정하는 연구로 발전해 나가고자 한다. 이는 개별 기업의 부도를 넘어, 거래 네트워크 내에서 부실의 파급 효과를 분석하는 데 중요한 기초 자료가 될 수 있다. 이를 통해 특정 기업의 부실이 거래 관계를 통해 다른 기업들에게 어떻게 전파되는지, 그리고 어떤 기업이나 산업이 연쇄 부도에 더 취약한지를 심층적으로 분석할 수 있다.

참고문헌

- [1] 오세경, 최정원, 장재원, 빅데이터를 이용한 딥러닝 기반의 기업 부도예측 연구. 한국금융연구원, (2017)
- [2] 박중현, "딥러닝을 이용한 재무와 비재무 정보 기반 기업부도 예측 분석에 관한 연구", 멀티미디어 학회 논문지, 제 26 권, 제 8 호. (2023)
- [3] 이준학, 정다연, 노승민, 문지훈, "GAN 과 LGBM 을 이용한 불균형 데이터 셋 문제 해결 및 신용카드 사기거래 탐지," 금융보안원 학술발표논문집, (2021).
- [4] 허선우, 백동현, "데이터 불균형을 고려한 설명 가능한 인공지능 기반 기업부도예측 방법론 연구," 한국산업경영시스템학회지, 제 45 권, 제 2 호. (2022)
- [5] J. Miao and W. Zhu, "Precision-Recall Curve (PRC) Classification Trees," Preprint, Stony Brook University, Nov. 2020.