

논문 리뷰

딥러닝 기반 부실기업 예측 모형에 관한 연구

목차

1. 팀 소개

2. 논문 소개

3. 선행 연구

4. 데이터 소개 및 전처리

5. 분석 방법

6. 결과 및 검증

7. 기여효과

8. 한계점

팀 소개

- 팀 이름 : FBI (Financial Bankruptcy Investigation)

- 팀원 소개:

서수아 - 팀장❤️

서선우

김정우

고원태

논문 소개

- 딥러닝 기반 부실 기업 예측 모형에 관한 연구 (2021)

저자 : 조재혁, 안은주, 김성수

목적 : 재무제표 내 재무 비율 이용, 머신 러닝과 딥러닝 기반 한계 기업 예측

Keyword : 기업 부실화, 딥러닝 기법, 머신 러닝, 앙상블 모형, 한계 기업 예측

논문 연구 배경

- 부실기업 예측은 회계와 재무 분야에서 중요하게 다루어져 온 연구 주제다. 그러나 부도기업은 사실상 사업활동을 중단한 기업으로 계속기업 간에 어떠한 기업이 부실 징후를 보이는지를 판단하는 기준으로 부적합하다는 한계점이 존재한다. 따라서 부도기업이 아닌 한계기업을 예측대상으로 선정하였으며 한계기업 예측을 위한 방법으로 머신 러닝 기법과 딥러닝 기법을 활용하였다.
 - 한계기업이란? (이기영, 우석진, 2015)
 - 최종 3개년 연속 총차입금이 매출액을 초과하는 기업
 - 최종 2개년 연속 자기자본 전액 잠식 기업
- ♥ 3개년도 연속 이자보상비율이 1 미만인 기업 (한국은행 기준이며 저자가 연구에 사용)

논문 선정 이유

- 딥러닝 기법의 사용 사례와 최적 파라미터 설계 요령
- 딥러닝 모델을 활용한 한계 기업 분류 방법 탐색
- 피쳐 선정에서의 인사이트 획득

선행 연구

저자	연구내용	기법
이기영, 우석진(2015)	공적 신용보증을 받은 한계기업(3개년 연속 이자보상비율 1미만)의 생존패턴을 분석 / 한계기업의 수명이 37.9% 더 짧음	생존분석 (사건의 발생 여부, 사건이 발생하기까지 소요 시간 분석)
Odom and Sharda(1990)	전통적 기법인 다중판별분석과 인공신경망의 기업부실예측 비교 / ANN 우세	ANN
Shin and Lee(2005)	서포트벡터머신 활용한 기업 부도 예측 / 딥러닝 모델에 비해 SVM의 예측력 우세	SVM
김승혁, 김종우(2007)	배깅 구성 모델 중 예측 상위 3개를 voting해서 SOHO(small office, home office)부도예측 / 수정 배깅 모델이 일반적 배깅 모델보다 예측력 우수	Bagging
차성재, 강정석(2018)	시간의 흐름에 따라 영향 있는 시계열 데이터 분석 통한 부도 예측 / ROC AUC 기반, 표본 내 우수 모형은 RF, RNN, LSTM 딥러닝 시계열 모델의 유용성 확인	선형판별분석, LR, DT, RF, KNN, SVM, MLP, RNN, LSTM
박종원, 안성만(2014)	외부감사대상 기업 대상으로 재무비율을 이용, 부도 예측 / 총자본회전률, 금융비용대비부채비율, 유보이익대총자산비율, 현금흐름대총자본비율, 총자산순이익률이 부도확률에 많은 영향 미치는 변수들이며 산업은 건설업과 제조업이 부도확률이 높은것을 확인 / ROC AUC (85%) 기반 모델이 변별력있는 예측력 (82%)을 보이며 극단치에도 영향을 적게 받아 모델의 강건성 입증	다항 로지스틱 회귀

데이터 소개

- KISVALUE에서 제공하는 기업 재무 데이터로 KOSPI, KOSDAQ, KONEX 상장 기업 및 외부감사 대상 33,916개 기업 데이터
 - I. 2016년 이후 설립 및 적정 감사의견 외 기업을 제외한 비금융업 업종의 12월 결산 법인
 - II. 금융업은 재무 기준 달라 제외
 - III. 결산월의 차이에 따른 영향이 데이터에 왜곡을 일으키는 것을 방지하기 위해 12월 결산 법인으로 통일
 - IV. 최종 선택된 데이터
- 2017 ~ 2019 KOSPI, KOSDAQ, KONEX 상장 기업 및 외부감사 대상 기업
 - > 16,813개 기업의 재무정보

데이터 전처리

〈표 1〉 수집 데이터 및 전처리 과정

구 분	설 명
분석대상	<ul style="list-style-type: none"> • 2017~2019년 KOSPI, KOSDAQ, KONEX 상장기업 및 외부감사 대상기업 16,813개 기업의 재무정보 • 16년 이후 설립 및 적정 감사의견 외 기업을 제외한 비금융업 업종의 12월 결산법인
데이터 전처리	<ul style="list-style-type: none"> • 결측치 처리 : 기업정보 결측치에 대해 중앙값(Median) 대체 적용 ※ 영업이익이자보상비율에서 결측치가 존재하는 경우에는 해당 기업 데이터 제거 • 이상치 처리 : 윈저라이징(Winsorizing)을 통해 0.01, 0.99에 해당하는 값으로 변환
표본	<ul style="list-style-type: none"> • 한계기업 : 2017~2019년 영업이익이자보상비율이 1 미만인 기업 • 정상기업 : 한계기업 이외의 기업 <p>♥ 2019년을 기준으로 3개년도의 영업이익이자보상비율이 1 미만인지 여부 판단</p>
최종 분석 대상기업	<ul style="list-style-type: none"> • 정상기업 : 13,818개 • 한계기업 : 2,995개

피쳐 선정

- 16,813개 기업의 재무비율들을 독립 변수로 활용

Q. 재무비율을 피쳐로 선정한 이유?

- A.
1. 이미 부실기업 예측 연구에 다수 활용 (이인로, 김동철, 2015)
 2. 외부 감사를 받은 기업의 정보의 신뢰성 높고 모형 구축의 용이성과 비용 절감 (박종원, 안성만, 2014)
 3. 재무비율은 상장 기업과 비 상장 기업 가치 평가에 공통적으로 사용할 수 있음 (김선배 등, 2016)

피쳐 선정

- 이인로, 김동철(2015) - 43개, 박종원, 안성만(2014) - 117개, 김선배 등(2016) - 11개
 - > 22개의 재무비율 피쳐 선정 (T-test(Prob 0.05보다 큰 변수 제외 - 보편성 떨어짐))
 - 연도 별로 서로 다른 11~2개 설명력을 가지는 변수 확인 (LR의 후진선택법)
- 후진선택법 : 모든 독립변수를 포함한 모형에서 출발, 가장 적은 영향을 주는 변수부터 하나씩 제거하면서
 - 더 이상 제거할 변수가 없을 때 까지 진행. 변수의 개수가 많으면 사용이 어려우나
 - 전체 변수들의 정보를 이용할 수 있는 장점이 있음

최종 선정 피쳐

〈표 2〉 선정된 후보 재무비율

구분	분류	명칭	산식
1	성장성	총자산증가율	$(\text{기말총자산} - \text{기초총자산}) / \text{기초총자산} \times 100$
2		유동자산증가율	$(\text{기말유동자산} - \text{기초유동자산}) / \text{기초유동자산} \times 100$
3		매출액증가율	$(\text{당기매출액} - \text{전기매출액}) / \text{전기매출액} \times 100$
4		순이익증가율	$(\text{당기순이익} - \text{전기순이익}) / \text{전기순이익} \times 100$
5		영업이익증가율	$(\text{당기영업이익} - \text{전기영업이익}) / \text{전기영업이익} \times 100$
6	수익성	매출액순이익률	$\text{순이익} / \text{매출액} \times 100$
7		매출총이익률	$\text{매출총이익} / \text{매출액} \times 100$
8		자기자본순이익률	$\text{순이익} / \text{자기자본} \times 100$
9	활동성	매출채권회전율	$\text{매출액} / \text{매출채권}$
10		재고자산회전율	$\text{매출원가} / \text{재고자산}$
11		총자본회전율	$\text{매출액} / \text{총자본}$
12		유형자산회전율	$\text{매출액} / \text{총자산}$
13		매출액대매출원가	$\text{매출원가} / \text{매출액} \times 100$
14		매출액대판매관리비	$\text{판매관리비} / \text{매출액} \times 100$
15	안정성	부채비율	$\text{부채} / \text{자기자본} \times 100$
16		유동비율	$\text{유동자산} / \text{유동부채} \times 100$
17		자기자본비율	$\text{자기자본} / \text{총자산} \times 100$
18		당좌비율	$\text{당좌자산} / \text{유동부채} \times 100$
19		고정비율	$\text{고정자산} / \text{총자본} \times 100$
20		순운전자본비율	$\text{순운전자본} / \text{총자본} \times 100$
21		차입금의존도	$(\text{장기 및 단기차입금} + \text{사채}) / \text{총자본} \times 100$
22		현금비율	$\text{현금예금} / \text{유동부채} \times 100$

피처 선정

3개년에 모두 선정된 피처 : 순이익증가율, 자기자본순이익률, 고정비율

〈표 3〉 최종 선정된 연도별 재무비율

구분	2017년	2018년	2019년
1	총자산증가율	총자산증가율	총자산증가율
2	유동자산증가율*	유동자산증가율	유동자산증가율*
3	매출액증가율	매출액증가율*	매출액증가율*
4	순이익증가율*	순이익증가율*	순이익증가율*
5	영업이익증가율*	영업이익증가율*	영업이익증가율
6	매출액순이익률	매출액순이익률*	매출액순이익률
7	매출총이익률*	매출총이익률	매출총이익률
8	자기자본순이익률*	자기자본순이익률*	자기자본순이익률*
9	매출채권회전율	매출채권회전율	매출채권회전율
10	재고자산회전율	재고자산회전율	재고자산회전율
11	총자본회전율*	총자본회전율	총자본회전율*
12	유형자산회전율	유형자산회전율	유형자산회전율*
13	매출액대매출원가	매출액대매출원가*	매출액대매출원가
14	매출액대판매관리비*	매출액대판매관리비*	매출액대판매관리비
15	부채비율	부채비율*	부채비율
16	유동비율	유동비율*	유동비율*
17	자기자본비율	자기자본비율*	자기자본비율*
18	당좌비율*	당좌비율*	당좌비율
19	고정비율*	고정비율*	고정비율*
20	순운전자본비율	순운전자본비율	순운전자본비율*
21	차입금의존도*	차입금의존도	차입금의존도*
22	현금비율*	현금비율	현금비율

* 연도별 최종 선정된 변수

분석 방법

- 머신 러닝 앙상블 모형(Random Forest, SVM, KNN), 딥러닝 기법(RNN-LSTM, RNN-GRU, CNN)

- ✓ RF : 여러가지 의사결정나무들을 학습시킨 후, 분류 결과들을 취합해서 최종 분류. 배깅 방식 데이터 임의 분할 (중복 허용) .
- ✓ SVM : 분류를 위한 기준 선(결정 경계)을 정하는 모델
- ✓ KNN : 데이터로부터 거리가 가까운 K개의 데이터를 특성에 따라 분류해주는 모델
- ✓ RNN : 순환적으로 연결된 신경망으로 시간이나 공간의 흐름에 따른 데이터의 변동 양상을 연속적으로 계산하여 예측
- ✓ RNN-LSTM : 기울기 소실 방지 위해 어떠한 데이터를 버릴지 결정하는 망각게이트, 어떠한 데이터를 갱신할지를 결정하는 입력게이트, 어떠한 값을 출력할지를 결정하는 출력게이트로 구성되어 있으며 이들 3가지 게이트를 통해 최종 결과 값을 산출
- ✓ RNN-GRU : 이전 결과값을 얼마나 기억할지를 결정하는 갱신게이트, 이전 결과값과 새로운 입력 데이터를 결합하는 리셋게이트로 구성
- ✓ CNN : 기존 DNN이 데이터를 입력 받는 과정에서 flatten 절차를 거치며 데이터의 변형/손실 발생 가능함.
데이터의 특징을 추출 하는 필터(Filter)의 합성곱(Convolution)을 이용하여 데이터를 그대로 입력 받으며 이 부분 해결.
입력층과 출력층 사이에 1개 이상의 합성곱층 (Convolution Layer)과 풀링층(Pooling)층이 있다.
합성곱층은 입력된 데이터의 합성곱 연산을 실시하고 필터를 통해 데이터의 특징이 추출하며,
풀링층은 합성곱층의 결과값을 단순화하고 크기를 줄이는 역할을 수행

데이터 분리 및 정규화, 모델 별 하이퍼 파라미터 설정

- Train set과 Test set을 데이터 전체에서 무작위로 7:3 비율로 나눔
- 데이터 정규화 진행

모델	하이퍼 파라미터
RF	n_estimator : 50, max_depth : 5, min_samples_split : 10
SVM	n_estimator : 50, C : 2, Kernel : Linear Kernel
KNN	n_estimator : 50, distance : Euclidean distance, K : 30
RNN-LSTM	Hidden layers : 4, number of cells : 128, epoch : 200, Activation function : Relu, earlystopping, Loss function and Optimizer : Binary Crossentropy, Adam
RNN-GRU	Hidden layers : 4, number of cells : 128, epoch : 200, Activation function : Relu, earlystopping, Loss function and Optimizer : Binary Crossentropy, Adam
CNN	Hidden layer : 4, number of filters : 128, Maxpooling, Dropout : 0.25, Activation function : Relu, Loss function and Optimizer : Binary Crossentropy, Adam

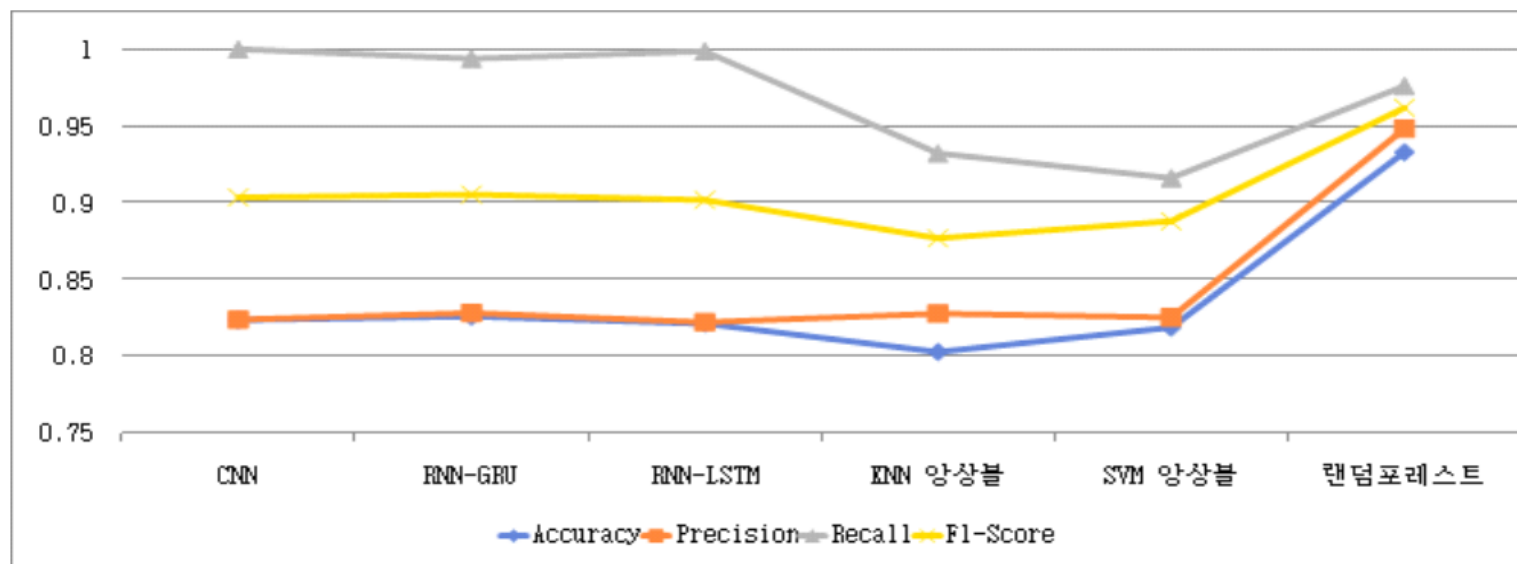
실증 결과 및 검증

- 모델 평가지표로 정확도, 정밀도, 재현율, F1 점수 사용
- 랜덤 포레스트 모형이 재현율 제외 가장 우수한 예측 성능
- 재현율에서는 딥러닝 모형인 RNN-LSTM, RNN-GRU, CNN 의 성과가 우세
- F1 스코어 관점으로는 랜덤 포레스트 제외한 ML Ensemble 모형들이 딥러닝 모형들에 우세
- 한계 기업 예측에서는 재현율이 중요 ($\text{True Positive} / (\text{TP} + \text{FP})$)
 - 1종오류 : 귀무가설이 옳은데 기각하는 경우 (실제 정상기업을 부실기업으로 예측하는 오류)
 - 2종오류 : 귀무가설이 틀린데 채택하는 경우 (실제 부실기업을 정상기업으로 예측하는 오류)
 - 2종 오류에 집중 하였으며 부실 기업을 정상 기업으로 예측할 경우 발생할 수 있는 부정적 결과를 방지 할 수 있다.

실증 결과 및 검증

〈표 6〉 모형별 정확도, 정밀도, 재현율, F1-점수

모형 \ 평가척도	정확도 (Accuracy)	정밀도 (Precision)	♥ 재현율 (Recall) ♥	F1-점수 (F1-Score)
Random Forest	0.9328	0.9481	0.9760	0.9618
SVM Ensemble	0.8186	0.8252	0.9610	0.8879
KNN Ensemble	0.8027	0.8277	0.9321	0.8768
RNN-LSTM	0.8212	0.8221	0.9986	0.9017
RNN-GRU	0.8259	0.8281	0.9938	0.9053
CNN	0.8236	0.8238	1.0000	0.9034



〈그림 2〉 모형별 정확도, 정밀도, 재현율, F1-점수

기여 효과

- 부실 기업 예측으로 부도 기업이 아닌, 한계 기업을 예측했다.
- 다양성 확보 측면으로 딥러닝 모형 연구를 수행했다.
- 재현율 측면으로 머신러닝 모형들보다 딥러닝 모형의 성과가 더 좋았다는걸 확인함으로 향후 관련 연구나 분류에서 딥러닝 모형들의 역할 확대 가능성 제시


한계점 (저자)

- ✓ 재무 비율만을 사용했다. 기업의 대외적 요인과 환경 요인을 고려하지 않았다.
- ✓ 회계 정보만을 활용했다. 즉, 적시성과 미래지향성 등이 부족했다.
(연간 재무데이터 공시되기 까지의 시간적 공백)
- ✓ 일정 기업 규모 이상의 외부 감사 대상기업 및 상장 기업들 만이 대상이었다. 기업 부실화는 중소기업이 더 비중이 높을 텐데, 분석대상에 포함되지 않아 결과의 편향성 가능하다.

한계점 (발표자)

- ✓ 정규화 과정이 드러나지 않았음
- ✓ 상관 계수 없음
- ✓ 기초 통계량 명시되지 않음
- ✓ 불확실한 용어
 - stepwise logistic regression's backward selection
 - OR logistic regression's backward elimination

한계점 (발표자)

표본	<ul style="list-style-type: none">• 한계기업 : 2017~2019년 영업이익이자보상비율이 1 미만인 기업• 정상기업 : 한계기업 이외의 기업  2019년을 기준으로 3개년도의 영업이익이자보상비율이 1 미만인지 여부 판단
----	--

- 최근 3개년 연속 이자보상비율이 1미만일 때 한계기업으로 판명
- 한계 기업 판명 시점은 2019년, 그 설정을 17, 18년 두 해에 모두 적용
- 이 경우, 실제 한계기업으로 2019년에 판명되나, 17년과 18년에 이미 한계 기업 레이블을 적용,
-> 모델에게 학습 시킴
- 전체 데이터를 무작위로 Train set과 Test set으로 분리
- 데이터 누수의 문제점