기계학습을 이용한 저축은행 부실 예측모형 검증

이경수*, 임희석**
*고려대학교 컴퓨터정보통신대학원 빅데이터융합학과
**고려대학교 컴퓨터정보통신대학원 디지털금융공학과 주임교수
e-mail : rudtn82@korea.ac.kr

Verification of insolvency prediction model for savings banks using machine learning

Kyoung-Soo Lee*, Heui-seok Lim**
*Dept of Graduate school of computer & information technology, Korea University
**Professor, Dept of Graduate school of computer & information technology, Korea University

요 약

본 연구의 목적은 저축은행 부실에 영향을 미치는 주요 변수를 선정하고, 기존 전통적인 통계기법에 국한된 국내 부실 예측 연구를 벗어나 기계학습을 활용하여 부실 예측모형에 대한 성능을 향상시키는 것이다. 이를 위해 본 연구는 2010년부터 2014년까지의 부실저축은행 29개사와 건전 저축은행 88개사의 재무정보 1,506개 분기자료를 기반으로 로지스틱회귀분석 뿐만 아니라, ANN, SVM 및 Decision Tree와 같은 알고리즘을 이용하여 보다 정교한 부실 예측 모형을 개발하고 활용함으로써 금융기관에 대한 리스크 상시 감시를 통해 부실을 사전에 예방하고 시장의 안정화 및 금융질서를 유지함을 목적으로 하고 있다.

1. 서론

서민에 대한 금융 편의 제공 목적으로 설립된 상호저축은행은 1997년 외환위기, 2000년대 초반 신용카드 위기를 거치며 상당수가 구조조정 되었다. 외환위기 직후 231개사에 달하였던 저축은행 수는 2010년 말 105개로 크게 감소하였음에도 불구하고 높은 금리의 예금 수신 및 거액여신을 통한 대출 확대로 저축은행의 자산은 빠르게 증가하였다. 그러나 이러한 외형 증가에도 불구하고 글로벌 금융위기에 따른 경기침체 등으로 부동산시장이 냉각되면서 부동산 Project Finance 대출의 부실이 급증하여 저축은행 건전성이 크게 악화되었다. 이에 따라 정부는 2011년 저축은행 부실에 따른 금융시장의 부정적인 영향을 차단하기 위하여 예금보험공사 내에 저축은행 구조조정 특별계정을 신설하고 2011년부터 2015년까지 총 30개의 저축은행을 제3자 매각, 계약이전 등의 형태로 구조 조정함으로써 예금자를 보호하고 금융시장을 안정시켰으며 2016년 말까지 27조원의 막대한 공적자금이 투입되었다.[1]

그 결과 2017년 12월말 현재 79개 저축은행의 당기순이익이 1조 674억원으로 흑자 전환하였고 BIS기준 자기자본비율도 14.31%로 대폭 개선되었지만[2], 현재까지도 자본잠식된 저축은행이 남아있어 추가 자본 증자 및 수익성이 개선되지 않을 경우 영업정지가 불가피할 것이라는 우려도 있다. 이에 따라 금융 당국은 저축은행에 대하여 부실 예측모형 등의 결과를 이용하여 공동검사, 단독조사 등 계획을 수립하고 점검을 실시함으로써 부실을 사전에 예방하고 적기 조치하고 있으며, 이러한 리스크 상시 감시 지표로 사용되는 모형에 대한 높은 신뢰도와 부실 분류 예측 정확도가 요구되고 있다.

본 연구 목적은 최근 구조조정 과정에서 퇴출된 저축은행과 생존하고 있는 저축은행의 재무자료 등을 이용하여 부실 발생에 영향을 미치는 주요 변수를 선정하고, 다양한 분석 기법을 통하여 분류 예측 정확도가 높은 부실 예측 모형을 구축함으로써 신뢰수준을 높여 실제 리스크 상시 감시에 활용하는데 그 목적이 있다.

2. 이론적 배경

2.1 부실 금융기관의 정의

본 연구에서 부실 금융기관의 정의는 예금자보호법 제2조제5호 및 금융산업의 구조개선에 관한 법률 제2조제2호에 따라 경영상태 실사 결과 부채가 자산을 초과하는 부보금융기관 또는 거액의 금융사고나 부실채권의 발생으로 부채가 자산을 초과하여 정상적인 경영이 어렵게 될 것이 명백한 부보금융기관으로서 금융위원회나 예금보험위원회가 결정한 부보금융기관 등을 의미한다.

2.2 관련 연구

Martin[3]은 1970~1976년 기간 중 도산 등 경영개선 조치를 받은 23개 은행 및 5,575개 건전은행에 대하여 분석하였으며, 자본적정성, 수익성, 자산 리스크, 유동성 부문 등 총 25개의 재무비율을 설명변수로 활용하였고 로지스틱 회귀분석 및 판별분석을 활용하였다. 주요 연구 결과는 로짓모형이 판별분석모형보다 예측력이 약간 우수하게 나타났다.

Jagtiani, Kolari, Lemieux and Shin[4]은 1988~1990년 기간 중 미국은행 177개를 대상으로 추정표본을 구성하였

으며, 테스트 표본은 동 기간 중 은행 499개를 대상으로 단순 로지스틱, 순위 로지스틱, 특성인식모형으로 분석하였다. 설명 변수는 자기자본비율 등 42개 변수를 활용하였으며 주요 연구 결과는 여러 분석모형 중 단순 로지스틱 회귀모형이 제일 우수하게 나타났다.

Boyacioglu, Kara, Raykan[5]는 1997~2004년 중 터키 예금보험기금으로 이전된 21개의 은행을 부실 집단, 정상 집단은 동 기간 내 정상 영업 44개 은행으로 구성하였다. 설명변수는 CAMELS부문 관련 20개 재무지표를 이용하고, 다변량 통계분석 및 인공지능 분석을 사용하였다. 주요 연구 결과는 신경망분석 및 SVM 예측력이 가장 높았다.

3. 연구 방법

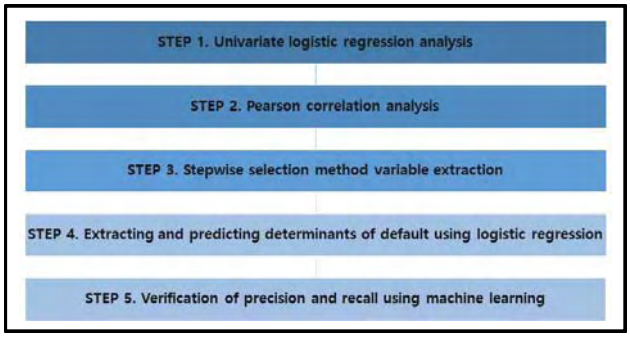
3.1 연구의 대상 및 자료

본 연구에서의 자료는 저축은행이 분기별 금융감독원 및 예금보험공사에 보고하는 재무보고서의 자본적정성, 자산건전성, 수익성, 유동성 및 예금동향보고서의 예금 현황 등 주요변수를 선정하여 총 12개의 후보변수로 구성하였으며 표본 수는 1,506개로 선정하였다.

3.2 연구 특징 및 단계

본 연구는 국내 상호저축은행을 대상으로 기계학습 등을 이용하여 부실 결정요인을 결정하고 분류 정확도를 측정하였으며, 이를 위하여 종속변수와 후보변수간의 일변량 로지스틱 회귀분석을 통해 1차 후보변수를 추출하고, 피어슨 상관계수를 이용하여 변수들 간 연관성이 높은 변수는 제외하였다. 마지막으로 단계적 선택방법을 통하여 최종 변수를 선정하고 Logistic regression analysis, Artificial Neural Network, Support Vector Machine 및 Decision Tree 분석 기법을 통하여 정확도를 검증하였다.

연구단계는 (그림 1)과 같으며, 분석도구는 IBM SPSS Statistics 23과 R3.3.3을 사용하였다.



(그림 1) Stage of research

3.3 종속변수와 설명변수

본 연구의 자료는 2010년 6월말부터 2014년 6월말까지 영업하였던 117개 저축은행이 금융감독원에 제공하는 1,506개 분기별 경영관리재무정보를 이용하였다. 종속변수는 해당 기간에 부실이 발생한 저축은행 29개사의 부실 발생시점(T) 시점이 아니라 저축은행이 부실 발생시점(T) 직전 분기를 부실 예측시점(T-1)으로 하였으며 부실이 발생하면 1, 정상인 경우에는 0으로 설정하였다. 설명변수는 저축은행의 재무정보인 CAEL부문에서 각 부문별 주요지표 2~3개를 선정하였으며 선정된 지표는 BIS자기자본비율(x1), 단순자기자본비율(x2), BIS기준 기본자본비율(x3), 고정이하 여신비율(x4), 연체대출비율(x5), Coverage Ratio(x6), PF대출비중 증감률_전반기대비(x7), 총자산영업이익률(x8), 총자산순이익률(x9), 수지비율(x10), 유동성비율(x11), 예금비중 증감률_전분기대비(x12)이며, 총 12개로 구성하였다.

<표 1> Variable Statistic

Name	N	Min	Max	Avg	S.D
x1	1,506	-41.45	138.64	13.1850	13.9924
x2		-94.84	84.55	8.3029	11.6143
x3		-134.88	136.98	9.8665	17.1875
x4		.98	79.60	17.5267	12.3856
x5		.76	69.11	19.4491	11.4778
x6		23.12	621.75	57.7193	33.4594
x7		-100.00	544.41	-1.9247	46.7566
x8		-42.88	32.79	-1.1476	4.0999
x9		-42.79	25.53	-1.3711	4.1845
x10		-702.47	1406.82	122.7892	80.5320
x11		66.71	2029.02	434.4270	238.0866
x12		-55.22	376.92	-.5530	15.3881

4. 연구 결과

4.1 저축은행 부실 결정요인 변수 추출

본 연구에서 저축은행 부실 결정요인을 추출하기 위한 데이터, 즉 추정용 데이터는 2010년 6월부터 2013년 3월까지 12개 분기 1,090건의 자료를 이용하였으며, 추후 검증용 데이터는 2013년 6월부터 2014년 6월까지 5개 분기데이터 416건을 이용하였다.

4.1.1 일변량 로지스틱 회귀 분석

일변량 로지스틱 회귀분석을 통해 설명변수 1개와 종속변수 간 P-value를 구하여 유의수준을 확인하는 절차 및 선행 연구 등을 통한 변수의 이론부호와 설명변수의 방향성 일치여부를 통해 변수를 선정하였으며, 결과는 <표 2>와 같다.

<표 2> Result of univariate logistic regression analysis

Name	Direction	β	P-value	Name	Direction	β	P-value
x1	-	-.100	.000***	x7	+	.007	.005***
x2	-	-.081	.000***	x8	-	-.141	.000***
x3	-	-.047	.000***	x9	-	-.139	.000***
x4	+	.048	.001***	x10	-	.003	.005***
x5	+	.071	.000***	x11	+	.001	.018**
x6	-	.001	.909	x12	+	-.033	.176

cf) *** is significant at 0.01 level, ** is 0.05 level

4.1.2 피어슨 상관분석

일변량 로지스틱 회귀분석 결과 이론부호와의 방향 및

P-value가 유의하지 않은 x6, x10, x12를 제외한 나머지 변수에 대하여 피어슨 상관분석을 통해 변수 간 상관성이 높은 변수를 제거하여 다중공선성이 발생하지 않도록 하였다. 분석결과는 <Table 3>과 같으며 중요도를 감안하여 변수 간 상관계수가 0.8을 초과하는 강한 상관관계를 가지는 x1, x2, x3 중에서 x2, x3을 제거하였고, x4, x5 중에서는 x4를, x8과 x9 중에서는 x8을 제거하였다.

<표 3> Result of Pearson correlation analysis

Name	x1	x2	x3	x4	x5	x7	x8	x9
x2	.829**	1	.951**	-.301**	-.249**	-.045	.375**	.311**
x3	.806**	.951**	1	-.301**	-.248**	-.066*	.364**	.310**
x4	-.279**	-.301**	-.301**	1	.874**	.065*	-.639**	-.663**
x5	-.253**	-.249**	-.248**	.874**	1	.080**	-.539**	-.549**
x7	.001	-.045	-.066*	.065*	.080**	1	-.040	-.036
x8	.312**	.375**	.364**	-.639**	-.539**	-.040	1	.965**
x9	.268**	.311**	.310**	-.663**	-.549**	-.036	.965**	1
x11	-.238**	-.147**	-.154**	-.143**	-.099**	.025	-.036	-.008

cf) *** is significant at 0.01 level, ** is 0.05 level

4.1.3 단계적 변수 선택법

선정된 변수 x1, x5, x7, x9를 단계적 변수 선택법을 이용하여 최적의 변수를 선택한 결과 변수 4개를 모두 선정할 경우 설명력을 나타내는 R^2가 22.4% 수준으로 가장 높게 나타났다.

<표 4> Stepwise Variable Selection Method Analysis

Name	β	S.E	P-value	R^2
x1	-.055	.023	.017**	0.224
x5	.038	.017	.025**	
x7	.008	.003	.009***	
x9	-.087	.044	.050**	
Constant	-4.518	.577	.000***	

cf) *** is significant at 0.01 level, ** is 0.05 level

4.2 저축은행 부실 예측 분류 정확도 분석

본 연구에서는 선정된 부실 결정 요인 변수를 이용하여 Logistic regression analysis, Artificial Neural Network, Support Vector Machine 및 Decision tree 알고리즘 분석 결과를 통해 예측 정확도를 비교 및 분석하여 정확도가 가장 높은 알고리즘을 확인하였다. 이에 대한 테스트 데이터는 2013년 6월부터 2014년 6월까지 5개 분기 저축은행 재무정보이며 테스트 데이터의 비율을 전체의 28% 수준으로 설정하였으며 정상 413개, 부실 3개로 구성하여 연구를 진행하였다.

4.2.1 Logistic regression analysis

Logistic regression analysis는 종속변수가 Binary 형태의 자료에 주로 사용되는 분석방법으로 독립변수들이 정규분포여야 한다는 가정이 전제될 필요가 없고, 결과의 확률 예측치 구간(0~1)도 이탈하지 않는 점 때문에 대부분의 연구들이 선택하고 있다.[6]

이 알고리즘을 통해 1,090개 데이터를 훈련시키고 선정된 주요 변수를 산식으로 나타내면 다음과 같다.

<산식 1> Verification formula for logistic regression analysis

$$P = \frac{1}{1 + e^{-(-4.518 - 0.055X_1 + 0.038X_5 + 0.008X_7 - 0.087X_9)}}$$

이 산식을 이용하여 416개의 테스트 데이터에 대한 결과값을 분석해보면 <표 5>와 같다. 분석 결과 정상 분류의 정밀도(Precision)는 99.5%이며, 정상 분류 재현율(Recall)은 99.8%이다. 또한 부실 분류 정밀도는 50%, 부실 분류 재현율은 33.3%로 나타냈으며, 전체 분류 정확도는 99.3%이다.

<표 5> Result of Logistic regression analysis

Test data Correct Set	Pre_Normal	Pre_Insolvency	Accuracy
Act_Normal	412	1	99.3%
Act_Insolvency	2	1	

4.2.2 Artificial Neural Network

Artificial Neural Network 알고리즘은 인간의 뇌에서 수행되는 정보 등의 처리 방식을 표방한 기계학습 알고리즘으로써 과거 수집된 데이터를 반복적으로 학습하는 과정을 통하여 일정한 패턴을 발견하고 새로운 데이터에 대한 예측을 하는 비선형 모델이다.[7] 이 알고리즘을 통해 1,090개의 데이터를 훈련시키고 나머지 416개의 테스트 데이터를 이용하여 분석하였다. 분석 결과 정상 분류 정밀도(Precision)는 99.5%이며, 정상 분류 재현율(Recall)은 100%이다. 또한, 부실 분류 정밀도는 100%, 부실 분류 재현율은 33.3%를 나타냈으며, 전체 분류 정확도는 99.5%로 우수하게 나타났다.

<표 6> Result of Artificial Neural Network

Test data Correct Set	Pre_Normal	Pre_Insolvency	Accuracy
Act_Normal	413	0	99.5%
Act_Insolvency	2	1	

4.2.3 Support Vector Machine

Support Vector Machine 알고리즘은 특정 공간에서 부실 여부 등 두 클래스 사이의 최상의 분리초평면을 찾는 알고리즘으로 일정 데이터를 훈련하고 테스트 데이터가 어느 클래스에 속할 것인지 판단하는 비확률적인 이진 선형 분류 모델이다.[8] 본 연구에서는 Radial Basis 커널을 사용하여 1,090개의 데이터를 훈련시키고 나머지 416개의 테스트 데이터를 이용하여 분석하였다. 분석 결과 정상 분류 정밀도(Precision)는 99.5%이며, 정상 분류 재현율(Recall)은 100%이다. 또한, 부실 분류 정밀도는 100%, 부실 분류 재현율은 33.3%를 나타냈으며, 전체 분류 정확도는 99.5%로 우수하게 나타났다.

<표 7> Result of Support Vector Machine

Test data Correct Set	Pre_Normal	Pre_Insolvency	Accuracy
Act_Normal	413	0	99.5%
Act_Insolvency	2	1	

4.2.4 Decision tree

Decision Tree는 나무구조의 형태로 의사결정규칙을 분류하고 예측하는 방법으로, 노드(Node)들로 이루어져 있으며, 적은 계산비용에 비해 정확도가 높아 여러 분야에서 사용되고 있다.[9] 이 알고리즘을 통해 1,090개의 데이터를 훈련시키고 나머지 416개의 테스트 데이터를 이용하여 분석하였다. 그 결과 정상 분류 정밀도(Precision)는 99.3%이며, 정상 분류 재현율(Recall)은 100%이다. 전체 분류 정확도는 99.3%이지만 부실 분류를 한 건도 예측하지 못하였다.

<표 8> Result of Decision Tree

Test data Correct Set	Pre_Normal	Pre_Insolvency	Accuracy
Act_Normal	413	0	99.3%
Act_Insolvency	3	0	

이상으로 본 연구에서 Logistic regression analysis, Artificial Neural Network, Support Vector Machine, Decision tree의 네 가지 분류 알고리즘을 이용하여 검증해 본 결과, <표 9>와 같이 모두 분류 정확도가 99% 이상의 높은 수준을 나타났으며, 그 중 Artificial Neural Network 및 Support Vector Machine의 분류 정확도가 99.5%로 가장 우수함을 알 수 있었다.

<표 9> Result of Four Algorithms

Division	Precision(%)		Recall(%)		Accuracy (%)
	Normal	Insolv.	Normal	Insolv.	
Logistic regression analysis	99.5	50.0	99.8	33.3	99.3
Artificial Neural Network	99.5	100.0	100.0	33.3	99.5
Support Vector Machine	99.5	100.0	100.0	33.3	99.5
Decision tree	99.3	-	100.0	-	99.3

5. 결론 및 제언

본 연구는 2010년부터 2014년까지의 저축은행 경영관리 재무정보 데이터를 1,506개를 활용하여 BIS자기자본비율, 연체대출비율, PF대출 비중 증감률(전반기 대비), 총자산 순이익률 4개의 주요 변수를 추출하였으며, 선정된 변수를 기반으로 Logistic regression analysis, Artificial Neural Network, Support Vector Machine, Decision tree의 네 가지 분류 알고리즘을 이용하여 저축은행에 대한 조기 부실 예측 분류 정확도를 검증 하였다. 이에 본 연구의 주요 결과를 크게 두 가지로 설명할 수 있다.

첫 번째는 저축은행 부실에 영향을 주는 주요 변수 중 PF대출 비중 증감률이 선정된 것으로 분석기간인 2010년부터 2014년 사이에 부동산시장이 냉각되면서 부동산PF 부실이 급증하여 저축은행 건전성이 크게 악화되었음을 실제 분석을 통해 확인할 수 있었다.

두 번째로 현재 국내 저축은행에 대한 금융당국의 부실 예측 모형의 경우 Logistic regression analysis 알고리즘을 적용하고 주기적으로 설명변수인 지표만 변경하여 사용하고 있는 실정이다. 이러한 틀에서 벗어나 인공신경망 등 다양한 기계학습 알고리즘을 이용한다면, 부실 예측에 대한 정확도 및 신뢰도가 높아져 저축은행에 대한 부실을 사전에 예방하고 적기 조치가 가능해져 안정적인 금융시장을 유지할 수 있을 것이라는 의미 있는 결과를 얻을 수 있었다. 따라서 본 연구 결과를 바탕으로 금융당국은 현재 구현되어 있는 조기 부실 예측 모형에 대한 새로운 알고리즘을 적용해야 할 필요성이 있다.

한편, 본 연구에서는 기계학습 등의 알고리즘을 이용하여 저축은행 부실 발생 여부를 1분기 전 시점에서 예측해 보았다. 그러나 실제 금융당국이 금융기관으로부터 설명변수로 사용되는 재무정보를 늦게 제공받는 것을 고려한다면 부실 발생 2분기 전, 또는 1년 전의 데이터를 이용하여 모형을 개발하는 것이 실용적으로 보인다. 또한, 다양한 설명변수를 구성하고 정량적인 정보가 아닌 정성적인 데이터를 추가하여 지표를 설정한 후 연구를 진행한다면 더 높은 분류 정확도를 가진 부실 예측 모형을 개발할 수 있을 것으로 기대된다.

참고문헌

[1] 「상호저축은행 구조조정 특별계정 관리백서」, KDIC, 2017

[2] 「2017년 저축은행 영업실적(잠정)」, 금융감독원 2018

[3] Daniel Martin, 「Early Warning of Bank Failure」, Journal of Banking and Finance(1977) pp. 249-276.

[4] Jagtiani, Julapa, James Kolari, Catharine Lemieux and Hwan Shin, 「Early Warning Models for Bank Supervision: Simpler Could Be Better」, Economic Perspectives, Vol.27, Federal Reserve Bank of Chicago, 2003, pp. 49-60.

[5] Boyacioglu, Melek Acar, Yakup Kara, Omer Kaan Baykan, 「Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey」. Expert Systems with Application 36, 2009, pp.3355-3366.

[6] 강병서·김계수, 「SPSS 17.0 사회과학 통계분석」,한나래 출판사, 2009

[7] Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., & Tourassi, G. D., "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance", Neural networks, 2008, Vol.21, No.2, 427-436

[8] S. Y. Choi, H. C. Ahn, "Optimized Bankruptcy Prediction through Combining SVM with Fuzzy Theory", Journal of digital Convergence, 2015, Vol.13, No.3, pp.155-165

[9] Pal, M., & Mather. P. M., "An assessment of the effectiveness of decision tree methods for land cover classification", Remote sensing of environment, 2003, Vol.86, No.4, pp.554-565