한국비즈니스리뷰
제 10 권 3 호 2017년 12월

기계학습 기법을 이용한 소상공인 신용평가모형 구축에 관한 연구

박주완* I 송창길** I 배진성***

본 논문은 로지스틱회귀모형, 의사결정나무모형, 신경망모형을 이용하여 소상공인 신용평가모형을 구축하고, 예측 성능이 가장 좋은 모형이 무엇인지를 확인하는 것이다. 모형 구축을 위한 분석 대상은 지역신용보증재단에서 보유하고 있는 자료이다. 이 자료를 이용하여 결측치와 특수값 등을 제거하고 통계적인 변수 선택 기법을 적용하여 최종적으로 15개의 독립변수와 67,308개의 차주의 자료가 모형 구축에 사용되었다. 구축된 세 가지 모형은 10중첩 교차타당법을 이용해 평가하였으며, 모형 평가 측도로는 오분류율, G-mean, F1측도와 반응률을 이용하였다.

지역신용보증재단의 자료와 세 가지 기계학습 기법을 이용하여 3개의 모형을 구축한 결과, 로지스틱회귀모형을 적용했을 경우 예측 성능이 가장 우수한 것으로 나타났다. 또한 계급불균형인 자료를 이용하여 기계학습 모형 구축 시 예측 성능이 저하될 수 있다는 사실을 발견하였다.

본 논문은 소상공인 신용평가모형 구축에 대해 자료의 양과 신뢰성 부족 문제로 비재무 자료 이용이라는 기존 연구의 틀에서 벗어나, 기계학습기법 적용 가능성을 확인하였다는 점에서 의의가 있다.

핵심주제어: 소상공인, 신용평가, 기계학습, 교차타당법

I. 서 론

경제활동에서 말하는 신용은 과거의 경제활동상태와 현재의 경제적인 능력을 살펴 미래에 빚을 상환할 수 있는 가능성을 판단하는 것이다(이명식·김정인, 2007). 이와 같은 개인이나 기업의 채무상환 능력의 판단 근거가 되는 신용도[1]는 신용상태에 영향을 미치는 많은 정보를 조사한 후 이를 분석하여 측정한다. 신용도는 재무 및 비재무 자료를 통계적인 기법에 적용하여 구축한 신용평가모형에 의해 산출된 신용점수나 등급으로 측정이 가능한데, 신용평가모형은 신용평가 전문기관[2]과 금융 관련 사기업 및 공공기관에서 금융상품의 목적과 평가 대상 등에 따라 자체적으로 구축하여 운용하고 있다(박주완·송창길, 2015).

각 기관에서 산출하는 신용평가의 결과는 분석대상, 자료, 방법론 등에 따라 상이할 수 있지만, 공급자와 수요자 간 정보 불균형(information asymmetry) 문제를 해결하는 데에 많은 도움을 주

1) 신용도(credibility)란 장래의 어느 시점에 그 대가를 지급할 것을 약속하고 경제적 가치를 획득할 수 있는 능력을 의미한다(이명식과 김정인, 2007).
2) 국내의 대표적인 기관으로는 나이스평가정보, 한국기업데이터, KCB, S&P, Moodys, Duff & Phelps 등이 있다(박주완과 송창길, 2015).
본 논문은 한국연구재단과 지식경영연구원에서 정한 연구윤리규정을 준수함
* 박주완(제1저자)_신용보증재단 중앙회 조사연구실, 연구위원, 통계학박사(jwan0217@koreg.or.kr)
** 송창길(제2저자)_국민연금연구원 재정추계분석실, 주임연구원, 이학박사(cgsong82@nps.or.kr)
***배진성(제3저자)_신용보증재단중앙회 조사연구실, 연구위원, 경제학박사(bjs0423@koreg.or.kr)
논문접수 : 2017. 11. 30 I 1차 심사일 : 2017. 12. 13 I 게재확정일 : 2017. 12. 18

고 있다. 그러나 정확한 신용등급을 산출하지 못할 경우 왜곡된 정보로 인해 오히려 시장 실패나 잘못된 대출의 가능성이 높아지게 되므로, 신용평가의 정확성과 신뢰성 확보는 매우 중요하다(김성환· 김태동, 2014). 이러한 이유로 신용평가를 수행하는 기관에서는 신용평가 결과의 정확성과 신뢰성 향상을 위해 새로운 기법 적용과 활용 가능한 자료 확장 등에 많은 노력을 기울이고 있다(박주완·송창길, 2015).

앞에서 언급한 노력들의 일환으로 금융 산업에서는 빅데이터(big data)와 기계학습(machine learning)[3]을 신용평가에 활용하기 위해 많은 연구와 노력을 하고 있으며, 실제 현업에서 빅데이터와 기계학습을 적용하는 사례들이 점차 증가하고 있다. 해외에서는 Kabbag[4], Zest Finance[5] 등의 P2P 대출업체들이 기계학습 기법과 빅데이터를 신용평가에 활용하고 있다(신윤재, 2016). 국내에서는 신한카드사가 신용도 판단이 어려운 사회 초년생과 중금리 대출 고객들을 대상으로 2017년 초 기계학습 기법을 적용한 신용평가시스템 개발을 완료하였다(서울경제신문, 2017). 이와 같은 사례를 통해 신용평가 시 빅데이터와 기계학습 기법 이용에 대한 관심과 중요성이 점차 높아지고 있음을 유추할 수 있다.

3) 기계학습은 종종 데이터마이닝과 혼용되고 있는데, 기계학습에 사용하는 분류나 군집 등의 방법을 데이터마이닝에서도 동일하게 사용하기 때문이다. 분류나 예측, 군집과 같은 기술, 모델, 알고리즘을 이용해 문제를 해결하는 것을 컴퓨터과학 관점에서는 기계학습이라 하고, 통계학 관점에서는 데이터마이닝이라고 한다. 데이터마이닝 관련 서적들은 전통적인 통계 분석 모형인 군집모형, 회귀모형 등과 기계학습 모형으로 알려져 있는 의사결정나무모형, 신경망모형, 랜덤포레스트 등이 모두 포함되어 설명되고 있다(김의중, 2016).

4) Kabbage는 소상공인 신용평가 시 기존의 재무적인 자료 이외의 배송, 회계, 인터넷 자료 등을 기계학습 기법에 적용하고 있다(신윤재, 2016).

5) Zest Finance의 경우 전통적인 신용정보 외에 직장정보, 고정수입, 인터넷 포스팅 내용 등이 포함된 7만개가 넘는 변수를 가진 빅데이터에 10개의 기계학습 모형을 적용하여 신용평가를 하고 있다(신윤재, 2016).

최근 소상공인을 대상으로 한 대출이 점차 활성화되면서, 소상공인 신용평가의 신뢰성과 정확성에 대한 요구가 커지고 있으며, 신뢰성 높은 자료 확보를 통한 신용평가모형 구축의 중요성은 점차 증가하고 있다. 그러나 기존의 부실기업 예측 등 신용평가 모형에 대한 연구는 주로 재무적인 자료가 충분한 어느 정도 규모가 있는 기업들을 대상으로 이루어지고 있으며, 소상공인 부도 예측 및 신용평가에 대한 연구는 많지 않은 것이 사실이다. 윤상용 외(2016)는 소상공인을 대상으로 한 신용평가 연구는 대기업이나 중소기업에 비해 상대적으로 그 수준이나 양적인 면에서 매우 미미한 수준이라고 언급하고 있다.

소상공인에 대한 신용평가 연구가 상대적으로 부족한 이유 중 하나는 신뢰성 있는 분석 자료의 부족이라는 한계에 기인한다. 소상공인은 개인적 특성과 기업적 특성이 혼재하고 있지만, 기업 신용정보를 대변하는 재무제표와 대차대조표 등 객관적인 정보가 부족(신용보증재단중앙회, 2016)하여 통계적인 기법을 적용한 신용평가 연구가 어려운 것이 현실이다. 이로 인해 소상공인 신용평가에 대한 연구는 기계학습 기법 적용보다 비재무적인 자료를 이용하는 등 사용 가능한 자료 활용에 대한 연구가 주를 이루고 있다. 실제로 소상공인 신용평가에 있어 기법 측면의 연구는 심사역의 경험에 의존하여 주관으로 점수를 산출하는 AHP (Analytic Hierarchy Process) 기법에 대한 연구가 주를 이루고 있다(윤종식·권영식, 2007). 이외에도 개인정보보호 등 정보의 사용 제약과 전통적 신용평가 기법의 고착화 등으로 인하여 새로운 기계학습 기법을 적용한 신용평가모형 연구는 한계가 많다(신윤재, 2016).

그러나 신용대출과 관련하여 향후에도 빅데이터와 기계학습을 활용한 신용평가 요구는 점차 늘어날 것으로 예측된다. 4차 산업혁명의 도래와 이를 선도하는 빅데이터 및 기계학습의 활용에 대한 요구가 금융 산업 분야에서도 점차 증가할 것으로 예측되는 가운데, 신용평가모형 구축에 있어 이를 적용하기 위한 노력이 어느 정도인지 고민해 볼 필요가 있다. 그리고 머지않은 미래에 빅데이터와 기계학습을 활

용한 신용평가가 보편화 될 때를 대비해 관련 기관에서는 자체적으로 충분한 데이터 확보와 모형 구축을 위한 분석 능력이나 기술력을 보유할 필요가 있을 것으로 사료된다.

본 논문은 이러한 문제의식 하에서 16개 지역신용보증재단이 보유한 다양한 내·외부 자료와 기계학습 기법을 이용하여 소상공인 신용평가모형을 구축하고 예측 성능이 좋은 모형이 무엇인지를 확인하는 것이 주된 목적이다. 분석 자료는 2013년 7월부터 2016년 12월까지 3년 6개월 동안 지역신용보증재단에서 신용보증을 받은 차주에 대한 내부자료와 NICE신용평가의 대표자 CB요약정보이다. 사용하고자 하는 기계학습 기법은 로지스틱회귀모형[6], 의사결정나무모형과 신경망모형이고, 모형의 평가 방법 및 측도(measure)는 10중첩 교차타당법과 오분류율, G-mean, F1측도, 반응률(percent of response)이다. 자료를 분석하고 모형을 구축하기 위한 통계프로그램은 SAS9.4와 R3.4.1 버전을 이용한다. 이 때 SAS로는 자료의 기초분석 및 데이터 정제(data cleaning) 작업을 수행하고, R은 "glm, rpart, nnet" 함수를 이용하여 세 가지 기계학습 모형을 구축하고 평가한다.

본 논문은 과거 비재무 자료 활용이라는 측면의 연구에서 벗어나, 4차 산업혁명 시대의 도래에 맞추어 기계학습 기법을 이용하여 소상공인 신용평가모형 구축의 가능성을 연구한다는 점에서 기존 연구들과 차별성이 있을 것으로 사료된다.

논문의 구성은 다음과 같다. Ⅰ장 서론에 이어, Ⅱ장에서는 기계학습 기법을 이용한 기업신용평가 관련 기존 연구사례들을 고찰하고, Ⅲ장에서는 예측모형 구축을 위한 알고리즘과 모형 평가 방법을 설명한다. Ⅳ장은 소상공인 신용평가모형을 구축하기 위한 연구 대상, 변수 선정, 모형 구축 과정에 대해 설명하며, Ⅴ장에서는 실제 자료를 이용하여 모형을 구축하여 예측 성능을 평가한다. 마지막으로 Ⅵ장에서는 결론 및 향후 과제에 대해 고찰한다.

Ⅱ. 기계학습과 신용평가

신용평가모형에 대한 연구는 기법적인 측면과 자료 활용적인 측면에서 고려할 수 있다. 먼저 기법적인 측면에서는 일정 규모 이상의 기업을 대상으로 전통적인 통계 모형 이외에도 신경망모형(neural network model) 등 다양한 기계학습 기법을 이용한 신용평가 방법들에 대한 연구가 활발히 진행되고 있다(강신형, 2016). 자료 활용 측면에서는 재무적 요인 이외에 비재무적 요인의 활용뿐 아니라(장원경·김연용, 2002), 근래에 들어 다양한 형태의 빅데이터를 사용하는 연구들이 활발히 진행되고 있다. 본 논문에서는 기법적인 측면에서 모형 구축과 비교를 수행하는 것이 주요 목적이기 때문에 신용평가모형 구축 시 다양한 기계학습 알고리즘을 비교한 연구 문헌에 대해 고찰하고자 한다.

기계학습 기법들을 이용한 부실기업 예측 및 기업신용평가모형 구축에 대한 연구 사례를 살펴보면 다음과 같다. 먼저 인공신경망 모형의 우수성을 검증한 연구 사례이다. 이건창(1993)은 1979~1992년 사이의 기업 자료를 이용하여 다변량판별분석, 인공신경망모형을 구축한 결과 인공신경망모형의 예측력이 높다고 하였다. 박정윤(2000)은 1991~1996년 자료로 기업 부실 예측을 실시한 결과 MDA모형, 확률모형, 인공신경망모형 중 인공신경망모형의 예측력이 우수하다고 하였다. 전성빈·김영일(2001)은 기업의 도산 예측력 시 인공신경망모형의 예측력이 가장 우월하였고 다변량판별분석, 로짓모형 등의 분류정확도는 비슷한 수준이라고 하였다. 정유석(2003)은 로짓모형, 다변량판별분석, 인공신경망모형을 이용하여 도산 기업을 예측한 결과 인공신경망모형의 예측력이 가장 우수하다고 하였다.

다음은 의사결정나무모형, 로지스틱회귀모형과 SVM(Support Vector Machine)모형의 판별력이 우수함을 실증 분석한 논문이다. 조준희와 강부식

6) 신용보증재단에서는 소상공인 신용평가모형 구축 시 로지스틱회귀모형을 사용하고 있다.

(2007)은 여러 가지 기계학습 기법을 이용하여 코스닥기업의 도산 예측 모형을 구축한 결과 의사결정나무모형이 신경망모형이나 로지스틱회귀모형 보다 좋은 예측 성능을 가지고 있다고 하였다. 박주완·송창길(2015)에서는 인적자본기업패널자료와 NICE자료를 이용하여 소기업 이상에 대해 로지스틱회귀모형, 신경망모형, 의사결정나무모형을 이용하여 신용평가모형 구축 결과 로지스틱회귀모형의 예측 성능이 더 우수함을 실증분석을 통해 검증하였다. 윤종식·권영식(2007)은 소상공인 부실예측모형 연구에서 로지스틱회귀모형, 다변량판별분석, CART, C5.0, 신경망 모형, SVM모형의 예측 성능을 비교한 결과, SVM모형을 이용하였을 때 예측 성능이 가장 우수함을 보였다.

마지막으로 앙상블 기법을 이용한 연구논문이다. 김승혁· 김종우(2007)은 SOHO 부도예측에 있어서 붓스트랩(Bootstrap) 방법으로 다수의 모델을 만들고 평균 이상의 예측 정확도를 가지는 모형들만을 선택해 투표(voting)하는 Modified Bagging Predictors가 인공신경망과 Bagging Predictors에 비해서 예측 성능이 향상됨을 확인하였다. 김명종·강대기(2010)는 기업 부실 예측을 위해 인공신경망과 부스팅 인공신경망 앙상블 기법을 적용한 결과 앙상블 학습은 기업부실 예측 문제에 있어 전통적인 인공신경망을 개선할 수 있음을 검증하였다. 김성진·안현철(2016)은 1,295개 국내 상장 기업을 대상으로 기업신용평가모형 구축 시 다변량판별분석, 인공신경망, 다분류 SVM모형, 랜덤포레스트모형을 비교한 결과 랜덤포레스트모형의 예측 성능이 더 우수함을 보였다.

앞에서 제시한 연구 사례들의 결과를 살펴보면 분석 자료에 따라 결과가 상이하게 나타나고 있으며, 대부분 일정 규모 이상의 기업을 대상으로 이루어지고 있다. 이에 반해 소상공인의 부도 예측이나 신용평가에 대한 연구는 소상공인의 특성과 자료의 부족 등으로 인해 모형 구축 연구가 많지 않고 제한적이다. 그러므로 소상공인을 대상으로 기계학습 기법을 이용한 신용평가모형 구축에 대한 연구는 합당한 시도이며 의미가 있는 연구로 사료된다.

III. 모형 구축 알고리즘 및 평가방법

3.1 모형 구축 알고리즘

기계학습 관점에서 신용평가모형의 개념을 살펴보면 다음과 같다. 차주의 우불량 여부를 판별하고 신용도를 예측하기 위한 신용평가모형은 기계학습 관점에서 지도학습(Supervised Learning) 중에서 분류(classification) 모형이다. 지도학습이란 훈련용 자료(training data)로부터 예측이나 분류를 위한 함수를 추정하는 것으로 독립변수(independent variable)를 이용하여 종속변수(dependent variable)의 값을 예측하거나 분류하는 것을 말한다. 그러므로 지도학습을 위한 자료에는 종속변수와 독립변수가 필요하다. 대표적인 지도학습 모형으로는 선형회귀모형, 로지스틱회귀모형, 의사결정나무모형, 신경망모형, 랜덤포레스트모형, SVM 등이 있다. 이 중에서 본 논문에서 사용할 예측 모형 구축 알고리즘은 보편적으로 많이 알려져 있고 비교적 사용이 용이한 로지스틱회귀모형, 의사결정나무 모형, 신경망모형 세 가지이다. 모형 구축을 위해 사용된 세 가지 모형에 대해 간략히 소개하면 다음과 같다.

로지스틱회귀모형은 종속변수 Y_i가 이진형(binary type)인 경우 반응함수 $E(Y_i \mid \boldsymbol{x_i}'s)$는 $\boldsymbol{x_i}'s$가 증가함에 따라 값이 1로 서서히 수렴하는 모형으로, 종속변수가 1이 될 확률인 $\Pr(Y=1|x_i's)$를 예측하는 모형이다(Hosmer와 Lemeshow, 2000). 즉, 로지스틱회귀모형은 관심의 대상에 대한 반응확률과 독립변수 사이의 관계를 분석하기 위한 회귀모형이다. 일반적으로 신용평가모형을 구축할 때 로지스틱회귀모형이 많이 선호되고 있으며 실제로 가장 많이 사용되고 있다. 그 이유는 첫째 모형 구축이 올바르다면 로지스틱회귀모형은 정확성이 우수하고, 둘째 구축 과정이 용이하고 해

석하기가 쉬우며, 셋째 과대 적합(over-fitting)할 가능성이 적고, 오차를 최소화하는 선형적인 관계를 찾는데 매우 우수한 기법이기 때문이다(이영섭, 2003). 본 논문에서는 통계프로그램 R의 "glm 함수"를 사용하여 분석을 수행한다. glm 함수는 선형회귀분석, 로지스틱회귀분석, 일반선형모형 등의 회귀분석을 수행하는 것으로써 R 프로그램에서 기본적으로 제공하고 있다.

의사결정나무모형은 의사결정 규칙(decision rule)을 나무 구조로 도표화하여 분류와 예측(prediction)을 수행하는 방법이다. 모형의 나무 형성 알고리즘은 CART(Classification And Regression Trees), CHAID(CHi-squared Automatic Interaction Detection), C4.5(또는 C5.0)가 대표적이다. 본 논문에서는 통계 프로그램인 R에서 "rpart 라이브러리"를 사용하여 분석을 수행한다. rpart 라이브러리는 CART 알고리즘을 구현한 패키지이다.

CART에 대해 좀 더 자세히 설명하면 다음과 같다. CART는 1984년에 L. Briemen에 의해 최초 발표된 기계학습 실험의 시초가 된 알고리즘이다. CART 알고리즘은 독립변수들과 종속변수로 이루어진 자료에서 독립변수의 특성에 따라 이진 분류(binary split)를 수행하며, 마디의 순수함을 나타내는 지니 지수(Gini index)에 의해 분리 여부를 결정한다. 특정 변수에 의해 집단이 구분되면, 구분된 하나의 집단에서 나머지 집단의 개체가 선택될 확률을 계산하여 집단을 분리하며 집단이 순수할수록 지니 지수의 값과 확률이 작아지게 된다.

의사결정나무모형의 가장 큰 장점은 분류 규칙을 나무구조의 도표를 통해 확인할 수 있으므로 이해가 쉽다는 것이다. 또한 연속형과 범주형 자료를 동시에 다룰 수 있고, 특정 변수에 결측치가 발생해도 이를 분석에 활용할 수 있다. 그러나 훈련용 자료에 대한 최적의 의사결정나무를 찾는 것은 쉽지 않으며, 훈련용 자료에 대해 매우 세밀하게 분류 및 예측을 수행할 경우 과대적합(over-fitting)의 가능성이 높아 새로운 자료에 대한 일반화 성능이 좋지 않을 수 있다(최종후·진서훈, 2005).

신경망모형은 인간의 뇌 기능에 착안하여 개발된 패턴인식의 한 분야로 과거의 경험이나 지식을 습득함으로써 오류를 최소화하는 과정들을 포함하고 있으며, 어떠한 통계적인 분포도 가정하고 있지 않다. 다양한 신경망 알고리즘 중 가장 널리 사용되는 모형은 다층인식자(Multi-Layer Perceptron, MLP) 신경망이다. 다층인식자 신경망은 입력층을 통해 자료를 입력받고, 은닉층에서는 입력층으로부터 전달되는 변수값들의 선형결합(linear combination)을 비선형함수로 처리하여 출력층 또는 다른 은닉층으로 전달하여 최종적으로는 출력층을 통해 예측 결과를 산출한다(강창완 외, 2007).

신경망모형의 장점은 자료들 간의 비선형적인 관계를 찾아 낼 수 있다는 것이다. 그러나 자료를 과대적합하는 경향이 있기 때문에 훈련을 통해 구축된 모형에 새로운 자료를 적용했을 때 예측 성능이 좋지 않을 수 있다는 단점이 있다. 또 다른 단점은 의사결정나무, 회귀분석 등의 기법에 비해 결과의 해석이 매우 어렵다는 것이다(Ripley, 1996). 일반적으로 신경망모형은 은닉층의 개수를 늘릴수록 예측 성능은 향상되지만, 과도하게 많을 경우 모형 실행 시간이 과다해지고, 다른 자료 적용 시 강건성(robustness)이 떨어지는 단점이 있다(김의중, 2016). 본 논문에서는 통계프로그램 R에서 "nnet 라이브러리"를 사용하여 분석을 수행한다. nnet 라이브러리에서 신경망의 파라미터(parameter)는 엔트로피(entrophy) 또는 오차제곱합(sum of squared error, SSE)을 고려해서 최적화된다.

3.2 모형 평가 방법

최적의 모형을 얻기 위해서는 여러 모형을 비교하여 가장 우수한 모형을 선택하여야 하는데, 이를 위한 과정이 모형 평가이다. 모형 평가는 예측을 위해 만든 여러 가지 모형의 예측과 분류 성능을 평가 및 비교하여, 가장 좋은 예측력을 보유하고 있는 모형을 선택하기 위한 필수 단계이다. 개발된 모형의

타당성을 검토하는 방법들로는 별도의 평가용(validation) 자료를 이용한 예비 방법(holdout method), k개의 분할된 자료를 이용하는 k-중첩 교차타당법(k-fold cross validation method)과 부스트랩 방법(bootstrap method) 등이 있다(Kohavi, 1995). 세 가지 모형 평가 방법 중 본 논문에서는 10중첩 교차타당법을 사용하여 구축된 모형의 성능을 비교한다.

교차타당법의 과정은 다음과 같다. 먼저 초기 자료를 크기가 유사하게 D_1, …, D_k인 k개의 상호 배반 부분집합으로 임의 분할한다. 분할된 자료를 이용하여 D_1, …, D_{k-1}를 모형 구축에 사용하고, D_k를 이용하여 검증한다. 분할된 자료가 k개이므로, 모두 k번의 모형 훈련과 평가가 발생한다. 교차타당법은 모형 평가를 위해 사용한 자료의 크기가 크지 않은 경우, 모형 평가에 소요되는 시간이 단축되는 장점이 있다. 일반적으로 10중첩 교차타당법이 예측 모형의 정확도를 추정하는데 많이 사용된다(강창완 외, 2007). 10회 중첩을 사용하는 이유는 상대적으로 작은 편향과 분산을 가지며, 시뮬레이션을 수행한 결과 10회 중첩이 최적의 오차 추정 값을 얻기 위해 필요한 횟수로 판명되었기 때문이다(이승현, 2014).

3.3 모형 평가 척도

일반적인 모형 평가는 종속변수가 범주형인 경우 실제값과 예측값 사이의 오분류 행렬(confusion matrix)을 통한 평가 측도인 오분류율, 민감도, 특이도와 ROC곡선 등으로 모형의 예측 성능을 평가할 수 있다. 이외에도 리프트(lift) 도표, 반응률(response rate) 도표 등이 각종 데이터마이닝 도구에서 많이 사용되고 있다(강현철 외, 1999). 다양한 모형 평가 측도 중 본 논문에서는 오분류율, G-Mean, F1측도, 반응률의 값으로 구축된 소상공인 신용평가모형의 예측 성능을 비교하고 평가한다.

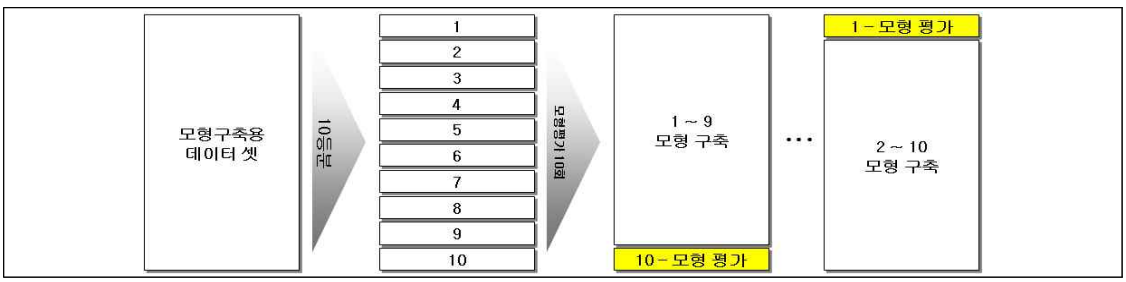
오분류율은 범주가 0과 1인 두 가지의 결과를 가질 경우 실제 0인 범주를 1 또는 실제 1인 범주를 0으로 오분류하는 확률이다. 오분류율은 전체 자료를 얼마나 잘못 분류하는가의 문제이므로 값이 적을수록 좋은 모형이다. 확률변수 X는 0 이나 1인 실제값이고, 확률변수 E는 0 이나 1의 값을 가지는 예측값이라고 할 때 오분류율은 (1)과 같이 나타낼 수 있다.

G-mean은 결과 범주가 0인 집단과 1인 집단을 동등하게 고려하는 측도로써 실제 범주가 0인 집단에 대한 정확도와 범주 1인 집단에 대한 정확도의 기하평균이다. 그러므로 G-mean의 값이 클수록 좋은 예측 모형이다. G-mean의 산식은 (2)와 같다.

$F1$측도는 어떤 특정한 계급의 성공적인 분류가 다른 계급의 분류에 비해 훨씬 중요한 경우 사용되는 측정 기준이다. $F1$측도는 특정계급, 특히 우량과 불량 간 불균형인 경우 소수계급에 주된 관심을 가지고 있으며, 이 값이 크다는 것은 특정계급에 대한

$$\text{오분류율} = \Pr(X=1, E=0) + \Pr(X=0, E=1) \quad (1)$$

<그림 1> 10중첩 교차타당법



예측 성능이 좋다는 것을 의미한다(Chawla 외, 2003). $F1$측도를 산출하기 위한 수식은 다음과 같다.

여기에서 p는 Precision이고 r은 Recall 값을 나타내며 민감도(sensitivity)라고도 불린다. p값이 크다는 것은 모형에 의해 예측된 자료 중 자료가 잘못 예측될 개체수가 적다는 것이며, r값이 크다는 것은 실제로 1의 값을 가지는 자료가 정분류될 가능성이 높음을 의미한다. 일반적으로 Precision이나 Recall 값이 클수록 구축된 모형의 특정계급에 대한 예측 성능이 좋음을 의미하므로, 이 두 가지 측도를 최대화하는 모형을 구축하여야 한다.

반응률은 모형을 통해 산출된 사후확률의 순서에 따라 전체 자료를 오름차순으로 정렬하고, 이를 N개의 집단으로 등분한 후 각 집단에서 반응변수의 특정 범주에 대한 빈도를 이용하여 계산한다. 반응률을 이용한 반응률 도표를 통해 점수가 높은 10분위수에서 높게 나타나다가 급격하게 감소하는 형태 또는 그 반대인 경우 좋은 예측 모형이다.

Ⅳ. 모형 구축

4.1 연구 대상

본 논문의 모형 구축 대상은 16개 지역신용보증재단에서 2013년 6월부터 2016년 12월까지 3년 6개월 동안 소상공인 신용평가모형을 통해 평가를 받은 차주 105,572개이다. 이중 결측치(missing value)가 발생하였거나 "-999,999,999" 등 특수값(special value)이 있는 표본은 분석에서 제외하였으며, 그 결과 총 67,308개 차주가 최종 모형 구축 대상이다.

소상공인 신용평가모형 구축을 위해 최초 총 50개의 변수를 이용한다. 독립변수는 총 49개인데 이 중 신용보증재단 내부정보는 27개, NICE CB요약정보는 22개이다. 모형 구축에 사용된 변수의 출처는 신용보증재단 내부 자료와 NICE CB 요약 정보인데, <표 1>은 본 논문에서의 모형 구축을 위해 최초로 사용하는 변수이다.

종속변수인 우불량(우량=0, 불량=1)은 차주의 신용평가 날짜를 기준으로 12개월 이내에 사고 30일 이상, 대위변제 발생, 신용정보원에 채무불이행 등재 경험이 있는 경우 불량으로 정의한다. 우불량 관측 기간은 금융감독원의 우불량 관측 기간에 대한 기준을 충족(최소 12개월)하는 범위 내에서 자료의 최신성과 안정성을 고려하여 12개월로 선정한다(신용보증재단중앙회, 2017).

$$G-mean=\sqrt{\frac{\text{실제}0,\text{예측}0\ \text{정분류 빈도}}{\text{실제 }0\text{ 빈도}}\times\frac{\text{실제}1,\text{예측}1\ \text{정분류 빈도}}{\text{실제 }1\text{ 빈도}}} \tag{2}$$

$$F1=\frac{2rp}{(r+p)}=\frac{2}{1/r+1/p}=\frac{2\cdot \text{실제}1\,\text{예측}1\,\text{정분류 빈도}}{\text{예측}1\text{빈도}+\text{실제}1\text{빈도}}$$
$$,\ p=\text{실제}1,\text{예측}1\,\text{정분류 빈도}/\text{예측}1\text{의 빈도} \tag{3}$$
$$,\ r=\text{민감도}=\text{실제}1,\text{예측}1\,\text{정분류 빈도}/\text{실제}1\text{의 빈도}$$

$$\text{반응율 값}\ =\ \frac{\text{일정 }N\text{ 등분내 범주 }1\text{ 빈도}}{\text{일정 }N\text{ 등분내 전체 빈도}} \tag{4}$$

4.2 모형 구축 절차

본 논문의 목적은 소상공인 신용평가 시 다양한 기계학습 모형을 구축한 후 그 결과를 비교하여 가장 성능이 좋은 모형이 무엇인지 알아보는 것이다. 이를 위한 모형 구축 절차는 다음의 <그림 2>와 같다.

첫째, 신용보증재단의 내부 자료와 NICE CB요약정보를 이용하여 모형 구축용 자료 세트를 구성하고 분석을 위한 종속변수인 차주의 우량과 불량을 정의

<표 1> 모형 구축을 위한 변수

구분		변수
종속변수(1개)		①우불량 여부 – 불량 기준 : 사고일수 30일 이상, 대위변제, 신용정보원채무불이행등재
독립변수 (49개)	신용보증 재단 내부정보 (27개)	①업력(개월), ②자가/임차여부, ③사업장임차보증금액(원), ④종업원수(명), ⑤차입금금액(원), ⑥차입기관수(개), ⑦년매출금액(원), ⑧여신거래실적금액(원), ⑨유가증권금액(원), ⑩현금서비스사용금액(원), ⑪거주기간(개월), ⑫직권말소여부, ⑬연체보유수량(개), ⑭소유부동산금액(원), ⑮주택임차보증금액(원), ⑯예금적금금액(원), ⑰기타현금금액(원), ⑱기차입금액(원), ⑲임대보증수입금액(원), ⑳기타고정수익금액(원), ㉑배우자소득금액(원), ㉒기타지출금액(원), ㉓기타가계소득금액(원), ㉔매출원가금액(원), ㉕개인법인구분, ㉖판매금액및일반관리금액(원), ㉗이자비용금액(원)
	NICE CB요약정보 (22개)	①채무불이행등록총건수(개), ②신용카드발급기간(일), ③3개월내신용카드의평균한도소진율(%), ④6개월내신용카드일시불이용률(%), ⑤12개월내신용카드현금서비스경과개월수(개월), ⑥대출총기관수(개), ⑦미상환대출개설일로부터의 기간(일), ⑧미상환대출총금액(원), ⑨캐피탈업권미상환대출총금액(원), ⑩저축은행업권미상환대출총금액(원), ⑪1년내연체총건수(개), ⑫연간원리금상환금액(기준1)_20150701금리기준(원), ⑬연간원리금상환금액(기준1)_20150701실상환금액미반영(원), ⑭연간이자금액상환금액1(원), ⑮주택/부동산담보대출제외연간원리금상환금액(원), ⑯월상환금액(SC저축은행별도기준)(원), ⑰연간원리금상환금액(기준2)_20140701금리기준(원), ⑱연간원리금상환금액(기준2)_20140701실상환금액미반영(원), ⑲연간원리금상환금액(전은연기준은행업권용)(원), ⑳연간원금상환금액(전은연기준은행업권용)(원), ㉑연간이자상환금액(전은연기준은행업권용)(원), ㉒미상환대출총금액중캐피탈&저축은행업권비중(%)

한다. 둘째, 모형 구축 단계에서는 모형 구축을 위한 변수를 선택하고 이를 이용하여 로지스틱회귀모형, 의사결정나무모형, 신경망모형을 이용하여 소상공인에 대한 신용평가모형을 구축한다.

모형 구축용 변수 선정의 첫 번째 과정은 Hosmer와 Lemshow(2000)에서 제안하는 방법인 각 개별 독립변수마다 단변량 로지스틱회귀분석을 통해 p-값이 0.05 이하인 변수를 1차적으로 선정한다. 다음으로 다중공선성이 있는 변수를 제거하기 위해 상관분석을 통해 상관계수가 0.7 이상인 변수 중 하나를 선택한다.

모형 구축을 위한 최종적인 변수 선택은 이영섭과 박주완(2007)에서 사용한 방법인 로지스틱회귀모형에서의 단계적 변수선택법(stepwise)과 최적조합 변수선택법(best selection)을 이용한다. 선택된 변수는 특이값(outlier) 등의 영향을 최소화하기 위해 최초-최대(mix-max) 표준화 기법[7]을 사용하여 자료를 표준화한다. 마지막 단계는 최종적으로 선택된 변수를 이용하여 구축된 세 가지 모형을 평가하고 비교하는 단계로써, 모형 평가 방법은 10중첩 교차타당법을 적용하며, 예측 성능 비교를 위한 측도로는 오분류율, G-mean, F1측도, 반응률을 이용한다.

V. 분석 결과

5.1 기초분석

소상공인 신용평가모형을 구축하기 위한 종속변수의 분포는 <표 2>와 같다. 분석대상 차주 총 67,308개 중에서 종속변수인 우량과 불량 차주는 각각 97.8%, 2.2%를 차지하고 있으며, 우량 차주가 불량 차주 보다 훨씬 많은 계급불균형(imbalanced)[8] 자료이다. 종속변수의 계급이 우량(Y=0), 불량(Y=1) 2개이므로 오분류표를 작성하기 위한 분류절단값(cut-off value)은 0.5를 기준으로 한다. 성웅현(2016)에 의하면 종속변수의 범주가 2개인 경우 분류기준값은 일반적으로 0.5를 사용하여 오분류표를 작성한 후 예측 성능을 비교할 수 있다고 언급되어 있다.

모형 구축을 위해 독립변수 선택 방법들을 단계적으로 적용한 결과 최종적으로 15개가 선정되었으며, 이중에서 신용보증재단의 내부정보는 7개, NICE CB요약정보는 8개이다. 최종적으로 선정된 독립변수와 종속변수 간의 관계 분포는 <표 3>과 같다.

<그림 2> 모형 구축 단계

7) 원래의 자료를 선형적(linear)으로 일정한 구간(일반적으로 0과 1사이)으로 변환하는 것을 말한다. 표준화를 위한 공식은 표준화값(A)=(x_i- 최소값)/(최대값-최소값)이다.

8) 계급불균형 자료는 종속변수가 이진형인 경우 각 계급 간 빈도의 차이가 매우 심한 경우를 의미한다. 이와 같은 계급불균형 자료를 이용하여 모형을 구축할 경우 문제점들이 발생할 수 있다(박주완, 2010).

<표 2> 종속변수 분포

(단위: 개, %)

구 분	우량(Y=0)	불량(Y=1)
빈도(비율)	65,806(97.8)	1,502(2.2)

<표 3> 독립변수 분포

변수(단위)	우량(Y=0)		불량(Y=1)	
	평균	표준편차	평균	표준편차
개인/법인 구분(이진형)-개인(1)	96%	-	92%	-
직권말소 여부(이진형)-말소(1)	4%	-	9%	-
업력(개월)	60.09	69.92	32.20	46.89
현금서비스 사용금액(원)	606,908.71	2,869,634.56	1,283,914.11	3,557,882.60
거주기간(개월)	143.14	151.49	116.02	138.08
매출원가금액(원)	14,067,167.15	38,463,621.60	19,710,463.10	151,768,807.00
기타지출금액(원)	2,018,138.08	1,095,229.66	1,829,032.71	1,128,680.36
채무불이행 등록 총 건수(개)	0.01	0.17	0.06	0.28
신용카드 발급일로부터 기간(일)	4,836.07	2,032.80	3,723.33	2,272.79
신용카드 평균 한도소진율(%)	23.67	17.15	37.40	21.67
신용카드 일시불 이용률(%)	62.35	26.98	53.38	27.32
대출 총 기관수(개)	1.63	1.21	2.41	1.50
1년 내 연체 총 건수(개)	0.32	1.25	1.03	2.81
연간원리금상환금액(기준1)(원)	8,737.15	12,592.54	14,836.40	17,068.64
미상환대출총금액 중 캐피탈 & 저축은행업권 비중(%)	13.91	29.10	28.67	36.41

먼저 개인/법인 구분에서 개인의 비율은 우량인 경우 96%, 불량인 경우 92%로 큰 차이가 없지만, 직권말소는 우량에서 4%, 불량에서는 9%로 불량 차주가 2배 이상 높게 나타나고 있다. 이외에도 우량인 차주의 평균이 더 높은 변수는 업력, 거주기간, 기타지출금액, 신용카드 발급일로부터 기간, 6개월 내 신용카드 일시불 이용률이다. 이에 반해 불량인 차주가 더 높은 평균을 나타내는 변수는 현금서비스 사용금액, 매출원가금액, 3개월 내 신용카드 평균 한도소진율, 대출 총 기관수, 1년 연체 총 건수, 부동산담보 제외 연간원리금상환금액, 미상환대출총금액 중 캐피탈 & 저축은행업권 비중으로 나타나고 있다. 이상의 결과에서 알 수 있듯이 우량 차주는 대체적으로 업력과 거주기간이 길며, 금전적으로 다소 여유가 있고 금융 거래 상태가 비교적 건전함을 알 수 있다.

5.2 최종 모형 구축 및 평가

최종 모형은 다음과 같이 구축한다. 세 가지 모형 모두 불량(Y=1)에 대한 확률을 산출하는 것으로써, 의사결정나무모형은 CART 알고리즘, 신경망모형은 다층신경망을 이용하여 은닉층은 1개, 은닉층의 노

드 수는 3개로 하여 소상공인 신용평가모형을 구축한다. 전술하였듯이 모형평가는 10중첩 교차타당법을 사용하며, 예측 성능을 비교하기 위한 평가 측도로는 오분류율, G-Mean, F1측도와 반응률을 이용한다.

10중첩 교차타당법에 의해 산출된 오분류율, G-Mean, F1측도에 대한 결과를 살펴보면 다음과 같다. 첫 번째로 오분류율을 살펴보면 로지스틱회귀모형, 의사결정나무모형, 신경망모형 모두 0.022(=2.2%) 근방으로 큰 차이가 없으며 오분류율이 매우 낮게 나타나고 있다. 두 번째로 G-mean은 로지스틱회귀모형 0.0794, 의사결정나무모형과 신경망모형은 0으로 나타나고 있어 로지스틱회귀모형이 비교적 향상된 예측 성능을 가지고 있다고 할 수 있으나 매우 낮게 나타나고 있다. 마지막으로 F1측도를 살펴보면 G-mean과 거의 유사한 결과를 나타내고 있다. 로지스틱회귀모형에 의한 결과가 0.016로 나타나 다른 두 가지 기계학습 모형에 비해 비교적 좋은 예측력을 가지고 있다고 말할 수 있지만, G-mean과 마찬가지로 전반적으로 F1측도가 매우 낮게 나타났다.

세 가지 평가 측도 중 단순히 오분류율만을 비교하였을 때에는 세 가지 모형 모두 오분류율이 매우 낮아 좋은 예측력을 보인다고 할 수 있다. 그러나 이렇게 단순히 오분류율로만 해석하기에는 자료의 구조 상 큰 문제점이 있다. 그것은 바로 자료의 심각한 계급불균형에 기인한 문제로써 소수계급인 불량 차주 전체를 잘못 분류해도 오분류율은 낮게 된다는 것이다. 본 분석에서도 의사결정나무모형과 신경망모형에서 소수계급인 불량 차주 전체를 우량으로 잘못 분류하고 있다.

계급불균형 자료와 관련하여 Chawla 외(2003)에 의하면 계급불균형 자료에서는 소수계급에 속한 자료를 전혀 분류하지 못하는 문제점이 발생하여 오분류율이 매우 낮은 현상을 보이므로 오분류율은 올바른 모형 평가 측도가 되지 못한다고 언급하고 있다. 예를 들어, 소수계급이 1%이고 다수계급이 99%인 자료를 이용하여 분류할 때 대부분의 자료를 다수계급으로 분류함으로써 낮은 오분류율을 얻게 된다. 이와 같은 계급불균형 자료에 일반적으로 많이 사용하는 오분류율 등의 단순 정확도를 모형 평가 측도로 사용하게 되면 관심의 대상인 소수계급에 속한 자료는 전혀 분류하지 못하는 문제점이 발생한다(박주완, 2010). 그러므로 G-Mean과 F1측도를 이용하여 모형의 판별력을 판단하는 것이 타당하다.

오분류율을 제외한 G-Mean과 F1측도를 비교하였을 때, 의사결정나무모형과 신경망모형으로 소상공인 신용평가모형을 구축할 경우 불량 차주를 전혀 판별하지 못하고 있지만, 로지스틱회귀모형을 이용할 경우 예측력이 향상됨을 확인할 수 있다. 그러나 G-Mean과 F1 측도만으로는 예측력이 매우 우수하다는 결론을 내리는 데에는 값이 크지 않아 한계가 있으므로, 추가적으로 반응률을 이용하여 모형의 예측력을 비교할 필요가 있다.

다음의 <그림 3>과 <표 5>는 반응률 그래프와 반응률을 나타낸 것이다. 그림과 표를 살펴보면 세 가지 모형 모두 반응률이 전반적으로 증가하는 경향을 보이고 있다. 반응률 값을 모형별로 자세히 살펴보면 의사결정나무모형과 신경망모형의 반응률은 2분위와 3분위, 4분위와 5분위, 6분위와 7분위 구간에서 역전현상이 발생하고 있고, 낮은 분위에서 높은 분위로 갈수록 서열화가 잘 이루어지지 않고 있다. 그러나 로지스틱회귀모형의 경우 불량일 사후확

<표 4> 오분류율, G-Mean, F1측도 비교

구분	오분류율	G-Mean	F1측도
로지스틱회귀모형	0.0224	0.0794	0.0160
의사결정나무모형	0.0223	0.0000	0.0000
신경망모형	0.0223	0.0000	0.0000

률이 낮은 구간에서 높은 구간으로 갈수록 실제 불량의 비율이 증가하고 서열화가 잘 이루어지고 있음을 확인할 수 있다. 반응률 비교 결과, 로지스틱회귀모형이 의사결정나무모형이나 신경망모형 보다 좋은 예측 성능을 가지고 있다는 결론을 얻을 수 있다.

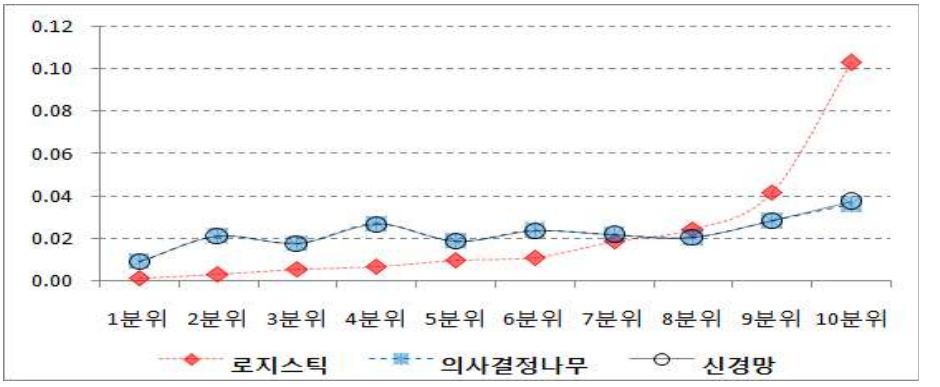
Ⅵ. 결론 및 향후 과제

본 논문은 소상공인 신용평가를 위해 신용보증재단이 보유한 내부정보와 NICE CB요약정보를 이용하여 로지스틱회귀모형, 의사결정나무모형, 신경망모형으로 신용평가모형을 구축하였을 때 예측 성능이 우수한 모형이 무엇인지를 확인하고 시사점을 찾아보는 것이다.

본 논문의 모형 구축 분석 대상은 16개 지역신용보증재단에서 2013년 7월~2016년 12월 보증을 받은 차주 105,572개 중 결측치, 특수값을 제외한 67,308개이다. 구축된 모형의 비교 및 평가는 오분류율, G-mean, F1측도, 반응률을 이용하였으며, 모형 구축 도구는 SAS 9.4와 R을 이용하였다.

소상공인 신용평가모형 구축 결과, 첫째, G-mean, F1측도를 비교한 결과, 로지스틱회귀모형이 비교적 좋은 예측 성능을 가지고 있으며 계급불균형자료에 대해 오분류율을 이용하여 모형을 평가하는 것은 적절하지 않다는 사실을 확인하였다. 둘

<그림 3> 반응률 도표



<표 5> 반응률표

구분	로지스틱회귀모형		의사결정나무모형		신경망모형	
Decile	예측확률	반응률	예측확률	반응률	예측확률	반응률
1분위	0.002	0.001	0.022	0.009	0.022	0.009
2분위	0.004	0.003	0.022	0.021	0.022	0.021
3분위	0.005	0.005	0.022	0.018	0.022	0.018
4분위	0.007	0.007	0.022	0.027	0.022	0.027
5분위	0.010	0.010	0.022	0.019	0.022	0.019
6분위	0.013	0.011	0.022	0.024	0.022	0.023
7분위	0.017	0.018	0.022	0.022	0.022	0.021
8분위	0.025	0.024	0.022	0.020	0.022	0.020
9분위	0.039	0.041	0.022	0.028	0.022	0.028
10분위	0.101	0.103	0.022	0.036	0.024	0.037

째, 반응률을 살펴보았을 때 의사결정나무모형이나 신경망모형 보다 로지스틱회귀모형이 불량일 사후확률이 낮은 구간에서 높은 구간으로 갈수록 실제 불량률의 점차 증가하고 서열화가 잘 이루어지고 있으므로 가장 좋은 예측 성능을 가지고 있다. 그러므로 신용보증재단의 자료를 이용하여 소상공인 신용평가모형을 구축할 경우 로지스틱회귀모형이 가장 좋은 방법으로 사료된다.

이와 같은 분석 결과를 통해 다음과 같은 결론을 내릴 수 있다. 기계학습 기법에서 정확한 예측을 위해서는 소수계급과 다수계급의 특성을 충분히 학습해야 한다. 그러나 계급불균형이 매우 심한 자료는 소수계급의 수가 충분하지 않기 때문에 예측 성능이 높은 모형의 구축이 어렵다는 것이다. 만약 신용평가의 목적이 불량인 차주에게 대출을 해주지 않은 대출 거절이라면, 이와 같은 계급불균형이 매우 심한 자료를 이용할 경우 불량을 대부분 우량으로 판별함으로써 위험관리(risk management)에 문제가 발생할 가능성이 매우 높을 수밖에 없다. 그러므로 본 논문에서 사용한 자료에 대해서는 로지스틱회귀모형을 이용하여 신용평가모형을 구축하는 것이 가장 타당하다는 결론을 내릴 수 있다.

물론 위의 결과는 본 논문에서 사용한 자료를 이용할 경우에 한정되는 것으로써, 다른 자료 이용, 자료의 표준화 방법 등 정제 기법 변경, 신경망모형의 은닉층 및 노드 개수 조정 랜덤포레스트나 SVM모형 등 다른 기계학습모형을 이용할 경우 다른 결과가 나타날 가능성을 배제할 수 없다.

소상공인에 대한 신용평가 연구 사례에서 살펴보았듯이 소상공인 신용평가에 대한 과거 많은 연구들은 객관적인 자료 부족이라는 한계상황으로 인해 재무 자료 이외의 다양한 비재무 자료 활용에 대한 연구가 많이 다루어지고 있다. 그러나 본 논문은 기계학습 기법을 이용하여 소상공인 신용평가모형 구축 가능성을 실증분석을 통해 탐색했다는 점에서 의의가 있다. 그러나 모형을 위한 데이터셋 구축 시 계급불균형 문제의 해결을 위한 오버샘플링(over-sampling) 방법 등의 고려, 세 가지 모형 이외의 기계학습 모형 적용 등이 부족했다는 점에서 한계가 있다. 또한 분석 자료 표준화를 위한 다양한 방법론이나 실제 신용평가모형 구축 시 많이 사용하는 독립변수의 계급화(classing)를 통한 변수 적용 등도 이루어지지 않았다는 한계점도 가지고 있다.

본 논문의 분석을 통한 향후 연구 방향은 다음과 같다. 첫째, 기법적인 측면에서 세 가지 모형 이외에 랜덤포레스트 등 다른 기계학습 방법론을 이용해 볼 필요가 있다. 둘째, 단순히 소상공인 등을 대상으로 한 모형을 구축 보다 소상공인 중에서도 특정 대상, 예를 들어 분석을 위한 정보가 부족한 신규 창업자, 우량과 불량의 구분이 모호한 판단미정 차주 등의 대상을 위한 기계학습 기법 적용 연구도 필요하다. 셋째, 불량 차주의 개수가 충분하다면 기업의 규모 및 업종을 구분한 모형의 구축을 고려해 볼 필요가 있다. 기업의 규모나 업종에 따라 기업의 특성에 차이가 있을 수 있으므로 규모와 업종을 고려한 연구가 필요하다. 넷째, 현재 빅데이터의 적용이 중요한 트랜드(trend)로 나타나고 있으며 빅데이터를 활용하는 결과가 실제 현업에서 일부 적용되어 있다. 그러므로 소상공인 신용평가모형 구축에서도 다양한 빅데이터 적용 가능성과 이를 적용하기 위한 제도적인 방법 등에 대해 연구해 볼만한 가치가 있다. 다섯째, 분석에 사용되는 변수들의 다양한 표준화 기법에 대한 고찰이 필요하다. 실제로 많은 모형 구축 시 분석에 적합하지 않은 결측치, 특이값, 특수값 등의 처리는 모형 구축 시 매우 중요하므로, 어떠한 표준화 방법을 이용하는 경우가 모형 구축에 적당한지에 대한 연구는 필수적이다. 여섯째, 불량 차주의 자료가 불충분한 계급불균형 자료인 경우 본 논문의 결과에서 살펴본 바와 같이 모형 구축 및 평가가 쉽지 않았다. 그러므로 계급불균형인 자료에 대한 모형 구축 방법론에 대해서도 추가적인 실증연구가 필요하다. 마지막으로 신경망모형 등을 이용할 경우 각 변수별로 구간을 나누어 평점화하기 위한 기법 등에서도 논의할 필요가 있다. 로지스틱회귀모형의 경우 평점 산출 로직을 전산적인 신용평가 시스템에 탑재하기 쉽지만, 신경망모형 등의 기계학습의 경우

산출된 결과를 평점화하여 시스템에 탑재하기 위해서는 로지스틱회귀모형 대비 수십에서 수백 배 이상 복잡한 로직이 필요하므로 이를 쉽게 탑재하기 위한 연구는 필수적일 것으로 보인다.

참고문헌

강신형(2016), Alternative Data 머신러닝을 이용한 새로운 평가 방법론, ORANGE REPORT VOL.2, KCB Research Center.

강창완, 강현철, 박우창, 승현우, 윤환승, 이동희, 이성건, 이영섭, 진서훈, 최종후, 한상태(2007), **데이터마이닝-개념과 기법**, 제2판, 사이플러스.

강현철, 한상태, 최종후, 김은석, 김미경(1999), **SAS Enterprise Miner를 이용한 데이터마이닝-방법론 및 활용-**, 자유아카데미.

김명종, 강대기(2010), "부스팅 인공신경망학습의 기업 부실 예측 성과 비교", **한국정보통신학회논문지**, 14(1), 63-69.

김성진, 안현철(2016), "기업 신용등급 예측을 위한 랜덤포레스트의 응용", **산업혁신연구**, 32(1), 187-211.

김성환, 김태동(2014), "신용평가사의 신용등급 고평가에 대한 연구", **회계연구**, 19(3), 27-49.

김승혁, 김종우(2007), "Modified Bagging Predictors를 이용한 SOHO 부도 예측", **지능정보연구**, 13(2), 15-26.

김의중(2016), **알고리즘으로 배우는 인공지능, 머신러닝,** 딥러닝 입문, 위키북스.

박정윤(2000), "재무정책과 기업부실 예측", **재무관리논총**, 6(1), 93-116.

박주완(2010), 로지스틱회귀모형 구축 시 오버샘플링효과에 관한 연구, **동국대학교 대학원 박사학위논문**, 서울.

박주완, 송창길(2015), 인적자원 변수를 이용한 기업신용평가모형 구축에 관한 연구, 인적자본기업패널학술대회.

서울경제신문(2017), "신한카드, 머신러닝 활용한 신용평가시스템 오픈", http://www.sedaily.com/NewsView/1OAXYYX4GJ/,.

성웅현(2016), **응용 로지스틱회귀분석-이론, 방법론, SAS 활용-**, 탐진.

신용보증재단중앙회(2016), 2016 소상공인 금융실태조사 보고서.

신용보증재단중앙회(2017), 2017 소상공인 신용평가모형 구축 최종보고서 - 내부자료.

신윤제(2016), 머신러닝을 활용한 신용평가모형의 개발 - 신용정보 부족군(Thin-File)을 대상으로, NICE Credit Insight Issue Report, NICE평가정보 CB연구소.

윤상용, 강만수, 이형탁(2016), "소상공인 신용평가에서 비재무적 정보는 중요한가", **경영컨설팅연구**, 16(2), 37-46.

윤종식, 권영식(2007), SVM을 이용한 소상공인 부실예측모형, 한국경영과학회 학술대회논문집, pp 826-833.

이건창(1993), "기업 도산 예측을 위한 귀납적 학습지원 인공신경망 접근방법 MDA, 귀납적 학습방법 인공신경망모형과의 성과 비교", **경영학연구**, 23(3), 109-144.

이승현(역)(2014), **데이터 마이닝**, 에이콘.

이명식, 김정인(2007). **개인신용평점제도**, 서울출판미디어.

이영섭, 박주완(2007), "인적자원관련 변수를 이용한 기업신용점수 모형 구축에 관한 연구", **응용통계연구**, 20(3), 1-19.

이영섭 역(2003), **데이터마이닝 Cookbook**, 교우사.

장원경, 김연용(2002), "중소기업에 대한 신용대출 의사결정 시 재무적 정보와 비재무적 정보의 상대적 중요성에 관한 연구", **중소기업연구**, 24(1), 235-255.

전성빈, 김영일(2001), "도산 예측 모형의 예측력 검증", **회계저널**, 10(1), 151-182.

정유석(2003), 인공신경망을 이용한 기업도산예측:IMF후 국내 상장회사를 중심으로, **경희대 대학원 박사학위 논문**, 서울.

조준희, 강부식(2007), "코스닥기업의 도산예측모형에 관한 연구", **산업경제연구**, 20(1), 141-160.

최종후, 진서훈(2005). **데이터마이닝의 현장**, 자유아카데미.

Chawla, N. V., Lazarevic, A., Hall, L. O. and Kegelmeyer, K. W.(2003), SMOTEBoost : Improving Prediction of the Minority Class in Boosting, Proceedings of Principles of Knowledge Discovery in Databases 2003, 107-119.

Hosmer, D. W., Lemeshow, S.(2000), *Applied Logistic Regression* Second Edition, New York: John Wiley and Sons.

Kohavi, R.(1995), A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1137-1143.

Ripley(1996). Pattern Recognition and Neural Networks, ISBN 0-521-46086-7, Cambridge University Press.

Abstract

A Study on the Construction of Credit Evaluation Model for Small Business Using Machine Learning Techniques

Park, Joo-wan* | Song, Chang-gil** | Bae, Jin-sung***

The purpose of this paper is to construct a credit scoring model for small businesses using logistic regression model, decision tree model and neural network model, and to identify the model with the best prediction performance. The data used to build the model are data used by the Korea Credit Guarantee Foundation for credit evaluation. From the data, We removed the missing values and special values, selected variables by statistical procedure, and finally used 15 independent variables and 67,308 borrower's data. The three models were evaluated using the 10-fold cross-validation method, and the error rate, G-mean, F1 measure, and reaction rate were used as the model evaluation measure.

As a result of constructing three models using the data of the Korea Credit Guarantee Foundation and three machine learning techniques, the predictive performance was the best when the logistic regression model was applied. And we found that the prediction performance can be degraded when constructing a machine learning model using class imbalance data.

Key words Small Business Owners, Credit Scoring, Machine Learning, Cross Validation

* Research Fellow, Korea Federation of Credit Guarantee Foundations(jwan0217@koreg.or.kr)

** Researcher, National Pension Research Institute(cgsong82@nps.or.kr)

*** Research Fellow, Korea Federation of Credit Guarantee Foundations(bjs0423@koreg.or.kr)