연구보고서 2019-02

빅데이터 분석 기법을 이용한 소상공인 신용평가 모형 구축 연구

박주완 • 배진성 • 윤혁준

KOREG 신용보증재단중앙회

머 리 말

현재는 4차 산업혁명의 시대라고 일컬어지고 있다. 이 시대의 핵심 기술은 빅데이터와 인공지능으로, 빅데이터를 활용한 기계학습 기반 알고리즘의 성능 향상이 매우 중요한 기술의 원천이다. 인공지능은 인간의 지능으로만 가능했던 인식, 판단, 추론, 문제 해결 등을 컴퓨터가 할 수 있도록 실현하는 기술인데, 이와 같은 일련의 과정을 수행하는 도구가 바로 기계학습이고 이를 구동하기 위해서는 양질의 데이터가 필수적이다.

많은 산업들 중에서 금융 산업은 빅데이터와 인공지능이 가장 빨리 적용되고 가장 널리 활용되는 산업 중 하나로, 실제로 국내외 금융 산업 분야의 현업에서 이를 적용하는 사례들이 점차 증가하고 있는데, 구체적으로 업무 자동화, 대고객 금융서비스, 신용평가, 트레이딩, 준법감시 등이 있다.

특히 신용대출과 관련하여 앞으로 빅데이터와 기계학습을 활용한 신용평가 요구는 점차 늘어날 것으로 예측된다. 그러므로 신용평가모형 구축에 있어 이를 적용하기 위한 노력이 어느 정도인지 고민해 볼 필요가 있다. 그리고 머지않은 미래에 빅데이터와 기계학습을 활용한 신용평가가 보편화 될 때를 대비해 관련 기관에서는 자체적으로 충분한 데이터 확보와 모형 구축을 위한 분석 능력이나 기술력을 보유할 필요가 있을 것으로 사료된다.

본 연구에서는 빅데이터 분석에서 가장 많이 사용되고 있는 기계학습 기법인 의사결정나무모형, 로지스틱회귀모형, 신경망모형, 랜덤포레스트모형, 서포트벡터머신을 이용하여 소상공인 신용평가모형 구축에 대해 연구하였다. 이와 같은 연구를 통해 소상공인 신용평가모형

구축 시 기존의 로지스틱회귀모형 이외에 다양한 기계학습 기법의 적용이 가능함을 살펴볼 수 있었지만, 본 자료의 경우 로지스틱회귀모형이 가장 우수한 성능이 있음을 보였다. 본 연구에서 제시된 결과는 향후 신용보증재단중앙회의 신용평가모형 구축에 많은 도움이 될 것으로 사료된다. 그리고 본 연구 보고서를 이용하여 분석을 수행하고자 하는 연구자 및 실무자를 위해 부록 부분에 소상공인 신용평가모형 구축을 위해 통계 패키지 R 및 SAS로 작성한 데이터 정제 및 분석 프로그램 전체를 수록하였다.

이 연구는 신용보증재단중앙회 교육연구부의 박주완 선임연구위원 책임 하에 배진성 선임연구위원과 윤혁준 연구원이 공동으로 수행하였다. 참여 연구진의 노고와 함께, 연구에 유익한 자문과 소중한 아이디어를 제공해 주신 내·외부 전문가, 이 보고서를 검독하고 소중한 의견을 주신 분들에게 이 자리를 빌려 감사의 뜻을 전한다.

끝으로 본 보고서에 수록된 모든 내용은 어디까지나 저자의 의견이며, 신용보증재단중앙회의 공식 견해가 아님을 밝혀 둔다.

2019년 12월

신용보증재단중앙회 회장 김병근

제목 차례

표 차례

그림 차례

요 약

Ⅰ. 서 론

□ 4차 산업혁명 시대의 핵심 기술은 빅데이터와 인공지능으로, 빅데이터를 활용한 기계학습 기반 알고리즘의 성능 향상이 매우 중요한 기술의 원천임

○ 4차 산업혁명 시대에서 빅데이터나 인공지능이 가장 빨리 적용되고 가장 널리 활용되는 산업 중 하나가 금융 분야임

○ 과거부터 지금까지 금융 관련 데이터는 폭발적으로 증가하고 있으며, 빅데이터에 새로운 가치를 창출할 수 있는 정보 분석에 활용되는 기계학습이 각광을 받게 됨

○ 빅데이터 시대에 금융 리스크 관리 능력 제고, 보안 기술 등의 효과를 높이기 위해서는 기계학습을 이용한 기술 개발과 과감한 투자가 필요

○ 해외에서는 Kabbag, Zest Finance 등의 P2P 대출업체들이 기계학습 기법과 빅데이터를 신용평가에 활용하고 있음

○ 일본의 요코하마은행과 지바은행에서는 인공지능을 이용하여 영세업체 및 개인사업자의 재무정보, 거래 결제정보 및 수익성 예측을 통해 대출 심사 및 금리를 결정

○ 국내에서는 신한카드사가 신용도 판단이 어려운 사회 초년생과 중금리 대출 고객들을 대상으로 2017년 초 기계학습 기법을 적

용한 신용평가시스템 개발을 완료

○ 케이뱅크는 KT의 통신요금 납부 실적, 비씨카드 신용카드 결제 정보를 가지고 자체적인 신용평가시스템을 만들어 중금리 대출 심사에 적용

□ 최근 소상공인을 대상으로 한 대출이 점차 활성화되면서, 소상공인 신용평가의 신뢰성과 정확성에 대한 요구가 커지고 있으며, 신뢰성 높은 자료 확보를 통한 신용평가모형의 중요성이 점차 증가

○ 그러나 소상공인을 대상으로 한 신용평가 연구는 대기업이나 중소기업에 비해 상대적으로 그 수준이나 양적인 면에서 매우 미미한 수준

○ 소상공인에 대한 신용평가 연구가 상대적으로 부족한 이유 중 하나는 신뢰성 있는 분석 자료의 부족이라는 한계에 기인

□ 머지않은 미래에 빅데이터와 기계학습을 활용한 신용평가가 보편화 될 때를 대비해 관련 기관에서는 자체적으로 분석 능력이나 기술력을 보유할 필요가 있을 것으로 판단됨

□ 이에 본 연구는 16개 지역신용보증재단이 보유한 자료와 기계학습 기법을 이용하여 소상공인 신용평가모형을 구축하고 예측 성능이 좋은 모형이 무엇인지를 확인하는 것이 주된 목적임

○ 사용하고자 하는 기계학습 기법은 의사결정나무모형, 로지스틱 회귀모형, 신경망모형, 랜덤포레스트모형, 서포트벡터머신임

□ 본 연구는 과거 비재무 자료 활용이라는 측면의 연구에서 벗어나, 4차 산업혁명 시대의 도래에 맞추어 기계학습 기법을 이용하여 소상공인 신용평가모형 구축의 가능성을 연구한다는 점에서 기존 연구들과 차별성이 있음

Ⅱ. 선행연구 고찰

□ 기계학습 기법 이용 연구 사례 고찰

○ 분류나 예측, 군집과 같은 기술, 모델, 알고리즘을 이용해 문제를 해결하는 것을 컴퓨터과학 관점에서는 기계학습이라 하고, 통계학 관점에서는 데이터마이닝이라고 함

○ 데이터마이닝에서는 전통적인 통계 분석 모형인 군집모형, 회귀모형 등과 기계학습 모형으로 알려져 있는 의사결정나무모형, 신경망모형, 랜덤포레스트 등이 모두 포함되어 설명

○ 다수의 연구 결과를 살펴보면 데이터의 유형에 따라 예측 성능이 우수한 기계학습 기법은 차이가 있음

□ 중소기업 및 소상공인 신용평가 연구 고찰

○ 객관적인 자료가 부족한 중소기업 및 소상공인의 신용평가모형 구축 시 객관성이 담보된 재무적인 요인 이외에도 비재무적 요인은 규모가 작은 기업에 대한 평가에서 매우 중요한 요인임

Ⅲ. 분류를 위한 기계학습 기법

□ 기계학습은 컴퓨터가 사전에 프로그램이 되어 있지 않고 데이터로부터 패턴을 학습하여 새로운 데이터에 대해 적절한 작업을 수행하는 일련의 알고리즘이나 처리 과정을 의미

□ 기계학습은 학습 시스템에 훈련 데이터 입력 형태에 따라 지도학습, 비지도학습과 강화학습(reinforcement learning)으로 구분

○ 지도학습은 입력변수와 목표변수가 존재한다는 것을 말함

○ 비지도학습은 입력변수만 존재하고 목표변수는 없는 경우로 입력 결과에 대한 답이 존재하지 않는 데이터를 학습하는 방법

○ 강화학습은 특정한 환경 안에서 정의된 에이전트가 현재의 상태를 인식하여 선택 가능한 행동들 중 보상을 최대화하는 행동 혹은 행동 순서를 선택하도록 하는 학습 기법

□ 기계학습 기법 개요

○ 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 모형

○ 의사결정나무모형은 의사결정 규칙(decision rule)을 나무 구조로 도표화하여 분류와 예측(prediction)을 수행하는 방법

○ 신경망모형은 인간 뇌 기능에 착안하여 개발된 패턴 인식의 한

분야로 과거의 경험이나 지식을 습득함으로써 오류를 최소화하는 과정으로 어떠한 통계적인 분포도 가정하고 있지 않음

○ 랜덤포레스트는 말 그대로 의사결정나무들이 많이 있는 모형으로, 의사결정나무 모형을 다수 만들어 예측력을 높이는 방법

○ SVM은 지도학습 기법 중 하나로 고차원의 벡터 공간 상에 존재하는 데이터를 가장 잘 분류하는 선 또는 초평면을 찾아 이를 이용하여 분류와 회귀를 수행

Ⅳ. 데이터 정제 및 모형 평가 방법

□ 데이터 정제 및 변수 선택

○ 통계 분석이나 기계학습 기법을 이용한 모형 구축의 성공을 위해서는 다양한 요소들이 있지만, 무엇보다 중요한 것은 질(quality) 좋은 데이터를 양(quantity)적으로 충분히 확보

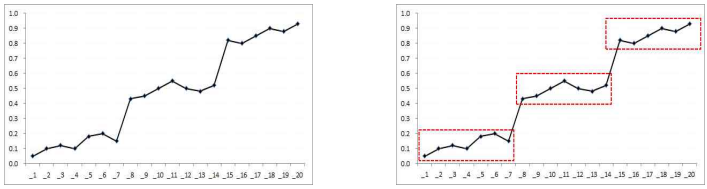
○ 데이터 표준화 기법으로는 최소-최대 표준화, z-스코어 표준화, 소수점으로의 변환에 의한 표준화가 대표적인 방법

○ 본 연구에서는 원천자료의 표준화를 위해 계급화(classing) 기법을 사용하는데, 크게 계급세분화(fine classing)과 성김화(coarse classing) 단계로 구분

- 계급세분화는 원래의 독립변수값을 종속변수인 불량과의 관계 분석을 통해 불량률이 유사한 범주를 하나의 범주로 묶은 후 계급화하여 분석에 사용하는 방법

- 성김화는 1차적으로 계급세분화에 의해 범주형으로 변환된 변수에 대해 동질적인 불량률을 보이는 구간을 재범주화

<요약 그림 1> fine & coarse classing 개요



(a) 계급세분화　　(b) 성김화 – 재범주화

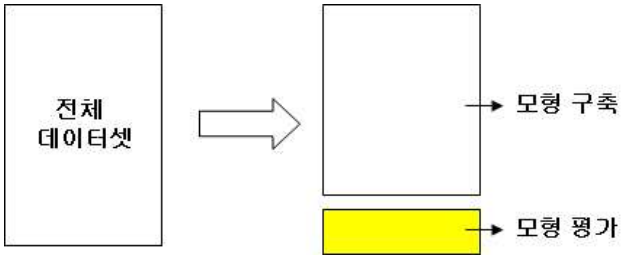
○ 본 연구에서는 신용평가모형 구축을 위한 변수 선택 기법으로 계급화, 단계적 선택법, 스피어만 상관계수에 의한 다중공선성 확인 등을 사용함

□ 모형 평가 방법 및 평가 측도

○ 모형 평가는 구축된 예측 모형이 좋은 예측력을 보유하고 있는가를 확인하기 위한 필수 단계임

- 본 연구에서는 자료의 개수가 충분히 크기 때문에 예비 방법을 이용하여 “훈련용 자료:평가용 자료 = 7:3” 구성

<요약 그림 2> 예비 방법의 개요



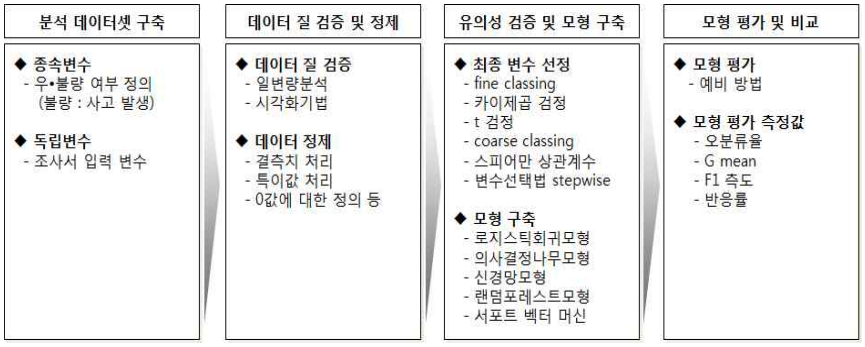
○ 본 연구에서는 오분류율, G-Mean, F1 측도, 반응률로 예측 성능 평가 및 비교에 사용

- 오분류율은 전체 자료를 얼마나 잘못 분류하는가의 문제이므로 값이 작을수록 좋은 모형
- G-mean은 결과 범주가 0인 집단과 1인 집단을 동등하게 고려하는 측도로써 실제 범주가 0인 집단에 대한 정확도와 범주 1인 집단에 대한 정확도의 기하평균
- *F*1 측도(measure)는 어떤 특정한 계급의 성공적인 분류가 훨씬 중요한 경우 사용되는 측정 기준
- 반응률은 훈련용 자료를 이용해 산출된 사후확률을 이용하는데, 사후확률이 가장 큰 구간에서 가장 낮은 구간으로 갈수록 급격하게 감소 또는 그 반대인 경우 좋은 예측 성능 가짐

Ⅴ. 분석 개요

□ 분석 과정은 다음의 그림을 참조

<요약 그림 3> 분석 및 모형 구축 과정



□ 분석 변수는 다음의 표 참조

<요약 표 1> 모형 구축을 위한 변수

변수종류	변수명	변수종류	변수명
종속변수	사고 여부	독립변수	현금서비스금액
독립변수	고객형태		보유부동산
	업종		업력
	종업원수		거주기간
	주사업장소유여부		월평균매출액
	주사업장임차보증 금액		월영업이익
	주사업장월세 금액		월배우자소득
	실거주지소유여부		월기타수익
	실거주지임차보증 금액		소유부동산금액
	실소유지월세 금액		임대보증금사업장
	차입금운전		임대보증금주택
	차입금시설		예적금금액
	차입금기타		유가증권금액
	기보증잔액재단		재고자산
	기보증잔액신보		고정자산
	기보증잔액기보		권리금
	기보증잔액개인		기타현금
	담보제외차입기관수		직권말소

□ 분석 변수 기초 분포

○ 종속변수인 사고 여부의 분포는 사고가 전혀 발생하지 않은 차주(사고 무)는 전체 분석 대상 중 98.1%(133,566개)이고 사고가 한 번이라도 발생한 차주(사고 유)는 1.9%(2,623개)로 계급불균형 자료(class imbalanced data)임

○ 차입금 시설, 차입금 기타, 기보증잔액 재단, 기보증잔액 신보, 기보증잔액 기보, 기보증잔액 개인, 현금서비스 금액, 월배우자소득, 월기타수익, 임대보증금사업장, 임대보증금주택, 예적금금액, 유가증권금액, 재고자산, 고정자산, 권리금, 기타현금은 0 또는 결측치의 비율이 90%를 상회하고 있음을 확인

○ 결측치 및 0값이 많다는 자료상의 한계점이 존재함에도 불구하고 모든 변수를 모형 구축에 활용

- 이유는 계급화(classing) 기법을 사용하여 변수 값을 표준화할 경우 결측치나 0값은 불량률의 기초하여 하나의 계급으로 표준화가 가능하며, 모형 구축을 위한 변수를 선택하는 다양한 분석 과정을 통해 통계적으로 우불량 여부에 유의미하지 않은 변수는 배제되기 때문임

Ⅵ. 분석 결과

□ 변수 선택

○ 계급 세분화, 카이제곱 검정, t 검정을 이용하여 1차적으로 변수를 선택한 결과 23개가 선정됨

- 고객 형태, 업종, 주사업장 소유 여부, 주사업장 임차보증 금액, 주사업장 월세 금액, 실거주지 소유 여부, 실거주지 임차보증 금액, 실소유지 월세 금액, 차입금 운전, 기보증잔액 재단, 기보증잔액 기보, 담보 제외 차입 기관 수, 현금서비스 금액, 보유 부동산, 업력, 거주 기간, 월평균 매출액, 월영업이익, 소유부동산 금액, 임대보증금 사업장, 임대보증금 주택, 재고자산, 직권말소

○ 성김화, 단계적 선택법, 다중공선성 확인 후 최종 변수 선택 결과 16개의 독립변수가 선정됨

- 고객 형태, 업종, 주사업장 임차보증 금액, 실거주지 소유 여부, 실소유지 월세 금액, 차입금운전, 기보증잔액 재단, 기보증

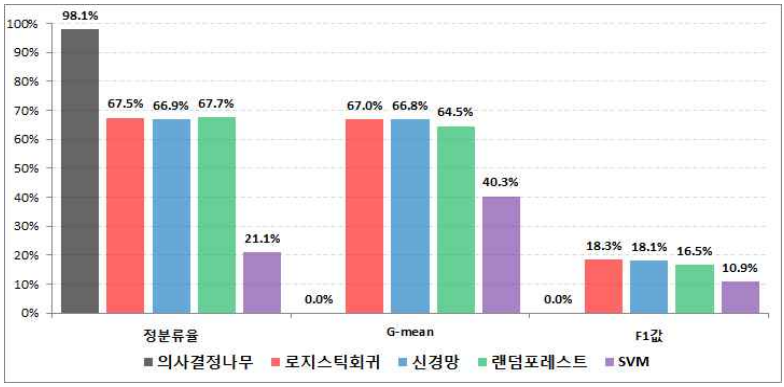
x **요약**

잔액 기보, 담보 제외, 차입 기관 수, 현금서비스 금액, 보유 부동산, 업력, 거주기간, 월평균 매출액, 재고자산, 직권말소

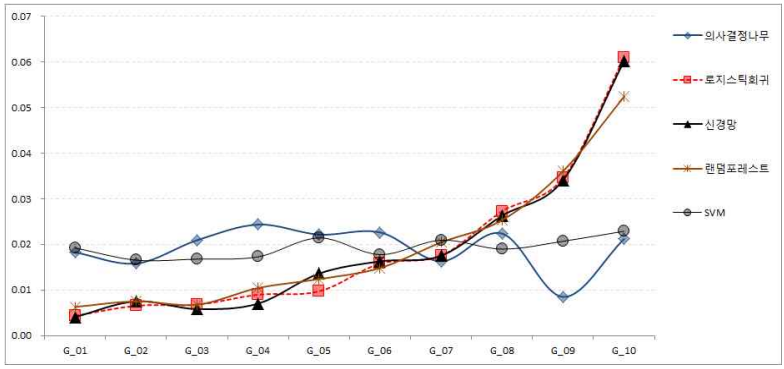
□ 모형 구축 평가 결과

○ 정분류율, G-mean, F1값, 반응률을 비교한 결과, 로지스틱회귀 모형이 분류를 위한 예측 성능과 안정성 측면에서 가장 우수한 신용평가모형임

<요약 그림 4> 정분류율, G-mean, F1값 비교



<요약 그림 5> 반응률 비교



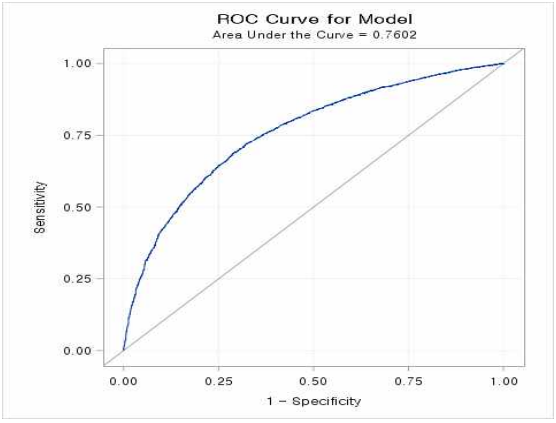
□ 최종적인 소상공인 신용평가모형 구축

○ 정분류율, 오분류율, c통계량, 반응률, ROC 곡선을 이용하여 분류 예측 성능을 평가한 결과 구축된 로지스틱회귀모형의 예측 성능이 매우 우수하다는 결론을 내릴 수 있음

<요약 표 2> 최종 로지스틱회귀모형 구축 결과

변수	더미	회귀계수	표준오차	왈드 카이제곱값	p-값	오즈비
절편		-2.389	0.400	35.77	<.0001	
고객형태	1	0.691	0.088	60.99	<.0001	2.00
업종	1	0.231	0.049	22.50	<.0001	1.26
	2	0.558	0.193	8.41	0.0037	1.75
주사업장임차보증금액	1	0.124	0.050	6.11	0.0135	1.13
실거주지소유여부	1	0.325	0.075	18.59	<.0001	1.38
실소유지월세 금액	1	0.309	0.055	31.05	<.0001	1.36
차입금운전	1	0.479	0.176	7.38	0.0066	1.61
기보증잔액재단	1	0.230	0.056	16.66	<.0001	1.26
기보증잔액기보	1	0.845	0.249	11.53	0.0007	2.33
담보제외차입기관수	1	0.659	0.066	100.08	<.0001	1.93
	2	1.185	0.070	290.81	<.0001	3.27
현금서비스금액	1	0.661	0.063	109.42	<.0001	1.94
보유부동산	1	0.453	0.083	29.62	<.0001	1.57
	2	0.692	0.079	76.07	<.0001	2.00
업력	1	0.708	0.052	184.26	<.0001	2.03
	2	1.184	0.091	167.78	<.0001	3.27
거주기간	1	0.223	0.050	19.77	<.0001	1.25
월평균매출액	1	0.379	0.053	50.74	<.0001	1.46
	2	0.659	0.086	58.44	<.0001	1.93
재고자산	1	0.701	0.148	22.52	<.0001	2.02
직권말소	1	1.027	0.261	15.55	<.0001	2.79
정분류율		**76.1%**				
오분류율		**23.9%**				
c통계량		**0.76**				
호스머-램쇼 적합도 검정		**13.70(0.09)**				

<요약 그림 6> ROC 곡선



○ 최종 모형 구축 결과 6개 등급으로 구분이 가능한데, 산출된 평점을 이용하여 계급세분화를 수행한 결과 평점이 낮아짐에 따라, 즉 신용등급이 나빠짐에 따라 불량률은 계속 증가하며 역전현상은 발생하지 않음

- 최종 신용등급은 1등급으로 994.7 초과, 2등급 990.4 초과, 3등급 983.9 초과, 4등급, 971.4 초과, 5등급 958.2 초과, 6등급 958.2 이하로 구분 가능

<요약 표 3> 최종 등급화

등급	구간	등급별 점수 구간		우량차주수	불량차주수	불량률
1등급	G_09, G_10	994.7	999.2	18,977	58	0.30%
2등급	G_07, G_08	990.4	994.7	18,912	111	0.58%
3등급	G_05, G_06	983.9	990.4	18,754	236	1.24%
4등급	G_03, G_04	971.4	983.9	18,678	421	2.20%
5등급	G_02	958.2	971.4	9,155	366	3.84%
6등급	G_01	557.9	958.2	8,867	640	6.73%

Ⅶ. 결론

□ 다양한 기계학습 기법을 이용하여 소상공인 신용평가모형을 구축한 결과를 요약하면 다음과 같음

○ G-mean, F1 측도를 살펴보았을 때 로지스틱회귀모형이 가장 우수한 예측 성능을 가지고 있음

○ 계급불균형 자료에 대해 오분류율을 이용하여 모형을 평가하는 것은 적절하지 않다는 사실을 확인

○ 반응률을 살펴보았을 때 의사결정나무모형, 신경망모형, 랜덤포레스트모형, SVM모형 보다 로지스틱회귀모형이 불량일 사후확률이 낮은 구간에서 높은 구간으로 갈수록 실제 불량률의 점차 증가하고 서열화가 잘 이루어지고 있으므로 가장 좋은 예측 성능을 가지고 있음

□ 결론적으로 신용보증재단의 자료를 이용하여 소상공인 신용평가모형을 구축할 경우 로지스틱회귀모형이 가장 좋은 방법임

Ⅰ. 서 론

현재는 4차 산업혁명(the fourth industrial revolution)의 시대라고 일컬어지고 있는데, 이는 2016년 다보스 세계경제포럼에서 언급된 이후 전 세계적으로 큰 관심을 받기 시작하였다. 4차 산업혁명은 정보통신기술 융합을 통한 기술혁신의 시대로 빅데이터(big data)를 통해 물리, 생물, 디지털 등의 다양한 분야를 통합하여 새로운 지식과 가치를 창출하는 것이 가능한 시대를 말한다.

4차 산업혁명 시대의 핵심 기술은 빅데이터와 인공지능(artificial intelligence, AI)으로, 빅데이터를 활용한 기계학습(machine learning) 기반 알고리즘의 성능 향상이 매우 중요한 기술의 원천이다. 여기에서 빅데이터는 기존의 데이터베이스(database) 관리 도구로 데이터를 수집·저장·관리·분석할 수 있는 역량을 넘어 대용량의 정형 및 비정형 데이터의 수집을 통해 가치를 창출하고 결과를 분석하는 것을 의미한다. 그리고 인공지능이란 그동안 인간의 지능으로만 가능했던 인식·판단·추론·문제 등의 해결을 컴퓨터가 할 수 있도록 실현한 기술인데, 이와 같은 일련의 과정을 수행하는 도구가 기계학습이다.

4차 산업혁명 시대에 접어든 이후 빅데이터나 인공지능이 가장 빨리 적용되고 가장 널리 활용되는 산업 중 하나가 금융 분야이다. 과거부터 지금까지 금융 관련 데이터는 폭발적으로 증가하고 있으며, 빅데이터에 새로운 가치를 창출할 수 있는 정보 분석에 활용되는 기계학습이 각광을 받게 되었다. 빅데이터 시대에 금융 리스크 관리 능력 제고, 보안 기술 등의 효과를 높이기 위해서는 기계학습을 이용한 기술 개발과 과감한 투자가 필요하다.

실제로 금융 산업에서는 빅데이터와 기계학습 등을 활용하기 위해

많은 연구와 노력을 하고 있으며, 국내외 금융 산업 분야의 현업에서 이를 적용하는 사례들이 점차 증가하고 있는데, 구체적으로 업무 자동화, 대고객 금융서비스, 신용평가, 트레이딩, 준법감시 등의 업무에 기계학습 기법을 도입하고 있다(김효진, 2018).

이와 같이 금융 산업 내의 다양한 업무에 기계학습 기법을 도입 후 적용하는 사례 중 신용평가 부문에 적용되는 국내외 사례를 살펴보면 다음과 같다. 먼저 해외에서 빅데이터와 기계학습을 신용평가에 적용한 사례이다. 해외에서는 Kabbag, Zest Finance 등의 P2P 대출업체들이 기계학습 기법과 빅데이터를 신용평가에 활용하고 있다. Kabbage는 소상공인 신용평가 시 기존의 재무적인 자료 이외의 배송, 회계, 인터넷 자료 등을 기계학습 기법에 적용하여 소상공인의 신용평가를 수행하고 있다. 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). 그리고 일본의 요코하마은행과 지바은행에서는 인공지능을 이용하여 영세업체 및 개인사업자의 재무정보, 거래 결제정보와 수익성 예측을 통해 대출 심사 및 금리를 결정하고 있다(김효진, 2018).

국내에서는 신한카드사가 신용도 판단이 어려운 사회 초년생과 중금리 대출 고객들을 대상으로 2017년 초 기계학습 기법을 적용한 신용평가시스템 개발을 완료하였다(서울경제신문, 2017). 케이뱅크는 빅데이터를 신용평가에 도입해 효과를 거두고 있다. 구체적으로 살펴보면 KT의 통신요금 납부 실적, 비씨카드 신용카드 결제 정보를 가지고 자체적인 신용평가시스템을 만들어 중금리 대출 심사에 적용하였는데, 통신 요금이나 단말기 대금 납부 실적, 통신 요금제 수준, 로밍 횟수 등의 데이터를 신용평가모형에 적용하였고, 그 결과 시중 은행

보다 연체율이 낮아진 효과를 거두고 있다고 발표하였다. 또한 가계나 자영업자의 신용대출 심사 시에도 카드 가맹점 정보를 활용해 평가를 수행하는데, 신용평가가 어려운 자영업자의 경우 비씨카드의 빅데이터에서 제공하는 매출 수준을 평가해 신용대출을 제공하는 방식이다. 카카오뱅크는 통신요금 납부 내역을 확인해 성실한 납부가 이어지는 경우 고정 수입이 확인되지 않더라도 금리 혜택을 주고 있으며, 대출고객을 대상으로 카카오택시 탑승기록 등을 수집하여 빅데이터 시스템을 구축하는 중이라고 밝히고 있다. 이와 같은 사례들을 통해 신용평가 시 빅데이터와 기계학습 기법 이용에 대한 관심과 중요성이 점차 높아지고 있음을 유추할 수 있다.

<표 Ⅰ-1> 기계학습 도입 사례 및 내용

업무	회사명	내용
신용평가	요코하마은행, 지바은행	인공지능 이용 영세업체 및 개인 사업자의 재무정보, 거래 결제정보 및 수익성 예측을 통해 대출 심사 및 금리 결정
	신한카드	신용도 판단이 어려운 사회 초년생과 중금리 대출 고객들을 대상으로 2017년 초 기계학습 기법을 적용한 신용평가시스템 개발
	Zest Finance	전통적인 신용정보 외에 직장정보, 고정수입, 인터넷 포스팅 내용 등이 포함된 7만개가 넘는 변수를 가진 빅데이터에 10개의 기계학습 모형을 적용하여 신용평가 수행
	Kabbage	소상공인 신용평가 시 기존의 재무적인 자료 이외의 배송, 회계, 인터넷 자료 등을 기계학습 기법에 적용
	케이뱅크	KT 통신요금 납부 실적, 비씨카드 신용카드 결제정보 등 빅데이터를 이용해 신용평가시스템 구축 후 중금리 대출 심사에 적용
	카카오뱅크	통신요금 납부 내역을 확인해 성실한 납부인 경우 고정 수입이 없더라도 금리 혜택 대출 고객 대상 카카오택시 탑승 기록 등을 수집하여 빅데이터 시스템 구축

업무	회사명	내용
업무 자동화	골드만삭스	자연어 분석을 통해 경제 동향 학습 후 보고서 자동 생성 시스템 구축
	AXA다이렉트	보험 가입자들의 자동차 사고 예측을 위한 기계학습 기법 도입
	JP모건체이스	COiN(Contract Intelligence) 도입으로 법률 문서에서 주요 조항 추출
	도쿄증권거래소	기계학습 기반 이상 거래 탐지 시스템 도입
	신한은행	딥러닝 기반 이상 금융 거래 탐지 시스템 도입
고객 서비스	미즈호 은행	IBM 왓슨 기반 콜센터 고객 대응 업무 수행
	뱅크 오브 아메리카	페이스북 메신저 기반의 챗봇 Erica 출시
	신한카드	소비 관리 서비스 FAN 페이봇으로 개인 맞춤형 서비스 제공
	삼성카드	딥러닝 기반 스마트 비주얼 시스템을 도입하여 고객의 결제 정보 및 주변 상권 정보로 맞춤형 할인 및 포인트 적립 혜택 제공
트레이딩	뱅가드	2015년 5월 고객 정보를 분석하고 자산을 운용하는 로보어드바이저 'Vanguard Personal Advisor Services'를 출시
	로열뱅크 오브 스코틀랜드	IBM의 왓슨 기반 로보어드바이저 및 챗봇인 '루보' 도입하여 고객의 자산, 투자 성향 정보를 학습하여 맞춤형 금융 상품 추천
	우리은행, KB금융, IBK	고객 정보를 분석하여 자산 배분을 수행하는 로보어드바이저 제공
준법 감시	홍콩증권 선물위원회	20개의 은행 데이터 분석 개선을 위한 파일럿 테스트 수행
	Droit	파생상품의 적법 여부 판단

출처 : 김효진(2017)과 신윤재(2016)의 내용을 재구성

최근 소상공인을 대상으로 한 대출이 점차 활성화되면서, 소상공인 신용평가의 신뢰성과 정확성에 대한 요구가 커지고 있으며, 신뢰성 높은 자료 확보를 통한 신용평가모형 구축의 중요성이 점차 증가하고 있다. 그러나 윤상용 외(2016)는 소상공인을 대상으로 한 신용평

가 연구는 대기업이나 중소기업에 비해 상대적으로 그 수준이나 양적인 면에서 매우 미미한 수준이라고 언급하고 있다. 이유는 기존의 신용평가모형에 대한 연구는 주로 재무적인 자료가 충분한 어느 정도 규모가 있는 기업들을 대상으로 이루어지고 있기 때문이다.

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 부족하여 통계적인 기법을 적용한 신용평가 연구가 어려운 것이 현실이다(신용보증재단중앙회, 2016).

이로 인해 기존에 소상공인을 대상으로 한 신용평가모형 구축 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(윤종식과 권영식, 2007). 이외에도 개인정보보호 등 정보의 사용 제약과 전통적 신용평가 기법의 고착화 등으로 인하여 새로운 기계학습 기법을 적용한 연구는 한계가 많다(신윤재, 2016).

그러나 전술하였듯이 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 활용한 신용평가가 보편화 될 때를 대비해 관련 기관에서는 자체적으로 충분한 데이터 확보와 모형 구축을 위한 분석 능력이

나 기술력을 보유할 필요가 있을 것으로 사료된다.

본 연구는 이러한 문제의식 하에서 16개 지역신용보증재단이 보유한 자료와 기계학습 기법을 이용하여 소상공인 신용평가모형을 구축하고 예측 성능이 좋은 모형이 무엇인지를 확인하는 것이 주된 목적이다. 사용하고자 하는 기계학습 기법은 의사결정나무모형(decision tree), 로지스틱회귀모형(logistic regression)[1], 신경망모형(neural network), 랜덤포레스트모형(random forest), 서포트벡터머신(support vector machine, SVM)이고, 모형의 평가 방법 및 측도(measure)는 예비 방법(holdout method)과 오분류율, G-mean, F1 측도, 반응률(percent of response)이다. 자료를 분석하고 모형을 구축하기 위한 통계프로그램은 SAS9.4와 R3.5.1 버전을 이용한다. 이 때 SAS로는 자료의 기초분석 및 데이터 정제(data cleaning), 모형 평가 작업을 수행하고, R은 "glm, rpart, nnet, rf, kernlab" 라이브러리를 이용하여 기계학습을 수행한다.

본 연구는 과거 비재무 자료 활용이라는 측면의 연구 틀에서 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. Ⅱ장에서는 신용평가모형 관련 기존 연구 사례들을 고찰하고, Ⅲ장에서는 예측 모형 구축을 위한 알고리즘, Ⅳ장은 데이터 정제와 모형 평가 방법을 설명한다. Ⅴ장은 소상공인 신용평가모형을 구축하기 위한 연구 대상, 변수 선정, 모형 구축 과정에 대해 설명하며, Ⅵ장에서는 모형을 구축하여 예측 성능을 비교 및 평가 후 최종 신용평가모형을 구축한다. 마지막으로 Ⅶ장에서는 결론 및 향후 과제에 대해 고찰한다.

---

1) 신용보증재단에서는 소상공인 신용평가모형 구축 시 로지스틱회귀모형을 사용하고 있다.

Ⅱ. 선행 연구 고찰

과거에는 전통적인 재무 정보 기반의 기업 부도 예측이나 신용평가에 대한 통계 방법론 적용 연구는 다변량판별분석(Altman, 1968)과 로짓모형(Ohlson, 1980)으로 대표되었는데, 데이터마이닝 기법 등이 연구된 이후 신경망모형, SVM 등 다양한 기법들을 적용하여 예측 모형의 성과를 높이는 방향으로 발전하여 왔다(강신형, 2016).

본 연구에서는 다양한 기계학습 기법을 이용하여 소상공인의 신용평가 모형을 구축하고 비교를 수행하는 것이 주요 목적이기 때문에 신용평가모형 구축 시 다양한 기계학습 알고리즘을 비교한 연구 문헌 위주로 고찰하고자 한다. 그리고 부가적으로 비교적 기업의 규모가 작고 신용평가모형 구축을 위한 데이터의 양과 질적인 면에서 부족하다고 알려진 중소기업 및 소상공인의 신용평가 방법론에 대한 기존 연구 사례에 대해서도 알아보고자 한다.

1. 기계학습 기법 이용 연구 사례 고찰

기계학습은 종종 데이터마이닝과 혼용되고 있는데, 이는 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, 랜덤포

레스트 등을 모두 포함하여 설명하고 있다(김의중, 2016).

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 가장 우수하다고 하였다. 전성빈과 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(2007)은 여러 가지 기계학습 기법을 이용하여 코스닥기업의 도산 예측 모형을 구축한 결과 의사결정나무모형이 신경망모형이나 로지스틱회귀모형 보다 좋은 예측 성능을 가지고 있다고 하였다. 박주완과 송창길(2015)에서는 인적자본기업패널자료와 NICE자료를 이용하여 소기업 이상에 대해 로지스틱회귀모형, 신경망모형, 의사결정나무모형을 이용하여 신용평가모형 구축 결과 로지스틱회귀모형의 예측성능이 더 우수함을 실증분석을 통해 검증하였다. 윤종식과 권영식(2007)은 소상공인 부실예측모형 연구에서 로지스틱회귀모형, 다변량판별분석, CART, C5.0, 신경망 모형, SVM모형의 예측 성능을 비교한 결과, SVM모형의 예측 성능이 가장 우수함을 보였다. 박주완 외(2017)은 지역신용보증재단의 자료와 로지스틱회귀모형, 의사결정나무

모형, 신경망모형을 이용하여 3개의 모형을 구축한 결과, 로지스틱회귀모형을 적용했을 경우 예측 성능이 가장 우수하며, 계급불균형인 자료를 이용하여 기계학습 모형 구축 시 예측 성능이 저하될 수 있다는 사실을 발견하였다.

마지막으로 앙상블 기법을 이용한 연구 논문이다. 김승혁과 김종우(2007)는 SOHO 부도예측에 있어서 부스트랩(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. 김명종과 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. 김성진과 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.

2. 중소기업 및 소상공인 신용평가 연구 고찰

신용평가 기관에서는 신경망 등의 기계학습 기법과 다양한 자료 활용을 통해 신용평가의 정확도와 신뢰성 향상을 위해 많은 연구를 하고 있다. 모형 구축 시 자료의 활용 측면에서는 재무적 요인 이외에 비재무적 요인을 고려하여 신용평가를 수행하고 있는데, 회계정보의 신뢰성이 낮은 중소벤처기업 등의 경우 정성적인 특성이 많은 비재무적 요인의 적용은 매우 중요하다(장원경과 김연용, 2002).

이러한 중요성에도 불구하고 비재무적 요인은 표준화 및 시스템화의 속도가 느린 편이고 여전히 평가자의 경험이나 역량에 많은 영향을 받고 있는 실정이다(이주민 외, 2007). 그리고 비재무적 요인은 항목의 방대성과 주관성으로 객관화와 계량화가 어렵고, 실무 적용을 위한 시스템과 인력 등 인프라 구축 등에 어려움이 많아, 체계적인 연구 및 실무에서의 활용도가 다소 미흡하며 통계 모형 등의 적용은 매우 드물고 쉽지 않다(유원종과 이철규, 2013).

실제로 소상공인 신용평가에 대한 연구는 주로 비재무적인 자료를 이용하는 등 사용 가능한 자료 활용에 대한 연구가 주를 이루고 있으며, 비재무적인 자료를 바탕으로 한 기법 측면의 연구는 심사역의 경험에 의존하여 주관적으로 점수를 산출하는 AHP(Analytic Hierarchy Process) 기법에 대한 연구가 주를 이루고 있고, 실무에도 적용되고 있다(윤종식과 권영식, 2007).

비재무적인 요인들을 신용평가에 적용한 연구 사례는 다음과 같다. Altman 외(2010)는 중소기업 신용평가 시 재무 및 비재무 항목을 기업도산의 예측변수로 동시에 사용할 때 모형의 예측 정확성이 13% 증가하였으며, 적정한 재무적 요인이 부족한 중소기업에 있어서 비재무적 요인의 가치는 더욱 중요하다고 하였다. 또한 Bhimani 외(2013)

는 비재무적 요인과 거시경제 지표들을 재무적 요인과 함께 평가하는 경우 부도예측 모형의 예측력을 25% 개선할 수 있다고 하였다. 그리고 이주민 외(2007)는 신용평가에서 비재무적 요인이 가지는 중요도를 34%로 제시하였으며, 이러한 수치는 신용평가등급을 한 등급 높이거나 낮출 수 있을 정도로 중요한 수준이라고 언급하였다.

이외에도 정성적인 설문조사 자료를 이용하여 기업의 신용평가모형에 대해 연구한 사례도 있다. 이영섭과 박주완(2007)은 한국직업능력개발원의 인적자본 기업패널(Human Capital Corporate Panel, HCCP) 설문조사 자료에서 기업의 인적자원관리와 개발과 관련한 설문문항을 이용하여 신용평가모형 구축을 시도하였다. 그 결과 기업 인적자원 관련 설문 문항을 이용해 기업 신용평가모형 구축이 가능하다는 결론을 내리고 있다.

박주완과 송창길(2015)은 NICE신용평가(주)에서 제공하는 재무변수와 2013년 인적자본 기업패널 설문조사 자료를 모두 이용하여 이를 로지스틱회귀모형에 적합한(adjusted) 기업 신용평가모형을 구축하였다. 연구 결과 설문 자료를 이용한 비재무적 요소인 인적자원 관련 변수들은 기업성과에 많은 영향을 주고 있는 것은 사실이지만, 기존의 재무변수를 이용한 기업 신용평가모형에 설문조사 변수를 정량화된 변수로써 직접 적용하는 것은 좋은 방법이 아닐 수 있으므로, 자료의 출처 및 유형별로 신용평점을 산출하고 이에 적절한 가중치를 적용하여 최종적인 평점을 구하는 방법 등을 모색하는 것이 타당하다고 주장하였다.

마지막으로 박주완(2018)은 소상공인의 신용평가를 위해 재무변수 등 객관적인 정보를 이용하지 않고 정성적인 자기기입식 설문조사 자료만을 이용하여 로지스틱회귀모형으로 신용평가모형의 구축 가능성 여부를 확인하였다. 모형 구축 결과 최종적으로 선택된 변수는 월세,

성장단계, 창업동기, 창업 전 직업, 경영상 애로사항, 현재 자금 운용 상황, 주된 차입기관, 대출 거절 이유, 평균 매출액, 업력 10개였다. 그리고 구축된 모형을 평가하기 위하여 10중첩 교차타당법을 통한 반응률을 확인한 결과, 불량일 사후확률이 낮은 구간에서 높은 구간으로 갈수록 불량률이 점차 증가하고 서열화가 잘 이루어지고 있음을 확인하였다. 결론적으로 소상공인 신용평가모형을 구축하기 위해 정성적인 설문조사 자료를 이용하는 것이 가능하다는 결론을 내리고 있으며, 신뢰성과 정확성이 높은 신용평가모형 구축을 위해서는 차주의 설문 응답이 성실해야 한다는 선행조건이 충족되어야 한다고 주장하였다.

앞에서 살펴본 연구사례들을 통해 객관성이 담보된 재무적인 요인 이외에도 비재무적 요인은 규모가 작은 기업에 대한 평가에서 매우 중요한 요인임을 확인할 수 있다.

Ⅲ. 분류를 위한 기계학습 기법

1. 기계학습의 개요

아서 사무엘(1959)은 “기계학습(machine learning)은 명시적인 프로그래밍 없이 컴퓨터가 학습하는 능력을 갖추게 하는 것”이라고 정의하였으며, 톰 미첼은 “어떤 작업 T에 대한 컴퓨터 프로그램의 성능을 P로 측정했을 때 경험 E로 인해 성능이 향상됐다면, 이 컴퓨터 프로그램은 작업 T와 성능 측정 P에 대해 경험 E로 학습한 것이다.”라고 기계학습을 공학적으로 풀이하였다(김효진, 2018). 즉, 기계학습은 컴퓨터가 사전에 프로그램이 되어 있지 않고 데이터로부터 패턴을 학습하여 새로운 데이터에 대해 적절한 작업을 수행하는 일련의 알고리즘이나 처리 과정을 의미한다(오미애 외, 2017).

기계학습 개념을 이해하기 위해서는 학습에 대한 개념의 이해가 선행되어야 한다. 실무적인 관점에서 학습은 표현(representation), 평가(evaluation), 최적화(optimization)의 합이다. 표현은 입력값을 처리해서 어떻게 결과를 유도하는지를 결정하는 방법이며, 평가는 업무를 잘 수행했는지를 판정하는 방법이고, 최적화는 평가 기준을 최적으로 만족하는 조건을 찾는 과정을 의미한다. 세 가지 과정이 끝났을 때 이를 학습이 완료되었다고 표현하며, 여러 가지 방법에 의해 학습이 완료된 후, 새로운 데이터에 대한 예측을 하는 것을 일반화(generalization)라고 한다. 예를 들어 설명하면, 표현은 컴퓨터가 글자를 인식하는 학습을 진행할 때, 글자를 인식하고 분류하는 논리모형이다. 평가는 논리모형에 의한 학습을 통해 인식한 글자가 어떤 형태에 가까운지를 나타내는 확률이라고 할 수 있다. 최적화는 글자를 인식하는 논리 모

형의 정확도 향상을 위해 사용된 가중치를 수정하여 최적의 가중치를 결정하는 것이다(김의중, 2016).

기계학습은 학습 시스템에 훈련 데이터를 입력하는 형태에 따라 지도학습(supervised learning), 비지도학습(unsupervised learning)과 강화학습(reinforcement learning)으로 나뉜다. 지도학습은 입력변수(input variable)와 목표변수(target variable)가 존재한다는 것을 말하는데, 통계학에서 입력변수는 독립변수(independent variable), 목표변수는 종속변수(dependent variable)라고 한다. 지도학습은 예측을 위한 입력변수와 각 개체에 대한 출력 결과인 목표변수가 하나의 세트가 된 훈련 데이터를 이용해 올바른 답이 나오도록 컴퓨터를 학습시키는 방법이다. 전형적인 지도학습으로는 분류와 회귀가 있다.

비지도학습은 입력변수만 존재하고 목표변수는 없는 경우로 입력 결과에 대한 답이 존재하지 않는 데이터를 학습하는 방법이다. 즉, 정답이 없이 목표만 주어지는 학습 기법으로 데이터 속에 있는 일정한 패턴이나 규칙을 추출하는 것이 목적이다. 군집분석(cluster analysis), 주성분 및 요인분석(principal component & factor analysis), 연관성 분석(association analysis) 등이 대표적인 비지도학습이다.

강화학습은 기계학습의 한 영역으로 어떤 환경 안에서 정의된 에이전트(agent)가 현재의 상태를 인식하여 선택 가능한 행동들 중 보상을 최대화하는 행동 혹은 행동 순서를 선택하도록 하는 학습 기법을 말한다. 예를 들면 강아지를 훈련시킬 때 잘한 행동에 보상을 주고 잘못된 행동에 벌칙을 주는 것과 같은 방식의 학습 기법이다. 강화학습은 입출력 쌍으로 이루어진 훈련 데이터를 제시하지 않으며 잘못된 행동에 대해서도 명시적으로 정정이 일어나지 않는다는 점에서 일반적인 지도학습과는 차이가 있다.

2. 기계학습 기법

기계학습 관점에서 신용평가모형의 개념을 살펴보면 다음과 같다. 차주의 우불량 여부를 판별하고 신용도를 예측하기 위한 신용평가모형은 기계학습 관점에서 지도학습 중에서 분류(classification) 모형이다.

전술하였듯이 지도학습이란 독립변수와 종속변수의 쌍이 주어진 훈련용 자료(training data)로부터 예측이나 분류를 수행하는 것을 말한다. 그러므로 지도학습을 위한 자료에는 종속변수와 독립변수가 필요하다. 대표적인 지도학습 모형으로는 선형회귀모형(linear regression), 로지스틱회귀모형(logistic regression), 의사결정나무모형(decision tree), 신경망모형(neural network), 랜덤포레스트모형(random forest), SVM (support vector machine) 등이 있다(박주완, 2017).

이 중에서 본 연구에서 사용할 분류 모형 구축 알고리즘은 보편적으로 많이 알려져 있고 비교적 사용이 용이한 로지스틱회귀모형, 의사결정나무모형, 신경망모형, 랜덤포레스트모형, SVM 5가지이다. 본 절에서는 연구에 사용된 5가지 모형에 대해 고찰해보고자 한다.

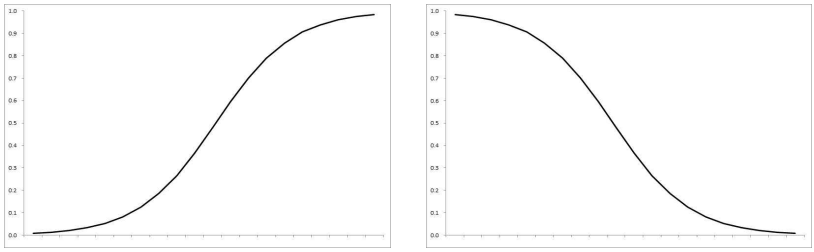
가. 로지스틱회귀모형

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 1 또는 0이 될 확률을 예측하는 모형을 말한다. 종속변수의 계급이 0과 1 두 가지 값을 가지고 관심의 대상이 되는 계급이 1이 될 확률을 예측하는 모형은 다음의 (식 1)과 같이 표현된다(Hosmer와 Lemeshow, 2000).

$$P(Y_i = 1 | \boldsymbol{x_i}) = p(\boldsymbol{x_i}) = \frac{\exp(\boldsymbol{\beta' x_i})}{1 + \exp(\boldsymbol{\beta' x_i})} \quad \text{(식 1)}$$

, $\boldsymbol{\beta'} = (\beta_0, \cdots, \beta_{p-1})$ 모수 벡터, $\boldsymbol{x}_i = (1, x_{1i}, \cdots, x_{p-1,i})$인 i번째 관찰 자료 벡터

<그림 Ⅲ-1> 로지스틱회귀모형에 의한 확률 분포



(a) $\beta > 0$ (b) $\beta < 0$

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 해석하기가 쉬우며, 셋째 과대 적합(over-fitting)할 가능성이 적고, 오차를 최소화하는 선형적인 관계를 찾는데 매우 우수한 기법이기 때문이다(이영섭, 2003).

이러한 로지스틱회귀모형은 $\boldsymbol{\beta'}$에 대해서 비선형 함수이나 이를 선형으로 변환시킬 수 있다. 기대 반응 $P(Y_i | \boldsymbol{x_i})$는 확률을 의미하므로, 다음과 같이 정의하자.

$$P(Y_i = 1 | \boldsymbol{x_i}) = p(x_i) \quad \text{(식 2)}$$

아래의 확률 $p(x_i)$에 대한 로짓 변환 (식 3)을 고려하자. (식 3)을

이용하면, (식 1)로부터 (식 4)를 구할 수 있다.

$$p'(x_i) = \log\left(\frac{p(x_i)}{1-p(x_i)}\right) \quad \text{(식 3)}$$

$$p'(x_i) = \log\left(\frac{p(x_i)}{1-p(x_i)}\right) = \boldsymbol{\beta' x_i} \quad \text{(식 4)}$$

(식 4)에서 $\log(p(x_i)/1-p(x_i))$은 로짓(logit)이라 부르며 (식 4)를 로짓함수라고 한다. 다음은 로지스틱회귀모형의 $\boldsymbol{\beta'}$의 최대우도추정량을 구하여 보자. Y_i는 변수가 0 또는 1을 가지는 가변수이므로, 각각의 관찰치는 베르누이(Bernoulli) 확률변수이다. 그러므로, 베르누이 분포를 이용하여 최대우도추정량(maximum likelihood estimator)을 구할 수 있다. $P(Y_i=1)=p(x_i)$, $P(Y_i=0)=1-p(x_i)$인 베르누이 분포는 다음과 같다.

$$f(y_i)=p(x_i)^{y_i}(1-p(x_i))^{1-y_i}, \quad y_i=0,1 \ ; \ i=1, \cdots, n \quad \text{(식 5)}$$

그리고 Y_i는 각각 독립이므로, 결합밀도 함수는 다음과 같다.

$$g(Y_1,\cdots,Y_n \mid \boldsymbol{x_i}) = \prod_{i=1}^{n} f_i(y_i \mid \boldsymbol{x_i}) = \prod_{i=1}^{n} p(x_i)^{y_i}(1-p(x_i))^{1-y_i} \quad \text{(식 6)}$$

위의 결합 밀도 함수는 로그의 성질을 이용하면 다음의 (식 7)로 표현할 수 있다.

$$\log g(Y_1,\cdots,Y_n \mid \boldsymbol{x_i})$$

$$= \sum_{i=1}^{n} y_i \log\left(\frac{p(x_i)}{1-p(x_i)}\right) + \sum_{i=1}^{n} \log\left(1-p(x_i)\right) \quad \text{(식 7)}$$

$$\log\left(\frac{p(x_i)}{1-p(x_i)}\right) = \boldsymbol{\beta}'\boldsymbol{x_i} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{p-1,i} \qquad \text{(식 8)}$$

(식 7)은 (식 8)에 의해 다음과 같은 (식 9)가 된다.

$$\log g(Y_1, \cdots, Y_n \mid \boldsymbol{x_i}) = \log L(\boldsymbol{\beta})$$
$$= \sum_{i=1}^{n} y_i(\boldsymbol{\beta}'\boldsymbol{x_i}) - \sum_{i=1}^{n} \log\{1 + \exp(\boldsymbol{\beta}'\boldsymbol{x_i})\} \qquad \text{(식 9)}$$

G를 (식 9) 우도함수의 이차(second order) 편미분의 행렬이라고 하자. 이때 각 편미분은 모수 $\beta_0, \cdots, \beta_{p-1}$에 대해서 이루어진다. 즉, $G_{pp} = [g_{jk}]$, $j = 0, 1, \cdots, p-1$, $k = 0, 1, \cdots, p-1$이다. 그러므로 g_{jk}에 대한 각각의 편미분 값을 0이라 한 후, 수리적 접근 절차(numerical search procedure)를 이용하여 풀면 $\boldsymbol{\beta}' = [\beta_0, \cdots, \beta_{p-1}]$에 대한 최대우도추정량은 $\boldsymbol{b}' = [b_0, \cdots, b_{p-1}]$이 된다. 그러므로, (식 1)의 로지스틱회귀모형에 대한 최대우도(maximum likelihood, ML) 추정식은 다음과 같이 표현된다.

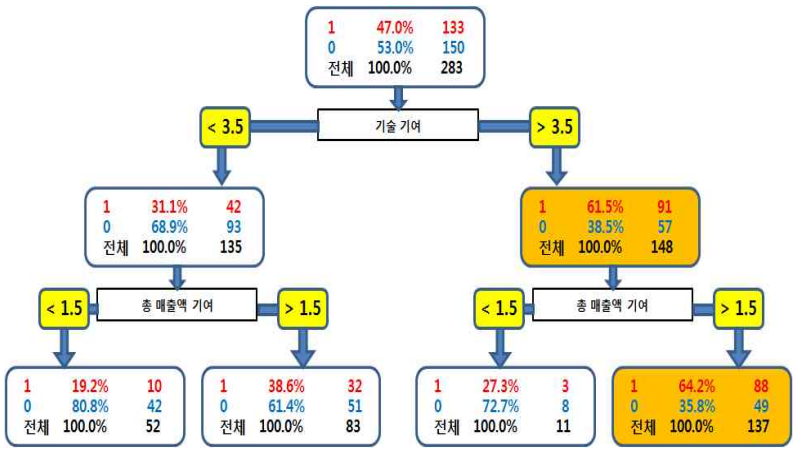
$$P(\widehat{Y}_i \mid \boldsymbol{x_i}) = \hat{p}(x_i) = \frac{\exp(\boldsymbol{b}'\boldsymbol{x_i})}{1 + \exp(\boldsymbol{b}'\boldsymbol{x_i})} \qquad \text{(식 10)}$$

본 연구에서는 통계 프로그램 R의 “glm 함수"를 사용하여 분석을 수행한다. glm 함수는 선형회귀분석, 로지스틱회귀분석, 일반선형모형 등의 회귀분석을 수행하는 것으로써 R 프로그램에서 기본적으로 제공하고 있다. 본 연구에서 이용한 로지스틱회귀모형 구축 R 프로그램은 부록에 제시한다.

나. 의사결정나무모형

의사결정나무모형은 의사결정 규칙(decision rule)을 나무 구조로 도표화하여 분류와 예측(prediction)을 수행하는 방법이다. 모형의 나무 형성 알고리즘은 CHAID(CHi-squared Automatic Interaction Detection), CART(Classification And Regression Trees), C4.5(또는 C5.0)가 대표적이다.

<그림 Ⅲ-2> 의사결정나무 구조



본 연구에서는 통계 프로그램인 R에서 “rpart 라이브러리"를 사용하여 분석을 수행한다. rpart 라이브러리는 CART 알고리즘을 구현한 패키지이다. CART 알고리즘에 대해 간략히 설명하면 다음과 같다. 1984년에 최초 발표된 CART는 기계학습(machine learning)의 시초가 되고 있다. CART 알고리즘은 독립변수들과 종속변수로 이루어진 자료에서 독립변수의 특성에 따라 이지 분리(binary split)를 수행한다. 이때 종속변수가 범주형인 경우 마디의 순수함을 나타내는 지니 지수

에 의해 분리 여부를 결정하고, 연속형인 경우는 분산의 감소량을 이용한다. 지니 지수는 불순도(impurity)를 측정하는 지수이다.

다음의 식은 지니 지수를 나타내는 것인데 n은 그 마디에 포함되어 있는 표본의 개수이고, n_j는 종속변수의 j번째 범주에 속하는 표본의 수를 말한다. 지니 지수는 n개의 표본 중에서 임의로 2개를 추출하였을 때, 추출된 2개가 서로 다른 그룹에 속해 있을 확률을 의미하며 Simpson의 다양도 지수(diversity index)로도 알려져 있다. 추출된 하나의 표본, 즉 개체가 특정 독립변수에 의해 집단이 구분되면, 구분된 하나의 집단에서 나머지 집단의 개체가 선택될 확률을 계산하여 집단을 분리하며 집단이 순수할수록 지니 지수의 값이 작아진다.

$$G = \sum_{j=1}^{c} P(j)(1-P(j)) = 1 - \sum_{j=1}^{c} P(j)^2 = 1 - \sum_{j=1}^{c} (n_j/n)^2 \quad \text{(식 11)}$$

, 여기에서 c은 종속변수의 범주 수

, $P(j)$는 주어진 자료 중 j범주에 분류될 확률

종속변수의 범주가 2개인 경우 지니 지수는 다음과 같이 표현될 수 있는데, 이는 카이제곱 통계량을 사용하는 것과 같다.

$$G = 2P(1)P(2) = 2(\frac{n_1}{n})(\frac{n_2}{n}) \quad \text{(식 12)}$$

CART는 지니 지수를 가장 작게 해주는 독립변수와 그 변수의 최적 분리를 자식마디로 선택하는데, 부모마디가 자식마디로 분리되었을 때 불순도가 가장 작도록 자식마디를 형성하는 것을 의미한다, 지니 지수의 감소량은 다음의 (식 13)으로 구할 수 있다. 이는 다음과 같은 자식 마디에서 불순도의 가중합을 최소화하는 것과 같다.

$$\Delta G = G - \frac{n_L}{n} G_L - \frac{n_r}{n} G_R \quad \text{(식 13)}$$

, 여기에서 n은 부모마디의 표본 수이며,

, n_L과 n_R은 각각 자식마디의 표본 수

$$P(L)G_L + P(R)G_R = \frac{n_L}{n} G_L + \frac{n_R}{n} G_R \quad \text{(식 14)}$$

의사결정나무모형의 장점은 분류 규칙을 나무구조의 도표를 통해 확인하므로 이해가 쉽다는 것이다. 또한 연속형과 범주형 자료를 동시에 다룰 수 있고, 변수에 결측치가 발생해도 이를 분석에 활용할 수 있다. 두 개 이상의 변수가 결합한 교효효과(interaction effect) 해석이 용이하며, 선형성, 정규성, 등분산성 등의 통계적인 가정이 필요하지 않다. 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. 또한 연속형 변수를 비연속적인 값으로 취급하여 분리의 경계점 근방에서 예측 오류 가능성이 높으며, 자료가 달라질 때마다 모형이 달라질 가능성이 높고, 분석용 자료에만 의존하므로 새로운 자료의 예측에서는 불안정성이 높다(최종후와 진서훈, 2005).

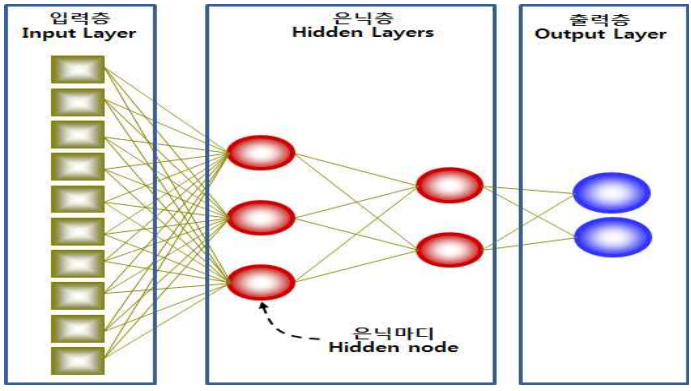
다. 신경망모형

인공신경망이라고 불리는 신경망은 주로 분류와 예측을 위해 사용되는 기계학습 기법 중 하나이다. 신경망모형은 인간이 학습하는 방식을 모방하여 고안되었는데, 뉴런(neuron)이 상호 연결되어 경험으로

부터 학습하는 두뇌의 생물학적 활동과 유사하다. 즉, 신경망모형은 인간의 뇌 기능에 착안하여 개발된 패턴 인식의 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 다음의 그림과 같이 입력층, 은닉층, 출력층으로 구성되어 있고 노드를 통해 연결되는 구조이다. 각 층의 역할을 살펴보면 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).

<그림 Ⅲ-3> 다층인식자 신경망 구조



앞의 그림에서 각 노드 간에는 연결강도인 가중치가 부여된다. 만약 입력층의 노드 i와 은닉층의 노드 j($node_j$)를 연결한 화살표에 부여된 가중치를 w_{ij}라 하고, 입력층에서의 입력을 $x_{1,},\cdots,x_n$라고 하면,

은닉층의 노드 j에서는 다음과 같은 값이 입력된다.

$$node_j = \sum x_i w_{ji} + w_{j0} \quad \text{(식 15)}$$

여기에서 $node_j$를 넷 활성화(net activation)라 하고, 이와 같은 계산식을 결합함수(combination function)라고 한다. 그리고 은닉층의 노드 j에서 다음의 은닉층이나 출력층으로 내보내어지는 값은 다음과 같은 활성함수(activation function)에 의해 계산되어지는데, 활성함수 중 가장 많이 사용되는 것은 부호(sign)함수와 시그모이드(sigmoid)함수이다. 시그모이드 함수는 0과 1 사이의 값을 출력하는 함수인데 로지스틱 활성함수라고도 한다. 활성함수의 주요 역할은 노드로 모인 신호를 좀 더 큰 변별력을 가지도록 전환하는 것이다.

$$\begin{aligned} sign(node_j) &= \ \ 1, node_j \geqq 0 \\ sign(node_j) &= -1, node_j < 0 \end{aligned} \quad \text{(식 16)}$$

$$sigmoid(node_j) = 1/(1+\exp(-x)) \quad \text{(식 17)}$$

신경망모형에 의한 학습은 연결 강도인 가중치를 조절하는 작업의 반복이다. 신경망모형에 의해 예측된 출력값과 원래의 결과값을 비교하여 값이 같은지를 확인한 후, 값이 같이 않다면 가중치를 조절한다. 여기에서 출력값과 원래의 결과값이 얼마나 같은지를 비교하는 기준에는 오차(error)의 제곱합(sum of square errer, SSE)과 정보 이론의 엔트로피(entropy)가 있다. 오차의 제곱합은 원래의 결과값 y_i와 예측된 출력값 $\hat{y_i}$의 차이에 대한 제곱의 합의 형태로 다음의 식과 같다.

$$SSE = \sum (y_i - \hat{y_i})^2 \quad \text{(식 18)}$$

(식 18)에 의해 차이가 발생하면 차이가 있는 출력 노드로부터 은

닉 노드로, 은닉 노드에서 입력 노드로 거슬러 올라가면서 가중치를 조절하여 값이 0에 가까워지도록 수회 반복하여 예측 성능을 최대화하는 적절한 가중치를 산출한다. 이와 같은 학습 방법을 역전파 알고리즘(back propagation algorithm)이라 하는데, 이 알고리즘은 홉필드(Hopfield)가 개발한 것으로 이 알고리즘이 개발된 이후 신경망모형은 널리 사용되기 시작하였다.

역전파는 전방향(feed forward) 수행, 오류의 역전파 계산, 가중치 조정이라는 세 가지 과정을 반복하여 사전에 목표한 정확도에 도달할 때까지 수행된다. 이와 같이 가중치를 조절하여 모형에 의한 예측 정확도를 높이다보면 과대적합이 발생할 수 있는데, 이를 피하기 위해서는 가중치를 조절할 때 마다 0과 1 사이의 값을 가중치에 곱해주는 가중치 감소(weight decay) 과정을 적용하여 신경망모형 구축을 수행하기도 한다.

$$adjw_j = w_j \times (1-\epsilon), \quad 0 < \epsilon < 1 \qquad \text{(식 19)}$$

신경망모형은 네트워크 구조가 복잡하여 학습 시 시간이 많이 소요된다. 신경망모형 학습을 위해서는 은닉층의 수와 각 은닉층에서의 노드 수를 결정하여야 한다. 일반적으로 가장 많이 사용하는 은닉층의 수는 한 개인데, 보통 하나의 은닉층만으로도 독립변수(예측변수)들 사이의 복잡한 관계를 알아내는 데 충분하기 때문이다. 은닉층에서의 노드 수는 그 수가 많을수록 과적합의 가능성이 높아진다. 일반적으로 가장 좋은 방법은 독립변수의 수를 적용하여 과적합의 여부를 점검하면서 점진적으로 줄이거나 늘려가는 방식을 이용한다.

신경망모형의 예측 정확도를 높이기 위한 중요한 요인 중 하나는 독립변수의 문제이다. 이는 신경망모형의 복잡성으로 인해 입력 자료에 매우 민감하기 때문인데, 일반적으로 신경망모형에 적합한 자료는

연속형 독립변수는 변수 간 범위에 큰 차이가 없으며, 범주형 독립변수인 경우 범주의 빈도가 비슷한 경우이다. 그러므로 연속형 독립변수의 경우 자료의 표준화 또는 범주화하는 방법을 이용하거나, 주성분 분석 기법 등을 사용해서 새로운 독립변수를 생성하여 사용한다. 그리고 범주형 독립변수는 더미변수(dummy variable)로 변형하여 사용한다. 그러나 더미 변수화 하는 경우 0과 1 더미와 -1과 1 더미를 사용할 때 회귀분석과는 달리 결과가 달라질 수 있다는 것을 유의하여 사용하여야 한다.

신경망모형의 장점은 자료들 간의 비선형적인 관계를 찾아 낼 수 있고 예측의 정확성이 매우 높다는 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).

신경망모형은 분석에 사용하는 표본으로부터 일반화하는 능력은 있지만 외삽추론은 심각한 위험요소이다. 즉, 신경망모형에서 사용한 독립변수가 특정한 범위 내에서 수행되었다면, 이 범위를 벗어나는 예측 결과들은 완전히 무의미할 수도 있다(조재희 외 역, 2018).

본 연구에서는 통계 프로그램 R에서 제공하는 “nnet 라이브러리" 를 사용하여 분석을 수행한다. nnet 라이브러리에서 신경망의 모수(parameter)는 엔트로피(entrophy) 또는 오차제곱합(sum of squared error, SSE)을 고려해서 최적화된다.

라. 랜덤포레스트모형

랜덤포레스트(random forest) 모형은 의사결정나무모형을 확장한 개념으로 앙상블(ensemble) 모형 중 하나이다. 즉 랜덤포레스트는 말 그대로 의사결정나무들이 많이 있는 모형으로, 의사결정나무 모형을 다수 만들어 예측력을 높이는 방법이다(Yoo, 2015). 일반적으로 의사결정나무모형은 특이값(outlier)을 하나의 노드로 구성할 수 있기 때문에 편향된 분포에 민감하지 않지만, 분포가 크기 때문에 노드에 미치는 영향은 크고 깊이가 깊어질수록 과적합의 위험이 커진다. 이와 같은 위험을 최소화하여 예측력을 높이고자 고안된 기법이 랜덤포레스트모형이다.

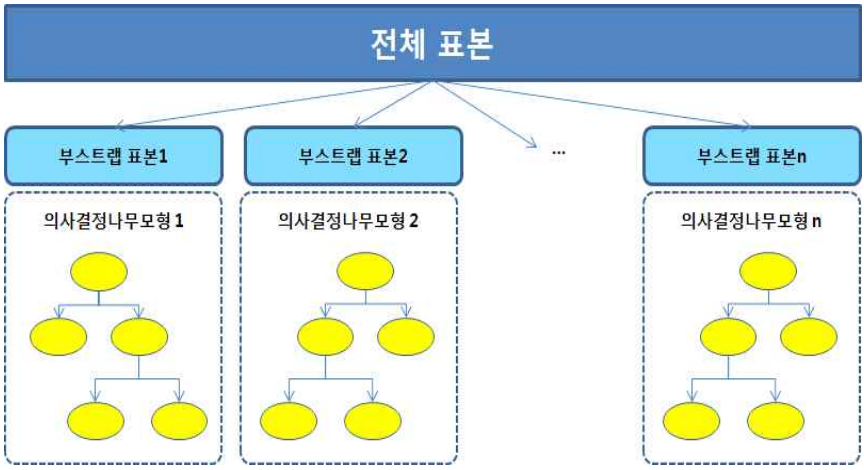
의사결정나무는 동일한 하나의 데이터 집합에서 한 번의 훈련용 데이터로 학습하고 하나의 의사결정나무를 생성하여 종속변수를 예측하고 분류하지만, 랜덤포레스트는 동일한 하나의 데이터 집합에서 훈련용 데이터를 여러 번 복원 추출하여 학습을 하는 방법이다. 다음의 그림과 같이 여러 개의 나무를 생성하고 이를 결합해 최종적으로 종속변수를 분류하거나 예측한다. 이 때 분류에 대해서는 투표(voting) 방법을 적용하고, 예측에 대해서는 평균화 방법을 적용한다. 랜덤포레스트모형의 실행 단계는 다음과 같이 정리할 수 있다.

① 주어진 표본으로부터 복원추출 방식으로 여러 개의 무작위 표본들을 생성하는데 이러한 추출 방법을 부스트랩(bootstrap)이라고 한다.

② 각 부스트랩 표본마다 무작위로 예측을 위한 독립변수들을 선택하여 서브셋을 만들고 이를 의사결정나무에 적합시킨다.

③ 예측력 향상을 위해 각 의사결정나무로부터 얻어진 결과를 결합한다.

<그림 Ⅲ-4> 랜덤포레스트 구조



전술하였듯이 랜덤포레스트는 부스트랩 표본[2]을 활용한 배깅(bagging)[3]을 이용해 나무를 만드는 작업을 반복하고, 이렇게 만들어진 다수의 의사결정나무들의 결과에 대한 투표로 최종 결과를 출력한다. 또 각 가지를 나누는 변수를 선택할 때 전체 변수를 매번 모두 고려하는 대신 변수의 일부를 임의로 선택하는 특징을 갖는다. 이때 각 나무들의 편향은 그대로 유지되지만 분산은 감소하기 때문에 더 안정적이고 예측의 정확도가 높다(Choi, 2017).

랜덤포레스트는 무작위성을 최대로 주기 위하여 부스트랩 표본과 더불어 독립변수들에 대한 무작위 추출을 결합한다. 만약 랜덤포레스

2) 부스트랩 표본이란 단순 복원 임의추출 방법을 통해 추출한 표본의 크기가 동일한 여러 개의 표본 데이터를 말한다.

3) 배깅은 데이터에서 여러 개의 부스트랩(bootstrap) 데이터를 생성하고 부스트랩된 각 데이터를 이용해 모형을 구축한 후 이를 결합하여 최종적인 예측 모형을 만드는 방법이다. 배깅은 예측 모형의 변동성을 감소시키기 위해 사용되는데, 원천자료(raw data)로부터 여러 번의 샘플링을 통해 예측 모형의 분산을 줄이므로 예측력이 향상된다.

트의 종속변수가 연속형인 경우 독립변수의 개수가 m개이면, 각 분할에서 랜덤으로 m/3개의 변수를 선택하여 나무(tree)를 구성한다(Breiman, 2001).

랜덤포레스트모형의 장점은 트리의 다양성을 극대화하여 예측력이 우수하고 많은 트리의 예측 결과를 종합하기 때문에 안정성이 높다는 것이다. 그러나 다수의 트리를 이용하여 결과를 종합하므로 의사결정나무모형의 장점인 설명력은 없다. 본 연구에서는 통계 프로그램인 R에서 제공하는 라이브러리인 "randomForest"를 사용하여 소상공인 신용평가 모형을 구축한다.

마. 서포트벡터머신(SVM)

SVM(support vector machine)의 이론적인 내용은 Lantz(2015)의 내용에서 발췌하여 정리한다. SVM은 분류 시 가장 보편적으로 사용되고 있는 기계학습 기법 중 하나이다. 이 기법은 주로 다루고자 하는 데이터의 종속변수가 2개의 범주 또는 계급으로 분류되어 있을 때 사용한다. 예를 들어, 신용평가 시 우량과 불량 기업 두 가지가 있을 때 어떠한 하나의 기업이 우량일 경우를 예측할 때 사용한다.

일반적으로 SVM은 다양한 학습 데이터의 분포에서도 매우 성능이 좋은 분류 기법으로 알려져 있다. SVM은 다른 분류 기법들 보다 정확도 측면에서 우수한 결과를 보여주고 있다는 장점이 있는 반면, 의사결정나무모형이나 로지스틱회귀모형과 같은 직관적인 해석이 불가능하다는 단점이 있다. 이와 같은 이유로 결과의 해석보다는 분류의 정확도가 중요한 경우 SVM을 사용하는 경우가 많다.

SVM은 용어의 의미 그대로 학습을 위한 데이터가 벡터 공간(vector space) 상에 존재한다고 가정한다. 즉, 학습을 위한 데이터는

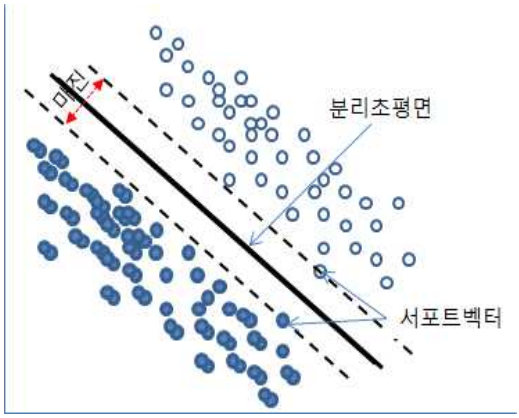
벡터 공간의 직각 좌표계에 위치해 있다. SVM은 벡터 공간 상에 존재하는 학습 데이터가 어떠한 그룹에 속하는지를 분류하기 위한 선형 분류자(linear classifier)를 찾는 기법이다. 이 때 차원은 데이터가 가지고 있는 변수의 개수이다.

SVM은 지도학습 기법 중 하나로 고차원의 벡터 공간 상에 존재하는 데이터를 가장 잘 분류하는 선 또는 초평면을 찾아 이를 이용하여 분류와 회귀를 수행하는데 선형 및 비선형 분류에 모두 사용할 수 있다. 그리고 이 기법은 모든 속성을 이용하는 전역(global) 분류 모형으로 비중첩(non-overlapping) 분할을 수행한다.

SVM의 이해를 위해서는 먼저 초평면(hyperplane), 분리 초평면(separating hyperplane), 최대 마진 분류기(maximal margin classifier)에 대한 설명이 필요하다. 먼저 초평면은 데이터를 분리하는 경계를 말하는데, 최대 마진 분류기의 선형 경계이며, 데이터가 d 차원이라면 초평면은 d-1차원을 가진다. 다음으로 분리 초평면이란 데이터를 완벽하게 분리할 수 있는 초평면을 말하는데, 분리 초평면이 여러 개 존재하는 경우 최적 분리 초평면을 선택하여야 한다. 마지막으로 최대 마진 분류기는 데이터를 선형으로 분류하는 것으로써, 직관적으로 이해하기 쉽지만 비선형 형태의 데이터에는 적용하기 힘들다.

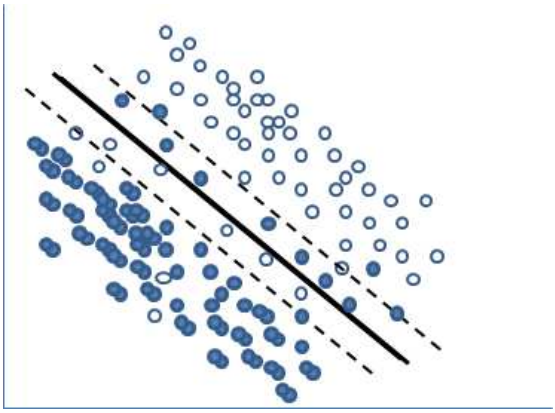
일반적으로 데이터를 분류하는 초평면은 무한개가 존재할 수 있다. 무한개의 분리 초평면 중에서 가장 최적의 초평면은 두 개의 집단으로 분류할 때, 각 집단에서 초평면까지의 수직 거리가 가장 짧은 마진(margin)이 가장 큰 최대 마진 초평면을 찾아야 한다. 이는 각 집단에서의 훈련 데이터까지의 최소 거리가 가장 큰 초평면이 데이터를 가장 잘 분류한다는 것을 의미한다.

<그림 Ⅲ-5> 서포트 벡터, 분리 초평면, 마진



위의 그림에서 점선에 걸쳐 있는 관측치를 “서포트 벡터”라고 하는데, 이 관측치들은 약간 이동하거나 다른 관측치가 서포트 벡터가 될 경우 초평면도 변경되어 최대 마진 초평면을 “서포트”한다는 의미에서 붙여진 명칭이다. 일반적으로 최대 마진 초평면은 서포트 벡터에 따라 변경된다.

<그림 Ⅲ-6> 초평면에 의해 분류가 되지 않는 데이터



만약 두 집단으로 분류가 가능한 분류 초평면이 존재한다면 최대 마진 분류기는 분류의 성능이 가장 높은 초평면이 된다. 그러나 실제 데이터를 이용하여 분류를 수행할 경우 초평면이 존재하지 않는 경우가 있다. 이러한 경우를 해결하기 위해 소프트 마진(soft margin)을 사용하는데, 이는 데이터 분류 시 약간의 오류를 허용하는 방법이다. 약간의 오류를 허용하여 분류를 수행하면 과대적합을 방지할 수 있다.

SVM에 대해 좀 더 자세히 살펴보면 다음과 같다. SVM은 최대 마진 분류기를 확장한 개념으로 분류 시 약간의 오류를 감안한 상태에서 최적의 분류가 가능하도록 하는 평면을 찾는 것이다. d 차원 공간상에 존재하는 n개의 데이터는 $D=\{(x_i,y_i), i=1,2,\cdots,n\}$이고, 종속변수 y는 두 개의 값(+1 또는 -1)만 가지고 있다고 하자. 이 때 다음의 초평면 $h(x)$는 d 차원에서 원래의 공간을 2개로 나누는 선형판별함수를 제공한다(나종화, 2017).

$$h(x)=w^Tx+b=w_1x_1+w_2x_2+\ \cdots\ +w_dx_d+b \qquad \text{(식 20)}$$

, 여기에서 w는 d 차원의 가중 벡터

, b는 편향(bias) 상수

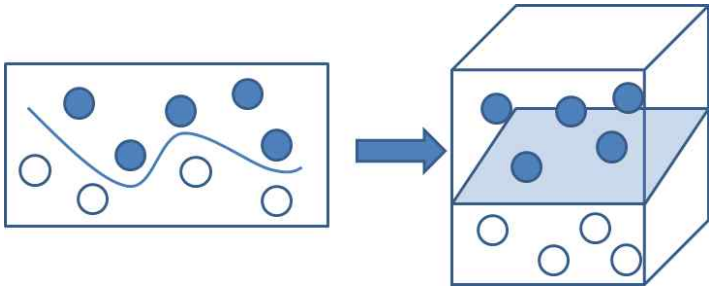
초평면 상의 점들은 $h(x)=0$으로, 초평면은 $w^Tx=-b$를 만족하는 모든 점들로 정의된다. 만약 모든 데이터가 선형으로 분리 가능하다면, $h(x)<0$인 데이터는 -1의 군집으로 $h(x)>0$인 데이터는 +1의 집단으로 분류될 수 있다. 이때 $h(x)$는 모든 데이터가 어떤 집단으로 분류되는지를 예측하는 선형 분류기의 역할을 한다.

만약 모든 데이터를 두 개의 집단으로 완벽하게 분류할 수 있는 초평면 $h(x)$를 찾게 된다면, 정확도는 높지만 과대적합이 발생할 가능성이 매우 크다. 전술하였듯이 이러한 경우에는 약간의 분류 오류

를 허용하게 되는데, SVM에서는 허용 가능한 오류를 설정하는 것이 매우 중요한 과정이다. SVM에서 허용 가능한 오류를 cost라고 하는데, cost와 마진은 서로 반비례 관계, 즉 cost가 커질수록 마진은 작아지게 된다. 실제로 분석 시 어떠한 cost가 좋은지는 데이터를 분석해 봐야 알 수 있는데, R에서는 "tune.svm"함수를 통해 적절한 cost를 산출할 수 있다.

SVM을 확장하여 두 집단 간 경계가 비선형인 경우의 분류가 가능한데, 선형 분류기를 비선형으로 변경함으로써 이와 같은 작업이 가능하게 된다. 이 때 커널 트릭(kernel trick)이라는 다차원 공간상으로의 맵핑(mapping) 기법을 사용하여, 선형 분류가 어려운 분포를 선형 분류가 가능하도록 변수 공간 또는 차원을 확장한 후, 이를 선형으로 변환하고 소프트 마진 방식으로 최종적으로 분류를 수행한다. 본 연구에서는 통계 프로그램인 R에서 제공하는 라이브러리인 "kernlab"를 사용하여 소상공인 신용평가 모형을 구축한다.

<그림 Ⅲ-7> 차원 확장을 통한 비선형 분류



Ⅳ. 데이터 정제 및 모형 평가 방법

1. 데이터 정제 및 변수 선택

통계 분석이나 기계학습 기법을 이용한 모형 구축의 성공을 위해서는 다양한 요소들이 있지만, 무엇보다 중요한 것은 질(quality) 좋은 데이터를 양(quantity)적으로 충분히 확보하는 것이다. 통계 분석의 경우 모집단의 특성을 추정하기 위해 표본 추출의 대표성이 담보되었다는 전제 하에 작은 수의 표본으로도 모집단의 특성을 찾는 것이 가능하지만, 만약 미래의 예측이 주된 목적이라면 데이터의 양과 질은 특히 중요하게 된다.

분석이나 모형 구축을 위한 양질의 원천자료(raw data)가 확보되어 있다고 하더라도, 성공적인 분석이나 모형 구축을 위해서는 통계적인 분석 기법을 이용하여 데이터를 탐색한 후, 다양한 방법을 이용하여 분석 가능한 형태로 데이터를 정제(cleaning)하여야만 한다. 데이터 정제 시에는 기본적으로 각 개별 변수에 대한 분석 즉, 단변량(uni-variate) 분석 기법을 이용하여 결측치(missing value)나 특이값(outlier value) 등에 대한 사항을 파악하고 이를 처리하여야 한다. 필요 시 불필요한 데이터를 제거하거나 대체하는 작업 등을 거치게 된다. 그리고 추가적으로 각 변수 간 논리적인 관계의 규명을 통해 변수 값들을 변환하는 작업을 수행하기도 한다.

이처럼 예측력이 우수한 모형 구축을 위한 요소 중 하나인 원천자료를 적절한 값으로의 변환하는 것은 결측치 대체, 특이값 처리, 표준화, 재범주화 등의 데이터 작업으로, 이는 주로 통계적인 기법을 이용하여 수행한다. 본 절에서는 안정성과 예측력이 우수한 모형 구축을

위한 데이터 변환 및 변수 선택 방법에 대해 기술한다.

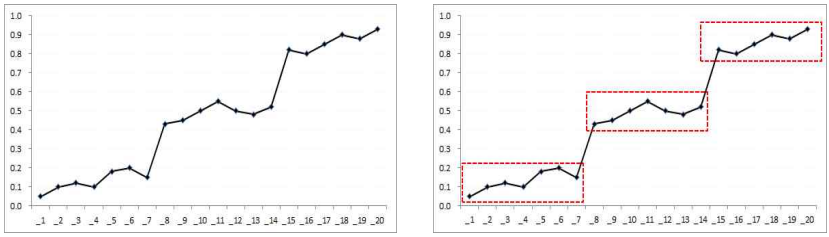
먼저 데이터 변환 방법에 대한 소개이다. 데이터 변환은 범위를 줄이기 위해 표준화를 하는 것이 일반적이다. 데이터 표준화란 원래 변수의 값을 -1.0~1.0 또는 0.0~1.0처럼 정해진 작은 구간 내에 들도록 척도화하며 정규화(normalizing)하는 것을 말한다. 원천자료에 대한 표준화를 통해 단위나 범위가 다른 변수들을 표준화함으로써 표준화된 변수들끼리 서로 상대적인 비교가 가능해진다는 장점이 있다. 일반적으로 소득이나 매출액 등의 데이터는 범위가 큰 구간을 가지는데, 표준화를 통해 원천자료의 범위가 큰 변수가 범위가 더 작은 구간을 가진 변수(예 : 0, 1만 가지는 이항변수)보다 중요해지는 것을 방지할 수 있다.

데이터 표준화를 위한 많은 방법들이 있는데, 이 중에서 3가지 방법, 최소-최대 표준화(min-max standardization), z-스코어 표준화(z-score standardization), 소수점으로의 변환에 의한 표준화(standardization by decimal scaling)가 대표적인 변수 표준화 방법이다. 첫째, 최대-최소 표준화는 원래의 자료를 선형적(linear)으로 일정한 구간(일반적으로 0과 1사이)으로 변환하는 방법이다. 둘째, z-스코어 표준화는 변수 A에서 산출한 평균과 표준편차를 이용하여 표준화(또는 정규화)하는 방법이다. 마지막으로 소수점으로의 변환에 의한 표준화는 변수 A값들의 소수값으로 변환시키는 방법이다.

본 연구에서는 위의 세 가지 방법 대신 원천자료의 표준화를 위해 계급화(classing) 기법을 사용하고자 한다. 계급화 기법은 크게 계급세분화(fine classing)과 성김화(coarse classing) 단계로 구분된다. 계급세분화는 원래의 독립변수값을 종속변수인 불량과의 관계 분석을 통해 불량률이 유사한 범주를 하나의 범주로 묶은 후 계급화하여 분석에 사용하는 방법이다. 자료의 크기나 변수의 척도에 따라 다소 차이는

있지만, 개별 변수를 구성비 5%를 기준으로 최대 20개의 구간으로 세분화하고 불량률을 기준으로 서열화한 후 변별력 지표인 KS 통계량(기준 KS≥0.1)을 이용하여 1차적으로 변수를 선정한다. 성김화는 1차적으로 계급세분화에 의해 범주형으로 변환된 변수에 대해 동질적인 불량률을 보이는 구간으로 다시 묶는 단계이다(Leung 외, 2008). 계급화 기법을 이용할 경우 결측치와 특이값의 사용이 가능해지는데, 그 이유는 결측치나 특이값이 불량률과 연관되어 하나 또는 그 이상의 구간으로 묶이기 때문이다.

<그림 Ⅳ-1> fine & coarse classing 개요



(a) 계급세분화 (b) 성김화 – 재범주화

신용평가모형 구축을 위한 독립변수를 선택할 때 통계적으로 선택된 결과를 바탕으로 대출 시 비즈니스 관점의 부합성을 고려하여 최적의 변수를 조합하여야 하는데, 통계적으로 유의한 변수를 선택하는 다양한 방법들이 있다. 변수 선택 기법은 변수의 개수 자체를 줄이는 방법과 서로 상관성이 높은 변수를 하나의 변수군으로 묶어서 사용하는 방법 크게 두 가지가 있다.

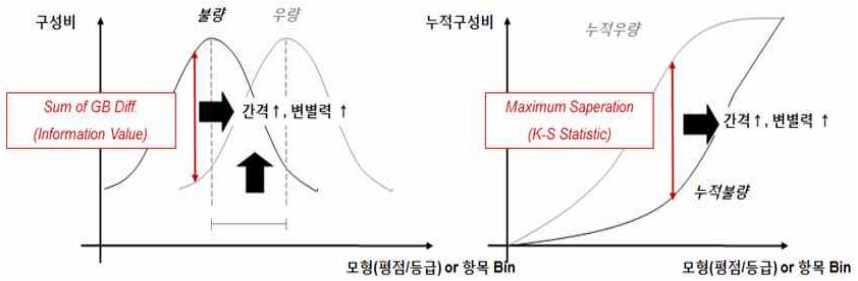
먼저 변수의 개수를 줄이는 방법으로는 평균 차이 검정, 범주 간 분포 차이 검정, 회귀분석 등을 이용하여 종속변수와 각 개별 독립변수 간 유의한 관계가 있는 변수를 선정하거나, 모든 독립변수를 동시에 적용하여 종속변수에 유의한 영향을 주는지를 회귀분석을 통해 분

석하는 방법 등이 있다. 그리고 유사한 변수를 하나의 변수군으로 묶어서 사용하는 방법으로는 주성분분석(principal analysis), 요인분석(factor analysis) 등이 있다.

본 연구에서는 박주완(2018)에서 사용한 변수 선택 기법을 적용한다. 박주완(2018)은 독립변수와 불량과의 관계를 이용한 계급세분화(fine classing)를 수행한 자료에 대해 KS 통계량 이용, 원 자료 중 범주형(categorical) 변수 대해 종속변수와 독립변수 간 카이제곱 통계량 이용, 연속형 변수에 대해서는 t-검정을 이용하여 1차적으로 변수를 선정하였다. KS 통계량은 0.1 이상인 경우를 선택하고, 카이제곱 검정과 t-검정 모두 p값이 유의수준 0.05 이하인 변수를 추출한다. 다음은 KS통계량, 카이제곱 검정, t-검정에 대한 내용이다.

여기에서 KS 통계량은 모형 및 평가항목의 변별력이 극대화되는 지점을 측정하여 평가하는 지표로써, 우량 집단과 불량 집단 간 누적분포 차이의 최대값을 말한다.

<그림 Ⅳ-2> KS 통계량의 개요



다음으로 카이제곱 검정은 2개 이상의 질적 범주로 구성되어 있는 명목 또는 순위 변수에 있어서 각 범주의 관찰빈도와 귀무가설에 의한 기대 빈도 사이에 유의한 차이가 있는가를 검정하는 방법이다. 예

를 들면, “주사위 한 면이 나올 확률은 1/6인데 실제 주사위를 던지는 실험을 한 후 각 면이 나오는 빈도를 관찰하여 실제로 주사위의 한 면이 1/6으로 나오는가?”, “종교에 따라 취미의 차이가 있는가?”, “여당과 야당에 따라 정치에 대한 긍정 부정의 시각적인 차이가 있는가?” 등에 대한 의문점을 통계적으로 해결할 때 사용한다. 이 때 자유도(degree of freedom)가 1인 경우는 전체 사례수가 30개 이상이면서 각 셀의 빈도가 5개 이상일 때 적용 가능하다. 그리고 자유도가 1보다 큰 경우는 사례수 30개 이상 5개 미만의 셀이 전체의 20% 미만이고 모든 셀이 1 이상의 기대빈도를 가질 때 카이제곱 검정 사용이 무난하다. 만약 변수가 비율 척도나 등간 척도인 경우에는 이를 명목척도로 변환하여 카이제곱 검정을 수행, 즉 일정 구간으로 나누어 각 계급 내 빈도를 구한 후 카이제곱 검정을 수행한다.

연속형 변수에 대해 사용하는 t-검정은 대표본이 아닌 소표본에 대한 모수의 추정치와 가설을 검정하는 방법으로 주로 두 집단 간 표본평균의 차이를 비교할 때 사용한다. 예를 들면, “성별에 따른 임금액수의 평균 차이가 있는가, 특정 질병 발생 집단과 정상 집단에 대한 약품 실험 시 효과에 차이가 있는가.” 등을 통계적으로 검정할 때 사용한다. t-검정은 독립인 두 표본 또는 쌍을 이룬 두 변수 값 사이의 평균 비교에 사용한다. 독립인 두 표본에 대한 평균 차이 비교는 두 집단의 평균의 차이를 검정하며, 독립변수 내에서 두 개의 집단 간 평균 점수의 차이가 유의미한지를 검정한다. 서로 쌍을 이룬 두 변수 값 사이의 평균 차이 비교는 동일 표본에서 측정된 두 변수 값의 평균의 차이를 검정하는 방법으로, 보통 통제 집단의 사전과 사후 검사의 차이를 검정할 때 사용한다.

최종적인 변수 선택은 계급세분화, 카이제곱 검정과 t-검정에 의해 1차적으로 선정된 변수에 대해 성김화(coarse classing) 방법을 이용해

계급세분화된 값을 축약하여 재범주화 한 후 더미변수를 생성하여 로지스틱회귀모형의 단계적 선택법(stepwise method)으로 변수를 선택한다. 로지스틱회귀모형을 이용한 변수 선택 방법은 전진 선택법, 후진 소거법, 단계적 선택법 등이 있다. 전진 선택법은 가장 유의미한 변수부터 하나씩 선택하여 더 이상 모형 설명력이 증가하지 않을 때까지 변수를 추가하고, 후진 소거법은 유의미한 p개의 변수 중 가장 설명력이 낮은 변수부터 하나씩 제거하여 더 이상 제거할 변수가 없다고 판단될 때 변수선택 과정을 중단하는 방법이다.

본 연구에서 사용하는 단계적 선택법은 전진 선택법과 후진 소거법의 단점을 보완한 방법으로 중요한 변수를 하나씩 선택하면서 이미 선택된 변수가 추가된 변수에 의해 설명력이 상실되는지 매 단계마다 검토하는 방법이다. 통상적으로 변수의 수가 많을 경우 후진 소거법, 적은 경우 전진 선택법이 적합하나 근래에는 통계 분석 툴 등의 성능 향상으로 인해 단계적 선택법을 주로 많이 사용하고 있다. 이 때 각 회귀계수의 p-값은 일반적으로 유의수준 0.05 이하인 경우 해당 항목이 통계적으로 유의미하다고 판단하며, 회귀계수 추정치 값의 부호가 음(-)의 값일 때 이는 다중공선성에 의한 결과일 가능성이 높으므로, 분석 결과에 의한 회귀계수가 음인 경우 다중공선성이 있는 것으로 판단하여 모형 구축용 변수에서 제외한다.

이와 같이 과정으로 선정된 변수는 스피어만 상관계수(Spearman's correlation coefficient)를 이용해 다중공선성(multi-collinearity) 여부를 점검한 후 다중공선성이 있는 변수 중 설명력이 약한 변수는 제거하고 모형 구축에 활용한다. 스피어만 상관계수는 순서형 변수의 상관계수를 산출하는 방법으로 독립변수 간 상관계수가 0.7 이상인 경우 다중공선성이 있다고 판단한다. 다중공선성이 있는 변수 중 설명력이 약한 변수는 제거하고 모형 구축에 활용하는데, 이유는 다중공선성이

있는 변수를 사용하여 분석을 수행할 경우 변수 간 간섭에 의해 회귀계수에 편의가 발생할 가능성이 높기 때문이다(박주완, 2018).

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.

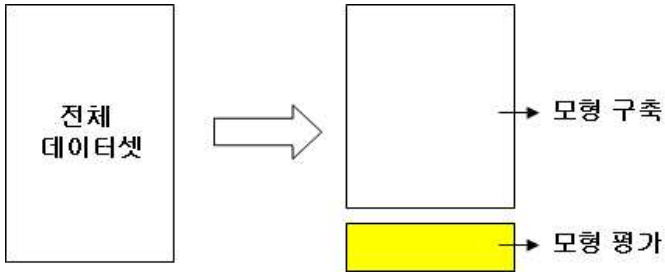
개발된 모형의 타당성을 검토하는 방법들로는 별도의 평가용(validation) 자료를 이용한 예비 방법(holdout method), k개의 분할된 자료를 이용하는 k-중첩 교차타당법(k-fold cross validation method)과 부스트랩 방법(bootstrap method) 등이 있다(Kohavi, 1995). 세 가지 모형 평가 방법에 대한 간략한 설명은 다음과 같다.

가. 예비 방법

예비 방법은 주어진 자료를 두 개의 독립된 집합인 훈련용 자료(training data)와 검증용 자료(validation data)로 임의 분할한 후 훈련용 자료는 모형을 구축하기 위해 사용하고 검증용 자료는 모형을 평가하기 위해 사용한다. 일반적으로 훈련용 자료로 2/3, 나머지 1/3은 검증용 자료로 할당한 후 모형 구축 및 검증 수행한다. 변형된 방법으로 무작위 부분 추출(random subsampling)이 있는데, 이는 예비 방법을 k번 반복한 후, 전체 정확도 추정은 반복으로 얻은 정확도의 평균으로 계산한다(박주완, 2010).

예비 방법은 평가를 위한 자료가 충분히 확보되어 있는 경우에 효과적인 방법으로 모형 구축을 위한 자료가 충분할 경우 평가의 정확성도 높고 평가에 소요되는 시간이 단축된다는 장점이 있다(강창완 외, 2007). 그러나 평가용 자료를 모형 개발에 사용할 수 없다는 것, 훈련용과 평가용 자료의 비율에 따라 다른 결과가 나타날 수 있다는 문제점과 개체수가 크지 않을 경우 불안정한 값을 제공한다는 단점이 있다(최종후 외, 2002; 박주완, 2010 재인용).

<그림 Ⅳ-3> 예비 방법의 개요



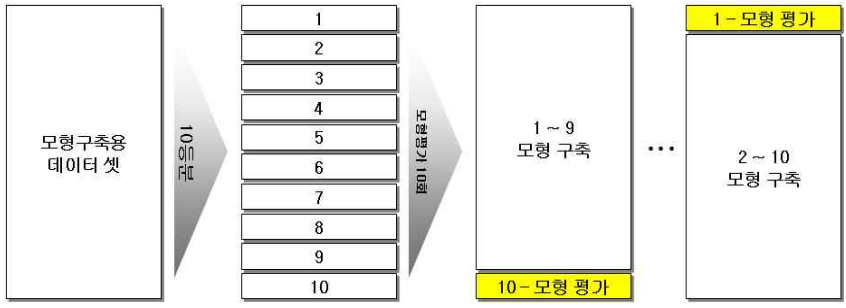
본 연구에서는 자료의 개수가 충분히 크기 때문에 예비 방법을 이용하여 “훈련용 자료:평가용 자료 = 7:3”으로 하여, 70%는 모형을 구축하는데 사용하고 나머지 30%의 자료를 이용하여 구축된 모형을 검증하고 평가한다.

나. 교차타당법

교차타당법의 과정은 다음과 같다. 먼저 초기 자료를 크기가 유사하게 D_1, …, D_k인 k개의 상호배반 부분집합으로 임의 분할한다. 분할된 자료를 이용하여 D_1, …, D_{k-1}는 모형 구축에 사용하고, D_k를

이용하여 검증한다. 분할된 자료가 k개이므로, 모두 k번의 모형 훈련과 평가가 발생한다. 이 방법은 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).

<그림 Ⅳ-4> 10중첩 교차타당법



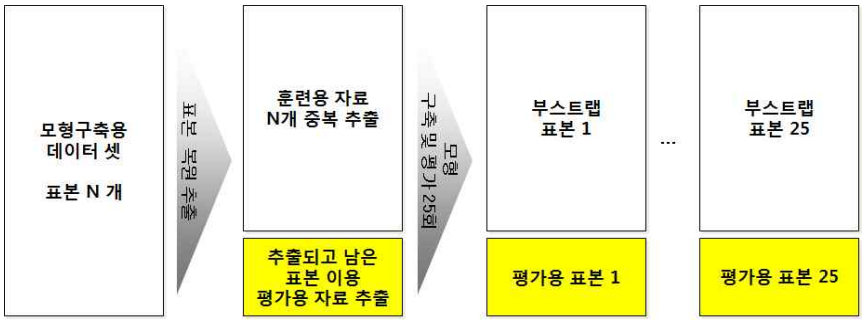
다. 부스트랩 방법

앞에서 설명한 예비 방법과 교차타당법은 구축된 모형을 검증하기 위한 표본을 분할(partitioning), 즉 검증용 표본이 훈련용 표본과 겹치지 않도록 데이터를 구성한다. 반면 부스트랩은 중복(replacement)을 허용하는 통계적 복원(replacement) 추출 절차를 기반으로 한다. 먼저 전체 N개인 분석용 자료에서 N개의 표본을 복원 추출(with replacement)하여 모형 구축용 데이

터 셋을 구성한 후, 추출되지 않은 나머지 개체를 이용하여 구축된 모형을 평가한다.

위의 과정을 B회 반복하여 총 B회의 모형 구축 및 평가를 수행하는데, 부스트랩 표본의 추출 반복 횟수(B)는 경험적으로 25번 이상의 반복은 결과에 큰 차이가 없다(이영섭, 2003). 부스트랩 방법은 분석용 데이터의 크기가 매우 작은 경우 오류 발생률을 추정하는 최적의 방법으로 알려져 있다.

<그림 Ⅳ-5> 부스트랩 방법



3. 모형 평가 측도

일반적인 모형 평가의 기준은 모형의 설명력의 측도인 결정계수(Cp), 멜로우스C_p 및 아카이케정보기준(Akaike Information Criterion, AIC) 등의 통계량을 통해 측정될 수 있으며, 특히 반응변수가 범주형인 경우 오분류 행렬(confusion matrix)을 통한 여러 가지 방법을 사용할 수 있다. 이외에도 리프트(lift) 도표, 반응률(response rate) 도표 등이 각종 데이터마이닝 도구에서 많이 사용되고 있다(강현철 외, 1999). 본 연구에서는 오분류율, G-Mean, F1 측도, 반응률로 예측 성

능을 평가한다.

정분류율, 오분류율, G-Mean, F1 측도에 대한 설명을 위해서는 오분류 행렬에 대한 이해가 선행되어야 한다. 먼저 분류 규칙 및 분류 모형의 정확도를 비교 및 평가할 때 오분류 행렬은 다음의 표와 같이 작성된다. 아래의 표에서 "실제"라고 되어 있는 부분은 실제 데이터에서 범주가 0 또는 1이라는 것을 의미하고, "예측"에서의 0과 1은 의사결정나무모형, 로지스틱회귀모형, 신경망모형, 랜덤포레스트모형 등 다양한 기계학습 모형에 의한 예측 결과에 의해 구분된 범주를 의미한다.

<표 Ⅳ-1> 오분류 행렬

구분		예 측		
		0	1	합계
실제	0	n_{00}	n_{01}	n_{0+}
	1	n_{10}	n_{11}	n_{1+}
	합계	n_{+0}	n_{+1}	n

위의 표를 이용해 정분류율, 오분류율에 대해 간략히 설명하면 다음과 같다. 정분류율은 실제 1 또는 0인 범주가 예측 결과 똑같은 범주로 분류되는 비율을 말하며, 오분류율은 실제 1이 0으로 실제 0이 1로 분류되는 비율을 의미한다. 위의 표에서 n_{00}, n_{11}이 정분류되는 개수이며, n_{01}과 n_{10}은 오분류가 되는 개수를 나타내고, 정분류율과 오분류율은 다음의 식으로 표현된다.

$$\text{정분류율} = (n_{00} + n_{11})/n \times 100 \qquad \text{(식 20)}$$

$$\text{오분류율} = (n_{10} + n_{01})/n \times 100 \quad \text{(식 21)}$$

오분류율은 전체 자료를 얼마나 잘못 분류하는가의 문제이므로 값이 작을수록 좋은 모형이다. 즉 설명력이나 예측력이 높은 모형은 정분류율은 높고 오분류율은 낮게 나타나는 것이 이상적이다. 오분류율은 다음과 같이 확률변수의 식으로 나타낼 수도 있다. 확률변수 X는 0 이나 1인 실제값이고, 확률변수 E는 0 이나 1의 값을 가지는 예측값이라고 할 때 오분류율은 다음과 같다.

$$\text{오분류율} = \Pr(X=1, E=0) + \Pr(X=0, E=1) \quad \text{(식 22)}$$

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 다음과 같다.

$$G-mean = \sqrt{\frac{n_{00}}{n_{0+}} \times \frac{n_{11}}{n_{1+}}} \quad \text{(식 23)}$$

$F1$ 측도(measure)는 어떤 특정한 계급의 성공적인 분류가 다른 계급의 분류에 비해 훨씬 중요한 경우 사용되는 측정 기준이다. $F1$ 측도는 특정 계급, 특히 우량과 불량 간 불균형인 경우 소수계급에 주된 관심을 가지고 있으며, 이 값이 크다는 것은 특정 계급에 대한 예측 성능이 좋다는 것을 의미한다(Chawla 외, 2003). $F1$측도를 산출하기 위한 수식은 다음과 같다.

$$F1 = \frac{2rp}{(r+p)} = \frac{2}{1/r+1/p} = \frac{2 \cdot n_{11}}{n_{1+}+n_{+1}} \quad \text{(식 24)}$$

, p = 실제1, 예측1 정분류 빈도/예측1의 빈도

, r = 민감도 = 실제1, 예측1 정분류 빈도/실제1의 빈도

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 특정 계급에 대한 예측 성능이 좋음을 의미하므로, 이 두 가지 측도를 최대화하는 모형을 구축하여야 한다.

전술하였듯이 Precision과 Recall은 그 값이 클수록 구축된 모형에 의한 소수계급의 예측 성능이 좋음을 의미한다. 그러나 두 값은 서로 상충관계(trade-off)를 가진다. Tan 외(2006)는 Precision과 Recall의 값을 모두 고려했을 때 최대화하는 모형을 구축하는 것이 좋다고 하였으며, 이 두 값을 모두 최대화하는 측도로, 다음과 같은 F_β 측도를 제안하였다.

$$F_\beta = \frac{(\beta^2+1)rp}{r+\beta^2 p} = \frac{(\beta^2+1)\times n_{11}}{(\beta^2+1)n_{11}+\beta^2 n_{01}+n_{10}} \qquad \text{(식 25)}$$

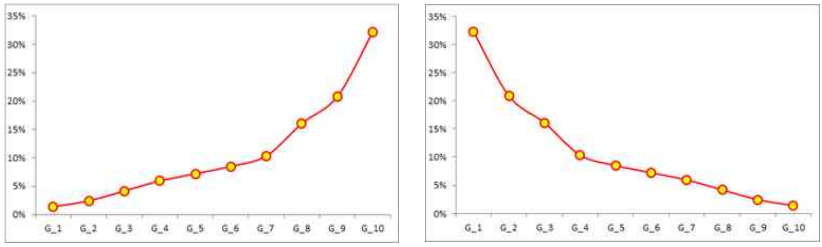
(식 25)에서 β값은 Precision과 Recall의 상대적인 중요도를 나타내는 것으로, β의 값이 작을수록 F_β는 Precision에 가까워지고, 반대로 β 값이 커질수록 F_β는 Recall에 가까워진다. 대부분의 분류 모형에서 Precision의 감소를 최소화하면서 Recall을 최대화하는 것이 중요하다. (식 25)에서 β값으로 1을 사용하는 경우를 $F1$측도라고 한다.

반응률은 훈련용 자료를 이용해 구축된 평가모형을 통해 산출된 사후확률을 정렬하여 N개의 구간으로 등분한 후, 각 구간에 포함된 종속변수의 특정 범주의 빈도를 이용해 산출한다. 이와 같이 계산된

반응률은 도표를 통해 모형의 성능을 명확히 확인할 수 있는데, <그림 Ⅳ-6>에서와 같이 사후확률이 가장 큰 구간에서 가장 낮은 구간으로 갈수록 반응률이 높게 나타나다가 급격하게 감소하는 형태 또는 그 반대인 경우 좋은 예측 판별력을 가진 모형이다(강현철 외, 1999; 박주완, 2018 재인용).

$$\text{반응률} = \frac{\text{일정 } N \text{ 등분내 범주 1 빈도}}{\text{일정 } N \text{ 등분내 전체 빈도}} \times 100 \qquad \text{(식 26)}$$

<그림 Ⅳ-6> 반응률 도표



Ⅴ. 분석 개요

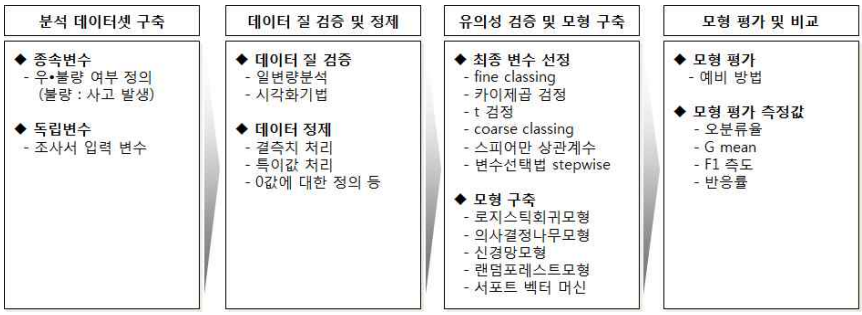
1. 분석 과정

본 연구는 다양한 빅데이터 분석 기법을 이용하여 소상공인의 신용평가모형을 구축하는 것이 주요 목적이다. 이 목적을 달성하기 위한 분석 과정은 분석 데이터셋(data set) 구축, 데이터 질 검증 및 정제, 변수 유의성 검증 및 모형 구축, 모형 평가 및 비교의 4단계로 진행한다. 각 단계별 분석 과정을 살펴보면 다음과 같다.

첫 번째 단계인 분석 데이터셋 구축 과정에서는 종속변수인 우불량에 대한 정의를 수행하고 차주가 신용보증을 받을 당시 조사서에 입력되는 자료들을 독립변수로 정의한다. 이 때 불량의 기준은 보수적인 모형 구축을 위해 사고 발생을 불량으로 정의한다. 두 번째 단계에서는 모형 구축에 사용할 전체 독립변수에 대한 질 검증을 수행하는데, 이때 빈도분석, 평균 분포 등 수치적인 일변량 통계 분석과 그래프를 이용한 시각화 기법을 이용한다. 이와 같은 분석을 통해 결측치, 특이값 등에 대한 정제 기준을 설정하고 데이터를 정제하며 모형 구축에 사용 가능한 독립변수를 선정한다. 세 번째 단계인 변수의 유의성 검정 및 모형 구축 단계에서는 모형 구축에 사용할 최종적인 변수를 선정하는데, 이때 fine 및 coarse classing, 카이제곱 검정, t 검정, 스피어만 상관계수, 단계적 선택법 등을 이용한다. 그리고 모형 구축은 전술한바와 같이 5가지의 대표적인 기계학습 기법인 로지스틱 회귀모형, 의사결정나무모형, 신경망모형, 렌덤포레스트모형, 서포트벡터머신을 이용하여 소상공인 신용평가모형을 구축한다. 마지막 단계인 모형 평가 및 비교에서는 예비 방법을 이용하여 각 구축된 모형의

오분류율, G-mean, F1 측도, 반응률을 비교한 후 예측 성능과 안정성이 가장 우수한 모형을 선택한다.

<그림 V-1> 분석 및 모형 구축 과정



2. 분석 변수

본 연구의 모형 구축 대상은 16개 지역신용보증재단에서 2017년 7월부터 2019년 6월까지 2년 동안 소상공인 신용평가모형을 통해 평가를 받은 차주 136,189개이다. 최종적인 분석 대상은 통상적으로 종속변수인 사고 여부의 판별이 불가능한 경우, 즉 종속변수가 결측치(missing value)이거나 “-999,999,999” 등 특수값(special value)이 있는 표본은 분석에서 제외하지만, 종속변수에 결측이 존재하지 않아 최종적인 분석 대상은 136,189개의 차주를 모두 활용한다.

그리고 소상공인 신용평가모형 구축을 위해 최초 총 37개의 변수를 이용한다. 이중에서 고객번호는 데이터 병합 및 구분 등을 위한 키변수(key variable)이므로 분석에는 활용하지 않으며, 종속변수는 사고 날짜를 이용하여 사고 여부를 생생하고, 독립변수는 총 35개인데 이는 모두 지역신용보증재단에서 신용보증을 받을 당시 전산에 입력되는 내부정보이다. 결과적으로 모형 구축에 사용된 변수의 출처는

신용보증재단 내부 자료만을 이용하며, NICE CB 요약 정보 등 외부 정보는 이용하지 않는다. <표 Ⅴ-1>은 본 논문에서의 모형 구축을 위해 최초로 사용하는 변수 목록이다.

<표 Ⅴ-1> 모형 구축을 위한 변수

N0	변수종류	변수명	변수척도	비고
1	키변수	고객번호	명목	
2	종속변수	사고 여부	이진	Y=0:사고무,Y=1사고유
3	독립변수	고객형태	명목	
4		업종	명목	업종 : 7개
5		종업원수	비율	명
6		주사업장소유여부	명목	
7		주사업장임차보증 금액	비율	만원
8		주사업장월세 금액	비율	천원
9		실거주지소유여부	명목	
10		실거주지임차보증 금액	비율	만원
11		실소유지월세 금액	비율	천원
12		차입금운전	비율	백만원
13		차입금시설	비율	백만원
14		차입금기타	비율	백만원
15		기보증잔액재단	비율	백만원
16		기보증잔액신보	비율	백만원
17		기보증잔액기보	비율	백만원
18		기보증잔액개인	비율	백만원
19		담보제외차입기관수	비율	개
20		현금서비스금액	비율	원
21		보유부동산	명목	
22		업력	비율	월
23		거주기간	비율	월
24		월평균매출액	비율	원
25		월영업이익	비율	원
26		월배우자소득	비율	원

N0	변수종류	변수명	변수척도	비고
27		월기타수익	비율	원
28		소유부동산금액	비율	원
29		임대보증금사업장	비율	원
30		임대보증금주택	비율	원
31		예적금금액	비율	원
32		유가증권금액	비율	원
33		재고자산	비율	원
34		고정자산	비율	원
35		권리금	비율	원
36		기타현금	비율	원
37		직권말소	이진	직권말소=0, 나머지 1

3. 분석 변수 기초 분포

종속변수인 사고 여부와 독립변수들 중 고객 형태, 업종, 주사업장 소유 여부, 실거주지 소유 여부, 보유 부동산, 직권말소 여부는 범주형 척도를 가진 자료이다. 그러므로 빈도표를 이용하여 자료의 기초분포를 확인하여야 한다.

먼저 종속변수인 사고 여부의 분포는 사고가 전혀 발생하지 않은 차주(사고 무)는 전체 분석 대상 중 98.1%(133,566개)이고 사고가 한 번이라도 발생한 차주(사고 유)는 1.9%(2,623개)로 계급불균형 자료(class imbalanced data)이다. 이와 같이 계급불균형인 자료는 오버샘플링(over-sampling)을 적용하여 계급 간 균형을 맞추어서 분석을 수행하거나, 분류절단값(cutoff value)을 조절하여 분석을 수행하는 것이 일반적인데, 본 연구에서는 분류절단값을 계급이 균형일 때 사용하는 50% 대신에 사고 유의 비중인 1.93%를 사용하여 정분류율, G-mean, F1값을 산출한다.

다음으로 범주형 척도를 가진 독립변수들의 기초분포를 살펴보면

다음과 같다. 고객 형태는 개인사업자 94.4%, 법인사업자 5.6%로 대부분 개인사업자로 나타나고 있는데, 이는 지역신용보증재단에서 신용보증을 받는 차주의 대부분이 개인사업자이기 때문이다. 업종은 제조업 7.4%, 서비스업 17.3%, 도소매업 32.5%, 음식숙박업 28.0%, 건설업 6.6%, 운수업 5.8%, 기타업 2.5%로 대부분 서비스업, 도소매업, 음식숙박업에 해당하고 있다. 이 또한 지역신용보증재단에서 신용보증을 받는 차주인 소상공인이 이 업종에 해당하기 때문이다. 주사업장 소유 여부는 임차인 경우가 81.6%, 자가는 13.5%로 나타나 대부분의 차주가 사업장을 임차하여 영업을 하고 있음을 알 수 있다. 실거주지 소유 여부는 임차 49.2%, 자가 43.2%의 순으로 임차와 자가 간 비중의 차이가 크지 않게 나타났다. 보유 부동산의 종류는 보유 부동산이 없는 경우가 48.9%로 가장 높은 가운데 아파트 28.9%, 단독주택 7.0%의 순이다. 직권말소는 직권말소가 아닌 경우(직권말소 부) 99.7%, 직권말소가 0.3%로 대부분 직권말소 상태가 아닌 것으로 나타났다.

<표 Ⅴ-2> 범주형(명목, 이진, 순위) 척도 변수 분포

변수	변수값	빈도(개)	비율(%)	결측치 수(개)
사고 여부	사고무(0)	133,566	98.1	0
	사고유(1)	2,623	1.9	
고객 형태	개인사업자	125,526	94.4	3,235
	법인사업자	7,428	5.6	
업종	1.제조업	10,064	7.4	0
	2.서비스업	23,513	17.3	
	3.도소매업	44,312	32.5	
	4.음식숙박업	38,067	28.0	
	5.건설업	9,029	6.6	
	6.운수업	7,871	5.8	
	7.기타업	3,333	2.5	

변수	변수값	빈도(개)	비율(%)	결측치 수(개)
주사업장 소유여부	가족소유	86	0.1	2
	기타	3,413	2.5	
	무점포	11	0.0	
	임차	111,108	81.6	
	임차(보증금20백만원초과)	3,012	2.2	
	자가	18,416	13.5	
	전차	141	0.1	
실거주지 소유여부	가족소유	93	0.1	2,754
	기타	8,887	6.7	
	임차	65,683	49.2	
	임차(보증금20백만원초과)	1,084	0.8	
	자가	57,685	43.2	
	전차	3	0.0	
보유 부동산	다가구	1,071	0.8	6
	다세대	4,645	3.4	
	단독주택	9,462	7.0	
	아파트	39,353	28.9	
	없음	66,645	48.9	
	임야 기타부동산	160	0.1	
	임야 혹은 기타부동산	14,847	10.9	
직권 말소	직권말소여(0)	368	0.3	0
	직권말소부(1)	135,821	99.7	

분석을 위한 독립변수 중 연속형 척도를 가진 변수는 종업원 수, 주사업장 임차보증 금액 등 총 29개 변수이다. 연속형 변수인 경우 수치적인 통계량인 평균, 표준편차, 최소값, 최대값을 이용해 기초분포를 살펴보는 것이 일반적이다. 분석 대상 차주의 종업원 수는 평균 1.1명인 것으로 나타나고 있지만 최대값이 48,857명인 경우가 존재하고 있는데 이는 입력 오류인 경우로 판단되며, 0 및 결측치의 비중이

71.0%로 나타났다. 주사업장 임차보증 금액의 평균은 1,429만원이며, 분석 자료 중 32.5%가 0 또는 결측치인 것으로 나타났다. 주사업장 월세 금액의 평균은 984천원이고, 분석 자료 중 0 또는 결측치의 비율은 31.5%이다. 실거주지 임차보증 금액의 평균은 597만원이며, 분석 자료 중 0 또는 결측치의 비율은 75.0%이고 이중 대부분이 0값이므로 실제 평균이라고 단언할 수 없다. 실소유지 월세 금액 평균은 99천원으로 나타나고 있는데, 이는 0 또는 결측치의 비율이 81.6%이기 때문이므로 실제 평균인지를 확인하기 어렵다. 담보 제외 차입기관 수는 평균 1,123.1개이고 최대값이 65,940,000개인 것으로 나타나고 있는데 이는 명백히 자료 입력의 오류인 것으로 판단된다. 그러므로 담보 제외 차입기관 수의 평균은 명백하게 오류이다. 업력은 평균 77.8개월인데 0 또는 결측치의 비율이 0.05%로 자료의 질이 양호한 것으로 사료된다. 거주기간은 평균 80.3개월로 0 또는 결측치의 비율이 0.3%로 분석에 활용하는데 큰 문제가 없는 것으로 판단된다. 월평균 매출액은 1천 8백 9십만원, 월 영업이익 평균은 3백 3십만원이고, 0 또는 결측치의 비율이 각각 1.0%, 1.4%로 비교적 자료의 질이 양호한 것으로 판단된다. 소유부동산 금액의 평균은 1천 6백 6십만원인데 0 또는 결측치의 비율이 52.9%인 것으로 나타나 자료의 질이 우수하다는 결론을 내기에는 무리가 있다.

<표 Ⅴ-3>을 보면 차입금 시설, 차입금 기타, 기보증잔액 재단, 기보증잔액 신보, 기보증잔액 기보, 기보증잔액 개인, 현금서비스 금액, 월 배우자 소득, 월 기타 수익, 임대보증금 사업장, 임대보증금 주택, 예적금 금액, 유가증권 금액, 재고자산, 고정자산, 권리금, 기타 현금은 0 또는 결측치의 비율이 90%를 상회하고 있음을 확인할 수 있다. 이 중에서 차입금 시설, 차입금 기타, 기보증잔액 재단, 기보증잔액 신보, 기보증잔액 기보, 기보증잔액 개인, 현금서비스 금액의 경우는

외부의 자료와 연계되어 있는 값이기 때문에 대부분의 0값이 실제 0일 가능성이 매우 높을 것으로 판단된다. 그러므로 분석에 활용하는 데에 있어 큰 문제가 없을 것으로 사료된다. 그러나 월 배우자 소득, 월 기타 수익, 임대보증금 사업장, 임대보증금 주택, 예적금 금액, 유가증권 금액, 재고자산, 고정자산, 권리금, 기타 현금의 경우 대부분의 값(약 98% 이상)이 0이나 결측치인데, 이 변수들의 실제 값이 0인지 그렇지 않으면 입력하지 않았기 때문에 0의 값을 가지고 있는지를 확인할 필요가 있다. 결론적으로 이 변수들에 대한 평균 등의 기초통계량은 큰 의미가 없을 가능성이 높으며, 신용평가모형을 구축하기 위한 변수로 활용하기에도 적절하지 않을 가능성이 높다.

이와 같은 자료상의 한계점이 존재함에도 불구하고 모든 변수를 모형 구축에 활용하고자 한다. 이유는 계급세분화(classing) 기법을 사용하여 변수 값을 표준화할 경우 결측치나 0값은 불량률의 기초하여 하나의 계급으로 표준화가 가능하며, 모형 구축을 위한 변수를 선택하는 다양한 분석 과정을 통해 통계적으로 우불량 여부에 대해 통계적으로 유의미하지 않은 변수는 배제되기 때문이다.

<표 Ⅴ-3> 연속형(구간, 비율) 척도 변수 분포

변수명	평균	표준편차	최소값	최대값	0 및 결측치 비율
종업원수	1.1	132.4	0	48,857	71.04%
주사업장임차보증금액	1,429.4	1,829.7	0	100,000	32.46%
주사업장월세금액	984.0	1,872.2	0	291,666	31.49%
실거주지임차보증금액	596.5	1,512.5	0	50,000	74.98%
실소유지월세금액	98.9	375.5	0	28,000	81.63%
차입금운전	7.2	43.0	0	2,859	86.46%
차입금시설	0.3	14.4	0	2,000	99.81%
차입금기타	1.7	27.8	0	3,607	99.07%

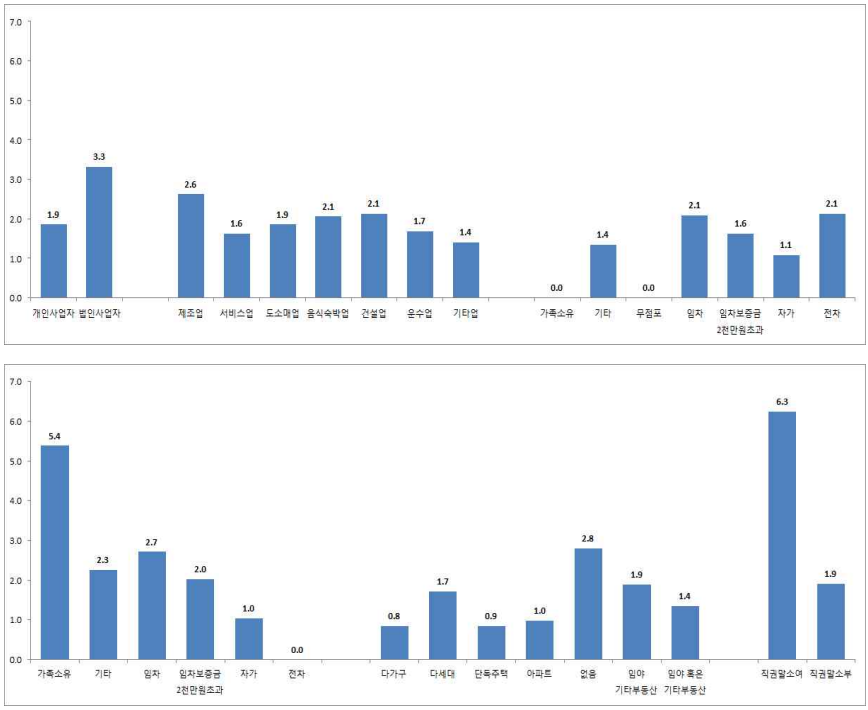
변수명	평균	표준편차	최소값	최대값	0 및 결측치 비율
기보증잔액재단	89,139.6	1,656,304.0	0	90,000,000	73.09%
기보증잔액신보	38,522.5	1,540,256.6	0	99,000,000	96.71%
기보증잔액기보	8,456.9	821,957.7	0	98,000,000	99.48%
기보증잔액개인	0.0	0.1	0	12	99.89%
담보제외차입기관수	1,123.1	226,787.0	0	65,940,000	25.90%
현금서비스금액	276,859.1	1,558,815.7	0	77,000,000	91.24%
업력	77.8	74.8	0	822	0.05%
거주기간	80.3	86.5	0	831	0.27%
월평균매출액	18,940,997.7	20,197,193.2	0	99,968,719	1.02%
월영업이익	3,311,958.0	4,899,291.7	-9,054,552	99,900,234	1.39%
월배우자소득	820.5	59,316.8	0	9,000,000	99.97%
월기타수익	21,733.1	351,034.5	0	30,257,993	98.84%
소유부동산금액	16,617,257.1	23,954,386.5	0	99,994,100	52.91%
임대보증금사업장	363,789.0	3,932,554.8	0	97,500,000	98.68%
임대보증금주택	1,029,182.4	6,780,627.8	0	99,800,000	96.47%
예적금금액	168,906.2	2,499,914.4	0	97,391,000	99.00%
유가증권금액	100,839.9	2,352,643.4	0	99,851,028	99.73%
재고자산	400,784.0	3,854,485.8	0	99,176,866	98.42%
고정자산	264,832.5	3,039,272.5	0	99,462,000	98.97%
권리금	17,791.4	823,150.1	0	95,000,000	99.94%
기타현금	722,992.0	5,543,927.7	0	99,810,000	97.25%

다음의 <그림 Ⅴ-2>와 <표 Ⅴ-4>는 모든 범주형 독립변수의 범주별 사고 비율, 즉 불량률을 나타낸 것이다. 고객 형태별로 사고 발생 비율을 살펴보면 개인사업자 1.9%, 법인사업자 3.3%로 법인사업자의 사고 발생 비율이 더 높게 나타나고 있다. 업종별 사고 발생 비율은 제조업 2.6%, 서비스업 1.6%, 도소매업 1.9%, 음식숙박업 2.1%, 건설업 2.1%, 운수업 1.7%, 기타업 1.4%로 제조업, 음식숙박업, 건설업의

사고 발생 비율이 상대적으로 높은 2% 대로 나타났다. 주사업장 소유에서는 임차와 전차인 경우의 사고 비율이 2.1%로 가장 높았으며, 다음으로 임차보증금 2천만원 초과 1.6%, 기타 1.4%, 자가 1.1% 등의 순으로 나타나고 있다. 실거주지 소유는 가족 소유인 경우 사고 비율이 5.4%로 가장 높았으며, 다음으로 임차 2.7%, 기타 2.3%, 임차보증금 2천만원 초과 2.0%, 자가 1.0%의 순으로 나타났다. 보유 부동산은 임야 및 기타 부동산 3.3%, 없는 경우가 2.8%로 상대적으로 높았으며, 다음으로 다세대 1.7%, 아파트 1.0%의 순이다. 직권말소는 직권말소 상태인 경우 사고 비율이 6.3%로 직권말소가 아닌 경우 보다 3.3배 높게 나타나고 있다.

<그림 V-2> 각 변수의 범주별 불량(사고 유) 비율

(단위 : %)



<표 Ⅴ-4> 범주형 독립변수별 사고 유무 분포

변수	범주	표본수	빈도(개)		비율(%)	
			사고무	사고유	사고무	사고유
고객 형태	개인사업자	125,526	123,188	2,338	98.1	1.9
	법인사업자	7,428	7,182	246	96.7	3.3
업종	1.제조업	10,064	9,800	264	97.4	2.6
	2.서비스업	23,513	23,133	380	98.4	1.6
	3.도소매업	44,312	43,489	823	98.1	1.9
	4.음식숙박업	38,067	37,283	784	97.9	2.1
	5.건설업	9,029	8,837	192	97.9	2.1
	6.운수업	7,871	7,738	133	98.3	1.7
	7.기타업	3,333	3,286	47	98.6	1.4
주사업장 소유여부	가족소유	86	86	0	100.0	0.0
	기타	3,413	3,367	46	98.7	1.4
	무점포	11	11	0	100.0	0.0
	임차	111,108	108,782	2,326	97.9	2.1
	임차보증금2천만원초과	3,012	2,963	49	98.4	1.6
	자가	18,416	18,217	199	98.9	1.1
	전차	141	138	3	97.9	2.1
실거주지 소유여부	가족소유	93	88	5	94.6	5.4
	기타	8,887	8,686	201	97.7	2.3
	임차	65,683	63,905	1,778	97.3	2.7
	임차보증금2천만원초과	1,084	1,062	22	98.0	2.0
	자가	57,685	57,087	598	99.0	1.0
	전차	3	3	0	100.0	0.0
보유 부동산	다가구	1,071	1,062	9	99.2	0.8
	다세대	4,645	4,565	80	98.3	1.7
	단독주택	9,462	9,382	80	99.2	0.9
	아파트	39,353	38,970	383	99.0	1.0
	없음	66,645	64,777	1,868	97.2	2.8
	임야 기타부동산	160	157	3	98.1	1.9
	임야 혹은 기타부동산	14,847	14,647	200	98.7	1.4
직원 말소	직권말소여(0)	368	345	23	93.8	6.3
	직권말소부(1)	135,821	133,221	2,600	98.1	1.9

연속형 독립변수는 사고 유무별로 평균을 비교하였는데, 종업원수, 주사업장 임차보증 금액, 업력, 월평균 매출액, 월 영업이익 등의 변수에서 사고가 없는 차주의 평균이 높게 나타났다. 이를 통해 보증사고가 없는 차주의 경영 및 재무 상태가 비교적 양호하다는 것을 유추할 수 있다.

<표 Ⅴ-5> 연속형 독립변수별 사고 유무 분포

변수	사고무		사고유	
	평균	표준편차	평균	표준편차
종업원수	1.1	133.7	0.7	2.0
주사업장임차보증금액	1,430.7	1,832.7	1,364.3	1,665.3
주사업장월세 금액	982.0	1,868.7	1,086.7	2,039.6
실거주지임차보증금액	595.6	1,512.9	644.7	1,493.3
실소유지월세 금액	97.5	375.0	171.4	395.4
차입금운전	7.2	43.2	5.0	25.1
차입금시설	0.3	14.4	0.2	11.9
차입금기타	1.6	27.4	2.0	43.4
기보증잔액재단	90,225.0	1,668,492.6	33,868.5	822,765.3
기보증잔액신보	38,066.1	1,523,019.2	61,762.7	2,250,010.1
기보증잔액기보	8,622.9	829,988.6	0.6	6.1
기보증잔액개인	0.0	0.1	0.0	0.2
담보제외차입기관수	1,145.1	229,003.0	2.3	1.8
현금서비스금액	268,871.1	1,537,681.9	683,599.3	2,365,481.6
업력	78.3	75.0	52.7	52.2
거주기간	80.3	86.5	79.3	89.5
월평균매출액	19,004,648.9	20,216,208.0	15,699,953.4	18,926,806.3
월영업이익	3,318,532.3	4,880,573.3	2,977,205.8	5,763,920.3
월배우자소득	769.2	54,600.4	3,431.2	175,729.0
월기타수익	21,709.8	348,863.9	22,920.3	447,945.1

변수	사고무		사고유	
	평균	표준편차	평균	표준편차
소유부동산금액	16,764,300.1	24,002,879.7	9,130,004.2	19,959,673.1
임대보증금사업장	368,679.8	3,959,973.3	114,754.1	2,093,947.1
임대보증금주택	1,040,191.3	6,818,493.3	468,623.7	4,409,552.4
예적금금액	170,086.9	2,509,447.7	108,788.5	1,953,375.1
유가증권금액	101,638.0	2,358,776.2	60,198.6	2,015,826.3
재고자산	394,707.5	3,814,983.3	710,195.1	5,495,214.2
고정자산	263,023.7	3,023,063.5	356,934.7	3,773,250.7
권리금	17,017.8	798,596.5	57,186.4	1,644,562.3
기타현금	720,082.8	5,532,118.1	871,126.1	6,114,519.4

Ⅵ. 분석 결과

1. 변수 선택

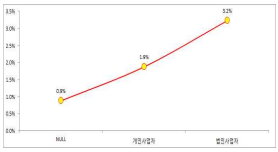
앞 장의 <그림 Ⅴ-1>에서 설명한 바와 같이 모형을 구축하는 단계에서는 모형 구축에 필요한 변수를 선정하여야 한다. 전술하였듯이 변수 선정의 첫 번째 단계는 계급세분화 수행 후 KS통계량이 0.1 이상인 값을 가지는 변수를 선정한다. 그러나 본 연구에서는 변수 선택 시 우불량 간 변별력이 크게 낮지 않고 좀 더 많은 분석 변수 사용을 위해 0.05를 기준으로 한다. 그리고 이를 보완하기 위해 범주형 변수의 원천자료에 대한 유의성 검정인 카이제곱 검정과 연속형 변수의 원천자료에 대한 유의성 검정인 t-검정을 수행하여 p-값이 0.05 이하인 변수를 1차적으로 선정한다. 세 가지 방법에 의해 1차로 선정된 독립변수는 다음 <표 Ⅵ-1>과 같다.

1차적으로 선정된 변수는 최초 변수 35개 중 23개로 KS통계량이 0.05 이상인 변수는 업종 외 15개, 카이제곱 및 t-검정 결과 통계적으로 유의한 변수는 고객 형태 외 19개로써 두 가지 방법 중 단 하나의 방법에서라도 유의하게 판명된 변수가 23개이다. 그 결과 1차적으로 고객 형태, 업종, 주사업장 소유 여부, 주사업장 임차보증 금액, 주사업장 월세 금액, 실거주지 소유 여부, 실거주지 임차보증 금액, 실소유지 월세 금액, 차입금 운전, 기보증잔액 재단, 기보증잔액 기보, 담보 제외 차입 기관 수, 현금서비스 금액, 보유 부동산, 업력, 거주 기간, 월평균 매출액, 월 영업이익, 소유 부동산 금액, 임대보증금 사업장, 임대보증금 주택, 재고자산, 직권말소이다. 다음 단계는 성김화에 의한 재범주화, 단계적 선택법 이용, 다중공선성을 확인하는 것이다.

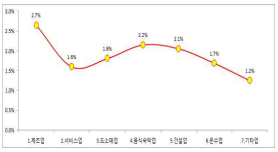
<표 Ⅵ-1> 1차 변수 선택 결과

변수	KS 통계량	카이제곱 & t 검정 p값	1차 변수 선택
고객형태	0.04	**<.0001**	O
업종	**0.07**	**<.0001**	O
종업원수	0.01	0.305	
주사업장소유여부	**0.07**	**<.0001**	O
주사업장임차보증 금액	**0.09**	**0.044**	O
주사업장월세 금액	**0.09**	**0.009**	O
실거주지소유여부	**0.22**	**<.0001**	O
실거주지임차보증 금액	**0.14**	0.101	O
실소유지월세 금액	**0.16**	**<.0001**	O
차입금운전	0.04	**<.0001**	O
차입금시설	0.00	0.754	
차입금기타	0.01	0.704	
기보증잔액재단	**0.06**	**0.001**	O
기보증잔액신보	0.01	0.591	
기보증잔액기보	0.01	**0.000**	O
기보증잔액개인	0.00	0.217	
담보제외차입기관수	**0.19**	0.068	O
현금서비스금액	**0.11**	**<.0001**	O
보유부동산	**0.24**	**<.0001**	O
업력	**0.22**	**<.0001**	O
거주기간	**0.10**	0.545	O
월평균매출액	**0.11**	**<.0001**	O
월영업이익	**0.11**	**0.003**	O
월배우자소득	0.00	0.438	
월기타수익	0.01	0.891	
소유부동산금액	**0.22**	**<.0001**	O
임대보증금사업장	0.01	**<.0001**	O
임대보증금주택	0.02	**<.0001**	O
예적금금액	0.00	0.114	
유가증권금액	0.00	0.299	
재고자산	0.01	**0.004**	O
고정자산	0.00	0.205	
권리금	0.00	0.212	
기타현금	0.01	0.210	
직권말소	0.01	**<.0001**	O

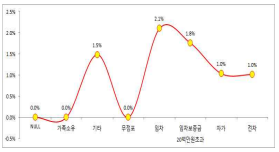
<그림 Ⅵ-1> 1차 선택 변수의 fine classing 결과 그래프



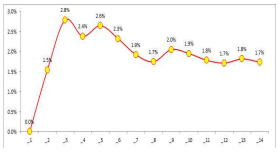
(1)고객형태



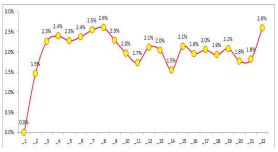
(2)업종



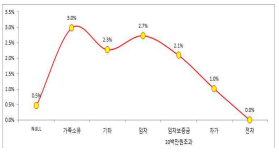
(3)주사업장소유여부



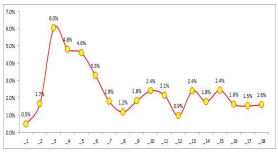
(4)주사업장임차보증 금액



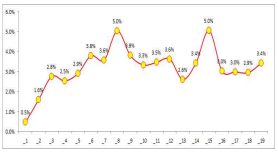
(5)주사업장월세 금액



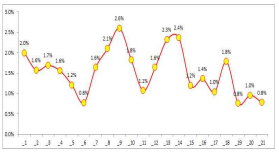
(6)실거주지소유여부



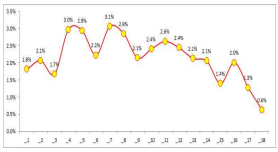
(7)실거주지임차보증 금액



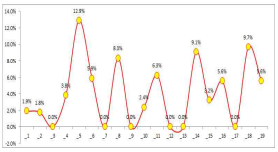
(8)실소유지월세 금액



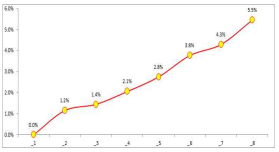
(9)차입금운전



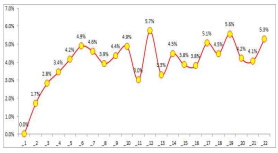
(10)기보증잔액재단



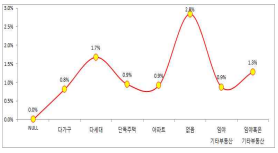
(11)기보증잔액기보



(12)담보제외차입기관수

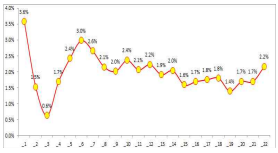


(13)현금서비스금액

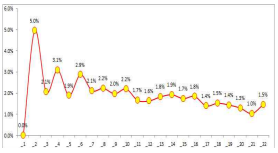


(14)보유부동산

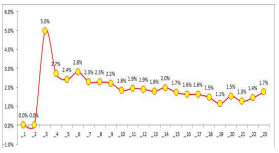
(15)업력



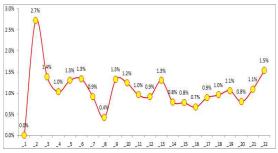
(16)거주기간



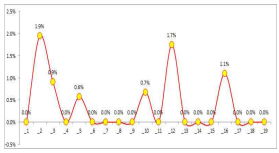
(17)월평균매출액



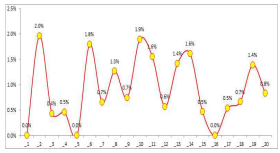
(18)월영업이익



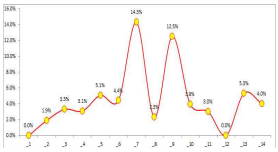
(19)소유부동산금액



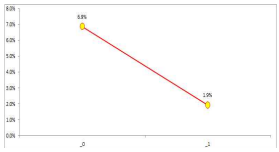
(20)임대보증금사업장



(21)임대보증금주택



(22)재고자산

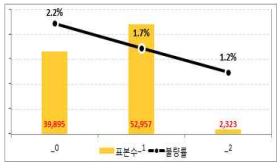


(23)직권말소

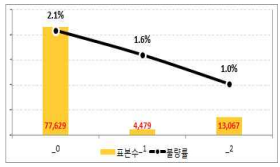
다음의 그림은 1차적으로 계급세분화에 의해 선택된 변수에 대해 성김화(coarse classing)를 수행하여 계급을 재범주화 한 후, 재범주화된 변수 23개의 범주별 불량률을 나타낸 것이다. 성김화에 의해 재범주화된 변수는 불량률에 기초하여 순서형 변수로 변환이 되어 _0, _1, _2로 갈수록 불량률이 점차 낮아지고 있음을 확인할 수 있다.

<그림 Ⅵ-2> coarse classing 결과 그래프

(1)고객형태

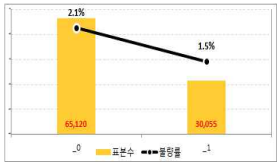


(2)업종

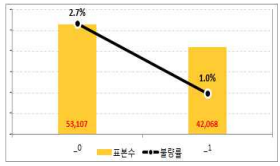


(3)주사업장소유여부

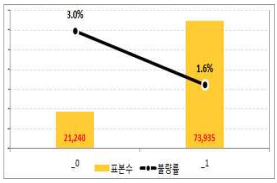
(4)주사업장임차보증 금액



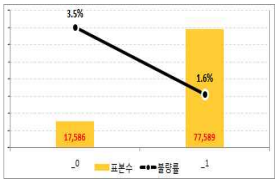
(5)주사업장월세 금액



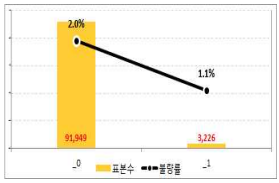
(6)실거주지소유여부



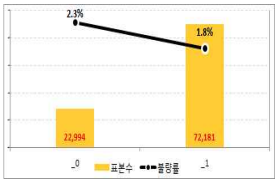
(7)실거주지임차보증 금액



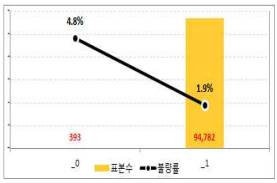
(8)실소유지월세 금액



(9)차입금운전



(10)기보증잔액재단



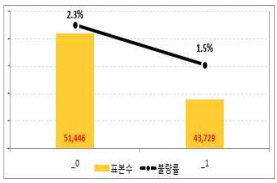
(11)기보증잔액기보

(12)담보제외차입기관수

(13)현금서비스금액

(14)보유부동산

(15)업력

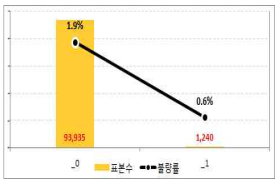


(16)거주기간

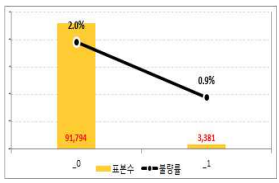
(17)월평균매출액

(18)월영업이익

(19)소유부동산금액

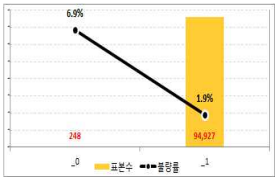


(20)임대보증금사업장



(21)임대보증금주택

(22)재고자산



(23)직권말소

성김화가 완료된 변수를 로지스틱회귀모형에서의 단계적 선택법에 적용한 결과 17개의 변수가 선택되었는데, 선택된 변수는 고객 형태, 업종, 주사업장 임차보증 금액, 실거주지 소유 여부, 실거주지 임차보증 금액, 실소유지 월세 금액, 차입금 운전, 기보증잔액 재단, 기보증잔액 기보, 담보 제외 차입 기관 수, 현금서비스 금액, 보유 부동산, 업력, 거주 기간, 월평균 매출액, 재고자산, 직권말소이다. 변수에 대한 회귀계수 추정치를 보면 실거주지 임차보증 금액의 회귀계수 추정치가 -0.32로 음(-)값을 가지고 있다. 이는 다중공선성이 의심되는 결과이므로 상관계수를 산출하여 다중공선성을 확인하여야 한다.

<표 Ⅵ-2> 단계적 선택법 적용 결과

변수	회귀계수 추정치	Standard Error	Wald Chi-Square	p값
Intercept	-2.25	0.40	32.07	<.0001
고객형태	0.70	0.09	62.22	<.0001
업종	0.23	0.05	27.06	<.0001
주사업장임차보증 금액	0.14	0.05	7.29	0.0069
실거주지소유여부	0.36	0.08	22.68	<.0001
실거주지임차보증 금액	-0.32	0.09	14.28	0.0002
실소유지월세 금액	0.55	0.09	41.93	<.0001
차입금운전	0.48	0.18	7.57	0.0059
기보증잔액재단	0.22	0.06	16.08	<.0001
기보증잔액기보	0.83	0.25	11.14	0.0008
담보제외차입기관수	0.58	0.03	285.54	<.0001
현금서비스금액	0.66	0.06	109.40	<.0001
보유부동산	0.35	0.04	82.49	<.0001
업력	0.64	0.04	272.88	<.0001
거주기간	0.23	0.05	20.79	<.0001
월평균매출액	0.34	0.04	78.45	<.0001
재고자산	0.70	0.15	22.28	<.0001
직권말소	1.01	0.26	15.06	0.0001

17개의 독립변수에 대해 스피어만 상관계수를 이용하여 다중공선성 여부를 확인한 결과, 2차적으로 선택된 17개의 변수 중 실거주지 임차보증 금액과 실소유지 월세 금액의 상관계수가 0.773으로 나타나 0.7을 초과하여 다중공선성이 존재하는 것으로 확인되었다.

<표 Ⅵ-3> 다중공선성 확인 결과

구분	상관계수		
고객형태	월평균매출액	담보제외차입기관수	기보증잔액기보
	-0.168	0.085	0.082
업종	월평균매출액	업력	기보증잔액기보
	-0.06713	0.05072	0.0447
주사업장 임차보증 금액	실소유지월세금액	실거주지임차보증금액	보유부동산
	0.091	0.091	0.087
실거주지 소유여부	보유부동산	실거주지임차보증금액	실소유지월세 금액
	0.641	0.468	0.415
실거주지 임차보증금액	**실소유지월세금액**	실거주지소유여부	보유부동산
	0.773	0.468	0.316
실소유지 월세 금액	**실거주지임차보증금액**	실거주지소유여부	보유부동산
	0.773	0.415	0.288
차입금운전	실거주지소유여부	실거주지임차보증금액	월평균매출액
	0.081	0.046	0.041
기보증잔액재단	담보제외차입기관수	업력	월평균매출액
	0.271	-0.112	-0.102
기보증잔액기보	고객형태	담보제외차입기관수	월평균매출액
	0.082	0.061	-0.045
담보제외 차입기관수	기보증잔액재단	현금서비스금액	고객형태
	0.271	0.155	0.085
현금서비스금액	담보제외차입기관수	월평균매출액	주사업장임차보증금액
	0.155	0.088	0.028
보유부동산	실거주지소유여부	실거주지임차보증금액	실소유지월세금액
	0.641	0.316	0.288
업력	실거주지소유여부	보유부동산	거주기간
	0.176	0.172	0.130
거주기간	업력	실거주지임차보증금액	실거주지소유여부
	0.130	0.097	0.085
월평균매출액	고객형태	기보증잔액재단	보유부동산
	-0.168	-0.102	0.093
재고자산	월평균매출액	차입금운전	실거주지소유여부
	-0.032	0.021	0.020
직권말소	실소유지월세 금액	업력	실거주지임차보증금액
	0.022	0.018	0.012

다중공선성이 존재하는 2개의 변수인 실거주지 임차보증 금액과 실소유지 월세 금액 각각에 대해 로지스틱회귀분석을 수행하여 회귀계수 추정치, 왈드 카이제곱 값, 정분류율, c통계량 값이 더 크게 나타난 변수의 설명력과 변별력이 더 크기 때문에, 이 값들이 큰 변수를 선택한다. 아래의 표를 보면 실소유지 월세 금액의 회귀계수 추정치, 왈드 카이제곱 값, 정분류율, c통계량이 더 크게 나타났으므로 최종 신용평가모형 구축 변수로는 실소유지 월세 금액을 선택하는 것이 타당하다.

<표 Ⅵ-4> 다중공선성 존재 변수 선택을 위한 회귀분석 결과

변수	회귀계수 추정치	Standard Error	Wald Chi-Square	p값	정분류율	c통계량
실거주지 임차보증 금액	0.6249	0.0497	157.8362	<.0001	27.0	0.563
실소유지 월세금액	**0.8319**	**0.0501**	**275.5236**	**<.0001**	**27.6**	**0.578**

이와 같은 과정을 통해 최종적으로 16개의 변수를 이용하여 의사결정나무모형, 로지스틱회귀모형, 신경망모형, 랜덤포레스트모형, SVM 모형을 구축한다. 다음의 표는 최종적으로 선택된 변수의 성김화에 의해 원자료의 계급화 결과인데 모든 변수에서 0, 1, 2로 갈수록 불량률이 낮아진다. 다음의 표에서 0, 1, 2는 성김화에 의해 원자료를 순위형으로 표준화한 것으로써 0~2로 갈수록 불량률은 낮아지고 우량률은 높아진다. 최종적으로 선택된 변수들 중 0, 1, 2 세 개의 계급으로 재범주화된 변수는 업종, 주사업장 임차보증 금액, 보유 부동산, 업력, 월평균 매출액이고, 2개의 계급으로 재범주화된 변수는 고객 형태, 실거주지 소유 여부, 실소유지 월세 금액, 차입금 운전, 기보증잔액 재단, 기보증잔액 기보, 담보 제외 차입 기관 수, 현금서비스 금액, 거주

기간, 재고자산, 직권말소이다. 표의 결과를 보는 방법을 예를 들어 설명하면, 계급이 2개인 고객 형태는 결측치 및 개인사업자인 경우 불량률이 낮아서 계급 1로 구분되었다. 계급이 3개인 업종의 경우 불량률이 가장 높은 0에 해당되는 범주는 제조업, 음식숙박업, 건설업이고 불량률이 가장 높고, 기타업의 불량률이 가장 낮다.

<표 Ⅵ-5> 성김화에 의한 재범주화 결과

구분	0 (더미0)	1 (더미1)	2 (더미2)
고객형태	법인사업자	결측치, 개인사업자	-
업종	제조업, 음식숙박업, 건설업	서비스업, 도소매업, 운수업	기타업
주사업장 임차보증 금액	임차	기타, 임차보증금 2천만원 초과	결측치, 가족소유, 무점포, 자가
실거주지 소유여부	가족소유, 기타, 임차, 임차보증금 2천만원 초과	결측치, 자가, 전차	-
실소유지 월세 금액	0원 초과	결측치, 0원 이하	-
차입금운전	56.6 이하	56.6 초과	-
기보증잔액재단	6 이하	6 초과	-
기보증잔액기보	15.8 이하	15.8 초과	-
담보제외 차입기관수	3 이하	3 초과	-
현금서비스금액	0 이하	0 초과	-
보유부동산	없음	다세대, 임야 및 기타 부동산	결측치, 다가구, 단독주택, 아파트
업력	15 초과 44 이하	0 초과 15 이하, 44 초과 124 이하	0 이하, 124 초과
거주기간	결측치, 16 초과 67 이하, 261 초과	0 이상, 16 이하, 67 초과 261 이하	-
월평균매출액	0 이상 8500000 이하	8500000 초과 38750000 이하	결측치, 38750000 초과
재고자산	0 초과	0 이하	-
직권말소	결측치, 직권말소 부	직권말소 여	-

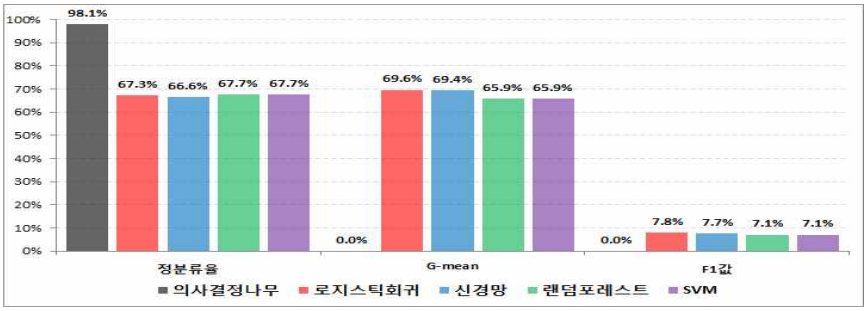
2. 기계학습 기법을 이용한 모형 구축 및 평가

본 절은 최종적으로 선택한 독립변수 16개를 이용하여 대표적인 기계학습 기법인 의사결정나무, 로지스틱회귀, 신경망, 랜덤포레스트, SVM을 이용하여 모형을 구축한 후, 구축된 모형에 대한 평가 결과를 비교한 것이다. 전술하였듯이 모형 구축을 위한 평가는 전체 자료를 2개로 분할하는 예비 방법을 이용하는데, 훈련용 자료(training data)와 평가용 자료(validation data)는 전체 자료에 대해 균등분포(uniform distribution)에 의한 난수를 생성한 후 '훈련용:평가용=7:3'으로 분할한다. 이 때 훈련용 자료는 총 95,147개로써 모형을 구축하는데 사용하며, 구축된 모형의 평가는 41,013개의 평가용 자료를 이용한다. 이 때 모형 구축을 위한 도구로는 기계학습에 가장 대표적으로 사용하고 있는 R을 이용하며, 평가 결과 분석은 SAS를 이용한다.

<그림 Ⅵ-3>과 <표 Ⅵ-6>는 훈련용 자료에 대한 정분류율, G-mean, F1값을 정리한 것이다. 먼저 정분류율은 의사결정나무 98.1%, 로지스틱회귀모형 67.3%, 신경망모형 66.6%, 랜덤포레스트모형 67.7%, SVM모형 21.3%로 의사결정나무모형에 의한 정분류율이 가장 높게 나타나고 있다. 그러나 결과 표를 보면 {실제 1, 예측 1}의 빈도가 0으로 나타나고 있는데, 이는 실제로 보증사고가 발생한 차주(실제 1)는 의사결정나무모형에 의해 보증사고가 없는 우량으로 잘못 분류되고 있음을 의미한다. 그러므로 의사결정나무모형에 의한 정분류율이 가장 높게 나타나고 있지만, 불량의 판별이 제대로 이루어지지 않고 있으므로 좋은 예측 모형이 아니다. 의사결정나무모형 이외에 로지스틱회귀모형, 신경망모형, 랜덤포레스트모형의 정분류율에 큰 차이가 없음을 확인할 수 있다. 다음으로 G-mean은 로지스틱회귀모형 69.6%, 신경망모형 69.4%, 램덤포레스트모형 65.9%, SVM모형 44.5%, 의사결

정나무모형 0.0%로 나타나고 있는데, G-mean은 결과 범주가 0인 집단과 1인 집단을 동등하게 고려하는 측도로써 실제 범주가 0인 집단에 대한 정확도와 범주 1인 집단에 대한 정확도의 기하평균이므로 값이 클수록 분류 및 예측 성능이 우수하다. 그러므로 로지스틱회귀모형과 신경망모형은 큰 차이가 나지 않는 가운데 값이 크므로 분류를 위한 예측 성능이 가장 우수하다. 마지막으로 F1값도 로지스틱회귀모형과 신경망모형이 다른 모형에 비해 상대적으로 높게 나타나고 있어 분류 성능이 좋음을 알 수 있다.

<그림 Ⅵ-3> 훈련용 자료에 대한 정분류율, G-mean, F1값



<표 Ⅵ-6> 훈련용 자료에 대한 분류 결과

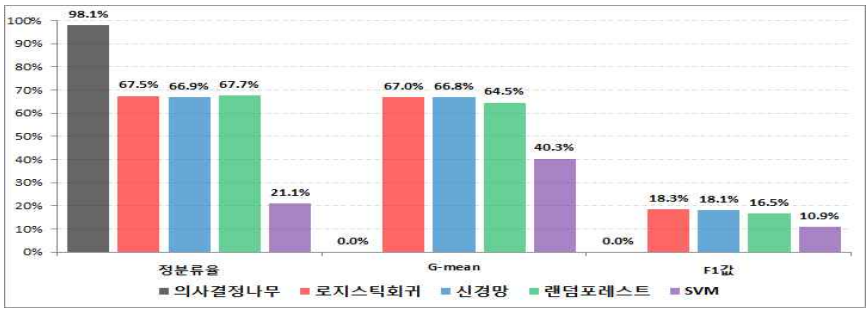
구분		훈련용자료				
		예측 0	예측 1	정분류율	G-mean	F1값
의사결정나무	실제 0	93,342	0	98.1%	0.0%	0.0%
	실제 1	1,832	0			
로지스틱회귀	실제 0	62,712	30,630	67.3%	69.6%	7.8%
	실제 1	511	1,321			
신경망	실제 0	62,058	31,284	66.6%	69.4%	7.7%
	실제 1	503	1,329			
랜덤포레스트	실제 0	63,215	30,127	67.7%	65.9%	7.1%
	실제 1	659	1,173			
SVM	실제 0	18,481	74,861	21.3%	44.5%	4.7%
	실제 1	0	1,832			

<그림 Ⅵ-4>와 <표 Ⅵ-7>은 평가용 자료에 대한 정분류율, G-mean, F1값을 정리한 것이다. 평가용 자료에 의한 결과는 훈련용 자료에 의한 결과와 유사함을 알 수 있다. 먼저 정분류율은 의사결정나무 98.1%, 로지스틱회귀모형 67.5%, 신경망모형 66.9%, 랜덤포레스트모형 67.7%, SVM모형 21.1%로 의사결정나무모형에 의한 정분류율이 가장 높게 나타나고 있다. 그러나 이 결과 또한 결과 표를 보면 {실제 1, 예측 1}의 빈도가 0으로 나타나고 있는데, 이는 실제로 보증사고가 발생한 차주(실제 1)는 의사결정나무모형에 의해 보증사고가 없는 우량으로 잘못 분류되고 있음을 의미한다. 그러므로 의사결정나무모형에 의한 정분류율이 가장 높게 나타나고 있지만 불량의 판별이 제대로 이루어지지 않고 있으므로 좋은 예측 모형이 아니라는 결론을 내릴 수 있다. 훈련용 자료에 의한 평가 비교 결과와 마찬가지로 평가용 자료를 이용한 결과 역시, 의사결정나무모형 이외에 로지스틱회귀모형, 신경망모형, 랜덤포레스트모형의 정분류율에 큰 차이가 없음을 확인할 수 있다. 다음으로 G-mean은 로지스틱회귀모형 67.0%, 신경망모형 66.8%, 램덤포레스트모형 64.5%, SVM모형 40.3%, 의사결정나무모형 0.0%로 나타나고 있는데, 로지스틱회귀모형과 신경망모형은 큰 차이가 나지 않는 가운데 분류 성능이 가장 우수하다. 마지막으로 F1값도 로지스틱회귀모형과 신경망모형이 다른 모형에 비해 상대적으로 높게 나타나고 있어 분류 성능이 좋음을 알 수 있다.

다음은 훈련용과 평가용 자료의 결과를 이용하여 구축된 모형의 안정성에 대해 살펴보고자 한다. 안정성이란 구축된 모형에 새로운 자료를 적용하였을 때, 분류 성능 측면에서 차이가 크지 않아야 한다는 조건을 만족하여야 한다. 앞의 결과들을 비교하였을 때, 구축된 기계학습 모형들의 훈련용 자료와 평가용 자료의 정분류율, G-mean, F1값을 살펴보면, 모든 모형에서 큰 차이가 나지 않음을 알 수 있다. 이

와 같은 결과를 통해 구축된 모형들은 새로운 자료를 적용하였을 때 분류 결과에 큰 차이가 없으며 안정성이 높은 굳건한 모형(robust model)임을 알 수 있다.

<그림 Ⅵ-4> 평가용 자료에 대한 정분류율, G-mean, F1값



<표 Ⅵ-7> 평가용 자료에 대한 분류 결과

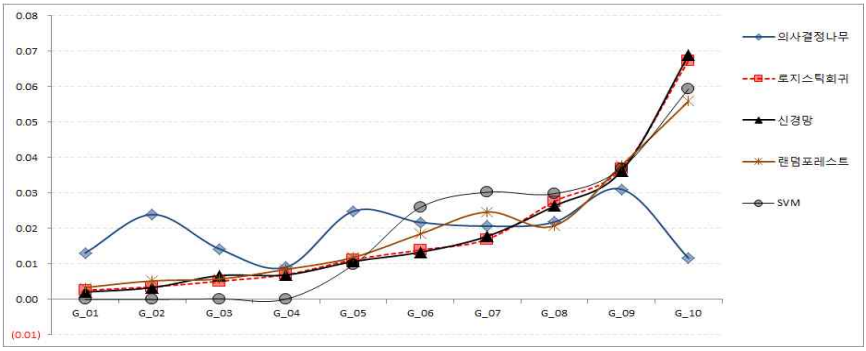
구분		평가용자료				
		예측 0	예측 1	정분류율	G-mean	F1값
의사결정나무	실제 0	40,222	0	98.1%	0.0%	0.0%
	실제 1	791	0			
로지스틱회귀	실제 0	27,138	13,084	67.5%	67.0%	18.3%
	실제 1	265	526			
신경망	실제 0	26,896	13,326	66.9%	66.8%	18.1%
	실제 1	263	528			
랜덤포레스트	실제 0	27,290	12,932	67.7%	64.5%	16.5%
	실제 1	306	485			
SVM	실제 0	8,008	32,214	21.1%	40.3%	10.9%
	실제 1	145	646			

다음은 구축된 모형의 반응률에 대한 결과이다. 전술하였듯이 반응률이란 구축된 평가모형을 통해 산출된 사후확률을 정렬하여 N개의 구간으로 등분한 후, 각 구간에 포함된 종속변수의 특정 범주의 빈도를 이용해 산출한 결과인데, 사후확률이 가장 큰 구간에서 가장 낮은 구간으로 갈수록 반응률이 높게 나타나다가 급격하게 감소하는

형태 또는 그 반대인 경우 좋은 분류 및 예측 성능을 가진 모형이다.

<그림 Ⅵ-5>와 <표 Ⅵ-8>은 훈련용 자료에 대한 반응률을 정리한 것이다. 이 결과를 보면 로지스틱회귀모형, 신경망모형, 랜덤포레스트 모형의 반응률이 점차 증가하고 있으며, G_09와 G_10에서 급격히 상승하고 있음을 알 수 있다. 그러나 의사결정나무모형과 SVM모형은 역전 현상이 발생하고 있다. 이 결과를 통해 로지스틱회귀모형, 신경망모형, 랜덤포레스트모형의 우불량 변별력이 좋음을 알 수 있다.

<그림 Ⅵ-5> 훈련용 자료에 대한 반응률 비교

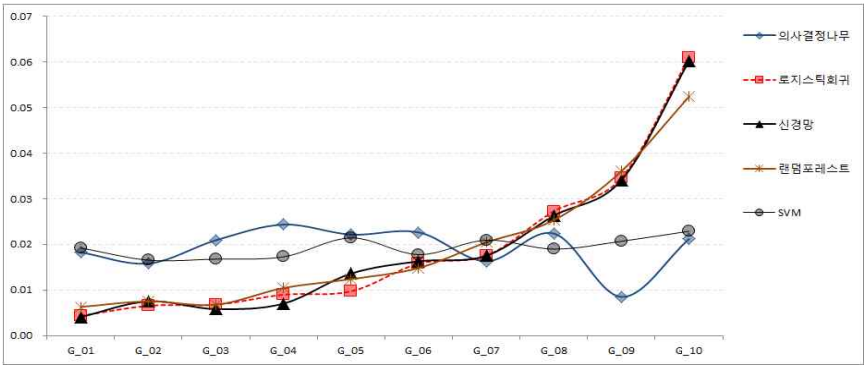


<표 Ⅵ-8> 훈련용 자료에 대한 반응률 결과

구분	의사결정나무	로지스틱회귀	신경망	랜덤포레스트	SVM
G_01	0.0131	0.0026	0.0021	0.0034	0.0000
G_02	0.0240	0.0036	0.0034	0.0053	0.0000
G_03	0.0143	0.0051	0.0067	0.0058	0.0002
G_04	0.0091	0.0070	0.0069	0.0085	0.0001
G_05	0.0249	0.0112	0.0107	0.0119	0.0098
G_06	0.0218	0.0140	0.0133	0.0185	0.0260
G_07	0.0207	0.0170	0.0179	0.0247	0.0303
G_08	0.0219	0.0276	0.0264	0.0208	0.0298
G_09	0.0311	0.0369	0.0362	0.0378	0.0370
G_10	0.0117	0.0674	0.0688	0.0559	0.0594

<그림 Ⅵ-6>과 <표 Ⅵ-9>는 평가용 자료에 대한 반응률을 정리한 것이다. 이 결과를 또한 훈련용 자료의 반응률 결과와 거의 유사하며, 로지스틱회귀모형, 신경망모형, 랜덤포레스트모형의 우불량 변별력이 좋다는 결론을 내릴 수 있다. 또한 모형의 안정성 측면에서도 훈련용 자료와 평가용 자료에 의한 반응률 결과가 거의 유사하기 때문에 새로운 자료를 적용해도 안정성이 높다(<그림 Ⅵ-7> 참조).

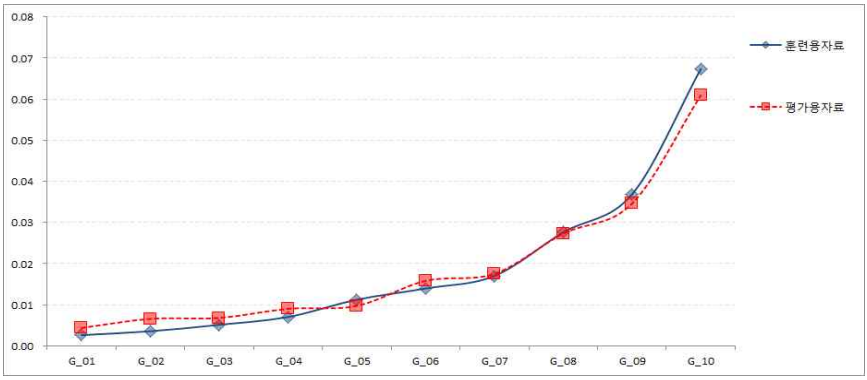
<그림 Ⅵ-6> 평가용 자료에 대한 반응률 비교



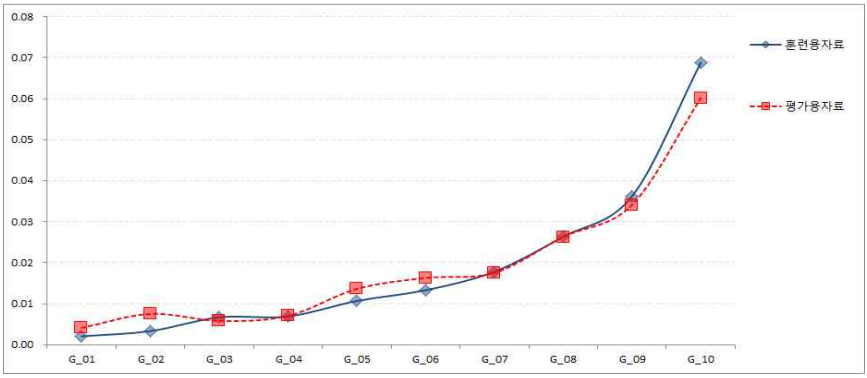
<표 Ⅵ-9> 평가용 자료에 대한 반응률 결과

구분	의사결정나무	로지스틱회귀	신경망	랜덤포레스트	SVM
G_01	0.0183	0.0044	0.0041	0.0063	0.0193
G_02	0.0158	0.0066	0.0076	0.0076	0.0166
G_03	0.0210	0.0068	0.0059	0.0068	0.0168
G_04	0.0244	0.0090	0.0071	0.0105	0.0173
G_05	0.0222	0.0098	0.0137	0.0124	0.0215
G_06	0.0227	0.0158	0.0163	0.0149	0.0178
G_07	0.0163	0.0176	0.0176	0.0205	0.0210
G_08	0.0224	0.0273	0.0263	0.0254	0.0190
G_09	0.0085	0.0346	0.0341	0.0361	0.0207
G_10	0.0212	0.0610	0.0602	0.0524	0.0229

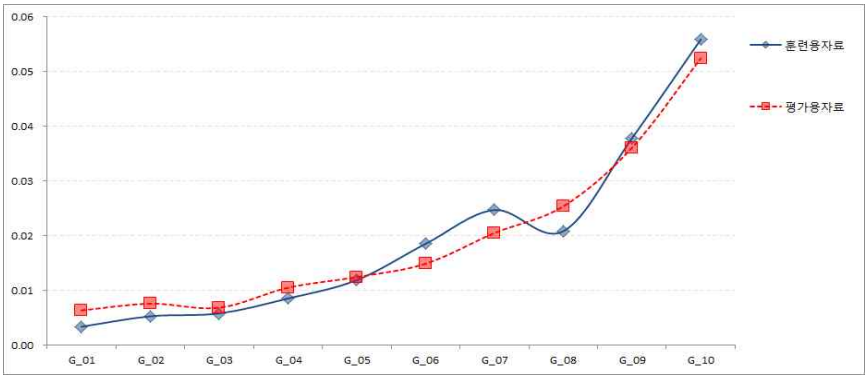
<그림 Ⅵ-7> 훈련용 및 평가용 자료의 반응률 비교



(1)로지스틱회귀모형



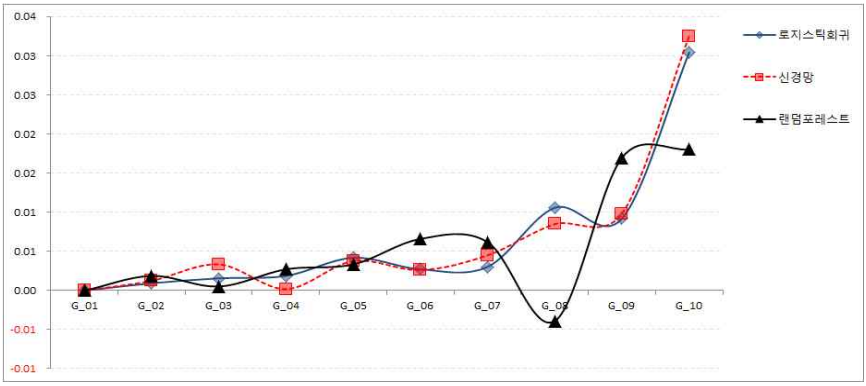
(2)신경망모형



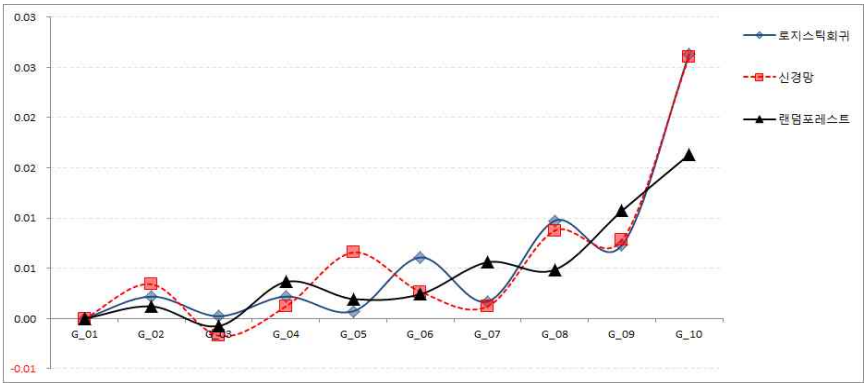
(3)랜덤포레스트모형

<그림 Ⅵ-8>, <그림 Ⅵ-9>와 <표 Ⅵ-10>은 훈련용 자료와 평가용 자료에 대한 반응률의 각 구간별 차이를 나타낸 것이다. 이 결과를 살펴보면, 로지스틱회귀모형의 반응률은 역전 현상이 발생하지 않는 반면, 신경망모형과 랜덤포레스트모형은 반응률 역전 현상이 발생하고 있음을 알 수 있다. 그러므로 로지스틱회귀모형의 분류 예측 성능 즉 등급 간 변별력이 가장 우수하다는 결론을 내릴 수 있다.

<그림 Ⅵ-8> 훈련용 자료에 대한 반응률 역전 현상 확인



<그림 Ⅵ-9> 평가용 자료에 대한 반응률 역전 현상 확인



<표 Ⅵ-10> 훈련 및 평가용 자료에 대한 구간별 반응률 비교

구분	의사결정나무	로지스틱회귀	신경망	의사결정나무	로지스틱회귀	신경망
G_01	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
G_02	0.0009	0.0013	0.0019	0.0022	0.0034	0.0012
G_03	0.0016	0.0034	0.0005	0.0002	**-0.0017**	**-0.0007**
G_04	0.0019	0.0002	0.0027	0.0022	0.0012	0.0037
G_05	0.0042	0.0038	0.0034	0.0007	0.0066	0.0020
G_06	0.0027	0.0026	0.0066	0.0061	0.0027	0.0024
G_07	0.0030	0.0045	0.0062	0.0017	0.0012	0.0056
G_08	0.0106	0.0085	**-0.0039**	0.0097	0.0088	0.0049
G_09	0.0092	0.0099	0.0170	0.0073	0.0078	0.0107
G_10	0.0305	0.0326	0.0181	0.0263	0.0261	0.0163

본 절에서 사용한 모든 평가 측도인 정분류율, G-mean, F1값, 반응률을 비교한 결과, 다음과 같은 결론을 내릴 수 있다. 계급화 과정을 통해 원자료를 표준화하고 이 자료를 의사결정나무모형, 로지스틱회귀모형, 신경망모형, 랜덤포레스트모형, SVM모형에 적용한 결과, 로지스틱회귀모형이 분류를 위한 예측 성능과 안정성 측면에서 가장 우수한 신용평가모형이다.

의사결정나무모형의 경우 정분류율은 가장 높았지만 사고가 발생한 차주 즉 불량인 차주의 분류 성능이 매우 나쁘며, 신경망과 랜덤포레스트모형은 로지스틱회귀모형과 분류 성능 면에서 큰 차이는 나지 않았지만, 등급화의 과정에서 일부 등급 구간에 불량률 역전 현상이 발생하였으므로 본 자료에 대해서는 로지스틱회귀모형을 사용하는 것이 가장 타당할 것으로 판단된다. 이에 본 장의 마지막 절에서는 분류 예측 성능이 가장 우수한 로지스틱회귀모형을 이용하여 소상공인의 신용평가모형을 구축해 보고자 한다.

3. 로지스틱회귀모형을 이용한 최종 신용평가모형

로지스틱회귀모형을 이용하여 우량(Y=0)에 대한 확률을 산출한 <표 Ⅵ-11>의 각 더미변수의 의미는 앞의 "<표 Ⅵ-5> 성김화에 의한 재범주화 결과"에 설명되어 있다. 즉 더미0은 <표 Ⅵ-11>에는 표시되어 있지 않지만 참조변수(reference variable)을 의미한다. <표 Ⅵ-11>에서 더미1은 각 변수별로 불량률이 가장 높은 구간이나 범주에 해당되는 회귀계수가 0인 더미0과 비교되는 회귀계수를 의미하는 것이다. 그리고 더미2 역시 더미0과 비교되는 회귀계수를 의미하는데, <표 Ⅵ-5>에서 0으로 표현된 구간에서 2로 갈수록 불량률은 감소하고 우량률은 점차 증가한다.

그러므로 더미0, 더미1, 더미2에 대한 회귀계수의 크기가 "더미0<더미1<더미2"의 순서가 아니거나 음(-)의 값이 존재한다면, 이는 다중공선성이 존재함을 의미하므로 잘못 구축된 모형임을 의미한다. 이런 경우가 발생하면 처음부터 다시 모형을 구축하여야 한다. 그러나 <표 Ⅵ-11>의 모든 독립변수에 대한 회귀계수를 보면 "더미0<더미1<더미2"의 순서가 지켜지고 있으며 음(-)의 값이 존재하지 않는다. 그리고 모든 회귀계수의 p-값과 모형에 대한 카이제곱 통계량의 p-값이 0.05 이하로 나타나고 있으므로, 모형 구축이 올바르며 신용평가모형 구축에 사용된 독립변수들은 모두 유의한 것으로 판단할 수 있다.

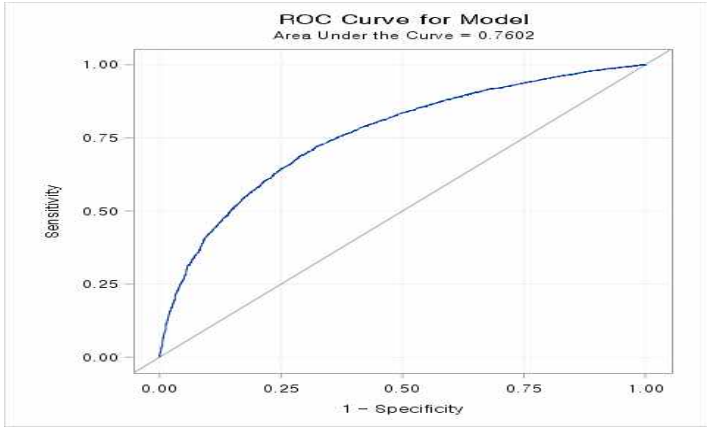
<표 Ⅵ-11>에서 로지스틱회귀모형을 이용하여 구축된 신용평가모형에 대한 예측 성능은 정분류율, 오분류율, c통계량으로 확인할 수 있으며, c통계량에 대한 <그림 Ⅵ-10>의 ROC 곡선을 이용하여 살펴볼 수 있다. 모형에 의한 정분류율은 76.1% 상당히 높고, c통계량 또한 0.76으로 매우 높기 때문에 구축된 평가모형의 예측 성능이 높다고 판단할 수 있다. c통계량은 ROC 곡선 아래의 면적으로 값이 1에

근접할수록 좋은 예측 성능을 가진 모형이다. 그리고 ROC 곡선을 보면 곡선이 위쪽으로 볼록하게 형성되어 있으므로 예측 성능이 매우 우수하다는 결론을 내릴 수 있다. 호스머-램쇼 검정통계량 값은 13.7이고, p-값은 0.09으로 로지스틱 회귀모형이 비교적 잘 적합되고 있다.

<표 Ⅵ-11> 최종 로지스틱회귀모형 구축 결과

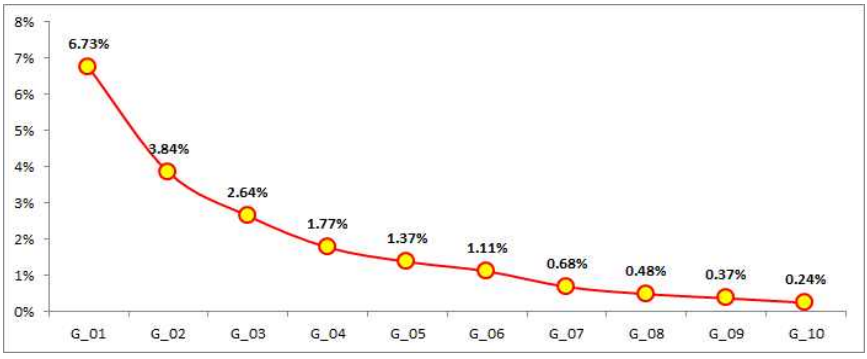
변수	더미	회귀계수	표준오차	왈드 카이제곱값	p-값	오즈비
절편		-2.389	0.400	35.77	<.0001	
고객형태	1	0.691	0.088	60.99	<.0001	2.00
업종	1	0.231	0.049	22.50	<.0001	1.26
	2	0.558	0.193	8.41	0.0037	1.75
주사업장임차보증금액	1	0.124	0.050	6.11	0.0135	1.13
실거주지소유여부	1	0.325	0.075	18.59	<.0001	1.38
실소유지월세금액	1	0.309	0.055	31.05	<.0001	1.36
차입금운전	1	0.479	0.176	7.38	0.0066	1.61
기보증잔액재단	1	0.230	0.056	16.66	<.0001	1.26
기보증잔액기보	1	0.845	0.249	11.53	0.0007	2.33
담보제외차입기관수	1	0.659	0.066	100.08	<.0001	1.93
	2	1.185	0.070	290.81	<.0001	3.27
현금서비스금액	1	0.661	0.063	109.42	<.0001	1.94
보유부동산	1	0.453	0.083	29.62	<.0001	1.57
	2	0.692	0.079	76.07	<.0001	2.00
업력	1	0.708	0.052	184.26	<.0001	2.03
	2	1.184	0.091	167.78	<.0001	3.27
거주기간	1	0.223	0.050	19.77	<.0001	1.25
월평균매출액	1	0.379	0.053	50.74	<.0001	1.46
	2	0.659	0.086	58.44	<.0001	1.93
재고자산	1	0.701	0.148	22.52	<.0001	2.02
직권말소	1	1.027	0.261	15.55	<.0001	2.79
정분류율		**76.1%**				
오분류율		**23.9%**				
카이제곱 p값, c통계량		**<.0001, 0.76**				
호스머-램쇼 적합도 검정		**13.70(0.09)**				

<그림 Ⅵ-10> ROC 곡선



<그림 Ⅵ-11>과 <표 Ⅵ-12>는 로지스틱회귀모형에 의한 사후확률을 이용한 반응률과 반응률 도표이다. 표와 그림에서 G_01은 우량일 사후확률이 낮은 구간이며 G_10은 우량일 사후확률이 높은 구간이다. 결과를 정리하면 반응률이 점차 감소하는 경향을 보이며 불량률 역전 현상이 관측되지 않고 있다. 그러므로 우불량을 변별하기 위한 분류 예측 성능이 우수하다는 결론을 내릴 수 있다.

<그림 Ⅵ-11> 사후확률 구간별 불량률



<표 Ⅵ-12> 사후확률 구간별 불량률 및 KS 통계량

구간	등급별 사후확률 구간		우량수	불량수	구간내불량률	KS 통계량
G_01	0.5579	0.9582	8,867	640	6.73%	0.394
G_02	0.9582	0.9714	9,155	366	3.84%	
G_03	0.9714	0.9789	9,255	251	2.64%	
G_04	0.9789	0.9839	9,423	170	1.77%	
G_05	0.9839	0.9876	9,331	130	1.37%	
G_06	0.9876	0.9904	9,423	106	1.11%	
G_07	0.9904	0.9928	9,626	66	0.68%	
G_08	0.9928	0.9947	9,286	45	0.48%	
G_09	0.9947	0.9965	9,494	35	0.37%	
G_10	0.9965	0.9992	9,483	23	0.24%	

앞에 절에서 구축한 신용평가모형이 통계적으로 유의미하기 때문에 이를 이용하여 최종적으로 평점표를 작성한다. 개인신용평가모형 구축 시 일반적으로 이용하는 평점표는 점수가 증가할 때마다 우량/불량 비율이 특정 배수(예를 들면 2배)로 커지도록 하고 있으나, 본 연구에서는 로지스틱회귀모형을 통해 산출한 업체마다의 사후확률을 평점으로 사용한다. 평점 산출 방법은 로지스틱회귀모형에서의 각 더미변수의 회귀계수를 배점으로 활용하는데 각 변수의 계급별 배점은 <표 Ⅵ-11>의 회귀계수와 같다. 기본 점수는 절편인 -2.389이고, 각 회귀계수가 해당 범주의 점수가 된다. 최종적인 평점은 각 차주별로 평가기준에 부합하는 배점 및 기본 배점을 모두 더해서 배점합계를 산출하고, 이 값을 (식 27)와 같은 로지스틱 회귀식에 대입하여 계산한다.

$$\text{평점} = \frac{\exp(\text{배점합계})}{1+\exp(\text{배점합계})} \times 1000 \qquad \text{(식 27)}$$

$$, \ 557 \leq \text{평점} \leq 1000$$

모형을 구축하는 데 사용한 훈련용 자료를 (식 27)에 적용하여 각 차주의 신용평점을 산출한 후, 평점 구간, 즉 신용등급을 설정하기 위해 산출된 평점의 계급세분화(fine classing)와 성김화(coarse classing) 과정을 수행한다. 성김화의 결과 6개 등급으로 구분이 가능한데, <표 Ⅵ-13>과 같이 산출된 평점을 이용하여 계급세분화를 수행한 결과 평점이 낮아짐에 따라, 즉 신용등급이 나빠짐에 따라 불량률은 계속 증가하며 역전 현상은 발생하지 않고 있다. 최종 신용등급은 1등급으로 994.7 초과, 2등급 990.4 초과, 3등급 983.9 초과, 4등급, 971.4 초과, 5등급 958.2 초과, 6등급 958.2 이하로 구분할 수 있다.

<표 Ⅵ-13> 최종 등급화

등급	구간	등급별 점수 구간		우량차주수	불량차주수	불량률
1등급	G_09, G_10	994.7	999.2	18,977	58	0.30%
2등급	G_07, G_08	990.4	994.7	18,912	111	0.58%
3등급	G_05, G_06	983.9	990.4	18,754	236	1.24%
4등급	G_03, G_04	971.4	983.9	18,678	421	2.20%
5등급	G_02	958.2	971.4	9,155	366	3.84%
6등급	G_01	557.9	958.2	8,867	640	6.73%

<그림 Ⅵ-12> 등급 구간별 불량률

Ⅶ. 결론 및 향후 과제

본 연구는 소상공인 신용평가를 위해 신용보증재단이 보유한 내부 정보를 이용하여 현재 현업에서 가장 많이 사용하고 있는 빅데이터 분석을 위한 기계학습 기법 중 의사결정나무모형, 로지스틱회귀모형, 신경망모형, 랜덤포레스트모형, SVM모형으로 신용평가모형을 구축하였을 때 분류를 위한 예측 성능이 우수한 모형이 무엇인지를 확인하고 최종적으로 신용평가모형을 구축한 후 시사점을 찾아보는 것이다.

본 논문의 모형 구축 분석 대상은 16개 지역신용보증재단에서 2017년 7월~2019년 6월 신용보증을 받은 차주 136,189개이다. 구축된 모형의 평가는 예비 방법을 적용하였으며, 평가 측도로는 오분류율, G-mean, F1 측도, 반응률을 이용하였고, 모형 구축 도구는 SAS 9.4와 R을 이용하였다.

다양한 기계학습 기법을 이용하여 소상공인 신용평가모형을 구축한 결과, 첫째, G-mean, F1 측도를 살펴보았을 때 로지스틱회귀모형이 가장 좋은 예측 성능을 가지고 있으며, 계급불균형 자료에 대해 오분류율을 이용하여 모형을 평가하는 것은 적절하지 않다는 사실을 확인하였다. 둘째, 반응률을 보았을 때 의사결정나무모형, 신경망모형, 랜덤포레스트모형, SVM모형 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 것으로 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 데이터 정제 기법을 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 방법론 적용 등도 이루어지지 않았다는 한계점도 가지고 있다.

현재 빅데이터(big data)가 4차 산업혁명의 주요 요소 중 하나로 부각되고 있고, 실제로 많은 국가와 기관에서 양질의 자료를 수집하고 이를 활용하기 위해 많은 노력을 하고 있다. 구글, 아마존, 페이스

북 등 세계적인 기업이 빅데이터의 중요성을 인식하고 다양한 종류의 데이터 수집을 위해 많은 투자를 하고 있다(박주완, 2018). 또한 금융산업에서도 빅데이터를 활용하기 위해 다양한 시도가 이루어지고 있으며, 이를 신용평가에 활용하기 위해 많은 연구와 노력을 하고 있고, 실제 현업에서 빅데이터를 적용하는 사례들이 점차 증가하고 있다(신윤재, 2016). 이처럼 현 시대에서는 양질의 자료 수집과 활용은 경쟁력 제고를 위한 필수적인 사항이 되었다.

데이터가 경쟁력이 되고 있는 현 시점에서 소상공인의 경우 일반적으로 양질의 객관적인 데이터 수집이 매우 힘들다고 알려져 있으며, 실제로 객관적인 자료의 수집을 통한 신용평가모형의 구축이나 다양한 분석이 매우 어려운 실정이다. 만약 자료가 부족하다는 이유만으로 외부 자료의 의존 비중을 점차 높일 경우, 자료 구입을 위한 추가적인 비용의 발생뿐 아니라, 평가모형 구축도 자료를 제공하는 기관에 종속될 수밖에 없으며 자체적인 자료 수집의 노력이 감소할 가능성 높아질 것이다. 물론 정확한 신용평가를 위해 외부자료를 이용하는 것이 문제가 되지 않을 수도 있다는 반론이 제기될 수 있고, 단순히 신뢰성과 정확성의 문제라면 외부자료를 이용하여 모형을 구축하는 것이 더욱 타당할 것이고 또 장려할만한 사항일 것이다.

그러나 모형의 신뢰성과 정확성 향상이라는 목적만으로 외부 자료 이용이라는 근시안적이고 단기적인 처방만 하는 것은 양질의 자료 확보가 경쟁력인 시대에서는 치명적인 문제가 될 수도 있다. 그러므로 단기적인 처방 이외에 장기적으로 자료 수집과 활용을 위한 투자와 노력이 필요하다. 다양한 경로를 통한 객관적인 자료의 수집과 더불어 이를 부가적 또는 심층적으로 설명할 수 있는 설문조사 자료를 수집하고 두 가지 자료의 장점 등을 활용한다면 자료의 활용도는 높아질 것이며, 자료 활용을 통한 소상공인에 대한 정책 수립 등에도 기

여할 것으로 판단된다.

본 연구를 통한 향후 연구 방향은 다음과 같다. 첫째, 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 연구가 필요하다. 둘째, 불량 차주의 개수가 충분하다면 기업의 규모 및 업종을 구분한 모형 구축을 고려해 볼 필요가 있다. 기업의 규모나 업종에 따라 기업의 특성에 차이가 있을 수 있으므로, 자료의 양과 질이 충분하다면 규모와 업종을 고려한 연구가 필요하다. 셋째, 현재 빅데이터가 중요한 트랜드(trend)로 나타나고 있으며 이를 활용하는 결과가 실제 현업에서 적용되어 있다. 그러므로 소상공인 신용평가모형 구축에서도 다양한 출처의 빅데이터 적용 가능성과 이를 적용하기 위한 제도적인 방법 등에 대해 연구해 볼만한 가치가 있다. 넷째, 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. 다섯째, 불량 차주의 자료가 불충분한 계급불균형 자료인 경우 본 논문의 결과에서 살펴본 바와 같이 모형 구축 및 평가가 쉽지 않다. 그러므로 계급불균형인 자료에 대한 모형 구축 방법론에 대해서도 추가적인 실증연구가 필요하다. 마지막으로 신경망모형 등을 이용할 경우 평점화하기 위한 기법 등에서도 논의할 필요가 있다. 로지스틱회귀모형의 경우 평점 산출 로직을 전산적인 신용평가 시스템에 탑재하기 쉽지만, 신경망모형 등의 경우 산출된 결과를 평점화하여 시스템에 탑재하기 위해서는 로지스틱회귀모형 대비 수십에서 수백 배 이상 복잡한 로직이 필요하므로 이에 연구는 필수적이다.

부 록

1. 기계학습 R 분석 프로그램

```
# 0. 분석 경로 설정
setwd("D:/@02.보고서/@2019년/01-02.(분석보고서)빅데이터분석기법이용소평모형구축
      /분석프로그램")

# 1. 패키지 설치
install.packages("rpart")
install.packages("rpart.plot")
install.packages("nnet")
install.packages("ROCR")
install.packages("devtools")
install.packages("randomForest")
install.packages("caret")
install.packages("party")
install.packages("Hmisc")
install.packages("readxl")
install.packages("corrgram")
install.packages("e1071")
install.packages("kernlab")

# 2. 라이브러리 설정
library(rpart)
library(rpart.plot)
library(nnet)
library(ROCR)
library(devtools)
library(clusterGeneration)
```

```
library(scales)
library(reshape)
library(party)
library(caret)
library(randomForest)
library(readxl)
library(Hmisc)
library(corrgram)
library(e1071)
library(kernlab)

# 3. 데이터 불러오기

## 3.1 훈련용 텍스트 자료 불러오기
tr_data <- read.table("tr_data.txt", header=T, sep="₩t")

## 3.2 평가용 텍스트 자료 불러오기
ts_data <- read.table("ts_data.txt", header=T, sep="₩t")

# 4. 모형 구축

## 4.1 CART
tr_tree<-rpart(Y~., data=tr_data, control=rpart.control(maxdepth=5),
               method='class', cp=0.00001, minsplit=5, xval=5)
printcp(tr_tree)
prp(tr_tree, type=0, extra=0, digits=2, split.font=1, varlen=-10)
plot(tr_tree, compress=TRUE, margin=0.1)
text(tr_tree, cex=0.5)
tr_tree_rs<-predict(tr_tree, type="prob")
ts_tree_rs<-predict(tr_tree, newdata=ts_data)
write.table(tr_data, "tree_tr1_1.txt", col.names=T, row.names=T, sep="₩t")
write.table(tr_tree_rs,  "tree_tr1_2.txt", col.names=T, row.names=T, sep="₩t")
```

```
write.table(ts_data, "tree_ts1_1.txt", col.names=T, row.names=T, sep="₩t")
write.table(ts_tree_rs, "tree_ts1_2.txt", col.names=T, row.names=T, sep="₩t")

## 4.2 LOGISTIC REGRESSION
tr_logis <- glm(Y~., data=tr_data, family="binomial")
tr_logis
options(scipen=999)
summary(tr_logis)
tr_logis_rs <- fitted(tr_logis)
ts_logis_rs <- predict(tr_logis, newdata=ts_data, type="response")

write.table(tr_data, "logis_tr1_1.txt", col.names=T, row.names=T, sep="₩t")
write.table(tr_logis_rs, "logis_tr1_2.txt", col.names=T, row.names=T, sep="₩t")
write.table(ts_data, "logis_ts1_1.txt", col.names=T, row.names=T, sep="₩t")
write.table(ts_logis_rs, "logis_ts1_2.txt", col.names=T, row.names=T, sep="₩t")

## 4.3 NEURAL NETWORK 1 -nnet
tr_net <- nnet(Y~., data=tr_data, size=3, linout=FALSE,
 decay=0.0005, rang=0.5, entropy=FALSE, maxit=1000)
tr_net
summary(tr_net)
garson(tr_net) # 신경망모형에서 변수 중요도 확인

tr_net_rs <- predict(tr_net)
ts_net_rs <- predict(tr_net, newdata=ts_data)

write.table(tr_data, "net_tr1_1.txt", col.names=T, row.names=T, sep="₩t")
write.table(tr_net_rs, "net_tr1_2.txt", col.names=T, row.names=T, sep="₩t")
write.table(ts_data, "net_ts1_1.txt", col.names=T, row.names=T, sep="₩t")
write.table(ts_net_rs, "net_ts1_2.txt", col.names=T, row.names=T, sep="₩t")

## 4.4 RANDOM FOREST
```

```
tr_for <- randomForest(Y~., data=tr_data, importance=F)
varImpPlot(tr_for)
tr_for
tr_for_rs <- predict(tr_for)
ts_for_rs <- predict(tr_for, newdata=ts_data)

write.table(tr_data, "for_tr1_1.txt", col.names=T, row.names=T, sep="₩t")
write.table(tr_for_rs, "for_tr1_2.txt", col.names=T, row.names=T, sep="₩t")
write.table(ts_data, "for_ts1_1.txt", col.names=T, row.names=T, sep="₩t")
write.table(ts_for_rs, "for_ts1_2.txt", col.names=T, row.names=T, sep="₩t")

## 4.5 SVM
#tune.svm(Y~., data=tr_data, gamma=10^(-1:1), cost=10^(1:2))
#tr_svm <- svm(Y~., data=tr_data, type="C-classification", kernal="radial", cost=10,
gamma=0.1)
tr_svm <- ksvm(Y~., data=tr_data, type="C-bsvc", kernal="rbfdot",
 kpar=list(sigma=0.1), C=10, prob.model=TRUE)
tr_svm_rs <- predict(tr_svm, tr_data, type="probabilities")
ts_svm_rs <- predict(tr_svm, ts_data, type="probabilities")

write.table(tr_data, "svm_tr1_1.txt", col.names=T, row.names=T, sep="₩t")
write.table(tr_svm_rs, "svm_tr1_2.txt", col.names=T, row.names=T, sep="₩t")
write.table(ts_data, "svm_ts1_1.txt", col.names=T, row.names=T, sep="₩t")
write.table(ts_svm_rs, "svm_ts1_2.txt", col.names=T, row.names=T, sep="₩t")
```

2. SAS 프로그램 - 데이터셋 구축 및 모형 평가

```
/*********************************************************************************/
/* 0. 환경 설정 및 원천 자료 불러오기  */
/*********************************************************************************/
```

```
* 0.1 옵션 설정;
OPTIONS NODATE NOCENTER COMPRESS='YES' FORMDLIM=' '
VALIDVARNAME=ANY DEBUG = 'NLSNAME=YES'; RUN;

* 0.2 라이브러리 설정;
LIBNAME BIG 'D:₩@02.보고서₩@2019년₩01-02.(분석보고서)빅데이터분석기법이용소
평모형구축₩분석프로그램'; RUN;

* 0.3 데이터 불러오기;
data BIGDATA.RAW_add_14 ;
%let _EFIERR_ = 0; /* set the ERROR detection macro variable */
infile 'D:₩2019년 업무₩01-06.(조사연구)내부연구₩1.빅데이터분석기법₩03.데이터작업
₩data₩빅데이터분석기법연구_14재단_추가자료.txt' delimiter='09'x MISSOVER DSD
lrecl=32767 firstobs=2 ;
informat 재단명 $600. ;
informat 고객번호 $600. ;
informat 고객형태 $600. ;
informat 심사번호 $600. ;
informat 보증일자 $600. ;
informat 보증번호 $600. ;
informat 조사번호 $600. ;
informat 자산총계_당기_백만원 $600. ;
informat 자산총계_전기_백만원 $600. ;
informat 자산총계_전전기_백만원 $600. ;
informat 유동자산_당기_백만원 $600. ;
informat 유동자산_전기_백만원 $600. ;
informat 유동자산_전전기_백만원 $600. ;
informat 부채총계_당기_백만원 $600. ;
informat 부채총계_전기_백만원 $600. ;
informat 부채총계_전전기_백만원 $600. ;
informat 유동부채_당기_백만원 $600. ;
informat 유동부채_전기_백만원 $600. ;
```

```
informat 유동부채_전전기_백만원 $600. ;
informat 자본총계_당기_백만원 $600. ;
informat 자본총계_전기_백만원 $600. ;
informat 자본총계_전전기_백만원 $600. ;
informat 매출액_당기_백만원 $600. ;
informat 매출액_전기_백만원 $600. ;
informat 매출액_전전기_백만원 $600. ;
informat 영업손익_당기_백만원 $600. ;
informat 영업손익_전기_백만원 $600. ;
informat 영업손익_전전기_백만원 $600. ;
informat 당기순이익_당기_백만원 $600. ;
informat 당기순이익_전기_백만원 $600. ;
informat 당기순이익_전전기_백만원 $600. ;
informat 업종구분 $600. ;
informat 직권말소여부 $600. ;
informat 성장성지수 $600. ;
informat 안정성지수 $600. ;
informat 밀집도지수 $600. ;
informat 구매력지수 $600. ;
informat 집객력지수 $600. ;
informat 최종사업성평가지수 $600. ;

format 재단명 $600. ;
format 고객번호 $600. ;
format 고객형태 $600. ;
format 심사번호 $600. ;
format 보증일자 $600. ;
format 보증번호 $600. ;
format 조사번호 $600. ;
format 자산총계_당기_백만원 $600. ;
format 자산총계_전기_백만원 $600. ;
format 자산총계_전전기_백만원 $600. ;
```

```
format 유동자산_당기_백만원 $600. ;
format 유동자산_전기_백만원 $600. ;
format 유동자산_전전기_백만원 $600. ;
format 부채총계_당기_백만원 $600. ;
format 부채총계_전기_백만원 $600. ;
format 부채총계_전전기_백만원 $600. ;
format 유동부채_당기_백만원 $600. ;
format 유동부채_전기_백만원 $600. ;
format 유동부채_전전기_백만원 $600. ;
format 자본총계_당기_백만원 $600. ;
format 자본총계_전기_백만원 $600. ;
format 자본총계_전전기_백만원 $600. ;
format 매출액_당기_백만원 $600. ;
format 매출액_전기_백만원 $600. ;
format 매출액_전전기_백만원 $600. ;
format 영업손익_당기_백만원 $600. ;
format 영업손익_전기_백만원 $600. ;
format 영업손익_전전기_백만원 $600. ;
format 당기순이익_당기_백만원 $600. ;
format 당기순이익_전기_백만원 $600. ;
format 당기순이익_전전기_백만원 $600. ;
format 업종구분 $600. ;
format 직권말소여부 $600. ;
format 성장성지수 $600. ;
format 안정성지수 $600. ;
format 밀집도지수 $600. ;
format 구매력지수 $600. ;
format 집객력지수 $600. ;
format 최종사업성평가지수 $600. ;

input
재단명 $
```

고객번호 $
고객형태 $
심사번호 $
보증일자 $
보증번호 $
조사번호 $
자산총계_당기_백만원 $
자산총계_전기_백만원 $
자산총계_전전기_백만원 $
유동자산_당기_백만원 $
유동자산_전기_백만원 $
유동자산_전전기_백만원 $
부채총계_당기_백만원 $
부채총계_전기_백만원 $
부채총계_전전기_백만원 $
유동부채_당기_백만원 $
유동부채_전기_백만원 $
유동부채_전전기_백만원 $
자본총계_당기_백만원 $
자본총계_전기_백만원 $
자본총계_전전기_백만원 $
매출액_당기_백만원 $
매출액_전기_백만원 $
매출액_전전기_백만원 $
영업손익_당기_백만원 $
영업손익_전기_백만원 $
영업손익_전전기_백만원 $
당기순이익_당기_백만원 $
당기순이익_전기_백만원 $
당기순이익_전전기_백만원 $
업종구분 $
직권말소여부 $

```
성장성지수 $
안정성지수 $
밀집도지수 $
구매력지수 $
집객력지수 $
최종사업성평가지수 $
;
if _ERROR_ then call symputx('_EFIERR_',1);  /* set ERROR detection macro variable
*/
run;

data BIGDATA.Raw_add_su ;
%let _EFIERR_ = 0; /* set the ERROR detection macro variable */
infile 'D:₩2019년 업무₩01-06.(조사연구)내부연구₩1.빅데이터분석기법₩03.데이터작업
₩data₩빅데이터분석기법연구_서울_추가자료.txt' delimiter='09'x MISSOVER DSD
lrecl=32767 firstobs=2 ;

informat 재단명 $600. ;
informat 고객번호 $600. ;
informat 고객형태 $600. ;
informat 심사번호 $600. ;
informat 보증일자 $600. ;
informat 보증번호 $600. ;
informat 조사번호 $600. ;
informat 자산총계_당기_백만원 $600. ;
informat 자산총계_전기_백만원 $600. ;
informat 자산총계_전전기_백만원 $600. ;
informat 유동자산_당기_백만원 $600. ;
informat 유동자산_전기_백만원 $600. ;
informat 유동자산_전전기_백만원 $600. ;
informat 부채총계_당기_백만원 $600. ;
informat 부채총계_전기_백만원 $600. ;
```

```
informat 부채총계_전전기_백만원 $600. ;
informat 유동부채_당기_백만원 $600. ;
informat 유동부채_전기_백만원 $600. ;
informat 유동부채_전전기_백만원 $600. ;
informat 자본총계_당기_백만원 $600. ;
informat 자본총계_전기_백만원 $600. ;
informat 자본총계_전전기_백만원 $600. ;
informat 매출액_당기_백만원 $600. ;
informat 매출액_전기_백만원 $600. ;
informat 매출액_전전기_백만원 $600. ;
informat 영업손익_당기_백만원 $600. ;
informat 영업손익_전기_백만원 $600. ;
informat 영업손익_전전기_백만원 $600. ;
informat 당기순이익_당기_백만원 $600. ;
informat 당기순이익_전기_백만원 $600. ;
informat 당기순이익_전전기_백만원 $600. ;
informat 업종구분 $600. ;
informat 직권말소여부 $600. ;
informat 성장성지수 $600. ;
informat 안정성지수 $600. ;
informat 밀집도지수 $600. ;
informat 구매력지수 $600. ;
informat 집객력지수 $600. ;
informat 최종사업성평가지수 $600. ;

format 재단명 $600. ;
format 고객번호 $600. ;
format 고객형태 $600. ;
format 심사번호 $600. ;
format 보증일자 $600. ;
format 보증번호 $600. ;
format 조사번호 $600. ;
```

```
format 자산총계_당기_백만원 $600. ;
format 자산총계_전기_백만원 $600. ;
format 자산총계_전전기_백만원 $600. ;
format 유동자산_당기_백만원 $600. ;
format 유동자산_전기_백만원 $600. ;
format 유동자산_전전기_백만원 $600. ;
format 부채총계_당기_백만원 $600. ;
format 부채총계_전기_백만원 $600. ;
format 부채총계_전전기_백만원 $600. ;
format 유동부채_당기_백만원 $600. ;
format 유동부채_전기_백만원 $600. ;
format 유동부채_전전기_백만원 $600. ;
format 자본총계_당기_백만원 $600. ;
format 자본총계_전기_백만원 $600. ;
format 자본총계_전전기_백만원 $600. ;
format 매출액_당기_백만원 $600. ;
format 매출액_전기_백만원 $600. ;
format 매출액_전전기_백만원 $600. ;
format 영업손익_당기_백만원 $600. ;
format 영업손익_전기_백만원 $600. ;
format 영업손익_전전기_백만원 $600. ;
format 당기순이익_당기_백만원 $600. ;
format 당기순이익_전기_백만원 $600. ;
format 당기순이익_전전기_백만원 $600. ;
format 업종구분 $600. ;
format 직권말소여부 $600. ;
format 성장성지수 $600. ;
format 안정성지수 $600. ;
format 밀집도지수 $600. ;
format 구매력지수 $600. ;
format 집객력지수 $600. ;
format 최종사업성평가지수 $600. ;
```

input
재단명 $
고객번호 $
고객형태 $
심사번호 $
보증일자 $
보증번호 $
조사번호 $
자산총계_당기_백만원 $
자산총계_전기_백만원 $
자산총계_전전기_백만원 $
유동자산_당기_백만원 $
유동자산_전기_백만원 $
유동자산_전전기_백만원 $
부채총계_당기_백만원 $
부채총계_전기_백만원 $
부채총계_전전기_백만원 $
유동부채_당기_백만원 $
유동부채_전기_백만원 $
유동부채_전전기_백만원 $
자본총계_당기_백만원 $
자본총계_전기_백만원 $
자본총계_전전기_백만원 $
매출액_당기_백만원 $
매출액_전기_백만원 $
매출액_전전기_백만원 $
영업손익_당기_백만원 $
영업손익_전기_백만원 $
영업손익_전전기_백만원 $
당기순이익_당기_백만원 $
당기순이익_전기_백만원 $
당기순이익_전전기_백만원 $

```
업종구분 $
직권말소여부 $
성장성지수 $
안정성지수 $
밀집도지수 $
구매력지수 $
집객력지수 $
최종사업성평가지수 $;
if _ERROR_ then call symputx('_EFIERR_',1);
run;

data BIGDATA.Raw_add_gg ;
%let _EFIERR_ = 0; /* set the ERROR detection macro variable */
infile 'D:₩2019년 업무₩01-06.(조사연구)내부연구₩1.빅데이터분석기법₩03.데이터작업
₩data₩빅데이터분석기법연구_경기_추가자료.txt' delimiter='09'x MISSOVER DSD
lrecl=32767 firstobs=2 ;

informat 재단명 $600. ;
informat 고객번호 $600. ;
informat 고객형태 $600. ;
informat 심사번호 $600. ;
informat 보증일자 $600. ;
informat 총자산금액 $600. ;
informat 자기자본금액 $600. ;
informat 납부자본금액 $600. ;
informat 매출금액 $600. ;
informat 당기순이익금액 $600. ;
informat 영업이익금액 $600. ;
informat 부채총계금액 $600. ;
informat 건물제곱미터단위면적 $600. ;
informat 건물평단위면적 $600. ;
informat 대지제곱미터단위면적 $600. ;
```

```
informat 대지평단위면적 $600. ;
informat 임차보증 금액 $600. ;
informat 월세 금액 $600. ;
informat 업종구분 $600. ;
informat 직권말소여부 $600. ;
informat 성장성지수 $600. ;
informat 안정성지수 $600. ;
informat 밀집도지수 $600. ;
informat 구매력지수 $600. ;
informat 집객력지수 $600. ;
informat 종사업성평가지수 $600. ;

format 재단명 $600. ;
format 고객번호 $600. ;
format 고객형태 $600. ;
format 심사번호 $600. ;
format 보증일자 $600. ;
format 총자산금액 $600. ;
format 자기자본금액 $600. ;
format 납부자본금액 $600. ;
format 매출금액 $600. ;
format 당기순이익금액 $600. ;
format 영업이익금액 $600. ;
format 부채총계금액 $600. ;
format 건물제곱미터단위면적 $600. ;
format 건물평단위면적 $600. ;
format 대지제곱미터단위면적 $600. ;
format 대지평단위면적 $600. ;
format 임차보증 금액 $600. ;
format 월세 금액 $600. ;
format 업종구분 $600. ;
format 직권말소여부 $600. ;
```

```
format 성장성지수 $600. ;
format 안정성지수 $600. ;
format 밀집도지수 $600. ;
format 구매력지수 $600. ;
format 집객력지수 $600. ;
format 종사업성평가지수 $600. ;

input
재단명 $
고객번호 $
고객형태 $
심사번호 $
보증일자 $
총자산금액 $
자기자본금액 $
납부자본금액 $
매출금액 $
당기순이익금액 $
영업이익금액 $
부채총계금액 $
건물제곱미터단위면적 $
건물평단위면적 $
대지제곱미터단위면적 $
대지평단위면적 $
임차보증 금액 $
월세 금액 $
업종구분 $
직권말소여부 $
성장성지수 $
안정성지수 $
밀집도지수 $
구매력지수 $
```

```
집객력지수 $
종사업성평가지수 $;
if _ERROR_ then call symputx('_EFIERR_',1);
run;

data Bigdata.Raw_add_14_su;
set Bigdata.Raw_add_14 Bigdata.Raw_add_su;
run;
proc contents data = Bigdata.Raw_add_14_su order = varnum; run;

proc sql;
create table Work.Raw_add_14_su_count as
select
count(재단명) as N001,
count(고객번호) as N002,
count(고객형태) as N003,
count(심사번호) as N004,
count(보증일자) as N005,
count(보증번호) as N006,
count(조사번호) as N007,
count(자산총계_당기_백만원) as N008,
count(자산총계_전기_백만원) as N009,
count(자산총계_전전기_백만원) as N010,
count(유동자산_당기_백만원) as N011,
count(유동자산_전기_백만원) as N012,
count(유동자산_전전기_백만원) as N013,
count(부채총계_당기_백만원) as N014,
count(부채총계_전기_백만원) as N015,
count(부채총계_전전기_백만원) as N016,
count(유동부채_당기_백만원) as N017,
count(유동부채_전기_백만원) as N018,
count(유동부채_전전기_백만원) as N019,
```

```
count(자본총계_당기_백만원) as N020,
count(자본총계_전기_백만원) as N021,
count(자본총계_전전기_백만원) as N022,
count(매출액_당기_백만원) as N023,
count(매출액_전기_백만원) as N024,
count(매출액_전전기_백만원) as N025,
count(영업손익_당기_백만원) as N026,
count(영업손익_전기_백만원) as N027,
count(영업손익_전전기_백만원) as N028,
count(당기순이익_당기_백만원) as N029,
count(당기순이익_전기_백만원) as N030,
count(당기순이익_전전기_백만원) as N031,
count(업종구분) as N032,
count(직권말소여부) as N033,
count(성장성지수) as N034,
count(안정성지수) as N035,
count(밀집도지수) as N036,
count(구매력지수) as N037,
count(집객력지수) as N038,
count(최종사업성평가지수) as N039
from Bigdata.Raw_add_14_su;
quit;

proc transpose data = Work.Raw_add_14_su_count out =
Work.Raw_add_14_su_count; run;

proc print data = Work.Raw_add_14_su_count; run;
proc delete data = Work.Raw_add_14_su_count; run;
proc contents data = Bigdata.Raw_add_gg order = varnum; run;

******* 변수의 데이터 유무 확인;
proc sql;
```

```
create table Work.Raw_add_gg_count as
select
count(재단명) as N001,
count(고객번호) as N002,
count(고객형태) as N003,
count(심사번호) as N004,
count(보증일자) as N005,
count(총자산금액) as N006,
count(자기자본금액) as N007,
count(납부자본금액) as N008,
count(매출금액) as N009,
count(당기순이익금액) as N010,
count(영업이익금액) as N011,
count(부채총계금액) as N012,
count(건물제곱미터단위면적) as N013,
count(건물평단위면적) as N014,
count(대지제곱미터단위면적) as N015,
count(대지평단위면적) as N016,
count(임차보증 금액) as N017,
count(월세 금액) as N018,
count(업종구분) as N019,
count(직권말소여부) as N020,
count(성장성지수) as N021,
count(안정성지수) as N022,
count(밀집도지수) as N023,
count(구매력지수) as N024,
count(집객력지수) as N025,
count(종사업성평가지수) as N026
from Bigdata.Raw_add_gg;
quit;

proc transpose data = Work.Raw_add_gg_count out = Work.Raw_add_gg_count;
```

```
run;

proc print data = Work.Raw_add_gg_count; run;
proc delete data = Work.Raw_add_gg_count; run;

******* 변수 내 0값의 개수 확인;
proc freq data = Bigdata.Raw_add_gg;
tables
총자산금액
자기자본금액
납부자본금액
매출금액
당기순이익금액
영업이익금액
부채총계금액
건물제곱미터단위면적
건물평단위면적
대지제곱미터단위면적
대지평단위면적
임차보증 금액
월세 금액
/nopercent nocum;
run;

******* '0'값만 있는 변수 삭제;
data Bigdata.Raw_add_gg_drop;
set Bigdata.Raw_add_gg;
drop
총자산금액
자기자본금액
납부자본금액
당기순이익금액
```

```
영업이익금액
건물제곱미터단위면적
건물평단위면적
대지제곱미터단위면적
대지평단위면적;
run;

******* 변수레이아웃 확인;
proc contents data = Bigdata.Raw_add_14_su order = varnum; run;
proc contents data = Bigdata.Raw_add_gg_drop order = varnum; run;

******* 5개 재단 변수 최대 길이 확인;
proc sql;
create table Work.Raw_add_14_su_length as
select
max(length(재단명)) as v001,
max(length(고객번호)) as v002,
max(length(고객형태)) as v003,
max(length(심사번호)) as v004,
max(length(보증일자)) as v005,
max(length(보증번호)) as v006,
max(length(조사번호)) as v007,
max(length(자산총계_당기_백만원)) as v008,
max(length(자산총계_전기_백만원)) as v009,
max(length(자산총계_전전기_백만원)) as v010,
max(length(유동자산_당기_백만원)) as v011,
max(length(유동자산_전기_백만원)) as v012,
max(length(유동자산_전전기_백만원)) as v013,
max(length(부채총계_당기_백만원)) as v014,
max(length(부채총계_전기_백만원)) as v015,
max(length(부채총계_전전기_백만원)) as v016,
max(length(유동부채_당기_백만원)) as v017,
```

```
max(length(유동부채_전기_백만원)) as v018,
max(length(유동부채_전전기_백만원)) as v019,
max(length(자본총계_당기_백만원)) as v020,
max(length(자본총계_전기_백만원)) as v021,
max(length(자본총계_전전기_백만원)) as v022,
max(length(매출액_당기_백만원)) as v023,
max(length(매출액_전기_백만원)) as v024,
max(length(매출액_전전기_백만원)) as v025,
max(length(영업손익_당기_백만원)) as v026,
max(length(영업손익_전기_백만원)) as v027,
max(length(영업손익_전전기_백만원)) as v028,
max(length(당기순이익_당기_백만원)) as v029,
max(length(당기순이익_전기_백만원)) as v030,
max(length(당기순이익_전전기_백만원)) as v031,
max(length(업종구분)) as v032,
max(length(직권말소여부)) as v033,
max(length(성장성지수)) as v034,
max(length(안정성지수)) as v035,
max(length(밀집도지수)) as v036,
max(length(구매력지수)) as v037,
max(length(집객력지수)) as v038,
max(length(최종사업성평가지수)) as v039
from Bigdata.Raw_add_14_su;
run;

proc transpose data = Work.Raw_add_14_su_length out =
      Work.Raw_add_14_su_length; run;
proc print data = Work.Raw_add_14_su_length; run;
proc delete data = Work.Raw_add_14_su_length; run;

******* 경기재단 변수 최대길이 확인;
proc sql;
```

```
create table Work.Raw_add_gg_length as
select
max(length(재단명)) as v001,
max(length(고객번호)) as v002,
max(length(고객형태)) as v003,
max(length(심사번호)) as v004,
max(length(보증일자)) as v005,
max(length(부채총계금액)) as v014,
max(length(매출금액)) as v023,
max(length(업종구분)) as v032,
max(length(직권말소여부)) as v033,
max(length(성장성지수)) as v034,
max(length(안정성지수)) as v035,
max(length(밀집도지수)) as v036,
max(length(구매력지수)) as v037,
max(length(집객력지수)) as v038,
max(length(종사업성평가지수)) as v039,
max(length(임차보증 금액)) as v040,
max(length(월세 금액)) as v041
from Bigdata.Raw_add_gg_drop;
run;

proc transpose data = Work.Raw_add_gg_length out = Work.Raw_add_gg_length;
run;
proc print data = Work.Raw_add_gg_length; run;
proc delete data = Work.Raw_add_gg_length; run;

******* 15재단 데이터 클리닝;
data Bigdata.Raw_add_14_su;
set Bigdata.Raw_add_14_su;
if 성장성지수 = '-' then 성장성지수 = '';
if 안정성지수 = '-' then 안정성지수 = '';
```

```
if 밀집도지수 = '-' then 밀집도지수 = '';
if 구매력지수 = '-' then 구매력지수 = '';
if 집객력지수 = '-' then 집객력지수 = '';
if 최종사업성평가지수 = '-' then 최종사업성평가지수 = '';
run;

******* 재단 변수 유형 수정;
data Work.Raw_add_14_su;
set Bigdata.Raw_add_14_su;
v001 = put(재단명, $4.);
v002 = put(고객번호, $8.);
v003 = put(고객형태, $10.);
v004 = put(심사번호, $12.);
v005 = input(보증일자, yymmdd10.); format v005 yymmdd10.;
v006 = put(보증번호, $12.);
v007 = put(조사번호, $6.);
v008 = input(자산총계_당기_백만원, 8.);
v009 = input(자산총계_전기_백만원, 8.);
v010 = input(자산총계_전전기_백만원, 8.);
v011 = input(유동자산_당기_백만원, 8.);
v012 = input(유동자산_전기_백만원, 8.);
v013 = input(유동자산_전전기_백만원, 8.);
v014 = input(부채총계_당기_백만원, 8.);
v015 = input(부채총계_전기_백만원, 8.);
v016 = input(부채총계_전전기_백만원, 8.);
v017 = input(유동부채_당기_백만원, 8.);
v018 = input(유동부채_전기_백만원, 8.);
v019 = input(유동부채_전전기_백만원, 8.);
v020 = input(자본총계_당기_백만원, 8.);
v021 = input(자본총계_전기_백만원, 8.);
v022 = input(자본총계_전전기_백만원, 8.);
v023 = input(매출액_당기_백만원, 8.);
```

```
v024 = input(매출액_전기_백만원, 8.);
v025 = input(매출액_전전기_백만원, 8.);
v026 = input(영업손익_당기_백만원, 8.);
v027 = input(영업손익_전기_백만원, 8.);
v028 = input(영업손익_전전기_백만원, 8.);
v029 = input(당기순이익_당기_백만원, 8.);
v030 = input(당기순이익_전기_백만원, 8.);
v031 = input(당기순이익_전전기_백만원, 8.);
v032 = put(업종구분, $49.);
v033 = put(직권말소여부, $1.);
v034 = input(성장성지수, 8.);
v035 = input(안정성지수, 8.);
v036 = input(밀집도지수, 8.);
v037 = input(구매력지수, 8.);
v038 = input(집객력지수, 8.);
v039 = input(최종사업성평가지수, 8.);

keep
v001 v002 v003 v004 v005 v006 v007 v008 v009 v010
v011 v012 v013 v014 v015 v016 v017 v018 v019 v020
v021 v022 v023 v024 v025 v026 v027 v028 v029 v030
v031 v032 v033 v034 v035 v036 v037 v038 v039;
run;

******* 경기재단 변수 유형 수정;
data Work.Raw_add_gg_drop;
set Bigdata.Raw_add_gg_drop;
v001 = put(재단명, $4.);
v002 = put(고객번호, $6.);
v003 = put(고객형태, $1.);
v004 = put(심사번호, $12.);
v005 = input(보증일자, yymmdd10.); format v005 yymmdd10.;
```

```
v014 = input(부채총계금액, 8.);
v023 = input(매출금액, 8.);
v032 = put(업종구분, $49.);
v033 = put(직권말소여부, $1.);
v034 = input(성장성지수, 8.);
v035 = input(안정성지수, 8.);
v036 = input(밀집도지수, 8.);
v037 = input(구매력지수, 8.);
v038 = input(집객력지수, 8.);
v040 = put(종사업성평가지수, $3.);
v041 = input(임차보증 금액, 8.);
v042 = input(월세 금액, 8.);

keep
v001 v002 v003 v004 v005 v014 v023 v032 v033 v034
v035 v036 v037 v038 v040 v041 v042;
run;

******* 경기재단 금액단위 수정;
data Work.Raw_add_gg_drop;
set Work.Raw_add_gg_drop;
if v014 ^= . then v014 = v014/1000000;
if v023 ^= . then v023 = v023/1000000;

if v014 ^= . then v014 = round(v014, 1);
if v023 ^= . then v023 = round(v023, 1);
run;

******* 경기&15개재단 데이터 열병합;
data Work.Raw_add;
set Work.Raw_add_14_su Work.Raw_add_gg_drop;
run;
```

```
proc datasets library = Work kill; run; quit;

******* 전체 데이터 변수명 수정;
data Bigdata.Raw_add;
set Work.Raw_add;
rename v001 = 재단명;
rename v002 = 고객번호;
rename v003 = 고객형태;
rename v004 = 심사번호;
rename v005 = 보증일자;
rename v006 = 보증번호;
rename v007 = 조사번호;
rename v008 = 자산총계_당기_백만원;
rename v009 = 자산총계_전기_백만원;
rename v010 = 자산총계_전전기_백만원;
rename v011 = 유동자산_당기_백만원;
rename v012 = 유동자산_전기_백만원;
rename v013 = 유동자산_전전기_백만원;
rename v014 = 부채총계_당기_백만원;
rename v015 = 부채총계_전기_백만원;
rename v016 = 부채총계_전전기_백만원;
rename v017 = 유동부채_당기_백만원;
rename v018 = 유동부채_전기_백만원;
rename v019 = 유동부채_전전기_백만원;
rename v020 = 자본총계_당기_백만원;
rename v021 = 자본총계_전기_백만원;
rename v022 = 자본총계_전전기_백만원;
rename v023 = 매출액_당기_백만원;
rename v024 = 매출액_전기_백만원;
rename v025 = 매출액_전전기_백만원;
rename v026 = 영업손익_당기_백만원;
rename v027 = 영업손익_전기_백만원;
```

```
rename v028 = 영업손익_전전기_백만원;
rename v029 = 당기순이익_당기_백만원;
rename v030 = 당기순이익_전기_백만원;
rename v031 = 당기순이익_전전기_백만원;
rename v032 = 업종구분;
rename v033 = 직권말소여부;
rename v034 = 성장성지수;
rename v035 = 안정성지수;
rename v036 = 밀집도지수;
rename v037 = 구매력지수;
rename v038 = 집객력지수;
rename v039 = 최종사업성평가지수;
rename v040 = 최종사업성평가지수_등급;
rename v041 = 임차보증 금액;
rename v042 = 월세 금액;
run;

proc datasets library = Work kill; run; quit;

******* 기존 데이터셋 정렬;
proc sort data = Bigdata.Raw_bigdata_fin out = Work.Raw_bigdata_fin; by 재단명 고객번호 보증일자 심사번호; run;

******* 추가 데이터셋 정렬;
proc sort data = Bigdata.Raw_add out = Work.Raw_add; by 재단명 고객번호 보증일자 심사번호; run;

******* 추가 데이터셋 불필요 변수 제거;
data Work.Raw_add;
set Work.Raw_add;
drop 고객형태 보증번호 조사번호 업종구분;
run;
```

```
******* 데이터셋 행병합;
data Bigdata.Raw_00;
merge Work.Raw_bigdata_fin Work.Raw_add;
by 재단명 고객번호 보증일자 심사번호;
run;

******* 최종 데이터셋;
data Bigdata.Raw_01;
set Bigdata.Raw_00;
if 직권말소여부 in ('Y', 'N');
run;

proc contents data = Bigdata.Raw_00 order = varnum; run;
proc contents data = Bigdata.Raw_01 order = varnum; run;

PROC IMPORT OUT= BIG.RAW_00 DATAFILE= "D:₩@02.보고서₩@2019년₩01-02.(분
석보고서)빅데이터분석기법이용소평모형구축₩분석프로그램₩full_data.txt"  DBMS=TAB
REPLACE; GETNAMES=NO; DATAROW=2;
RUN;

PROC IMPORT OUT= BIG.TR_00 DATAFILE= "D:₩@02.보고서₩@2019년₩01-02.(분석
보고서)빅데이터분석기법이용소평모형구축₩분석프로그램₩tr_data.txt"       DBMS=TAB
REPLACE; GETNAMES=NO; DATAROW=2;
RUN;

PROC IMPORT OUT= BIG.TS_00 DATAFILE= "D:₩@02.보고서₩@2019년₩01-02.(분석
보고서)빅데이터분석기법이용소평모형구축₩분석프로그램₩ts_data.txt"       DBMS=TAB
REPLACE; GETNAMES=NO; DATAROW=2;
RUN;
PROC CONTENTS DATA=BIG.RAW_00 VARNUM; RUN;
```

```
/************************************************/
/* 1. 최종 데이터셋 구축 */
/************************************************/

* 1.1 소평 등급이 있는 차주만 선택 ;
PROC FREQ DATA=BIG.RAW_00; TABLES 소평_CB등급 소평_최종신용등급; RUN;
PROC FREQ DATA=BIG.RAW_00; TABLES 소평_전략등급*소평_최종신용등급
/NOPERCENT NOROW NOCOL; RUN;

********** 임차 및 임대금액 관련 변수 비교 및 확인;
DATA BIG._0TMP01;
SET BIG.RAW_00;
KEEP
주사업장_임차보증금액_만원
주사업장_월세금액_천원
대표자실거주지_임차보증금_만원
대표자실거주지_월세금액_천원
거주_주택임차보증금_원단위
사업장임차보증금_원단위
소유부동산금액_원단위
임차보증금_원단위
임대보증금_사업장_원단위
임대보증금_주택_원단위;
RUN;

********** 종업원 수 관련 변수 비교 및 확인;
DATA BIG._0TMP02;
SET BIG.RAW_00;
KEEP
종업원수_명
상시종업원수_명;
RUN;
```

```
********** 차입 관련 변수 비교 및 확인;
DATA BIG._0TMP03;
SET BIG.RAW_00;
KEEP
차입금운전_백만원
차입금시설_백만원
차입금기타_백만원
차입금합계_백만원
기보증잔액_재단_백만원
기보증잔액_신보_백만원
기보증잔액_기보_백만원
기보증잔액_개인보증_백만원
차입금합계_백만원
기보증잔액_재단_백만원
기보증잔액_신보_백만원
기보증잔액_기보_백만원
기보증잔액_개인보증_백만원
차입금_담보제외_원단위
현금서비스금액_원단위;
RUN;

********** 매출 관련 변수 비교 및 확인;
DATA BIG._0TMP04;
SET BIG.RAW_00;
KEEP
매출액_원단위
전기매출액_백만원
당기매출액_백만원
매출액_월평균_백만원
연간신고매출액_원단위
연간실제매출액_원단위
평균매출액_월_원단위;
```

```
RUN;

********** 사업장 및 주택 소유 여부 관련 변수 비교 및 확인;
DATA BIG._0TMP05;
SET BIG.RAW_00;
KEEP
주사업장_소유여부
대표자실거주지_소유여부
사업장_자가구분
주택_자가구분
보유부동산;
RUN;

* 1.2 변수 레이아웃 1 ;
PROC CONTENTS DATA=BIG.RAW_01 VARNUM; RUN;

* 1.3 기초분포 확인 ;
PROC FREQ DATA=BIG.RAW_01; TABLES 소평_CB등급 소평_최종신용등급; RUN;
PROC FREQ DATA=BIG.RAW_01; TABLES 고객형태; RUN;
PROC FREQ DATA=BIG.RAW_01; TABLES 기업형태; RUN;
PROC FREQ DATA=BIG.RAW_01; TABLES 기업규모; RUN;
PROC FREQ DATA=BIG.RAW_01; TABLES 주사업장_소유여부; RUN;

* 1.4 분석 변수 선택 - 1차;
DATA BIG.RAW_02;
SET BIG.RAW_01;
KEEP
고객번호
고객형태
업종코드
종업원수_명
매출액_원단위
```

주사업장_소유여부
주사업장_임차보증금액_만원
주사업장_월세금액_천원
대표자실거주지_소유여부
대표자실거주지_임차보증금_만원
대표자실거주지_월세금액_천원
전기매출액_백만원
당기매출액_백만원
매출액_월평균_백만원
상시종업원수_명
차입금운전_백만원
차입금시설_백만원
차입금기타_백만원
차입금합계_백만원
기보증잔액_재단_백만원
기보증잔액_신보_백만원
기보증잔액_기보_백만원
기보증잔액_개인보증_백만원
연간신고매출액_원단위
연간실제매출액_원단위
사업장_자가구분
주택_자가구분
차입기관수_담보제외_개수
차입금_담보제외_원단위
현금서비스금액_원단위
거주_주택임차보증금_원단위
사업장임차보증금_원단위
보유부동산
업력_개월
거주기간_개월
평균매출액_월_원단위
영업이익_월_원단위

```
배우자소득_월_원단위
기타수익_월_원단위
소유부동산금액_원단위
임차보증금_원단위
임대보증금_사업장_원단위
임대보증금_주택_원단위
금융자산_예금적금금액_원단위
금융자산_유가증권금액_원단위
기타현금금액_재고자산_원단위
기타현금금액_고정자산_원단위
기타현금금액_권리금_원단위
기타현금금액_기타_원단위
소평_CB등급
소평_최종신용등급
사고일자
직권말소여부;
RUN;

********** 임차 및 임대금액 관련 변수 비교 및 확인;
DATA BIG._TMP01;
SET BIG.RAW_02;
KEEP
주사업장_임차보증금액_만원
주사업장_월세금액_천원
대표자실거주지_임차보증금_만원
대표자실거주지_월세금액_천원
거주_주택임차보증금_원단위
사업장임차보증금_원단위
소유부동산금액_원단위
임차보증금_원단위
임대보증금_사업장_원단위
임대보증금_주택_원단위;
```

```
RUN;
********** 종업원 수 관련 변수 비교 및 확인;
DATA BIG._TMP02;
SET BIG.RAW_02;
KEEP
종업원수_명
상시종업원수_명;
RUN;

********** 차입 관련 변수 비교 및 확인;
DATA BIG._TMP03;
SET BIG.RAW_02;
KEEP
차입금운전_백만원
차입금시설_백만원
차입금기타_백만원
차입금합계_백만원
기보증잔액_재단_백만원
기보증잔액_신보_백만원
기보증잔액_기보_백만원
기보증잔액_개인보증_백만원
차입금합계_백만원
기보증잔액_재단_백만원
기보증잔액_신보_백만원
기보증잔액_기보_백만원
기보증잔액_개인보증_백만원
차입금_담보제외_원단위
현금서비스금액_원단위;
RUN;

********** 매출 관련 변수 비교 및 확인;
DATA BIG._TMP04;
```

```
SET BIG.RAW_02;
KEEP
매출액_원단위
전기매출액_백만원
당기매출액_백만원
매출액_월평균_백만원
연간신고매출액_원단위
연간실제매출액_원단위
평균매출액_월_원단위;
RUN;

********** 사업장 및 주택 소유 여부 관련 변수 비교 및 확인;
DATA BIG._TMP05;
SET BIG.RAW_02;
KEEP
주사업장_소유여부
대표자실거주지_소유여부
사업장_자가구분
주택_자가구분
보유부동산;
RUN;

* 1.5 분석 변수 선택 - 2차;
DATA BIG.RAW_03;
SET BIG.RAW_02;
KEEP
고객번호
고객형태
업종코드
종업원수_명
주사업장_소유여부
주사업장_임차보증금액_만원
```

주사업장_월세금액_천원
대표자실거주지_소유여부
대표자실거주지_임차보증금_만원
대표자실거주지_월세금액_천원
차입금운전_백만원
차입금시설_백만원
차입금기타_백만원
기보증잔액_재단_백만원
기보증잔액_신보_백만원
기보증잔액_기보_백만원
기보증잔액_개인보증_백만원
차입기관수_담보제외_개수
현금서비스금액_원단위
보유부동산
업력_개월
거주기간_개월
평균매출액_월_원단위
영업이익_월_원단위
배우자소득_월_원단위
기타수익_월_원단위
소유부동산금액_원단위
임대보증금_사업장_원단위
임대보증금_주택_원단위
금융자산_예금적금금액_원단위
금융자산_유가증권금액_원단위
기타현금금액_재고자산_원단위
기타현금금액_고정자산_원단위
기타현금금액_권리금_원단위
기타현금금액_기타_원단위
소평_CB등급
소평_최종신용등급
사고일자

```
직권말소여부;
RUN;

* 1.6 변수 레이아웃 2 ;
PROC CONTENTS DATA=BIG.RAW_03 VARNUM; RUN;

* 1.7 사고여부(Y) 변수 생성 ;
DATA BIG.RAW_04;
SET BIG.RAW_03;
IF 사고일자 = . THEN Y = 0; ELSE Y = 1;
DROP 사고일자;
RUN;

PROC FREQ DATA=BIG.RAW_04; TABLES Y; RUN;

* 1.8 업종 변수 변환 ;
DATA WORK.INDUS1;
SET BIG.RAW_04;
INDUS1 = SUBSTR(업종코드,1,1);
RUN;

PROC FREQ DATA=WORK.INDUS1; TABLES INDUS1; RUN;

DATA WORK.INDUS1;
SET WORK.INDUS1;
IF INDUS1 = 'C' THEN 업종 = '1.제조업 '; ELSE
IF INDUS1 = 'J' THEN 업종 = '2.서비스업  '; ELSE
IF INDUS1 = 'M' THEN 업종 = '2.서비스업  '; ELSE
IF INDUS1 = 'N' THEN 업종 = '2.서비스업  '; ELSE
IF INDUS1 = 'P' THEN 업종 = '2.서비스업  '; ELSE
IF INDUS1 = 'Q' THEN 업종 = '2.서비스업  '; ELSE
IF INDUS1 = 'R' THEN 업종 = '2.서비스업  '; ELSE
```

```
IF INDUS1 = 'S' THEN 업종 = '2.서비스업  '; ELSE
IF INDUS1 = 'G' THEN 업종 = '3.도소매업  '; ELSE
IF INDUS1 = 'I' THEN 업종 = '4.음식숙박업'; ELSE
IF INDUS1 = 'F' THEN 업종 = '5.건설업 '; ELSE
IF INDUS1 = 'H' THEN 업종 = '6.운수업 '; ELSE 업종 = '7.기타업 ';
RUN;

DATA BIG.RAW_05;
SET WORK.INDUS1;
DROP 업종코드 INDUS1;
RUN;

PROC FREQ DATA=BIG.RAW_05; TABLES 업종; RUN;

* 1.9 직권말소 변수 변환 ;
PROC FREQ DATA=BIG.RAW_05; TABLES 직권말소여부; RUN;

DATA BIG.RAW_06;
SET BIG.RAW_05;
IF 직권말소여부 = 'N' THEN 직권말소 = 1; ELSE 직권말소 = 0;
DROP 직권말소여부;
RUN;

* 1.10 소평등급 변환 ;
DATA BIG.RAW_07;
SET BIG.RAW_06;
IF 소평_최종신용등급 = 'AAA' THEN 소평등급 = 1; ELSE
IF 소평_최종신용등급 = 'AA'  THEN 소평등급 = 2; ELSE
IF 소평_최종신용등급 = 'A' THEN 소평등급 = 3; ELSE
IF 소평_최종신용등급 = 'BBB' THEN 소평등급 = 4; ELSE
IF 소평_최종신용등급 = 'BB'  THEN 소평등급 = 5; ELSE
IF 소평_최종신용등급 = 'B' THEN 소평등급 = 6; ELSE
```

```
IF 소평_최종신용등급 = 'CCC' THEN 소평등급 = 7; ELSE 소평등급 = 8;
DROP 소평_최종신용등급;
RUN;

* 1.11 변수명 변환 ;
DATA BIG.RAW_FIN;
SET BIG.RAW_07;
RENAME 고객형태 = 고객형태 ;
RENAME 종업원수_명 = 종업원수 ;
RENAME 주사업장_소유여부 = 주사업장소유여부 ;
RENAME 주사업장_임차보증금액_만원 = 주사업장임차보증금액 ;
RENAME 주사업장_월세금액_천원 = 주사업장월세금액 ;
RENAME 대표자실거주지_소유여부 = 실거주지소유여부 ;
RENAME 대표자실거주지_임차보증금_만원 = 실거주지임차보증금액 ;
RENAME 대표자실거주지_월세금액_천원 = 실소유지월세금액 ;
RENAME 차입금운전_백만원 = 차입금운전 ;
RENAME 차입금시설_백만원 = 차입금시설 ;
RENAME 차입금기타_백만원 = 차입금기타 ;
RENAME 기보증잔액_재단_백만원 = 기보증잔액재단 ;
RENAME 기보증잔액_신보_백만원 = 기보증잔액신보 ;
RENAME 기보증잔액_기보_백만원 = 기보증잔액기보 ;
RENAME 기보증잔액_개인보증_백만원 = 기보증잔액개인 ;
RENAME 차입기관수_담보제외_개수 = 담보제외차입기관수 ;
RENAME 현금서비스금액_원단위 = 현금서비스금액 ;
RENAME 보유부동산 = 보유부동산 ;
RENAME 업력_개월 = 업력 ;
RENAME 거주기간_개월 = 거주기간 ;
RENAME 평균매출액_월_원단위 = 월평균매출액 ;
RENAME 영업이익_월_원단위 = 월영업이익 ;
RENAME 배우자소득_월_원단위 = 월배우자소득 ;
RENAME 기타수익_월_원단위 = 월기타수익 ;
RENAME 소유부동산금액_원단위 = 소유부동산금액 ;
```

```
RENAME 임대보증금_사업장_원단위 = 임대보증금사업장 ;
RENAME 임대보증금_주택_원단위 = 임대보증금주택 ;
RENAME 금융자산_예금적금금액_원단위 = 예적금금액 ;
RENAME 금융자산_유가증권금액_원단위 = 유가증권금액 ;
RENAME 기타현금금액_재고자산_원단위 = 재고자산 ;
RENAME 기타현금금액_고정자산_원단위 = 고정자산 ;
RENAME 기타현금금액_권리금_원단위 = 권리금 ;
RENAME 기타현금금액_기타_원단위 = 기타현금 ;
RENAME 소평_CB등급 = CB등급 ;
RENAME Y = 사고여부 ; RUN;

/************************************************/
/* 2. 기초통계 분석 */
/************************************************/

* 2.1 단변량 분석 ;
PROC FREQ  DATA=BIG.RAW_FIN; TABLES 사고여부 ; RUN;
PROC FREQ  DATA=BIG.RAW_FIN; TABLES 고객형태 ; RUN;
PROC FREQ  DATA=BIG.RAW_FIN; TABLES 업종 ; RUN;
PROC FREQ  DATA=BIG.RAW_FIN; TABLES 주사업장소유여부 ; RUN;
PROC FREQ  DATA=BIG.RAW_FIN; TABLES 실거주지소유여부 ; RUN;
PROC FREQ  DATA=BIG.RAW_FIN; TABLES 보유부동산 ; RUN;
PROC FREQ  DATA=BIG.RAW_FIN; TABLES 직권말소 ; RUN;
PROC FREQ  DATA=BIG.RAW_FIN; TABLES CB등급 ; RUN;
PROC FREQ  DATA=BIG.RAW_FIN; TABLES 소평등급 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR  종업원수 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR  주사업장임차보증금액; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR  주사업장월세금액 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR  실거주지임차보증금액; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR  실소유지월세금액 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR  차입금운전 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR  차입금시설 ; RUN;
```

```
PROC MEANS DATA=BIG.RAW_FIN; VAR 차입금기타 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 기보증잔액재단 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 기보증잔액신보 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 기보증잔액기보 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 기보증잔액개인 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 담보제외차입기관수 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 현금서비스금액 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 업력 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 거주기간 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 월평균매출액 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 월영업이익 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 월배우자소득 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 월기타수익 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 소유부동산금액 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 임대보증금사업장 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 임대보증금주택 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 예적금금액 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 유가증권금액 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 재고자산 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 고정자산 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 권리금 ; RUN;
PROC MEANS DATA=BIG.RAW_FIN; VAR 기타현금 ; RUN;

* 2.2 연속형 변수 0 또는 결측치 비율 산출 ;
DATA WORK.RAW_FIN;
SET BIG.RAW_FIN;
IF 종업원수 <= 0 THEN 종업원수 = 1 ; ELSE 종업원수 = 0;
IF 주사업장임차보증금액 <= 0 THEN 주사업장임차보증금액 = 1 ; ELSE 주사업장임
차보증금액= 0;
IF 주사업장월세금액 <= 0 THEN 주사업장월세금액 = 1 ; ELSE 주사업장월세금액
= 0;
IF 실거주지임차보증금액 <= 0 THEN 실거주지임차보증금액 = 1 ; ELSE 실거주지임
```

```
차보증금액= 0;
IF 실소유지월세금액  <= 0 THEN 실소유지월세금액  = 1 ; ELSE 실소유지월세금액
= 0;
IF 차입금운전  <= 0 THEN 차입금운전 = 1 ; ELSE 차입금운전 = 0;
IF 차입금시설 <= 0 THEN 차입금시설 = 1 ; ELSE 차입금시설 = 0;
IF 차입금기타  <= 0 THEN 차입금기타 = 1 ; ELSE 차입금기타 = 0;
IF 기보증잔액재단  <= 0 THEN 기보증잔액재단  = 1 ; ELSE 기보증잔액재단 = 0;
IF 기보증잔액신보 <= 0 THEN 기보증잔액신보  = 1 ; ELSE 기보증잔액신보 = 0;
IF 기보증잔액기보  <= 0 THEN 기보증잔액기보 = 1 ; ELSE 기보증잔액기보 = 0;
IF 기보증잔액개인 <= 0 THEN 기보증잔액개인  = 1 ; ELSE 기보증잔액개인  = 0;
IF 담보제외차입기관수 <= 0 THEN 담보제외차입기관수  = 1 ; ELSE 담보제외차입기
관수 = 0;
IF 현금서비스금액 <= 0 THEN 현금서비스금액 = 1 ; ELSE 현금서비스금액 = 0;
IF 업력  <= 0 THEN 업력  = 1 ; ELSE 업력 = 0;
IF 거주기간  <= 0 THEN 거주기간  = 1 ; ELSE 거주기간 = 0;
IF 월평균매출액 <= 0 THEN 월평균매출액 = 1 ; ELSE 월평균매출액 = 0;
IF 월영업이익  <= 0 THEN 월영업이익 = 1 ; ELSE 월영업이익 = 0;
IF 월배우자소득 <= 0 THEN 월배우자소득 = 1 ; ELSE 월배우자소득  = 0;
IF 월기타수익 <= 0 THEN 월기타수익 = 1 ; ELSE 월기타수익 = 0;
IF 소유부동산금액  <= 0 THEN 소유부동산금액 = 1 ; ELSE 소유부동산금액 = 0;
IF 임대보증금사업장 <= 0 THEN 임대보증금사업장 = 1 ; ELSE 임대보증금사업장 =
0;
IF 임대보증금주택 <= 0 THEN 임대보증금주택 = 1 ; ELSE 임대보증금주택 = 0;
IF 예적금금액 <= 0 THEN 예적금금액 = 1 ; ELSE 예적금금액 = 0;
IF 유가증권금액 <= 0 THEN 유가증권금액 = 1 ; ELSE 유가증권금액  = 0;
IF 재고자산 <= 0 THEN 재고자산  = 1 ; ELSE 재고자산 = 0;
IF 고정자산  <= 0 THEN 고정자산 = 1 ; ELSE 고정자산 = 0;
IF 권리금  <= 0 THEN 권리금 = 1 ; ELSE 권리금 = 0;
IF 기타현금 <= 0 THEN 기타현금 = 1 ; ELSE 기타현금 = 0;
RUN;

PROC MEANS DATA=WORK.RAW_FIN; VAR  종업원수 ; RUN;
```

```
PROC MEANS DATA=WORK.RAW_FIN; VAR 주사업장임차보증금액; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 주사업장월세금액 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 실거주지임차보증금액; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 실소유지월세금액 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 차입금운전 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 차입금시설 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 차입금기타 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 기보증잔액재단 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 기보증잔액신보 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 기보증잔액기보 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 기보증잔액개인 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 담보제외차입기관수 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 현금서비스금액 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 업력 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 거주기간 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 월평균매출액 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 월영업이익 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 월배우자소득 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 월기타수익 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 소유부동산금액 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 임대보증금사업장 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 임대보증금주택 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 예적금금액 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 유가증권금액 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 재고자산 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 고정자산 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 권리금 ; RUN;
PROC MEANS DATA=WORK.RAW_FIN; VAR 기타현금 ; RUN;

* 2.3 이변량 분석 ;
PROC FREQ DATA=BIG.RAW_FIN; TABLES 고객형태 * 사고여부/NOPERCENT
NOROW NOCOL; RUN;
```

```
PROC FREQ DATA=BIG.RAW_FIN; TABLES 업종 * 사고여부/NOPERCENT NOROW NOCOL; RUN;
PROC FREQ DATA=BIG.RAW_FIN; TABLES 주사업장소유여부 * 사고여부 /NOPERCENT NOROW NOCOL; RUN;
PROC FREQ DATA=BIG.RAW_FIN; TABLES 실거주지소유여부 * 사고여부 /NOPERCENT NOROW NOCOL; RUN;
PROC FREQ DATA=BIG.RAW_FIN; TABLES 보유부동산 * 사고여부/NOPERCENT NOROW NOCOL; RUN;
PROC FREQ DATA=BIG.RAW_FIN; TABLES 직권말소 * 사고여부/NOPERCENT NOROW NOCOL; RUN;
PROC FREQ DATA=BIG.RAW_FIN; TABLES 고객형태 * 사고여부/NOFREQ NOPERCENT NOCOL; RUN;
PROC FREQ DATA=BIG.RAW_FIN; TABLES 업종 * 사고여부/NOFREQ NOPERCENT NOCOL; RUN;
PROC FREQ DATA=BIG.RAW_FIN; TABLES 주사업장소유여부 * 사고여부/NOFREQ NOPERCENT NOCOL; RUN;
PROC FREQ DATA=BIG.RAW_FIN; TABLES 실거주지소유여부 * 사고여부/NOFREQ NOPERCENT NOCOL; RUN;
PROC FREQ DATA=BIG.RAW_FIN; TABLES 보유부동산 * 사고여부/NOFREQ NOPERCENT NOCOL; RUN;
PROC FREQ DATA=BIG.RAW_FIN; TABLES 직권말소 * 사고여부/NOFREQ NOPERCENT NOCOL; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 종업원수 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 주사업장 임차보증금액; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 주사업장 월세금액 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 실거주지 임차보증금액; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 실소유지 월세금액 ; WHERE 사고여부 = 0; RUN;
```

```
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 차입금운전 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 차입금시설 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 차입금기타 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 기보증잔액재단 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 기보증잔액신보 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 기보증잔액기보 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 기보증잔액개인 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 담보제외차입기관수 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 현금서비스금액 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 업력 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 거주기간 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 월평균매출액 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 월영업이익 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 월배우자소득 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 월기타수익 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 소유부동산금액 ; WHERE 사고여부 = 0; RUN;
```

```
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 임대보증금사업장 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 임대보증금주택 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 예적금금액 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 유가증권금액 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 재고자산 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 고정자산 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 권리금 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 기타현금 ; WHERE 사고여부 = 0; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 종업원수 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 주사업장임차보증금액; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 주사업장월세금액 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 실거주지임차보증금액; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 실소유지월세금액 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 차입금운전 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 차입금시설 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 차입금기타 ; WHERE 사고여부 = 1; RUN;
```

PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 기보증잔액재단 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 기보증잔액신보 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 기보증잔액기보 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 기보증잔액개인 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 담보제외차입기관수 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 현금서비스금액 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 업력 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 거주기간 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 월평균매출액 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 월영업이익 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 월배우자소득 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 월기타수익 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 소유부동산금액 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 임대보증금사업장 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 임대보증금주택 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR 예적금금액 ; WHERE 사고여부 = 1; RUN;

```
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR  유가증권금액  ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR  재고자산 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR  고정자산 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR  권리금 ; WHERE 사고여부 = 1; RUN;
PROC MEANS DATA=BIG.RAW_FIN N MEAN STD; CLASS 사고여부; VAR  기타현금 ; WHERE 사고여부 = 1; RUN;

/************************************************/
/* 3. 데이터 분할 및 분석 변수 선택 */
/************************************************/

* 3.1 데이터 분할;
DATA BIG.RAW_FIN;
SET BIG.RAW_FIN;
RAN = RANUNI(1234567);
RUN;

PROC SQL;
CREATE TABLE BIG.RAW_FIN AS
SELECT
고객번호, 고객형태, 업종, 종업원수, 주사업장소유여부, 주사업장임차보증금액, 주사업장월세금액, 실거주지소유여부,
실거주지임차보증금액, 실소유지월세금액, 차입금운전, 차입금시설, 차입금기타, 기보증잔액재단, 기보증잔액신보,
기보증잔액기보, 기보증잔액개인, 담보제외차입기관수, 현금서비스금액, 보유부동산, 업력, 거주기간, 월평균매출액,
월영업이익, 월배우자소득, 월기타수익, 소유부동산금액, 임대보증금사업장, 임대보증금주택, 예적금금액, 유가증권금액,
```

```
재고자산, 고정자산, 권리금, 기타현금, 직권말소, CB등급, 소평등급, 사고여부, RAN
FROM BIG.RAW_FIN;
RUN; QUIT;

DATA BIG.TR_DATA BIG.TS_DATA;
SET BIG.RAW_FIN;
IF RAN <= 0.3 THEN OUTPUT BIG.TS_DATA; ELSE OUTPUT BIG.TR_DATA;
RUN;

DATA BIG.TR_DATA; SET BIG.TR_DATA; DROP RAN; RUN;
DATA BIG.TS_DATA; SET BIG.TS_DATA; DROP RAN; RUN;
* 3.2 분석 변수 선택;
********** FINE CLASSING ;
%include "D:₩@02.보고서₩@2019년₩01-02.(분석보고서)빅데이터분석기법이용소평모
형구축₩분석프로그램₩01 기본
Macro_Scoring_Macro_TwoTarget_pdo40_20120524.sas";
%include "D:₩@02.보고서₩@2019년₩01-02.(분석보고서)빅데이터분석기법이용소평모
형구축₩분석프로그램₩02 기본Macro_FineClassing_Macro(V).sas";

DATA BIG.TR_DATA; SET BIG.TR_DATA; WW=1; RUN;

%FINECLASS_V(BIG.TR_DATA,사고여부,사고여부,WW,고객형태,직권말
소,20,0,BIG.FINECLASS,FINECLASS,2);

* 매크로 적용 예시;
/*fineclass_v(data, target1, target2, ww, first, last, grp, psi_grp, file_name,
result_name, output_type);*/

* DATA = DATA SET NAME, TARGET1/TARGET2=TARGET 변수, 2개 까지 넣을 수
있는 MACRO WW=WEIGHT;
* FIRST=FINE CLASSING을 시작할 변수, LAST=FINE CLASSING을 그만 둘 변수,
GRP=FINE CLASSING 구분하는 구간수;
```

```
* PSI_GRP : PSI를 보고 싶은 대상 구분 COLUMN.  EX) 개발 & 검증 DATASET이 하
   나에 있을 경우 사용, 없을 경우 그냥 0 입력 0이 PSI를 볼 대상자, 그 뒤로 1, 2,
   3 등 계속 연속형으로 나가면 됨;
* FILE_NAME  EXPORT해서 생성할 파일 이름;
* RESULT_NAME FINECLASSING OUTPUT SAS FILE 이름;
* OUTPUT_TYPE 1: HTML OUTPUT - (1) &FILE_NAME._상세.HTML, (2)
&FILE_NAME._요약.HTML 등 2개 파일 생성
 2: CSV OUTPUT - (1) &FILE_NAME._상세.CSV, (2) &FILE_NAME._요약.CSV 등 2개 파
    일 생성
 3: EXCEL OUTPUT - (1) &FILE_NAME._상세.EXCEL, (2) &FILE_NAME._요약.EXCEL 등
    2개 파일 생성;

********** 범주형 독립변수 : 카이제곱 검정 ;
PROC FREQ DATA=BIG.RAW_FIN; TABLES 고객형태 * 사고여부/NOFREQ
NOPERCENT NOROW NOCOL CHISQ; RUN;
PROC FREQ DATA=BIG.RAW_FIN; TABLES 업종 * 사고여부/NOFREQ NOPERCENT
NOROW NOCOL CHISQ; RUN;
PROC FREQ DATA=BIG.RAW_FIN; TABLES 주사업장소유여부 * 사고여부/NOFREQ
NOPERCENT NOROW NOCOL CHISQ; RUN;
PROC FREQ DATA=BIG.RAW_FIN; TABLES 실거주지소유여부 * 사고여부/NOFREQ
NOPERCENT NOROW NOCOL CHISQ; RUN;
PROC FREQ DATA=BIG.RAW_FIN; TABLES 보유부동산 * 사고여부/NOFREQ
NOPERCENT NOROW NOCOL CHISQ; RUN;
PROC FREQ DATA=BIG.RAW_FIN; TABLES 직권말소 * 사고여부/NOFREQ
NOPERCENT NOROW NOCOL CHISQ; RUN;

********** 연속형 독립변수 : t 검정 ;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 종업원수 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 주사업장임차보증금액  ;
RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 주사업장월세금액  ; RUN;
```

```
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 실거주지임차보증금액 ;
RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 실소유지월세금액 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 차입금운전 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 차입금시설 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 차입금기타 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 기보증잔액재단 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 기보증잔액신보 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 기보증잔액기보 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 기보증잔액개인 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 담보제외차입기관수 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 현금서비스금액 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 업력 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 거주기간 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 월평균매출액 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 월영업이익 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 월배우자소득 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 월기타수익 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 소유부동산금액 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 임대보증금사업장 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 임대보증금주택 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 예적금금액 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 유가증권금액 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 재고자산 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 고정자산 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 권리금 ; RUN;
PROC TTEST DATA=BIG.RAW_FIN; CLASS 사고여부; VAR 기타현금 ; RUN;

* 3.3 COARSE CLASSING;

DATA BIG.TR_DATA_COARSE;
SET BIG.TR_DATA;
```

```
IF 고객형태 = ' ' THEN R1_고객형태=1; ELSE
IF 고객형태 = '개인사업자' THEN R1_고객형태=1; ELSE R1_고객형태=0;
IF 업종 = '1.제조업' THEN R1_업종=0; ELSE
IF 업종 = '2.서비스업' THEN R1_업종=1; ELSE
IF 업종 = '3.도소매업' THEN R1_업종=1; ELSE
IF 업종 = '4.음식숙박업' THEN R1_업종=0; ELSE
IF 업종 = '5.건설업' THEN R1_업종=0; ELSE
IF 업종 = '6.운수업' THEN R1_업종=1; ELSE R1_업종=2;
IF 주사업장소유여부 = ' ' THEN R1_주사업장소유여부=2; ELSE
IF 주사업장소유여부 = '가족소유' THEN R1_주사업장소유여부=2; ELSE
IF 주사업장소유여부 = '기타' THEN R1_주사업장소유여부=1; ELSE
IF 주사업장소유여부 = '무점포' THEN R1_주사업장소유여부=2; ELSE
IF 주사업장소유여부 = '임차' THEN R1_주사업장소유여부=0; ELSE
IF 주사업장소유여부 = '임차(보증금20백만원초과)' THEN R1_주사업장소유여부=1;
ELSE
IF 주사업장소유여부 = '자가' THEN R1_주사업장소유여부=2; ELSE R1_주사업장소유
여부=2;
IF 주사업장임차보증금액 <= 0 THEN R1_주사업장임차보증금액=1; ELSE
IF 0 < 주사업장임차보증금액 <= 1000 THEN R1_주사업장임차보증금액=0; ELSE
IF 1000 < 주사업장임차보증금액 THEN R1_주사업장임차보증금액=1;
IF 주사업장월세금액 <= 0 THEN R1_주사업장월세금액=1; ELSE R1_주사업장월세금액
=0;
IF 실거주지소유여부 = ' ' THEN R1_실거주지소유여부=1; ELSE
IF 실거주지소유여부 = '가족소유' THEN R1_실거주지소유여부=0; ELSE
IF 실거주지소유여부 = '기타' THEN R1_실거주지소유여부=0; ELSE
IF 실거주지소유여부 = '임차' THEN R1_실거주지소유여부=0; ELSE
IF 실거주지소유여부 = '임차(보증금20백만원초과)' THEN R1_실거주지소유여부=0;
ELSE
IF 실거주지소유여부 = '자가' THEN R1_실거주지소유여부=1; ELSE R1_실거주지소유
여부=1;
IF 실거주지임차보증금액 <= 0 THEN R1_실거주지임차보증금액=1; ELSE
IF 실거주지임차보증금액 <= 5000 THEN R1_실거주지임차보증금액=0; ELSE R1_실
```

```
거주지임차보증금액=1;
IF 실소유지월세금액  <= 0 THEN R1_실소유지월세금액=1; ELSE R1_실소유지월세금액=0;
IF 차입금운전 <= 56.6 THEN R1_차입금운전=0; ELSE R1_차입금운전=1;
IF 기보증잔액재단 <= 6 THEN R1_기보증잔액재단=1; ELSE R1_기보증잔액재단=0;
IF 기보증잔액기보 <= 15.8  THEN R1_기보증잔액기보=1; ELSE R1_기보증잔액기보=0;
IF 담보제외차입기관수 <= 1 THEN R1_담보제외차입기관수=2; ELSE
IF 담보제외차입기관수 <= 3 THEN R1_담보제외차입기관수=1; ELSE R1_담보제외차입기관수=0;

IF 현금서비스금액  <= 0 THEN R1_현금서비스금액=1; ELSE R1_현금서비스금액=0;
IF 보유부동산 = ' '  THEN R1_보유부동산=2; ELSE
IF 보유부동산 = '다가구' THEN R1_보유부동산=2; ELSE
IF 보유부동산 = '다세대' THEN R1_보유부동산=1; ELSE
IF 보유부동산 = '단독주택' THEN R1_보유부동산=2; ELSE
IF 보유부동산 = '아파트' THEN R1_보유부동산=2; ELSE
IF 보유부동산 = '없음'  THEN R1_보유부동산=0; ELSE
IF 보유부동산 = '임야기타부동산' THEN R1_보유부동산=2; ELSE R1_보유부동산=1;
IF 업력 <= 0 THEN R1_업력=2; ELSE
IF 업력 <=  15 THEN R1_업력=1; ELSE
IF 업력 <=  44 THEN R1_업력=0; ELSE
IF 업력 <= 124 THEN R1_업력=1; ELSE R1_업력=2;
IF 거주기간 <= . THEN R1_거주기간 = 0; ELSE
IF 거주기간 <=  16 THEN R1_거주기간 = 1; ELSE
IF 거주기간 <=  67 THEN R1_거주기간 = 0; ELSE
IF 거주기간 <= 261 THEN R1_거주기간 = 1; ELSE R1_거주기간=0;
IF 월평균매출액 <= . THEN R1_월평균매출액 = 2; ELSE
IF 월평균매출액 <=  8500000 THEN R1_월평균매출액 = 0; ELSE
IF 월평균매출액 <= 38750000 THEN R1_월평균매출액 = 1; ELSE R1_월평균매출액=2;
IF 월영업이익 <=  . THEN R1_월영업이익=2; ELSE
```

```
IF 월영업이익 <=  957105 THEN R1_월영업이익=0; ELSE
IF 월영업이익 <= 5959233 THEN R1_월영업이익=1; ELSE R1_월영업이익=2;
IF 소유부동산금액 <= . THEN R1_소유부동산금액=1; ELSE
IF 소유부동산금액 <= 0 THEN R1_소유부동산금액=0; ELSE R1_소유부동산금액=1;
IF 임대보증금사업장 <= . THEN R1_임대보증금사업장=1; ELSE
IF 임대보증금사업장 <= 0 THEN R1_임대보증금사업장=0; ELSE R1_임대보증금사업장
=1;
IF 임대보증금주택 <= . THEN R1_임대보증금주택=1; ELSE
IF 임대보증금주택 <= 0 THEN R1_임대보증금주택=0; ELSE R1_임대보증금주택=1;
IF 재고자산 <= 0 THEN R1_재고자산=1; ELSE R1_재고자산=0;
IF 직권말소 <= 0 THEN R1_직권말소=0; ELSE R1_직권말소=1;
RUN;

DATA BIG.TR_DATA_COARSE;
SET BIG.TR_DATA_COARSE;
KEEP
고객번호
R1_고객형태
R1_업종
R1_주사업장소유여부
R1_주사업장임차보증금액
R1_주사업장월세금액
R1_실거주지소유여부
R1_실거주지임차보증금액
R1_실소유지월세금액
R1_차입금운전
R1_기보증잔액재단
R1_기보증잔액기보
R1_담보제외차입기관수
R1_현금서비스금액
R1_보유부동산
R1_업력
```

R1_거주기간
R1_월평균매출액
R1_월영업이익
R1_소유부동산금액
R1_임대보증금사업장
R1_임대보증금주택
R1_재고자산
R1_직권말소
사고여부;
RUN;

PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_고객형태 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_업종 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_주사업장소유여부 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_주사업장임차보증금액 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_주사업장월세금액 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_실거주지소유여부 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_실거주지임차보증금액 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_실소유지월세금액 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_차입금운전 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_기보증잔액재단 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_기보증잔액기보 ; VAR 사

```
고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_담보제외차입기관수 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_현금서비스금액 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_보유부동산 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_업력 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_거주기간 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_월평균매출액 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_월영업이익 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_소유부동산금액 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_임대보증금사업장 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_임대보증금주택 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_재고자산 ; VAR 사고여부; RUN;
PROC MEANS DATA=BIG.TR_DATA_COARSE; CLASS R1_직권말소 ; VAR 사고여부; RUN;

DATA BIG.TS_DATA_COARSE;
SET BIG.TS_DATA;
IF 고객형태 = ' ' THEN R1_고객형태=1; ELSE
IF 고객형태 = '개인사업자' THEN R1_고객형태=1; ELSE R1_고객형태=0;
IF 업종 = '1.제조업' THEN R1_업종=0; ELSE
IF 업종 = '2.서비스업' THEN R1_업종=1; ELSE
```

```
IF 업종 = '3.도소매업' THEN R1_업종=1; ELSE
IF 업종 = '4.음식숙박업' THEN R1_업종=0; ELSE
IF 업종 = '5.건설업' THEN R1_업종=0; ELSE
IF 업종 = '6.운수업' THEN R1_업종=1; ELSE R1_업종=2;
IF 주사업장소유여부 = ' ' THEN R1_주사업장소유여부=2; ELSE
IF 주사업장소유여부 = '가족소유' THEN R1_주사업장소유여부=2; ELSE
IF 주사업장소유여부 = '기타' THEN R1_주사업장소유여부=1; ELSE
IF 주사업장소유여부 = '무점포' THEN R1_주사업장소유여부=2; ELSE
IF 주사업장소유여부 = '임차' THEN R1_주사업장소유여부=0; ELSE
IF 주사업장소유여부 = '임차(보증금20백만원초과)' THEN R1_주사업장소유여부=1;
ELSE
IF 주사업장소유여부 = '자가' THEN R1_주사업장소유여부=2; ELSE R1_주사업장소유
여부=2;
IF 주사업장임차보증금액 <= 0 THEN R1_주사업장임차보증금액=1; ELSE
IF 0 < 주사업장임차보증금액 <= 1000 THEN R1_주사업장임차보증금액=0; ELSE
IF 1000 < 주사업장임차보증금액 THEN R1_주사업장임차보증금액=1;
IF 주사업장월세금액 <= 0 THEN R1_주사업장월세금액=1; ELSE R1_주사업장월세금액
=0;
IF 실거주지소유여부 = ' ' THEN R1_실거주지소유여부=1; ELSE
IF 실거주지소유여부 = '가족소유' THEN R1_실거주지소유여부=0; ELSE
IF 실거주지소유여부 = '기타' THEN R1_실거주지소유여부=0; ELSE
IF 실거주지소유여부 = '임차' THEN R1_실거주지소유여부=0; ELSE
IF 실거주지소유여부 = '임차(보증금20백만원초과)' THEN R1_실거주지소유여부=0;
ELSE
IF 실거주지소유여부 = '자가' THEN R1_실거주지소유여부=1; ELSE R1_실거주지소유
여부=1;
IF 실거주지임차보증금액 <= 0 THEN R1_실거주지임차보증금액=1; ELSE
IF 실거주지임차보증금액 <= 5000 THEN R1_실거주지임차보증금액=0; ELSE R1_실
거주지임차보증금액=1;
IF 실소유지월세금액 <= 0 THEN R1_실소유지월세금액=1; ELSE R1_실소유지월세금
액=0;
IF 차입금운전 <= 56.6 THEN R1_차입금운전=0; ELSE R1_차입금운전=1;
```

```
IF 기보증잔액재단 <= 6 THEN R1_기보증잔액재단=1; ELSE R1_기보증잔액재단=0;
IF 기보증잔액기보 <= 15.8  THEN R1_기보증잔액기보=1; ELSE R1_기보증잔액기보
=0;
IF 담보제외차입기관수 <= 1 THEN R1_담보제외차입기관수=2; ELSE
IF 담보제외차입기관수 <= 3 THEN R1_담보제외차입기관수=1; ELSE R1_담보제외차입
기관수=0;
IF 현금서비스금액  <= 0 THEN R1_현금서비스금액=1; ELSE R1_현금서비스금액=0;
IF 보유부동산 = ' '  THEN R1_보유부동산=2; ELSE
IF 보유부동산 = '다가구' THEN R1_보유부동산=2; ELSE
IF 보유부동산 = '다세대' THEN R1_보유부동산=1; ELSE
IF 보유부동산 = '단독주택' THEN R1_보유부동산=2; ELSE
IF 보유부동산 = '아파트' THEN R1_보유부동산=2; ELSE
IF 보유부동산 = '없음'  THEN R1_보유부동산=0; ELSE
IF 보유부동산 = '임야기타부동산' THEN R1_보유부동산=2; ELSE R1_보유부동산=1;
IF 업력 <= 0 THEN R1_업력=2; ELSE
IF 업력 <=  15 THEN R1_업력=1; ELSE
IF 업력 <=  44 THEN R1_업력=0; ELSE
IF 업력 <= 124 THEN R1_업력=1; ELSE R1_업력=2;
IF 거주기간 <= . THEN R1_거주기간 = 0; ELSE
IF 거주기간 <=  16 THEN R1_거주기간 = 1; ELSE
IF 거주기간 <=  67 THEN R1_거주기간 = 0; ELSE
IF 거주기간 <= 261 THEN R1_거주기간 = 1; ELSE R1_거주기간=0;
IF 월평균매출액 <= . THEN R1_월평균매출액 = 2; ELSE
IF 월평균매출액 <=  8500000 THEN R1_월평균매출액 = 0; ELSE
IF 월평균매출액 <= 38750000 THEN R1_월평균매출액 = 1; ELSE R1_월평균매출액
=2;
IF 월영업이익 <=  . THEN R1_월영업이익=2; ELSE
IF 월영업이익 <=  957105 THEN R1_월영업이익=0; ELSE
IF 월영업이익 <= 5959233 THEN R1_월영업이익=1; ELSE R1_월영업이익=2;
IF 소유부동산금액 <= . THEN R1_소유부동산금액=1; ELSE
IF 소유부동산금액 <= 0 THEN R1_소유부동산금액=0; ELSE R1_소유부동산금액=1;
IF 임대보증금사업장 <= . THEN R1_임대보증금사업장=1; ELSE
```

```
IF 임대보증금사업장 <= 0 THEN R1_임대보증금사업장=0; ELSE R1_임대보증금사업장
=1;
IF 임대보증금주택 <= . THEN R1_임대보증금주택=1; ELSE
IF 임대보증금주택 <= 0 THEN R1_임대보증금주택=0; ELSE R1_임대보증금주택=1;
IF 재고자산 <= 0 THEN R1_재고자산=1; ELSE R1_재고자산=0;
IF 직권말소 <= 0 THEN R1_직권말소=0; ELSE R1_직권말소=1;
RUN;

DATA BIG.TS_DATA_COARSE;
SET BIG.TS_DATA_COARSE;
KEEP
고객번호
R1_고객형태
R1_업종
R1_주사업장소유여부
R1_주사업장임차보증금액
R1_주사업장월세금액
R1_실거주지소유여부
R1_실거주지임차보증금액
R1_실소유지월세금액
R1_차입금운전
R1_기보증잔액재단
R1_기보증잔액기보
R1_담보제외차입기관수
R1_현금서비스금액
R1_보유부동산
R1_업력
R1_거주기간
R1_월평균매출액
R1_월영업이익
R1_소유부동산금액
R1_임대보증금사업장
```

```
R1_임대보증금주택
R1_재고자산
R1_직권말소
사고여부;
RUN;

* 3.4 단계적선택법에 의한 변수 선택;
PROC LOGISTIC DATA=BIG.TR_DATA_COARSE;
MODEL
사고여부 =
R1_고객형태 R1_업종 R1_주사업장소유여부 R1_주사업장임차보증금액
R1_주사업장월세금액  R1_실거주지소유여부 R1_실거주지임차보증금액
R1_실소유지월세금액  R1_차입금운전 R1_기보증잔액재단
R1_기보증잔액기보 R1_담보제외차입기관수 R1_현금서비스금액
R1_보유부동산  R1_업력 R1_거주기간
R1_월평균매출액 R1_월영업이익 R1_소유부동산금액
R1_임대보증금사업장  R1_임대보증금주택  R1_재고자산
R1_직권말소/SELECTION = STEPWISE;
RUN;

DATA BIG.TR_DATA_COARSE_01;
SET BIG.TR_DATA_COARSE;
KEEP
고객번호
사고여부
R1_고객형태 R1_업종 R1_주사업장임차보증금액
R1_실거주지소유여부 R1_실거주지임차보증금액
R1_실소유지월세금액 R1_차입금운전
R1_기보증잔액재단 R1_기보증잔액기보
R1_담보제외차입기관수 R1_현금서비스금액
R1_보유부동산 R1_업력
R1_거주기간 R1_월평균매출액
```

```
R1_재고자산 R1_직권말소;
RUN;

DATA BIG.TS_DATA_COARSE_01;
SET BIG.TS_DATA_COARSE;
KEEP
고객번호
사고여부
R1_고객형태 R1_업종 R1_주사업장임차보증금액
R1_실거주지소유여부 R1_실거주지임차보증금액
R1_실소유지월세금액 R1_차입금운전
R1_기보증잔액재단 R1_기보증잔액기보
R1_담보제외차입기관수 R1_현금서비스금액
R1_보유부동산 R1_업력
R1_거주기간 R1_월평균매출액
R1_재고자산 R1_직권말소;
RUN;

* 3.5 다중공선성 확인;
PROC CORR DATA = BIG.TR_DATA_COARSE_01 BEST = 5 SPEARMAN;
VAR
R1_고객형태 R1_업종 R1_주사업장임차보증금액
R1_실거주지소유여부 R1_실거주지임차보증금액
R1_실소유지월세금액 R1_차입금운전
R1_기보증잔액재단 R1_기보증잔액기보
R1_담보제외차입기관수 R1_현금서비스금액
R1_보유부동산 R1_업력
R1_거주기간 R1_월평균매출액
R1_재고자산 R1_직권말소;
RUN;

PROC LOGISTIC DATA=BIG.TR_DATA_COARSE; MODEL 사고여부 = R1_실거주지임차
```

```
보증금액; RUN;
PROC LOGISTIC DATA=BIG.TR_DATA_COARSE; MODEL 사고여부 = R1_실소유지월세
금액;  RUN;
DATA BIG.TR_DATA_COARSE_02;
SET BIG.TR_DATA_COARSE_01;
DROP R1_실거주지임차보증금액;
RUN;

DATA BIG.TS_DATA_COARSE_02;
SET BIG.TS_DATA_COARSE_01;
DROP R1_실거주지임차보증금액;
RUN;

PROC CONTENTS DATA=BIG.TR_DATA_COARSE_02 ORDER=VARNUM; RUN;

DATA BIG.TR_FIN;
SET BIG.TR_DATA_COARSE_02;
RENAME 고객번호 = ID ;
RENAME 사고여부 = Y ;
RENAME R1_고객형태 = X01 ;
RENAME R1_업종 = X02 ;
RENAME R1_주사업장임차보증금액 = X03 ;
RENAME R1_실거주지소유여부 = X04 ;
RENAME R1_실소유지월세금액 = X05 ;
RENAME R1_차입금운전 = X06 ;
RENAME R1_기보증잔액재단 = X07 ;
RENAME R1_기보증잔액기보 = X08 ;
RENAME R1_담보제외차입기관수 = X09 ;
RENAME R1_현금서비스금액 = X10 ;
RENAME R1_보유부동산 = X11 ;
RENAME R1_업력 = X12 ;
RENAME R1_거주기간 = X13 ;
```

```
RENAME R1_월평균매출액 = X14 ;
RENAME R1_재고자산 = X15 ;
RENAME R1_직권말소 = X16 ;
D='TR';
NO=_N_;
RUN;

DATA BIG.TS_FIN;
SET BIG.TS_DATA_COARSE_02;
RENAME 고객번호 = ID ;
RENAME 사고여부 = Y ;
RENAME R1_고객형태 = X01 ;
RENAME R1_업종 = X02 ;
RENAME R1_주사업장임차보증금액 = X03 ;
RENAME R1_실거주지소유여부 = X04 ;
RENAME R1_실소유지월세금액 = X05 ;
RENAME R1_차입금운전 = X06 ;
RENAME R1_기보증잔액재단 = X07 ;
RENAME R1_기보증잔액기보 = X08 ;
RENAME R1_담보제외차입기관수 = X09 ;
RENAME R1_현금서비스금액 = X10 ;
RENAME R1_보유부동산 = X11 ;
RENAME R1_업력 = X12 ;
RENAME R1_거주기간 = X13 ;
RENAME R1_월평균매출액 = X14 ;
RENAME R1_재고자산 = X15 ;
RENAME R1_직권말소 = X16 ;
D='TS';
NO=_N_;
RUN;

PROC SQL;
```

```
CREATE TABLE BIG.TR_FIN_1 AS
SELECT X01, X02, X03, X04, X05, X06, X07, X08, X09, X10, X11, X12, X13, X14, X15,
X16, Y
FROM BIG.TR_FIN;
QUIT; RUN;

PROC SQL;
CREATE TABLE BIG.TS_FIN_1 AS
SELECT X01, X02, X03, X04, X05, X06, X07, X08, X09, X10, X11, X12, X13, X14, X15,
X16, Y
FROM BIG.TS_FIN;
QUIT; RUN;

PROC EXPORT DATA= BIG.Tr_fin_1 OUTFILE= "C:₩Users₩dw₩Desktop₩tr_data.txt"
DBMS=TAB REPLACE; PUTNAMES=YES; RUN;
PROC EXPORT DATA= BIG.Ts_fin_1 OUTFILE= "C:₩Users₩dw₩Desktop₩ts_data.txt"
DBMS=TAB REPLACE; PUTNAMES=YES; RUN;

/*************************************************/
/* 4. 분석 결과 데이터셋 불러오기 */
/*************************************************/

* 4.1 의사결정나무;
***** 4.1.1 훈련용 자료 불러오기;
PROC IMPORT OUT= BIG.tree_tr1_1 DATAFILE= "D:₩@02.보고서₩@2019년₩01-02.
(분석보고서)빅데이터분석기법이용소평모형구축 ₩분석프로그램₩tree_tr1_1.txt"
DBMS=TAB REPLACE; GETNAMES=NO; DATAROW=2;
RUN;

***** 4.1.2 훈련용 자료 결과 불러오기;
PROC IMPORT OUT= BIG.tree_tr1_2 DATAFILE= "D:₩@02.보고서₩@2019년₩01-02.
(분석보고서)빅데이터분석기법이용소평모형구축 ₩분석프로그램₩tree_tr1_2.txt"
```

```
DBMS=TAB REPLACE; GETNAMES=NO; DATAROW=2;
RUN;

***** 4.1.3 평가용 자료 불러오기;
PROC IMPORT OUT= BIG.tree_ts1_1 DATAFILE= "D:₩@02.보고서₩@2019년₩01-02.
(분석보고서)빅데이터분석기법이용소평모형구축 ₩분석프로그램₩tree_ts1_1.txt"
DBMS=TAB REPLACE; GETNAMES=NO; DATAROW=2;
RUN;

***** 4.1.4 평가용 자료 결과 불러오기;
PROC IMPORT OUT= BIG.tree_ts1_2 DATAFILE= "D:₩@02.보고서₩@2019년₩01-02.
(분석보고서)빅데이터분석기법이용소평모형구축 ₩분석프로그램₩tree_ts1_2.txt"
DBMS=TAB REPLACE; GETNAMES=NO; DATAROW=2;
RUN;

***** 4.1.5 분석 결과 조합하기;
DATA WORK.TREE_TR1_1; SET BIG.TREE_TR1_1;  KEEP VAR1 VAR13; RUN;
DATA WORK.TREE_TR1_1; SET WORK.TREE_TR1_1; RENAME VAR13=Y; DT="TR";
RUN;

DATA WORK.TREE_TR1_2; SET BIG.TREE_TR1_2;  KEEP VAR1 VAR3; RUN;
DATA WORK.TREE_TR1_2; SET WORK.TREE_TR1_2; RENAME VAR3=TREE_P; DT="TR";
RUN;

DATA WORK.TREE_TS1_1; SET BIG.TREE_TS1_1;  KEEP VAR1 VAR13; RUN;
DATA WORK.TREE_TS1_1; SET WORK.TREE_TS1_1; RENAME VAR13=Y; DT="TS";
RUN;

DATA WORK.TREE_TS1_2; SET BIG.TREE_TS1_2;  KEEP VAR1 VAR3; RUN;
DATA WORK.TREE_TS1_2; SET WORK.TREE_TS1_2; RENAME VAR3=TREE_P; DT="TS";
RUN;
```

```
PROC SORT DATA=WORK.TREE_TR1_1; BY VAR1; RUN;
PROC SORT DATA=WORK.TREE_TR1_2; BY VAR1; RUN;
DATA WORK.TREE_TR; MERGE WORK.TREE_TR1_1 WORK.TREE_TR1_2; BY VAR1;
RUN;

PROC SORT DATA=WORK.TREE_TS1_1; BY VAR1; RUN;
PROC SORT DATA=WORK.TREE_TS1_2; BY VAR1; RUN;
DATA WORK.TREE_TS; MERGE WORK.TREE_TS1_1 WORK.TREE_TS1_2; BY VAR1; RUN;
DATA WORK.TREE_RES; SET WORK.TREE_TR WORK.TREE_TS; RUN;
DATA BIG.RES_TREE; SET WORK.TREE_RES; IF TREE_P<0.5 THEN TREE_Y=0; ELSE
TREE_Y=1; RUN;
* 4.2 로지스틱회귀;

***** 4.2.1 훈련용 자료 불러오기;
PROC IMPORT OUT= BIG.LOGIS_tr1_1 DATAFILE= "D:₩@02.보고서₩@2019년
₩01-02.(분석보고서)빅데이터분석기법이용소평모형구축₩분석프로그램
₩LOGIS_tr1_1.txt" DBMS=TAB REPLACE; GETNAMES=NO; DATAROW=2;
RUN;

***** 4.2.2 훈련용 자료 결과 불러오기;
PROC IMPORT OUT= BIG.LOGIS_tr1_2 DATAFILE= "D:₩@02.보고서₩@2019년
₩01-02.(분석보고서)빅데이터분석기법이용소평모형구축₩분석프로그램
₩LOGIS_tr1_2.txt" DBMS=TAB REPLACE; GETNAMES=NO; DATAROW=2;
RUN;

***** 4.2.3 평가용 자료 불러오기;
PROC IMPORT OUT= BIG.LOGIS_ts1_1 DATAFILE= "D:₩@02.보고서₩@2019년
₩01-02.(분석보고서)빅데이터분석기법이용소평모형구축₩분석프로그램
₩LOGIS_ts1_1.txt" DBMS=TAB REPLACE; GETNAMES=NO; DATAROW=2;
RUN;

***** 4.2.4 평가용 자료 결과 불러오기;
```

```
PROC IMPORT OUT= BIG.LOGIS_ts1_2 DATAFILE= "D:₩@02.보고서₩@2019년
₩01-02.(분석보고서)빅데이터분석기법이용소평모형구축₩분석프로그램
₩LOGIS_ts1_2.txt" DBMS=TAB REPLACE; GETNAMES=NO; DATAROW=2;
RUN;

***** 4.2.5 분석 결과 조합하기;
DATA WORK.LOGIS_TR1_1; SET BIG.LOGIS_TR1_1;  KEEP VAR1 VAR13; RUN;
DATA WORK.LOGIS_TR1_1; SET WORK.LOGIS_TR1_1; RENAME VAR13=Y; DT="TR";
RUN;

DATA WORK.LOGIS_TR1_2; SET BIG.LOGIS_TR1_2;  KEEP VAR1 VAR2; RUN;
DATA WORK.LOGIS_TR1_2; SET WORK.LOGIS_TR1_2; RENAME VAR2=LOGIS_P;
DT="TR"; RUN;

DATA WORK.LOGIS_TS1_1; SET BIG.LOGIS_TS1_1;  KEEP VAR1 VAR13; RUN;
DATA WORK.LOGIS_TS1_1; SET WORK.LOGIS_TS1_1; RENAME VAR13=Y; DT="TS";
RUN;

DATA WORK.LOGIS_TS1_2; SET BIG.LOGIS_TS1_2;  KEEP VAR1 VAR2; RUN;
DATA WORK.LOGIS_TS1_2; SET WORK.LOGIS_TS1_2; RENAME VAR2=LOGIS_P;
DT="TS"; RUN;

PROC SORT DATA=WORK.LOGIS_TR1_1; BY VAR1; RUN;
PROC SORT DATA=WORK.LOGIS_TR1_2; BY VAR1; RUN;
DATA WORK.LOGIS_TR; MERGE WORK.LOGIS_TR1_1 WORK.LOGIS_TR1_2; BY VAR1;
RUN;

PROC SORT DATA=WORK.LOGIS_TS1_1; BY VAR1; RUN;
PROC SORT DATA=WORK.LOGIS_TS1_2; BY VAR1; RUN;
DATA WORK.LOGIS_TS; MERGE WORK.LOGIS_TS1_1 WORK.LOGIS_TS1_2; BY VAR1;
RUN;
DATA WORK.LOGIS_RES; SET WORK.LOGIS_TR WORK.LOGIS_TS; RUN;
```

```
DATA BIG.RES_LOGIS; SET WORK.LOGIS_RES; IF LOGIS_P<0.5 THEN LOGIS_Y=0;
ELSE LOGIS_Y=1; RUN;

* 4.3 신경망;
***** 4.3.1 훈련용 자료 불러오기;
PROC IMPORT OUT= BIG.NET_tr1_1 DATAFILE= "D:₩@02.보고서₩@2019년₩01-02.
(분석보고서)빅데이터분석기법이용소평모형구축 ₩분석프로그램₩NET_tr1_1.txt"
DBMS=TAB REPLACE; GETNAMES=NO; DATAROW=2;
RUN;

***** 4.3.2 훈련용 자료 결과 불러오기;
PROC IMPORT OUT= BIG.NET_tr1_2 DATAFILE= "D:₩@02.보고서₩@2019년₩01-02.
(분석보고서)빅데이터분석기법이용소평모형구축 ₩분석프로그램₩NET_tr1_2.txt"
DBMS=TAB REPLACE; GETNAMES=NO; DATAROW=2;
RUN;

***** 4.3.3 평가용 자료 불러오기;
PROC IMPORT OUT= BIG.NET_ts1_1 DATAFILE= "D:₩@02.보고서₩@2019년₩01-02.
(분석보고서)빅데이터분석기법이용소평모형구축 ₩분석프로그램₩NET_ts1_1.txt"
DBMS=TAB REPLACE; GETNAMES=NO; DATAROW=2;
RUN;

***** 4.3.4 평가용 자료 결과 불러오기;
PROC IMPORT OUT= BIG.NET_ts1_2 DATAFILE= "D:₩@02.보고서₩@2019년₩01-02.
(분석보고서)빅데이터분석기법이용소평모형구축 ₩분석프로그램₩NET_ts1_2.txt"
DBMS=TAB REPLACE; GETNAMES=NO; DATAROW=2;
RUN;

***** 4.3.5 분석 결과 조합하기;
DATA WORK.NET_TR1_1; SET BIG.NET_TR1_1;  KEEP VAR1 VAR13; RUN;
DATA WORK.NET_TR1_1; SET WORK.NET_TR1_1; RENAME VAR13=Y; DT="TR"; RUN;
```

```
DATA WORK.NET_TR1_2; SET BIG.NET_TR1_2;  KEEP VAR1 VAR2; RUN;
DATA WORK.NET_TR1_2; SET WORK.NET_TR1_2; RENAME VAR2=NET_P; DT="TR";
RUN;

DATA WORK.NET_TS1_1; SET BIG.NET_TS1_1;  KEEP VAR1 VAR13; RUN;
DATA WORK.NET_TS1_1; SET WORK.NET_TS1_1; RENAME VAR13=Y; DT="TS"; RUN;

DATA WORK.NET_TS1_2; SET BIG.NET_TS1_2;  KEEP VAR1 VAR2; RUN;
DATA WORK.NET_TS1_2; SET WORK.NET_TS1_2; RENAME VAR2=NET_P; DT="TS";
RUN;

PROC SORT DATA=WORK.NET_TR1_1; BY VAR1; RUN;
PROC SORT DATA=WORK.NET_TR1_2; BY VAR1; RUN;
DATA WORK.NET_TR; MERGE WORK.NET_TR1_1 WORK.NET_TR1_2; BY VAR1; RUN;

PROC SORT DATA=WORK.NET_TS1_1; BY VAR1; RUN;
PROC SORT DATA=WORK.NET_TS1_2; BY VAR1; RUN;
DATA WORK.NET_TS; MERGE WORK.NET_TS1_1 WORK.NET_TS1_2; BY VAR1; RUN;
DATA WORK.NET_RES; SET WORK.NET_TR WORK.NET_TS; RUN;

DATA BIG.RES_NET; SET WORK.NET_RES; IF NET_P<0.5 THEN NET_Y=0; ELSE
NET_Y=1; RUN;

* 4.4 랜덤포레스트;
***** 4.4.1 훈련용 자료 불러오기;
PROC IMPORT OUT= BIG.FOR_tr1_1 DATAFILE= "D:₩@02.보고서₩@2019년₩01-02.
(분석보고서)빅데이터분석기법이용소평모형구축 ₩분석프로그램₩FOR_tr1_1.txt"
DBMS=TAB REPLACE; GETNAMES=NO; DATAROW=2;
RUN;

***** 4.4.2 훈련용 자료 결과 불러오기;
```

```
PROC IMPORT OUT= BIG.FOR_tr1_2 DATAFILE= "D:₩@02.보고서₩@2019년₩01-02.
(분석보고서)빅데이터분석기법이용소평모형구축 ₩분석프로그램₩FOR_tr1_2.txt"
DBMS=TAB REPLACE; GETNAMES=NO; DATAROW=2;
RUN;

***** 4.4.3 평가용 자료 불러오기;
PROC IMPORT OUT= BIG.FOR_ts1_1 DATAFILE= "D:₩@02.보고서₩@2019년₩01-02.
(분석보고서)빅데이터분석기법이용소평모형구축 ₩분석프로그램₩FOR_ts1_1.txt"
DBMS=TAB REPLACE; GETNAMES=NO; DATAROW=2;
RUN;

***** 4.4.4 평가용 자료 결과 불러오기;
PROC IMPORT OUT= BIG.FOR_ts1_2 DATAFILE= "D:₩@02.보고서₩@2019년₩01-02.
(분석보고서)빅데이터분석기법이용소평모형구축 ₩분석프로그램₩FOR_ts1_2.txt"
DBMS=TAB REPLACE; GETNAMES=NO; DATAROW=2;
RUN;

***** 4.4.5 분석 결과 조합하기;
DATA WORK.FOR_TR1_1; SET BIG.FOR_TR1_1;  KEEP VAR1 VAR13; RUN;
DATA WORK.FOR_TR1_1; SET WORK.FOR_TR1_1; RENAME VAR13=Y; DT="TR"; RUN;
DATA WORK.FOR_TR1_2; SET BIG.FOR_TR1_2;  KEEP VAR1 VAR2; RUN;
DATA WORK.FOR_TR1_2; SET WORK.FOR_TR1_2; RENAME VAR2=FOR_P; DT="TR";
RUN;
DATA WORK.FOR_TS1_1; SET BIG.FOR_TS1_1;  KEEP VAR1 VAR13; RUN;
DATA WORK.FOR_TS1_1; SET WORK.FOR_TS1_1; RENAME VAR13=Y; DT="TS"; RUN;

DATA WORK.FOR_TS1_2; SET BIG.FOR_TS1_2;  KEEP VAR1 VAR2; RUN;
DATA WORK.FOR_TS1_2; SET WORK.FOR_TS1_2; RENAME VAR2=FOR_P; DT="TS";
RUN;

PROC SORT DATA=WORK.FOR_TR1_1; BY VAR1; RUN;
PROC SORT DATA=WORK.FOR_TR1_2; BY VAR1; RUN;
```

```
DATA WORK.FOR_TR; MERGE WORK.FOR_TR1_1 WORK.FOR_TR1_2; BY VAR1; RUN;

PROC SORT DATA=WORK.FOR_TS1_1; BY VAR1; RUN;
PROC SORT DATA=WORK.FOR_TS1_2; BY VAR1; RUN;
DATA WORK.FOR_TS; MERGE WORK.FOR_TS1_1 WORK.FOR_TS1_2; BY VAR1; RUN;
DATA WORK.FOR_RES; SET WORK.FOR_TR WORK.FOR_TS; RUN;

DATA BIG.RES_FOR; SET WORK.FOR_RES; IF FOR_P<0.5 THEN FOR_Y=0; ELSE
FOR_Y=1; RUN;

* 4.5 서포트벡터머신;

***** 4.5.1 훈련용 자료 불러오기;
PROC IMPORT OUT= BIG.svm_tr1_1 DATAFILE= "D:₩@02.보고서₩@2019년₩01-02.
(분석보고서)빅데이터분석기법이용소평모형구축 ₩분석프로그램₩svm_tr1_1.txt"
DBMS=TAB REPLACE; GETNAMES=NO; DATAROW=2;
RUN;

***** 4.5.2 훈련용 자료 결과 불러오기;
PROC IMPORT OUT= BIG.svm_tr1_2 DATAFILE= "D:₩@02.보고서₩@2019년₩01-02.
(분석보고서)빅데이터분석기법이용소평모형구축 ₩분석프로그램₩svm_tr1_2.txt"
DBMS=TAB REPLACE;
  GETNAMES=NO; DATAROW=2;
RUN;

***** 4.5.3 평가용 자료 불러오기;
PROC IMPORT OUT= BIG.svm_ts1_1 DATAFILE= "D:₩@02.보고서₩@2019년₩01-02.
(분석보고서)빅데이터분석기법이용소평모형구축 ₩분석프로그램₩svm_ts1_1.txt"
DBMS=TAB REPLACE;
  GETNAMES=NO; DATAROW=2;
RUN;
```

```
***** 4.5.4 평가용 자료 결과 불러오기;
PROC IMPORT OUT= BIG.svm_ts1_2 DATAFILE= "D:₩@02.보고서₩@2019년₩01-02.
(분석보고서)빅데이터분석기법이용소평모형구축 ₩분석프로그램₩svm_ts1_2.txt"
DBMS=TAB REPLACE;
  GETNAMES=NO; DATAROW=2;
RUN;

***** 4.5.5 분석 결과 조합하기;
DATA WORK.svm_TR1_1; SET BIG.svm_TR1_1;  KEEP VAR1 VAR13; RUN;
DATA WORK.svm_TR1_1; SET WORK.svm_TR1_1; RENAME VAR13=Y; DT="TR"; RUN;
DATA WORK.svm_TR1_2; SET BIG.svm_TR1_2;  KEEP VAR3; RUN;
DATA WORK.svm_TR1_2; SET WORK.svm_TR1_2; RENAME VAR3=svm_P; DT="TR";
RUN;

DATA WORK.svm_TS1_1; SET BIG.svm_TS1_1;  KEEP VAR1 VAR13; RUN;
DATA WORK.svm_TS1_1; SET WORK.svm_TS1_1; RENAME VAR13=Y; DT="TS"; RUN;
DATA WORK.svm_TS1_2; SET BIG.svm_TS1_2;  KEEP VAR3; RUN;
DATA WORK.svm_TS1_2; SET WORK.svm_TS1_2; RENAME VAR3=svm_P; DT="TS";
RUN;

DATA WORK.svm_TR; MERGE WORK.svm_TR1_1 WORK.svm_TR1_2; RUN;
PROC SORT DATA=WORK.svm_TR; BY VAR1; RUN;

DATA WORK.svm_TS; MERGE WORK.svm_TS1_1 WORK.svm_TS1_2; RUN;
PROC SORT DATA=WORK.svm_TS; BY VAR1; RUN;

DATA WORK.svm_RES; SET WORK.svm_TR WORK.svm_TS; RUN;

DATA BIG.RES_svm; SET WORK.svm_RES; IF svm_P<0.5 THEN svm_Y=0; ELSE
svm_Y=1; RUN;

* 4.6 결과 합치기;
```

```
PROC SORT DATA=BIG.RES_TREE; BY VAR1; RUN;
PROC SORT DATA=BIG.RES_LOGIS; BY VAR1; RUN;
PROC SORT DATA=BIG.RES_NET; BY VAR1; RUN;
PROC SORT DATA=BIG.RES_FOR; BY VAR1; RUN;
PROC SORT DATA=BIG.RES_SVM; BY VAR1; RUN;

DATA BIG.RES_ALL;
MERGE BIG.RES_TREE BIG.RES_LOGIS BIG.RES_NET BIG.RES_FOR BIG.RES_SVM;
BY VAR1;
RUN;

/*****************************************************************************/
/* 5. 결과 분석 */
/*****************************************************************************/

/* 오분류표 */
PROC FREQ DATA=BIG.RES_ALL; TABLES Y*TREE_Y/NOROW NOCOL NOPERCENT;
WHERE DT="TR"; RUN;
PROC FREQ DATA=BIG.RES_ALL; TABLES Y*TREE_Y/NOROW NOCOL NOPERCENT;
WHERE DT="TS"; RUN;

PROC FREQ DATA=BIG.RES_ALL; TABLES Y*LOGIS_Y/NOROW NOCOL NOPERCENT;
WHERE DT="TR"; RUN;
PROC FREQ DATA=BIG.RES_ALL; TABLES Y*LOGIS_Y/NOROW NOCOL NOPERCENT;
WHERE DT="TS"; RUN;

PROC FREQ DATA=BIG.RES_ALL; TABLES Y*NET_Y/NOROW NOCOL NOPERCENT;
WHERE DT="TR"; RUN;
PROC FREQ DATA=BIG.RES_ALL; TABLES Y*NET_Y/NOROW NOCOL NOPERCENT;
WHERE DT="TS"; RUN;

PROC FREQ DATA=BIG.RES_ALL; TABLES Y*FOR_Y/NOROW NOCOL NOPERCENT;
```

```
WHERE DT="TR"; RUN;
PROC FREQ DATA=BIG.RES_ALL; TABLES Y*FOR_Y/NOROW NOCOL NOPERCENT;
WHERE DT="TS"; RUN;

/* 의사결정나무 반응률 */
PROC SORT DATA=BIG.RES_ALL OUT=WORK.TREE_SORT(KEEP=VAR1 Y DT TREE_P);
BY TREE_P; RUN;
DATA WORK.TREE_SORT_TR WORK.TREE_SORT_TS;
SET WORK.TREE_SORT;
IF DT='TR' THEN OUTPUT WORK.TREE_SORT_TR;
IF DT='TS' THEN OUTPUT WORK.TREE_SORT_TS;
RUN;

DATA WORK.TREE_SORT_TR; SET WORK.TREE_SORT_TR;NUM=_N_; RUN;
DATA WORK.TREE_SORT_TS; SET WORK.TREE_SORT_TS;NUM=_N_; RUN;

DATA WORK.TREE_SORT_TR; SET WORK.TREE_SORT_TR; P=NUM/281*100; RUN;
DATA WORK.TREE_SORT_TS; SET WORK.TREE_SORT_TS; P=NUM/120*100; RUN;

DATA WORK.TREE_SORT_TR;
SET WORK.TREE_SORT_TR;
IF P < 10 THEN G1='01'; ELSE
IF P < 20 THEN G1='02'; ELSE
IF P < 30 THEN G1='03'; ELSE
IF P < 40 THEN G1='04'; ELSE
IF P < 50 THEN G1='05'; ELSE
IF P < 60 THEN G1='06'; ELSE
IF P < 70 THEN G1='07'; ELSE
IF P < 80 THEN G1='08'; ELSE
IF P < 90 THEN G1='09'; ELSE G1 = '10';
RUN;
```

```
DATA WORK.TREE_SORT_TS;
SET WORK.TREE_SORT_TS;
IF P < 10 THEN G1='01'; ELSE
IF P < 20 THEN G1='02'; ELSE
IF P < 30 THEN G1='03'; ELSE
IF P < 40 THEN G1='04'; ELSE
IF P < 50 THEN G1='05'; ELSE
IF P < 60 THEN G1='06'; ELSE
IF P < 70 THEN G1='07'; ELSE
IF P < 80 THEN G1='08'; ELSE
IF P < 90 THEN G1='09'; ELSE G1 = '10';
RUN;

PROC MEANS DATA=WORK.TREE_SORT_TR; CLASS G1; VAR TREE_P; RUN;
PROC MEANS DATA=WORK.TREE_SORT_TS; CLASS G1; VAR TREE_P; RUN;

PROC MEANS DATA=WORK.TREE_SORT_TR; CLASS G1; VAR Y; RUN;
PROC MEANS DATA=WORK.TREE_SORT_TS; CLASS G1; VAR Y; RUN;

/* 로지스틱 반응률 */
PROC SORT DATA=BIG.RES_ALL OUT=WORK.LOGIS_SORT(KEEP=VAR1 Y DT
LOGIS_P); BY LOGIS_P; RUN;
DATA WORK.LOGIS_SORT_TR WORK.LOGIS_SORT_TS;
SET WORK.LOGIS_SORT;
IF DT='TR' THEN OUTPUT WORK.LOGIS_SORT_TR;
IF DT='TS' THEN OUTPUT WORK.LOGIS_SORT_TS;
RUN;

DATA WORK.LOGIS_SORT_TR; SET WORK.LOGIS_SORT_TR; NUM=_N_; RUN;
DATA WORK.LOGIS_SORT_TS; SET WORK.LOGIS_SORT_TS; NUM=_N_; RUN;

DATA WORK.LOGIS_SORT_TR; SET WORK.LOGIS_SORT_TR; P=NUM/281*100; RUN;
```

```
DATA WORK.LOGIS_SORT_TS; SET WORK.LOGIS_SORT_TS; P=NUM/120*100; RUN;
DATA WORK.LOGIS_SORT_TR;
SET WORK.LOGIS_SORT_TR;
IF P < 10 THEN G1='01'; ELSE
IF P < 20 THEN G1='02'; ELSE
IF P < 30 THEN G1='03'; ELSE
IF P < 40 THEN G1='04'; ELSE
IF P < 50 THEN G1='05'; ELSE
IF P < 60 THEN G1='06'; ELSE
IF P < 70 THEN G1='07'; ELSE
IF P < 80 THEN G1='08'; ELSE
IF P < 90 THEN G1='09'; ELSE G1 = '10';
RUN;

DATA WORK.LOGIS_SORT_TS;
SET WORK.LOGIS_SORT_TS;
IF P < 10 THEN G1='01'; ELSE
IF P < 20 THEN G1='02'; ELSE
IF P < 30 THEN G1='03'; ELSE
IF P < 40 THEN G1='04'; ELSE
IF P < 50 THEN G1='05'; ELSE
IF P < 60 THEN G1='06'; ELSE
IF P < 70 THEN G1='07'; ELSE
IF P < 80 THEN G1='08'; ELSE
IF P < 90 THEN G1='09'; ELSE G1 = '10';
RUN;

PROC MEANS DATA=WORK.LOGIS_SORT_TR; CLASS G1; VAR LOGIS_P; RUN;
PROC MEANS DATA=WORK.LOGIS_SORT_TS; CLASS G1; VAR LOGIS_P; RUN;

PROC MEANS DATA=WORK.LOGIS_SORT_TR; CLASS G1; VAR Y; RUN;
PROC MEANS DATA=WORK.LOGIS_SORT_TS; CLASS G1; VAR Y; RUN;
```

```
/* 신경망 반응률 */
PROC SORT DATA=BIG.RES_ALL OUT=WORK.NET_SORT(KEEP=VAR1 Y DT NET_P);
BY NET_P; RUN;
DATA WORK.NET_SORT_TR WORK.NET_SORT_TS;
SET WORK.NET_SORT;
IF DT='TR' THEN OUTPUT WORK.NET_SORT_TR;
IF DT='TS' THEN OUTPUT WORK.NET_SORT_TS;
RUN;

DATA WORK.NET_SORT_TR; SET WORK.NET_SORT_TR; NUM=_N_; RUN;
DATA WORK.NET_SORT_TS; SET WORK.NET_SORT_TS; NUM=_N_; RUN;

DATA WORK.NET_SORT_TR; SET WORK.NET_SORT_TR; P=NUM/281*100; RUN;
DATA WORK.NET_SORT_TS; SET WORK.NET_SORT_TS; P=NUM/120*100; RUN;

DATA WORK.NET_SORT_TR;
SET WORK.NET_SORT_TR;
IF P < 10 THEN G1='01'; ELSE
IF P < 20 THEN G1='02'; ELSE
IF P < 30 THEN G1='03'; ELSE
IF P < 40 THEN G1='04'; ELSE
IF P < 50 THEN G1='05'; ELSE
IF P < 60 THEN G1='06'; ELSE
IF P < 70 THEN G1='07'; ELSE
IF P < 80 THEN G1='08'; ELSE
IF P < 90 THEN G1='09'; ELSE G1 = '10';
RUN;

DATA WORK.NET_SORT_TS;
SET WORK.NET_SORT_TS;
IF P < 10 THEN G1='01'; ELSE
```

```
IF P < 20 THEN G1='02'; ELSE
IF P < 30 THEN G1='03'; ELSE
IF P < 40 THEN G1='04'; ELSE
IF P < 50 THEN G1='05'; ELSE
IF P < 60 THEN G1='06'; ELSE
IF P < 70 THEN G1='07'; ELSE
IF P < 80 THEN G1='08'; ELSE
IF P < 90 THEN G1='09'; ELSE G1 = '10';
RUN;
PROC MEANS DATA=WORK.NET_SORT_TR; CLASS G1; VAR NET_P; RUN;
PROC MEANS DATA=WORK.NET_SORT_TS; CLASS G1; VAR NET_P; RUN;

PROC MEANS DATA=WORK.NET_SORT_TR; CLASS G1; VAR Y; RUN;
PROC MEANS DATA=WORK.NET_SORT_TS; CLASS G1; VAR Y; RUN;

/* 랜덤포레스트 반응률 */
PROC SORT DATA=BIG.RES_ALL OUT=WORK.FOR_SORT(KEEP=VAR1 Y DT FOR_P);
BY FOR_P; RUN;
DATA WORK.FOR_SORT_TR WORK.FOR_SORT_TS;
SET WORK.FOR_SORT;
IF DT='TR' THEN OUTPUT WORK.FOR_SORT_TR;
IF DT='TS' THEN OUTPUT WORK.FOR_SORT_TS;
RUN;

DATA WORK.FOR_SORT_TR; SET WORK.FOR_SORT_TR; NUM=_N_; RUN;
DATA WORK.FOR_SORT_TS; SET WORK.FOR_SORT_TS; NUM=_N_; RUN;

DATA WORK.FOR_SORT_TR; SET WORK.FOR_SORT_TR; P=NUM/281*100; RUN;
DATA WORK.FOR_SORT_TS; SET WORK.FOR_SORT_TS; P=NUM/120*100; RUN;

DATA WORK.FOR_SORT_TR;
SET WORK.FOR_SORT_TR;
```

```
IF P < 10 THEN G1='01'; ELSE
IF P < 20 THEN G1='02'; ELSE
IF P < 30 THEN G1='03'; ELSE
IF P < 40 THEN G1='04'; ELSE
IF P < 50 THEN G1='05'; ELSE
IF P < 60 THEN G1='06'; ELSE
IF P < 70 THEN G1='07'; ELSE
IF P < 80 THEN G1='08'; ELSE
IF P < 90 THEN G1='09'; ELSE G1 = '10';
RUN;
DATA WORK.FOR_SORT_TS;
SET WORK.FOR_SORT_TS;
IF P < 10 THEN G1='01'; ELSE
IF P < 20 THEN G1='02'; ELSE
IF P < 30 THEN G1='03'; ELSE
IF P < 40 THEN G1='04'; ELSE
IF P < 50 THEN G1='05'; ELSE
IF P < 60 THEN G1='06'; ELSE
IF P < 70 THEN G1='07'; ELSE
IF P < 80 THEN G1='08'; ELSE
IF P < 90 THEN G1='09'; ELSE G1 = '10';
RUN;

PROC MEANS DATA=WORK.FOR_SORT_TR; CLASS G1; VAR FOR_P; RUN;
PROC MEANS DATA=WORK.FOR_SORT_TS; CLASS G1; VAR FOR_P; RUN;
PROC MEANS DATA=WORK.FOR_SORT_TR; CLASS G1; VAR Y; RUN;
PROC MEANS DATA=WORK.FOR_SORT_TS; CLASS G1; VAR Y; RUN;

/* SVM 반응률 */

PROC SORT DATA=BIG.RES_ALL OUT=WORK.SVM_SORT(KEEP=VAR1 Y DT SVM_P);
BY SVM_P; RUN;
```

```
DATA WORK.SVM_SORT_TR WORK.SVM_SORT_TS;
SET WORK.SVM_SORT;
IF DT='TR' THEN OUTPUT WORK.SVM_SORT_TR;
IF DT='TS' THEN OUTPUT WORK.SVM_SORT_TS;
RUN;

DATA WORK.SVM_SORT_TR; SET WORK.SVM_SORT_TR; NUM=_N_; RUN;
DATA WORK.SVM_SORT_TS; SET WORK.SVM_SORT_TS; NUM=_N_; RUN;

DATA WORK.SVM_SORT_TR; SET WORK.SVM_SORT_TR; P=NUM/281*100; RUN;
DATA WORK.SVM_SORT_TS; SET WORK.SVM_SORT_TS; P=NUM/120*100; RUN;

DATA WORK.SVM_SORT_TR;
SET WORK.SVM_SORT_TR;
IF P < 10 THEN G1='01'; ELSE
IF P < 20 THEN G1='02'; ELSE
IF P < 30 THEN G1='03'; ELSE
IF P < 40 THEN G1='04'; ELSE
IF P < 50 THEN G1='05'; ELSE
IF P < 60 THEN G1='06'; ELSE
IF P < 70 THEN G1='07'; ELSE
IF P < 80 THEN G1='08'; ELSE
IF P < 90 THEN G1='09'; ELSE G1 = '10';
RUN;

DATA WORK.SVM_SORT_TS;
SET WORK.SVM_SORT_TS;
IF P < 10 THEN G1='01'; ELSE
IF P < 20 THEN G1='02'; ELSE
IF P < 30 THEN G1='03'; ELSE
IF P < 40 THEN G1='04'; ELSE
IF P < 50 THEN G1='05'; ELSE
```

```
IF P < 60 THEN G1='06'; ELSE
IF P < 70 THEN G1='07'; ELSE
IF P < 80 THEN G1='08'; ELSE
IF P < 90 THEN G1='09'; ELSE G1 = '10';
RUN;

PROC MEANS DATA=WORK.SVM_SORT_TR; CLASS G1; VAR SVM_P; RUN;
PROC MEANS DATA=WORK.SVM_SORT_TS; CLASS G1; VAR SVM_P; RUN;
PROC MEANS DATA=WORK.SVM_SORT_TR; CLASS G1; VAR Y; RUN;
PROC MEANS DATA=WORK.SVM_SORT_TS; CLASS G1; VAR Y; RUN;

/*****************************************************************************/
/* 6. 최종 로지스틱 회귀모형 구축              */
/*****************************************************************************/
PROC LOGISTIC DATA = BIG.TR_DATA_COARSE_02
                OUTMODEL = BIG._LOGIT_MODEL_FIN;
CLASS
R1_고객형태 R1_업종 R1_주사업장임차보증금액 R1_실거주지소유여부
R1_실소유지월세금액 R1_차입금운전 R1_기보증잔액재단 R1_기보증잔액기보
R1_담보제외차입기관수 R1_현금서비스금액 R1_보유부동산 R1_업력
R1_거주기간 R1_월평균매출액 R1_재고자산 R1_직권말소
/PARAM = REFERENCE REF = FIRST;
MODEL 사고여부 =
R1_고객형태 R1_업종 R1_주사업장임차보증금액 R1_실거주지소유여부
R1_실소유지월세금액 R1_차입금운전 R1_기보증잔액재단 R1_기보증잔액기보
R1_담보제외차입기관수 R1_현금서비스금액 R1_보유부동산 R1_업력
R1_거주기간 R1_월평균매출액 R1_재고자산 R1_직권말소 / OUTROC =
BIG._LOGIT_RES_ROC_1 LACKFIT;
OUTPUT OUT = BIG._LOGIT_RES_1_1 P=PHAT;
RUN;

********** 사후확률에 대한 FINE CLASSING ;
```

```
%include  "D:₩@02.보고서₩@2019년₩01-02.(분석보고서)빅데이터분석기법이용소평모
형구축₩분석프로그램₩01기본Macro_Scoring_Macro_TwoTarget_pdo40_20120524.sas";

%include  "D:₩@02.보고서₩@2019년₩01-02.(분석보고서)빅데이터분석기법이용소평모
형구축₩분석프로그램₩02 기본Macro_FineClassing_Macro(V).sas";

DATA BIG._LOGIT_RES_1_1; SET BIG._LOGIT_RES_1_1; WW=1; RUN;

%FINECLASS_V(BIG._LOGIT_RES_1_1,사고여부,사고여부,
                WW,PHAT,PHAT,10,0,BIG._LOGIT_FINECLASS,_LOGIT_FINECLASS,1);
```

참고 문헌

강신형 (2016). Alternative Data 기계학습을 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, pp 63-69.

김성진, 안현철 (2016). 기업 신용등급 예측을 위한 랜덤포레스트의 응용, 산업혁신연구, 제32권 1호, pp 187-211.

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

김승혁, 김종우 (2007). Modified Bagging Predictors를 이용한 SOHO 부도예측, 지능정보연구, 13(2), pp 15-26.

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

나종화 (2017). R 데이터마이닝, 자유아카데미.

김효진 (2018). 머신러닝에 대한 이해, 주택금융리서치.

박정윤 (2000). 재무정책과 기업부실 예측, 재무관리논총, pp 93-116.

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

박주완 (2018), 소상공인 신용평가모형 구축에 관한 연구-설문조사 자료를 이용하여- , 제350호, 중소기업금융연구.

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

박주완, 송창길, 배진성 (2017). 기계학습 기법을 이용한 소상공인 신용평가 모형 구축에 관한 연구, 제10권, 3호, 한국비즈니스리뷰.

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

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

오미애, 최현수, 김수현, 장준혁, 진재현, 천미경 (2017). 기계학습(Machine Learning)기반 사회보장 빅데이터 분석 및 예측모형 연구, 한국보건사회연구원

유원종, 이철규 (2013). 비재무적 요인이 중소벤처기업의 신용평가에 미치는 영향, 대한경영학회지, 제28권 제12호, pp 3191-3210.

윤상용, 강만수, 이형탁 (2016). 소상공인 신용평가에서 비재무적 정보는 중요한가, 경영컨설팅연구 , 제16권 , pp 37-46.

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

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

이명식, 김정인 (2007). 개인신용평점제도, 서울출판미디어.

이승현(역) (2014). 데이터 마이닝, 에이콘.

이영섭 역 (2003). 데이터마이닝 Cookbook, 교우사.

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

이주민, 김승연, 하은호, 노태협 (2007). AHP 모형을 활용한 소상공인 신용평

가시스템 구축, 정보시스템연구, 16(3), pp 109-132.

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

전성빈, 김영일 (2001). 도산 예측 모형의 예측력 검증, 회계저널, pp 151-182.

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

조재희, 조성배, 이성임, 신현정, 김성범 역 (2019). 비즈니스 애널리틱스를 위한 데이터마이닝 R edition, 이앤비플러스.

조준희, 강부식 (2007). 코스닥기업의 도산예측모형에 관한 연구, 산업경제연구, 제20권 제1호.

최종후, 진서훈 (2005). 데이터마이닝의 현장, 자유아카데미.

Altman, E. I., Sabato, G., & Wilson, N. (2010). "The value of non-financial information in small and medium-sized enterprise risk management," The Journal of Credit Risk, 6(2), pp 95-127.

Bhimani, A., Gulamhussen, M. A., & Lopes, S. R. (2013). The role of financial, macro- economic, and non-financial information in bank loan default timing prediction, European Accounting Review, 22(4), pp 739-763.

Breiman, L. (2001). Random Forests. Machine Learning, Vol. 45, No. 1, pp 5-32.

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, pp 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.

B. Lantz (2015). Machine learning with R second edition, O'reilly.

Leung, K., Cheong, F., Cheong, C., O'Farrell, S. Tissington, R. (2008). Building a scorecard in practice, Proceedings of the 7th International Conference on Computational Intelligence in Economics and Finance (CIEF 2008).

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

Tan, P. N., Steinbach, M. and Kumar, V. (2006). Introduction to Data Mining. Pearson.

Yoo, J.E. (2015). Random forests, an alternative data mining technique to decision tree. Journal of Educational Evaluation, Vol. 28, No. 2, pp. 427-448.

저자 약력

• 박 주 완
- 통계학 석사, 경영학 석사, 통계학 박사
- 동국대학교 통계학과 시간강사
- 세종대학교 통계학과 시간강사
- 한양대학교 의과대학 역학연구소 연구원
- 한국직업능력개발원 동향데이터분석센터 연구원
- 한국지식재산연구원 동향분석센터 전문위원
- 국민연금연구원 재정추계분석실 부연구위원
- 現 신용보증재단중앙회 교육연구부 선임연구위원

• 배 진 성
- 경제학 석사, 경제학 박사
- 전남대학교 경제학과 시간강사
- 전남대학교 BK21플러스 글로벌창의경제인력양성팀 박사후연구원
- 現 신용보증재단중앙회 교육연구부 선임연구위원

• 윤 혁 준
- 경제학 석사, 경제학 박사 과정
- 한국직업능력개발원 진로교육센터 연구원
- 現 신용보증재단중앙회 교육연구부 연구원

연구보고서 2019-02

빅데이터 분석 기법을 이용한 소상공인 신용평가모형 구축 연구

2019년 11월 인쇄
2019년 12월 발행
발행인 김병근
발행처 신용보증재단중앙회
대전광역시 서구 한밭대로 713, 12~15층(나라키움대전센터)
TEL 1588-7365
FAX 042-715-5124

ISBN 979-11-954584-8-6

빅데이터 분석 기법을 이용한 소상공인 신용평가모형 구축 연구

KOREG 신용보증재단중앙회

비매품/무료
93320
9 791195 458486
ISBN 979-11-954584-8-6