ISSN 2288-5536

KOREG RESEARCH

제8권 1호 (통권8호)

KOREG
신용보증재단중앙회

[차 례]

연구논문

슈퍼마켓의 점포성과에 미치는 영향요인에 관한 연구

김정일*

본 연구는 한국의 슈퍼마켓 업태에서 경영성과에 미치는 영향요인을 내적요인과 외적요인으로 분류하고, 각 요인이 경영성과에 미치는 영향을 검증하는데 목적을 두고 있다. 연구 결과 슈퍼마켓의 외적요인(상권요인)에서는 '상권내 인구수', '소비자 소득수준', '경쟁점 가격' 등이, 내적요인(역량요인)에서는 '매장면적', 상품구색', '종업원 역량' 등이 슈퍼마켓의 경영성과에 영향을 미치는 것으로 분석되었다.

Keywords: *슈퍼마켓, 입지, 소비자, 경쟁자, 하드웨어, 소프트웨어, 휴먼웨어, 경영성과*

* 신용보증재단중앙회 경영본부 전무이사(경영학박사)

Ⅰ. 서 론

슈퍼마켓을 경영하는 데 있어서 외부요인과 내부요인 중 무엇이 점포성과에 더 크게 영향을 미칠까? 점포성과에 영향을 미치는 외적요인, 즉 입지, 소비자, 경쟁자 중에서 어떤 변수들이 영향을 더 미칠까? 소매점의 내부요인인 하드웨어, 소프트웨어, 휴먼웨어 측면에서는 어떤 측면이 점포성과에 영향을 더 미칠까? 라는 관점에서 연구가 시작되었다.

따라서 본 연구의 목적은 한국 슈퍼마켓의 외적•내적 요인의 다양한 독립변수가 점포성과에 미치는 영향관계의 메커니즘(mechanism)을 규명하기 위한 것이다. 구체적인 변수로는 외적 측면에서 입지, 소비자, 경쟁자 요인과, 내적 측면에서 하드웨어, 소프트웨어, 휴먼웨어 요인 등 크게 6개의 요인(factors)으로 분류하고, 이들 개별 요인에 속하는 총 48개의 세부 요인을 독립변수로 설정하여 성과에 미치는 영향을 분석하였다. 조사대상 표본 점포는 전국의 나들가게 회원 슈퍼마켓 중 무작위 표본추출 과정을 거쳐 총 105개를 선정하였다.

Ⅱ. 이론적 배경

Hortman 외(1990)는 소비자의 사회경제적 세분화(Demographic & Socioeconomic Segmentation)를 연령, 소득, 인종별로 소비자 집단을 구분하였다. 또한 Darden(1980)은 그의 소비자 행동모델 연구에서 소비자 가치, 라이프스타일, 사회계층, 가족생활주기 단계 등의 외생적 소비자 속성변수와 도구적 가치, 쇼핑습관, 쇼핑지향성과 같은 내재적 소비자 속성변수를 기초적인 독립변수로 사용하였다.

Arnold 외(1983)는 1974년부터 1981년까지 미국, 캐나다, 영국, 네덜란드에 위치한 176,447개의 슈퍼마켓을 대상으로 소비자의 슈퍼마켓 이용행동에 대한 조사를 통해 소비자가 슈퍼마켓을 선택할 때에 가장 중요하게 생각하는 것은 입지특성이라는 점을 밝혔다. 또한 Huff(1963)은 접근성이 좋은 점포는 소비자에

의해서 선호도가 높은 경향을 갖는다고 하였다.

Mason과 Mayer(1981)는 경쟁요인을 분석하기 위해서는 경쟁점포에 대한 자료가 필요하다고 하였으며, 즉 경쟁점포의 수, 매장면적, 지점과의 거리, 상품구색, 가격수준, 브랜드 인지도 및 선호도 등의 자료가 여기에 해당된다. 경쟁점포의 수, 매장면적, 지점과의 거리는 공간적 경쟁(Spatial Competition)과 관련된다고 하였다.

Bates(1979)는 슈퍼마켓 업태에서 상품배치는 노출가치를 극대화하고, 집기비품, 곤돌라, 통로 배치 등에서 고객의 체류 시간(Shopping Time)을 길게 할수록 고객의 구매단가가 높아진다는 사실을 발견하였다. 서상윤과 차제빈(2013)은 매장의 핵심은 상품이며 머천다이징은 상품화를 의미하는 데 상품을 보다 매력적이고 노출적(visible)으로 만들어 구매를 자극하기 위한 것으로 점내배치 면적배분과 선반 상품화, 상품진열을 포괄하는 개념이라고 정의하였다.

Berman과 Evans(1976)는 점포분위기 요소로 외장, 인테리어, 점포배치, 내부진열 등 4가지 차원으로 구분하였고, Kotler(1994)는 점포분위기를 소비자의 감각기관에 따라 4가지 차원(시각적, 청각적, 후각적, 촉각적)으로 분류하였다.

Cathy 외(1994)는 슈퍼마켓에서 매장면적, 점포 입지, 고객의 구매 패턴에 맞는 최적의 점포 레이아웃이 적용되면, 합리적인 상품 배치와 식품·비식품 연관성 제고 등의 성과를 극대화할 수 있다고 주장하였다. Berman과 Evans(1976)는 소매 마케팅을 소매점이 목표고객을 대상으로 고지, 설득 또는 상기시키기 위해 수행하는 정보제공 활동으로 규정하고, 전통적인 마케팅에서처럼 소매촉진 역시 광고, 홍보, 판매촉진, 인적판매로 이루어진다고 하였다.

Mason과 Mayer(1981)에 의하면 판매원의 성과에 영향을 미치는 요인과 직무만족과의 관계를 보면 동기유발(motivation), 숙련도자질(aptitude), 개인적·조직적·환경적 변수(personal organizational, and environmental variables), 역할지각(role perception) 등이 상호 작용하여 영향을 미치게 되며, 이는 보수, 직무만족 수준과 관련되어 있음을 분석하였다.

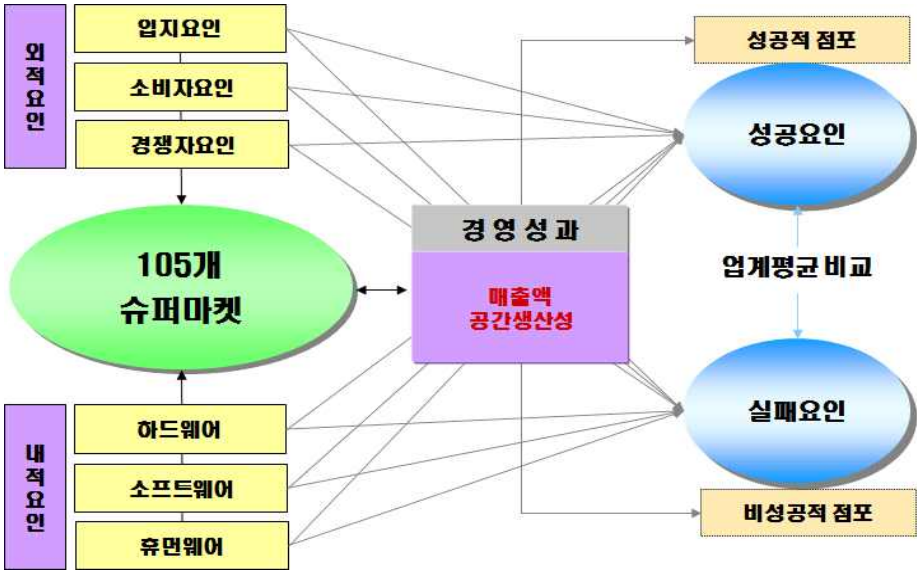
Ⅲ. 연구방법

1. 연구모형

1장에서 밝힌 바와 같이 본 논문의 연구목적은 점포성과에 영향을 주는 독립변수를 분석하는 것이다. 이러한 연구목적을 달성하기 위해, 2장의 기존연구 결과를 바탕으로 다음 그림과 같은 연구모형(research model)을 설정하였다.

먼저 점포의 환경적 요인인 외적요인과 점포의 역량 요인인 내적요인으로 구분하였다. 외적요인은 입지, 소비자, 경쟁자요인에 대해 8개의 독립변수를 두어 총 24개의 독립변수가 슈퍼마켓의 경영성과에 영향을 미치는 것으로 가정하였다. 또한, 내적요인은 하드웨어, 소프트웨어, 휴먼웨어 측면에 대해 각각 8개씩의 독립변수를 두어 24개의 독립변수가 점포성과에 영향을 미치는 것으로 가정하였다.

<그림 1> 연구 모형



2. 표본의 특성

본 연구에 사용된 표본 점포(sample store)는 전국의 나들가게와 농협 하나로마트 중 규모가 큰 상위 105개 점포를 대상으로 하였다. 나들가게와 하나로마트를 표본으로 선정한 이유는 표본 점포의 이질성(heterogeneity) 문제를 극복하기 위해서이다. 이들 표본은 점포이미지, 점포형태, 취급품목, 가격수준 측면에서 슈퍼마켓 업태로서 표본의 동질성(homogeneity)을 확보할 수 있을 것으로 판단된다. 또한, 지리적으로 전국의 대도시 및 읍면단위에 고루 분포하고 있다는 측면에서 공간적 불안정성(spatial non-stationarity) 문제를 보완할 것으로 보인다.

표본의 평균 매장면적은 1,428.2㎡(429평), 평균 종사자 32명, 1일 평균 매출은 1,716.7만원으로 조사되었다. 지역별로는 서울 및 경기 등 수도권이 35개, 경상도 29개, 충청도 13개, 전라도 13개, 강원도 10개, 제주도 5개 등이다.

3. 변수의 산정

본 연구에서 종속변수와 독립변수는 앞장의 기존연구결과를 토대로 본 연구의 연구목적에 맞도록 선정되었으며, 구체적인 세부 변수는 다음과 같다. 우선 종속변수는 매출액, 평당 생산성, 인당 생산성 등을 조사하였으나, 객관적인 측정을 위하여 평당 생산성을 활용하였다. 독립변수는 외적요인 3개(입지, 소비자, 경쟁)와 내적요인 3개(하드웨어, 소프트웨어, 휴먼웨어)에 대해 각각 8개씩 총 48개를 분석대상으로 하였다. 이들 세부 변수들은 다음 표와 같다.

<표 1> 측정 변수의 구성

구분	변 수		변수 속성
종속변수	점포성과		105개 표본점포의 경영성과로 매출액, 평당 생산성, 인당 생산성으로 설정하였으나, 본 연구에서는 평당 생산성을 종속변수로 활용함.
독립변수	외적요인	입 지	대상점포의 입지특성(locational attributes)을 나타내는 지표로서 입지의 유불리에 의해 점포성과가 달라짐. 본 연구에서는 '지형지세', '입지유형', '도로조건', '차량통행량', '보행자통행량', '대중교통수단', '상업집적도', '향후 성장잠재력' 등 8개 독립변수를 설정하였음.
		소비자	상권내 거주하는 소비자(customer)와 관련한 지표로 양적측면과 질적측면으로 구분이 가능함. 본 연구에서는 '상권내 인구수', '인구증가율', '연령별 분포', '인구밀도', '소득수준', '주택유형', '직업분포', '라이프스타일' 등 8개 독립변수를 설정하였음.
		경쟁자	경쟁관련 지표로서 경쟁이 심할수록 대상점포의 점포성과에는 부정적(negative)인 영향을 미치게 됨. 본 연구에서는 경쟁점의 '경쟁점포수' '매장면적', '지점과의 거리', '주차대수', '상품구색', '가격수준', '신선도', '브랜드 인지도/선호도' 등 8개 독립변수를 설정하였음.
	내적요인	하드웨어	대상점포의 하드웨어(hardware) 측면의 점포속성(store attributes)으로서 본 연구에서는 '대지면적', '용도지역', '매장면적', '점포이미지', '주차대수', '인테리어', '집기/비품', 'POS대수' 등 8개 독립변수를 설정하였음.
		소프트웨어	대상점포의 소프트웨어(software) 측면의 점포속성으로서 '상품구색', '주력상품군', '브랜드 인지도', '브랜드 선호도', '신선도품질', '가격수준', '광고판촉', '서비스 수준', '정보시스템' 등을 독립변수를 설정하였음.
		휴먼웨어	대상점포의 휴먼웨어(humanware) 측면의 점포속성으로서 '종업원 역량', '점장의 리더십', '업무몰입도', '종업원만족도', '급여수준', '동기부여', '판매숙련도', '배달서비스' 등 8개 독립변수를 설정하였음.

4. 분석방법

1) 변수의 상관성 분석

슈퍼마켓의 경영성과에 영향을 미치는 외적요인(입지, 소비자, 경쟁자)과 내적요인(하드웨어, 소프트웨어, 휴먼웨어)에 대해 각 요인별로 다른 요인들의 영향을 통제하고 상관분석을 실시하는 편상관분석(partial correlation analysis)을 실시하였다.

2) 경영성과에 미치는 영향변수 측정

상관성이 있는 변수들을 중심으로 다중회귀분석(multiple regression analysis)을 실시하였다. 다중회귀분석은 두 개 이상 복수의 독립변수들이이 하나의 종속변수에 미치는 영향의 정도를 파악하는 통계분석기법이다.

3) 성공 및 실패점포의 영향요인 평가

슈퍼마켓 성공 및 실패 점포의 영향요인을 분석하기 위한 로지스틱 회귀분석을 실시하였다. 다중회귀분석을 통해 유의성이 있는 변수들을 활용하여 점포성과(공간 생산성)와의 영향관계를 분석하였다.

Ⅳ. 연구결과

1. 변수의 상관성 측정 결과

48개의 독립변수들 중 종속변수와 높은 상관관계를 보이는 변수들을 선별하여 다중회귀분석을 실시하고자 하며, 이를 위해 피어슨 상관계수 0.1을 기준으로 설정하였다. 이는 일반적으로 사회과학에서 의미가 있는 계수로 인식되고 있다.

1) 외적요인 측면

입지요인에 대한 상관분석 결과 상관계수가 0.1이상인 요인은 입지유형, 도로조건, 차량통행량, 보행자통행량, 그리고 향후의 성장잠재력이 점포성과와 상관관계가 높은 것으로 나타났다. 소비자요인 분석 결과 상관계수가 높은 변수는 상권내 인구수, 인구증가율, 소득수준, 주택유형, 라이프스타일, 직업유형, 그리고 인구밀도가 선정되었다. 경쟁자요인에서는 경쟁점 수, 경쟁점 매장면적, 경쟁점과의 거리, 경쟁점 주차대수, 경쟁점 가격수준 등이 상관성이 높은 것으로 분석되었다.

2) 내적요인 측면

하드웨어 측면에서 상관계수가 0.1이상인 변수는 매장면적, 대지면적, 용도지역 등의 변수가 상관성이 있는 것으로 분석되었다. 소프트웨어 측면에서 상관분석 결과 상품구색, 품질, 주력상품(야채/청과, 가공식품), 브랜드 인지도, 브랜드 선호도, 서비스수준 등이 선정되었다. 휴먼웨어 측면에서 상관계수가 0.1이상인 변수는 종업원수, 종업원 직무만족도, 종업원 판매역량, 점장의 리더십 등이 선정되었다.

2. 경영성과에 영향을 미치는 변수 분석 결과

1) 경영성과에 영향을 미치는 요인추출

상관분석에 의해 선정된 변수들을 독립변수로 설정하여 다중회귀분석을 실시한 결과 외적요인 중 소비자요인에서 '상권내 인구수'와 '소득수준'이, 경쟁자요인에서 '경쟁점 가격수준'이 경영성과에 영향을 미치는 것으로 분석되었다. 내적요인 중에서는 하드웨어 측면에서 '매장면적', 소프트웨어 측면에서 '상품구색', 그리고 휴먼웨어 측면에서는 '종업원 역량'이 점포성과에 영향을 미치는 것으로 분석되었다. 이들 변수들로 회귀분석한 결과는 다음 표와 같다.

<표 2> 회귀분석 결과

구 분			비표준화 계수 β	표준 오차	표준화 계수 β	t값	유의 확률 (P-value)	F값	유의 확률
(상수)			-23,772	5,845		-4.067	0.000	17.482	0.000
외부	소비자	상권내 인구수	3,972	1,080	0.282	3.678	0.000		
	소비자	소득수준	1,844	824	0.172	2.239	0.027		
	경쟁자	경쟁점 가격수준	2,651	911	0.222	2.911	0.004		
내부	하드웨어	매장면적	2,530	1,043	0.175	2.425	0.017		
	소프트웨어	상품구색	4,074	1,874	0.157	2.174	0.032		
	휴먼웨어	종업원역량	2,670	510	0.397	5.237	0.000		
종속변수 : 평당 생산성, R^2 = 0.52									

소비자 요인의 '상권내 인구수'가 슈퍼마켓의 점포성과에 미치는 영향은 t값=3.678, p값=0.000, '소득수준'이 점포성과에 미치는 영향은 t값=2.239, p값=0.027, 경쟁자요인의 '경쟁점 가격수준'이 점포성과에 미치는 영향은 t값=2.911, p값=0.004로 나타났다. 외적요인 중에서 '상권내 인구수', '소득수준', '경쟁점 가격수준'이 유의수준 0.05에서 통계적으로 유의하기 때문에 슈퍼마켓의 점포성과에 영향을 미치고 있음을 알 수 있다.

내적요인의 하드웨어에 해당하는 '매장면적'이 슈퍼마켓 점포성과에 미치는 영향은 t값=2.425, p값=0.017, 소프트웨어의 '상품구색'이 점포성과에 미치는 영향은 t값=2.174, p값=0.032, 그리고 휴먼웨어의 '종업원 역량'이 점포성과에 미치는 영향은 t값=5.237, p값=0.000으로 나타나 유의수준 0.05에서 통계적으로 유의한 것으로 분석되었다.

외적요인에서 '상권내 인구수', '소득수준', '경쟁점 가격수준', 그리고 내적요인에서 '매장면적', '상품구색', '종업원 역량'의 비표준화 계수 값 B가 양(+)의 값을 갖기 때문에 슈퍼마켓 점포의 성과에 긍정적인 영향을 미치고 있다. 표준화 계수를 통해 슈퍼마켓 점포의 성과에 미치는 영향의 정도를 살펴보면 '종업원 역량'(베타 값=0.397), '상권내 인구수'(베타 값=0.282), '경쟁점 가격수준'(베타 값=0.222), '매장면적'(베타 값=0.175), '소득수준'(베타 값=0.172), '상품구색'(베타 값=0.157)의 순으로 나타난다.

슈퍼마켓의 점포성과에 미치는 영향을 보면 내적요인의 '종업원 역량'이 가장 큰 영향을 미치고 있는 것으로 나타나서, 슈퍼마켓 경영에서 인적자원(human resource)의 역량강화가 가장 중요한 요인인 것으로 볼 수 있다. 외적요인에서는 '상권내 인구수'가 큰 영향을 미치고 있어서, 슈퍼마켓 출점에서 상권의 양적지표인 인구통계적 지표를 활용한 상권분석이 이루어져야 함을 알 수 있다. '경쟁점 가격수준'도 중요한 영향을 미치고 있다. 소비자요인에서 상권내 인구의 질적 지표인 '소득수준' 역시 점포성과에 영향을 미치고 있다. '상품구색' 또한 점포성과에 영향을 미치는 요인으로 나타났다.

2) 성공 및 실패에 영향을 미치는 변수 측정

상관분석 결과 선정된 변수들을 독립변수로 설정하여 성공과 실패에 영향을 미치는 정도를 로지스틱 회귀분석을 통해 살펴보았다. 여기서 종속변수로 사용된 성공과 실패의 기준에 수익률은 큰 차이가 없으므로 평균매출액을 기준으로 구분하였다. 분석 결과 소비자요인에서 '상권내 인구수', 경쟁자요인에서 '경쟁점포 수', 내부의 소프트웨어 요인에서 '브랜드 선호도'가 성공 및 실패에 영향을 미치는 것으로 조사되었다.

<표 3> 성공 및 실패에 영향을 미치는 변수별 측정 결과

구 분		B	S.E.	Wald	P-value	Exp(B)
외부	상권내 인구수	1.042	0.473	4.865	0.027	2.836
	경쟁점포수	-0.484	0.196	6.097	0.014	0.617
내부	브랜드선호	0.533	0.229	5.402	0.020	1.704
상수		-1.048	1.230	0.726	0.394	0.350

분류정확 %	-2 Log 우도	Cox와 Snell의 R^2	Nagelkerke R^2
73.333	120.934	0.146	0.199

	단계	블록	모형
카이제곱	5.187	16.512	16.512
유의확률	0.023	0.001	0.001

외부요인인 '상권내 인구수'가 슈퍼마켓 점포의 성공과 실패에 미치는 영향은 wald값=4.865, p값=0.027로 나타났으며, '경쟁점포수'가 점포의 성공과 실패에 미치는 영향은 wald값=6.097, p값=0.014로 나타나 통계적 유의수준 0.05에서 성공과 실패에 영향을 미치는 것으로 분석되었다. 그리고 내부요인의 소프트웨어 측면에서 '브랜드 선호도' 역시 슈퍼마켓 점포의 성공과 실패에 미치는 영향이 wald값=5.402, p값=0.020로 나타나 유의수준 0.05에서 통계적으로 유의한 것으로 나타났다.

요컨대, 슈퍼마켓의 성공과 실패에 영향을 미치는 요소로 외부요인에서는 '상권내 인구수', '경쟁점포수', 그리고 내부요인에서 '브랜드 선호도'인 것으로 판명되었다. 즉, 슈퍼마켓 점포의 선택특성인 외부·내부요인 중에서 '상권내 인구수', '경쟁점포수', '브랜드 선호도'의 비표준화 계수 값 B가 양(+)의 값을 갖기 때문에 슈퍼마켓 점포의 성공과 실패에 긍정적인 영향을 미치고 있음을 알 수 있다.

Exp(B)는 성공에 대한 승산비(odds ratio)를 나타내는 것으로서 '상권내 인구수'가 1단위 증가 할 때마다 슈퍼마켓 점포가 실패할 확률보다는 성공할 확률이 2.836배 높아진다는 것을 알 수 있다. 이는 슈퍼마켓 출점에 있어서 '상권내 인구수'가 가장 중요한 요인임을 시사한다. 또한, '경쟁점포수'의 의미는 상권 내에서

'경쟁점포 수'가 한 단위 개선될(적어질) 때 마다 슈퍼마켓 점포가 실패할 확률보다는 성공할 확률이 1.622배 높아진다는 것을 의미한다. 따라서 슈퍼마켓 경영에서 경쟁관계가 중요한 성공 및 실패요인으로 작용하고 있다고 볼 수 있다.

'브랜드 선호도'의 경우는, 소비자의 점포선택에 있어서 브랜드 선호도가 1단위 증가할 때마다 실패할 확률보다 성공할 확률이 1.704배 증가함을 의미한다. 이는 슈퍼마켓 경영에서 슈퍼마켓 업체의 브랜드 혹은 업체의 평판이 성공 및 실패를 가늠하는 중요한 요소(element)임을 시사한다. Fortheringham(1993)이 소비자의 슈퍼마켓 선택에 있어서 슈퍼마켓의 체인이미지(chain image)가 점포 크기나 점포 간의 경쟁만큼 중요하다는 주장을 뒷받침하는 결과로 볼 수 있다. 성공점포 및 실패점포로 분류된 결과를 비교하여 해당모형이 얼마나 잘 부합되는지를 살펴보았다. 로지스틱 회귀모형의 적합성 판정 결과 73.333%의 분류 정확도를 보여 어느 정도 적합한 모형임을 입증해 주고 있다.

Ⅴ. 결론

본 연구의 목적은 외적·내적 요인에서 설정된 여러 독립변수가 점포성과에 미치는 영향관계의 메커니즘(mechanism)을 규명하는 것이다. 분석결과 '상권내 인구수', '소득수준', '경쟁점 가격수준', '매장면적', '상품구색', '종업원 역량' 등 6개 변수가 슈퍼마켓의 점포성과에 영향을 미치는 것으로 분석되었다. 슈퍼마켓의 성공 및 실패에 영향을 미치는 변수로는 '상권내 인구수', '경쟁점포 수', '브랜드 선호도'가 슈퍼마켓의 성공 및 실패에 영향을 미치는 것으로 조사되었다. 따라서 신규출점을 계획 중인 자영업자나 기존에 슈퍼마켓을 운영 중이나 성과가 부진한 점주들은, 위와 같은 관점에서 자신의 점포를 면밀하게 분석한다면 좋은 성과를 얻을 수 있을 것이다.

끝으로 본 연구는 매출액을 성과변수로 사용하였으나, 향후에는 시장점유율, 매출액성장률, 투자수익률(ROI) 등과 같은 성과변수들도 포함하여 다각적으로 연구한다면 보다 의미 있는 결과를 얻을 수 있을 것이다.

참고문헌

서상윤, 차재빈(2013), 경영컨설팅연구 제13권 제1호 2013년 3월

Amold S.J, & Oum T.H., & Tigert D.J., "Determinant attributes in retail patronage seasonal, temporal, regional and international comparisons", Journal of Marketing Research, 20(2), 1983

Bates, A.D., Retailing and its Environment, 1979

Berman, R. & Evans,]. R.(1976), Retail Management, A Strategic Approach.

Cathy H., & Mark D., The location and merchandising of non-food insupermarkets, International Journal of Retail and Distribution Management, Vol.24, No.3, 1996, pp. 17-25.

Darden W. R., "A Percentage Model of Consumer Behavior", in R. W. Stamfl and E. Hivschman, eds., Competitive Structure in Retail Markets; the Department Store Perspective, Chicago: AMA, 1980, pp. 45-47.

Forthergham A.S., "Market Share Analysis Techniques a Review and Illustration of Current US Practice", in Wrigley, N., eds., Store Choice, Store Locarion and Market Analysis, Routledge, New York, pp.120-159.

Hortman S,M., & Allaway A.W., & Marson J.G., & Rasp, J., 1990. Multi-segment analysis of supermarket patronage. Jounal of Business Research 21, 209-223.

Huff, D. L. (1963), "A probabilistic analysis of shopping center trade areas," Land Economics, 39.

Kotler P., Marketing Management, 8th ed., Prectice-Hall, 1994, p.342.

Mason, J. B. & Meyer, M. L.(1981), Modern Retailing, Theory and Practice, 3rd ed

연구논문

기계학습과 인적자원 변수를 활용한 신용평가모형 구축 연구

박주완*

본 논문의 목적은 기업의 인적자원 관리 및 개발 요소를 다양한 기계학습 기법에 적용한 후 예측 성능과 안정성이 가장 우수한 기법이 무엇인지 확인하는 것이다. 모형 구축을 위해 2017년 한국직업능력개발원의 인적자본 기업패널(HCCP) 설문조사 자료와 (주)나이스신용평가의 기업신용등급을 이용하였다.

모형 구축을 위한 독립변수는 인적자원 바퀴모델의 구성요소에 인적자본 기업패널 설문조사 문항을 연결하여 사용하였으며, 종속변수는 기업신용평가등급을 이용하여 우량과 불량을 정의하였다. 또한 모형 구축을 위한 기계학습 알고리즘으로는 가장 많이 알려져 있는 의사결정나무, 로지스틱회귀, 신경망, 랜덤포레스트, SVM을 이용하였다.

모형 구축 결과, 본 논문에서 사용한 자료에 대해서는 기계학습 알고리즘 중 로지스틱회귀모형의 분류 정확도와 안정성이 가장 높게 나타났으므로, 이를 이용하여 기업신용평가모형을 구축하는 것이 가장 타당하다는 결론을 내릴 수 있다.

* 신용보증재단중앙회 교육연구부 선임연구위원(통계학박사)

Ⅰ. 서론

기업신용평가(corporate credit evaluation)는 기업의 채무 상환 능력 등을 종합적으로 판단하기 위해 기업의 내·외부 경영환경 등을 조사 및 분석하여 신용도[1]를 산출하는 것이다. 이를 위해 기업신용평가는 일정 시점에 기업이 내부적으로 보유한 재무 및 비재무적 요인과 외부 환경 등을 조사, 분석, 검토하여 종합하는 과정을 거치게 된다. 이러한 과정을 통해 산출된 신용평가 결과는 기업과 투자자 간에 발생하는 정보 불균형(information asymmetry) 문제를 해결하는데 도움이 된다고 알려져 있다(김성환과 김태동, 2014).

4차 산업혁명(the fourth industrial revolution) 시대에 접어들면서 다양한 산업 분야에서 빅데이터나 인공지능의 활용을 위한 기술 개발에 많은 투자를 하고 있다(박주완, 2019). 그리고 실제로 기업이나 개인의 신용평가에 빅데이터와 다양한 기계학습을 적용하는 사례들이 점차 증가하고 있다(김효진, 2018). 금융산업에서 빅데이터와 기계학습을 신용평가나 대출심사 등에 적용하는 국내외 사례를 살펴보면 다음과 같다.

먼저 해외 사례로는 Kabbag, Zest Finance, 요코하마은행과 지바은행 등이 있다. Kabbage는 소상공인 신용평가 시 기존의 재무자료 이외의 배송, 회계, 인터넷 자료 등을 기계학습에 적용하여 소상공인의 신용평가를 수행하고 있다. Zest Finance는 전통적인 신용정보 외에 직장정보, 고정수입, 인터넷 포스팅 내용 등이 포함된 7만개가 넘는 변수를 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 2017년에 신용도 판단이 어려운 사회 초년생과 중금리 대출 고객들을 대상으로 기계학습을 적용한 신용평가시스템 개발을 완료하였다(서울경제신문, 2017). 케이뱅크는 가계나 자영업자의 신용대출 심사 시

1) 신용도(creditability)란 장래의 어느 시점에 그 대가를 지급할 것을 약속하고 경제적 가치를 획득할 수 있는 능력을 의미한다.

KT의 통신요금 납부 실적, 비씨카드 신용카드 결제 정보를 중금리 대출 심사에 적용하여 연체율 감소 효과를 거두고 있다(연합인포믹스, 2018). 카카오뱅크는 이상 거래를 탐지하기 위해 지도학습 기계학습 모형을 활용하고 있는데, 이는 다양한 사기 거래 데이터를 통해 '정상 데이터와 달리 사기 데이터는 이런 특성이 있어'라는 걸 학습시킨 후 이상 거래를 탐지하는 방식이다(블로터, 2019). 이와 같은 사례들을 통해 국내외 여러 기관에서의 신용평가 시 빅데이터와 기계학습 이용에 대한 관심과 중요성이 점차 높아지고 있음을 유추할 수 있다(박주완, 2019). 그러므로 인자원관리 및 개발 등 인적자원 활동 관련 요인을 기계학습 알고리즘에 적용하여 신용평가모형을 구축하는 연구는 매우 의미 있는 작업으로 판단된다.

기업 신용평가 모형을 구축하기 위해 2017년 한국직업능력개발원의 인적자본 기업패널(Human Capital Corporate Panel, HCCP) 7차년도 조사 자료와 NICE 신용평가(주)의 KIS-신용평점모델에서 생성된 2017년 기업 신용평가등급을 이용한다. 여기에서 인적자본 기업패널 자료는 2017년에 조사되었지만 조사된 내용들은 2016년도의 기업 인적자원에 관한 사항들이기 때문에 2017년 기업 신용평가 점수와 자료를 합쳐서 사용해도 자료 통합 시점 상에는 큰 문제가 없는 것으로 사료된다.

본 연구에서 사용하고자 하는 기계학습 분류 기법은 대표적으로 잘 알려져 있는 의사결정나무모형(decision tree), 로지스틱회귀모형(logistic regression), 신경망모형(neural network), 서포트벡터머신(support vector machine, SVM), 랜덤포레스트모형(random forest)이다. 모형의 평가는 예비 방법(holdout method), 구축된 모형의 분류 정확도를 평가하기 위한 측도(measure)로는 정분류율, G-mean, F1 측도, 반응률(percent of response)을 이용한다. 자료를 분석하고 모형을 구축하기 위한 통계 프로그램은 SAS9.4와 R3.5.1 버전을 이용하는데, 이 때 R로는 기계학습을 수행할 수 있는 함수 및 라이브러리인 "glm, rpart, nnet, rf, kernlab"을 이용하여 분류 모형을 생성한다.

논문의 구성은 다음과 같다. 2장에서는 신용평가모형 관련 연구 문헌들을 고찰하고, 3장에서는 모형 구축에 사용한 알고리즘, 데이터 정제와 모형 평가

방법을 설명한다. 4장은 기업 신용평가모형을 구축하기 위한 표본 및 변수, 모형 구축 과정에 대해 설명하며, 5장은 모형을 구축하여 예측 성능을 비교 및 평가 후 예측 성능이 가장 우수한 모형을 선택한다. 마지막으로 6장에서는 결론에 대해 고찰한다.

Ⅱ. 선행연구 고찰

기업 부도 예측 연구의 초창기에는 다변량판별분석(Altman, 1968)과 로지스틱회귀모형(Ohlson, 1980) 등의 전통적인 통계 방법론을 적용하는 연구가 주를 이루었는데, 이후 신경망모형, SVM 등 다양한 데이터마이닝 기법들을 이용하여 예측 성능을 향상시키는 방향으로 발전하여 왔다(강신형, 2016). 본 절에서는 기계학습들을 이용한 부실기업 예측 및 기업신용평가모형 구축에 대한 연구 사례를 살펴보고자 한다.

먼저 기업의 부도 예측 시 인공신경망모형의 우수성을 검증한 연구 사례로는 이건창(1993), 박정윤(2000), 전성빈 · 김영일(2001), 정유석(2003) 등이 있다. 이건창(1993)은 다변량판별분석, 인공신경망모형으로 기업 부도 예측을 수행한 후 이를 비교하여 인공신경망 모형의 예측 성능이 우수하다고 하였다. 박정윤(2000)은 1991~1996년 자료로 기업 부도 예측을 실시한 결과 MDA모형, 확률모형, 인공신경망모형 중 인공신경망모형의 예측 성능이 가장 우수하다고 하였다. 전성빈 · 김영일(2001)은 기업 부도 예측 시 인공신경망모형의 예측 성능이 가장 우수하였고 다변량판별분석, 로지스틱회귀모형 등의 분류 정확도는 비슷한 수준이라고 하였다. 정유석(2003)은 로지스틱회귀모형, 다변량판별분석, 인공신경망모형을 이용하여 부도 기업을 예측한 결과 인공신경망모형의 예측력이 가장 우수하다고 하였다.

다음은 의사결정나무모형, 로지스틱회귀모형과 SVM의 예측 성능이 우수함을 실증 분석한 연구 사례이다. 조준희 · 강부식(2007)은 코스닥기업의 부도 예측 시 의사결정나무모형이 신경망모형이나 로지스틱회귀모형 보다 좋은 예측

성능을 가지고 있다고 하였다. 박주완 · 송창길(2015)은 인적자본기업패널과 NICE 자료를 이용하여 로지스틱회귀모형, 신경망모형, 의사결정나무모형으로 소기업 이상에 대해 신용평가모형을 구축한 결과 로지스틱회귀모형의 예측 성능이 가장 우수함을 실증 분석하였다. 윤종식 · 권영식(2007)은 소상공인 부실 예측모형 연구에서 로지스틱회귀모형, 다변량판별분석, CART, C5.0, 신경망 모형, SVM 중 SVM의 예측 성능이 가장 우수함을 보였다. 박주완(2017)은 소상공인 신용평가 시 로지스틱회귀모형이 의사결정나무모형이나 신경망모형 보다 예측 성능이 우수하며, 계급불균형 자료를 이용하여 신용평가모형 구축 시 예측 성능이 저하될 수 있다고 밝히고 있다.

마지막으로 앙상블 기법을 이용한 연구 사례이다. 김승혁 · 김종우(2007)는 SOHO 부도 예측 시 수정된 배깅 예측자(Modified Bagging predictors)[2]가 인공신경망과 배깅 예측자(Bagging predictors) 보다 예측 성능이 향상된다고 하였다. 김명종 · 강대기(2010)는 기업 부실 예측을 위해 인공신경망과 부스팅 인공신경망 앙상블 기법을 적용한 결과 앙상블 학습은 기업 부실 예측 문제에 있어 전통적인 인공신경망을 개선할 수 있다고 하였다. 김성진 · 안현철(2016)은 1,295개 국내 상장 기업을 대상으로 기업신용평가모형 구축 시 다변량판별분석, 인공신경망, SVM, 랜덤포레스트모형을 비교한 결과 랜덤포레스트모형의 예측 성능이 가장 우수함을 보였다.

이상의 연구를 살펴보면 연구자에 따라 결과가 상이하게 나타나고 있는데 이는 분석 자료에 따른 차이일 가능성이 높다. 여기에서 중요한 함의는 특정한 하나의 알고리즘이 가장 우수하다는 결론을 내릴 수 없다는 것이다. 그리고 인적자원 관련 변수를 이용한 부도 예측 등에 대한 연구는 많지 않다는 것이다. 그러므로 인적자원 관련 변수와 기계학습 알고리즘을 이용한 기업 신용평가모형 구축 연구는 합당한 시도이며 의미가 있는 작업으로 사료된다.

2) 부스트랩(bootstrap) 방법으로 다수의 모델을 만들고 평균 이상의 예측 정확도를 가지는 모형들만을 선택해 투표(voting)하는 방법

Ⅲ. 모형 구축 및 평가 방법론

1. 모형 구축 알고리즘

차주의 우불량 여부를 판별하고 신용도를 예측하기 위한 신용평가모형은 기계학습 관점에서 지도학습 중에서 분류(classification) 모형이다(오미애 외, 2017). 지도학습을 위한 자료에는 종속변수와 독립변수가 필요하다. 대표적인 지도학습 모형으로는 선형회귀모형(linear regression), 로지스틱회귀모형, 의사결정나무모형, 신경망모형, 랜덤포레스트모형, SVM 등이 있다(박주완, 2017). 본 연구에서 사용할 분류 모형 구축 알고리즘은 보편적으로 많이 사용하고 있는 로지스틱회귀모형, 의사결정나무모형, 신경망모형, 랜덤포레스트모형, SVM 5가지인데, 본 절에서는 연구에 사용된 5가지 모형에 대해 고찰한다.

로지스틱회귀모형은 종속변수의 계급이 0과 1 두 가지 값을 가지고 관심의 대상이 되는 계급이 1이 될 확률을 예측하는 모형이다(Hosmer and Lemeshow, 2000; 성웅현, 2001). 실제로 현업에서 신용평가모형을 구축할 때 로지스틱회귀모형이 가장 많이 사용되고 있는데, 이유는 다음과 같다. 첫째 모형 구축이 올바르다면 로지스틱회귀모형은 정확성이 우수하고, 둘째 구축 과정이 용이하고 해석하기가 쉬우며, 셋째 과대적합(over-fitting)할 가능성이 적고, 오차를 최소화하는 선형적인 관계를 찾는데 매우 우수한 기법이기 때문이다(이영섭, 2003). 본 연구에서는 통계 프로그램인 R에서 제공하는 기본 함수인 “glm”이다.

의사결정나무모형은 나무 구조로 도표화하여 의사결정 규칙(decision rule)을 찾고 분류와 예측을 수행하는 방법으로 대표적인 알고리즘으로는 CHAID, C5.0, CART가 있다. 이 모형의 장점은 나무구조로 분류 규칙을 도표화하기 때문에 이해가 쉽다는 것이다. 또한 연속형과 범주형 자료를 동시에 다룰 수 있고, 결측치를 분석에 활용할 수 있다. 그리고 교효효과(interaction effect) 해석이 쉬우며, 선형성, 정규성, 등분산성 등 통계적인 가정이 필요하지 않다. 그러나 최적의 의사결정나무를 찾는 것은 쉽지 않으며, 매우 세밀하게 분류 및 예측을

수행할 경우 과대적합(over-fitting)의 가능성이 높아 새로운 자료에 대한 일반화 성능이 좋지 않을 수 있다(최종후 · 진서훈, 2005). 본 연구에서는 R 라이브러리 중 CART를 수행하는“rpart”를 사용한다.

신경망모형은 과거의 경험이나 지식을 습득하여 모형화하면서 오류 최소화 과정을 통해 예측 및 분류를 수행하며, 어떠한 통계적인 분포도 가정하지 않는다. 신경망모형 중 가장 널리 사용되는 다층인식자(multi-layer perceptron, MLP) 신경망은 입력층, 은닉층, 출력층으로 구성되어 있고 노드를 통해 연결되는 구조이다. 먼저 입력층을 통해 자료를 입력받고, 은닉층에서는 입력층으로부터 전달되는 변수값들의 선형결합(linear combination)을 비선형함수로 처리하여 출력층 또는 다른 은닉층으로 전달하여, 최종적으로는 출력층을 통해 예측 결과를 산출한다(강창완 외, 2007). 신경망모형의 장점은 비선형적인 관계를 찾아낼 수 있고 예측의 정확성이 매우 높다는 것이다. 그러나 과대적합(over-fitting)하는 경향이 있으며 결과의 해석이 매우 어렵다(Ripley, 1996). 본 연구에서는 R 라이브러리 중 “nnet”를 사용한다.

랜덤포레스트모형은 의사결정나무를 확장한 개념으로 앙상블(ensemble) 모형 중 하나이다. 랜덤포레스트는 다수의 의사결정나무모형을 만들어 예측 성능을 높이는 방법이다(Yoo, 2015). 일반적으로 의사결정나무모형은 특이값(outlier)을 하나의 노드로 구성할 수 있기 때문에 편향된 분포에 민감하지 않지만, 깊이가 깊어질수록 과적합의 위험이 커진다. 이와 같은 위험을 최소화하여 예측 성능을 높이고자 고안된 기법이 랜덤포레스트모형이다(Breiman, 2001). 이 모형의 장점은 트리의 다양성을 극대화하여 예측력이 우수하고 많은 나무의 예측 결과를 종합하기 때문에 안정성이 높다는 것이다. 그러나 의사결정나무모형의 장점인 설명력은 없다. 본 연구에서는 랜덤포레스트 모형을 수행할 수 있는 R 라이브러리 중 “randomForest”를 사용한다.

SVM은 분류 및 예측 시 가장 보편적으로 사용되고 있는 기계학습 알고리즘 중 하나이다. 일반적으로 SVM은 벡터 공간에 존재하는 학습 데이터가 어떠한 그룹에 속하는지를 분류하기 위한 선형 분류자(linear classifier)를 찾는 기법이다(Lantz, 2015). 이 모형은 다양한 학습 데이터의 분포에서도 정확도 측면

에서 우수하다는 장점이 있지만, 직관적인 해석이 불가능하다는 단점이 있다. 이와 같은 이유로 결과의 해석보다는 분류의 정확도가 중요한 경우 SVM을 사용하는 경우가 많다(김의중, 2016). 본 연구에서는 R 라이브러리 중“kernlab”을 사용한다.

2. 자료 변환 및 변수 선택

성공적인 모형 구축을 위해서는 양질의 원천 자료(raw data)가 확보되어야 하며 다양한 방법을 이용하여 분석 가능한 형태로 데이터를 정제(cleaning)하여야만 한다. 데이터 정제 시에는 기본적으로 분석 변수에 대해 결측치(missing value)나 특이값(outlier value) 등에 대한 사항을 파악한 후, 이를 제거하거나 대체하는 작업 등을 거치게 된다. 그리고 신용평가모형 구축을 위한 독립변수를 선택할 때 통계적으로 선택된 결과를 바탕으로 대출 시 비즈니스 관점의 부합성을 고려하여 최적의 변수를 조합하여야 한다(신용보증재단중앙회, 2017).

본 연구에서는 박주완(2018)에서 사용한 변수 정제 및 선택 기법을 이용한다. 변수 정제 및 선택 방법을 구체적으로 설명하면 다음과 같다. 먼저 독립변수와 불량과의 관계를 이용한 변수 계급화(classing) 기법 이용, 원천 자료 중 범주형(categorical) 변수는 종속 및 독립변수 간 카이제곱 통계량, 연속형(continuous) 변수에 대해서는 t-검정을 이용하여 1차적으로 변수를 선정한다. 다음 단계는 1차로 선택된 변수 풀(pool)에 대해 성김화(coarse classing) 기법으로 계급 세분화된 값을 축약하여 재범주화를 수행하고 신용평가모형 구축에 활용한다.

본 연구에서 사용하는 변수 선택 기법인 계급화는 원천 자료의 표준화를 위해 실제 현업에서 신용평가모형 구축 시 사용하는 방법이다. 이는 원래의 독립변수들과 종속변수인 불량 여부와의 관계 분석을 통해 불량률이 유사한 독립변수의 범주를 하나의 계급으로 묶은 후 순위의 의미를 가지도록 변환하는 기법이다. 계급화 기법은 크게 계급세분화(fine classing)과 성김화(coarse classing) 단계로 구분된다. 먼저 계급세분화 단계에서는 자료의 크기나 변수의 척도에

따라 다소 차이는 있지만, 개별 독립변수의 값을 정렬(sorting)한 후, 이를 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(신용보증재단중앙회, 2017). 성김화는 1차적으로 계급세분화에 의해 범주형으로 변환된 변수를 동질적인 불량률을 보이는 구간으로 다시 묶는 단계이다(Leung et al., 2008). 계급화 기법을 이용할 경우 결측치와 특이값의 사용이 가능해지는데, 그 이유는 결측치나 특이값이 불량률과 연관되어 하나 또는 그 이상의 구간으로 묶이기 때문이다

3. 모형 평가

모형 평가는 구축된 모형의 예측 성능과 안정성을 확인하는 과정으로, 여러 가지 모형의 예측과 분류 성능을 평가 및 비교하여 예측 성능과 안정성이 가장 우수한 모형을 선택하는 필수 단계이다(강현철 외, 1999).

개발된 모형을 평가하는 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). 본 논문에서는 예비 방법을 이용하여 “훈련용 자료:평가용 자료 = 7:3”, 즉 모형을 구축 하는 데에 전체 자료의 70%를 사용하고 나머지 30%로는 구축된 모형을 검증하고 평가하는 데에 사용한다. 예비 방법의 설명은 다음과 같다.

예비 방법은 난수(random number)를 이용하여 전체 분석용 자료를 두 개의 배타적인(exclusive) 훈련용(training data)과 검증용 자료(validation data)로 임의(randomly) 분할한 후, 모형 구축을 위해서는 훈련용 자료를 사용하고 검증용 자료는 모형 평가에 사용한다. 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 방

법으로 평가의 정확성도 높고 평가에 소요되는 시간이 단축된다는 장점이 있다(강창완 외, 2007). 그러나 평가용 자료를 모형 개발에 사용할 수 없고, 훈련용과 평가용 자료의 비율에 따라 다른 결과가 나타날 수 있다는 문제점과 개체수가 크지 않을 경우 불안정한 값을 제공한다는 단점이 있다(최종후 · 진서훈, 2002).

일반적인 종속변수가 범주형인 모형을 평가하는 경우의 측도는 오분류 행렬(confusion matrix)을 통한 여러 가지 방법을 사용할 수 있다(성웅현, 2001). 이외에도 리프트(lift) 도표, 반응률(response rate) 도표 등이 많이 사용되고 있다(강현철 외, 1999). 본 연구에서는 정분류율, G-Mean, F1 측도, 반응률을 이용하여 구축된 신용평가모형의 예측 성능을 비교하고 평가한다.

정분류율은 전체 자료를 얼마나 제대로 분류하는가의 문제이므로 값이 클수록 좋은 모형이다. 정분류율은 실제 0이 0으로 실제 1이 1로 분류되는 비율을 의미한다(강현철 외, 1999). 다음의 식에서 n_{00}과 n_{11}은 정분류가 되는 개수, n은 전체 표본수를 나타내고, 다음의 식으로 표현된다.

$$\text{정분류율} = (n_{00} + n_{11})/n \times 100 \qquad \text{(식 1)}$$

G-mean은 결과 범주가 0인 집단과 1인 집단을 동등하게 고려하는 측도로써 실제 범주가 0인 집단에 대한 정확도와 범주 1인 집단에 대한 정확도의 기하평균이다(Kubat et al. 1998). 그러므로 G-mean의 값이 클수록 좋은 예측 모형이다. G-mean의 산식은 다음과 같다(박주완, 2010).

$$G-mean = \sqrt{\frac{n_{00}}{n_{0+}} \times \frac{n_{11}}{n_{1+}}} \qquad \text{(식 2)}$$

$$, \; n_{0+} = \text{실제0인 개수}, n_{1+} = \text{실제1인 개수}$$

$F1$ 측도(measure)는 어떤 특정한 계급의 성공적인 분류가 다른 계급의 분류에 비해 훨씬 중요한 경우 사용되는 측정 기준이다. $F1$ 측도 값이 크다는 것은 특정 계급에 대한 예측 성능이 좋다는 것을 의미한다(Chawla et al., 2003). $F1$측도를 산출하기 위한 수식은 다음과 같다.

$$F1 = \frac{2rp}{(r+p)} = \frac{2}{1/r+1/p} = \frac{2 \cdot n_{11}}{n_{1+}+n_{+1}} \quad \text{(식 3)}$$

, p = 실제1, 예측1 정분류 빈도/예측1의 빈도

, r = 민감도 = 실제1, 예측1 정분류 빈도/실제1의 빈도

, n_{11} = 실제1이 예측1로 분류되는 개수

반응률은 훈련용 자료를 이용해 산출된 사후확률을 정렬하여 N 등분한 후, 각 등분에 포함된 종속변수의 특정 범주, 즉 불량의 빈도를 이용해 산출한다. 이와 같이 계산된 반응률은 도표를 통해 모형의 성능을 명확히 확인할 수 있는데, 모형에 의해 산출된 우량일 사후확률이 가장 큰 구간에서 가장 낮은 구간으로 갈수록 반응률, 즉 불량률이 낮게 나타나다가 급격하게 증가하는 형태인 경우 좋은 예측 판별력을 가진 모형이다(강현철 외, 1999).

$$\text{반응률} = \frac{\text{일정 } N \text{ 등분내 범주 1 빈도}}{\text{일정 } N \text{ 등분내 전체 빈도}} \times 100 \quad \text{(식 4)}$$

Ⅳ. 모형 구축

1. 분석 표본 및 변수

먼저 본 논문의 분석 표본은 2017년도 한국직업능력개발원 인적자본 기업패널 표본 474개 기업체 중 2017년의 신용등급이 있는 기업체 465개를 모형 구축을 위한 대상으로 한다.

분석 변수 중 종속변수는 기업의 우불량 여부는 2017년 ㈜NICE신용정보에서 제공한 신용평가등급 자료를 이용하여 정의한다. ㈜NICE신용정보에 의한 기업신용평가 등급은 크게 재무 평가와 비재무 평가로 구분되어 있는데, 재무평가는 기업의 재무제표를 기반으로 한 재무상태에 대한 평가이고, 비재무 평가는 재무적인 요인 이외의 항목을 이용한 기업의 대내외적 여건을 평가하는 것이다. 이를 이용하여 최종적으로 산출하는 기업의 신용평가 등급은 불량률에

기초하여 기업의 우불량 간 변별력과 안정성이 가장 우수한 재무 평가와 비재무 평가 점수 간 가중치를 적용하여 산출한다. 보통 각 기업 신용평가 등급은 불량률에 기초하여 10개의 신용등급으로 변환하는데 등급에 대한 의미는 다음의 표와 같다.

<표 4-1> 기업 신용등급의 구분

신용상태	신용등급	신용등급의 정의
우수	AAA	상거래 위한 신용능력 최우량급, 환경변화에 충분한 대처 가능
	AA	상거래 위한 신용능력 우량, 환경변화에 적절한 대처 가능
	A	상거래 위한 신용능력 양호, 환경변화에 대한 대처 능력 제한적
양호	BBB	상거래 위한 신용능력 양호, 경제 및 환경 악화에 따라 거래 안정성 저하 가능성
보통	BB	상거래 위한 신용능력 보통, 경제 및 환경 악화에 따라 거래 안정성 저하가 우려
	B	상거래 위한 신용능력 보통, 경제 및 환경 악화 시 거래 안정성 저하 가능성 높음
열위	CCC	상거래를 위한 신용능력이 보통 이하로, 거래 안정성 저하 예상되어 주의 요함
	CC	상거래를 위한 신용능력이 매우 낮으며 , 거래 안정성이 낮음
	C	상거래를 위한 신용능력이 최하위 수준이며 , 거래위험 발생 가능성 매우 높음
부실	D	현재 신용위험이 실제 발생하였거나 신용위험에 준하는 상태에 처함
평가 제외	R	1년 미만 재무제표, 합병, 영업양수도, 업종 변경 등 기업신용평가 등급 부여 유보
	NR	우편번호, 표준산업코드 등 누락으로 평가 유보

본 논문에서는 모형을 구축하기 위해 신용등급이 양호 이상인 경우를 우량기업(Y=1)으로 정의하고, 보통, 열위, 부실인 경우는 불량 기업(Y=0)으로 변환하여 사용한다. 평가 제외인 경우는 결측치 보정 방법 중 하나인 콜덱방법을 이용하여 2016년 또는 2017년 이후 신용평가 등급으로 대체하며, 이외의 결측자료는 분석에서 제외한다.

분석에 사용할 인적자원 활동 관련 독립변수는 1989년 맥라겐(McLagan)이 제시한 인적자원 바퀴(HR Wheel) 모형을 이용하여 설정하는데, 이 모형은 기업의 인적자원 실무자들에게 일반적으로 많이 수용되고 있는 것이다(김미숙 외, 2005). 인적자원 바퀴모형은 크게 인적자원개발(Human Resource Development, HRD), 인적자원관리(Human Resource Management, HRM) 영역으로 구성되어 있다. 먼저 인적자원개발은 개인개발, 경력개발, 조직개발의 3개의 세부 영역으로

구분되며, 인적자원관리는 직무설계, 인적자원계획, 선발 및 임명, 인적자원정보체제, 보상 및 장려, 근로자 복지후생, 노조근로자 관계로 7개의 영역으로 세분화된다.

<표 4-2>에 나타난 것처럼 모형 구축에 사용할 독립변수는 맥라겐이 제시한 인적자원관리와 인적자원개발 영역의 항목별로 2017년도 한국직업능력개발원의 인적자본 기업패널의 설문 문항과 연계하여 선정한다. 선정 결과 1차적으로 사용할 설문 문항은 총 63개로, 대분류 영역별로는 HRD 영역 32개, HRM 영역 30개, 기타 1개 문항이다.

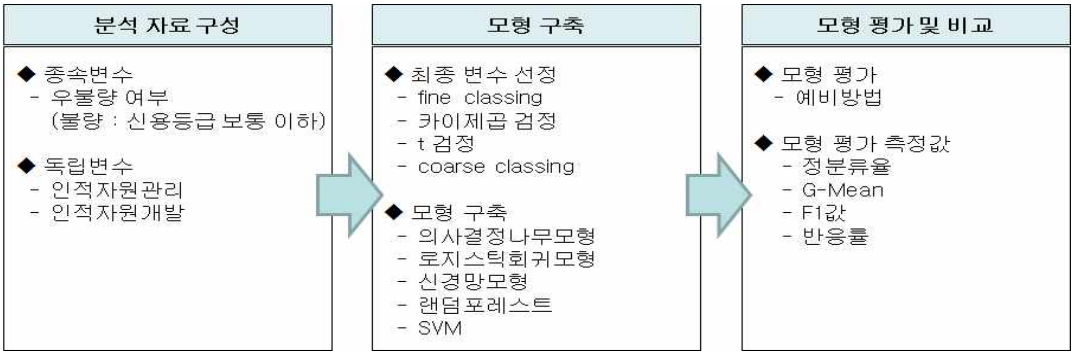
<표 4-2> 인적자원 바퀴모델과 기업패널 설문 문항 연계

인적자원바퀴 모델 영역		설문 문항 연계
대분류	중분류	
인적자원개발(HRD)	개인개발	집체식사내교육훈련여부, 집체식사외교육훈련여부, 인터넷학습여부, 우편통신훈련여부, 국내연수여부, 해외연수여부, OJT여부
	경력개발	사내공모제여부, 경력개발제도여부, 교육훈련휴가제여부, 멘토링또는코칭여부, 학원수강료지원여부, 국내대학등록금지원여부, 국내대학원등록금지원여부, 해외대학원학위과정지원여부, 직무순환여부, 국가자격수당제도유무, 국가기술자격수당제도유무, 공인민간자격수당제도유무, 민간자격수당제도유무, 고용노동부인증사내자격자격제도운용여부, 해외자격수당제도유무, 사내자격자격제도운용여부, 사내자격수당제도유무
	조직개발	균형평가표(BSC)_실시여부, 목표에의한관리(MBO)_실시여부, 역량평가_실시여부, 리더십평가_실시여부, 다면평가_실시여부, 승계계획_실시여부, QC(품질분임조)_실시여부, 6-시그마_실시여부
인적자원관리(HRM)	조직직무설계	직무분석여부, 직무재설계, 직무표준화, 직급단순화, 직급폐지, 직급다단계화, 직급직책분리, 직급직책연계
	인적자원계획	HR전담조직유무, 매년인력계획수립여부
	선발및임명	사내공모제충원비율, 전환배치여부
	인적자원정보체계	인사정보시스템발전단계, 인사정보시스템운용여부
	보상및장려	호봉제여부, 연봉제여부, 직무급여부, 직능급여부, 개인성과급여부, 팀성과급여부, 사업부성과급여부, 전사성과급여부, 이윤배분제도여부, 우리사주제도여부, 스톡옵션여부
	복리후생	사원1년차복리후생수준, 과장1년차복리후생수준, 부장1년차복리후생수준, 선택적복리후생여부
	노사관계	노사관계
기타	기업규모	기업규모

2. 모형 구축 절차

모형 구축은 “분석 자료 구성, 모형 구축, 모형 평가 및 비교” 3단계로 진행된다. 첫째, 인적자본 기업패널 설문조사와 ㈜NICE신용정보의 신용평가등급 자료를 이용하여 모형 구축용 자료 세트를 구성하고 우량과 불량을 정의한다. 둘째, 모형 구축 단계에서는 변수를 선택하고, 신용평가모형을 구축한다. 마지막 단계는 구축된 모형을 평가하고 비교하는 단계로써, 모형 평가 방법은 예비방법을 적용하며, 예측 성능 비교를 위한 측도로는 정분류율 G-Mean, F1값, 반응률을 이용한다.

[그림 4-1] 모형 구축 단계



Ⅴ. 분석 결과

1. 변수 선택 및 기초 분포

독립변수 선정의 첫 번째 단계는 계급세분화를 수행한 후 KS통계량이 0.1 이상인 값을 가진 변수와 이를 보완하기 위해 원천자료(raw data)가 범주형인 경우는 카이제곱 검정과 연속형인 경우는 t-검정을 이용하여 p-값이 0.05 이하인 변수를 선택한다. 세 가지 방법에 의해 1차적으로 선정된 독립변수는 최초 변수 63개 중 32개를 제외한 총 31개로 다음의 <표 5-1>과 같다. 여기에서 KS 통계량은 각 독립변수의 범주별 누적 우량 비율과 누적 불량 비율을 구한 후

각 범주 또는 구간별 누적 우량 비율과 누적 불량 비율의 차이 값 중 가장 큰 값인데, 이 값이 클수록 변별력이 크다는 것을 의미한다.

KS통계량 기준으로 변별력이 가장 큰 독립변수는 기업 규모(0.27), 해외연수 여부(0.21), 인터넷학습 여부(0.20), 국내 대학원 등록금 지원 여부(0.20), 직무순환 여부(0.19), 우편통신훈련(0.17) 등의 순으로 나타나고 있어 이 변수들이 종속변수인 우불량 여부에 많은 영향을 주고 있음을 유추할 수 있다.

<표 5-1> 1차 변수 선택 결과

변수	K-S 통계량	p값	1차 선택	변수	K-S 통계량	p값	1차 선택
기업규모	0.27	<.0001	O	6-시그마_실시여부	0.03	0.5533	
집체식사내교육훈련여부	0.02	0.3062		직무분석여부	0.04	0.4967	
집체식사외교육훈련여부	0.07	0.1492		직무재설계	0.06	0.1579	
인터넷학습여부	0.20	0.0012	O	직무표준화	0.02	0.8905	
우편통신훈련여부	0.17	0.0012	O	직급단순화	0.07	0.0017	O
국내연수여부	0.11	0.0412	O	직급폐지	0.04	0.6309	
해외연수여부	0.21	<.0001	O	직급다단계화	0.03	0.2055	
OJT여부	0.00	0.9768		직급직책분리	0.04	0.4818	
사내공모제여부	0.13	0.0077	O	직급직책연계	0.04	0.4818	
경력개발제도여부	0.13	0.0045	O	HR전담조직유무	0.12	0.0216	O
교육훈련휴가제여부	0.08	0.0477		매년인력계획수립여부	0.12	0.0156	O
멘토링또는코칭여부	0.15	0.0093	O	사내공모제충원비율	0.15	0.1199	O
학원수강료지원여부	0.15	0.0099	O	전환배치여부	0.09	0.1214	
국내대학등록금지원여부	0.09	0.0506	O	인사정보시스템_발전단계	0.17	0.0272	O
국내대학원등록금지원여부	0.20	<.0001	O	인사정보시스템운용여부	0.06	0.019	O
해외대학원학위과정지원여부	0.03	0.1454		호봉제여부	0.12	0.0403	O
직무순환여부	0.19	0.0011	O	연봉제여부	0.09	0.0434	O
국가자격수당제도유무	0.03	0.6078		직무급여부	0.06	0.102	
국가기술자격수당제도유무	0.01	0.8027		직능급여부	0.03	0.4505	
공인민간자격수당제도유무	0.05	0.1413		개인성과급여부	0.17	0.0024	O
민간자격수당제도유무	0.01	0.5672		팀성과급여부	0.10	0.0473	O
고용노동부인증사내자격자격제도여부	0.00	0.8644		사업부성과급여부	0.01	0.7736	
해외자격수당제도유무	0.04	0.196		전사성과급여부	0.14	0.0145	O
사내자격자격제도운용여부	0.01	0.8356		이윤배분제도여부	0.10	0.0409	O
사내자격수당제도유무	0.01	0.7011		우리사주제도여부	0.05	0.2465	
균형평가표_실시여부	0.05	0.225		스톡옵션여부	0.02	0.4472	
목표에의한관리_실시여부	0.15	0.0073	O	사원1년차복리후생수준	0.16	0.0011	O

역량평가_실시여부	0.16	0.0012	O	과장1년차복리후생수준	0.16	0.0028	O
리더십평가_실시여부	0.07	0.2035		부장1년차복리후생수준	0.14	0.0146	O
다면평가_실시여부	0.15	0.0059	O	선택적복리후생여부	0.07	0.0767	
승계계획_실시여부	0.06	0.0237	O	노사관계	0.03	0.9122	
품질분임조_실시여부	0.03	0.6369		-			

다음의 <표 5-2>는 독립변수 계급 세분화된 변수값을 성김화에 의해 재범주화하고 범주별 불량률을 정리한 것이다. 표에서 '성김화 범주'는 재범주화된 결과 범주를 나타내는 것인데, 재범주 값이 '0'에서 '1', '2'로 갈수록 불량률이 점차 감소하는 것이 가장 적절한 성김화이다. 표에서 모든 변수에 대해 성김화된 범주가 0에서 2로 갈수록 불량률이 감소하고 있으므로 선택된 변수에 대해 계급화가 잘 이루어졌다는 결론을 내릴 수 있다. 또한 표를 보면 성김화에 의해 2개의 범주로 재범주화가 이루어진 변수는 29개, 3개의 범주로 재범주화된 변수는 기업규모와 사원 1년차 복리후생 수준 2개이다. 특이한 사항은 직급단순화의 경우 다른 변수들과는 달리 이를 실시한 기업들이 실시하지 않은 기업에 비해 불량률이 약 4배가 높다.

<표 5-2> 성김화(coarse classing)에 의한 재범주화 결과

변수	원변수범주	성김화범주	불량률	변수	원변수범주	성김화범주	불량률
인터넷학습 여부	0	0	53.7	기업규모	1	0	55.8
	결측, 1	1	34.1		2	1	32.7
우편통신훈련 여부	0	0	47.8		3	2	22.2
	결측, 1	1	23.1	HR전담조직 유무	0	0	52.8
국내연수 여부	0	0	45.9		1	1	38.6
	결측, 1	1	30.3	매년인력계획 수립여부	0	0	54.7
해외연수 여부	0	0	48.8		1	1	38.8
	결측, 1	1	13.0	사내공모제 충원비율	결측, 1% 이하	0	46.9
사내공모제 여부	0	0	46.5		응답거절, 1% 초과	1	14.6
	1	1	28.4	인사정보시스템 발전단계	결측	0	50.3
경력개발 제도여부	0	0	46.8		1단계 이상	1	33.3
	1	1	27.3	인사정보시스템 운용여부	0	0	68.4
멘토링또는 코칭여부	0	0	50.0		1	1	41.0
	1	1	35.4	호봉제	0	0	47.5

학원수강료 지원여부	0	0	49.4	여부	1	1	35.9
	1	1	34.9	연봉제 여부	0	0	54.1
국내대학등록금 지원여부	0	0	31.8		1	1	39.8
	1	1	45.4	개인성과급 여부	0	0	50.6
국내대학원등록금 지원여부	0	0	48.8		1	1	33.6
	1	1	17.7	팀성과급 여부	0	0	45.6
직무순환 여부	0	0	51.2		1	1	32.4
	1	1	32.9	전사성과급 여부	0	0	49.1
목표에의한관리 실시여부	0	0	52.1		1	1	35.4
	1	1	36.7	이윤배분제도 여부	0	0	45.7
역량평가 실시여부	0	0	58.4		1	1	31.9
	1	1	37.5	사원1년차 복리후생수준	1, 2	0	55.8
다면평가 실시여부	0	0	48.3		3	1	44.1
	1	1	32.1		4, 5	2	23.9
승계계획 실시여부	0	0	19.1	과장1년차 복리후생수준	1, 2, 3	0	47.4
	1	1	44.3		4, 5	1	23.0
직급단순화	결측, 1	0	44.6	부장1년차 복리후생수준	1, 2, 3	0	47.0
	0	1	11.1		4, 5	1	24.6

신용평가모형을 구축하기 위한 훈련용 및 평가용 자료의 종속변수에 대한 분포는 <표 5-3>과 같다. 먼저 모형 구축을 위한 훈련용 자료는 분석대상 표본 총 312개 중에서 우량과 불량의 비율이 각각 57.4%, 42.6%를 차지하고 있고 평가용 자료는 우량과 불량의 비율이 각각 64.7%, 35.3%이다. 실제로 신용평가 시 불량의 정의는 부도, 휴업, 폐업, 대출 연체 등을 사용하는 것이 타당하지만, 본 자료에서는 분석대상 기업들의 휴폐업 정보 등이 존재하지 않아 대리변수로 신용등급을 사용하였다는 한계가 존재한다.

<표 5-3> 훈련용 및 평가용 자료의 종속변수 분포

(단위 : 개, %)

구 분	우량(Y=0)	불량(Y=1)
훈련용자료	179(57.4%)	133(42.6%)
평가용자료	99(64.7%)	54(35.3%)

2. 최종 모형 구축 및 평가

5가지 기계학습을 이용하여 신용평가모형을 구축한 후 우불량 분류의 정확도와 안정성을 평가 결과는 <표 5-4>에 정리되어 있다. 표를 보면 훈련용 자료와 평가용 자료에 대한 각 기계학습모형의 정분류율, G-mean, F1값이 나타나 있는데 이 값은 1에 가까울수록 우불량의 분류 정확도가 높은 모형이다. 그리고 표의 우측 부분에 나타나 있는 (훈련용-평가용)의 값은 훈련용 자료와 평가용 자료에 대해 산출된 정분류율, G-mean, F1값의 차이인데 이는 안정성에 대한 부분으로 이 값은 0에 가까울수록 안정성이 높음을 의미한다.

표에 나타난 결과를 세부적으로 살펴보면 다음과 같다. 첫째 훈련용 자료의 정분률율은 SVM이 1.0으로 100%의 분류 정확도를 나타내고 있는 가운데 신경망모형 0.929, 로지스틱회귀모형 0.721, 의사결정나무모형 0.696, 랜덤포레스트모형 0.606으로 나타나고 있어 전반적으로 모든 모형의 분류 정확도가 높은 것으로 나타났다. 두 번째로 G-mean과 F1값 또한 정분류율과 마찬가지로 SVM, 신경망모형, 로지스틱회귀모형, 의사결정나무모형, 랜덤포레스트모형의 순으로 분류 정확도가 나타나고 있다.

그러나 이에 반해 평가용 자료의 경우는 대체적으로 랜덤포레스트모형, 로지스틱회귀모형, 신경망모형, SVM, 의사결정나무모형의 순으로 분류 정확도가 나타나고 있다. 이는 훈련용 자료에 의해 구축된 신용평가모형에 새로운 자료인 평가용 자료를 적용하였을 때 안정성이 매우 떨어지는 모형이 있다는 것을 말하는데, 이는 훈련용 자료에 의한 평가모형이 과대추정(overestimate) 되었다는 것을 의미한다. 그러므로 훈련용 자료와 평가용 자료의 차이를 통해 과대추정이 되지 않고 모형의 안정성이 높은 모형을 찾을 필요가 있다. 즉, 분류의 정확도가 높으면서도 안정성이 높은 모형이 좋은 모형이다.

표에 나타난 것처럼 모형의 안정성에 대한 결과인 (훈련용-평가용)의 값들을 살펴보면, 랜덤포레스트가 세 가지 측도 모두 0.05 미만의 값을 가지고 있어 차이가 가장 작게 나타나고 있으며, 다음으로 로지스틱회귀모형, 의사결정나무모형, 신경망모형, SVM으로 나타나고 있다. SVM과 신경망모형의 경우 훈련용

자료에서는 90% 이상의 매우 높은 분류 정확도를 나타내고 있지만, 평가용 자료에서는 분류 정확도가 매우 낮게 나타나 안정성 측면에서 사용하기 어려운 모형으로 판단된다. 이결과를 통해 로지스틱회귀모형, 의사결정나무모형, 랜덤포레스트모형이 비교적 분류 정확도와 안정성 측면에서 인적자원 관련 변수를 이용한 기업 신용평가모형으로 사용하는 것이 타당한 것으로 사료된다.

그러나 기계학습을 이용하여 신용평가모형 등 분류를 목적으로 하는 모형을 구축할 때에는 오분류표에 의한 우불량의 단순 분류 정확도 이외에 모형에 의해 예측된 사후확률과 사후확률을 오름차순으로 정렬한 후 일정 구간으로 나누어 각 구간별 실제 불량률인 반응률을 확인하여야 한다. 이유는 실제 등급을 산출한 후 신용등급이 낮은 구간에 해당될수록 불량률이 점차 높아져야 하기 때문이다. 즉 등급이 낮아질수록 불량률의 역전 현상이 관측되지 않아야 한다.

<표 5-4> 우불량 분류 성능 및 안정성 비교

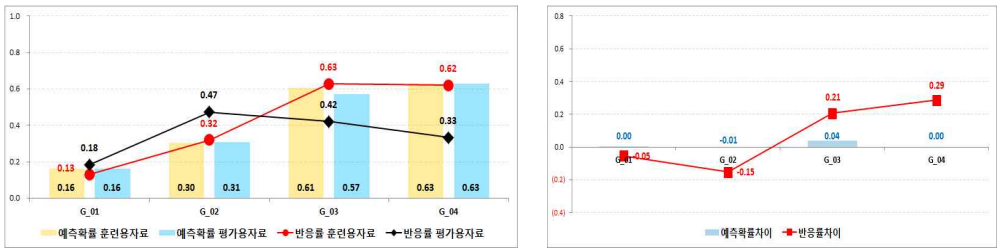
구분		훈련용자료					평가용자료					훈련용-평가용		
		예측 0	예측 1	정분류율	G-mean	F1값	예측 0	예측 1	정분류율	G-mean	F1값	정분류율	G-mean	F1값
의사결정나무	실제 0	123	56	0.696	0.697	0.664	58	41	0.549	0.531	0.430	0.146	0.166	0.235
	실제 1	39	94				28	26						
로지스틱회귀	실제 0	131	48	0.721	0.719	0.684	65	34	0.595	0.562	0.456	0.126	0.157	0.227
	실제 1	39	94				28	26						
신경망	실제 0	170	9	0.929	0.926	0.916	59	40	0.556	0.536	0.433	0.374	0.390	0.483
	실제 1	13	120				28	26						
랜덤포레스트	실제 0	117	62	0.606	0.595	0.539	70	29	0.647	0.616	0.518	-0.041	-0.021	0.021
	실제 1	61	72				25	29						
SVM	실제 0	179	0	1.000	1.000	1.000	76	23	0.595	0.462	0.326	0.405	0.538	0.674
	실제 1	0	133				39	15						

[그림 5-1]과 <표 5-5>는 반응률과 반응률 도표를 나타낸 것이다. 표와 그림에서 G_01은 불량일 사후확률이 낮은, 즉 우량일 사후확률이 높은 구간이며 G_04는 불량일 사후확률이 높은 구간이다. G_01~G_04의 구분은 각 기계학습 알고리즘에 의한 불량일 확률을 산출한 후, 이를 오름차순으로 정렬하고 각 구간별로 기업의 수가 유사하게 배정한다.

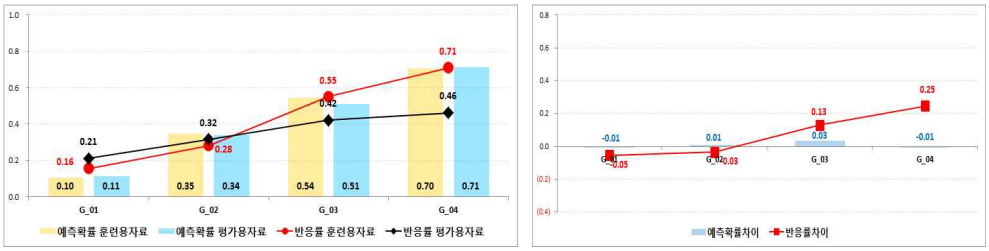
반응률과 반응률 도표의 결과를 정리하면 다음과 같다. 첫 번째로 의사결정나무모형에 대한 불량일 사후확률인 예측확률의 각 구간별 평균은 훈련용 및 평가용 자료에서 모두 G01에서 G04로 갈수록 점차 증가하고 있지만, 반응률은 훈련용 자료는 G03과 G04에서 불량률 역전현상이 발생하고 있고 평가용 자료에서는 G02에서 G04로 갈수록 점차 감소하고 있다. 그러므로 불량일 사후확률이 높은 구간에서는 불량에 대한 예측력이 다소 떨어지며, 안정성 또한 높지 않음을 있음을 알 수 있다.두 번째로 로지스틱회귀모형에 대한 예측확률의 각 구간별 평균은 훈련용 및 평가용 자료에서 모두 G01에서 G04로 갈수록 점차 증가하고 있고, 반응률 또한 훈련용 과 평가용 자료 모두 G01에서 G04로 갈수록 점차 증가하는 추세를 보이고 있다. 그러므로 좋은 예측 판별력을 가지고 있는 것으로 판단된다. 세 번째로 신경망모형에 대한 예측확률은 훈련용 및 평가용 자료에서 모두 G01과 G02에서는 0으로 나타나고 있으며 G03과 G04에서 급격히 증가하고 있다. 반응률은 훈련용 자료에서는 G01에서 G02 구간에서 역전현상이 발생한 이후 급격히 증가하며, 평가용 자료에서는 완만하게 증가하는 모습을 보이고 있다. 이와 같은 결과를 통해 안정적인 예측판별력을 가지고 있지 않은 것으로 사료된다. 네 번째로 랜덤포레스트모형에 대한 예측확률은 훈련용 및 평가용 자료에서 모두 G01에서 G04로 갈수록 점차 증가하고 다. 반응률은 훈련용 자료에서는 G03과 G04에서 역전현상이 발생하였고, 평가용 자료에서는 비교적 안정적으로 증가하는 추세를 보이고 있다. 마지막으로 SVM의 경우 훈련용과 평가용 자료에 대한 예측확률과 훈련용 자료에 대한 반응률은 G01에서 G04로 갈수록 점차 증가하는 모습이지만 평가용 자료에 대한 반응률은 G03과 G04에서 역전현상이 발생하고 있다. 그러므로 불량일 가능성이 높은 구간에서의 예측 성능이 다소 떨어지고 있는 것으로 판단된다.

각 모형의 안정성은 훈련용 자료와 평가용 자료에 의한 예측확률과 반응률의 각 구간별 차이를 통해서 확인할 수 있는데 그림 (a)~(e)의 우측 그림이 이에 해당한다. 그림을 보면 훈련용 자료와 평가용 자료의 차이는 랜덤포레스트모형, 로지스틱회귀모형, 의사결정나무모형, 신경망모형, SVM의 순으로 차이가 점차 커짐을 알 수 있다.

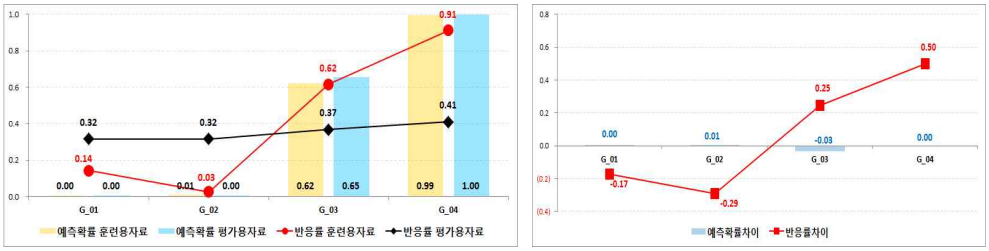
[그림 5-1] 반응률 도표 및 안정성 확인



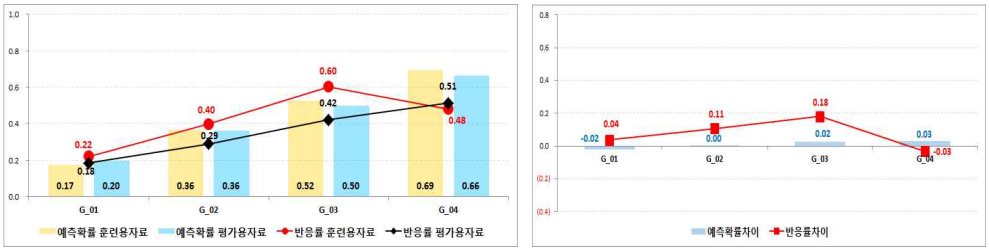
(a) 의사결정나무모형의 예측확률, 반응률 및 안정성



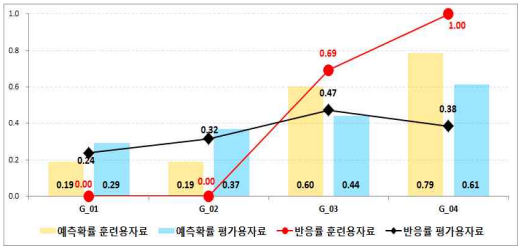
(b) 로지스틱회귀모형의 예측확률, 반응률 및 안정성



(c) 신경망모형의 예측확률, 반응률 및 안정성



(d) 랜덤포레스트의 예측확률, 반응률 및 안정성



(e) SVM의 예측확률, 반응률 및 안정성

<표 5-5> 반응률 비교

모형	구간	예측확률		반응률		예측확률 차이	반응률 차이
		훈련용자료	평가용자료	훈련용자료	평가용자료		
의사결정 나무	G_01	0.16	0.16	0.13	0.18	0.00	-0.05
	G_02	0.30	0.31	0.32	0.47	-0.01	-0.15
	G_03	0.61	0.57	0.63	0.42	0.04	0.21
	G_04	0.63	0.63	0.62	0.33	0.00	0.29
로지스틱 회귀	G_01	0.10	0.11	0.16	0.21	-0.01	-0.05
	G_02	0.35	0.34	0.28	0.32	0.01	-0.03
	G_03	0.54	0.51	0.55	0.42	0.03	0.13
	G_04	0.70	0.71	0.71	0.46	-0.01	0.25
신경망	G_01	0.00	0.00	0.14	0.32	0.00	-0.17
	G_02	0.01	0.00	0.03	0.32	0.01	-0.29
	G_03	0.62	0.65	0.62	0.37	-0.03	0.25
	G_04	0.99	1.00	0.91	0.41	0.00	0.50
랜덤 포레스트	G_01	0.17	0.20	0.22	0.18	-0.02	0.04
	G_02	0.36	0.36	0.40	0.29	0.00	0.11
	G_03	0.52	0.50	0.60	0.42	0.02	0.18
	G_04	0.69	0.66	0.48	0.51	0.03	-0.03
SVM	G_01	0.19	0.29	0.00	0.24	-0.10	-0.24
	G_02	0.19	0.37	0.00	0.32	-0.18	-0.32
	G_03	0.60	0.44	0.69	0.47	0.16	0.22
	G_04	0.79	0.61	1.00	0.38	0.17	0.62

오분류표에 의한 정분류율, G-mean, F1값, 예측확률과 반응률에 대한 결과를 종합할 때, 우불량에 대한 분류 성능, 각 등급 구간별 예측 성능과 안정성 측면에서 로지스틱회귀모형이 본 자료에 대해 가장 우수한 기계학습 알고리즘이라는 결론을 내릴 수 있다. 그리고 우불량에 대한 분류 성능과 불량일 사후확률이 높은 구간에서의 예측 성능이 다소 떨어지지만 랜덤포레스트모형이 인적

자원 관련 변수를 이용한 기업 신용평가모형 구축을 위한 기계학습 알고리즘으로 추천할 수 있는 것으로 판단된다.

Ⅵ. 결론

본 논문은 2017년도 한국직업능력개발원 인적자본 기업패널에 응답한 기업체 표본 474개 중 2017년의 기업 신용등급이 있는 465개를 대상으로 대표적으로 많이 사용하는 기계학습 알고리즘인 의사결정나무모형, 로지스틱회귀모형, 신경망모형, 랜덤포레스트모형, SVM에 적용하여 기업 신용평가모형을 구축하였을 때, 예측 성능과 안정성 측면에서 가장 우수한 모형이 무엇인지를 확인한 후 인적자원 관련 변수를 신용평가모형에의 적용 가능성을 확인해 보는 것이다. 그리고 구축된 모형의 비교 및 평가는 정분류율, G-mean, F1측도, 반응률이며, 모형 구축 도구는 SAS 9.4와 R을 이용하였다.

인적자원 관련 변수를 5가지의 기계학습 알고리즘에 적용하여 기업 신용평가모형을 구축한 결과를 요약하면 다음과 같다. 첫째, 오분류표를 이용한 정분류율, G-mean, F1값을 비교한 결과, 분류 성능과 안정성 측면에서 로지스틱회귀모형, 의사결정나무모형, 랜덤포레스트모형이 다른 두 모형에 비해 상대적으로 좋은 모형인데 특히 로지스틱회귀모형이 가장 좋은 모형으로 확인되었다. 둘째, 불량일 사후확률인 예측확률과 사후확률을 오름차순으로 정렬한 후 산출하는 실제 불량의 비율인 반응률을 살펴보았을 때, 로지스틱회귀모형과 랜덤포레스트모형이 다른 세 가지 모형에 비해 비교적 좋은 예측 성능과 안정성을 가지고 있는데, 로지스틱회귀모형이 가장 우수한 예측 성능과 안정성을 가지고 있었다. 그러므로 인적자원 관련 변수를 이용하여 기업 신용평가모형을 구축할 경우 로지스틱회귀모형이 가장 좋은 방법으로 판단된다. 그러므로 오분류표와 반응률의 결과를 통해 본 논문에서 사용한 자료에 대해서는 기계학습 알고리즘 중 로지스틱회귀모형을 이용하여 기업 신용평가모형을 구축하는 것이 가장 타당하다는 결론을 내릴 수 있다.

물론 위의 결론은 본 논문에서 사용한 자료를 이용할 경우에 한정되는 것으로써, 다른 자료 이용, 자료의 표준화 방법 등 정제 기법 변경, 신경망모형의 은닉층 및 노드 개수 조정 랜덤포레스트나 SVM모형의 다양한 세부 옵션 선택 등에 따라 다른 결과가 나타날 가능성을 배제할 수 없다.

이러한 문제점에도 불구하고 본 논문에서의 의의는 다음과 같다. 인적자원 관련 변수가 기업의 신용등급에 영향을 주고 있으며 다양한 기계학습 알고리즘을 사용하여 기업 신용평가가 가능하고 재무정보 등 객관적인 자료를 이용하지 않고 객관성이 다소 부족한 것으로 인식되고 있는 설문조사만으로도 신용평가가 가능하다는 것을 확인하였다는 점이다. 마지막으로 기존의 기업 인적자원 요소들을 이용한 기업 신용평가모형 구축의 연구를 더욱 발전시켜 인적자원변수를 정량화하여 기업 신용평가모형 구축에의 적용 가능성과 방법론을 실증분석을 통해 탐색했다는 점이다. 하지만 여러 가지 측면에서 한계를 내포하고 있는 것이 사실이다.

본 논문이 가지는 한계점과 향후 연구 방향은 다음과 같다. 첫째, 본 논문에서 사용된 인적자원관련 활동 변수들은 일반화를 시키기에는 표본의 크기가 부족하다는 것이다. 그러므로 더욱 큰 표본을 이용하거나 전체 패널자료를 풀링(pooling)하여 사용하는 등의 방법을 고려할 필요가 있다. 둘째 투입된 인적자원 관련 독립변수들 역시 기업의 인적자원 활동을 충분히 평가할 수 있는지와 맥라겐의 인적자원 바퀴모형 이외의 다른 즉 BSC, 7S 등 기업 인적자원을 평가하고 설명하는 다른 모형을 통해 변수를 선정하는 방법에 대해서도 고려해야 할 필요성이 있다. 따라서 향후의 연구를 위해서는 표본이나 독립변수에 대한 보다 치밀한 논의가 필요하다. 셋째, 만약 분석을 위한 자료의 양이 충분하다면 기업의 규모 및 업종에 따른 세분화된 모형의 구축 연구를 고려해보아야 한다. 즉 기업의 규모와 업종에 따른 인적자원개발 및 관리 기법, 기업의 문화에 차이가 있을 수 있으므로 규모와 업종에 따라 어떠한 인적자원관련 항목들이 영향을 주는지 부가적인 연구가 필요하다.

참고문헌

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, 한국정보통신학회 논문지, pp 63-69.

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

김미숙, 김안국, 이기성, 김재구, 이석재, 김태준(2005). 인적자원개발 우수기관 인증제도 도입을 위한 심사지표 및 메뉴얼 개발 연구, 한국직업능력개발원.

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

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

김효진(2018). 머신러닝에 대한 이해, 주택금융리서치.

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

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

박주완, 송창길(2015). 인적자원 변수를 이용한 기업신용평가모형 구축에 관한 연구, 인적자본 기업패널 학술대회 논문집.

박주완(2017). 데이터마이닝 기법을 이용한 소상공인 신용평가모형 구축에 관한 연구, KOREG Research, 제5권 1호.

박주완(2018). 소상공인 신용평가모형 구축에 관한 연구-설문조사 자료를 이용

하여-, 중소기업금융연구, 제350호, pp 38-65.

박주완(2019). 빅데이터 분석 기법을 이용한 소상공인 신용평가모형 구축 연구, 신용보증재단중앙회.

블로터(2019). http://www.bloter.net/archives/351562, 카카오뱅크 시스템은 진화 중.

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

성웅현(2001). 응용 로지스틱 회귀분석-이론, 방법론, SAS 활용-, 탐진.

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

신윤제(2016). 기계학습을 활용한 신용평가모형의 개발 - 신용정보 부족군을 대상으로, NICE Credit Insight Issue Report, NICE평가정보 CB연구소.

연합인포믹스(2018). http://news.einfomax.co.kr/news/articleView, 카뱅·케뱅, 자체 신용평가 모형 구축 박차…당국도 힘 실어.

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

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 학습 방법 인공신경망모형과의 성과 비교, 경영학연구, pp 109-144.

이영섭 역(2003). 데이터마이닝 Cookbook, 교우사.

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.

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), 95-127.

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.

Kubat, M., Holte, R. C., and Matwin, S.(1998). Machine Learning for the Detection of Oil Spills in Satellite Radar Images, Machine Learning, 30, pp 195 – 215.

Lantz. B.(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).

Ohlson, J. A.(1980). "Financial Ratios and the Probabilistic Prediction of Bankruptcy," Journal of Accounting Research(spring), pp 109-131.

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

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.

연구논문

소상공인 신용평가를 위한 기계학습 적용 연구

배진성* 박주완

본 논문의 목적은 다양한 기계학습으로 소상공인 신용평가모형 구축한 후 예측 성능이 가장 우수한 모형이 무엇인지를 알아보는 것이다. 이를 위해 지역신용보증재단에서 보증을 받은 차주의 보증을 받을 당시 입력된 자료를 이용하였다.

분석 결과를 정리하면 다음과 같다. 다양한 기계학습 알고리즘 중 로지스틱회귀모형에 의한 우불량 판별의 정분류율, G-mean, F1측도값이 가장 양호하였으며, 반응률을 보았을 때 우량일 사후확률이 높은 구간에서 낮은 구간으로 갈수록 불량률이 점차 증가하며, 역전현상이 발견되지 않고, 서열화가 잘 이루어지고 있음을 확인하였다. 또한 훈련용 자료와 평가용 자료에 의해 산출된 다양한 평가 측도를 비교하였을 때, 값들의 차이가 거의 나지 않아 모형의 안정성이 확인되었다.

* 신용보증재단중앙회 교육연구부 선임연구위원(경제학박사)

Ⅰ. 서론

4차 산업혁명(the fourth industrial revolution) 시대에 접어들면서 금융 산업 분야에서는 금융 리스크 관리 능력 제고 등을 위해 빅데이터나 인공지능의 활용을 위한 기술 개발에 많은 투자를 하고 있다(박주완, 2019). 실제로 국내외 금융 산업에서는 신용평가 및 위험관리, 업무 자동화, 고객 금융서비스, 트레이딩, 준법감시 등의 업무에 빅데이터와 기계학습을 적용하는 사례들이 점차 증가하고 있다(김효진, 2018).

그러나 소상공인을 대상으로 한 신용평가는 기업 신용정보를 대변하는 재무제표와 대차대조표 등 정량적이고 객관적인 분석 자료의 부족이라는 한계로 말미암아 빅데이터 기반 기계학습 이용은 고사하고 전통적인 통계 기법을 적용한 신용평가모형 구축도 쉽지 않은 현실이다(윤상용 외, 2016).

소상공인을 대상으로 한 신용보증에 의한 보증부대출이 점차 증가하고 있는 현재의 상황에서 소상공인 신용평가의 신뢰성과 정확성에 대한 요구는 점차 커지고 있다(박주완 외, 2017). 그러나 소상공인을 대상의 신용보증 기관인 지역신용보증재단 등에서는 재무정보 등이 비교적 잘 갖추어져 있는 일부 소상공인에게는 자체 소상공인신용평가모형을 이용하여 신용등급을 산출 후 일반보증 상품용 신용보증서를 제공하고 있지만, 신용보증의 대부분을 차지하고 있는 특례 및 협약보증 상품에 대한 신용보증서 발급 시 국내 신용정보회사에서 제공하는 대표자 개인신용등급에 의존하는 실정이다(박주완, 2019).

그러므로 4차 산업혁명 시대를 맞이하여 소상공인의 신용보증 대출 시 빅데이터와 기계학습을 활용한 신용평가 요구가 점차 증가할 것으로 예측되는 가운데, 소상공인을 대상으로 한 신용평가모형 구축에 있어 빅데이터와 기계학습을 적용하기 위한 노력이 어느 정도인지 고민해 볼 필요가 있다.

기존에 소상공인을 대상으로 한 신용평가모형 구축 연구는 기계학습 등의 적용보다 비재무적인 자료를 이용하는 등 사용 가능한 자료 활용에 대한 연구가 주를 이루고 있다(박주완, 2018). 실제로 소상공인 신용평가에 있어 기법 측

면의 연구는 심사역의 주관적인 경험에 의존하여 점수를 산출하는 AHP(Analytic Hierarchy Process) 기법에 대한 연구가 주를 이루고 있다(윤종식 · 권영식, 2007). 이외에도 개인정보보호 등 정보의 사용 제약과 전통적 신용평가 기법의 고착화 등으로 인하여 기계학습을 적용한 연구는 한계가 많다(신윤재, 2016).

이에 본 논문에서는 16개 지역신용보증재단이 보유한 차주의 신용보증을 받을 당시의 자료를 현재 많이 활용되고 있는 기계학습 분류 기법에 적합하여 (adjusted) 소상공인 신용평가모형을 구축하고 예측 성능을 비교한 후, 기계학습을 이용한 소상공인 신용평가모형의 구축 가능성을 확인하고자 한다.

본 논문은 과거 비재무 자료 활용과 AHP 기법 적용 측면의 소상공인 신용평가모형 연구 틀에서 벗어나, 4차 산업혁명 시대에 맞추어 기계학습 분류 기법을 이용하여 소상공인 신용평가모형 구축의 가능성을 연구한다는 점에서 기존 연구들과 차별성이 있을 것으로 사료된다.

본 연구에서 사용하고자 하는 기계학습 분류 기법은 대표적으로 잘 알려져 있는 의사결정나무모형(decision tree), 로지스틱회귀모형(logistic regression), 신경망모형(neural network), 서포트벡터머신(support vector machine, SVM), 랜덤포레스트모형(random forest)이다. 모형의 평가는 예비 방법(holdout method), 구축된 모형의 분류 정확도를 평가하기 위한 측도(measure)로는 정분류율, G-mean, F1 측도, 반응률(percent of response)을 이용한다. 자료를 분석하고 모형을 구축하기 위한 통계 프로그램은 SAS9.4와 R3.5.1 버전을 이용하는데, 이 때 R로는 기계학습을 수행할 수 있는 함수 및 라이브러리인 “glm, rpart, nnet, rf, kernlab”을 이용하여 분류 모형을 생성한다.

논문의 구성은 다음과 같다. Ⅱ장에서는 신용평가모형 관련 실제 사례와 연구 문헌들을 고찰하고, Ⅲ장에서는 모형 구축에 사용한 알고리즘, 데이터 정제와 모형 평가 방법을 설명한다. Ⅳ장은 소상공인 신용평가모형을 구축하기 위한 표본 및 변수, 모형 구축 과정에 대해 설명하며, Ⅴ장은 모형을 구축하여 예측 성능을 비교 및 평가 후 예측 성능이 가장 우수한 모형을 선택한다. 마지막으로 Ⅵ장에서는 결론에 대해 고찰한다.

Ⅱ. 선행 연구 고찰

1. 신용평가 시 기계학습 적용 실제 사례

금융 산업에서 빅데이터와 기계학습을 신용평가나 대출심사 등에 적용하는 국내외 사례를 살펴보면 다음과 같다. 먼저 해외 사례로는 Kabbag, Zest Finance, 요코하마은행과 지바은행 등이 있다. Kabbage는 소상공인 신용평가 시 기존의 재무자료 이외의 배송, 회계, 인터넷 자료 등을 기계학습에 적용하여 소상공인의 신용평가를 수행하고 있다. Zest Finance는 전통적인 신용정보 외에 직장정보, 고정수입, 인터넷 포스팅 내용 등이 포함된 7만개가 넘는 변수를 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 2017년에 신용도 판단이 어려운 사회 초년생과 중금리 대출 고객들을 대상으로 기계학습을 적용한 신용평가시스템 개발을 완료하였다(서울경제신문, 2017). 케이뱅크는 가계나 자영업자의 신용대출 심사 시 KT의 통신요금 납부 실적, 비씨카드 신용카드 결제 정보를 중금리 대출 심사에 적용하여 연체율 감소 효과를 거두고 있다(연합인포믹스, 2018). 카카오뱅크는 이상 거래를 탐지하기 위해 지도학습 기계학습 모형을 활용하고 있는데, 이는 다양한 사기 거래 데이터를 통해 '정상 데이터와 달리 사기 데이터는 이런 특성이 있어'라는 걸 학습시킨 후 이상 거래를 탐지하는 방식이다(블로터, 2019). 이와 같은 사례들을 통해 국내외 여러 기관에서의 신용평가 시 빅데이터와 기계학습 이용에 대한 관심과 중요성이 점차 높아지고 있음을 유추할 수 있다(박주완, 2019). 그러나 사례를 통해서 살펴보았을 때에도 여전히 소상공인에 대한 특화된 신용평가모형 개발은 부족하다.

2. 기계학습 이용 신용평가모형 연구 사례

기업 부도 예측 연구의 초창기에는 다변량판별분석(Altman, 1968)과 로지스틱회귀모형(Ohlson, 1980) 등의 전통적인 통계 방법론을 적용하는 연구가 주를 이루었는데, 이후 신경망모형, SVM 등 다양한 데이터마이닝 기법들을 이용하여 예측 성능을 향상시키는 방향으로 발전하여 왔다(강신형, 2016). 본 절에서는 기계학습들을 이용한 부실기업 예측 및 기업신용평가모형 구축에 대한 연구 사례를 살펴보고자 한다.

먼저 기업의 부도 예측 시 인공신경망모형의 우수성을 검증한 연구 사례로는 이건창(1993), 박정윤(2000), 전성빈 · 김영일(2001), 정유석(2003) 등이 있다. 이건창(1993)은 다변량판별분석, 인공신경망모형으로 기업 부도 예측을 수행한 후 이를 비교하여 인공신경망 모형의 예측 성능이 우수하다고 하였다. 박정윤(2000)은 1991~1996년 자료로 기업 부도 예측을 실시한 결과 MDA모형, 확률모형, 인공신경망모형 중 인공신경망모형의 예측 성능이 가장 우수하다고 하였다. 전성빈 · 김영일(2001)은 기업 부도 예측 시 인공신경망모형의 예측 성능이 가장 우수하였고 다변량판별분석, 로지스틱회귀모형 등의 분류 정확도는 비슷한 수준이라고 하였다. 정유석(2003)은 로지스틱회귀모형, 다변량판별분석, 인공신경망모형을 이용하여 부도 기업을 예측한 결과 인공신경망모형의 예측력이 가장 우수하다고 하였다.

다음은 의사결정나무모형, 로지스틱회귀모형과 SVM의 예측 성능이 우수함을 실증 분석한 연구 사례이다. 조준희 · 강부식(2007)은 코스닥기업의 부도 예측 시 의사결정나무모형이 신경망모형이나 로지스틱회귀모형 보다 좋은 예측 성능을 가지고 있다고 하였다. 박주완 · 송창길(2015)은 인적자본기업패널과 NICE 자료를 이용하여 로지스틱회귀모형, 신경망모형, 의사결정나무모형으로 소기업 이상에 대해 신용평가모형을 구축한 결과 로지스틱회귀모형의 예측 성능이 가장 우수함을 실증 분석하였다. 윤종식 · 권영식(2007)은 소상공인 부실예측모형 연구에서 로지스틱회귀모형, 다변량판별분석, CART, C5.0, 신경망 모형, SVM 중 SVM의 예측 성능이 가장 우수함을 보였다. 박주완 외(2017)는 소

상공인 신용평가 시 로지스틱회귀모형이 의사결정나무모형이나 신경망모형 보다 예측 성능이 우수하며, 계급불균형 자료를 이용하여 신용평가모형 구축 시 예측 성능이 저하될 수 있다고 밝히고 있다.

마지막으로 앙상블 기법을 이용한 연구 사례이다. 김승혁・김종우(2007)는 SOHO 부도 예측 시 수정된 배깅 예측자(Modified Bagging predictors)[1]가 인공신경망과 배깅예측자(Bagging predictors) 보다 예측 성능이 향상된다고 하였다. 김명종・강대기(2010)는 기업 부실 예측을 위해 인공신경망과 부스팅 인공신경망 앙상블 기법을 적용한 결과 앙상블 학습은 기업 부실 예측 문제에 있어 전통적인 인공신경망을 개선할 수 있다고 하였다. 김성진・안현철(2016)은 1,295개 국내 상장 기업을 대상으로 기업신용평가모형 구축 시 다변량판별분석, 인공신경망, SVM, 랜덤포레스트모형을 비교한 결과 랜덤포레스트모형의 예측 성능이 가장 우수함을 보였다.

이상의 연구를 살펴보면 연구자에 따라 결과가 상이하게 나타나고 있는데 이는 분석 자료에 따른 차이일 가능성이 높다. 여기에서 중요한 함의는 특정한 하나의 알고리즘이 가장 우수하다는 결론을 내릴 수 없다는 것이다. 그리고 연구 대상은 대부분 일정 규모 이상의 기업으로 소상공인 등 규모가 매우 작은 기업의 부도 예측 등에 대한 연구는 자료의 부족 등으로 인해 연구가 많지 않다는 것이다. 그러므로 소상공인을 대상으로 기계학습을 이용한 신용평가모형 구축 연구는 합당한 시도이며 의미가 있는 작업으로 사료된다.

Ⅲ. 기계학습 알고리즘 및 모형 평가

1. 기계학습 알고리즘

기계학습 관점에서 신용평가모형의 개념을 살펴보면 다음과 같다. 차주의 우불량 여부를 판별하고 신용도를 예측하기 위한 신용평가모형은 기계학습 관

1) 부스트랩(bootstrap) 방법으로 다수의 모델을 만들고 평균 이상의 예측 정확도를 가지는 모형들만을 선택해 투표(voting)하는 방법

점에서 지도학습 중에서 분류(classification) 모형이다(오미애 외, 2017). 지도학습을 위한 자료에는 종속변수와 독립변수가 필요하다. 대표적인 지도학습 모형으로는 선형회귀모형(linear regression), 로지스틱회귀모형, 의사결정나무모형, 신경망모형, 랜덤포레스트모형, SVM 등이 있다(박주완 외, 2017).

본 연구에서 사용할 분류 모형 구축 알고리즘은 보편적으로 많이 사용하고 있는 로지스틱회귀모형, 의사결정나무모형, 신경망모형, 랜덤포레스트모형, SVM 5가지인데, 본 절에서는 연구에 사용된 5가지 모형에 대해 고찰한다.

로지스틱회귀모형은 종속변수의 계급이 0과 1 두 가지 값을 가지고 관심의 대상이 되는 계급이 1이 될 확률을 예측하는 모형이다(Hosmer · Lemeshow, 2000). 실제로 현업에서 신용평가모형을 구축할 때 로지스틱회귀모형이 가장 많이 사용되고 있는데, 이유는 다음과 같다. 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). 본 연구에서는 통계 프로그램인 R에서 제공하는 기본 함수인 “glm”을 사용한다.

의사결정나무모형은 나무 구조로 도표화하여 의사결정 규칙(decision rule)을 찾고 분류와 예측을 수행하는 방법으로 대표적인 알고리즘으로는 CHAID (chisquared automatic interaction detection), C5.0, CART(classification and regression Trees)가 있다. 이 모형의 장점은 나무구조로 분류 규칙을 도표화하기 때문에 이해가 쉽다는 것이다. 또한 연속형과 범주형 자료를 동시에 다룰 수 있고, 결측치를 분석에 활용할 수 있다. 그리고 교효효과(interaction effect) 해석이 쉬우며, 선형성, 정규성, 등분산성 등 통계적인 가정이 필요하지 않다. 그러나 최적의 의사결정나무를 찾는 것은 쉽지 않으며, 매우 세밀하게 분류 및 예측을 수행할 경우 과대적합(over-fitting)의 가능성이 높아 새로운 자료에 대한 일반화 성능이 좋지 않을 수 있다(최종후 · 진서훈, 2005). 본 연구에서는 R 라이브러리 중 CART를 수행하는“rpart”를 사용한다.

신경망모형은 과거의 경험이나 지식을 습득하여 모형화하면서 오류 최소화 과정을 통해 예측 및 분류를 수행하며, 어떠한 통계적인 분포도 가정하지 않는다.

신경망모형 중 가장 널리 사용되는 다층인식자(multi-layer percept -ron, MLP) 신경망은 입력층, 은닉층, 출력층으로 구성되어 있고 노드를 통해 연결되는 구조이다. 먼저 입력층을 통해 자료를 입력받고, 은닉층에서는 입력층으로부터 전달되는 변수값들의 선형결합(linear combination)을 비선형함수로 처리하여 출력층 또는 다른 은닉층으로 전달하여, 최종적으로는 출력층을 통해 예측 결과를 산출한다(강창완 외, 2007). 신경망모형의 장점은 비선형적인 관계를 찾아낼 수 있고 예측의 정확성이 매우 높다는 것이다. 그러나 과대적합(over-fitting)하는 경향이 있으며 결과의 해석이 매우 어렵다(Ripley, 1996). 본 연구에서는 R 라이브러리 중 “nnet”를 사용한다.

랜덤포레스트모형은 의사결정나무를 확장한 개념으로 앙상블(ensemble) 모형 중 하나이다. 랜덤포레스트는 다수의 의사결정나무모형을 만들어 예측 성능을 높이는 방법이다(Yoo, 2015). 일반적으로 의사결정나무모형은 특이값(outlier)을 하나의 노드로 구성할 수 있기 때문에 편향된 분포에 민감하지 않지만, 깊이가 깊어질수록 과적합의 위험이 커진다. 이와 같은 위험을 최소화하여 예측 성능을 높이고자 고안된 기법이 랜덤포레스트모형이다(Breiman, 2001). 이 모형의 장점은 트리의 다양성을 극대화하여 예측력이 우수하고 많은 나무의 예측 결과를 종합하기 때문에 안정성이 높다는 것이다. 그러나 의사결정나무모형의 장점인 설명력은 없다. 본 연구에서는 랜덤포레스트 모형을 수행할 수 있는 R 라이브러리 중 “randomForest”를 사용한다.

SVM은 분류 및 예측 시 가장 보편적으로 사용되고 있는 기계학습 알고리즘 중 하나이다. 일반적으로 SVM은 벡터 공간에 존재하는 학습 데이터가 어떠한 그룹에 속하는지를 분류하기 위한 선형 분류자(linear classifier)를 찾는 기법이다(Lantz, 2015). 이 모형은 다양한 학습 데이터의 분포에서도 정확도 측면에서 우수하다는 장점이 있지만, 직관적인 해석이 불가능하다는 단점이 있다. 이와 같은 이유로 결과의 해석보다는 분류의 정확도가 중요한 경우 SVM을 사용하는 경우가 많다(김의중, 2016). 본 연구에서는 R 라이브러리 중“kernlab”을 사용한다.

2. 변수 정제 및 모형 평가 기법

성공적인 모형 구축을 위해서는 양질의 원천 자료(raw data)가 확보되어야 하며 다양한 방법을 이용하여 분석 가능한 형태로 데이터를 정제(cleaning)하여야만 한다. 데이터 정제 시에는 기본적으로 분석 변수에 대해 결측치(missing value)나 특이값(outlier value) 등에 대한 사항을 파악한 후, 이를 제거하거나 대체하는 작업 등을 거치게 된다. 그리고 신용평가모형 구축을 위한 독립변수를 선택할 때 통계적으로 선택된 결과를 바탕으로 대출 시 비즈니스 관점의 부합성을 고려하여 최적의 변수를 조합하여야 한다(신용보증재단중앙회, 2017).

본 연구에서는 신용보증재단중앙회(2017)과 박주완(2018)에서 사용한 변수 정제 및 선택 기법을 이용한다. 변수 정제 및 선택 방법을 구체적으로 설명하면 다음과 같다. 먼저 독립변수와 불량과의 관계를 이용한 변수 계급화(classing) 기법 이용, 원천 자료 중 범주형(categorical) 변수는 종속 및 독립변수 간 카이제곱 통계량, 연속형(continuous) 변수에 대해서는 t-검정을 이용하여 1차적으로 변수를 선정한다. 다음 단계는 1차로 선택된 변수 풀(pool)에 대해 성김화(coarse classing) 기법으로 계급세분화된 값을 축약하여 재범주화 한 후 단계적 선택법(stepwise method)으로 변수를 선택한다. 마지막 단계에서는 스피어만 상관계수(Spearman's correlation coefficient)를 이용해 다중공선성(multi-collinearity) 여부를 점검한 후 다중공선성이 있는 변수 중 설명력이 약한 변수는 제거하고 신용평가모형 구축에 활용한다.

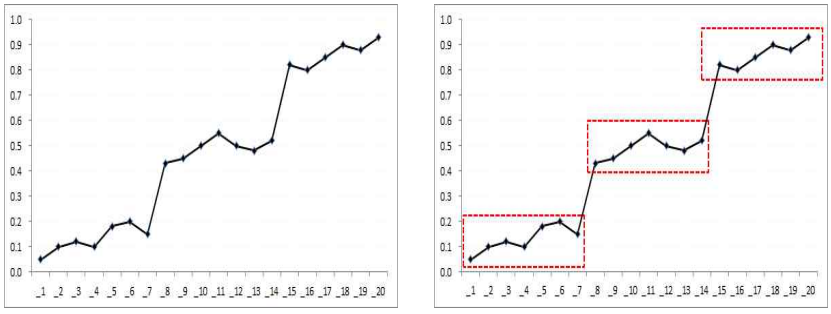
다음은 계급화에 대한 설명이다. 계급화는 원천 자료의 표준화를 위해 실제 현업에서 신용평가모형 구축 시 사용하는 방법이다. 이는 원래의 독립변수들과 종속변수인 불량 여부와의 관계 분석을 통해 불량률이 유사한 독립변수의 범주를 하나의 계급으로 묶은 후 순위의 의미를 가지도록 변환하는 기법이다.

계급화 기법은 크게 계급세분화(fine classing)과 성김화(coarse classing) 단계로 구분된다. 먼저 계급세분화 단계에서는 자료의 크기나 변수의 척도에 따라 다소 차이는 있지만, 개별 독립변수의 값을 정렬(sorting)한 후, 이를 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(신용보증재단중앙회, 2017). 성김화는 1차적으로 계급세분화에 의해 범주형으로 변환된 변수에 대해 동질적인 불량률을 보이는 구간으로 다시 묶는 단계이다(Leung et al., 2008).

계급화 기법을 이용할 경우 결측치와 특이값의 사용이 가능해지는데, 그 이유는 결측치나 특이값이 불량률과 연관되어 하나 또는 그 이상의 구간으로 묶이기 때문이다. 다음의 그림은 계급화에 대한 설명이다. 그림에서 x축은 세분화된 독립변수의 범주이며 y축은 각 세분화된 범주별 불량률을 의미한다.

[그림 1] 계급화의 개요



(a) 계급세분화 (b) 성김화 – 재범주화

로지스틱회귀모형을 이용한 변수 선택 방법은 전진 선택법, 후진 소거법, 단계적 선택법, 최적 조합법 등이 있다(Hosmer · Lemeshow, 2000). 본 연구에서 사용하는 단계적 선택법은 전진 선택법과 후진 소거법의 단점을 보완한 방법으로 변수 선택 시 이미 선택된 변수가 추가된 변수에 의해 설명력이 상실되는지 매 단계마다 검토한다(박주완, 2018). 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(신용보증재단중앙회, 2017). 이 때 각 회귀계수의 p-값은 일반적으로 유의수준 0.05 이하인 경우 해당 항목이 통계적으로 유의미하다고 판단한다(성웅현, 2001).

마지막 단계인 다중공선성 확인은 스피어만 상관계수를 이용한다. 보통 순서형 변수의 상관계수를 산출하기 위해 스피어만 상관계수를 사용하는데 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). 그리고 앞의 단계에서 단계적 선택법에 의한 회귀계수 추정치 값이 음(-)일 때, 이는 다중공선성의 영향일 가능성이 매우 높으므로, 이와 같은 경우 다중공선성이 있는 것으로 판단하여 상관계수 등을 통해 다중공선성이 확인되면 모형 구축용 독립변수에서 제외한다.

3. 모형 평가 방법 및 평가 측도

모형 평가는 구축된 모형의 예측 성능과 안정성을 확인하는 과정이다. 좀 더 자세히 설명하면, 모형 평가는 예측을 위해 만든 여러 가지 모형의 예측과 분류 성능을 평가 및 비교하여, 가장 좋은 예측 성능을 보유하고 있는 모형을 선택하기 위한 필수 단계이다(강현철 외, 1999).

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). 본 연구에서는 분석 표본의 개수가 충분하므로 예비 방법을 이용하여 “훈련용 자료:평가용 자료 = 7:3”, 즉 모형을 구축 하는 데에 전체 자료의 70%를 사용하고 나머지 30%로는 구축된 모형을 검증하고 평가하는 데에 사용한다. 예비 방법의 설명은 다음과 같다.

예비 방법은 난수(random number)를 이용하여 전체 분석용 자료를 두 개의 배타적인(exclusive) 훈련용(training data)과 검증용 자료(validation data)로 임의(randomly) 분할한 후, 모형 구축을 위해서는 훈련용 자료를 사용하고 검증용 자료는 모형 평가에 사용한다. 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 평가의 정확성도 높고 평가에 소요되는 시간이 단축된다는 장점이 있다(강창완 외, 2007). 그러나 평가용 자료를 모형 개발에 사용할 수 없고, 훈련용과 평가용 자료의 비율에 따라 다른 결과가 나타날 수 있다는 문제점과 개체수가 크지 않을 경우 불안정한 값을 제공한다는 단점이 있다(최종후 · 진서훈, 2002). 본 연구의 분석 자료는 13만개 이상으로 충분하기 때문에 예비 방법을 이용하는 것이 타당한 것으로 판단된다.

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(성웅현, 2001). 이외에도 리프트(lift) 도표, 반응률(response rate) 도표 등이 많이 사용되고 있다(강현철 외, 1999). 본 연구에서는 정분류율, G-Mean, F1 측도, 반응률을 이용하여 구축된 신용평가모형의 예측 성능을 비교하고 평가한다.

정분류율은 전체 자료를 얼마나 제대로 분류하는가의 문제이므로 값이 클수록 좋은 모형이다. 정분류율은 실제 0이 0으로 실제 1이 1로 분류되는 비율을 의미한다(강현철 외, 1999). 다음의 식에서 n_{00}과 n_{11}은 정분류가 되는 개수, n은 전체 표본수를 나타내고, 다음의 식으로 표현된다.

$$\text{정분류율} = (n_{00} + n_{11})/n \times 100 \qquad \text{(식 1)}$$

G-mean은 결과 범주가 0인 집단과 1인 집단을 동등하게 고려하는 측도로써 실제 범주가 0인 집단에 대한 정확도와 범주 1인 집단에 대한 정확도의 기하평균이다(Kubat et al. 1998). 그러므로 G-mean의 값이 클수록 좋은 예측 모형이다. G-mean의 산식은 다음과 같다(박주완, 2010).

$$G-mean=\sqrt{\frac{n_{00}}{n_{0+}}\times\frac{n_{11}}{n_{1+}}} \quad \text{(식 2)}$$

, n_{0+} = 실제0인 개수, n_{1+} = 실제1인 개수

$F1$ 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 et al., 2003). $F1$측도를 산출하기 위한 수식은 다음과 같다.

$$F1=\frac{2rp}{(r+p)}=\frac{2}{1/r+1/p}=\frac{2\cdot n_{11}}{n_{1+}+n_{+1}} \quad \text{(식 3)}$$

, p = 실제1, 예측1 정분류 빈도/예측1의 빈도

, r = 민감도 = 실제1, 예측1 정분류 빈도/실제1의 빈도

, n_{11} = 실제1이 예측1로 분류되는 개수

반응률은 훈련용 자료를 이용해 산출된 사후확률을 정렬하여 N 등분한 후, 각 등분에 포함된 종속변수의 특정 범주, 즉 불량의 빈도를 이용해 산출한다. 이와 같이 계산된 반응률은 도표를 통해 모형의 성능을 명확히 확인할 수 있는데, 모형에 의해 산출된 우량일 사후확률이 가장 큰 구간에서 가장 낮은 구간으로 갈수록 반응률, 즉 불량률이 낮게 나타나다가 급격하게 증가하는 형태인 경우 좋은 예측 판별력을 가진 모형이다(강현철 외, 1999).

$$\text{반응률}=\frac{\text{일정 } N \text{ 등분내 범주 1 빈도}}{\text{일정 } N \text{ 등분내 전체 빈도}}\times 100 \quad \text{(식 4)}$$

Ⅳ. 분석 개요

1. 분석 과정

본 연구는 다양한 기계학습을 이용하여 소상공인의 신용평가모형을 구축하고 예측 성능이 가장 우수한 모형을 선택하여 기계학습을 이용한 소상공인 신용평가모형 구축 가능성을 타진하는 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 기초 분석을 수행한다. 세 번째 단계인 변수의 유의성 검정 및 모형 구축 단계에서는 모형 구축에 사용할 최종적인 변수를 선정하는데, 계급화, 카이제곱 검정, t 검정, 스피어만 상관계수, 단계적 선택법 등을 이용한다. 그리고 모형 구축은 로지스틱회귀모형, 의사결정나무모형, 신경망모형, 렌덤포레스트모형, SVM을 이용하여 소상공인 신용평가모형을 구축한다. 마지막 단계인 모형 평가 및 비교에서는 예비 방법을 이용하여 각 구축된 모형의 정분류율, G-mean, F1 측도, 반응률을 비교한 후 예측 성능과 안정성이 가장 우수한 모형을 선택한다.

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 판

별이 불가능한 경우, 즉 종속변수가 결측치(missing value)인 경우 표본은 분석에서 제외하는 것이 원칙이지만, 종속변수에 결측이 존재하지 않아 최종적인 분석 대상으로 136,189개의 차주를 모두 활용한다.

그리고 소상공인 신용평가모형 구축을 위해 최초 총 36개의 변수를 이용한다. 이중에서 종속변수는 사고 발생 날짜를 이용하여 사고 여부를 생생하고, 독립변수는 총 35개이다. 다음의 표는 모형 구축을 위해 최초로 사용하는 변수 목록이다. 독립변수를 살펴보면 월 평균 매출액, 월 영업 이익을 제외한 나머지 변수는, 재무적인 변수가 아닌 기업의 일반 현황, 대표자 자금 상황, 보증 현황과 관련된 비재무적인 변수임을 알 수 있다.

<표 1> 모형 구축을 위한 변수

NO	변수명	비고	NO	변수명	비고
1	사고 여부	사고무(0), 사고유(1)	19	현금서비스금액	원
2	고객형태	개인, 법인	20	보유부동산	아파트, 단독주택 등
3	업종	업종 : 7개	21	업력	월
4	종업원수	명	22	거주기간	월
5	주사업장소유여부	소유, 임대 등	23	월평균매출액	원
6	주사업장임차보증금액	만원	24	월영업이익	원
7	주사업장월세금액	천원	25	월배우자소득	원
8	실거주지소유여부	소유, 임대 등	26	월기타수익	원
9	실거주지임차보증금액	만원	27	소유부동산금액	원
10	실소유지월세금액	천원	28	임대보증금사업장	원
11	차입금운전	백만원	29	임대보증금주택	원
12	차입금시설	백만원	30	예적금금액	원
13	차입금기타	백만원	31	유가증권금액	원
14	기보증잔액재단	백만원	32	재고자산	원
15	기보증잔액신보	백만원	33	고정자산	원
16	기보증잔액기보	백만원	34	권리금	원
17	기보증잔액개인	백만원	35	기타현금	원
18	담보제외차입기관수	개	36	직권말소	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(사고무)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, F1 값을 산출한다.

<표 2> 종속변수 분포

변수	변수값	빈도(개)	비율(%)	결측치 수(개)
사고 여부	사고무(0)	133,566	98.1	0
	사고유(1)	2,623	1.9	

다음의 <표 3>은 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.

<표 3> 사고 유무별 범주형 독립변수 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	제조업	10,064	9,800	264	97.4	2.6
	서비스업	23,513	23,133	380	98.4	1.6
	도소매업	44,312	43,489	823	98.1	1.9
	음식숙박업	38,067	37,283	784	97.9	2.1
	건설업	9,029	8,837	192	97.9	2.1
	운수업	7,871	7,738	133	98.3	1.7
	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	15,007	14,804	203	98.6	1.4
직권 말소	직권말소여(0)	368	345	23	93.8	6.3
	직권말소부(1)	135,821	133,221	2,600	98.1	1.9

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.

<표 4> 사고 유무별 연속형 독립변수 분포

변수	사고무		사고유	
	평균	표준편차	평균	표준편차
종업원수	1.1	133.7	0.7	2.0
주사업장임차보증 금액	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
실거주지임차보증 금액	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
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

연속형 독립변수는 사고 유무별로 평균을 비교하였는데, 종업원 수, 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.

V. 분석 결과

1. 변수 선택

본 절에서는 모형 구축에 필요한 독립변수를 선정한다. 변수 선정의 첫 번째 단계는 계급세분화 수행 후 KS통계량이 0.1 이상인 값을 가지는 변수를 선정한다.

<표 5> 1차 변수 선택 결과

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

그러나 본 연구에서는 변수 선택 시 좀 더 많은 분석 변수 사용을 위해 0.05를 기준으로 한다. 그리고 이를 보완하기 위해 원천 자료에 대해 카이제곱 검정과 t-검정을 수행하여 p-값이 0.05 이하인 변수를 1차적으로 선정한다.

세 가지 방법에 의해 1차로 선정된 독립변수는 다음 <표 5>와 같은데, 1차적으로 선정된 변수는 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 최초 변수 35개 중 23개이다. 그 결과 1차적으로 고객 형태 외 22개이다.

다음으로 성김화와 로지스틱회귀모형에서의 단계적 선택법에 적용한 결과 선택된 변수는 고객 형태 외 16개이다. 변수의 회귀계수를 보면 실거주지 임차보증 금액의 회귀계수가 -0.32로 음(-)값을 가지고 있다. 이는 다중공선성이 의심되는 결과이므로 스피어만상관계수로 다중공선성을 확인한다.

<표 6> 성김화 및 단계적 선택법 적용 결과

변수	회귀계수	표준오차	p값	변수	회귀계수	표준오차	p값
Intercept	-2.25	0.40	<.0001	기보증잔액기보	0.83	0.25	0.0008
고객형태	0.70	0.09	<.0001	담보제외차입기관수	0.58	0.03	<.0001
업종	0.23	0.05	<.0001	현금서비스금액	0.66	0.06	<.0001
주사업장임차보증금액	0.14	0.05	0.0069	보유부동산	0.35	0.04	<.0001
실거주지소유여부	0.36	0.08	<.0001	업력	0.64	0.04	<.0001
실거주지임차보증금액	**-0.32**	**0.09**	**0.0002**	거주기간	0.23	0.05	<.0001
실소유지월세금액	0.55	0.09	<.0001	월평균매출액	0.34	0.04	<.0001
차입금운전	0.48	0.18	0.0059	재고자산	0.70	0.15	<.0001
기보증잔액재단	0.22	0.06	<.0001	직권말소	1.01	0.26	0.0001

스피어만 상관계수를 이용하여 각 독립변수별로 다중공선성을 확인한 결과, 실거주지 임차보증 금액과 실소유지 월세 금액의 상관계수가 0.773으로 나타나 0.7을 초과하여 다중공선성이 존재하는 것으로 확인되었다.

다중공선성이 존재하는 2개의 변수인 실거주지 임차보증 금액과 실소유지 월세 금액 각각에 대해 종속변수를 불량 여부로 하여 단변량 로지스틱회귀분석을 수행한 후 회귀계수 추정치, 왈드 카이제곱 값, 정분류율, c통계량 값이

더 크게 나타난 변수의 설명력과 변별력이 더 크기 때문에, 이 값들이 큰 변수를 최종적으로 선택한다.

<표 7> 다중공선성 확인 결과

변수	상관계수	변수	상관계수	변수	상관계수
고객형태	월평균매출액	차입금운전	실거주지 소유여부	업력	실거주지 소유여부
	-0.168		0.081		0.176
업종	월평균매출액	기보증잔액 재단	담보제외 차입기관수	거주기간	업력
	-0.06713		0.271		0.130
주사업장 임차보증금액	실소유지월세금액	기보증잔액 기보	고객형태	월평균 매출액	고객형태
	0.091		0.082		-0.168
실거주지 소유여부	보유부동산	담보제외 차입기관수	기보증잔액재단	재고자산	월평균매출액
	0.641		0.271		-0.032
실거주지 임차보증금액	**실소유지 월세금액**	현금서비스 금액	담보제외 차입기관수	직권말소	실소유지 월세금액
	0.773		0.155		0.022
실소유지 월세 금액	**실거주지 임차보증금액**	보유 부동산	실거주지 소유여부		
	0.773		0.641		

다음의 표를 보면 실소유지 월세 금액의 회귀계수 추정치, 왈드 카이제곱값, 정분류율, c통계량이 더 크게 나타났으므로, 최종 신용평가모형 구축 변수로는 실소유지 월세 금액을 선택하는 것이 타당하다는 결론을 내릴 수 있다.

<표 8> 다중공선성 존재 변수 선택을 위한 회귀분석 결과

변수	회귀계수 추정치	Standard Error	Wald Chi-Square	p값	정분류율	c통계량
실거주지임차보증금액	0.6249	0.0497	157.8362	<.0001	27.0	0.563
실소유지월세금액	**0.8319**	**0.0501**	**275.5236**	**<.0001**	**27.6**	**0.578**

다음의 표는 최종적으로 선택된 변수 16개 독립변수의 성김화에 의한 원천자료 재범주화 결과이다. 최종적으로 선택된 변수들 중 계급0, 계급1, 계급2 세 개의 계급으로 재범주화된 변수는 업종 포함 5개이고, 2개의 계급으로 재범주화된 변수는 고객 형태 포함 11개이다.

<표 9> 성김화에 의한 재범주화 결과

구분	계급0	계급1	계급2
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	-
담보제외차입기관수	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	16 초과 67 이하, 261 초과, 결측치	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	-

다음의 표는 각 독립변수별로 계급화 과정에 의해 원천 자료를 계급화하여 표준화 한 후 각 계급별 불량률을 산출한 결과이다. 전술하였듯이 계급화는 원천 자료에서 각 독립변수의 값이 계급이 높아질수록 불량률이 감소하도록 표준화하는 방법으로, 계급화가 제대로 이루어졌다면 계급이 높을수록 차주가 우량일 확률이 증가하여 신용평가점수가 높아지게 된다. 그러므로 계급화가 잘 이루어진 경우 계급0, 계급1, 계급2로 갈수록 불량률은 감소해야 한다.

<표 10>을 보면 모든 변수에서 계급0에서 계급1 또는 계급2로 갈수록 계급별 불량률이 점차 감소하고 있음을 알 수 있다. 그러므로 계급화에 의한 자료 표준화가 잘 이루어졌다고 결론내릴 수 있다. 다음 절에서는 성김화된 16개의 독립변수값을 이용하여 의사결정나무모형, 로지스틱회귀모형, 신경망모형, 랜덤포레스트모형, SVM을 이용하여 신용평가모형을 구축하고 비교하여 예측 성능이 가장 우수한 모형을 선택한다.

<표 10> 성김화에 의한 재범주화 결과

구분	계급0 불량률	계급1 불량률	계급2 불량률
고객형태	3.2%	1.8%	-
업종	2.2%	1.7%	1.2%
주사업장임차보증금액	2.4%	1.7%	-
실거주지소유여부	2.7%	1.0%	-
실소유지월세금액	3.5%	1.6%	-
차입금운전	2.0%	1.1%	-
기보증잔액재단	2.3%	1.8%	-
기보증잔액기보	4.8%	1.9%	-
담보제외차입기관수	4.2%	2.3%	1.3%
현금서비스금액	4.2%	1.7%	-
보유부동산	2.8%	1.4%	0.9%
업력	3.1%	1.5%	0.8%
거주기간	2.3%	1.5%	-
월평균매출액	2.4%	1.7%	1.3%
재고자산	3.4%	1.9%	-
직권말소	6.9%	1.9%	-

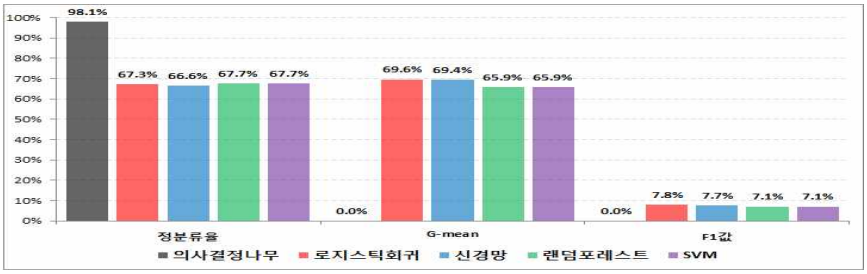
2. 신용평가모형 구축 및 비교

본 절은 최종적으로 선택된 독립변수 16개를 이용하여 대표적인 기계학습 기법인 의사결정나무, 로지스틱회귀, 신경망, 랜덤포레스트, SVM을 이용하여 모형을 구축한 후, 구축된 모형에 대한 평가 결과를 비교한 것이다. 모형 구축을 위한 평가는 예비 방법을 이용하는데, 훈련용 자료와 평가용 자료는 전체 자료에 대해 균등분포(uniform distribution)에 의한 난수를 생성한 후 '훈련용:평가용=7:3'으로 분할한다. 분할 결과 모형 구축을 위한 훈련용 자료는 총 95,147개이고, 구축된 모형의 평가는 41,013개의 평가용 자료를 이용한다. 이 때 모형 구축을 위한 도구로는 기계학습에 가장 대표적으로 사용하고 있는 R을 이용하며, 평가 결과 분석은 SAS9.4를 이용한다.

[그림 3]과 <표 11>는 훈련용 자료에 대한 정분류율, G- mean, F1값을 정리한 것이다. 먼저 정분류율은 의사결정나무 98.1%, 로지스틱회귀모형 67.3%, 신경망모형 66.6%, 랜덤포레스트모형 67.7%, SVM 21.3%으로 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) 전체가 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.

[그림 3] 훈련용 자료에 대한 정분류율, G-mean, F1값



<표 11> 훈련용 자료에 대한 분류 결과

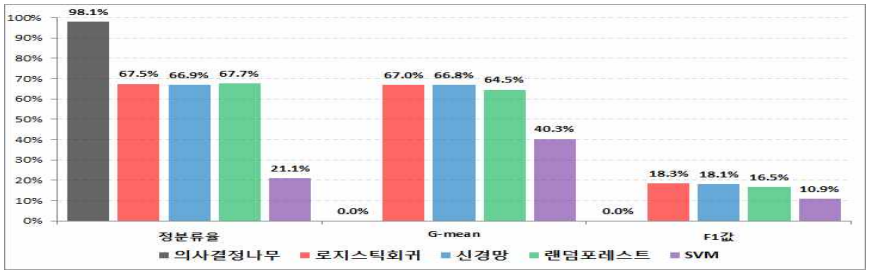
구분		훈련용자료				
		예측 0	예측 1	정분류율	G-mean	F1값
의사결정나무	실제 0	93,342	0	98.1%	0.0%	0.0%
	실제 1	1,832	0			
로지스틱회귀	실제 0	62,712	30,630	67.3%	69.6%	7.8%
	실제 1	511	1,321			
신경망	실제 0	62,058	31,284	66.6%	69.4%	7.7%
	실제 1	503	1,329			
랜덤포레스트	실제 0	63,215	30,127	67.7%	65.9%	7.1%
	실제 1	659	1,173			
SVM	실제 0	18,481	74,861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

다음으로 G-mean은 로지스틱회귀모형 69.6%, 신경망모형 69.4%, 램덤포레스트모형 65.9%, SVM 44.5%, 의사결정나무모형 0.0%로 나타나고 있는데, G-mean은 결과 범주가 0인 집단과 1인 집단을 동등하게 고려하는 측도로써 값이 클수록 분류 및 예측 성능이 우수하다. 로지스틱회귀모형과 신경망모형은

큰 차이가 나지 않는 가운데 다른 모형에 비해 값이 크므로 분류를 위한 예측 성능이 가장 우수하다고 할 수 있다. 마지막으로 F1값도 로지스틱회귀모형과 신경망모형이 다른 모형에 비해 상대적으로 높게 나타나고 있어 예측 성능이 좋음을 알 수 있다.

[그림 4]와 <표 12>는 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%, SVM 21.1%로 의사결정나무모형에 의한 정분류율이 가장 높게 나타나고 있다.

[그림 4] 평가용 자료에 대한 정분류율, G-mean, F1값



<표 12> 평가용 자료에 대한 분류 결과

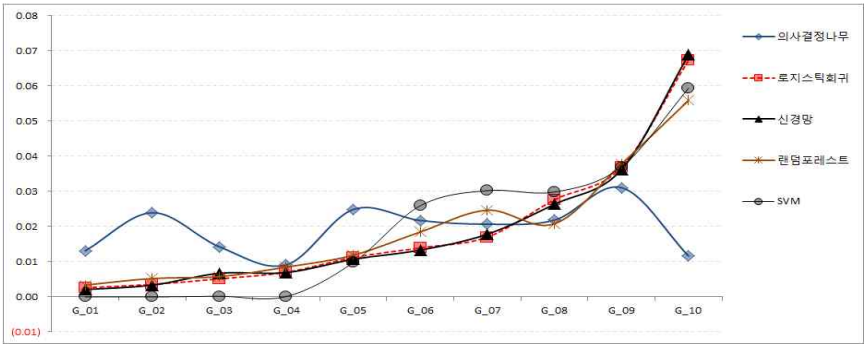
구분		평가용자료				
		예측 0	예측 1	정분류율	G-mean	F1값
의사결정나무	실제 0	40,222	0	98.1%	0.0%	0.0%
	실제 1	791	0			
로지스틱회귀	실제 0	27,138	13,084	67.5%	67.0%	18.3%
	실제 1	265	526			
신경망	실제 0	26,896	13,326	66.9%	66.8%	18.1%
	실제 1	263	528			
랜덤포레스트	실제 0	27,290	12,932	67.7%	64.5%	16.5%
	실제 1	306	485			
SVM	실제 0	8,008	32,214	21.1%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

그러나 이 결과 또한 결과 표를 보면 {실제 1, 예측 1}의 빈도가 0으로 나타나고 있는데, 이는 의사결정나무모형은 불량인 차주를 모두 우량으로 분류하

고 있음을 나타내는 것이다. 그러므로 의사결정나무는 불량인 차주의 판별이 제대로 이루어지지 않아 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%, SVM 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.

[그림 5] 훈련용 자료에 대한 반응률 비교



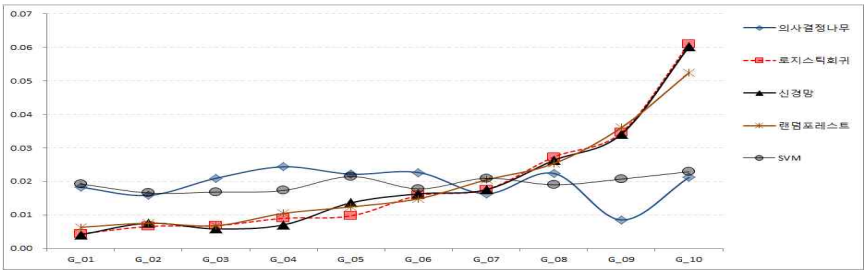
<표 13> 훈련용 자료에 대한 반응률 결과

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

다음은 구축된 모형의 반응률에 대한 결과이다. 전술하였듯이 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 5]와 <표 13>은 훈련용 자료에 대한 반응률을 정리한 것이다. 이 결과를 보면 로지스틱회귀모형, 신경망모형, 랜덤포레스트모형의 반응률이 점차 증가하고 있으며, G_09와 G_10에서 급격히 상승하고 있음을 알 수 있다. 그러나 의사결정나무모형과 SVM은 역전 현상이 발생하고 있다. 이 결과를 통해 로지스틱회귀모형, 신경망모형, 랜덤포레스트모형의 우불량 변별력이 좋음을 알 수 있다.

[그림 6]과 <표 14>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.

[그림 6] 평가용 자료에 대한 반응률 비교



<표 14> 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

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

지금까지의 분석 결과를 토대로 본 연구에서 사용한 자료에 대해 예측 성능이 가장 우수하고 안정성이 뛰어난 소상공인 신용평가모형을 선택한다. 본 절의 결과를 살펴보면, 로지스틱회귀모형, 신경망모형과 랜덤포레스트모형에 의한 소상공인 신용평가모형의 예측 성능과 안정성이 우수함을 알 수 있다. 그러나 반응률 값을 살펴보면 로지스틱회귀모형은 역전 현상이 발생하지 않고 G_01~G10으로 갈수록 점차 증가하는 반면, 신경망모형과 랜덤포레스트모형은 반응률 역전 현상이 발생하고 있음을 알 수 있다. 그러므로 로지스틱회귀모형의 분류 예측 성능 즉 등급 간 변별력이 가장 우수한 것으로 판단된다.

결론적으로 의사결정나무모형의 경우 정분류율은 가장 높았지만 사고가 발생한 차주 즉 불량인 차주의 예측 성능이 매우 나쁘며, 신경망과 랜덤포레스트모형은 로지스틱회귀모형과 분류 성능 면에서 큰 차이는 나지 않고 비교적 우수하였지만, 등급화의 과정에서 일부 등급 구간에 불량률 역전 현상이 발생하였으므로 본 자료에 대해서는 로지스틱회귀모형을 사용하는 것이 가장 타당한 것으로 판단된다.

VI. 결론 및 향후 과제

본 연구는 소상공인 신용평가를 위해 일반적인 재무정보 대신 신용보증재단이 보유한 내부정보를 이용하여 실제 현업에서 신용평가모형 구축 시 가장 많

이 사용하고 있는 빅데이터 분석을 위한 기계학습 중 의사결정나무모형, 로지스틱회귀모형, 신경망모형, 랜덤포레스트모형, SVM으로 신용평가모형을 구축하였을 때 분류를 위한 예측 성능이 우수한 모형이 무엇인지를 확인하는 것이다.

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

다양한 기계학습 기법을 이용하여 소상공인 신용평가모형을 구축한 결과는 다음과 같이 요약할 수 있다. 첫째, G-mean, F1 측도를 살펴보았을 때 로지스틱회귀모형이 가장 좋은 예측 성능을 가지고 있으며, 계급불균형 자료에 대해 정분류율을 이용하여 모형을 평가하는 것은 적절하지 않다는 사실을 확인하였다. 둘째, 반응률을 보았을 때 의사결정나무모형, 신경망모형, 랜덤포레스트모형, SVM 보다 로지스틱회귀모형이 우량일 사후확률이 높은 구간에서 낮은 구간으로 갈수록 불량률이 증가 추세를 보이고 서열화가 매우 잘 이루어지고 있으므로 가장 좋은 예측 성능을 가지고 있다. 그러므로 신용보증재단의 자료를 이용하여 소상공인 신용평가모형을 구축할 경우 로지스틱회귀모형이 가장 좋은 것으로 판단된다.

이상의 분석 결과를 통한 본 논문의 결론은 다음과 같다. 기계학습을 이용하여 정확한 분류나 예측을 위해서는 소수계급과 다수계급의 특성을 충분히 학습해야 한다. 그러나 계급불균형이 매우 심한 자료는 소수계급의 수가 부족하기 때문에 예측 성능이 높은 모형을 구축하기 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.

소상공인 신용평가에 대한 과거 많은 연구들은 객관적인 자료 부족이라는 한계상황으로 인해 재무 자료 이외의 다양한 비재무 자료 활용에 대한 연구가 많이 다루어지고 있다(박주완, 2018). 그러나 본 논문은 과거의 연구와는 달리 기계학습을 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 적용 등이 부족했다는 점에서 한계가 있다. 또한 분석 자료 표준화를 위한 다양한 방법론 적용 등도 다루지 않았다는 한계점도 가지고 있다.

현재 빅데이터와 이를 이용한 기계학습은 4차 산업혁명의 주요 요소가 되었고, 금융 산업에서는 신용평가 등 다양한 업무 영역에 빅데이터나 기계학습을 적용하기 위해 시도가 이루어지고 있으며, 실제 현업에서 이를 적용하는 사례들이 점차 늘어나고 있다(신윤재, 2016). 이처럼 현 시대는 양질의 자료 수집과 가치창출을 위한 기계학습의 적용은 국가나 기업의 경쟁력 향상을 위한 꼭 필요한 사항이 되었다.

데이터 자체가 중요한 경쟁력 제고의 요소가 되고 있는 현 시점에서, 일반적으로 객관적인 데이터가 부족하다고 알려져 있는 소상공인의 경우 실제로 객관적인 양질의 자료를 통한 연체 및 신용평가모형 구축과 현황 분석이 매우 어려운 실정이다. 그러므로 다양한 데이터 수집과 이를 활용하기 위한 방법론의 연구는 앞으로도 계속 추진되어야 할 중요한 과제일 것이다.

본 연구를 통한 향후 연구 방향은 다음과 같다. 첫째, 우량과 불량의 구분이 모호한 판단미정 차주, 분석을 위한 기본적인 정보가 부족한 신규 창업자 등의 대상을 위한 기계학습 적용 연구가 필요하다. 둘째, 기업의 규모나 업종 등에 특성에 차이가 있을 수 있으므로, 자료의 양과 질이 충분하다면 규모와 업종 등을 세분화한 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.

참고문헌

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.

김승혁 · 김종우(2007). 「Modified Bagging Predictors를 이용한 SOHO 부도 예측」, 『지능정보연구』, 13(2), pp 15-26.

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

김효진(2018). 『머신러닝에 대한 이해』, 주택금융리서치.

박정윤(2000). 「재무정책과 기업부실 예측」, 『재무관리논총』, pp 93-116.

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

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

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

박주완 · 송창길 · 배진성(2017). 「기계학습 기법을 이용한 소상공인 신용평가모형 구축에 관한 연구」, 『한국비즈니스리뷰』, 제10권, 3호.

박주완(2019),『빅데이터 분석 기법을 이용한 소상공인 신용평가모형 구축 연구』, 신용보증재단중앙회.

블로터(2019). http://www.bloter.net/archives/351562, 카카오뱅크 시스템은 진화 중.

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

성웅현(2001).『응용 로지스틱 회귀분석-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 CB 연구소.

연합인포믹스(2018). http://news.einfomax.co.kr/news/ articleView.html?idxno=346331, 카뱅·케뱅, 자체 신용평가 모형 구축 박차…당국도 힘 실어.

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

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.

이영섭(역)(2003).『데이터마이닝 Cookbook』, 교우사.

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.

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.

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

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

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, pp 1137-1143.

Kubat, M., Holte, R. C., Matwin, S.(1998), “Machine Learning for the Detection of Oil Spills in Satellite Radar Images,” Machine Learning, 30, pp 195–215.

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).

Ohlson, J. A.(1980). “Financial Ratios and the Probabilistic Prediction of Bankruptcy,” Journal of Accounting Research (spring), pp 109-131.

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

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.

소상공인 및 자영업자 경영환경 변화에 따른 대응

윤혁준*

□ 경제위기는 주로 금융과 같은 경제적 요인으로 발생하였으나, '20년 국내외 경제위기는 코로나19 감염병 위협에서 비롯되어 일부 경제활동을 마비시켰으며, 경제에 악영향과 장기적 영향을 미칠 것으로 전망

- 소상공인 및 자영업자는 매출감소에 따른 위기상황을 인식하고 현 사업장의 재정 상황 및 현금흐름에 대한 면밀한 검토와 함께 정부 및 지자체의 적극적인 지원 사업을 잘 활용할 필요가 있음
- 또한 코로나19 이후 소비행태 변화에 맞춰 온라인 판매 채널을 강화하고, 데이터 기반의 소비트렌드 분석, 마케팅 등의 노력을 경주해야 할 것임

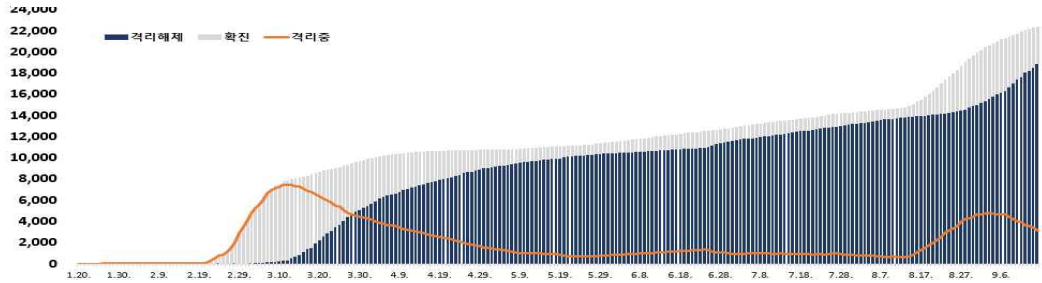
* 신용보증재단중앙회 교육연구부 과장(경제학석사)

□ 경제위기는 주로 금융과 같은 경제적 요인으로 발생하였으나, ‘20년 국내외 경제위기는 코로나19 감염병 위협에서 비롯되어 일부 경제활동을 마비시켰으며, 경제에 악영향과 장기적 영향을 미칠 것으로 전망

○ 코로나19 확진자는 ‘20년 3월 급격히 증가한 이후 정부의 철저한 방역과 전 국민적인 노력으로 증가세가 둔화되었으나 ’20년 8월 재확산

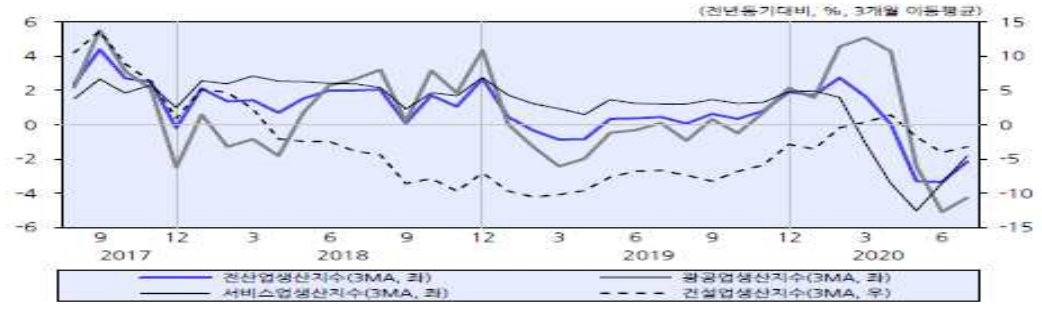
- 산업생산지수는 ‘20년 3월 전후로 급격한 하락을 보이다 6월 전후 반등을 보이나, 8월 중순 이후 코로나19 재확산으로 경기 하방압력이 확대

<코로나19 일일 신규 확진자 현황>



자료 : 질병관리청(2020.9.15.) 코로나바이러스감염증-19 국내 발생 현황 정례 브리핑

<산업별 생산지수>

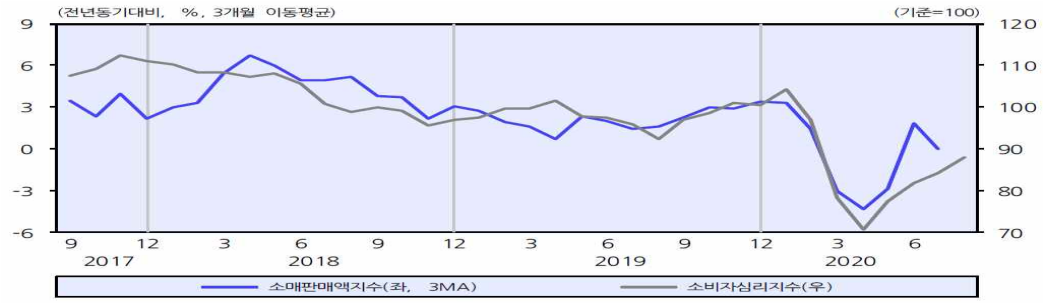


자료 : KDI 경제동향(2020.9월) 인용

- ‘20년 7월 소매판매액의 증가세가 둔화되고 서비스업생산 감소폭이 확대된 가운데, 코로나19 재확산으로 소비가 크게 위축

- 한국개발연구원이 분석한 신용카드 매출액 증가율을 보면 수도권 방역이 2단계로 격상된 8월 중순 이후(8월19~30일) -12.1%를 기록하면서 신천지를 중심으로 확진자가 발생하여 사회적 거리두기가 처음 시행되었던 5월 이전 수준(2월 19일~5월 5일, -14.2%)으로 낮아짐

<소매판매액 및 소비자심리지수>



자료 : KDI 경제동향(2020.9월) 인용

<코로나19 확산 시기와 신용카드 매출액>

자료 : KDI 경제동향(2020.9월), 신한카드 추정치

□ 서민경제의 주축인 소상공인 등 자영업자 또한 '19년 말부터 시작된 코로나19의 확산 및 재확산의 여파로 경영상황이 매우 악화

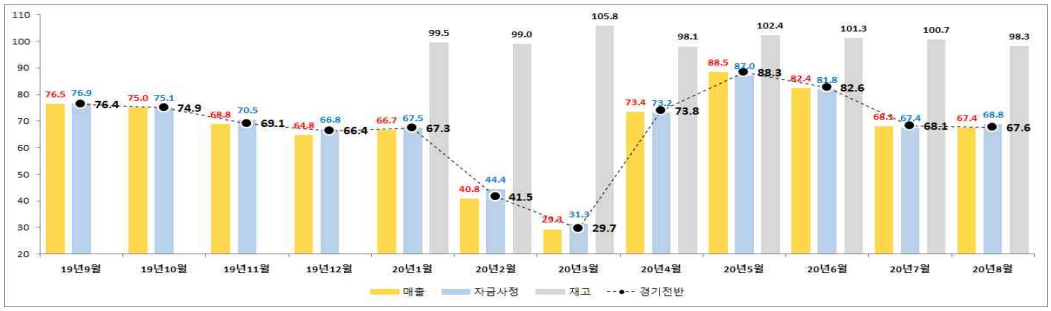
○ 소상공인의 경기 및 매출 관련 체감 지수는 코로나19 확산 초기 급감하였으나, 정부의 경기 부양을 위한 자금 지원, 동행세일 등 다양한 정책 시행으로 다소 반등을 보였으나,

- 코로나19 재확산에 따른 강도 높은 사회적 거리두기 시행으로 체감 경기 및 매출에 대한 지수는 감소 추이로 사업운영에 전반적으로 어려움이 가중되고 있음을 시사

  · 소상공인 체감경기지수(기준 100) : ('19.9월) 76.4 → ('20.8월) 67.6

  · 소상공인 체감매출지수(기준 100) : ('19.9월) 76.5 → ('20.8월) 67.4

<소상공인 부문별 체감 지수-소진공 BSI>



자료 : 소상공인시장진흥공단,「소상공인 시장 경기동향 조사」자료를 이용하여 재구성

<소상공인 업종별 체감 지수-소진공 BSI>

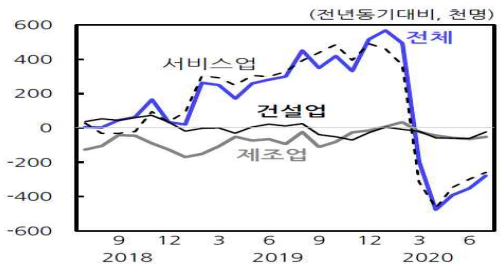
업종	'19.1	'19.2	'19.3	'19.4	'20.1	'20.2	'20.3	'20.4	'20.5	'20.6	'20.7	'20.8	추세
제조업	85.5	72.6	75.0	66.4	78.9	46.5	32.2	72.2	77.5	79.7	67.0	63.9	
소매업	80.1	64.2	63.3	73.0	64.6	38.5	35.4	72.9	88.2	88.2	65.5	63.6	
음식점업	67.8	71.6	65.7	56.8	60.9	29.8	24.2	77.0	98.5	78.9	68.4	68.9	
부동산중개업	70.6	74.5	58.9	64.5	59.2	58.5	40.5	75.7	79.1	78.6	72.1	61.9	
전문기술사업서비스업	71.4	79.2	74.5	62.8	79.3	36.6	19.5	58.5	79.9	87.2	70.7	62.8	
교육서비스업	84.9	87.3	79.4	72.4	87.4	62.1	30.8	76.1	93.1	95.0	85.6	84.5	
스포츠및오락관련서비스업	72.4	67.8	56.6	80.0	62.0	35.8	22.5	68.7	74.9	65.8	65.0	69.0	
수리업	66.3	71.9	72.2	62.5	54.5	41.1	25.3	68.8	81.5	91.4	69.9	70.5	
개인서비스업	85.5	92.3	81.0	65.3	69.5	39.3	31.8	76.5	94.0	80.6	54.7	61.2	

자료 : 소상공인시장진흥공단,「소상공인 시장 경기동향 조사」자료를 이용하여 재구성

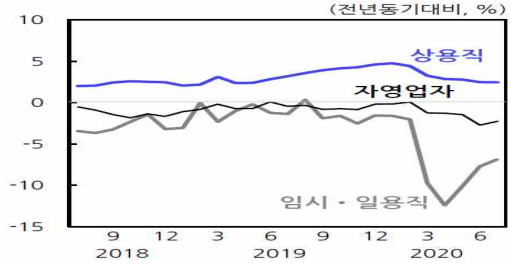
□ ‘20년 고용시장 통계를 통해서도 소상공인 및 자영업자의 경영활동 전반에 어려움이 가중되고 있음을 유추

○ 취업자는 대면 접촉이 많은 서비스업을 중심으로 큰 폭으로 감소한 후 5월부터는 고용 부진이 일부 완화되기도 하였으나, 코로나19 재확산으로 고용시장이 다시 위축될 가능성이 상존

<산업별 취업자 증감>



<종사상 지위별 취업자 증가율>

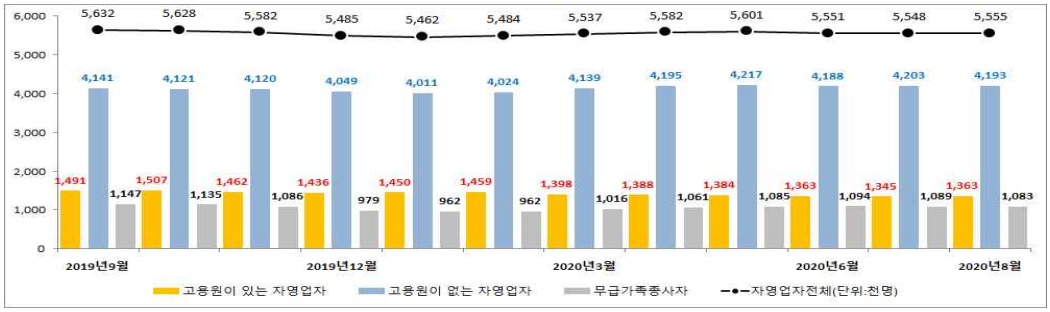


자료 : 통계청(KOSIS) 인용

○ 산업별로는 대면 접촉이 많은 서비스업에, 종사상 지위별로는 임시·일용직 근로자와 자영업자에 고용 충격이 집중되는 등 코로나19는 경제 주체별로 불균등한 영향

- 다만, 코로나19 확진자가 급증한 ‘20.3월에 비해 ’20.8월의 전체 자영업자(+18천명)와 고용원이 없는 자영업자(+54천명)는 오히려 증가한 수치로, 이는 유급 종업원 대신 비용이 들지 않는 가족을 동원하여 영업 활동을 유지하거나 연명하고 있는 것으로 보임

<자영업자 및 무급가족종사자 추세>



자료 : 통계청(KOSIS), 경제활동인구조사

□ 코로나19 확산 이후 고강도 사회적 거리두기의 장기화로 외부활동이 크게 감소하면서 소상공인 등 자영업자는 매출 감소로 인한 어려움이 커지고 있으며, 유동인구 감소는 핵심 상권에 더 큰 영향

○ 소상공인은 코로나19 이전과 대비해 매출액 감소폭이 6월 4주차 33.4%로 6월 2~3주차보다 증가

<코로나19 발생 이전(평소) 대비 소상공인 매출액 감소 비율>

(단위 : %)

조사일	3.23	3.30	4.6	4.13	4.20	4.27	5.4	5.11	5.18	5.25	6.1	6.8	6.15	6.22	6.29
소상공인	66.8	66.9	**69.2**	65.4	64.5	56.7	55.0	54.6	51.3	45.3	38.7	32.0	31.6	31.6	33.4
전통시장	**65.8**	65.5	65.0	65.4	61.1	55.8	56.4	52.6	51.6	39.6	32.5	27.1	26.5	26.6	28.5

* 질문내용 : 평소(코로나19 확산 상황 이전) 대비 매출액 변화가 어떻게 되는가?

자료 : 중소벤처기업부, 소상공인 매출액 조사(22차, 6월 29일) 결과 인용

○ 근거리에 거주하는 고객의 수요 보다 멀리서 해당 상권을 방문하는 고객이나 관광객에 대한 매출 의존도가 높은 핵심 상권이 코로나19에 따른 영향을 더 크게 받는 모습

- 서울시와 서울연구원의 보도자료('20.6.2일)에 따르면 코로나19 영향이 없던 2020년 1월 평시 대비 5월 넷째 주 생활인구 수는 서울의 주거중심지역 자치구(강동, 성북, 도봉, 광진, 금천, 은평 등)는 평시('20.1월) 생활인구 수를 넘어선 반면, 업무 및 상업 중심과 관광객이 많은 중구, 종로 등은 '20.1월 대비 인구수가 낮게 조사

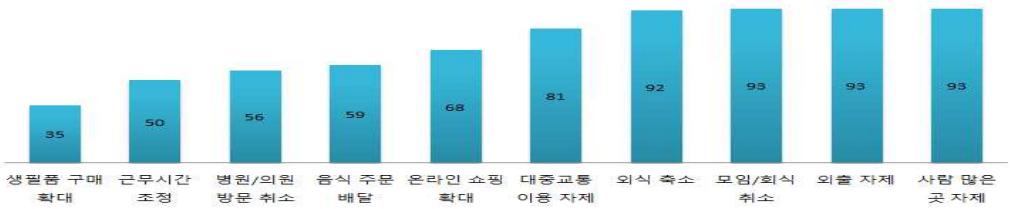
- 서울 소재 상점매출액은 4개월간(2.10.~5.24, 15주) 약 3조2천억 원 감소하였으며, 특히 삼성1동, 서교동, 신촌동, 명동에서 1천억원 이상의 매출이 감소. 반포4동, 소공동, 역삼1동, 종로 1·2·3·4가동, 한강로동, 잠실3동은 7백억원 이상의 매출 감소 등 상업 및 업무중심 지역에서 매출 감소가 큰 것 조사

□ 코로나19 확산에 따른 소상공인 및 자영업자의 매출 감소와 함께 경영 및 영업환경에 대한 급격한 변화가 발생

○ 많은 사람들이 집에 머무는 시간이 늘어나면 다양한 활동을 집 안에서 해결하는 홈코노미(Homeconomy)가 빠르게 확산

- 한국리서치 조사에서 응답자의 90% 이상이 코로나 이후 사람이 많은 곳의 출입과 외출을 자제하고 모임 및 회식을 취소하거나 외식을 줄인 것으로 조사

<코로나19 이후 일상생활 변화>

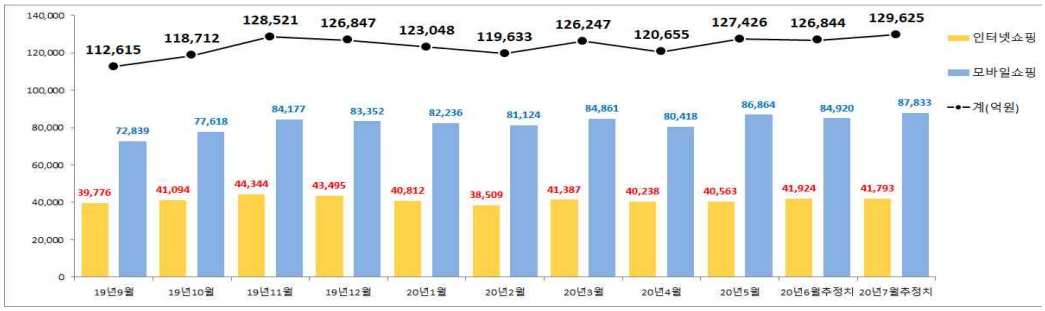


자료 : 전국 만18세 이상 남녀 1,000명 조사, 한국리서치 코로나19 인식조사('20.5.6일자) 재가공

○ 또한, 사회적 거리두기로 인해 대면 접촉을 최소화하고 외부 활동을 자제하면서 집에 머무는 시간이 길어지고 온라인을 통한 소비 활동이 증가

- 코로나19 시작 이후 소상공인 및 자영업자 영업환경의 가장 큰 변화 중 하나는 비대면 온라인 거래가 증가하였다는 것

<온라인쇼핑 거래액 증감 추세>

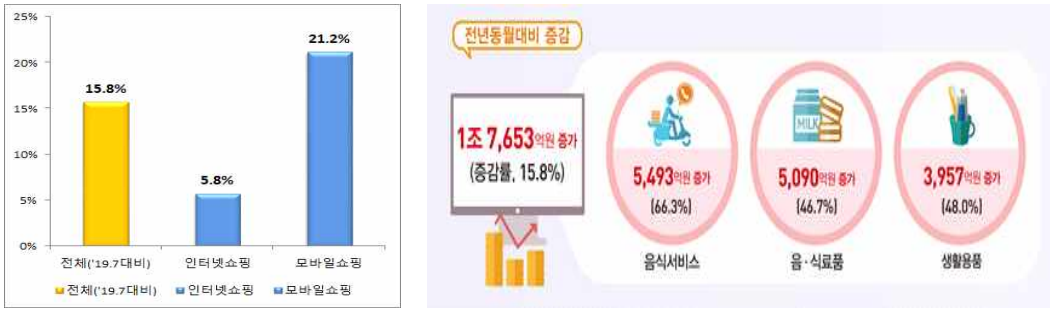


자료 : 통계청, 「온라인쇼핑 동향 조사」를 이용하여 재구성

- '20년 7월 온라인쇼핑거래액 증감률은 전년 동월대비 15.8% 증가, 세부 상품군별 온라인쇼핑거래액은 전년 동월대비 문화 및 레저서비스(-67.8%), 여행 및 교통서비스(-51.6%) 등에서 감소, 반면 음식서비스(66.3%), 생활용품(48.0%), 음·식료품(46.7%) 등은 증가

- '20년 7월 모바일쇼핑거래액 증감률은 전년 동월대비 21.2% 증가, 온라인쇼핑 총 거래액 중 모바일쇼핑 비중 67.8%로 전년 동월대비 3.1%p 상승, 특히 음식서비스, 음·식료품, 생활용품 등에서 증가

<온라인쇼핑거래액 증가>



주 : 통계청, 2020년 7월 온라인쇼핑 동향 재가공 및 인용

□ 코로나19 확산 이후 지역신보 보증 이용 소상공인의 체감 경기 및 전망 역시 부정적으로 인식되고 있으며, 이에 따른 필요자금 수요 및 대출규모는 급격하게 증가된 모습

○ 신용보증을 받은 차주의 현재 경기, 매출에 대한 체감은 코로나19 이전에 비해 매우 악화되었으며 향후 전망도 매우 부정적임

* 보증이용업체 체감경기지수(기준치 100) : ('19.2Q) 57.4 → ('20.2Q) 47.4 → (3Q) 51.5^{P}

* 보증이용업체 체감매출지수(기준치 100) : ('19.2Q) 54.1 → ('20.2Q) 32.8 → (3Q) 36.5^{P}

○ 소상공인 및 자영업자의 대출은 코로나19 이후 불확실성이 커지면서 대출 수요 및 정책금융기관 등의 금융지원이 이어지면서 증가세는 지속

* 개인사업자 은행 원화대출(한은, 조원) :
('18년 연중) 25.0 → ('19년 연중) 24.7 → ('20년 1~8월중) 34.0
* 지역신보 신규보증공급(신보중앙회, 조원) :
('18년) 7.2 → ('19년) 9.0 → ('20년 1~8월중) 20.5

□ '21년 소상공인 및 자영업자 경영환경 변화에 따른 대응 방향

○ 감염병 위협 해소의 관건인 백신 및 치료제 개발, 보급이 지연될 경우 경기 침체 장기화가 불가피하며, 이러한 가능성에도 대비가 필요

- 이번 경제 위기에서 침체의 심도와 길이를 결정하는 가장 중요한 요소는 경제정책의 효과가 아닌 감염병 위협의 해소 여부가 관건

○ 심각한 경기침체 속에 일부 IT 업종 등 소위 비대면 관련 업종은 매출과 순익이 급증하는 호황을 구가

- 이는 보편적 지원보다 주요 피해업종과 취약계층에 지원을 집중하는 대응이 필요함을 시사

○ 중장기적으로 산업구조의 변화, 경제정책 및 정부 역할의 변화, 감염병 위협을 포함한 환경 생태적 위기에 대한 인식 변화 등으로 이러한 변화에 민관 모두 선제적으로 대비하기 위한 방안을 강구해함

○ 소상공인 및 자영업자는 단기적인 매출감소에 따른 위기상황을 인식하고 현 사업장의 재정 상황 및 현금 흐름에 대한 면밀한 검토와 함께 정부 및 지자체의 적극적인 지원 사업을 잘 활용할 필요가 있음

- 또한 코로나19 이후 소비행태 변화에 맞춰 온라인 판매 채널을 강화하고, 데이터 기반의 소비트렌드 분석, 마케팅 등의 노력을 경주해야 할 것임

참고 자료

서울연구원, “코로나19 영향, 생활인구는 지역별·상점 매출액은 업종별 차이 커”, 2020.6.2. 보도자료.

소상공인시장진흥공단, 각 년도「소상공인 시장 경기동향 조사」.

신용보증재단중앙회, 2020년 2/4분기 보증이용업체 경기실사지수.

중소벤처기업부, 소상공인 매출액 조사(22차, 2020년 6월 29일).

질병관리청(2020.9.15.) 코로나바이러스감염증-19 국내 발생 현황 정례 브리핑.

통계청,「온라인쇼핑동향조사」.

통계청(KOSIS).

한국개발연구원(KDI), 경제동향(2020.9월).

한국리서치, 코로나19 인식조사(‘20.5.6일자).

한국은행, 2020년 8월중 금융시장 동향.

KB지식 비타민, 포스트 코로나 시대, 자영업 시장의 변화, KB금융지주 경영연구소(2020.6월).

KIET 산업정책 리포트, 이번 위기는 다르다 : 코로나발 경제위기의 특이성과 정책적 함의, 산업연구원(2020.8월).

KOSBI, 포스트 코로나 시대 자영업 생태계 변화 전망과 대응전략, 중소기업연구원(2020.6월).

KOREG RESEARCH

2020년 10월 31일 인쇄 · 발행
제8권 1호(통권 8호)
ISSN 2288-5536 (비매품)
편집처 | 신용보증재단중앙회
발행처 | 신용보증재단중앙회
발행인 | 김병근

편집 및 발간 | 교육연구부 박주완, 배진성
구독 및 게재 문의 | 042-480-4025~6

www.koreg.or.kr

KOREG

신용보증재단중앙회

대전광역시 서구 한밭대로 713 나라키움대전센터
TEL.1588-7365 FAX.(042)480-4007~8