에너지경제연구 제 19 권 제 1 호
Korean Energy Economic Review
Volume 19, Number 1, March 2020 : pp. 117~152

Logistic Lasso를 이용한 에너지·환경산업 기업부도 예측

정기호* · 임희준**

요 약

공공성이 높은 에너지산업과 환경산업(이하 에너지환경산업)에서 기업부도의 사전 예측은 갑작스러운 공급 충격의 가능성을 예측하고 대비할 수 있게 하며 이를 통해 경제 전반의 충격과 국민 후생 저하를 완화시킬 수 있다는 점에서 중요하다. 본 연구는 에너지환경산업에 대한 기업부도 예측모형을 분석한다. 추정모형으로는 로짓모형을 이용하고 부도예측 선행연구에서 자주 이용되는 51개 재무 변수를 초기 설명변수로 고려하였다. 그리고 이들 설명변수에 에너지환경산업 더미변수를 곱하여 추가한 경우와 추가하지 않은 경우의 예측력을 비교하였다. 한편 설명변수 수가 많기 때문에 예상되는 모형의 과다적합(over-fitting)과 다중공선성 문제를 완화하기 위해 모형축소방법인 lasso(Tibshirani, 1996)를 적용하여 로짓 모형과 예측력을 비교하였다. 분석결과, 에너지환경산업 더미변수를 고려한 로짓 모형에 lasso를 적용한 logistic lasso모형의 예측력이 가장 우수한 것으로 나타났다.

주요 단어 : 에너지환경산업, 부도예측, Logistic Lasso
경제학문헌목록 주제분류 : C1

* 경북대학교 경제통상학부(제1저자, 교신저자). khjeong@knu.ac.kr
** 경북대학교 경제학과 대학원생(공동저자). hjlim8623@naver.com

www.kci.go.kr

Ⅰ. 서 론

1997년 외환위기와 2008년 금융위기를 거치면서 우리나라 기업의 부도 확률은 크게 높아졌다. 기업부도는 재무적으로 부실한 기업을 시장에서 퇴출시키고 건실한 기업을 남게 함으로써 경제 전반의 경쟁력과 효율성 그리고 시장 참여자들의 신뢰를 제고하여 시장경제 체제가 건전하게 유지되는데 필요한 제도적 장치이지만 또 다른 한편으로는 지역, 산업, 국가경제 등에 충격과 피해를 미치게 된다.

에너지 산업과 환경 산업(이하 에너지환경산업으로 표기)에서 예상하지 못한 기업 퇴출은 공급 지장을 통해 국가경제와 국민후생 전반에 큰 손실을 미칠 수 있다. 에너지환경산업은 공공성이 높으며 또한 공급이 원활할 때의 경제 기여효과보다 공급이 원활하지 못할 때의 공급지장효과가 비대칭적으로 큰 특징을 갖는다. 동 산업을 구성하는 광산품, 석탄 및 석유제품, 전력·가스·수도·폐기물 산업 등은 대체로 전후방파급효과가 상대적으로 낮은 산업이다(한국은행, 2019). 그러나 이들 산업의 공급이 부족하면 경제나 국민의 일상생활이 중단되거나 큰 불편이 초래될 수 있다. 산업 생산의 1단위 감소에 따른 타 산업의 생산 감소를 나타내는 공급지장효과를 보면 오폐수처리의 경우는 0.8417(박소연 외, 2015)로서 1보다 작지만 신재생에너지발전은 1.7860 그리고 화력발전 1.5966(강지은 외, 2017), 원자력발전 1.747(최용석,조창익, 2019), 석유산업 1.1279(박중구, 2012) 등으로 높다. 2011년 전력 순환단전으로 국민생활과 국가경제 전반에 미쳤던 충격 그리고 2018년 수도권 지역에서의 폐비닐, 혼합플라스틱, 재활용품 수거 중단 등으로 겪었던 시민들의 불편은 단적으로 이들 산업에 공급지장이 발생할 경우의 부정적인 효과를 보여준다. 따라서 공급의 안정성 확보는 이들 산업에 대해 가장 중요하게 고려되어야 할 사안이다. 갑작스러운 기업부도는 조업 중단을 통해 공급의 안정성에 큰 영향을

www.kci.go.kr

미칠 수 있는 요인으로서, 기업부도에 영향을 미치는 요인들을 파악하고 사전에 기업부도를 예측하는 연구는 다른 일반 산업들에 비해 상대적으로 에너지환경산업에서 더 요구된다. 정확한 부도예측을 통해서 기업의 부도 전후에 문제를 인식하고 적절한 대비가 가능해진다. 기업이 부도날 가능성이 있다고 판단하면 손실이 크지 않을 때 여러 대비책을 강구하여 기업을 다시 정상적으로 돌려놓을 수 있고 부도가능성이 높다고 판단하는 경우 사전에 공급지장에 대한 대비를 함으로써 충격을 완화시킬 수 있다.

본 연구는 에너지환경산업에 대한 기업부도 예측모형을 분석한다. 기업부도를 사전에 예측할 수 있다면 예방조치를 통해 부도를 막거나 준비를 통해 부도에 따른 충격과 손실을 최소화하는데 도움이 되는 이유로 기업부도예측 연구는 오랜 기간에 걸쳐 많은 연구가 수행되어왔다(Beaver, 1966; Altman, 1968; Ohlson, 1980; Zmijewski, 1984; Campbell et al., 2008; Lu et al., 2015; Amendola et al., 2015).[1] 그러나 에너지환경산업에 대해서는 저자들의 제한된 검색 범위에서 해외문헌으로 석유가스산업에 대해 Eldahrany(1986)와 Platt et al. (1994)가 있을 뿐이며 국내문헌에서는 없다. 본 논문은 동 산업에 대한 기업부도 예측모형을 연구함으로써 이러한 문헌의 갭을 메꾸고자 한다. 에너지산업과 환경산업의 정의는 각각 통계청에서 제공하는 특수산업분류를 따른다.

본 연구에서 제일 먼저 고려해야 할 이슈는 일반적인 산업들과는 차별적으로 에너지환경산업의 기업들에 대한 부도예측 모형을 별도로 연구할 필요가 있는지 여부이다. 본 연구에서 이 이슈는 기업부도예측 선행연구들에서 자주 이용되는 설명변수들에 추가해서 이들 변수들에 에너지환경산업 더미변수를 곱해서 모형에 포함시키고 이렇게 더미변수를 고려한 모형과 고려하지 않는 모형 간에 에너지환경산업 기업에 대한 부도예측력을 비교하는 방식으로 접근한다.[2]

---

1) 구글스칼라에서 논문으로 한정한 검색결과는 “기업부도예측” 검색어 4,400개, “기업도산예측” 검색어 4140개, “bankruptcy prediction” 검색어 106,000개임(2020년 1월23일 검색기준).

2) Platt and Platt(1990)은 부도예측 선행연구들의 예측정확도가 떨어지는 이유로서 산업간 차이를 지적하고 한 산업의 기업 자료만 분석에 사용할 것을 제시하였으나 해외연구

www.kci.go.kr

한편 이러한 접근 방식은 경험적 분석에서 다른 이슈를 발생시킨다. 본 연구는 로짓(logit) 모형을 이용하고 기업부도예측 선행연구에서 자주 이용되는 51개 설명변수를 고려하는데 여기에 더미변수를 곱해서 추가하면 모형에서 다루어야 하는 설명변수 수는 모두 102개가 된다. 이렇게 많은 수의 설명변수를 모형에 포함시키면 주어진 표본 자료에 대한 설명력은 좋지만 새로운 표본 자료에 대한 예측력은 떨어지는 모형의 과다적합(over-fitting) 문제가 발생한다. 또한 상관관계가 높은 기업 재무변수들의 특성으로 다중공선성 문제가 존재하는 경우에는 t-검정을 통해 유의성이 없는 설명변수를 순방향이나 역방향으로 제거하더라도 모형의 선정 결과나 계수의 추정값이 모두 불안정하다는 점이 잘 알려져 있다(Hastie et al., 2009; Tutz et al., 2012). 본 연구는 모형축소(shrinkage) 방법인 lasso(Least Absolute Shrinkage and Selection Operator)(Tibshirani, 1996)를 적용하여 모형 차원을 축소시킴으로써 모형 과다적합과 다중공선성 문제를 보완한다. lasso는 최소자승추정법(OLS)나 최우추정법(MLE)와 같은 기존의 추정방법에서 사용되는 목적함수에 계수 차원에 대한 벌칙항을 추가함으로써 불필요하게 설명변수가 많아서 계수 차원이 커지는 것을 방지하고 다중공선성이 높은 경우에는 상관성인 높은 변수들을 제거한다(Tian et al, 2015). 적절하게 모형 차원을 축소시키면 예측력이 향상되기 때문에 2015년 이후 기업부도예측 연구에서도 lasso를 사용하는 국외 문헌들이 나타나고 있으며 국내에서도 한편의 선행연구가 발견된다(Tian et al., 2015; Pereira et al., 2016; Tian and Yu, 2017; 차성재·강정석, 2018).

이 논문은 다음과 같이 구성된다. 다음 장에서는 본 연구의 수리적 분석모형인 logistic lasso을 소개하고, 제Ⅲ장에서는 경험적 분석의 세부 내용을 구체적으로 살펴본다. 제Ⅳ장에서는 분석결과를 제시한다. 마지막 제Ⅴ장에서는 논문을 요약하고 결론을 내린다.

에서도 많은 연구들이 자료 부족으로 모든 산업들의 자료를 모두 사용하며(Altman, 1968; Zavgren, 1985), 특히 우리나라 에너지환경산업은 기업 수가 적기 때문에 본 연구는 모든 산업의 자료를 이용하되 산업 더미변수를 고려하는 접근을 채택함.

www.kci.go.kr

Ⅱ. 수리적 분석모형

1. Lasso

lasso(Least Absolute Shrinkage and Selection Operator)는 설명변수가 많은 회귀함수의 문제점을 해결하기 위해 Tibshirani(1996)에 의해 제시된 모형 축소방법이다. 설명변수가 많을수록 주어진 표본자료에 대한 설명력은 높지만 새로운 표본자료에 대한 예측력은 떨어지게 된다. 이 경우에 중요하지 않은 설명변수들의 계수를 0으로 강제적으로 설정하면 약간의 편의(bias)를 희생하는 대신에 예측 분산이 감소해서 모형의 전반적인 예측력을 개선시킬 수 있다(Bickel and Li, 2006). t-검정을 통해 통계적 유의성이 없는 설명변수를 제거할 수도 있지만 설명변수가 많을 경우 예상되는 다중공선성이 존재하면 이론적으로 중요한 변수도 추정량 분산의 증가로 제거될 수 있다(Hill et al., 2011). lasso는 식 (1)과 같이 최소자승추정법의 목적함수에 계수 절대값의 합을 벌칙항(penalty)으로 추가함으로써 불필요한 설명변수들의 계수를 0으로 추정하고 이에 따라 불필요한 모형의 차원 증가를 방지하며 다중공선성이 높은 경우에는 상관성인 높은 변수들을 제거하는 효과를 갖는다(Tian et al, 2015).

$$\hat{\beta}^{lasso} = \arg\min \left\{ \sum_{i=1}^{n} (y_i - \beta_o - \sum_{j=1}^{k} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{k} |\beta_j| \right\} \quad (1)$$

n : 표본자료 크기
k : 설명변수 수

식 (1) 우변의 중괄호에서 첫 번째 항은 최소자승추정법의 목적함수인 오차제곱합이고 두 번째 항이 계수차원에 대한 벌칙항을 나타낸다. λ는 벌칙 정

www.kci.go.kr

도를 제어하는 모수로서 λ 값이 클수록 차원축소가 커지고 많은 수의 계수들이 0의 값을 갖게 된다. 반대로 λ의 값이 작을수록 차원축소는 적게 이루어지며 0이면 통상적인 최소자승추정법과 추정결과가 같다.

lasso의 목적이 예측력 향상이므로 이론적으로 λ 값은 예측오차제곱 기댓값(expected squared prediction error)을 최소화하도록 결정하며, 경험 분석에서는 예측오차제곱 기댓값을 보통 k겹 교차검증법(k-fold cross-validation)에 의해 추정한다(Efron and Tibshirani, 1993; Hastie et al., 2001).[3] k겹 교차검증법은 표본자료를 동일한 크기의 k개 소그룹으로 랜덤하게 구분하여 (k-1)개 소그룹(훈련자료, training set)을 추정에 사용하고 나머지 1개 소그룹(평가자료, validation set)을 추정결과의 예측오차 계산에 사용한다. 평가자료로 이용되는 소그룹은 k개 소그룹에 대해 순차적으로 적용되며 예측오차제곱 기댓값은 k개의 예측오차 그룹에 대한 산술평균으로 추정된다.[4]

2. Logistic Lasso

종속변수가 0 혹은 1의 값을 갖는 이항변수인 경우에 선형회귀모형은 예측값이 [0,1] 범위를 벗어나거나 이분산 문제 등 여러 문제를 갖는다(Hill et al., 2011). 로짓 모형은 이러한 상황에서 이용될 수 있는 모형으로서 기업부도예측 연구에서 많이 이용된다.[5] 설명변수가 많을 경우에 모형 과다적합과 다중공선성 문제는 로짓 모형에서도 발생하는데 lasso는 이 경우에도 선형회귀모형에서와 유사하게 적용될 수 있으며 lasso가 적용된 로짓 모형이 logistic

3) 종속변수가 Y이고 회귀모형이 $\eta(X)$, 추정결과가 $\hat{\eta}(X)$일 때 오차제곱 기댓값은 $E\{\eta(X)-\hat{\eta}(X)\}^2$이고 예측오차제곱 기댓값은 $E\{Y-\hat{\eta}(X)\}^2$임(Tibshirani, 1996).

4) lasso의 λ와 같은 hyperparameter의 결정 외에도 최종모형의 예측과 같은 성능을 최종 평가하기 위해서는 전체 표본자료를 훈련자료(training set), 평가자료(validation set), 검정자료(test set)로 세분할 필요가 있음. 본 연구에서 이러한 표본자료의 구분과 k겹 교차검증법의 구체적인 적용과정은 제 Ⅲ장에서 설명됨.

5) 구글스칼라에서 논문으로 한정하여 “bankruptcy prediction logit” 검색어로 검색한 결과는 24,300개임(2020년 1월23일 검색기준).

www.kci.go.kr

lasso이다. 아래 식 (2)와 식 (3)은 각각 로짓 모형과 logistic lasso 모형에서 사용되는 로그우도함수를 보여준다(Pereira et al., 2016).

$$l(\beta) = \sum_{i=1}^{n} \{ y_i \log(\pi_i) + (1-y_i) \log(1-\pi_i) \} \quad (2)$$
$$= \sum_{i=1}^{n} \left\{ y_i x_i \beta - \log(1+e^{x_i \beta}) \right\}$$

$$l(\beta, \lambda) = \sum_{i=1}^{n} \left\{ y_i x_i \beta - \log(1+e^{x_i \beta}) \right\} - \lambda \sum_{j=1}^{k} |\beta_j| \quad (3)$$

단, $\pi_i = \Pr(y_i = 1) = e^{x_i \beta} / (1+e^{x_i \beta})$

식 (3)에서 두 번째 항이 계수차원에 대한 벌칙항을 나타낸다. 최소자승법에 적용되는 식 (1)에서는 목적함수가 최소화되기 때문에 벌칙항을 더하지만 최우추정법이 적용되는 식 (3)에서는 목적합수가 최대화되기 때문에 벌칙항을 제하는 형태로 추가한다.

www.kci.go.kr

Ⅲ. 경험적 분석모형

1. 부도 정의

기업부도예측 연구에서 가장 먼저 결정해야할 것은 부도의 정의이다. 부도에 관한 해외와 국내 간의 법적 차이가 있기 때문에 본 연구에서는 국내 문헌에서 주로 사용되는 정의를 따른다. 금융투자업규정(제8-19조의9제3항제2호)에 따른 법률적 부도의 정의는 원리금의 적기상환이 이루어지지 않거나 기업회생절차 파산절차의 개시가 있는 경우이다.[6] 그러나 기업부도예측의 선행연구에서는 주로 한국거래소의 규정에 따라 관리종목에 지정되거나(이계원, 1993; 장휘용, 1998; 남재우·이회경·김동석, 2000; 전성빈·김영일, 2001), 상장폐지된 기업(이건창·김명종·김혁, 1994; 이인로·김동철, 2015; 권혁건·이동규·신민수, 2017; 차성재·강정석, 2018)을 부도기업에 포함시킨다. 본 연구는 이인로·김동철(2015), 권혁건·이동규·신민수(2017), 차성재·강정석(2018) 등을 따라 실적부진의 사유로 인해 상장폐지된 기업 또는 관리종목으로 지정된 기업을 부도기업으로 정의한다. 표본자료에서 부도기업의 수가 상대적으로 크게 적은 경우에는 부도확률이 과소 추정되는 문제가 있는데 부도의 범위를 상장폐지와 관리종목 지정을 모두 포함하도록 확대하면 이러한 문제를 완화시키는 효과를 갖는다. 본 연구의 경우 상장폐지만을 고려하면 부도기업 사례가 420건(에너지환경산업: 46, 기타산업: 374)이지만 관리종목지정까지 포함하면 712건(에너지환경산업: 86, 기타산업: 626)으로 증가한다.

6) http://www.law.go.kr/%ED%96%89%EC%A0%95%EA%B7%9C%EC%B9%99/%EA%B8%88%EC%9C%B5%ED%88%AC%EC%9E%90%EC%97%85%EA%B7%9C%EC%A0%95

www.kci.go.kr

2. 설명변수

경험적 부도예측 연구에서 다음으로 결정해야 할 사항은 예측 모형에서 사용될 설명변수이다. 선행연구에서 사용되는 설명변수의 수와 변수내용은 문헌마다 차이가 있다. 국내외 선행연구에서 설명변수의 선정 기준으로 자주 인용되는 Beaver(1966), Altman(1968), Zmijewski(1984), Platt and Platt(1990), Shumway(2001), Chava and Jarrow(2004), Campbell et al.(2008) 등의 문헌들은 설명변수 후보로서 대체로 30개 이하의 변수들을 고려한다. 한편 해외 선행연구(19개)와 국내 선행연구(28개)에서 설명변수로 일회 이상 고려된 설명변수들의 수는 모두 358개에 달한다. 이것은 그만큼 문헌마다 사용하는 변수들에 큰 차이가 있음을 나타내는데 이러한 358개 변수들을 모두 고려할 수 없기 때문에 설명변수의 선택은 불가피하다. 본 논문은 이들 358개 변수들의 사용횟수를 측정해서 5회 이상 사용된 총 51개 변수를 초기 변수후보로 고려한다.[7] <표 1>은 이렇게 선정된 변수들을 보여준다.

〈표 1〉 본 연구의 설명변수 초기 후보

번호	변수 이름	변수 정의
1	총자산 증가율	(총자산(t기)-총자산(t-1기))/총자산(t-1기)
2	유형고정자산 증가율	(유형자산(t기)-유형자산(t-1기))/(유형자산(t-1기)
3	매출액 증가율	(매출액(t기)-매출액(t-1기))/매출액(t-1기)
4	경상이익 증가율	(경상이익(t기)-경상이익(t-1기))/경상이익(t-1기) 단, 경상이익 = 영업이익(영업수익-영업비용)+비영업수익-비영업비용.
5	순이익 증가율	(당기순이익(t기)-당기순이익(t-1기))/당기순이익(t-1기)
6	고정자산 증가율	(비유동자산(t기)-비유동자산(t-1기))/비유동자산(t-1기)
7	총자본 경상이익율	경상이익/총자본
8	총자본 순이익율	당기순이익/총자본
9	매출액 경상이익율	경상이익/매출액

7) 해외와 국내 선행연구의 선정기준은 구글 스칼라에서 검색한 타 문헌에서의 인용횟수임 (2019년 2월 17일 검색 기준).

www.kci.go.kr

번호	변수 이름	변수 정의
10	매출액 순이익율	당기순이익/매출액
11	매출액 영업이익율	영업이익/매출액
12	매출액 이자비용율	이자비용/매출액
13	매출액 총이익율	매출총이익/매출액
14	EBIT/총자산[1)]	EBIT/총자산
15	총자산 경상이익율	경상이익/총자산
16	총자산 순이익율	당기순이익/총자산
17	금융비용/총부채	이자비용/총부채
18	영업활동이익/총부채	영업활동이익/총부채 단, 영업활동이익 = 세전순이익(pretax income)+감가상각비
19	자기자본비율	자기자본(총자본)/총자산
20	고정비율(fixed ratio)	고정자산(비유동자산)/자기자본(총자본)
21	고정장기적합율	고정자산(비유동자산)/(자기자본(총자본)+장기부채(비유동부채))
22	총부채/총자산	총부채/총자산
23	부채비율	총부채/자기자본(총자본)
24	유동부채비율	유동부채/자기자본(총자본)
25	고정부채비율	고정부채(=장기부채=비유동부채)/자기자본(총자본)
26	이자보상비율	EBIT/이자비용
27	이익잉여금/총자산	이익잉여금/총자산
28	유동부채/총자산	유동부채/총자산
29	현금비율(cash ratio)	현금및현금성자산/유동부채
30	현금/총자산	현금/총자산
31	유동부채/총부채	유동부채/총부채
32	유동비율(current ratio)	유동자산/유동부채
33	당좌비율(quick ratio)	당좌자산/유동부채
34	운전자본/총자산	운전자본/총자산 단, 운전자본=유동자산-유동부채.
35	주식시장가치/총부채	시가총액/총부채
36	유동자산/총자산	유동자산/총자산
37	현금흐름/부채	현금흐름/총부채 단, 현금흐름=당기순이익+유무형자산상각비
38	현금흐름/총자본	현금흐름/총자본

www.kci.go.kr

번호	변수 이름	변수 정의
39	현금흐름/매출액	현금흐름/매출액
40	현금흐름/총자산	현금흐름/총자산
41	총자본회전율	매출액/총자본
42	순운전자본회전율	매출액/운전자본
43	고정자산회전율 (fixed assets turnover)	매출액/고정자산(비유동자산)
44	재고자산회전율 (inventories turnover)	매출액/재고자산
45	총자산회전율 (total assets turnover)	매출액/총자산
46	매출채권회전율	매출액/매출채권
47	매입채무회전율 (payable turnover)	매출액/매입채무
48	log(매출액)	log(매출액)
49	log(총자산)	log(총자산)
50	Shumway(2001)의 relative size	기업시가총액/KOSPI와 KOSDAQ 시장시가총액
51	당좌자산/총자산	당좌자산/총자산

주: 1) EBIT(Earnings Before Interest and Tax) : 이자 및 법인세를 제하기 이전 이익

3. k겹 교차검증법

모형은 특정한 자료에서만 우연히 좋은 성과를 나타낼 수도 있기 때문에 다양한 자료에 적용해서 일반적인 성과를 평가할 필요가 있다. 이런 이유로 최종적으로 추정된 모형의 예측력을 평가하기 위해서는 일반적으로 전체 표본자료를 훈련자료(training set)와 검정자료(test set)로 분할하고 모형 추정은 전자에 대해서 그리고 추정된 모형의 성능 평가는 후자에 대해서 각각 다른 자료를 이용하여 한다. 로짓 모형의 경우 추정과 최종모형의 예측력 평가 작업만 필요하기 때문에 표본자료를 훈련자료와 검정자료로 구분하면 된다.

그러나 lasso의 경우에는 축소척도를 제어하는 λ의 값을 일반적인 계수들의 추정에 앞서 결정할 필요가 있으며 이 경우에는 훈련자료를 다시 일반적

www.kci.go.kr

인 계수들의 추정에 이용하는 자료와 λ값의 결정에 이용하는 평가자료(validation set)로 세분할 필요가 있다. [그림 1]에서 로짓 모형의 경우 전체 표본자료를 훈련자료와 검정자료의 2개 소그룹으로 분할하는 반면에 logistic lasso 모형에서는 전체 표본자료를 훈련자료, 평가자료, 검정자료의 3개 소그룹으로 분할하는 것을 시각적으로 보이고 있다.

[그림 1] 표본자료의 구분

<table>
<tr><td></td><td colspan="3">전체 표본자료</td></tr>
<tr><td>로짓</td><td colspan="2">훈련자료(training data)</td><td>검정자료(test data)</td></tr>
<tr><td>logistic lasso</td><td>훈련자료(training data)</td><td>평가자료(validation data)</td><td>검정자료(test data)</td></tr>
</table>

한편 평가와 검정은 다양한 자료에 대해서 수행되어야 하므로 표본자료의 분할 역시 한 번에 그치지 않고 여러 번 수행할 필요가 있다. 이런 상황에서 사용되는 기법이 k겹 교차검증법(k-fold cross-validation)이다(Efron and Tibshirani, 1993; Hastie et al., 2001). k겹 교차검증법은 전체 자료를 동일한 크기의 k개 소그룹으로 랜덤하게 구분한 다음에 (k-1)개 소그룹을 추정에 사용하고 나머지 1개 소그룹을 평가나 검정의 목적에 사용하고 후자에 사용되는 1개 소그룹을 k개 소그룹에 대해 순차적으로 적용하는 방법이다. 보통 k값으로는 5나 10을 사용하며(Hastie et al., 2001) 본 연구에서는 10겹 교차검증법을 사용한다. 다음은 본 연구의 경험적 분석에서 logistic lasso에 대해 10겹 교차검증법을 적용하는 과정이다.

(1) 전체 표본자료를 랜덤하게 10등분한다.

(2) 첫 번째 소그룹을 검정 목적으로 사용하고 나머지 9개 소그룹을 훈련과 평가에 사용한다.

(3) λ를 특정한 값으로 고정한다.

① (2)에서 결정된 9개 소그룹을 합친 다음에 다시 랜덤하게 10개의 소그룹으로 균등하게 분할한다.

www.kci.go.kr

② ①에서 구한 10개 소그룹 중 9개를 이용해서 λ를 제외한 일반 계수들을 식 (3)을 극대화하도록 추정하고 나머지 1개 소그룹에 대해 종속변수를 예측하고 실제 값과 비교해서 예측오차제곱 평균값을 계산한다.

③ 위 ②의 작업을 나머지 9개 소그룹의 각각에 대해 순차적으로 적용하여 모두 10개의 예측오차제곱 평균값을 도출한다.

④ ③의 10개 예측오제곱 평균값을 산술 평균해서 λ의 고정된 값에 대한 예측오차제곱 기댓값을 추정한다.

(4) (3)의 작업을 λ의 다른 값들에 대해 반복 적용하고 예측오차제곱 기댓값의 추정값을 가장 최소화하는 λ의 최적값을 찾는다.

(5) (4)에서 결정된 최적값으로 λ값을 고정하고 (2)의 9개 소그룹 전체 자료를 이용해서 일반 계수를 추정한다. 이 단계에서는 λ값에 따라서 얼마나 많은 설명변수들의 계수가 0으로 설정되는지가 결정된다.

(6) (5)에서 추정된 모형을 (2)에서 검정목적으로 남겨진 1개 소그룹 자료에 적용해서 종속변수의 값을 예측하고 실제 값과 비교해서 최종 모형의 예측력 평가지표를 계산한다.

(7) (3)~(6) 과정을 (2)의 나머지 9개 소그룹에 대해서도 순차적으로 적용하여 예측력 평가지표의 10개 값을 구하고 이들 값을 평균하여 최종적인 예측력 평가지표 값을 계산한다.

4. 모형 예측력 평가지표

본 연구의 경험적 분석에서 모두 5개 유형의 모형을 고려하는데 에너지환경산업의 기업부도에 대한 이들 모형들의 예측력을 비교하기 위해서는 올바른 모형예측력 평가지표가 중요하다. 본 연구에서는 Accuracy, ROC AUC, PRAUC, MCC 등 4개의 예측력 평가지표를 이용한다.

예측력 평가지표에서 중요한 역할을 하는 혼동행렬(confusion matrix)은 이진 분류(binary classification)에서 실제 소속 집단과 예측결과를 비교하여 정

www.kci.go.kr

리한 2x2 행렬이다. 부도기업 집단을 Positive로 그리고 정상기업 집단을 Negative로 표시하면 [그림 2]는 실제 그룹과 예측결과를 비교하여 가능한 네 개의 경우의 혼동행렬을 보여준다. 표에서 대각항은 소속 집단이 올바르게 예측된 경우이며 비대각항은 틀리게 예측된 경우이다.

[그림 2] 혼동행렬

		실제 그룹	
		Positive	Negative
예측 그룹	Positive	True Positive(TP)	False Positive(FP)
	Negative	False Negative(FN)	True Negative(TN)

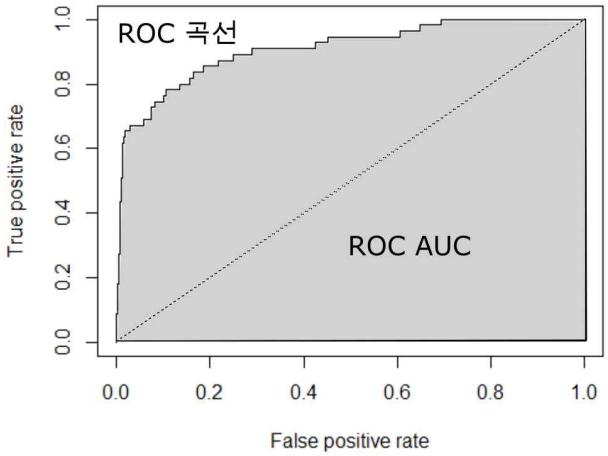
Accuracy는 분류문제에서 가장 많이 사용되는 예측평가지표이며 기업부도 예측 문헌에서도 많이 사용되어왔다(Altman, 1968; Ohlson, 1980; Zmijewski, 1984; Barboza et al., 2017). Accuracy는 전체 표본자료 중 소속 집단이 올바르게 예측된 비율로서 정의되며 계산식은 식 (4)와 같다.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

현실에서는 일반적으로 부도기업의 관측자료보다 정상기업의 관측자료가 훨씬 많기 때문에 Accuracy만을 이용하면 예측력을 과대평가할 수 있다. 예를 들어 본 연구의 분석 자료에서는 부도에 해당하는 관측 자료는 전체 표본자료에서 약 3.8%에 불과하며 이것은 모형이 기업을 전부 부도가 발생하지 않은 정상기업으로 예측해도 정확도가 96%로 높게 되는 것을 의미한다.

www.kci.go.kr

[그림 3] ROC곡선과 ROC AUC



ROC AUC(Receiver Operating Characteristics Area Under Curve)도 역시 이진분류에서 예측평가지표로 많이 사용되는 평가지표 중 하나이다(Tian et al., 2015; Barboza et al., 2017). ROC AUC를 정의하기 위해서는 먼저 참양성비율(True Positive Rate, TPR)과 허위양성비율(False Positive Rate, FPR)을 정의할 필요가 있다.

$$\text{TPR} = \frac{TP}{TP+FN} \tag{5}$$

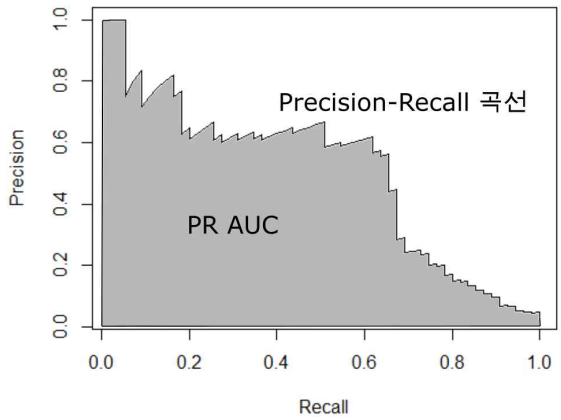
$$\text{FPR} = \frac{FP}{TN+FP} \tag{6}$$

TPR은 실제로 부도기업인 집단 중에서 모형이 부도기업을 얼마나 잘 찾아내는지를 측정하며 통계학에서는 민감도(sensitivity)라고 한다. 1-FPR은 진짜 정상기업 집단 중에서 모형이 정상기업을 얼마나 잘 분류하는지를 측정하며 통계학에서 특이도(specificity)라고 한다. 부도예측모형에서 분류 임계값을 낮출수록 더 많은 기업이 부도기업으로 분류되므로 민감도는 증가하고 특이도는 감소해서 두 값은 반대 방향으로 움직인다. ROC는 민감도인 TPR을 종축에

www.kci.go.kr

놓고 FPR을 횡축에 놓고 그린 곡선으로서 양(+)의 기울기를 갖도록 횡축을 특이도인 1-FPR대신에 FPR을 놓았다. ROC AUC는 ROC곡선 아래의 면적을 나타내며, 이 값이 클수록 주어진 특이도에 대해 민감도가 높은 것을 의미한다.

[그림 4] PR곡선과 PR AUC



민감도는 검출율(recall)로도 불리는데 부도예측모형의 목적이 실제 부도기업을 잘 검출하는 점에서 후자의 용어가 사용된다. 한편 만약 모형이 모든 기업을 부도기업으로 예측하면 FP와 TP 모두 증가하므로 검출율이 증가하며, 이 경우에는 식 (6)에서 정의되는 정밀도(precision)를 같이 고려한다.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{7}$$

정밀도는 모형이 부도기업으로 예측한 것 중에서 실제로 부도기업인 경우의 비율이다. 식 (5)와 식 (7)을 비교하면 검출율과 정밀도는 분모에서 분류 오류에 해당하는 FN과 FP를 사용하는 차이만 있다. 통계학에서 오류는 제1종 오류와 제2종 오류로 구분하며 두 오류는 동시에 최소화할 수 없으므로 검출율과 정밀도 역시 동시에 최대화할 수 없다. Precision-Recall(PR) 곡선은 검출율을 횡축에 그리고 정밀도를 종축에 표시하여 두 값 간의 트레이드오프 관계를

www.kci.go.kr

보여주는 곡선이다. Saito and Rehmsmeier(2015)은 PR곡선이 올바르게 부도기업을 예측한 경우(TP)를 2개의 방향에서 평가하므로 ROC곡선보다 더 효과적이라고 주장하고 PR곡선 아래의 면적인 PR AUC를 제시하였다.

정밀도와 검출율(민감도) 모두 부도예측의 관심대상인 올바른 부도기업 예측(TP)에만 초점을 맞추고 정상기업의 올바른 예측(TN)은 고려하지 않는다. 식 (4)의 Accuracy는 부도기업과 정상기업 모두에 대한 올바른 예측(TP, TN)을 고려하지만 정상기업의 수가 부도기업의 수보다 월등 큰 표본자료에서는 정상기업의 예측 정확도만 크게 높고 부도기업의 예측 정확도는 크게 낮은 경우에도 Accuracy의 값이 높게 계산될 수 있다. 메튜상관계수(Matthews correlation coefficient, MCC)는 다중분류문제에서 실제값과 예측값 간의 상관도를 계산하는 통계량으로서 부도예측과 같은 이진분류에서는 아래 식 (8)과 같이 계산된다.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{8}$$

이진분류에서 그룹 간의 관측자료 크기가 크게 불균형인 경우에는 MCC가 다른 이진분류 평가지표보다 좋은 것으로 알려져 있다(Chicco,2017).

본 연구에서는 위에서 논의된 Accuracy, ROC AUC, PRAUC, MCC 등 4개의 예측력 평가지표를 이용해서 부도기업예측 모형들을 평가한다.

www.kci.go.kr

Ⅳ. 분석 결과

1. 자료

본 연구에서 분석대상은 2000-2018년 기간에 우리나라 코스피 및 코스닥 시장에 상장된 비금융기업들로서 자료는 에프엔가이드(FnGuide)의 금융데이터베이스인 Dataguide 5.0에서 추출한 분기별 자료이다.[8] 금융기업을 포함시키지 않은 이유는 비금융기업과 부도가 발생하는 환경이 다르며(Ohlson, 1980), 재무제표 양식 또한 다르기 때문이다(이인로·김동철, 2015).

본 연구에서 에너지환경산업에 포함되는 에너지산업과 환경산업은 각각 통계청이 제공하는 특수분류를 따른다.[9] 통계청에 의한 에너지산업의 범위는 '에너지원의 획득, 발전 및 송배전, 전기 가스 사업'이며, 환경산업은 '물·공기·토양의 환경적 유해요인과 폐기물·소음·환경시스템과 관련된 문제를 측정, 방지, 조절 및 최소화할 수 있는 재화 및 서비스를 생산하는 산업'이면서 '환경위험을 감소시키고 오염 및 자원이용을 최소화시키는 청정기술, 재화 및 서비스'를 포함한다. 이들 산업에 속하는 기업은 통계청에서 제공하는 특수분류와 제10차 한국표준산업분류의 연계표를 참고하여 FnGuide의 Dataguide 5.0에서 추출했다.

이상치 처리는 적은 수의 이상치들이 모형의 추정결과에 큰 영향을 미치는 것을 막기 위해 필수적인 데이터 전처리 과정 중 하나이다. 본 연구에서는 Shumway(2001), Wu *et al.*(2010), Foster and Zurada(2013) 등을 따라 설명

8) Dataguide 5.0의 설치와 사용법은 http://www.dataguide.co.kr/DG5web/intro_how.asp을 참조.
9) 통계청의 특수분류에서 에너지산업과 환경산업의 정의는 다음 사이트를 참조.
http://kssc.kostat.go.kr/ksscNew_web/kssc/common/selectIntroduce.do?gubun=2&bbsId=energe_ug
http://kssc.kostat.go.kr/ksscNew_web/kssc/common/selectIntroduce.do?gubun=2&bbsId=env_ug

www.kci.go.kr

변수별로 자료의 양 극단 1%에 대해 절단을 실시했다.

최종적으로 얻어진 분석대상 기업 수는 2,290개이고 이들 기업에 대한 관측자료 수는 총 104,486개이다. 이 중 실적부진의 사유로 상장폐지된 경우의 관측자료 수는 420개이고 관리종목지정된 경우의 관측자료 수는 3,739개이며 상장폐지와 관리종목 지정이 동시에 된 경우는 130개이다. 본 연구의 부도 정의는 실적부진의 사유로 인한 상장폐지와 관리종목지정을 모두 포함하므로 부도에 해당하는 관측자료 수는 총 4,029개가 되며 부도가 발생하지 않은 관측자료 수는 총 100,457개이다.

분석대상인 2,290개 기업 중 에너지환경산업에 포함되는 기업은 348개이며 이중에서 환경산업에 포함되는 기업은 320개이고 에너지산업에 포함되는 기업은 39개이며 두 산업에 동시에 포함되는 기업은 11개이다. 분석 기간 동안 이들 에너지환경산업 기업에 관한 관측자료 수는 총 16,570개이다. 이 중 관리종목지정된 경우의 관측자료 수는 483개이고 상장폐지된 경우의 관측자료 수는 46개이므로 에너지환경산업에서 본 연구의 부도 정의에 해당하는 관측자료 수는 총 520개이다. 세부적으로는 환경산업 기업에 대해서 관리종목지정의 관측자료 수는 433개이고 상장폐지의 관측자료 수는 45개이며 에너지산업 기업에 대해서 관리종목지정의 관측자료 수는 50개이고 상장폐지의 관측자료 수는 1개이다. 경험적 분석에서 에너지환경산업에 대한 더미변수는 에너지환경산업에 속하면 1의 값을 갖고 그렇지 않으면 0의 값을 갖도록 정의된다.

〈표 2〉 분석대상 기업

		에너지환경산업			다른 산업	계
		에너지 (C)	환경 (D)	C+D		
부도(O)	관리종목지정(A)	7	64	71	527	598
	상장폐지(B)	1	45	46	374	420
	A+B	7	79	86	626	712
부도(X)		32	241	262	1,316	1,578
계		39	320	348	1,942	2,290

www.kci.go.kr

〈표 3〉 분석대상 관측자료

		에너지환경산업			다른 산업	계
		에너지 (C)	환경 (D)	C+D		
부도(O)	관리종목지정(A)	50	432	482	3,256	3,728
	상장폐지(B)	1	45	46	374	420
	A+B	51	469	520	3,509	4,029
부도(X)		2,117	14,580	16,050	84,407	100,457
계		2,168	15,049	16,570	87,916	104,486

2. 경험적 모형의 유형

lasso의 적용 여부와 에너지환경산업의 더미변수의 적용 여부에 따라 모두 4개의 예측모형을 고려할 수 있다.

[그림 5] 예측모형의 유형

	에너지환경산업 더미변수(O)	에너지환경산업 더미변수(X)
로짓	모형 1	모형 2
logistic lasso	모형 3	모형 4

모형 1은 에너지환경산업 더미변수를 고려하는 로짓 모형이다. 모형 1에 대해서는 초기 변수집합인 51개 설명변수를 사용해서 로짓 모형을 추정한 후 t-검정을 실시하여 10% 유의수준에서 유의성이 없는 변수를 제거한다. 유의한 변수들과 그 변수들에 에너지환경산업 더미변수를 곱한 변수들을 추가하여 다시 로짓 모형을 추정한다. 그리고 10% 유의수준의 t-검정을 통해 유의한 변수들만으로 구성된 최종 모형을 추정한 후 에너지환경산업 기업에 대해 부도예측을 실시한다.

www.kci.go.kr

모형 2는 모형 1과 유사하되 에너지환경산업 더미변수를 고려하지 않은 로짓 모형이다. 모형 2에 대해서는 51개 설명변수를 사용해서 로짓 모형을 추정한 후 10% 유의수준의 t-검정을 통해 유의한 변수들만으로 구성된 최종 로짓 모형을 추정하여 에너지환경산업 기업에 대해 부도예측을 실시한다. 따라서 모형2는 에너지환경산업과 다른 산업의 부도예측모형이 동일하다는 가정을 반영한 모형이다.

모형 3은 에너지환경산업 더미변수를 고려하는 logistic lasso 모형이다. 모형 3에 대해서는 먼저 초기 변수집합인 51개 설명변수로 구성된 logistic lasso를 추정하여 변수를 선택한다. lasso의 λ는 R 패키지인 glmnet을 이용하였으며 10겹-교차검증법에 의해 결정한다. logistic lasso에 의해 선택된 변수와 그 변수들에 에너지환경산업 더미변수를 곱한 변수들로 구성된 모형에 다시 logistic lasso를 적용하여 최종적으로 변수를 선정하고 에너지환경산업 기업에 대해 부도예측을 실시한다.

모형 4는 모형 3과 유사하되 에너지환경산업 더미변수를 고려하지 않은 logistic lasso 모형이다.

본 연구는 위의 4개 모형에 추가해서 혼합모형을 하나 더 고려한다. 모형 5는 에너지환경산업 더미변수를 고려하는 logistic lasso 모형에서 출발하는 것은 모형 3과 동일하다. 그러나 초기 변수집합인 51개 설명변수로 logistic lasso를 추정하여 변수를 선택하고 선택된 변수와 그 변수들에 에너지환경산업 더미변수를 곱한 변수들로 구성된 모형을 logistic lasso가 아니라 로짓 모형으로 추정한 다음에 10% 유의수준의 t-검정을 통해 유의한 변수들만으로 구성된 최종 모형을 추정한 후 에너지환경산업 기업에 대해 부도예측을 실시하는 점에서 모형 3과 차이가 있다.[10)]

10) Logistic Lasso의 변수선택 기준은 예측력인 반면에 로짓 모형에서는 t-검정을 통한 통계적 유의성이 변수선택 기준이다. 설명변수가 많을 경우 t-검정을 통한 변수 선정 과정은 불안정하고 이에 따라 추정결과도 불안정하다는 단점이 있다. 한편 Logistic Lasso도 훈련자료와 평가자료를 분할할 때 보통 5개나 10개 소그룹만을 고려하는 k겹 교차검증법을 사용하므로 분할에 대한 모든 경우를 고려하지는 않는 문제가 있다. 혼합모형 접근은 51개의 많은 설명변수 후보를 줄이는 첫 단계에서는 Logistic Lasso를

www.kci.go.kr

3. 예측력 분석결과

<표 4>~<표 8>은 에너지환경산업 기업의 부도 예측에 대한 5개 모형의 예측력 성과를 보여준다. 제 Ⅲ장 3절에서 소개한 10겹-교차검증법에 의해 전체 표본자료를 랜덤하게 10개로 균등하게 구분하고 9개 소그룹을 훈련자료(training set)로 그리고 나머지 1개 소그룹을 검정자료(test set)를 사용하여 평가지표를 계산한 다음에 검정자료를 순차적으로 나머지 9개 소그룹에 대해서도 적용한 결과이다. <표 9>는 각 모형에 대해 각 지표의 평균값만을 따로 정리하였다.

〈표 4〉 모형1의 예측평가지표

10-fold	Accuracy	ROC AUC	PR AUC	MCC
1	0.900	0.763	0.156	0.104
2	0.969	0.857	0.341	0.352
3	0.861	0.676	0.048	0.001
4	0.967	0.865	0.172	0.171
5	0.891	0.735	0.100	0.149
6	0.933	0.785	0.069	0.038
7	0.871	0.759	0.164	0.085
8	0.970	0.910	0.409	0.311
9	0.958	0.860	0.149	0.171
10	0.860	0.733	0.199	0.079
평균	0.918	0.794	0.181	0.146

사용해서 다중공선성과 과적합 문제를 완화시키고 어느 정도 줄어든 설명변수들에 더미변수를 곱해 새로운 변수를 추가한 이후 둘째 단계에서는 로짓 모형을 적용함으로써 두 모형의 장단점을 보완할 수 있을 것으로 기대된다.

www.kci.go.kr

〈표 5〉 모형2의 예측평가지표

10-fold	Accuracy	ROC AUC	PR AUC	MCC
1	0.961	0.892	0.310	0.156
2	0.970	0.917	0.467	0.338
3	0.974	0.932	0.291	0.226
4	0.972	0.928	0.274	0.130
5	0.965	0.897	0.353	0.271
6	0.973	0.874	0.271	0.206
7	0.969	0.885	0.399	0.351
8	0.969	0.926	0.499	0.285
9	0.968	0.934	0.277	0.169
10	0.966	0.913	0.400	0.236
평균	0.969	0.910	0.354	0.237

〈표 6〉 모형3의 예측평가지표

10-fold	Accuracy	ROC AUC	PR AUC	MCC
1	0.962	0.884	0.310	0.173
2	0.970	0.913	0.462	0.361
3	0.972	0.929	0.287	0.123
4	0.970	0.927	0.255	0.086
5	0.964	0.898	0.305	0.241
6	0.974	0.888	0.276	0.231
7	0.968	0.898	0.408	0.320
8	0.968	0.931	0.483	0.237
9	0.967	0.932	0.300	0.193
10	0.966	0.900	0.332	0.181
평균	0.968	0.910	0.342	0.215

www.kci.go.kr

〈표 7〉 모형4의 예측평가지표

10-fold	Accuracy	ROC AUC	PR AUC	MCC
1	0.961	0.889	0.312	0.142
2	0.971	0.917	0.454	0.349
3	0.974	0.927	0.291	0.131
4	0.971	0.925	0.267	0.088
5	0.963	0.890	0.293	0.160
6	0.974	0.875	0.257	0.195
7	0.969	0.888	0.375	0.329
8	0.969	0.927	0.497	0.256
9	0.968	0.934	0.277	0.179
10	0.966	0.907	0.350	0.204
평균	0.968	0.908	0.337	0.203

〈표 8〉 모형5의 예측평가지표

10-fold	Accuracy	ROC AUC	PR AUC	MCC
1	0.962	0.893	0.312	0.268
2	0.970	0.910	0.484	0.390
3	0.974	0.941	0.335	0.289
4	0.973	0.936	0.282	0.272
5	0.964	0.896	0.377	0.300
6	0.974	0.888	0.315	0.237
7	0.970	0.894	0.438	0.396
8	0.969	0.928	0.488	0.304
9	0.969	0.931	0.339	0.290
10	0.967	0.906	0.404	0.293
평균	0.969	0.912	0.377	0.304

www.kci.go.kr

〈표 9〉 모형들의 예측평가지표 평균

모형	Accuracy	AUC	PRAUC	MCC
1	0.918	0.794	0.181	0.146
2	0.969	0.910	0.354	0.237
3	0.968	0.910	0.342	0.215
4	0.968	0.908	0.337	0.203
5	0.969	0.912	0.377	0.304

<표 9>에서 모형 5가 모든 예측평가지표에서 가장 우수한 것으로 나타났다. 모형 5는 초기 변수집합인 51개 설명변수로 logistic lasso를 추정하여 변수를 선택하고 선택된 변수와 그 변수들에 에너지환경산업 더미변수를 곱한 변수들로 구성된 모형을 logistic lasso가 아니라 로짓 모형으로 추정한 다음에 10% 유의수준의 t-검정을 통해 유의한 변수들만으로 구성된 최종 모형을 추정한 후 에너지환경산업 기업에 대해 부도예측을 실시하는 모형이다. 모형 5는 특히 본 연구의 분석자료와 같이 부도기업과 정상기업 간 관측자료의 불균형이 큰 경우에 올바르게 예측력을 평가하는 지표로 알려진 메튜상관계수(Matthews correlation coefficient, MCC)가 다른 모형에 비해 크게 높다.

한편 로짓 모형 중에서 에너지환경산업 더미변수를 사용하는 모형 1과 사용하지 않은 모형 2를 비교하면 후자의 모형이 더 우수한 것으로 분석되는 반면에, logistic lasso 모형인 모형 3과 4를 비교하면 에너지환경산업 더미변수를 사용한 모형 3이 사용하지 않은 모형 4보다 더 우수하여서 분석 방법에 따라서 에너지환경산업 더미변수의 사용 효과가 다르게 나왔다.

에너지환경산업 더미변수를 사용한 경우에 로짓 모형인 모형 1과 logistic lasso 모형인 모형 3을 비교하면 후자의 예측력이 크게 우수한 것으로 분석된다. 반면에 더미변수를 사용하지 않은 경우에는 로짓 모형인 모형 2가 logistic lasso 모형인 모형 4보다 우수하였다. 따라서 로짓모형으로 추정하고 t-검정에 의해 변수를 선정하는 방법과 logistic lasso 모형에 의해 추정과 변수 선정을 동시에 하는 방법 중 어느 한 방법이 더 우월하다고는 말할 수 없는 것으로 보인다.

www.kci.go.kr

마지막으로 초기 변수집합인 51개 설명변수에 대한 변수 선정과 더미변수를 곱한 변수를 포함한 경우의 변수 선정에 모두 logistic lasso를 적용한 모형 3과 첫 번째 변수 선정에서는 logistic lasso를 적용하고 두 번째 변수 선정에서는 logit과 t-검정을 적용한 모형 5를 비교하면 후자가 예측력이 더 우수하였다. 이것은 logistic lasso와 로짓 모형 중 하나를 택일하지 않고 두 모형을 혼용하거나 가중평균(앙상블)하는 접근을 다양한 상황에서 더 깊게 연구할 필요성을 제시한다.

4. 모형 추정결과

본 절에서는 모든 예측평가지표에서 가장 우수한 것으로 나타난 모형 5를 중심으로 모형 추정 결과를 분석한다. <표 10>은 모형 5의 1단계에 적용한 logistic lasso의 추정결과이다.

〈표 10〉 logistic lasso 추정결과

변수 이름	계수
상수항	-3.358228
유형고정자산 증가율	-0.093800
매출액 증가율	0.022021
순이익 증가율	-0.000946
매출액 영업이익율	-0.139732
매출액 이자비용율	0.173343
매출액 총이익율	-0.594714
금융비용/총부채	14.909464
영업활동이익/총부채	-1.067900
고정장기적합율	0.129366
총부채/총자산	4.774285
유동부채비율	0.001219

www.kci.go.kr

변수 이름	계수
고정부채비율	0.087438
이익잉여금/총자산	-1.035039
유동부채/총자산	-1.205405
현금/총자산	-0.791323
유동비율(current ratio)	0.027989
운전자본/총자산	0.302001
현금흐름/부채	-0.086312
현금흐름/총자산	-0.979556
총자산회전율	-1.761894
log(매출액)	-0.120900
Shumway(2001)의 relative size	-300.846859
당좌자산/총자산	0.008170
λ	0.001004
Pseudo-R^2 1)	0.210986

주 : 1) Efron(1978)의 Pseudo-R^2임.

logistic lasso에서 모형복잡도에 대한 벌칙 정도를 제어하는 모수인 λ는 약 0.001로 추정되었으며 처음 51개 설명변수 후보 중 23개 변수만 선택되었다.[11]

<표 11>은 위 23개 변수들에 더미변수를 곱한 변수들을 추가하고 로짓 모형으로 추정한 다음에 t-검정을 통해서 통계적 유의성이 있는 변수들만을 선택해서 구성된 최종 로짓 모형을 추정한 결과이다. 변수이름에서 'D'는 에너지환경산업 더미변수를 의미한다. 1단계의 logistic lasso에서 예측력을 기준으로

11) <표 10>에서 표준편차를 제시하지 않은 이유는, 여러 학자들(Fan and Li, 2001; Zou, 2006; Osborne et al., 2000)이 제시한 lasso추정량의 공분산행렬에 대한 대표본 추정량에 문제가 있고(Potscher and Leeb, 2009) 부츠트랩에 의한 추정량도 불안정하거나 일치성을 갖지 않는 것으로 알려져 있기 때문임(Kyung et al., 2010).

www.kci.go.kr

선택된 23개 변수들 중에서 15개 변수들이 2단계에서 통계적 유의성을 기준으로 한 t-검정에서 선택되었고 더미변수를 곱한 23개 변수들 중에서는 8개 변수들이 선택되었다. 계수의 절대값이 가장 큰 변수는 Schumway(2001)의 기업상대규모로서 기업시가총액을 주식시장시가총액으로 나눈 값이다. 이 변수는 더미변수를 곱한 변수도 역시 절대값이 크며 모두 음(-)의 부호를 갖는다. 로짓 모형에서 계수값은 종속변수가 1일 확률과 비례하며 본 연구에서 종속변수가 1의 값을 갖는 것은 부도가 나는 경우이므로 기업상대규모가 클수록 부도확률이 감소한다고 해석할 수 있다. 더미변수의 계수가 음(-)이므로 에너지환경산업에서는 기업상대규모가 클수록 다른 산업들에서보다 부도확률이 더 크게 감소하는 것으로 분석된다. 다음으로 계수 절대값이 큰 변수는 금융비용/총부채 변수이며 부호는 양(+)이다. 따라서 총부채 중 금융비용이 차지하는 비중이 클수록 부도확률은 증가하는 것으로 분석된다. 이 변수의 경우는 더미변수를 곱한 변수가 t-검정에서 제거되었으므로 에너지환경산업과 다른 산업들 간에 영향력 차이가 없다. 이외에 계수의 부호가 음(-)인 변수들은 유동부채/총자산, 총자산회전율, 영업활동이익/총부채 등이며 이들 변수들의 값이 클수록 부도확률은 감소한다. 한편 총부채/총자산과 당좌자산/총자산 변수는 양(+)의 계수를 가지며 따라서 변수값이 클수록 부도확률이 증가한다. 더미변수를 살펴보면, 음(-)의 계수를 갖는 경우는 당좌자산/총자산, 영업활동이익/총부채, 매출액 총이익율 등이다. 이중에서 영업활동이익/총부채와 매출액 총이익율은 더미변수가 없는 변수들의 계수도 음(-)이므로 부도확률에 대한 음(-)의 영향력이 에너지환경산업에서 더 큰 것으로 나타났고, 당좌자산/총자산은 더미변수가 없는 변수는 양(+)의 부호를 가지므로 에너지환경산업에서는 부도확률에 대한 양(+)의 영향력이 다른 산업들에 비해 줄어드는 것으로 나타났다. 더미변수들 중 양(+)의 계수를 갖는 경우는 운전자본/총자산과 총부채/총자산 등이다. 먼저 운전자본/총자산 변수는 더미변수가 없는 변수는 제외되었기 때문에 다른 산업들에서는 부도확률에 영향을 미치지 않는 반면에 에너지환경산업에서는 부도확률에 양(+)의 영향력을 갖는 것으로 분석된다. 총부채/총자산 변수는 더미변수가 없는 변수는 양(+)의 부호를 가지므로

www.kci.go.kr

에너지환경산업에서 부도확률에 대한 양(+)의 영향력이 더 증가하는 것으로 나타났다.

〈표 11〉 더미변수 추가 후 최종 로짓 모형의 추정결과

변수 이름	계수	표준오차	Pr(>|z|)
상수항	-4.599837	0.324617	0.000000 (***)[1)]
유형고정자산 증가율	-0.216421	0.087970	0.013887 (**)
매출액 영업이익율	-0.218825	0.044784	0.000001 (***)
매출액 총이익율	-0.594672	0.087856	0.000000 (***)
금융비용/총부채	15.191784	1.791591	0.000000 (***)
영업활동이익/총부채	-1.457413	0.253556	0.000000 (***)
고정장기적합율	0.513474	0.062922	0.000000 (***)
총부채/총자산	5.341186	0.118992	0.000000 (***)
이익잉여금/총자산	-1.089256	0.034504	0.000000 (***)
유동부채/총자산	-2.493134	0.180401	0.000000 (***)
현금/총자산	-1.779639	0.307690	0.000000 (***)
유동비율	0.027032	0.008293	0.001116 (***)
총자산회전율	-2.148089	0.205209	0.000000 (***)
log(매출액)	-0.069208	0.020555	0.000760 (***)
Shumway(2001)의 relative size	-1073.406427	142.711653	0.000000 (***)
당좌자산/총자산	1.334958	0.168372	0.000000 (***)
Dx(매출액 증가율)	0.196238	0.079622	0.013716 (**)
Dx(매출액 영업이익율)	0.327943	0.129088	0.011071 (**)
Dx(매출액 총이익율)	-1.058985	0.313975	0.000744 (***)
Dx(영업활동이익/총부채)	-2.428349	0.688870	0.000423 (***)
Dx(총부채/총자산)	1.980427	0.227614	0.000000 (***)
Dx(운전자본/총자산)	3.500486	0.356427	0.000000 (***)
Dx(Shumway(2001)의 relative size)	-1295.655314	467.420379	0.005573 (***)
Dx(당좌자산/총자산)	-3.266574	0.432694	0.000000 (***)
R^2		0.222880	

주 : 1) '***'와 '**'는 각각 1%, 5% 수준에서 유의함을 의미함

www.kci.go.kr

Ⅴ. 요약 및 결론

에너지환경산업에서 기업부도 예측은 갑작스러운 공급 지장의 가능성을 대비하고 경제 전반의 충격과 국민 후생 저하를 완화시킬 수 있다는 점에서 중요하다.

본 연구에서는 우리나라 코스피와 코스닥 시장에 상장된 비금융기업들의 분기별 재무재표 자료를 이용해서 에너지환경산업에 대한 기업부도 예측모형을 분석하였다. 분석에서는 기업부도예측 연구에서 많이 사용되는 로짓 모형에 대해 모형축소방법인 lasso를 적용한 여부와 에너지환경산업의 더미변수의 적용 여부에 따라 모두 5개의 예측모형을 고려하였다.

분석 결과, 초기 설명변수로 구성된 모형의 추정과 설명변수 선정에서는 logistic lasso를 사용하고 에너지환경산업 더미변수를 반영한 다음에는 로짓 모형에 의한 추정과 t-검정에 의한 설명변수 선정 과정을 사용한 혼합모형이 모든 예측력 평가지표에서 가장 우수한 것으로 나타났다. 특히 이러한 혼합모형은 관측자료가 불균형인 경우에 모형예측력을 올바르게 평가하는 것으로 알려진 MCC 평가지표에서 다른 모형들과 큰 차이를 나타냈다. 한편 로짓 모형에 대해 모형축소방법인 lasso를 적용하는 것과 에너지환경산업의 더미변수의 적용은 모형에 따라 예측력 성과가 혼재되어 나타났다. 또한 로짓 모형과 logistic lasso 중 택일하는 것보다는 혼합하거나 앙상블을 취하는 접근에 대한 더 깊은 연구의 필요성이 제시되었다.

본 논문에서 제시한 예측모형 정교화 절차와 전체 산업의 표본자료를 이용하되 특정 산업에 대한 더미변수를 이용하는 접근이 에너지환경산업 이외에 다른 산업들에서도 과연 예측력을 향상시키는지 여부는 흥미로운 향후 연구 주제로 보인다.

다른 산업들에 비해 에너지환경산업에 대한 부도예측 연구는 해외에서도

www.kci.go.kr

많지 않고 국내에서는 없는 것으로 보인다. 따라서 일반적인 부도예측 연구에서 다루어지는 많은 이슈들이 에너지환경산업에 대해 추후 연구를 통해 추가적으로 분석될 필요가 있다. 여기에는 분류 집단 간 자료의 불균형 문제, 결측치의 보완, 다양한 예측모형의 고려 등이 포함된다. 특히 최근 빠르게 발전하고 있는 기계학습(machine learning)의 방법들을 에너지환경산업에 적용하는 연구 역시 흥미있는 향후 연구주제로 보인다.

접수일(2020년 2월 3일), 수정일(2020년 3월 17일), 게재확정일(2020년 3월 27일)

www.kci.go.kr

◎ 참 고 문 헌 ◎

강지은·이중호·박중구, 2017.「한국 신재생에너지발전과 화력발전의 경제적 파급효과 비교분석」에너지공학 제26권 제3호: pp51-63.

권혁건·이동규·신민수, 2017.「RNN(Recurrent Neural Network)을 이용한 기업부도예측모형에서 회계정보의 동적 변화 연구」 지능정보연구 제23권 제3호: pp139-153.

남재우·이회경·김동석, 2000.「기업 도산 예측을 위한 생존분석 기법의 응용」금융학회지 제5권 제3호: pp29-61.

박소연·임슬예·유승훈, 2015.「산업연관분석을 활용한 하수처리 부문의 경제적 파급효과 분석」 상하수도학회지 제29권 제2호: pp171-182.

박중구, 2012.「석유정보 - 한국석유산업의 산업연관효과 분석」Korea Petroleum Association Journal, 284: pp49-56.

이건창·김명종·김혁, 1994. 「기업도산예측을 위한 귀납적 학습지원 인공신경망 접근방법: MDA, 귀납적 학습방법, 인공신경망 모형과의 성과비교」 경영학연구 제23권 제3호: pp109-144.

이계원, 1993.「회계정보에 의한 기업부실예측과 시장반응」회계학연구 제16호: pp49-77.

이인로·김동철, 2015.「회계정보와 시장정보를 이용한 부도예측모형의 평가 연구」재무연구 제28권 제4호: pp625-665.

장휘용, 1998.「비금융 상장기업의 부실예측모형」재무관리연구 제15권 제16호: pp299-327.

전성빈·김영일, 2001.「도산예측모형의 예측력 검증」회계저널 제10권 제1호: pp151-182.

차성재·강정석, 2018.「딥러닝 시계열 알고리즘 적용한 기업부도예측모형 유용성 검증」지능정보연구 제24권 제4호: pp1-32.

최용석·조창익, 2019.「산업연관관계에 의한 원자력발전의 경제적 효과」사회과학연구 제45권 제3호: pp125-149.

www.kci.go.kr

Altman, E. 1968. “Financial Ratios, Discriminant Analysis and Prediction of Corporate Bankruptcy.” *Journal of Finance* 23 : pp58-610.

_____. 1971. “Railroad Bankruptcy Propensity.” *Journal of Finance* 26 : pp333-346.

_____. 1977. “Predicting Performance in the Savings and Loan Association Industry.” *Journal of Monetary Economics* 10 : pp443-466.

Amendola, A., Restaino, M., and Sensini, L. 2015. “An Analysis of the Determinants of Financial Distress in Italy: A Competing Risks Approach.” *International Review of Economics & Finance* 37 : pp33-41.

Barboza, F., Kimura, H. and Altman, E. “Machine Learning Models and Bankruptcy Prediction” *Expert Systems with Applications* 83 : pp405-417.

Beaver, H. 1966. “Financial Ratios As Predictors of Failure.” *Journal of Accounting Research* 4 : pp71-111

Bickel, P. and Li, B. 2006. “Regularization in Statistics.” *Sociedad de Estad´ıstica e Investigacio´n Operativa Test* 15(2) : pp27-344

Campbell, J., Hilscher, J., and Szilagyi, J. 2008. “In Search of Distress Risk.” *Journal of Finance* 63 : pp2899-2939.

Chava, S. and Jarrow, R. 2004. “Bankruptcy Prediction with Industry Effects.” *Review of Finance* 8 : pp537-569.

Chicco, D. 2017. “Ten Quick Tips for Machine Learning in Computational Biology.” *BioData Mining* 10(35).

Eldahrany, K. 1986. “The Effect of SFAS No. 69 Signals on the Discriminant and Predictive Ability of Financial Reporting for Business Failure in the Oil and Gas Industry.” *Journal of Petroleum Accounting* 5 : pp77-88.

Efron, B. 1978. “Regression and ANOVA with Zero-One Data: Measures of Residual Variation.” *Journal of the American Statistical Association* 73: 113-212.

Efron, B. and Tibshirani, R. 1993. An Introduction to the Bootstrap. London : Chapman and Hall.

Fan, J. and Li, R. 2001. “Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties” *Journal of the American Statistical Association* 96 : pp1348-1360.

www.kci.go.kr

Foster, B. and Zurada, J. 2013. “Loan Defaults and Hazard Models for Bankruptcy Prediction” *Managerial Auditing Journal* 28(6) : pp.516-541.

Hastie, T., Tibshirani, R. and Friedman, J. 2001. The Elements of Statistical Learning. New York : Springer.

Hill, R., Griffiths, W. and Lim, G. 2011. Principles of Econometrics, 4th ed. Wiley.

Kyung, M., Gill, J., Ghosh, M. and Casella, G. 2010. “Penalized Regression, Standard Errors, and Bayesian Lassos” *Bayesian Analysis* 5(2) : pp369-412.

Lu, C., Wei, C., and Chang, Y. 2015. “The Effects and Applicability of Financial Media Reports on Corporate Default Ratings.” *International Review of Economics & Finance* 36 : pp69-87.

Odom, M. D. and Sharda, R. 1990. “A Neural Network Model for Bankruptcy Prediction.” *IJCNN International Joint Conference on Neural Networks.*

Ohlson, S. 1980. “Financial Ratios and the Probabilistic Prediction of Bankruptcy.” *Journal of Accounting Research* 18(1) : pp109-131.

Osborne, M. R., Presnell, B. and Turlach, B. A. 2000. “On the Lasso and Its Dual.” *Journal of Computational and Graphical Statistics* 9(2) : pp319-337.

Pereira, M., Bastoa, M. and da Silva, A. 2016. “The Logistic Lasso and Ridge Regression in Predicting Corporate Failure.” *Procedia Economics and Finance* 39 : pp634-641.

Platt, A., Platt, M. and Pedersen, J. 1994. “Bankruptcy Discrimination with Real Variables.” *Journal of Business Finance & Accounting* 21(4) : pp491-510

Platt, A. and Platt, M. 1990. “Developing A Stable Class of Predictive Variables: The Case of Bankruptcy Prediction.” *Journal of Business Finance & Accounting* 17(1) : pp31－51.

Potscher, B. M. and Leeb, H. 2009. “On The Distribution of Penalized Maximum Likelihood Estimators: The LASSO, SCAD, and Thresholding” *Journal of Multivariate Analysis* 100(9) : pp2065-2082.

Saito, T. and Rehmsmeier, M. 2015. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets” *PLoS ONE* 10(3).

www.kci.go.kr

Shumway, T. 2001. “Forecasting Bankruptcy More Accurately: A Simple Hazard Model.” *The Journal of Business* 74(1) : pp101-124.

Tian, S., Yu, Y., and Guo, H. 2015. “Variable Selection and Corporate Bankruptcy Forecasts.” *Journal of Banking and Finance* 52 : pp89-100.

Tian, S. and Yu, Y. 2017. “Financial Ratios and Bankruptcy Predictions: An International Evidence.” *International Review of Economics and Finance* 51 : pp510-526

Tibshirani, R. 1996. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society* 58(1) : pp267-288.

Tutz, G., Possnecker, W. and Uhlmann, L. 2012. “Variable Selection in General Multinomial Logit Models.” Technical Report 126, Department of Statistics, University of Munich.

Wu, Y., Gaunt, C., and Gray, S. 2010. “A Comparison of Alternative Bankruptcy Prediction Models” *Journal of Contemporary Accounting & Economics* 6(1) : pp.34-45.

Zavgren, C. 1985. “Assessing the Vulnerability to Failure of American Industrial Firms: A Logistic Analysis.” *Journal of Business Finance & Accounting* 12(1) : pp19-45.

Zmijewski, M. 1984. “Methodological Issues Related to the Estimation of Financial Distress Prediction Models.” *Journal of Accounting Research* 22 : pp59-82.

Zou, H. 2006. “The Adaptive Lasso and Its Oracle Properties.” *Journal of the American Statistical Association* 101 : pp1418-1429.

www.kci.go.kr

ABSTRACT

Analyzing Bankruptcy Prediction in Energy and Environment Industries Using Logistic Lasso

Ki-Ho Jeong* and Hee-Jun Lim**

Since energy and environment industries have much publicity, predicting the industries' corporate bankruptcy in advance is important in that it can predict and prepare for the possibility of sudden supply shocks, thereby alleviating the overall economic shock and deteriorating public welfare. This study analyzes prediction of corporate bankruptcy in energy and environment industries. The logit model is used as basic estimation model, in which 51 financial variables frequently used in the previous studies are considered as initial explanatory variables. And we compare the predictive power between two models, one with and another without the variables multiplied by the industries' dummy variable as additional variables. On the other hand, Lasso(Tibshirani, 1996), a model shrinkage method, is applied to alleviate the problem of over-fitting and multicollinearity due to many explanatory variables used in the model. The results showed that the logistic lasso model including the dummy variables of the energy environment industry was the best in all predictive measures.

Key Words : Energy and environment industries, Bankruptcy prediction, Logistic lasso

* Professor, Kyungpook National University(main and corresponding author). khjeong@knu.ac.kr
** Graduate Student, Kyungpook National University(co-author). hjlim8623@naver.com

www.kci.go.kr