The Korean Journal of Applied Statistics (2018) DOI: http://dx.doi.org/10.5351/KJAS.2018.31.3.367
31(3), 367–382

Comparative study of prediction models for corporate bond rating

Hyeongkwon Park[a] · Junyoung Kang[a] · Sungwook Heo[a] · Donghyeon Yu[a,1]

[a]Department of Statistics, Inha University

(Received March 29, 2018; Revised April 30, 2018; Accepted May 2, 2018)

Abstract

Prediction models for a corporate bond rating in existing studies have been developed using various models such as linear regression, ordered logit, and random forest. Financial characteristics help build prediction models that are expected to be contained in the assigning model of the bond rating agencies. However, the ranges of bond ratings in existing studies vary from 5 to 20 and the prediction models were developed with samples in which the target companies and the observation periods are different. Thus, a simple comparison of the prediction accuracies in each study cannot determine the best prediction model. In order to conduct a fair comparison, this study has collected corporate bond ratings and financial characteristics from 2013 to 2017 and applied prediction models to them. In addition, we applied the elastic-net penalty for the linear regression, the ordered logit, and the ordered probit. Our comparison shows that data-driven variable selection using the elastic-net improves prediction accuracy in each corresponding model, and that the random forest is the most appropriate model in terms of prediction accuracy, which obtains 69.6% accuracy of the exact rating prediction on average from the 5-fold cross validation.

Keywords: Corporate bond rating, linear regression, ordered logit, random forest, elastic-net

1. 서론

회사채(corporate bond)는 기업이 자금을 조달하는 방법 중 하나로 액면 금액, 만기, 금리 등이 명시되어 있는 회사가 발행하는 채권이다. 이러한 회사채는 주로 기업의 자금 조달의 목적으로 발행되지만 시장의 거래를 통하여 기업의 채무 상환 능력 및 신용 위험(credit risk) 등이 회사채 수익률로 반영되어 채무 기업의 현행 시장 이자율(market interest rate) 측정에도 활용된다 (Kim과 Choi, 2006). 회사채의 발행은 기업의 입장에서 자금 조달의 유용한 수단이며 기업 회계 기준에 의한 채권, 채무의 현재 가치 측정에 활용되는 이점을 지니나 제약 없는 회사채의 발행은 과도한 차입과 과잉 투자를 이끌 수 있으며 대규모의 기업 부도와 채권자의 손실 등을 야기하여 국가 경제에 심각한 영향을 줄 수 있다. 이에 따라 회사채 발행 적격에 대한 평가 기준이 요구 되었으며 1986년 3월 일반 회사채 및 전환사채 발행 적격 기준에 기업의 신용을 평가하는 신용평가제도가 부분적으로 도입 되었다 (Ko와 Kim, 2002).

국내의 회사채 신용 등급(corporate bond rating)은 금융위원회에 허가를 받은 한국기업평가, NICE평가정보, 한국신용평가의 3개 신용 평가 기관에 의해 공시되고 있으며 각 신용 평가 기관의 내부 평가 기

This work was supported by Inha Universitiy Research Grant (INHA-55456).
[1]Corresponding author: Department of Statistics, Inha University, 100 Inharo, Nam-gu, Incheon 22212, Korea. E-mail: dyu@inha.ac.kr

준에 따라 AAA, AA+, . . . , B+, B, B−, CCC, CC, CC, D의 20개 등급으로 표현된다. 여기서 +, 0, − 표현은 AA∼B의 등급에 적용되며 AAA등급부터 BBB등급까지를 투자등급으로 BB등급부터 D등급까지를 투기등급으로 분류하고 있다 (Ko와 Kim, 2002). 이러한 회사채의 등급 결정은 각 산업별 평가자에 의해 작성된 현재 및 미래 시장 상황에 대한 보고서 뿐만 아니라 다양한 전문가가 참여한 여러 위원회의 협의를 통하여 이루어진다 (Kim 등, 2006).

신용 평가 기관에 의해 공시된 회사채 신용 등급 정보를 활용하는 것에 그치지 않고 회사채 신용 등급을 예측하고자 하는 이유로는 Kim과 Choi (2006)에서 서술한 바와 같이 회사채를 발행하지 않는 회사에 대한 신용 등급 예측 등 통하여 채무 기업에 대한 시장이자율을 측정하는데 활용할 수 있으며 회사채 발행 회사의 경우에도 자금 조달을 원활하게 하기 위해 회사채 신용 등급 관리를 위한 중요 지표를 식별 관리할 수 있게 하는 이점, 채권자가 채무 회사의 신용 등급 조정을 사전에 예측하여 대응 할 수 있는 점 등을 들 수 있다. 회사채 신용 등급의 예측 모형은 전통적인 선형 회귀(linear regression) 기반의 예측 모형 (Horrigan, 1966; West, 1970; Kim과 Kim, 2002; Kim과 Choi, 2006)을 시작으로 판별분석(discriminant analysis) 기반의 예측 모형 (Pinches와 Mingo, 1973; Altman과 Katz, 1976), 순위 로짓(ordered logit) 기반의 예측 모형 (Kim 등, 2006; Jeong, 2011), 순위 프로빗(ordered probit) 기반의 예측 모형 (Kaplan과 Urwitz, 1979; Ederington, 1986), 서포트 벡터 기계(support vector machine) 기반의 예측 모형 (Huang 등, 2004) 등 여러 예측 모형이 제안되었으며 최근에는 앙상블(ensemble) 기반의 예측 모형 (Kim, 2012; Kim과 Ahn, 2016) 중심으로 진행되고 있다.

기존의 회사채 신용 등급의 예측 모형들은 예측 모형의 우수성을 입증하기 위하여 각 연구에 따라 수집된 자료를 기반으로 예측 성능을 비교 평가하여 보고하였다. 하지만 연구에 따라 수집된 자료의 기간과 예측 모형에 실제 적용한 회사채 신용 등급 구간이 상이하여 각 연구에 따라 보고된 예측 정확도(prediction accuracy)를 기반으로 단순히 비교하기에는 어려움이 있다. 특히, 회사채 신용 등급 구간은 수집된 자료 내의 회사채 신용 등급의 불균형 문제 (Kim, 2012) 또는 각 등급에 따른 표본 수의 부족 (Kim 등, 2006; Kim과 Choi, 2006) 등을 이유로 회사채 신용 등급을 적게는 5등급 (Kim, 2012)에서 최대 20등급 (Seo, 2015)으로 고려되었다.

따라서 본 연구에서는 기존 예측 모형들의 공정한 비교를 위하여 동일한 기간에 수집된 자료를 기반으로 AAA, CCC 이하인 5개 등급을 제외한 AA+∼B−의 15개 등급을 대상으로 예측 모형을 적용하여 성능 비교를 진행하였다. 여기서 AAA, CCC 이하인 등급은 표본 수의 이유로 많은 연구에서 제외 되었으며 (Kim, 2012), AAA등급의 경우 공기업을 포함 소수의 기업이 해당 등급으로 평가 되며 CCC 이하의 등급은 투기 등급으로 부실 채권으로 분류되고 발행 대상 기업의 수 또한 적으므로 비교 분석에서 제외하였다 (Kim과 Choi, 2006). 예측 성능의 평가를 위하여 수집된 자료를 각 등급별 임의 치환(random permutation)을 적용한 뒤 5개의 조각으로 균등하게 나누어 5-fold 교차검증(cross validation)을 수행하였다. 예측 정확도는 15개의 등급에 대한 정확한 예측 및 1등급 오차 허용 예측 정확도로 구분하여 성능 비교를 진행하였다. 추가적으로 본 연구에서는 기존의 예측 모형에 포함된 변수들 중 상관성이 높은 변수들이 포함된 것을 확인하여 이에 대한 보완책으로 회귀 모형, 순위 로짓 모형, 순위 프로빗 모형에 대하여 Elastic-net 벌점을 고려하여 예측 모형을 수립하고 예측 성능이 향상됨을 확인하였다.

본 논문의 2절에서는 국내의 회사채 신용 등급 예측과 관련한 기존 연구 결과를 개략적으로 살펴보고 예측 모형의 공정한 비교를 위하여 수집된 국내 회사채 신용 등급 자료 및 재무적/비재무적 특성 변수들에 대한 현황을 3절에 정리하였다. 4절에서는 본 연구에서 비교하고자 하는 기존 예측 모형을 요약하고 통합된 변수를 기반으로 Elastic-net 벌점 모형의 적용에 대하여 서술하였다. 5절에서는 기존의 예측 모형들을 수집된 자료에 적합하여 예측 성능을 비교하였으며 Elastic-net 기반의 변수 선택을 통한 예측 모형의 성능 개선 정도를 살펴보았다. 끝으로 6절에서 본 연구의 결과를 요약하고 결론을 정리하였다.

2. 국내 회사채 신용 등급 예측의 기존 연구

국내 기업에 대한 신용 평가와 관련한 연구는 크게 기업의 부도 예측과 회사채에 대한 신용 등급 예측으로 구분할 수 있으며 기업의 부도 예측을 기반으로 한 신용 평가 모형에 대한 연구가 주로 이루어지고 있다. 본 논문은 국내 회사채의 신용 등급 예측 모형의 비교를 목적으로 하여 기업의 부도와 관련한 예측 및 신용 평가 모형에 관한 기존 연구의 현황은 생략하였다. 또한 1절에서 언급한 바와 같이 회사채 신용 등급의 예측에 관한 연구는 다양한 통계 모형을 통하여 이루어져 왔으며 국내·외에 많은 연구 결과가 보고 되어 있다. 본 절에서는 국내 회사채 신용 등급 예측 연구 중 대표적인 연구 결과들을 살펴보고 국내외 여러 연구에 대한 현황은 Kim과 Choi (2006) 및 Kim과 Ahn (2016)을 참조하는 것으로 대신한다.

국내의 대표적 연구들을 살펴 보면, Kim과 Choi (2006)는 Kaplan과 Urwitz (1979)의 연구 결과를 기반으로 국내의 회사채 신용 등급 예측 모형에 활용하여 선형 회귀 모형과 순위 프로빗 모형에 기반한 예측 모형을 제안하였다. Kim과 Choi (2006)에서는 B−등급부터 AA+까지 15개의 신용 등급을 1부터 15까지 수치로 재정의하여 적합한 모형과 등급의 세부 등급$(-, 0, +)$을 구분하지 않고 B, BB, BBB, A, AA의 5등급을 순차적으로 1부터 5의 수치로 재정의하여 적합한 모형을 고려하였다. 예측 모형에 사용된 자료는 2000년부터 2004년까지의 기간에 회사채 등급이 존재하는 510개 거래소 상장 및 KOSDAQ 등록 기업을 대상으로 하였으며 이 중 340개 기업을 훈련 자료(training set)로 이용하여 예측 모형을 추정하고 나머지 170개의 기업을 검증 자료(testing set)로 사용하여 예측 모형의 정확도 평가를 진행하였다. 회사채 신용 등급 예측을 위해 선형 회귀 모형 및 순위 프로빗 모형의 독립 변수로 자산총계, 매출액, 부채비율, 자본이익률, 베타계수, 누적이익률, 누적시장조정수익률, 주가순자산비율을 고려하였으며 선형 회귀 모형 및 순위 프로빗 모형의 15등급의 예측 정확도는 각각 29.4%와 30.0%로 낮은 정확도를 지니나 한 등급 차이 내의 예측 정확도는 상대적으로 크게 증가하여 각각 74.1% 및 76.5%로 보고 되었다.

Kim 등 (2006)은 순위 로짓 모형을 기반으로 예측 모형을 제안하였으며 Kim과 Choi (2006)와 다르게 실질적 부도 상태에 돌입하는 단계로 판단되는 CCC등급 이하의 기업도 포함하여 예측 모형을 적합하였다. 또한 회사채 신용 등급을 AA, AA+를 하나의 등급으로 B−, B, B+를 하나의 등급, D∼CCC를 하나의 등급으로 재정의 하여 높은 등급부터 순차적으로 1부터 14의 수로 정의 하였다. 예측 모형에 사용된 자료는 1999년부터 2003년까지의 기간에 회사채 등급 공시가 존재하는 991개 기업을 대상으로 하였으며 이 중 600개의 기업을 훈련 자료로 이용하여 예측 모형을 추정하고 391개의 기업을 검증 자료로 사용하여 예측 모형의 정확도를 평가하였다. 모형에 포함되는 변수의 선택을 위하여 기업의 신용도를 우량(BBB− 이상)과 불량(D, 신용불량 정보)로 구분 한 뒤 개별 후보 변수들에 대한 평균 검정을 진행하고 유의한 차이가 나타난 변수들을 대상으로 요인분석, 단변량 로짓을 통한 요인별 변수의 선정 등을 토대로 최종 모형을 식별하였다. 최종 모형에는 순금융비용부담율, 단기차입금/총차입금, EBIDA/매출액(이자, 세금, 감가상각비, 무형자산상각비 차감 전 이익(earnings before interest, tax, depreciation, and amortization; EBITDA)), 자기자본비율, 총자산(자연대수)의 5개 변수가 선택되었다. 순위 로짓 모형의 검증 자료에 대한 예측 정확도는 36.33%이며 한 등급 차이 내의 예측 정확도는 71.76%로 보고되었다.

Kim과 Ahn (2016)은 단기 신용 등급에 대한 예측 모형으로 Breiman (1996)의 랜덤 포레스트(random forest) 모형을 제안하였으며 모형의 비교, 평가를 위하여 다분류 서포트 벡터 기계 모형, 다중 판별 분석 등을 고려하였다. Kim과 Ahn (2016)은 A1, A2, A3 및 B와 C의 4개의 등급으로 재정의하여 단기 회사채 신용 등급에 대한 예측 모형을 제안하였다. 예측 모형에 사용된 자료는 2002년에 공시된 1,295개 거래소 상장 및 KOSDAQ 등록 기업을 대상으로 하였으며 이 중 1,036개 기업을 훈련 자료로

Table 3.1. Summary of observed corporate bond ratings from 2013 to 2017

Rating	Year					Sum	Ratio (%)
	2013	2014	2015	2016	2017	(rating)	(rating)
AA+	11	10	9	12	12	54	7.88
AA	15	19	17	18	19	88	12.85
AA−	21	18	22	20	20	101	14.74
A+	22	19	16	14	16	87	12.70
A	18	15	14	19	12	78	11.39
A−	14	17	24	25	29	109	15.91
BBB+	10	4	4	6	6	30	4.38
BBB	8	8	8	8	10	42	6.13
BBB−	2	5	4	1	1	13	1.90
BB+	10	6	4	2	4	26	3.80
BB	3	2	3	3	1	12	1.75
BB−	3	1	3	5	5	17	2.48
B+	4	4	0	2	1	11	1.61
B	1	1	4	1	2	9	1.31
B−	1	2	1	2	2	8	1.17
Sum(year)	143	131	133	138	140	685	100.00

이용하여 예측 모형을 추정하고 259개 기업을 검증 자료로 사용하여 예측 모형의 정확도를 평가하였다. 서포트 벡터 기계 모형과 랜덤 포레스트 모형의 설명변수로 자산총계, 단기차입금/총차입금, 자기자본, 부채총계, 업력, 주당순이익, 유보액대총자산비율, 금융비용부담률, 금융비용대총비용비율, 고정자산구성비율, 재고자산대유동자산비율, 현금흐름대총자본비율, 1인당매출액, (영업활동으로인한현금흐름 − 현금배당)/(고정자산 + 운전자본), 매출액의 14개 변수를 고려하였으며 4등급 기준 정확한 등급 예측의 정확도는 각각 67.60%, 72.79%으로 보고 되었다.

3. 분석 대상 자료

3.1. 분석 대상 자료의 수집 및 현황

본 비교 연구를 위하여 2013년부터 2017년의 각 연도 말을 기준으로 한국증권거래소 상장기업과 KOSDAQ 등록기업 중 12월 결산 법인으로 비금융 업종에 속한 기업들을 대상 표본으로 수집하였다. 금융업종의 경우 비금융 업종과의 자금 조달 구조, 신용 평가시 고려하는 재무 변수의 종류 및 의미의 차이 등을 이유로 기존 연구들 (Ko와 Kim, 2002; Kim과 Choi, 2006)에서도 제외되어 본 연구에서도 제외하였다. 또한 서론에서 언급한 바와 같이 회사채 신용 등급이 AAA등급의 경우 공기업을 포함 소수의 기업이 해당 등급으로 평가 되며 CCC 이하의 등급은 투기 등급으로 부실 채권으로 분류되고 발행 대상 기업의 수 또한 적으므로 해당 등급의 기업-년 표본도 분석 대상 자료에서 제외하였다. 위의 절차를 통하여 수집된 최종 대상 표본은 685개의 기업-년으로 Table 3.1에 정리하였으며 연도별로 143개(2013), 131개(2014), 133개(2015), 138개(2016), 140(2017)개로 구성된다. Table 3.1에서 확인할 수 있듯이 신용 등급에 따라 관측된 표본의 기업-년 자료의 비율이 변화한다. 따라서 예측 정확도의 비교를 위해 본 연구에서는 685개의 기업-년 자료에서 각 신용 등급에 따라 임의 치환을 적용한 뒤 균일하게 5개의 조각으로 나누어 5-fold 교차 검증을 수행하였다.

회사채 신용 등급 예측 모형의 비교를 위하여 요구되는 회사채 신용 등급 자료, 대상 기업의 재무제표 자료, 기타 기업 관련 자료는 아래의 절차를 통하여 수집되었다.

(1) 회사채 신용 등급: 2013∼2017년에 대한 회사채 신용 등급 자료는 신용 평가 기관에서 발표된 신용 등급을 기반으로 KIS 채권평가(www.bond.co.kr)에서 제공하는 채권분석정보에 게시된 신용등급 종합표에서 수집되었다. 사업 보고서 공시 후 회계 정보가 회사채 신용 등급에 반영 되기까지의 기간을 고려한 기존 연구 (Kim과 Choi, 2006)와 동일하게 각 회사채 발행 기업의 결산월로부터 6개월이 지난 시점에 공시된 신용 등급을 당해 기업의 신용 등급으로 조사하였다.

(2) 기업의 재무제표 자료: 효율적인 자료의 수집을 위하여 표본 기업에 대한 재무제표 자료는 네이버 금융(finance.naver.com) 내의 기업별 종목 분석 항목에서 기업의 연도별 결산 재무제표 자료를 웹 스크래핑(web scraping)을 활용하여 수집하였다. 종목 분석 항목 내의 기업의 재무제표 정보는 DART 전자공시시스템(dart.fss.or.kr) 내에 공시된 자료를 기반으로 WISEfn(www.wisefn.com)에서 취합 및 정리하여 제공하는 정보이다.

(3) 기타 기업 관련 자료: 기업의 베타 계수(beta coefficient) 및 주가순자산비율(price to book-value ratio; PBR)을 포함한 기업 정보 및 주가 자료는 코스콤(www.koscom.co.kr)에서 제공하는 실시간 주가 정보와 국내시세정보와 WISEfn에서 제공하는 투자정보를 참고한 네이버 금융에서 자료를 수집하였고, 사원 수 자료의 경우는 DART 전자정보공시시스템에서 제공하는 간편 검색 서비스를 활용하여 수집하였다.

3.2. 독립 변수의 선정 및 측정

기존 모형의 비교와 각 연구의 결과를 통합한 분석을 진행하기 위하여 국내 회사채 신용 등급 예측 모형 연구에서 고려된 적이 있는 변수들을 통합하여 독립 변수로 선정하고 각 변수의 값을 산출하였다. 보다 상세히 살펴 보면, Kim과 Choi (2006)는 최종 예측 모형에 자산총계, 매출액, 부채비율, 자본이익률, 기업위험(베타계수), 누적이익률 및 기업의 수익성에 대한 시장의 기대를 대표하는 변수로 주가순자산비율과 시장지수조정수익률의 8개의 변수를 고려하였고 Kim과 Ahn (2016)의 연구에서는 단기차입금/총차입금, 자기자본, 부채총계, 업력, 주당순이익, 유보액대총자산비율, 금융비용부담율, 금융비용대총비용비율, 고정자산구성비율, 재고자산대유동자산비율, 현금흐름대총자본비율, 매출액, 1인당 매출액, (영업활동으로 인한 현금흐름 − 현금배당)/(고정자산 + 운전자본)의 14개의 변수를 고려하였다. 이 외에도 여러 연구 (Kim 등, 2006; Jeong, 2011)에서 적용된 재무 비율들을 참고하여 추가적으로 독립 변수를 선정하였다. 최종적으로 고려한 독립 변수는 총 60개의 변수로 아래와 같이 규모 지표(5개), 비재무 지표(4개), 생산성 지표(2개), 수익성 지표(18개), 안정성 지표(17개), 현금흐름 지표(10개), 활동성 지표(4개)로 구분하여 정리하였다. 보다 자세한 독립 변수의 정의 및 계산식은 Kim과 Choi (2006) 및 Kim과 Ahn (2016)을 참조하길 바란다.

- 규모 지표:
  자산총계(X_1), 매출액(X_2), 자기자본(X_3), 부채총계(X_4), 사원수(X_5)
- 비재무 지표:
  베타계수(X_6), 누적시장조정 수익률(X_7), 업력(X_8), 주가순자산비율(X_9)
- 생산성 지표:
  1인당 매출액(X_{10}), 경영자본회전율(X_{11})
- 수익성 지표:
  자기자본이익률(X_{12}), 누적이익률(X_{13}), EBITDA/매출액(X_{14}), 순금융비용부담율(X_{15}), 영업이익/총자산(X_{16}), 이자이익/총자산(X_{17}), 주당순이익(X_{18}), 금융비용부담률(X_{19}), 금융비용대총비용

비율(X_{20}), 이자보상배율(X_{21}), 매출액경상이익률(X_{22}), 수익성비율(X_{23}), 매출액총이익률(X_{24}), EBIT/매출액(X_{25}), 매출액순이익률(X_{26}), 매출원가율(X_{27}), 판관비율(X_{28}), 경상이익/자기자본(X_{29})

- 안정성 지표:
  부채비율(X_{30}), 자기자본비율(X_{31}), 유보액대총자산비율(X_{32}), 고정자산구성비율(X_{33}), 재고자산대유동자산비율(X_{34}), 단기차입금대총차입금비율(X_{35}), 당좌비율(X_{36}), 장기부채/유동부채(X_{37}), 유동비율(X_{38}), 고정장기적합률(X_{39}), 순차입금비율(X_{40}), 유동부채비율(X_{41}), 비유동부채비율(X_{42}), 단기차입금의존도(X_{43}), 차입금의존도(X_{44}), 순자산/부채총계(X_{45}), 순운전자본구성비율(X_{46})
- 현금흐름 지표:
  현금흐름대총자본비율(X_{47}), (영업활동으로 인한 현금흐름 − 현금배당)/(고정자산 + 운전자본) (X_{48}), 영업활동으로 인한 현금흐름(X_{49}), EBITDA/순금융비용(X_{50}), 영업활동 후의 현금흐름 (X_{51}), OCF대부채비율(X_{52}) (영업활동으로 인한 현금흐름(operating cash flow; OCF)), OCF대총자본비율(X_{53}), NCF대부채비율(X_{54}) (순영업활동 현금흐름(net cash flow from operating activities; NCF)), NCF대매출액비율(X_{55}), 총차입금/EBITDA(X_{56})
- 활동성 지표:
  재고자산회전율(X_{57}), 총자산회전율(X_{58}), 매출채권회전율(X_{59}), 매입채무회전율(X_{60})

각 독립 변수의 기술 통계량인 표본 평균(sample mean), 표본 표준편차(sample standard deviation), 최솟값(minimum) 및 최댓값(maximum)은 Table 3.2에 정리하였다. 자산총계(X_1), 매출액(X_2), 주가순자산비율(X_9)의 경우 기존 연구에서 log 변환을 적용하여 예측 모형을 수립하여 Table 3.2에도 해당 변수에 대하여 log 변환된 값을 보고 하였다. 기존 연구에서 최솟값과 최댓값의 특이값에 대한 처리를 위하여 윈저화(Winsoriation) 방법을 적용하였으므로 본 논문에서도 변수들의 상위, 하위 각각 5%에 해당하는 값을 기준값으로 하여 기준값 이상 또는 이하의 값을 각각 기준값으로 대체하였다 (Kim과 Choi, 2006; Kim과 Ahn, 2016). 또한 Table 3.2에 제시된 바와 같이 X_3, X_4, X_5, X_7, X_8, X_{10}, X_{18}, X_{49}, X_{50}, X_{57}의 10개 변수는 다른 변수와 비교하여 분산이 매우 큰 변수들로 분산 안정화 변환이 요구된다. 하지만 4.1절에서 나타낸 기존 연구의 예측 모형에서 변환을 적용하지 않은 변수를 적용한 경우 비교의 목적으로 원래의 모형에서 적합한 변수를 그대로 사용하였다. 4.2절에서는 추가적인 분석을 목적으로 위의 변수들에 대하여 분산 안정화 변환을 적용하고 회귀 모형 하에서 변수들 사이의 높은 상관성 또는 그룹 성질이 존재하는 경우 예측 성능이 우수하다고 알려진 Elastic-net 벌점화 모형을 적용하여 예측 모형을 수립하고 평가하였다.

4. 회사채 신용 등급 예측 모형

4.1. 기존 연구의 예측 모형

본 절에서는 기존 연구의 예측 모형들을 간략히 소개하고자 한다. 기존 연구에서 적용된 변수들을 통합하여 표현하기 위하여 3절에서 나타낸 $X_1, \ldots, X_{60}$의 변수명 표기를 활용한다. 먼저 Kim과 Choi (2006)에 의해 제안된 모형을 살펴보면 X_1, X_2, X_6, X_7, X_9, X_{13}, X_{29}, X_{30}의 8개의 독립 변수를 고려한 회귀 모형 및 순위 프로빗 모형을 예측 모형으로 고려하였다. 적용된 회귀 모형은 B−를 1로 시작하여 AA+를 15로 할당하여 정의한 종속 변수를 고려한

$$Y_i = \beta_0 + \sum_{j \in \{1,2,6,7,9,13,29,30\}} \beta_j X_{ij} + \epsilon_i. \tag{4.1}$$

Table 3.2. Summary statistics of explanatory variables used in the prediction models

Variable	Mean	Std.	Min	Max	Variable	Mean	Std.	Min	Max
$\log(X_1)$	4.40	0.64	3.21	5.51	X_{31}	0.42	0.14	0.17	0.69
$\log(X_2)$	4.33	0.65	3.15	5.46	X_{32}	0.33	0.17	0.03	0.62
X_3	25598.12	36999.05	637.20	136090.40	X_{33}	0.59	0.15	0.32	0.83
X_4	39241.82	57412.30	982.60	215979.60	X_{34}	0.26	0.14	0.02	0.54
X_5	3259.12	4585.93	72.60	18154.60	X_{35}	0.30	0.20	0.01	0.68
X_6	0.90	0.39	0.24	1.61	X_{36}	0.88	0.38	0.33	1.69
X_7	−0.16	31.92	−56.80	61.11	X_{37}	0.64	0.62	0.05	2.54
X_8	40.65	16.98	13.00	70.80	X_{38}	1.21	0.48	0.49	2.23
$\log(X_9)$	0.08	0.63	−1.18	1.16	X_{39}	0.94	0.24	0.56	1.47
X_{10}	33.26	55.71	4.06	227.75	X_{40}	1.06	0.70	0.17	2.81
X_{11}	1.06	0.44	0.44	2.07	X_{41}	1.07	0.75	0.27	3.12
X_{12}	0.02	0.13	−0.36	0.18	X_{42}	0.58	0.36	0.11	1.49
X_{13}	0.23	0.17	−0.07	0.53	X_{43}	0.10	0.08	0.00	0.28
X_{14}	0.08	0.08	−0.10	0.25	X_{44}	0.32	0.13	0.12	0.57
X_{15}	0.02	0.02	0.00	0.06	X_{45}	0.76	0.54	0.09	2.09
X_{16}	0.03	0.04	−0.05	0.11	X_{46}	0.05	0.15	−0.24	0.30
X_{17}	0.01	0.01	0.00	0.04	X_{47}	0.00	0.03	−0.05	0.06
X_{18}	3188.94	6196.80	−5421.20	20219.20	X_{48}	0.07	0.10	−0.13	0.26
X_{19}	0.03	0.02	0.01	0.09	X_{49}	3713.04	6911.79	−1045.00	26141.60
X_{20}	0.03	0.02	0.01	0.08	X_{50}	6.78	16.72	−32.85	47.47
X_{21}	2.82	3.78	−2.05	13.84	X_{51}	0.06	0.07	−0.08	0.20
X_{22}	0.02	0.07	−0.16	0.14	X_{52}	0.09	0.11	−0.09	0.32
X_{23}	0.03	0.05	−0.08	0.12	X_{53}	0.05	0.06	−0.06	0.15
X_{24}	0.19	0.14	0.02	0.57	X_{54}	0.01	0.05	−0.10	0.12
X_{25}	0.04	0.06	−0.12	0.15	X_{55}	0.00	0.03	−0.06	0.07
X_{26}	0.01	0.06	−0.16	0.11	X_{56}	4.26	5.60	−8.54	16.79
X_{27}	0.81	0.14	0.43	0.98	X_{57}	12.41	15.18	2.19	67.48
X_{28}	0.15	0.13	0.03	0.53	X_{58}	0.93	0.41	0.34	1.91
X_{29}	0.01	0.01	0.00	0.04	X_{59}	6.48	3.43	2.21	14.57
X_{30}	1.71	1.13	0.45	4.93	X_{60}	6.72	2.96	2.49	12.78

Mean and Std. denote a sample mean and a sample standard deviation for each variable, respectively.

모형을 제안하였다. 회사채 신용 등급 예측을 위한 회귀 모형의 적용은 회귀 모형의 기본 가정인 오차항의 정규성에 대한 위배가 큰 단점이며 등급 r_i의 예측에 있어서 $\hat{r_i} = j$ 만약 $j - 0.5 \le \hat{Y}_i \le j + 0.5$와 같이 구간을 정하여 등급을 할당하는 단점이 있다. 하지만 예측 모형에 대한 설명과 해석이 용이하여 기존 예측 모형 연구에서 많이 고려된 모형이다. 이러한 회귀 모형의 단점을 보완하기 위하여 Kim과 Choi (2006)는 신용 등급이 순서를 가진 범주형 자료임을 고려하여 동일한 독립 변수를 기반으로 순위 프로빗 모형을 비교 모형으로 함께 제시하였다. 순위 프로빗 모형에서도 회귀 모형과 동일한 종속 변수를 사용하여

$$\Phi\big(P(Y_i \le j)\big)^{-1} = \mu_j - \sum_{k\in\{1,2,6,7,9,13,29,30\}} \beta_k X_{ik}, \quad \text{for } 1 \le j \le 14. \tag{4.2}$$

모형을 제안하였다. 여기서 $\Phi(x)$는 표준 정규분포의 누적분포함수(cumulative distribution function)를 나타낸다.

순위 로짓 모형을 고려한 Kim 등 (2006)에서는 53개의 재무 관련 변수를 기반으로 t-test, 요인분석 및 단변량 로지스틱 모형을 활용하여 최종 모형에 X_1, X_{14}, X_{15}, X_{31}, X_{35}의 5개의 독립 변수를 고려하였다. Kim 등 (2006)에서는 다른 연구와는 다르게 다중공선성 문제를 피하기 위해 요인 분석을 활용하는 등 여러 절차를 고려하였으며 최종 모형으로

$$\log \frac{P(Y_i \le j)}{1 - P(Y_i \le j)} = \mu_j - \sum_{k \in \{1,14,15,31,35\}} \beta_k X_{ik}, \quad \text{for } 1 \le j \le 14 \tag{4.3}$$

를 고려하였다. 원래 제안된 모형은 AAA를 1, AA+, AA를 2, . . . , B+, B, B−를 13, CCC 이하를 14로 정의하여 예측 모형을 수립하였으나 비교 목적을 위하여 본 연구에서는 Kim과 Choi (2006)과 동일하게 반응 변수를 정의하였다. 순위 로짓 및 순위 프로빗 모형의 적합을 위해 R의 `ordinal` 패키지(github.com/runehaubo/ordinal) 내의 `clm` 함수를 활용하였다. 또한 `clm` 함수의 적용으로 식 (4.2) 및 (4.3)의 표현에서 모수에 대한 표현을 $+\beta$ 대신 $-\beta$의 형태로 사용하였다.

마지막으로 Kim과 Ahn (2016)에서는 회사채 신용등급 예측을 위하여 랜덤 포레스트를 고려하였으며 성능의 비교를 위하여 다중 판별분석, 인공신경망, 서포트 벡터 기계를 적용하였다. 하지만 실제 자료 분석 시 단기 신용 등급(short-term rating)을 고려하여 A1, A2, A3, B와 C를 각각 1∼4의 값으로 재정의하여 예측 모형을 수립하여 기존에 제안된 예측 모형과의 직접적인 성능 비교에는 어려움이 있다. 따라서 본 논문에서는 Kim과 Choi (2006)에서 고려한 15등급 체계를 기준으로 Kim과 Ahn (2016)에서 고려한 커널 서포트 벡터 기계 및 랜덤 포레스트를 적용하고자 한다. Kim과 Ahn (2016)에서는 39개의 재무 변수를 기반으로 일원배치 분산분석(one-way ANOVA)과 다중 판별 분석을 기반으로 $X_2, X_3, X_4, X_8, X_{10}, X_{18}, X_{19}, X_{20}, X_{32}, \ldots, X_{35}, X_{47}, X_{48}$의 14개 변수를 선정하여 예측 모형을 수립하였다. 본 비교 연구에서는 Kim과 Ahn (2016)의 결과에서 우수하게 나타난 one versus one (OVO) 기반 radial basis function (RBF) 커널 서포트 벡터 기계 및 랜덤 포레스트 모형만을 고려하였다. 따라서 OVO기반 RBF 커널 서포트 벡터 기계에서는 $(15 \times 14/2)$개의 두 신용 등급의 조합에 대하여 아래의 이진 분류에 대한 RBF 커널 서포트 벡터 기계 분류기를 적합한 뒤 가장 많은 투표(vote)를 얻은 등급을 이용하여 예측한다.

$$\begin{aligned} &\max_{\alpha_i \ge 0} \ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{N_{kl}} \sum_{j=1}^{N_{kl}} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ &\text{subject to } 0 \le \alpha_i \le C, \ \sum_{i=1}^{N_{kl}} \alpha_i y_i = 0, \end{aligned} \tag{4.4}$$

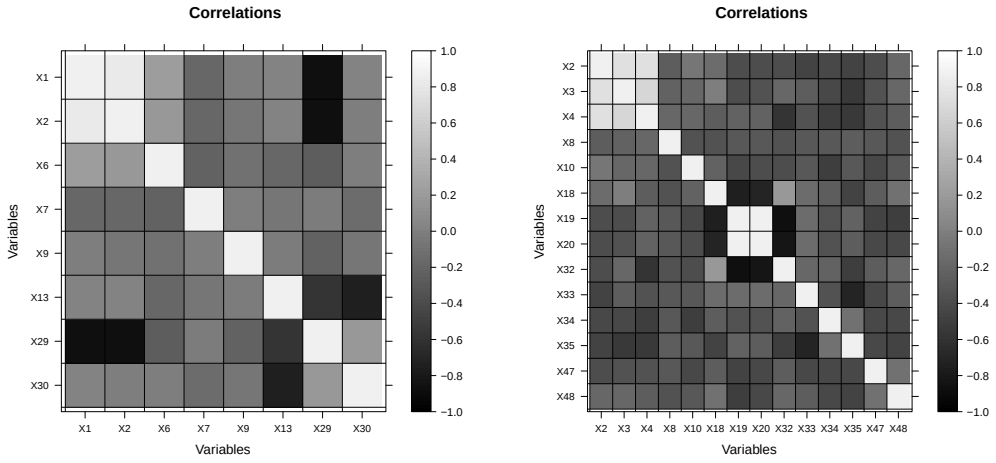
여기서 $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|_2^2)$의 RBF 커널을 나타내며 $\mathbf{x}_i = (X_{i2}, X_{i3}, \ldots, X_{i,47}, X_{i,48})^T \in \mathbb{R}^{14}$, $y_i, y_j \in \{k, l\}$, $1 \le k \ne l \le 15$, $N_{kl} = |\{i | y_i = k\}| + |\{i | y_i = l\}|$. 서포트 벡터 기계는 R의 `e1071` 패키지(cran.r-project.org/web/packages/e1071)를 이용하여 적합하였다.

랜덤 포레스트는 표본에 대한 부트스트랩(bootstrap)과 변수에 대한 임의 선택을 통하여 다수의 의사결정 나무를 생성하고 전체 의사 결정 나무에 의해 예측된 결과를 투표를 통하여 결합하는 기법으로 Breiman (1996)에 의해 제안된 앙상블 기법이다. 본 비교 연구에서는 Kim과 Ahn (2016)에서 제안한 14개의 변수에 대한 랜덤 포레스트 모형의 적용과 4.2절에서 설명하는 Elastic-net 벌점화를 고려한 모형에서 선택된 변수를 이용한 랜덤 포레스트의 적용을 진행하였다. 랜덤 포레스트의 적용을 위하여 의사 결정 나무 생성에 고려하는 변수의 수와 전체 의사 결정 나무의 생성 수에 대한 조율은 교차 검증을 통하여 진행하였다. 랜덤 포레스트는 R의 `randomForest` 패키지(cran.r-project.org/web/packages/randomForest)를 이용하여 적합하였다.

Table 4.1. Summary statistics of explanatory variables after log transformation

Variable	Mean	Standard errors	Min	Max
X_3	9.18	1.48	6.46	11.82
X_4	9.55	1.52	6.89	12.28
X_5	7.17	1.46	4.28	9.81
X_7	3.74	1.08	0.00	4.78
X_8	3.60	0.50	2.56	4.26
X_{10}	2.76	1.07	1.40	5.43
X_{18}	8.47	2.06	0.00	10.15
X_{49}	7.49	2.02	0.00	10.21
X_{50}	3.50	0.90	0.00	4.40
X_{57}	2.09	0.84	0.78	4.21

Mean and Std. denote a sample mean and a sample standard deviation for each variable, respectively.



(a) OLS/Probit (b) SVM/RF

Figure 4.1. Correlation maps of variables used in (a) OLS/Probit models and (b) SVM/RF models. OLS = ordinary least squares; SVM = support vector machine; RF = random forest.

4.2. Elastic-net 벌점화 모형의 적용

먼저 3절의 마지막 부분에서 설명한 바와 같이 X_3, X_4, X_5, X_7, X_8, X_{10}, X_{18}, X_{49}, X_{50}, X_{57}의 10개 변수는 다른 변수와 비교하여 분산이 매우 큰 변수들로 분산 안정화를 위하여 각 변수에 각 변수의 최솟값의 절댓값 + 1의값을 더한 뒤 log 변환을 하여 새로운 독립 변수로 정의하였다. 분산 안정화 변환 결과는 Table 4.1에 정리하였으며 log 변환 결과 표준편차가 0.5∼2.06으로 안정화된 것을 확인할 수 있다. 또한 기존의 예측 모형에서 적용한 독립 변수들을 살펴보면 순위 로짓 모형을 제안한 Kim 등 (2006)의 연구를 제외 할 경우 4.1절에서 언급한 모형의 독립 변수들은 Figure 4.1의 상관 계수 맵(correlation map)에서 확인할 수 있듯이 높은 상관 계수를 갖는 변수들을 포함한다. 이러한 변수들은 다중공선성(multicolinearity)으로 인하여 모형의 추정 시 추정량의 분산을 매우 크게 증가 시킬 수 있으므로 분석에 주의가 요구된다. 이에 대한 해결을 위해서 능형 추정량(ridge estimator) 또는 주성분(principal component)을 이용한 모형의 적합을 고려할 수 있으나 본 연구에서는 상관성이

높은 변수들을 그룹화하는 성질을 지닌 Elastic-net 벌점화 모형 (Zou와 Hastie, 2005)을 고려하였다. Elastic-net은 모수의 ℓ_1 노음(norm)에 벌점을 고려한 LASSO (least absolute shrinkage and selection operator) 벌점 (Tibshirani, 1996)과 ℓ_2 노음의 제곱에 벌점을 고려한 능형 벌점의 볼록 조합(convex combination)으로

$$\lambda\alpha\|\beta\|_2^2 + \lambda(1-\alpha)\|\beta\|_1 \tag{4.5}$$

로 정의되는 벌점항을 고려한 모형으로 LASSO 벌점을 고려한 모형과 비교하여 예측력이 우수함이 수치 실험을 통하여 보고되었다 (Zou와 Hastie, 2005). 여기서 λ는 모수의 희박함을 조절하는 조율 모수(tuning parameter)이며 $0 \leq \alpha \leq 1$인 조율 모수로 LASSO ($\alpha = 0$)와 능형 ($\alpha = 1$) 벌점의 조합의 정도를 결정한다. 본 연구에서는 통합된 60개의 독립 변수들에 대한 Elastic-net 벌점화 회귀 모형의 적합을 위하여 R의 `elasticnet` 패키지 (Zou와 Hastie, 2005)를 활용하였으며 순위 로짓 및 순위 프로빗 모형의 적합을 위하여 R의 `ordinalNet` (Wurm 등, 2017)을 사용하였다. 회귀 모형을 적합하는 경우, 모형의 가정에 대한 영향을 고려하여 조율 모수의 선택은 제곱 손실 함수를 고려하여 예측 오차의 제곱합을 최소화하는 조율 모수를 선택하였으며 순위 로짓 및 순위 프로빗 모형의 경우 가능도 함수(likelihood function) 기반으로 Akaike 정보 기준(Akaike information criterion; AIC)을 최소화하는 조율 모수를 선택하였다.

5. 회사채 신용 등급 예측 모형의 성능 비교

본 절에서는 앞 서 살펴본 기존의 예측 모형 및 Elastic-net 벌점을 적용한 회귀 모형, 순위 로짓, 순위 프로빗 모형에 대한 결과를 정리하고 예측 성능의 비교를 위하여 D-등급 차이 예측 정확도(accuracy) ACC_D를 아래와 같이 정의한다.

$$\text{ACC}_D = \frac{\sum_{i=1}^{n_{\text{test}}} I(|y_i - \hat{y}_i| \leq D)}{n_{\text{test}}} \times 100, \tag{5.1}$$

여기서 n_{test}는 검증 자료의 수를 나타내며 $I(\cdot)$는 지시함수(indicator function)을 나타낸다. 보다 정확한 비교를 위하여 서론에서 언급한 바와 같이 수집된 자료를 5개의 조각으로 나누어 약 80%는 훈련 자료로 활용하여 모형의 추정 및 조율 모수 선택에 활용하고 나머지 20%를 이용하여 모형을 검증, 예측 정확도를 계산한다. 이러한 과정을 5회 반복하는 5-fold 교차 검증을 통하여 예측 정확도의 평균 및 표준 오차를 계산하였다.

5.1. 기존 연구의 예측 모형 결과 비교

본 절에서는 기존 연구에서 적용한 예측 모형을 수집한 자료에 적합하고 예측 정확도의 비교를 진행하고자 한다. 앞서 살펴본 회귀 모형, 순위 로짓 모형, 순위 프로빗 모형, SVM 모형, RF 모형을 각각 제안된 변수에 대하여 적합하고 5-fold 교차 검증 기반의 예측 정확도의 평균 및 표준 오차를 Table 5.1에 나타내었다. 비교 결과, 회사채 신용 등급을 정확하게 ($D = 0$) 예측하는 정도를 측정하는 ACC_0의 경우 랜덤 포레스트가 69.64%로 가장 높게 나타났으며 SVM 모형, 순위 프로빗, 순위 로짓, 선형 회귀 모형 순으로 나타났다. 1등급 오차 범위 ($D = 1$)를 고려할 경우 여전히 랜덤 포레스트와 SVM 모형의 예측 정확도가 각각 87.74%, 80.58%로 높게 나타났으며 회귀 모형의 정확도도 75.33%로 크게 증가하였다.

다음으로 연구의 재현성(reproducibility)에 대하여 살펴보기 위해 추정량의 크기 및 방향성을 비교하고자 한다. 기존 연구들과 비교하여 본 연구는 조사된 자료의 기간과 대상의 차이가 존재하고 추정량 자체의 임의성(randomness)으로 인하여 연구 결과의 추정값은 기존의 연구 결과와 차이가 존재할 수 있

Table 5.1. Summary of prediction accuracies for OLS, OL, OP, SVM, and RF

Prediction model	Type	ACC_D	
		$D = 0$	$D = 1$
Kim & Choi (2006)	OLS	31.97 (0.78)	75.33 (0.67)
Elastic-net	OLS-EN	33.87 (0.91)	79.71 (0.88)
Kim *et al.* (2006)	OL	32.41 (1.32)	67.45 (1.17)
OrdinalNet	OL-EN	43.50 (1.21)	77.77 (1.00)
Kim & Choi (2006)	OP	38.25 (1.49)	71.68 (1.35)
OrdinalNet	OP-EN	42.04 (1.23)	77.37 (1.13)
Kim & Ahn (2016)	SVM	59.42 (1.17)	80.58 (2.10)
Kim & Ahn (2016)	RF	69.64 (0.59)	87.74 (0.43)

OLS = ordinary least squares; OL = ordered logistic; OP = ordered probit; SVM = support vector machine; RF = random forest. OLS-EN and OL-EN (OP-EN) denote the OLS and the OL (OP) models with the elastic-net penalty, respectively. Numbers in parenthesis denote the standard errors.

Table 5.2. Summary of estimates of coefficients of the ordinary least squares

Variable	Estimate	Standard errors	p-value	VIF	Estimate (Kim & Choi, 2006)
Intercept	6.021	1.069	<0.00001	-	−3.619
X_1	−0.077	0.361	0.8310	14.173	0.341
X_2	1.642	0.306	<0.00001	10.424	0.966
X_6	−1.370	0.168	<0.00001	2.118	−1.332
X_7	0.001	0.002	0.5397	1.039	−0.209
X_9	−0.746	0.099	<0.00001	1.086	−0.728
X_{13}	5.601	0.513	<0.00001	2.118	3.412
X_{29}	−149.300	14.350	<0.00001	6.087	1.217
X_{30}	−0.147	0.083	0.077	2.474	−7.299

VIF = variance inflation factor.

다. 하지만, 연구의 재현성 측면에서 신용 등급에 대한 독립 변수의 영향의 정도 및 방향성은 어느 정도 유지 되어야 할 것이다. 회귀 계수의 비교를 위하여 5회의 반복 중 마지막 반복에 대한 훈련 자료의 회귀 모형의 회귀 계수 추정값을 Table 5.2에 정리하였다. Kim과 Choi (2006)의 결과와 비교하면 8개의 변수 중에서 X_2, X_6, X_9, X_{13}, X_{30}의 5개의 변수의 방향성이 일치하였다. 크게 차이가 나타난 변수는 X_{29}인 경상이익/자기자본으로 기존 연구의 경상이익/평균자기자본 변수와 대응하는 변수이나 기존에 발표된 논문 (Kim과 Choi, 2006)에 제시된 기술통계의 값을 비교하면 본 연구에서 수집된 X_{29}의 경우 평균 0.01, 표준편차 0.01, 최솟값/최댓값 0.00/0.04로 작은 범위의 값으로 관측되었으나 Kim과 Choi (2006)의 연구에서는 평균 0.062, 최솟값/최댓값 −2.016/1.557로 상대적으로 큰 범위의 값이 이용되어 회귀 계수 차원의 상대적인 크기 차이는 발생할 수 있다. 추가로 독립 변수들의 종속 변수에 대한 영향의 방향성에 대한 차이가 발생 할 수 있는 원인으로는 다중공선성을 생각할 수 있다. 적합된 모형의 분산 팽창 인자(variance inflation factor; VIF)를 확인하면 X_1, X_2, X_{29}가 상대적으로 큰 값을 지니며 X_1과 X_2의 VIF 값이 10이상으로 공선성이 존재한다고 판단할 수 있다. 기존의 연구 결과와 유사한 결과를 갖는 변수로는 log 매출액(X_2), 베타계수(X_6), 주가순자산비율(X_9), 누적이익률(X_{13})이 있으며 기존 연구에서는 누적시장조정 수익률(X_7)에 대한 회귀 계수의 부호가 (−)로 기대와 반대로 추정되었다고 언급 하였으나 본 연구에서는 유의한 영향을 주지 않는 것으로 판단되었다.

추가적으로 순위 로짓 모형에서 추정된 회귀 계수 추정값을 Table 5.3에 보고 하였다. 상대적인 크기에

Table 5.3. Summary of estimates of coefficients of the ordered logistic

Variable	Estimate	Standard errors	p-value	Estimate (Kim *et al.*, 2006)
X_1	3.088	0.180	<0.00001	1.55
X_{14}	5.618	1.095	<0.00001	0.06
X_{15}	−65.725	6.090	<0.00001	−0.37
X_{31}	6.907	0.758	<0.00001	0.04
X_{35}	−1.283	0.436	0.00328	−0.02

Table 5.4. Summary of the chosen tuning parameters (α, λ)

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
OLS	(0, 0.48)	(0, 0.33)	(0, 0.33)	(0, 0.28)	(0, 0.39)
OL	(0, 0.00017)	(0, 0.00020)	(0, 0.0003)	(0, 0.00013)	(0, 0.00033)
OP	(0, 0.00063)	(0, 0.00042)	(0, 0.00042)	(0, 0.00028)	(0, 0.00063)

OLS = ordinary least squares; OL = ordered logistic; OP = ordered probit.

는 차이가 존재하나 기존 연구와의 방향성은 모두 일치함을 확인하였다. 단, 본 연구에서 적용한 등급의 순서는 Kim 등 (2006)에서 적용한 등급의 순서와 반대이며 Kim 등 (2006)에서의 β 모수를 $-\beta$의 형태로 표현하여 Table 5.3의 부호의 일치는 동일한 방향성을 나타냄을 밝혀둔다.

5.2. Elastic-net 변수 선택 기반 예측 모형 결과 비교

앞서 살펴본 바와 같이 기존의 예측 모형에 사용 된 몇몇 변수들 사이에는 강한 상관성이 존재하며 이는 예측 모형의 모수 추정을 불안정하게 하며 모형의 신뢰도 및 예측력을 낮추는 요인이 될 수 있다. 따라서 본 절에서는 변수들 사이의 상관성이 높은 경우 회귀 계수의 차이를 줄여주는 그룹화 효과를 지닌 Elastic-net 벌점 (Zou와 Hastie, 2005)을 고려하여 회귀 모형, 순위 로짓, 순위 프로빗 모형에 적용하고 예측 성능의 향상에 대하여 살펴 보고자 한다. 먼저 본 연구에서 적용되는 자료는 회귀 모형의 기본 분포 가정에 위배되므로 Elastic-net의 조율 모수의 선택의 기준으로 가능도 함수 기반의 모형 선택이 아닌 교차 검증을 이용하여 예측 오차를 최소화하는 조율 모수 λ와 α를 최적의 조율 모수로 정하였다. 순위 로짓 및 순위 프로빗 모형의 경우, 관측 자료에 타당한 분포 가정이 가능하므로 효율적인 계산과 예측력이 높은 모형을 선택하기 위하여 AIC (Akaike, 1974)을 적용하여 최적의 조율 모수를 선택하였다.

Elastic-net의 조율 모수는 5번의 반복마다 각각 그리드 탐색(grid search)를 이용하였으며 선택된 조율 모수는 Table 5.4에 정리한 것 처럼 $\alpha = 0$으로 선택되었다. 단, `ordinalNet` (Wurm 등, 2017)의 경우 Elastic-net 벌점이 $\lambda(1-\alpha)/2\|\beta\|_2^2 + \lambda\alpha\|\beta\|_1$로 정의되나 표기의 편의성을 위하여 $(\lambda\alpha/2)\|\beta\|_2^2 + \lambda(1-\alpha)\|\beta\|_1$로 표현하였다. 본 연구에서는 Elastic-net 벌점을 고려하였으나 관측된 자료에 기반한 조율 모수의 선택에서 $\alpha = 0$으로 선택 되어 최종적으로는 LASSO 벌점을 적용한 모형을 적용한 결과와 동일하게 되었음을 참고하기를 바란다. 선택된 조율 모수를 토대로 각각의 예측 정확도를 계산하였고 이를 Table 5.1에 OLS-EN, OL-EN, OP-EN으로 요약하였다. Table 5.1에 나타난 것처럼 LASSO 벌점(elastic-net with $\alpha = 0$)을 적용한 예측 모형의 정확도가 회귀 모형, 순위 로짓, 순위 프로빗 모형과 비교 시 약 2%∼10% 정도 향상된 것을 확인할 수 있다. 또한 LASSO 벌점이 적용되어 모수의 추정 뿐만 아니라 변수 선택을 동시에 진행한 효과를 얻을 수 있는 이점이 있다. Table 5.5에 Elastic-net 벌점을 고려한 3가지 모형의 마지막 반복에서의 추정값을 정리하였다. 각 모형에서 60개의 회귀 계수 중에서 약 40∼50개의 회귀 계수가 0이 아닌 값으로 추정되었으며 Table 5.2에 있는 결과와 비교 시 추정값이 유사하게 나타남을 확인할 수 있다. 또한 순위 로짓과 순위 프로빗의 경우 0이 아닌 추정값의 패턴이

Table 5.5. Summary of the estimated coefficients of OLS-EN, OL-EN, OP-EN ($\alpha = 0$, LASSO)

Variable	OLS-EN	OL-EN	OP-EN	Variable	OLS-EN	OL-EN	OP-EN
X_1	0.81	-	-	X_{31}	−1.48	-	-
X_2	−1.44	3.52	1.75	X_{32}	1.92	−3.46	−1.80
X_3	2.13	−3.47	−2.03	X_{33}	1.93	−0.48	−0.20
X_4	−1.14	0.67	0.47	X_{34}	1.56	−0.25	−0.29
X_5	0.34	−0.91	−0.42	X_{35}	1.27	−0.52	−0.29
X_6	−1.05	1.76	0.98	X_{36}	0.52	−0.49	−0.36
X_7	0.02	0.02	-	X_{37}	−0.08	0.12	0.09
X_8	−0.11	0.29	0.21	X_{38}	−0.06	-	0.08
X_9	−0.66	1.22	0.65	X_{39}	−1.05	−0.05	-
X_{10}	0.55	−1.31	−0.61	X_{40}	−0.72	1.40	0.78
X_{11}	−0.22	0.20	-	X_{41}	1.29	−1.44	−0.81
X_{12}	1.34	-	-	X_{42}	2.06	−3.13	−1.80
X_{13}	4.17	−7.03	−4.16	X_{43}	−4.68	1.49	0.78
X_{14}	1.84	−2.24	−1.32	X_{44}	-	-	-
X_{15}	−7.27	3.25	-	X_{45}	−0.98	1.06	0.71
X_{16}	-	-	-	X_{46}	−0.58	-	-
X_{17}	-	-	-	X_{47}	-	-	-
X_{18}	−0.04	−0.03	−0.01	X_{48}	1.00	−0.74	-
X_{19}	−7.61	11.67	6.08	X_{49}	−0.08	0.10	0.04
X_{20}	-	-	1.44	X_{50}	−0.10	0.16	0.10
X_{21}	−0.04	0.06	0.03	X_{51}	1.06	-	-
X_{22}	10.83	−15.77	−8.23	X_{52}	−1.21	-	0.09
X_{23}	−15.63	21.00	10.62	X_{53}	-	-	-
X_{24}	−0.14	2.17	1.03	X_{54}	-	-	-
X_{25}	-	-	-	X_{55}	-	-	-
X_{26}	0.22	-	-	X_{56}	0.01	−0.01	0.00
X_{27}	-	-	-	X_{57}	0.24	−0.29	−0.18
X_{28}	1.29	−3.53	−1.88	X_{58}	1.45	−2.08	−1.06
X_{29}	−39.44	-	-	X_{59}	0.02	−0.13	−0.07
X_{30}	−0.62	0.90	0.49	X_{60}	−0.04	0.11	0.06

OLS-EN = ordinary least squares with the elastic-net penalty, OL-EN = ordered logistic with the elastic-net penalty; OP-EN = ordered probit with the elastic-net penalty.

유사한 반면 회귀 모형 기반의 회귀 계수 추정값은 상대적으로 0이 아닌 추정값을 더 많이 갖는다. 단, Table 5.5에서 순위 로짓 및 순위 프로빗 모형의 회귀 계수는 회귀 모형에 적용된 회귀 계수 β와는 부호가 반대인 $-\beta$로 정의되었음에 유의하길 바란다.

6. 결론

본 논문은 회사채 신용 등급 예측 모형과 관련하여 기존의 예측 모형을 동일한 자료와 신용 등급 구간을 고려하여 예측 성능을 비교하였으며 기존에 제안된 예측 모형에서의 다중공선성 문제를 확인하였다. 다중공선성 문제의 해결과 회귀 모형, 순위 로짓, 순위 프로빗 모형의 예측 성능 향상을 위하여 Elastic-net 벌점을 고려한 예측 모형을 적용하였으며 수집된 자료를 토대로 예측 성능이 향상됨을 확인하였다.

회귀 모형 및 순위 로짓, 순위 프로빗 모형은 앙상블 모형과 비교하여 이해하기 쉽고 해석이 용이하다는

장점을 지니나 본 연구의 결과로 미루어 볼 때 신용 등급 예측의 다중 클래스 분류 문제에서는 앙상블 방법이 가장 뛰어난 예측 정확도를 지님을 확인하였다. 따라서 회사채 신용 등급의 예측력이 우선 시 되는 경우 랜덤 포레스트 기반의 예측 모형을 적용하는 것이 유리할 것이며 신용 등급의 유지 및 관리 관점에서 접근 시 Elastic-net 벌점을 고려한 회귀 모형 또는 순위 로짓, 프로빗 모형을 적용하는 것도 충분히 의미가 있을 것이다.

본 연구는 국내 회사채 신용 등급 예측 모형의 비교에 중점을 두어 상대적으로 국제 신용 평가사에서 공시하는 기업의 신용 등급 및 국외 기업의 회사채 신용 등급에 대한 부분을 다루지 못하였다. 추후 연구에서는 국제 회계 기준을 따르는 국내외의 기업을 대상으로 재무 정보를 수집하고 국제 신용 평가사와 국내 신용 평가사의 공시 자료를 기반으로 국내외 자료를 통합한 신용 등급 또는 부실 예측 모형을 개발하고자 한다. 추가적으로 본 비교 연구에 사용된 685개의 기업-년 자료는 186개의 기업에 대한 연도별 신용 등급과 재무 자료로 이루어져 있으나 기존 연구들의 비교 목적으로 본 연구의 모형들은 자료의 시계열적 특성을 고려하지 않았다. 자료의 시계열적 특성을 반영한 모형은 추후 연구에서 다루고자 한다.

References

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**, 716–723.

Altman, E. I. and Katz, S. (1976). Statistical bond rating classification using financial and accounting data. In *Proceeding of the Conference on Topical Research in Accounting*, NYU Press, 205–239.

Breiman, L. (1996). Bagging predictors, *Machine Learning*, **24**, 123–140.

Ederington, L. H. (1986). Why split ratings occur, *Financial Management*, **15**, 37–47.

Horrigan, J. O. (1966). The determination of long-term credit standing with financial ratios, empirical research in accounting: selected studies, *Supplement to Journal of Accounting Research*, **4**, 44–62.

Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., and Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study, *Decision Support Systems*, **37**, 534–558.

Jeong, C. J. (2011). *The empirical study on factors affecting corporate credit ratings* (Unpublished master's thesis), Kyonggi University, Suwon, Korea.

Kaplan, R. S. and Urwitz, G. (1979). Statistical models of bond ratings: a methodological inquiry, *Journal of Business*, **52**, 231–261.

Kim, J. S. and Choi, Y. M. (2006). Development of a bond rating prediction model based on financial and stock price-based variables, *Study on Accounting, Taxation & Auditing*, **43**, 185–217.

Kim, K. J. and Kim, J. S. (2002). Development of bond rating prediction model for effective interest rate estimation, *Korean Accounting Journal*, **11**, 81–100.

Kim, M. J. (2012). Ensemble learning with support vector machines for bond rating, *Journal of Intelligence and Information System*, **18**, 29–45.

Kim, S. J. and Ahn, H. (2016). Application of random forests to corporate credit rating prediction, *The Journal of Business and Economics*, **32**, 187–211.

Kim, S. T., Lee, J. J., and Hong, J. B. (2006). The prediction model of bond-rating with ordered logit analysis, *Journal of the Korean Data Analysis Society*, **8**, 641–654.

Ko, D. P. and Kim, H. M. (2002). Using financial health index approach to credit analysis, *Industrial Management Review*, **25**, 231–254.

Pinches, G. E. and Mingo, K. A. (1973). A multivariate analysis of industrial bond ratings, *Journal of Finance*, **28**, 1–18.

Seo, Y. H. (2015). *The effect of revenue-expense matching on corporate bond ratings* (Unpublished master's thesis), Chung-Ang University, Seoul, Korea.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **58**, 267–288.

West, R. R. (1970). An alternative approach to predicting corporate bond ratings, *Journal of Accounting Research*, **8**, 118–125.

Wurm, M. J., Rathouz, P. J., and Hanlon B. M. (2017). Regularized ordinal regression and the ordinalNet R package, *arXiv preprint arXiv:1706.05003*.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society. Series B*, **67**, 301–320.

국내 회사채 신용 등급 예측 모형의 비교 연구

박형권[a] · 강준영[a] · 허성욱[a] · 유동현[a],[1]

[a]인하대학교 통계학과

(2018년 3월 29일 접수, 2018년 4월 30일 수정, 2018년 5월 2일 채택)

요 약

회사채 신용 등급 예측 모형에 대한 연구는 신용 평가 기관이 회사채 신용 등급 평가에 사용될 것이라 예상 되는 여러 재무적 특성 변수들을 기반으로 진행되었으며 선형 회귀 모형(linear regression), 순위 로짓(ordered logit), 순위 프로빗(ordered probit), 서포트 벡터 기계(support vector machine), 랜덤 포레스트(random forest) 등 다양한 모형들을 적용하여 개발되었다. 하지만 기존 연구들에서 고려한 회사채 신용 등급은 연구에 따라 5등급에서 20등급까지 다른 등급 구간을 적용하였으며 분석에 이용된 표본 자료의 기간 및 대상도 상이하여 예측 성능의 공정한 비교에 어려움이 있다. 따라서 본 연구에서는 2013년부터 2017년까지의 회사채 신용 등급 자료와 기존 연구들에서 사용된 재무 지표들을 통합하여 기존에 발표된 예측 모형들을 동일한 자료에 적용하고 예측 성능을 비교하였다. 추가적으로 Elastic-net 벌점화 회귀 모형 및 순위 로짓, 순위 프로빗 모형을 적합하여 LASSO 벌점이 선택됨을 확인하였으며 LASSO 벌점을 고려한 예측 모형이 대응하는 기존의 예측 모형들보다 향상된 성능을 보임을 확인하였다. 본 연구의 수행 결과, 랜덤 포레스트를 이용한 예측 모형이 15등급 기준 검증 자료에서 정확한 등급 예측률이 69.6%로 다른 모형과 비교하여 높은 예측 성능을 나타내었다.

주요용어: 회사채 신용 등급, 선형 회귀, 순위 로짓, 랜덤 포레스트, Elastic-net

이 논문은 2017년도 인하대학교의 지원에 의하여 연구되었음 (INHA-55456).

[1]교신저자: (22212) 인천광역시 남구 인하로 100, 인하대학교 통계학과. E-mail: dyu@inha.ac.kr