財務管理硏究 第四十卷 第五號(2023. 10), 213~243

기계학습을 이용한 회사채 발행금리 예측*

안지영** · 임병권***

〈요 약〉

본 연구는 머신러닝 알고리즘을 이용하여 회사채 발행금리 예측의 유용성에 대해 고찰하였다. 구체적으로 회사채 특성과 함께 거시경제적 요인, 채권 및 주식시장 전반에 관한 정보와 함께 회사채 발행기업의 재무적 특성 등을 종합적으로 고려하여 회사채 스프레드에 영향을 미치는 요인을 규명한 후 발행금리 예측 모형을 설계하였다.

본 연구의 주요 분석결과를 요약하면 다음과 같다. 첫째, 회사채 발행금리 예측에는 신용등급, 신용스프레드, 장단기금리차, 기준금리, GDP 등이 주요한 예측 변수로 작용하는 것으로 나타난다. 또한, 최적의 예측 모델은 모델의 형태나 변수의 개수 그리고 표본의 크기 등에 따라 다양하여 분석 데이터셋에 따라 예측 모델의 성능이 달라짐이 확인된다. 이는 회사채 발행금리 예측에 있어 머신러닝 알고리즘의 활용이 유용하나, 데이터에 따라 최적 모형의 선택이 달라질 수 있음을 의미한다. 둘째, 본 연구에서는 데이터 기반의 예측 모형을 활용하는 경우 발행 실적이 상대적으로 저조한 ESG 채권 발행금리 예측에도 회사채 데이터 기반 모델이 유용하게 활용될 수 있음이 확인된다.

전체적으로 회사채 발행금리 예측에 있어 머신러닝 기법은 유용하게 활용될 수 있을 것으로 판단된다. 따라서 회사채 발행금리 결정에 있어 본 연구는 새로운 시각을 제공하며, 채권 발행기업과 투자자에게 유용한 정보를 제공할 수 있을 것으로 기대된다.

주제어 : 기계학습, Elastic Net, 회사채, ESG 채권, 발행금리

논문접수일 : 2023년 08월 06일 논문수정일 : 2023년 10월 11일 논문게재확정일 : 2023년 10월 24일

* 이 논문은 2023년도 과학기술정보통신부의 재원으로 과학기술사업화진흥원의 지원을 받아 수행된 연구임 (1711195821, 지역 과학기술 성과 실용화 지원사업(충남대학교)).

** 제1저자, 에너지경제연구원 부연구위원, E-mail: ajy4129@gmail.com

*** 교신저자, 충남대학교 대학원 기술실용화융합학과 산학협력교수, E-mail: bk81.lim@gmail.com

Ⅰ. 서 론

기업의 자금조달은 대표적으로 주식이나 채권발행 그리고 금융기관의 차입을 통해 이루어지는데, 직접금융시장을 통한 회사채 발행은 주요 외부 자본조달 방안 중 하나이다. 여기서 회사채 발행 시 결정되는 발행금리는 발행자인 기업 입장에서는 자금조달비용이며, 투자자인 채권자에게는 위험에 대한 대가인 리스크 프리미엄(risk premium)이 된다. 회사채 발행금리는 발행 당시의 시장금리 수준이나 채권시장의 수급 그리고 발행기업의 신용도 등의 다양한 요인에 영향을 받을 수 있다. 따라서 회사채 발행금리에 영향을 미칠 수 있는 다양한 요인을 규명한 후, 이를 토대로 적정한 발행금리를 추정하는 것은 회사채시장의 발전을 위해 필수적인 요소로 볼 수 있다.

한편, 회사채 발행금리에 미치는 영향에 대해 국내의 경우 제한적인 수준에서 연구가 진행된 상황이다. 기존연구에서 코로나19나 수요예측제도 도입 그리고 경제정책 불확실성 정도가 회사채 가격에 어떠한 영향을 미치는지를 규명하고 있으며(정희준 외 2인, 2021; 채병권, 한재현, 2020; 황광숙, 이준희, 2022), 기업의 재무정보 또는 회계처리 방법과 회사채수익률 간의 관련성에 대해서도 분석하고 있다(윤윤석 외 2인, 2005; 최보람 외 2인, 2012). 추가적으로 거시적 측면에서 안전자산 선호현상이나 거시경제에 대한 시장참여자들의 이질적인 기대가 채권수익률에 어떠한 영향을 미치는지에 대해서도 고찰하고 있다(김도완, 2018; 배광일, 이순희, 2020). 하지만, 채권의 발행금리에 미치는 영향은 전술한 요인들뿐만 아니라 채권의 고유 특성이나 자본시장(채권시장과 주식시장) 전반의 상황 그리고 시장금리 등과 같은 다양한 측면에 영향을 받을 가능성이 있다. 따라서 이와 같은 요인들을 종합적으로 고려하여 회사채 발행금리가 주로 어떠한 측면에 영향을 받는지를 고찰해 볼 필요가 있다.

이와 같은 배경하에 본 연구는 2001년부터 2022년까지 기간을 대상으로 국내 채권시장에서 발행된 회사채를 이용하여 발행금리에 미치는 요인들을 규명하고 예측 모형을 제시하고자 한다. 이를 위해, 최근 예측을 위해 널리 활용하고 있는 머신러닝 알고리즘 모델을 이용하여 회사채 발행금리를 예측할 수 있는 모형을 설계하고자 한다. 이후 머신러닝 기반의 모형과 기존의 선형회귀모형과의 예측 정합성을 비교・분석하여 어떠한 모형이 회사채 발행금리 예측에 보다 유용하게 활용될 수 있는지를 고찰한 후 관련 시사점을 제시하고자 한다.

보다 구체적인 연구의 목적은 다음과 같다. 우선 회사채 발행금리 추정을 위해 발행시점의 회사채가 갖는 고유한 특성 변수(신용등급, 발행금액, 발행만기, 보증채 여부 등)와 거시변수(경기지수, 산업생산지수, 장단기금리차, GDP, 소비자물가지수 등), 채권시장 변수(국고채 금리, 회사채 어음부도율, 신용스프레드, 안전자산 발행금액 및 거래대금 수준, 채권지수

등), 주식시장 변수(주가지수, 주식시장 거래대금, 변동성지수 등), 채권 발행기업의 재무변수(자산, 부채비율, 영업이익률 등)를 종합적으로 이용한다. 또한, 발행금리 예측을 위해 전통적으로 활용되는 다중선형회귀모형과 함께 머신러닝은 LASSO(Least Absolute Shrinkage and Selection Operator), Ridge 및 Elastic net의 3가지 기법을 활용한다.[1] 이상과 같이 다양하게 선정된 변수들과 모형(전통적인 선형회귀모형과 3가지 머신러닝 기법)을 기초로 어떠한 변수와 모형이 회사채 발행금리 예측에 적절하게 이용될 수 있는지를 분석하고자 한다.

본 연구의 주요 분석결과를 요약하면 다음과 같다. 첫째, 회사채 발행금리 예측에 있어 회사채 특성 외에도 채권 및 주식시장, 거시경제 상황, 기업의 재무 변수 등을 포함하여 분석 데이터가 갖는 정보가 클수록 머신러닝 알고리즘을 통한 예측 모형 개발에 정합성이 높은 것으로 나타났다. 특히, 예측 변수의 개수가 충분히 많은 경우에 일반 선형회귀모형과 머신러닝 모형간의 예측 성과 차이가 두드러지게 나타나는 것으로 확인되었다.

둘째, 데이터의 구성에 따라서 최적 예측 모형은 달라질 수 있음이 확인된다. 즉, 예측에 적합한 모형은 절대적이지 않으며, 분석 데이터 내 변수의 개수나 형태, 표본 크기 그리고 표본 구성 등에 따라 가변적으로 나타난다. 신용등급, 신용스프레드, 장단기금리차, 기준금리, GDP, 국고채 1년물 금리 등은 데이터셋에 관계 없이 일관적으로 중요한 예측 요인으로 확인된다. 또한, 신용스프레드, 기준금리와 같이 시장 상황을 나타내는 변수가 기업의 재무상태를 나타내는 총자산, 부채비율과 상호작용함을 고려하여 예측 모형을 구성하는 경우 예측 기여도를 더 높일 수 있음을 확인하였다. 반면, 실업률, 채권시장 전체 거래대금, 안전자산 거래대금, 코스피지수, 국고채 3년물, 5년물, 10년물 금리 등의 변수는 모형에 따라 예측 중요도가 매우 낮거나 유의미하지 않은 것으로 나타났다. 셋째, 일반 회사채 데이터를 바탕으로 ESG 채권의 발행금리를 예측하는 것도 예측 성과가 높을 수 있음을 확인되었다. 다만 예측 성과 지표를 RMSE(Root Mean Squared Error) 기준으로 모형을 선택하는 경우가 MAE(Mean Absolute Error) 기준보다 ESG 채권 발행금리를 예측하는데 보다 유의미한 것으로 나타났다.

본 연구가 갖는 기존연구와의 차별성 및 연구의 중요성은 다음과 같다. 첫째, 회사채 발행가격에 미치는 영향에 대해 국내연구에서는 정책적 측면 또는 발행기업의 재무적 측면에 중점을 두고 연구가 진행되었다(정희준 외 2인, 2021; 채병권, 한재현, 2020; 황광숙, 이준희, 2022; 윤윤석 외 2인, 2005). 또한, 머신러닝 기법을 이용한 예측은 주가나 기업의 부도

1) OLS(Ordinary Least Squares)를 이용한 선형회귀분석에서는 평균제곱오차(Mean Square Error, MSE)의 최소값을 통해 최적의 모델을 찾아낸다. 하지만, 다중회귀모형에서는 복잡도가 높아 과대적합 될 수 있다. 따라서 이를 통제하기 위해 모델 자체적으로 규제를 부여하는 방법이 LASSO와 Ridge이다.

또는 신용정보 등을 대상으로 분석을 행하고 있다(권혁건 외 2인, 2017. 김경목 외 2인, 2021; 송민찬, 류두진, 2021; 이현상, 오세환, 2020). 따라서 머신러닝 기법을 이용하여 회사채 발행금리를 예측하고자 하는 본 연구는 기존연구와 차별성이 존재한다. 한편, Binanch et al.(2021)과 Kim et al.(2021)은 미국 시장을 대상으로 머신러닝 기법을 이용하여 미국 국채(Treasure bond) 또는 회사채 수익률을 분석하고 있으며, 머신러닝 기법의 예측 유용성을 제시하고 있다. 따라서 국내 회사채 시장에서도 머신러닝 기법이 채권 수익률 예측에 유용한지를 검증해 볼 필요성이 있으며, 해당 내용을 분석하고자 하는 본 연구는 이론적으로 중요성이 존재한다.

둘째, 본 연구는 최근 사회과학 분야에서 중요성이 증대되고 있는 머신러닝 예측 모델을 이용하여 회사채 발행금리 예측에 유의미한 변수를 찾고 예측성과를 높일 수 있는 분석 알고리즘을 제안한다. 즉, 본 연구는 전통적인 선형회귀모형과의 비교를 통해 머신러닝 기법이 회사채 발행금리 예측에 유용하게 활용될 수 있는지 여부를 실증적으로 규명함으로써, 주가 또는 기업부도뿐만 아니라 회사채 시장에서도 머신러닝 기법이 유용하게 활용될 수 있는지를 고찰한다. 이를 토대로 적절한 회사채 발행금리 예측을 위한 변수 및 모델 등을 제시함으로써, 실무적으로도 중요한 시사점을 전달해 줄 수 있을 것으로 판단된다.

이하 본 논문은 다음과 같이 구성된다. 우선 제Ⅱ장에서는 회사채 스프레드에 영향을 미치는 요인 또는 머신러닝 기법을 이용한 예측에 관한 기존연구를 살펴보고, 제Ⅲ장에서는 본 연구의 내용 및 분석방법에 관해 설명한다. 그리고 제Ⅳ장에서는 실증분석 결과를 제시하고 마지막 제Ⅴ장 결론에서는 연구결과를 요약하고 시사점을 제시하고자 한다.

Ⅱ. 기존연구

기업의 자금조달 수단에 있어 발행하는 채권의 발행금리는 조달비용이므로 발행금리가 어떠한 요인들에 영향을 받는지에 대해 국외에서는 다양한 연구가 진행되고 있다. 해외의 기존연구에 의하면, 거시경제적 요인이나 채권 발행기업의 고유 특성 그리고 채권의 유동성 등 다양한 요인이 채권의 수익률에 영향을 미치고 있다는 결과를 보여주고 있다(Athanassakos and Carayannopoulos, 2001; Bernoth and Erdogan, 2012; Chen et al., 2007; Favero et al., 2010; Han and Zhou, 2014; Huang et al., 2015; Jubinski and Lipton, 2011; Mayberger et al., 2014; Wang et al., 2008; Saltzman and Yung, 2018). 하지만, 국내시장의 경우 제한된 범위 내에서 채권가격에 영향을 미치는 요인에 대해 연구가 진행된 상황이다. 채권의 발행금리에 관한 국내 주요 연구를 소개하면 다음과 같다.

우선, 김도완(2018)은 신용등급이 양호한 BBB-등급 이상에서는 안전자산 선호현상이 채권발행에 미치는 영향이 제한적이나, BBB- 등급 미만의 경우 회사채 신용 스프레드가 높아짐을 확인하였다. 즉, 안전자산 선호현상은 신용등급이 낮은 기업의 자금조달에 부정적 요인으로 작용함을 제시하고 있다. 그리고 양철원(2013)은 통화정책과 같은 외부의 요인은 채권시장과 주식시장의 유동성에 일련의 영향력을 미친다는 결과를 보여주고 있다. 또한, 두 시장의 수익률은 반대방향으로 움직여 주식시장에 부정적인 충격이 존재하면 안전자산인 채권으로 자금이 이동한다는 결과를 보여주고 있다.

다음으로 윤윤석 외 2인(2005)은 회계정보가 회사채수익률에 어떠한 영향을 미치는지에 대해 분석하였다. 분석결과 안정성과 관련된 재무비율인 자기자본비율, 부채비율, 유동비율과 함께 수익성 지표인 총자산수익률이 회사채수익률에 영향을 미친다는 결과를 보여주고 있다.

채병권, 한재현(2020)은 2012년 도입된 채권의 수요예측(book building)제도가 채권발행가격에 어떠한 영향을 미치는지 분석하였다. 분석결과 수요예측 제도 도입 이후에 발행된 채권들은 발행가격 고평가현상이 완화되어 가격 형성을 효율성이 증대된다는 시사점을 제시하고 있다. 최보람 외 2인(2012)은 보수주의 회계처리 방법이 기업의 채권발행금리에 어떠한 영향을 미치는지 검증하였다. 분석결과, 보수주의 회계처리 측정방법에 따라 채권 발행금리에 미치는 영향은 혼재되는 것으로 나타났다. 한편, 공모사채의 경우에는 보수주의 회계처리가 발행금리를 낮추는 역할을 수행하여 회계처리 방법에 따라 채권발행금리에 영향을 미친다는 결과를 제시하고 있다.

마지막으로 황광숙, 이준희(2022)는 미국에서 발표하는 경제정책불확성지수(EPU)가 국내의 회사채 신용스프레드에 어떠한 영향을 미치는지 분석하였다. 분석결과에 의하면, 미국 EPU(3 components)는 국내의 신용스프레드와 유의성이 매우 높으며, EPU(재정정책)의 경우에도 유의성이 높은 것으로 나타났다. 즉, 경제정책불확실성은 국내 회사채 스프레드에 일련의 영향을 미치고 있음을 시사하고 있다.

한편, 최근의 연구에서는 개별기업의 주식수익률이나 국가별 주가지수, 기업의 신용등급과 부도예측 등 재무금융 분야의 다양한 영역에서 기계학습(머신러닝, 딥러닝)을 이용하여 제반 유용성을 분석하고 있다. 기존연구에 의하면, 주가를 예측하거나(김경목 외 2인, 2021; Chen et al., 2023; Gu et al., 2020; Mishra and padhy, 2019), 기업의 부도를 예측하거나(권혁건 외 2인, 2017; 송민찬, 류두진, 2021; Moscatelli et al., 2020; Zhang et al., 1999), 또는 기업의 신용평점을 예측하는데 머신러닝 기법이 유용한지를 검증하고 있다(이현상, 오세환, 2020; Golbayani et al., 2020), 또한, 머신러닝 알고리즘을 이용하여 주택저당증권(Mortgage-Backed Securities; MBS)의 기초자산인 주택담보대출의 조기상환율도 예측하고 있으며

(안지영, 임병권, 2020). 증권발행신고서나 미국 연준의 베이지북 등을 대상으로 비정형화된 텍스트에 대해 머신러닝 기법을 이용하여 텍스트 분석도 이루어지고 있다(김용석, 조성욱, 2019; Li, 2008; Saltzman and Yung, 2018).

한편, 본 연구와 직접적으로 관련된 채권의 수익률 예측에 대해 Bianchi et al.(2021)과 Kim et al.(2021)은 머신러닝 기법을 이용하여 분석하고 있다. 우선, Bianchi et al.(2021)은 미국 국채를 대상으로 위험 프리미엄에 대해 주성분 회귀(Principal Components Regression; PCR) 또는 부분 최소제곱법(Partial Least Squares; PLS), 규제 선형 모형(Ridge, Lasso, Elastic net), 그리고 비선형 머신 러닝 기법인 회귀 트리(regression tree)와 인공신경망(neural network)을 이용하여 분석하였다. 분석결과, 회귀 트리와 인공신경망 기법이 국채 예측에 가장 유용한 것으로 나타났다. 또한, 인공신경망 예측의 경우 경기 역행적이며, 거시적인 불확실성과 시간 가변적인 위험회피 변수와 관련성이 존재한다는 결과를 제시하고 있다.

다음으로 Kim et al.(2021)은 OLS, 주성분회귀(PCR), 부분 최소제곱법(PLS), Gaussian copula marginal regression. 다양한 머신러닝 모델(Ridge, MARS, SVM, Random forest, Neural network) 등을 이용하여 회사채 수익률 스프레드를 분석하였다 있다. 이를 위해 유동성, 주식수익률 변동성, 표면금리, 장단기금리차, 이자율 변동성, 단기 금리, 채권 만기, 신용등급 등의 변수를 이용하였는데, 딥러닝 기법인 인공신경망의 경우에 예측 정합성이 가장 높으며, 회사채 수익률 스프레드 예측에 있어 기업의 신용등급 보다는 주식수익률 변동성이 가장 유용한 변수임을 제시하고 있다.

추가적으로 Kim(2021)은 환율, 주가지수, LIBOR 금리, S&P500 변동성, WTI 지수, 금 가격, 국내 및 해외 주요국가의 GDP와 CPI 등의 경제지표를 종합적으로 이용하여 국내 국고채 장단기금리차(10년-3년)를 예측하였다. 다양한 머신러닝 기법을 이용한 분석결과에서 AdaBoost 기법의 예측 정합성이 가장 높다는 결과를 제시하고 있다.

이상과 같이 최근 연구에서는 머신러닝 기법이 재무금융 분야의 다양한 영역에 활용하고 있으며, 예측의 유용성이 제시되고 있다. 따라서 국내 회사채 시장에 있어서도 머신러닝 기법이 유용성 여부에 대해 고찰해 볼 필요성이 있다.

Ⅲ. 연구내용 및 방법

1. 표본의 구성

본 연구는 2001년부터 2022년까지 기간 동안 발행된 채권 중 금융채, 특수채, 사모발행

및 유동화증권(ABS, MBS 등)을 제외한 후 공모로 발행된 회사채를 분석 대상으로 한다. 본 연구기간 동안에 분석 표본에 포함된 회사채는 총 6,924건인데 이를 유형별로 구분하는 경우 일반 회사채 6,797건과 ESG 채권 127건을 포함한다. 본 연구의 분석 표본은 독립적으로 발행된 6,924건의 채권에 대한 발행정보와 발행시점의 채권 및 주식시장 환경, 거시경제적 환경, 발행 기업의 재무정보와 같은 고유 특성 정보 등을 포함하는 횡단면 자료로 구성되어 있다.

회사채 발행금리 예측을 위해 본 연구에서는 전체 표본을 분석 표본과 예측 표본으로 각각 구분한다. 분석 표본은 일반 회사채 6,797건의 발행정보 및 발행 당시의 거시변수 등을 포함한다. 반면, 예측 표본은 ESG 채권 127건에 대한 동일 정보를 포함하도록 표본을 구성한다. 이렇게 표본을 구분한 것은 기존의 채권시장에서 회사채를 중심으로 설계된 발행금리 예측 모형이 ESG 채권 발행금리 예측에도 동일하게 적용될 수 있는지를 파악하기 위함이다. 또한, 머신러닝 알고리즘을 통해 채권 발행금리 예측 모형을 설계하고 예측성과를 검증한 후, 도출된 최적 모형을 이용하여 ESG 채권 발행금리 예측에도 활용함으로써 예측 정합성을 고찰하고자 한다. 추가적으로 ESG 채권 정보만을 포함하는 예측 표본은 2019년 9월부터 발행된 ESG 채권 127건에 대한 변수로 구성되어 있다.[2)]

실증분석을 위한 회사채 발행정보는 FnGuide의 Dataguide Pro와 연합인포맥스를 통해 추출 및 가공하여 분석에 이용한다. 그리고 한국은행 경제통계시스템, 금융투자협회 채권정보센터, 한국거래소를 통해 채권 및 주식시장 관련 지표와 거시경제 변수를 추가적으로 추출한 후 가공하여 이용한다. 한편, 채권시장 및 거시경제 변수 지표는 회사채 발행시점(연월)을 기준으로 매칭하여 분석에 활용하고자 한다.

2. 연구방법

1) 회사채 발행금리 예측 모형

채권의 발행금리를 예측하기 위해서는 앞서 기존연구에서 살펴보았듯 채권 자체에 대한 발행정보, 채권 및 주식시장 상황, 거시 경제 요인 등 다수의 변수를 활용할 필요가 있다. 본 연구에서는 선행연구에서 발행금리 예측 모형에 활용된 변수들을 종합적으로 검토하여 어떤 요인이 발행금리 예측에 주된 예측 요인으로 작용하였는지를 파악하기 위해서 고차원데이터의 축소(shrinkage)와 선택(selection)을 기반으로 한 발행금리 예측 모형을

2) 공공기관을 제외한 상장사 등의 ESG 채권 발행은 2019년부터 본격적으로 발행하기 시작하였다.

설계하고자 한다.

기본적으로 선형 회귀모형과 유사한 발행금리 예측 모형을 상정하고, 각 변수의 계수 추정치를 제한하거나 정규화(regularize)하는 머신러닝 기법을 활용하여 신규 채권 발행에 대한 발행금리 예측 성과를 높여주는 주요 변수를 파악한다.

일반 선형 회귀모형과 정규화 모형은 둘 다 회귀 분석에 사용되는 방법이나, 주요한 차이점이 있다. 일반 선형 회귀모형은 최소자승법(Ordinary Least Squares)을 이용하여 모델의 파라미터를 추정하는 방법으로, 오차의 제곱합을 최소화하여 각 예측 변수의 가중치(계수 추정치)를 결정하며 이렇게 추정된 가중치를 사용하여 종속 변수와 독립 변수들 간의 선형 관계를 모델링한다. 수많은 예측변수를 포함하는 고차원 데이터의 경우에는 전술한 일반 선형 회귀모형을 적용하였을 때 변수들 사이의 다중공선성 등의 문제가 있을 수 있으며, 과적합(overfitting) 문제를 다루기 어려울 수 있어 예측과 관련된 연구에서는 적합하지 않을 수 있다.[3)]

반면, 정규화 모형은 모델의 복잡성을 제어하고 과적합 문제를 완화하며 변수 선택과 다중공선성 감소에 효과적이라 알려져 있다. 정규화 모형은 고차원 데이터에서 일반 선형 회귀모형의 단점을 보완하기 위해 제안된 모형으로, 목적식에 추가적인 제약 조건을 추가하여 주어진 데이터를 가장 잘 설명하는 계수 추정치가 아니더라도 새로운 데이터에도 적용할 수 있는 보다 일반적인 계수 추정치를 추정한다(Hastie et al., 2009). 이때, 추가적인 제약 조건은 모형의 계수추정치의 크기에 적용되는데, 이러한 제약 조건의 형태에 따라 크게 LASSO 모형과 능형회귀(Ridge regression)로 구분된다. LASSO 모형은 예측모형의 파라미터의 크기를 L1 Norm을 이용하여 제한하는 모형이고, 능형회귀의 경우 L2 Norm을 이용하여 제한한다. 두 제약식을 혼합한 하이브리드 모형으로 Elastic Net 모형을 활용할 수도 있다.

L1 norm과 L2 norm은 둘 다 Regularization 기법으로 사용되며, 각각 다른 방식으로 모델의 계수 추정치를 축소시킨다. 우선, L1 norm은 모델의 계수를 0으로 강제로 수축시키는 특징을 갖고 있다. 이는 불필요한 변수들을 선택적으로 제거하여 모델을 간단하게 만들어준다는 의미이다. 따라서, L1 norm을 적용하면 변수 선택(feature selection)이 자동으로 이루어지는데, 중요한 변수만이 모델에 남게 된다. 이는 해석력을 높여주고, 모델의 설명력을 개선하는데 도움이 된다. 하지만, 변수 선택과 동시에 계수 추정치를 0으로 만드는

3) 모형의 과적합 이란 모델이 특정 데이터에 지나치게 적합되어 새로운 데이터에 대해 일반화하지 못하는 것을 의미한다. 즉, 기존데이터를 가장 잘 설명하는 모형은 새로운 데이터에 대한 예측에는 효과적이지 않음을 의미한다.

특성으로 인해 특정 변수들의 중요성이 과도하게 부각되는 현상이 발생할 수도 있다(Tibshirani, 1996). 반면, L2 norm은 모델의 계수를 0에 가깝게 수축시키지만, 0이 되지는 않는 모형이다. L2 norm은 계수들의 크기를 제한하여 모델의 복잡성을 감소시키고, 과적합을 방지하는데 주로 활용되며, L2 norm을 적용하면 변수들이 모두 모델에 기여하게 된다. 따라서, L2 norm은 모든 변수들을 고려하고자 할 때 유용한 방식이다(Hoerl and Kennard, 1970).

본 연구에서는 기존에 수집된 회사채 발행 정보를 바탕으로 향후 발행될 회사채 발행금리를 예측하는 모형을 개발하는 것이 주요한 연구목적이므로, 일반 선형회귀 모형이 아닌 정규화 모형을 주 분석 모형으로 삼는다. 본 연구에서는 두 정규화 방식을 결합한 Elastic Net를 활용한다. Elastic Net은 L1 norm과 L2 norm을 결합한 형태로, 두 정규화 기법의 장점을 합쳐서 사용하는 방법이다. 이를 통해 L1 norm만큼 변수 선택의 효과를 갖으면서, L2 norm처럼 모든 변수들을 고려하는 데에도 유리한 정규화 방식이다. 발행금리 스프레드를 예측하기 위한 Elastic Net 정규화 모형은 아래 식 (1)과 같이 표현할 수 있다.

$$\min_{(\beta_0\beta)} \frac{1}{2N}\sum_{N}^{i=1}(y_i - g(x_i))^2 + \lambda[(1-\alpha)\frac{\| \beta \|_2^2}{2} + \alpha \| \beta \|_1] \quad (1)$$

여기서, $g(x_i) = x_i\beta,\ \lambda \geq 0\ \text{and} 0 \leq \alpha \leq 1$

식 (1)에서 y_i는 종속변수(목표변수) 이며 i번째 채권의 발행금리 스프레드(bp)이다. x_i는 i번째 채권의 발행금리를 예측하기 위한 예측 변수 벡터이며, 다수의 예측 변수를 포함한다. $g(x_i)$는 x_i를 활용하여 발행금리 스프레드(y_i)를 예측하기 위해 설정된 선형 회귀모형을 의미한다. β는 x_i가 발행금리 스프레드(y_i)에 미치는 영향을 나타내는 계수추정치 벡터이다. α는 L1 Norm 형태의 정규화 제약의 크기를 조절하는 하이퍼파라미터로 0에서 1 사이의 값을 가진다. α가 0인 경우 식 (1)에서 제약 조건은 L2 Norm만 적용되는 것을 뜻하며, α가 1인 경우에는 L1 Norm만 제약 조건으로 적용된다는 의미이다. α가 0과 1 사이 값인 경우에는 두 형태의 제약 조건이 적절히 혼합된다는 것이며, 본 연구에서는 α을 조정하여 Elastic Net 모형을 구현하고자 한다(Zou and Hastie, 2005).

하이브리드 정규화 모형인 Elastic Net을 활용하면 L1 norm과 L2 norm의 장점을 조합하여 더욱 강력한 모델을 구축할 수 있는 이점이 있다. 특히, 고차원 데이터에서 변수 선택과 과적합 방지를 동시에 고려해야 할 때 유용하며, α를 조정하여 원하는 정규화 정도를 조절할

수 있다. 마지막으로 λ의 경우 α를 조정하여 설정된 제약 조건을 모형에 얼마나 반영할 것인지를 결정하는 조율 파라미터(tuning parameter)로 제약의 강도를 의미한다. λ가 0인 경우 제약 조건이 전혀 작용하지 못하기 때문에 식 (1)은 일반 선형 회귀모형인 OLS 모형과 동일한 목적식이 된다. 반면, λ가 커지면 계수 추정치에 대한 정규화 정도가 강해져 OLS 계수 추정치에서 점차 멀어지게 된다.

단, 분석을 위해서는 모형에 포함된 모든 변수를 표준화하는 작업이 선행되어야 한다. 왜냐하면 각 변수가 표준화되어야 동일한 기준으로 조율 파라미터의 제약을 적용 받기 때문이다. 본 연구에서는 머신러닝 알고리즘을 활용하여 새로운 데이터에 대한 예측성과를 가장 높여주는 정규화 모형과 조율 파라미터를 찾고자 한다.

2) 예측 변수 구성

데이터의 구성에 따라 모형의 예측력이 영향을 받을 수 있으므로 분석 모형을 다양한 유형으로 구성하는 것이 중요하다. 이러한 이유로 데이터셋에 따라 5가지 유형의 분석 모형을 구성하고자 한다. 구체적인 분석 데이터셋은 아래의 <표 1>과 같다.

<표 1> 분석 데이터셋 구분

본 연구에서 사용한 데이터셋의 설명이다.

구분	설명
데이터셋1 (D1)	회사채 신용등급, 발행만기, 보증채 여부, 발행금액
데이터셋2 (D2)	데이터셋 1 + 채권 및 주식시장 변수 추가
데이터셋3 (D3)	데이터셋 2 + 거시경제 변수 추가
데이터셋4 (D4)	데이터셋 3 + 기업의 재무적 상황 등 추가
데이터셋5 (D5)	데이터셋 4 + 기업의 재무적상황과 시장요인 간 교호항 추가

우선, 데이터셋 1(D1)은 회사채의 자체적인 정보만을 이용하여 분석하는 모형으로 신용등급, 만기, 보증여부, 발행금액 등의 변수를 포함한다. 다음으로 데이터셋 2(D2)는 채권시장과 함께 주식시장 전반의 변수를 추가하여 분석한다. 이를 위해 데이터셋 1(D1)에서 사용한 변수에 국고채금리(1년, 3년 5년, 10년), 회사채어음부도율, 신용스프레드(국고채 3년물-회사채BBB-), 안전자산(국채+통안채+지방채) 발행금액, 채권시장 전체 거래대금, 안전자산 거래대금, KRX 채권지수(시장가격 지수 및 총수익지수) 코스피지수, 주식시장 전체 거래대금, 코스피 변동성지수(VKOSPI) 등을 포함한다.

데이터셋 3(D3)은 데이터셋 2(D2)에 거시경제 변수를 추가한 것이다. 이를 위해 경기선행

지수, 경기동행지수, 장단기금리차(국고채 10년-국고채 1년), 산업생산지수, GDP, 소비자물가지수, 생산자물가지수, 실업률, 기준금리 등을 반영한다. 데이터셋 4(D4)는 데이터셋 3(D3)에 회사채 발행 기업의 재무 상태 및 추가적인 특성을 반영한다. 이를 위해 기업규모(총자산), 부채비율, 총자산영업이익률, 영업활동현금흐름, 최대주주 및 특수관계인 지분율, 회사채 상장여부, 발행기업 소속시장 등의 변수를 추가적으로 활용한다. 마지막으로 데이터셋 5(D5)는 데이터셋 4(D4)에 신용스프레드와 총자산 및 부채비율, 기준금리와 총자산 및 부채비율을 각각 곱하여 생성한 교호항들을 추가적으로 활용한다.[4)]

본 연구에서는 전술한 5개의 데이터셋을 각각 학습 데이터와 검증 데이터로 구분하고, 이를 기반으로 5개의 분석 모형을 학습 및 검증에 반복적으로 적용하여 최적의 발행금리 예측 모형을 개발하고자 한다. 이후, 예측 결과를 이용하여 예측용 외표본(out-of-sample prediction data)의 ESG 채권의 발행금리를 예측하고, 각 모형의 예측 성과를 비교 분석하여 어떤 요인이 발행금리 예측에 가장 영향력을 미치는지를 파악하고자 한다. 이후 분석 결과를 통해 향후 ESG 채권 발행 시 어떤 요인을 고려해야 할지에 대한 시사점을 제시하고자 한다.

3) 예측성과 평가 지표

본 연구에서 예측성과를 평가하기 위해 활용하는 지표는 2가지이다. 첫 번째는 RMSE(Root Mean Squared Error)이고, 두 번째는 MAE(Mean Absolute Error)이다. 2가지 모두 연속형 변수에 대한 예측성과를 평가하는 지표로 주로 활용되며 각각의 특성과 장단점이 있다.

$$RMSE = \sqrt{(1/N)\sum_{i}^{N}(\hat{y_i} - y_i)^2} \tag{2}$$

$$MAE = (1/N)\sum_{i}^{N}|\hat{y_i} - y_i| \tag{3}$$

RMSE는 예측 값과 실제 값 사이의 오차를 제곱하여 평균한 값에 루트를 취한 지표로 제곱을 함으로써 오차가 크게 튀는 값들에 패널티를 주는 특징이 있다. 따라서, 큰 오차가 존재할 경우 RMSE는 더 큰 값이 나오게 된다. RMSE의 장점은 큰 오차를 더 크게 반영해주기 때문에 이상치(outlier)에 민감하게 반응하고, 모델의 안정성을 평가하는데 유용하다. 하지만, 제곱을 취하므로 오차가 커질수록 평가 지표의 값도 더 커지기 때문에 오차가 작은 경우에는 상대적으로 민감하지 않을 수 있다.

4) 데이터셋 별로 포함된 변수에 결측치가 발생하는 경우가 있어 분석 표본 개수에서 차이가 존재한다.

한편, MAE는 예측 값과 실제 값 사이의 오차의 절대값의 평균을 나타내는 지표로 제곱이 없기 때문에 오차의 크기를 그대로 반영하고, 이상치에 대해 민감하지 않다. 따라서, RMSE보다 이상치에 덜 민감하게 작동하며, 큰 오차들이 평가 지표에 덜 영향을 미친다. MAE의 장점은 이상치에 덜 민감하게 반응하고, 예측 성과를 모델에 대한 실제 오차와 더 가깝게 평가할 수 있다는 점이다. 하지만, 제곱이 없기 때문에 오차가 작은 경우에도 상대적으로 값이 작아져서 모델의 민감도가 낮아질 수 있다.

RMSE와 MAE 외에 MAPE(Mean Absolute Percentage Error)를 예측 성과 지표로 활용할 수 있다. MAPE는 예측 값과 실제 값 간의 평균 백분율 차이를 측정하는 지표로, 예측 정확도를 이해하는데 직관적이라는 장점이 있으나 실제값이 0에 가까운 작은 수인 경우 미세한 오차에도 민감할 수 있다. 본 연구에서 분석 대상으로 삼은 발행금리 스프레드의 경우 0의 가까운 수치이며 0의 값을 가지는 경우도 많아 MAPE를 계산함에 있어서 예측 성과가 무한대(Infinity)로 추정되는 경우가 빈번할 수 있다. 따라서 본 연구는 RMSE와 MAE를 주된 평가 지표로 활용하고자 한다.

채권 발행금리 예측에는 RMSE와 MAE 등 2개 지표 중 어떤 것을 사용하는 것이 효과적인지는 상황과 목적에 따라 다를 수 있다. 채권 발행금리 예측에서는 예측 성과의 평가 방식을 결정할 때, 데이터의 특성, 이상치의 존재 여부, 모델의 안정성 평가 등을 고려하여 적절한 평가 지표를 선택하는 것이 중요하다. 이는 모델의 신뢰성과 예측 정확도를 높이는데 기여할 수 있다.

4) 분석 알고리즘 설계

전체 분석 데이터셋 중 회사채에 해당하는 데이터를 랜덤으로 추출하여 6:2:2의 비율로 학습데이터(Training set), 검증데이터(Validation set), 평가데이터(Test set)로 구분한다. 학습데이터는 모델의 학습에 사용되는 데이터로 모델은 학습데이터를 통해 변수들의 관계를 학습하고, 예측을 위한 파라미터를 조정한다. 검증데이터는 모델의 1차적인 성능 평가에 사용되는 데이터로 모델의 일반화 능력을 평가하며, 모델의 하이퍼파라미터를 조정하거나 모델의 형태를 결정하는데 활용된다. 검증데이터를 사용하여 모델의 예측 성과를 측정하고 최적의 모델을 선택할 수 있다. 테스트데이터는 최종적으로 선택된 모델의 성능을 평가하기 위한 데이터로 모델이 완전히 학습되지 않은 새로운 데이터를 사용하여 모델의 예측 성과를 최종적으로 평가한다. 테스트데이터는 개발된 모델의 일반화 능력을 실제 상황에 가깝게 평가하는데 사용된다.

학습데이터에서 모델을 학습한 후, 학습된 모형을 적용하여 검증데이터의 발행금리를

예측하고 예측성과를 바탕으로 모델에서 필요한 파라미터를 결정한다. 학습데이터에서 모델을 학습할 때, K-fold Cross Validation을 적용하여 모델의 파라미터인 계수추정치(beta)와 정규화 수준을 결정하는 파라미터인 λ를 결정한다. K-fold Cross Validation은 데이터를 나누어서 모델의 성능을 평가하는 간단하고 유용한 방법으로, 예측 문제에서 모델의 일반화 능력을 평가하는데 효과적으로 활용된다. 데이터를 k개의 부분집합으로 나누어서 k번의 학습과 검증을 반복한다. 이렇게 함으로써 더 많은 데이터를 활용하여 모델의 일반화 능력을 제고하고 최적의 하이퍼파라미터를 결정할 수 있다. K-fold Cross Validation을 통해 여러 번의 검증을 수행하므로 모델의 성능이 특정 검증데이터에 의존하지 않고 전체 데이터에 대해 일반화 능력을 평가할 수 있다. 이를 통해 모델의 과적합 여부를 판단할 수 있다. K번의 검증 결과를 평균하여 최종 모델의 성능을 평가함으로써 보다 안정적인 예측 성과 평가를 반영할 수 있다. 본 연구에서는 4개의 fold로 데이터를 구분하여 교차검증을 수행한다.

검증데이터에서는 모델의 형태를 결정하는 α를 0.001 간격으로 Grid Search를 통해 결정한다. 예측 성과 평가 지표로는 RMSE와 MAE를 사용하여 각 모델의 성능을 검토한다.

<표 2> 분석 알고리즘 단계별 설명

본 연구에서 수행한 발행금리 스프레드 예측을 위한 알고리즘에 대한 단계적 설명이다.

단계	설명
1. 데이터 준비	① 6,924건의 데이터를 분석표본과 예측용 외표본으로 구분 ② 분석표본 6,797을 무작위 랜덤 추출방식으로 6:2:2로 구분 (학습데이터 60%, 검증데이터 60%, 평가데이터 20%)
2. 모형 학습	① 학습데이터를 4개의 fold로 구분하여 3개 fold 로만 모형 학습(계수 추정) ② 나머지 1개 fold의 발행금리 예측 ③ 예측오차 확인 및 예측성과 지표(RMSE, MAE) 계산 ④ 학습 fold와 예측 fold를 바꿔가며 1)~3) 단계 반복하고 평균 예측성과 지표 확인 (교차검증) ⑤ 조율 파라미터(λ) 조정 후 예측성과 지표의 개선이 없을 때까지 ①~④ 과정 반복 ⑥ 최종 계수 및 조율파라미터 결정
3. 모형 선택	① 학습데이터에서 결정된 계수추정치 및 조율파라미터로 검증데이터 발행금리 예측 ② 예측 오차 및 예측성과 지표 계산 ③ 정규화 파라미터(α) 조정 후 2.-1)~3.-2) 과정 반복 ④ 정규화 파라미터(α) 별 예측 성과지표 비교하여 최적 정규화 파라미터(α) 결정
4. 모형 평가	① 최적 파라미터에 따른 모형 및 계수추정치 활용하여 평가데이터 발행금리 예측 ② 예측성과 지표 최종 확인
5. 표본 외 예측	① 최적 파라미터에 따른 모형 및 계수추정치를 활용하여 예측용 외표본(ESG 채권) 발행금리 예측 및 결과 확인

최적의 모형과 파라미터가 결정되면, 평가데이터(test set)의 발행금리를 예측하고 최종적으로 예측 성과를 확인한다. 예측 성과 평가에는 RMSE와 MAE를 비교하여 모델의 예측 성능을 평가한다.

마지막으로, 회사채 데이터를 기반으로 학습된 예측 모델을 예측용 외표본(out-of-sample prediction data)인 ESG 채권 데이터에 적용해 보고, 동일한 방법으로 발행금리를 예측 가능한지를 확인한다. 이를 통해 개발된 모델이 ESG 채권 발행금리 예측에도 적용 가능한지를 검증한다.

Ⅳ. 분석 결과

<표 3>은 본 연구에서 사용한 변수들의 설명이다. 예측변수인 *Spread*는 회사채 발행금리 스프레드로 발행금리에서 동일 만기의 국고채 수익률을 차감한 것이다. 그리고 *Credit*은 AAA부터 B까지 총 14개 등급으로 구분되며, 상위등급부터 점수를 부여한 것이다(AAA는 14점, B는 1점). 거시변수는 불안정 시계열이 존재할 수 있어 로그차분(log difference)하여 이용하며, 채권 및 주식시장 관련 변수들의 변동률 계산 또한 로그차분한 것이다. 마지막으로 기업요인에서 계산한 부채비율(LEV) 등의 재무비율이나 지분율은 채권발행 기업의 전년도 재무자료를 이용하였다.

<표 3> 변수설명

본 연구에서 사용한 변수들에 대한 설명이다.

구분	변수명	변수설명
회사채 요인	Spread	회사채 발행금리 스프레드(bp)
	Credit	신용등급(1~14점 scale)
	Amount	발행금액(백만원)
	Maturity	발행만기(년)
	Coupon	표면금리
	Guarantee	보증채는 1, 아니면 0
	ESG	ESG채권은 1, 아니면 0
거시변수	CLI	경기선행지수(전월대비)
	CI	경기동행지수(전월대비)
	INTSP	장단기금리차(국고10년-국고1년)
	IPI	산업생산지수(계절조정, 전월대비))
	GDP	GDP(동분기)
	CPI	소비자물가지수(계절조정, 전월대비)
	PPI	생산자물가지수(계절조정, 전월대비)
	UNE	실업률(계절조정, 전월대비)
	BaseRate	기준금리

<표 3> 변수설명(계속)

구분	변수명	변수설명
채권시장	KTB1	국고채금리(1년, 전월대비)
	KTB3	국고채금리(3년, 전월대비)
	KTB5	국고채금리(5년, 전월대비)
	KTB10	국고채금리(10년, 전월대비)
	Default	회사채어음부도율(전월대비)
	CreSpread	신용스프레드(국고채3년물-회사채BBB-)
	RfAmount	안전자산(국채+통안채+지방채) 발행금액(전월대비)
	BondTV	채권시장 전체 거래대금(전월대비)
	RfTV	안전자산 거래대금(전월대비)
	Index1	KRX채권지수_시장가격(전월대비)
	Index2	KRX채권지수_총수익(전월대비)
주식시장	KOSPI	코스피지수(전월대비)
	StockTV	주식시장 전체 거래대금(전월대비)
	VKOSPI	변동성지수(2003년부터 제공, 전월대비))
기업요인	Assets	ln(총자산)
	LEV	부채 / 총자산
	ROA	영업이익 / 총자산
	CFO	영업현금흐름 / 총자산
	OWN	최대주주 및 특수관계인 지분율
	Listing	상장은 1, 비상장은 0
	Market	상장기업이 코스피에 속하면1, 코스닥에 속하면 0

<표 4>는 본 연구에서 사용한 변수들의 기초통계량이다. 주요 변수의 분석결과를 보면, *Spread*의 평균은 120.39bp로 나타나 동일 만기의 국고채 대비 약 1.2% 높게 발행되고 있다. 한편, *Spread*의 최소값은 -452.60bp를 보여 국고채보다 더 낮은 수준으로 회사채가 발행되는 경우도 일부 확인된다.[5] *Credit*의 평균은 9.80으로 회사채 발행기업의 신용등급은 평균적으로 약 A로 나타나 양호한 수준이며, *ESG*의 평균은 0.02로 전체 표본의 약 2% 정도가 ESG 채권에 해당된다.

<표 4> 기초통계량

본 연구에서 사용한 변수들의 기초통계량이다. 회사채 요인과 관련된 변수는 $Spread \sim ESG$ 이며, 거시경제적 요인을 반영하는 변수는 $CLI \sim BaseRate$이다. 채권시장과 관련된 변수는 $KTB1 \sim Index2$이고, 주식시장에 관한 변수는 $KOSPI \sim VKOSPI$ 이며, 기업 특성과 관련된 변수는 $Assets \sim Market$이다.

Variable	Obs	Mean	Std. dev.	Min	Max
Spread	6,924	120.39	137.01	-452.60	814.40
Credit	6,924	9.80	2.56	1.00	14.00

5) 글로벌 금융위기 이전 발행된 일부 최우량등급 회사채에서 표면금리가 동일 만기의 국고채보다 더 낮게 발행된 사례가 존재한다.

<표 4> 기초통계량(계속)

Variable	Obs	Mean	Std. dev.	Min	Max
Amount	6,924	88479.71	75169.23	4.00	912291.00
Maturity	6,924	3.62	2.15	0.25	30.00
Coupon	6,924	4.46	2.04	1.00	13.00
Guarantee	6,924	0.04	0.19	0.00	1.00
ESG	6,924	0.02	0.13	0.00	1.00
CLI	6,924	0.00	0.00	-0.01	0.02
CI	6,924	0.00	0.00	-0.01	0.01
INTSP	6,924	0.86	0.61	-0.16	2.88
IPI	6,924	0.00	0.01	-0.04	0.05
GDP	6,924	3.47	2.22	-2.50	8.20
CPI	6,924	0.00	0.00	-0.01	0.01
PPI	6,924	0.00	0.01	-0.02	0.02
UNE	6,924	-0.01	0.21	-0.80	1.10
BaseRate	6,924	2.58	1.28	0.50	5.25
KTB1	6,924	-0.02	0.19	-1.27	0.55
KTB3	6,924	-0.01	0.21	-1.00	0.75
KTB5	6,924	-0.01	0.22	-0.87	0.84
KTB10	6,924	-0.01	0.21	-0.86	0.82
Default	6,924	0.00	0.07	-0.26	0.33
CreSpread	6,924	6.03	1.25	3.10	8.62
RfAmount	6,924	-0.07	0.87	-2.06	2.03
BondTV	6,924	0.00	0.12	-0.34	0.50
RfTV	6,924	0.00	0.14	-0.39	0.50
Index1	5,914	0.00	0.01	-0.02	0.03
Index2	5,914	0.00	0.01	-0.02	0.04
KOSPI	6,924	0.01	0.06	-0.26	0.20
StockTV	6,924	0.01	0.17	-0.37	0.69
VKOSPI	6,250	0.00	0.18	-0.46	0.79
Assets	4,875	21.93	1.30	15.88	25.08
LEV	4,865	0.56	0.16	0.04	0.98
ROA	4,875	0.05	0.05	-0.29	0.41
CFO	4,875	0.06	0.07	-0.36	0.41
OWN	4,875	0.36	0.18	0.00	0.97
Listing	6,924	0.66	0.47	0.00	1.00
Market	4,875	0.96	0.18	0.00	1.00

1. 회사채 발행금리 예측 결과

본 절에서는 회사채 발행금리 예측을 위한 기계학습의 결과와 이를 통해 도출된 최적 모형의 예측 성과를 제시하고자 한다. 본 연구에서는 <표 1>에 제시된 5개의 데이터셋(D1~D5)에 대해 발행금리 예측을 위한 머신러닝 분석을 수행하였다. 학습데이터에서

벤치마크 모형인 OLS 모형(λ=0)과 α=0.000~α=1.000까지 총 1,002개 모형을 이용하여 예측 모형을 추정하고, 이를 검증 데이터에 적용하여 예측 성과를 비교하였다. <표 5>는 RMSE 기준으로 예측 성과를 비교한 결과이다.

<표 5> 모형 선택 결과 및 최적 모형의 예측 성과(RMSE 기준)

Best model은 1,001개의 머신러닝 모형과 OLS 모형의 검증데이터 예측 성과 중 가장 작은 값을 가지는 경우를 의미한다.

RMSE	D1	D2	D3	D4	D5
OLS	121.46	92.14	87.12	89.55	83.01
Avg. of ML	121.47	91.96	87.02	89.62	82.73
Worst model	α=0.001	α=0.001	α=0	α=0.806	α=0
Optimal λ	6.27	7.22	7.25	1.58	7.67
Best Model	α=1	α=0.997	α=0.975	α=0.002	α=0.878
Optimal λ	1.57	0.58	0.89	3.72	0.50
Test Set (Best Model)	112.91	97.21	86.27	82.53	87.41
Obs. of Training Set	4078	3472	3472	2397	2397
Obs. of Validation Set	1360	1158	1158	799	799
Obs. of Test set	1359	1157	1157	799	799

총 1,002개의 예측 모형의 예측 성과(RMSE)를 비교하여, 가장 낮은 예측 성과를 보이는 모형(α) 및 최적 조율 파라미터 값(λ)과 가장 높은 예측 성과를 보이는 모형(α)과 최적 조율 파라미터 값(λ)을 도출하였다. D1의 경우 최적 모형이 LASSO(α=1)인 것으로 확인되고, 그 외 데이터셋에서는 LASSO와 Ridge Regression을 적절히 결합하는 Elastic Net 모형이 최적 모형인 것으로 확인된다. 각 데이터셋에서의 최적 모형 하에서 검증 데이터의 예측 성과를 비교해 보면, 많은 표본에도 불구하고 적은 변수를 활용한 경우 예측성과가 낮은 것으로 나타난다.

이 중 D4의 경우 α=0.002인 경우 예측 성과가 높게 가장 나타나며, 이는 L1 Norm보다는 L2 Norm으로 정규화 하는 경우의 예측력이 더 높음을 시사한다. 즉, D4에 포함된 예측변수를 최대한 많이 모형에 포함하는 것이 예측 성과를 높이는데 도움이 된다는 것으로 해석해 볼 수 있다. 이때 최적 조율 파라미터 λ는 3.72로 비교적 높게 추정된다. 이를 통해 D4의 경우 능형회귀를 통해 계수를 제한하되 조율 파라미터 값을 높여 제약 정도를 강하게 가져가는 것이 예측 성과를 높일 수 있음을 파악해 볼 수 있다. 가장 많은 변수가 포함된 D5의 경우 α=0.878로 두 정규화 방식을 D4 보다 고루 반영하는 모형이 가장 최적인 것으로 확인되고, 이때의 조율 파라미터는 0.5로 비교적 작게 추정된다.

검증데이터의 예측 성과를 가장 높여주는 최적 모형을 이용하여 평가데이터를 예측한 결과 D4에서 가장 좋은 예측 성과가 확인된다. D5에는 D4보다 많은 예측변수를 포함하였으나, 예측 성과 개선에 크게 기여하지 못한 것으로 나타난다. 이는 제한된 표본 크기에서 무조건적인 예측변수 셋의 확장은 예측성과 개선에 크게 기여하지 않음을 시사한다. 해당 결과는 또한 예측 변수 후보가 많이 포함될수록 학습 모형의 예측력이 좋아질 수 있으나, 그 과정에서 표본의 수가 많이 탈락되는 경우에는 분석데이터에 예측변수가 추가되는 것이 예측 성과에 도움이 되지 않을 가능성도 존재할 수 있다고 해석가능하다. 이를 통해 머신러닝을 이용한 예측 모형 설계 시 고차원 데이터 확보뿐만 아니라 데이터의 표본 수 확보도 중요한 것으로 추론 가능하다.

다음으로 <표 6>에서는 <표 5>와 동일한 분석을 예측 성과 지표를 MAE 기준으로 바꾸어 수행한 결과이다. 전술한 바와 같이, MAE 기준으로 예측 성과를 비교하는 경우에는 오차의 정도에 따라 가중치를 부여하지 않기 때문에 RMSE를 기준으로 결정한 최적 모형 및 최적 파라미터와 상이하게 나타날 수 있다.

<표 6>의 분석결과를 보면, D1부터 D5까지 모든 데이터셋에서 Elastic Net이 최적 모형으로 선택되었음을 알 수 있다. 예측 변수가 가장 적은 D1의 경우 α=0.906인 모형이 가장 최적인 것으로 추정되고, 이는 정규화 형태가 능형회귀 보다는 LASSO에 가깝다는 것을 의미한다. 다만, 조율 파라미터 λ가 상대적으로 낮게 추정되어 변수 선택 기능이 크게 작용하지는 않았다. D2의 경우 α=0.083일 때가 가장 최적인 것으로 확인되는데, 이는 LASSO 보다는 능형회귀와 가까운 정규화 형태일 때 가장 높은 예측 성과를 보인다는 것을 의미한다. 즉 모든 변수를 적절히 정규화하여 모형에 반영할 때 예측성과가 가장 높을 수 있음을 시사한다. 이때, 조율 파라미터 λ가 상대적으로 높게 추정되었다는 점에서 계수에 대한 제약이 크게 발생하였을 것으로 판단된다. D3와 D4, D5의 경우 α=0.986, α=0.938, α=0.857가 최적 모형인 것으로 추정되어, Elastic Net 중 능형회귀와 유사한 특성을 보이는 모형이 최적 모형인 것으로 확인된다. 다만, MAE 기준의 예측 성과 지표로 분석했을 경우 α값이 1에 가깝더라도 RMSE 기준 분석결과에 비해 조율 파라미터 λ가 높게 추정되고 있음을 알 수 있다. 이는 MAE 기준 분석에서 최적 정규화 파라미터가 더 크게 추정되어 RMSE 기준 모형 보다 많은 변수를 모형에 포함하고 있기는 하나 정규화 수준을 강화하면서 모형 예측력을 높이기 위한 노력을 하고 있음을 의미한다. MAE 기준 분석에서도 RMSE 분석과 유사하게 D4를 활용한 예측 성과가 가장 높음을 알 수 있다.

전체적으로 볼 때, 분석 데이터를 어떻게 구성하고 어떤 예측 성과 지표를 활용하느냐에 따라 주어진 데이터에 가장 적합한 예측 모형이 달라질 수 있다. 이와 같은 차이는 각각의

데이터 및 예측 성과 지표별로 예측 모형에 포함된 변수 선택 결과로도 나타난다.

<표 6> 모형 선택 결과 및 최적 모형의 예측 성과 (MAE 기준)

Best model은 1,001개의 머신러닝 모형과 OLS 모형의 검증데이터 예측 성과 중 가장 작은 값을 가지는 경우를 의미한다.

MAE		D1	D2	D3	D4	D5
OLS		82.88	65.57	59.34	61.51	58.03
Avg. of ML		82.91	65.25	58.61	60.21	58.09
Worst model		α=0	ols	ols	ols	α=0.004
	Optimal λ	6.82	0.00	0.00	0.00	1.94
Best Model		α=0.906	α=0.083	α=0.986	α=0.938	α=0.857
	Optimal λ	2.17	6.37	2.80	0.29	3.47
Test Set (Best Model)		80.61	69.13	58.91	55.37	56.07
Obs. of Training Set		4078	3472	3472	2397	2397
Obs. of Validation Set		1360	1158	1158	799	799
Obs. of Test set		1359	1157	1157	799	799

<표 7>은 각각의 모형 별 선택되지 않은 변수에 대한 결과이다. D1에서는 절편을 제외한 총 4개 변수를 활용하였고, D2에서는 18개, D3에서는 27개, D4에서는 34개 변수를 예측변수의 후보군으로 포함하였다. 마지막으로 D5는 예측변수 중 신용스프레드와 기업의 총자산 및 부채비율 간의 교호항, 기준금리와 기업의 총자산 및 부채비율간의 교호항 등 총 4개 변수를 추가하여 38개 변수를 예측 후보군으로 포함하였다.

<표 7> 모형별 미선택 변수

해당 표는 RMSE와 MAE를 기준으로 각각의 데이터셋별 선택되지 않은 변수이다.

모형		미선택 변수
RMSE	D1	None
	D2	KTB5, RfAmount, BondTV, StockTV
	D3	KTB3, KTB5, KTB10, RfAmount, RfTV, KOSPI
	D4	None
	D5	CPI, UNE, KTB5, KTB10, RfAmount, RfTV, (BaseRate X Asset)
MAE	D1	None
	D2	KTB3, BondTV, StockTV
	D3	CI, IPI, CPI, PPI, UNE, KTB3, KTB5, KTB10, Default, RfAmount, BondTV, RfTV, Index1, KOSPI, VKOSPI
	D4	CLI, IPI, CPI, PPI, UNE, KTB3, KTB5, KTB10, Default, RfAmount, BondTV, RfTV, Index1, Index2, KOSPI, StockTV, VKOSPI, CFO, OWN, Market, Listing
	D5	IPI, CPI, PPI, UNE, KTB5, KTB10, Default, RfAmount, BondTV, RfTV, Index1, KOSPI, StockTV, Asset, LEV, CFO, Market, Listing, (BaseRate X LEV)

RMSE를 기준으로 분석한 결과, 가장 낮은 예측 성과를 보였던 D1과 가장 높은 예측 성과를 보였던 D4 데이터셋을 이용하여 예측 모형을 분석한 경우 탈락되는 변수 없이 모든 변수가 모형에 포함하였다. 이를 통해 예측 변수의 셋이 어떻게 구성되느냐에 따라 예측 기여도가 높은 변수는 다르게 선택될 수 있음을 확인할 수 있다. D5에서 D4에 비해 4개 변수가 더 추가되었는데, 이러한 변수가 데이터셋에 추가됨으로 인해서 다른 변수들까지 모형에서 선택되지 않게 되었다. 이는 예측변수 간 상관관계를 고려하여 변수를 선택하는 선형 기반의 Elastic Net 모형의 특징에 기인한 것으로 판단된다. 반면, MAE 기준으로 예측 성과를 평가한 경우에는 RMSE 기준 보다 더 많은 변수를 탈락시킨 것으로 확인된다. D2에서 3개, D3에서 총 15개 변수, D4에서 21개 변수, D5에서 19개를 모형에서 제외시켰다는 점에서 RMSE 기준보다 훨씬 적은 예측 변수를 활용하였음을 알 수 있다.

2. 변수별 중요도 평가 결과

본 절에서는 예측 모형별로 예측 변수가 회사채 발행금리 예측에 어느 정도 기여를 했는지를 분석한다. 예측 변수의 중요도는 표준화 계수의 절대값을 통해 산정하였다.

우선, <표 8>은 RMSE 기준으로 선정된 각각의 데이터셋별 최적 예측 모형에서의 예측 변수와 중요도를 중요도 순으로 정렬한 결과이다. 중요도가 0의 값을 갖는 경우는 해당 모형에서 그 변수가 선택되지 않았음을 의미한다. 즉, 발행금리 예측에 어떠한 기여도 없었음을 의미한다.

<표 8>에서 확인되는 바와 같이 D1과 D4를 제외하고는 대부분의 모형에서 일부 변수를 예측 모형에 포함하지 않았다. 분석 결과를 보면, 회사채 요인 중에서는 모든 모형에서 신용등급(Credit)이 가장 중요한 예측 변수인 것으로 확인된다. 다음으로 채권 및 주식시장 요인 변수 중에서는 신용스프레드(CreSpread), KRX 채권지수(총수익)(Index2), 국고채 1년물 금리(KTB1), KRX 채권지수(시장가격)(Index1) 등의 중요도가 높게 나타난다.

그리고 거시변수가 추가되는 경우에도 신용스프레드(CreSpread)는 예측 중요도가 높은 변수로 확인된다. 다음으로 장단기금리차(INTSP), 기준금리(BaseRate) 및 분기별 GDP(GDP)와 같은 거시요인 변수의 예측 기여도가 높은 것으로 나타난다.

D4에서 회사채 발행기업의 재무적 특성과 같은 기업요인을 추가하는 경우에도 중요도 상위 변수는 신용등급(Credit), 신용스프레드(CreSpread), 장단기금리차(INTSP), 기준금리(BaseRate), 분기별 GDP(GDP), 국고채 1년물 금리(KTB1) 순으로 나타난다. 따라서 D3 모형과 유사한 결과를 보였고, 기업특성 요인은 상대적으로 낮은 중요도를 보였다. 기업요인 중에서는 총자산(Assets)의 중요도가 가장 높으며, 다음으로 부채비율(LEV),

<표 8> 모형별 예측변수 중요도(RMSE 기준)

중요도는 표준화 계수의 절대값을 이용하여 추정하였다. 모형에서 예측에 중요하지 않은 것으로 판단되어 중요도가 0 혹은 0에 근사한 값으로 추정된 변수의 경우 음영표기하였다. 절편의 표준화 계수는 0이므로, 중요도 또한 0으로 나타난다. 절편 아래에 위치한 변수들은 모형에서 선택되지 않은 것을 의미한다.

D1		D2		D3		D4		D5	
변수	중요도	변수	중요도	변수	중요도	변수	중요도	변수	중요도
Credit	0.439	Credit	0.604	Credit	0.578	Credit	0.570	Credit	0.597
Guarantee	0.131	CreSpread	0.359	CreSpread	0.344	CreSpread	0.306	CreSpread X LEV	0.430
Maturity	0.112	Index2	0.199	INTSP	0.260	INTSP	0.289	LEV	0.360
Amount	0.008	KTB1	0.168	BaseRate	0.260	BaseRate	0.225	INTSP	0.287
절편	0	Guarantee	0.158	GDP	0.119	GDP	0.168	GDP	0.181
		Index1	0.151	KTB1	0.118	KTB1	0.129	BaseRate X LEV	0.179
		KTB10	0.136	Guarantee	0.107	Assets	0.118	Assets	0.121
		Maturity	0.074	Amount	0.038	Guarantee	0.117	KTB1	0.110
		KOSPI	0.049	Index1	0.037	Amount	0.074	Guarantee	0.109
		Amount	0.033	CI	0.033	LEV	0.073	BaseRate	0.094
		RfTV	0.022	Index2	0.032	ROA	0.049	Amount	0.084
		Default	0.018	StockTV	0.020	CLI	0.045	CreSpread	0.052
		KTB3	0.011	BondTV	0.019	Index2	0.040	ROA	0.050
		VKOSPI	0.010	Default	0.011	Index1	0.040	Maturity	0.038
		절편	0	Maturity	0.009	Maturity	0.039	CLI	0.036
		KTB5	0	CLI	0.005	BondTV	0.033	Index2	0.033
		RfAmount	0	CPI	0.005	RfTV	0.032	VKOSPI	0.019
		BondTV	0	PPI	0.005	OWN	0.024	Listing	0.017
		StockTV	0	IPI	0.003	CI	0.022	Index1	0.011
				UNE	0.003	IPI	0.021	KTB3	0.009
				VKOSPI	0.001	CPI	0.016	IPI	0.008
				절편	0	Default	0.013	PPI	0.007
				KTB3	0	KTB10	0.012	CredSpread X Asset	0.005
				KTB5	0	KTB5	0.012	BondTV	0.004
				KTB10	0	VKOSPI	0.011	Market	0.004
				RfAmount	0	PPI	0.009	OWN	0.003
				RfTV	0	Market	0.009	KOSPI	0.003
				KOSPI	0	KTB3	0.007	CI	0.002
						KOSPI	0.006	StockTV	0.002
						RfAmount	0.006	Default	0.001
						StockTV	0.003	CFO	0.000
						Listing	0.003	절편	0
						UNE	0.002	CPI	0
						CFO	0.000	UNE	0
						절편	0	KTB5	0
								KTB10	0
								RfAmount	0
								RfTV	0
								BaseRate X Asset	0

총자산수익률(ROA) 등이 중요한 것으로 나타난다. 다만, 기업 요인의 중요도는 채권 고유 요인(보증채 여부, 발행금액, 만기 등)에 비해서 월등히 높은 중요도를 가진다고 보기에는 한계가 있다.

마지막으로 시장 상황과 기업의 특수한 상황의 교호항을 추가한 D5의 경우 기존의 신용스프레드(CreSpread)와 기준금리(BaseRate)의 중요도는 상대적으로 줄어들고, 신용스프레드와 기업의 부채비율의 교호항(CreSpread X LEV)과 기준금리와 기업의 부채비율의 교호항(CreSpread X LEV)의 중요도가 높게 나타난다. 또한, 기업의 부채비율(LEV)도 다른 데이터셋을 분석했을 때에 비해 높은 예측 기여도를 보여준다. 이를 통해 회사채 발행금리 예측에 신용스프레드와 기준금리와 같은 시장 상황 및 거시경제적 상황이 중요한 역할을 하지만 이는 회사채 발행 기업의 부채비율과 밀접한 관계가 있음을 시사한다.

<표 9>는 MAE 기준으로 선정한 최적 예측 모형에서 각 변수의 중요도를 보여준다.

중요도 분석 결과를 보면, 모든 모형에서 신용등급(Credit)이 가장 중요한 예측 변수인 것으로 확인된다. 다음으로 채권 및 주식시장 요인 변수 중에서는 신용스프레드(CreSpread), 국고채 1년물 금리(KTB1), KRX 채권지수(총수익, 시장가격)(Index2, Index1) 등의 중요도가 높게 나타난다.

거시변수가 추가되는 경우, 신용스프레드(CreSpread) 다음으로 장단기금리차(INTSP)가 중요한 예측변수로 나타난다. 또한, 채권 및 주식시장 요인에 앞서 기준금리(BaseRate) 및 분기별 GDP(GDP)와 같은 거시요인의 예측 기여도가 더 높은 것으로 확인된다. 기업요인을 추가하였을 경우 중요도 상위 변수에는 크게 변화가 없으나, 기업 요인 중에서는 부채비율(LEV), 총자산수익률(ROA), 총자산(Assets) 순으로 예측성과에 기여하고 있음이 확인된다.

마지막으로 시장 관련 변수와 기업 특성 변수의 교호항을 추가한 경우, 앞서 RMSE 기준 분석과 동일하게 신용스프레드(CreSpread)와 기준금리(BaseRate)의 중요도는 상대적으로 낮아지고 교호항이 예측 기여도가 높은 것으로 나타난다. 특히 MAE 기준에서는 기업의 총자산(Asset)과 신용스프레드(CreSpread) 및 기준금리(BaseRate) 간의 교호항 그리고 기업의 부채비율(LEV)와 신용스프레드(CreSpread) 간의 교호항이 예측 성과를 높이는 데 많은 기여를 하고 있는 것으로 확인된다.

앞서 설명한 바와 같이, MAE 기준 모형의 경우 D3, D4, D5에서 상당히 많은 변수를 발행금리 예측에 중요치 않다고 판단하여 모형에서 제외하였다. 제외된 변수는 주식 및 채권시장과 거시경제 관련 변수가 많았다. 이를 통해 채권과 기업의 특성, 그리고 시장 상황과 기업의 특성 간의 상호작용 등을 고려하여 모형을 구성하는 경우 기타 시장 변수는 기업의 회사채 발행금리에 큰 영향을 주지 못함이 확인된다.

<표 9> 모형별 예측변수 중요도(MAE 기준)

중요도는 표준화 계수의 절대값을 이용하여 추정하였다. 모형에서 예측에 중요하지 않은 것으로 판단되어 중요도가 0 혹은 0에 근사한 값으로 추정된 변수의 경우 음영표기하였다. 절편의 표준화계수는 0이므로, 중요도 또한 0으로 나타난다. 절편 아래에 위치한 변수들은 모형에서 선택되지 않았다.

D1		D2		D3		D4		D5	
변수	중요도	변수	중요도	변수	중요도	변수	중요도	변수	중요도
Credit	0.465	Credit	0.555	Credit	0.571	Credit	0.528	Credit	0.560
Maturity	0.106	CreSpread	0.332	CreSpread	0.279	CreSpread	0.307	INTSP	0.243
Guarantee	0.096	KTB1	0.208	INTSP	0.255	INTSP	0.257	CredSpread X Asset	0.190
Amount	0.022	Guarantee	0.152	BaseRate	0.203	BaseRate	0.220	BaseRate X Asset	0.139
절편	0	Index2	0.141	GDP	0.139	KTB1	0.136	GDP	0.128
		Index1	0.119	Guarantee	0.106	GDP	0.132	CreSpread X LEV	0.102
		Maturity	0.085	KTB1	0.095	Guarantee	0.069	KTB1	0.093
		KTB10	0.081	Amount	0.022	LEV	0.058	Guarantee	0.082
		Amount	0.067	Maturity	0.018	ROA	0.057	CreSpread	0.080
		KOSPI	0.049	CLI	0.004	Maturity	0.046	BaseRate	0.063
		KTB5	0.023	Index2	0.000	Amount	0.042	ROA	0.049
		RfTV	0.023	StockTV	0.000	Asset	0.027	Amount	0.045
		Default	0.021	절편	0	CI	0.015	Maturity	0.029
		RfAmount	0.008	CI	0	절편	0	VKOSPI	0.018
		VKOSPI	0.004	IPI	0	CLI	0	Index2	0.015
		절편	0	CPI	0	IPI	0	KTB3	0.009
		KTB3	0	PPI	0	CPI	0	CI	0.004
		BondTV	0	UNE	0	PPI	0	OWN	0.004
		StockTV	0	KTB3	0	UNE	0	CLI	0.000
				KTB5	0	KTB3	0	절편	0
				KTB10	0	KTB5	0	IPI	0
				Default	0	KTB10	0	CPI	0
				RfAmount	0	Default	0	PPI	0
				BondTV	0	RfAmount	0	UNE	0
				RfTV	0	BondTV	0	KTB5	0
				Index1	0	RfTV	0	KTB10	0
				KOSPI	0	Index1	0	Default	0
				VKOSPI	0	Index2	0	RfAmount	0
						KOSPI	0	BondTV	0
						StockTV	0	RfTV	0
						VKOSPI	0	Index1	0
						CFO	0	KOSPI	0
						OWN	0	StockTV	0
						Market	0	Asset	0
						Listing	0	LEV	0
								CFO	0
								Market	0
								Listing	0
								BaseRate X LEV	0

3. ESG 채권 발행금리 예측 결과

본 절에서는 2절에서 분석한 회사채 발행금리 예측 모형을 이용하여 예측용 외표본인 2019년 9월 이후 발행된 127건의 ESG 채권 데이터에 적용하여 분석한다. 즉, 본 절에서는 일반적인 회사채 발행금리 데이터를 토대로 만들어진 예측 모형을 바탕으로 ESG 채권의 발행금리를 예측하는 경우, 예측예 유용성이 존재한 수 있는지에 대해 추가적으로 분석한다.

이를 위해 분석에 활용된 5개의 데이터셋과 2개의 예측 지표에 대해 총 10개 모형의 127개 ESG 채권 발행금리에 대한 예측 결과를 비교한다. 다만, D4와 D5의 경우 결측치가 있는 변수가 존재함에 따라 총 84건의 ESG 채권 데이터에 대한 예측치만을 추정할 수 있었으며, <표 10>은 모형별 ESG 채권 발행금리 예측 결과이다. 각 데이터셋과 예측성과 지표별 최적 모형(Best Model)과 조율 파라미터(λ)를 통해 예측 변수와 그 계수를 추정하고, 이를 통해 ESG 채권의 발행금리를 예측하였다.

RMSE 기준으로 볼 때, 이전의 평가데이터 분석결과와는 상이하게 D3 데이터셋 기반의 예측 모형이 예측 성과가 가장 높은 것으로 나타난다. MAE 기준의 경우에도 동일하게 D3 데이터셋 기반의 예측 모형의 성과가 가장 높은 것이 확인된다.

<표 10> 모형별 ESG 채권 발행금리 예측 성과

OOP Test Set은 out-of-sample prediction test set을 의미한다.

		D1	D2	D3	D4	D5
RMSE	Best Model	α=1	α=0.997	α=0.975	α=0.002	α=0.878
	Optimal λ	1.57	0.58	0.89	3.72	0.50
	OOP Test Set Performance	36.89	59.43	42.78	46.92	47.00
	Obs. of OOP Test Set	127	127	127	84	84
MAE	Best Model	α=0.906	α=0.083	α=0.986	α=0.938	α=0.857
	Optimal λ	2.17	6.37	2.80	0.29	3.47
	OOP Test Set Performance	28.37	48.32	34.49	35.88	38.68
	Obs. of OOP Test Set	127	127	127	84	84

데이터셋과 예측 성과 지표가 달라짐에 따라 ESG 채권 발행금리 스프레드 예측 결과가 달라진다. 따라서 어떤 방식이 가장 적합한 것인가에 대한 명확한 결과를 도출하는데 한계가 있다. 이에 [그림 1]과 같이 가장 예측 성과가 높은 2개의 데이터셋 D3와 D4에 대하여 각 예측 성과 지표에 따른 발행금리 스프레드 예측 결과를 시각화하여 살펴보았다. 2개의 데이터셋에서 모두 예측된 발행금리 스프레드 값이 있는 데이터에 한정하여 그래프로

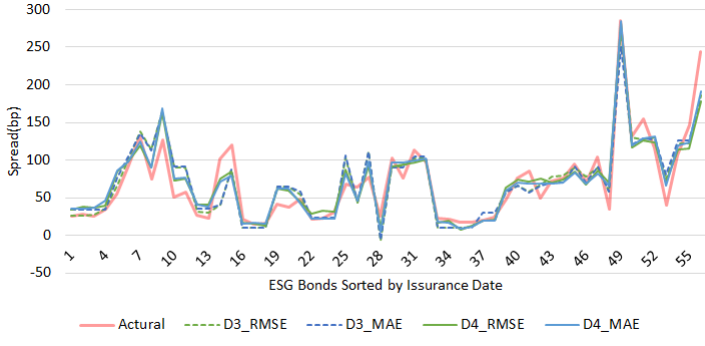
나타내었기 때문에 [그림 1]에서는 총 84개의 ESG 채권에 대한 실제 발행금리와 4개의 예측 값만 확인할 수 있다.

[그림 1]을 보면, 4개 모형을 통해 추정된 예측값이 크게 상이하지 않고 대부분 유사하게 나타남을 확인해 볼 수 있다. 이를 통해 본 연구에서 일반 회사채를 기반으로 학습한 채권 발행금리 예측 모형이 ESG 채권 발행금리 예측에도 충분히 적용될 수 있음을 추론해 볼 수 있다. 그러나, 일부 극단적으로 높거나 낮은 발행금리 스프레드를 보이는 ESG 채권의 경우 예측치 간의 차이가 발생하는 것으로 나타난다.

ESG 채권의 발행금리가 평균적인 트렌드에 비해 큰 차이가 발생하는 경우에는 MAE 기준([그림 1]의 점선 참조)을 적용하기 보다는 RMSE 기준([그림 1]의 실선 참조)으로 모형을 선택하고 예측하는 것이 실제 값과의 차이를 더 줄이는 것으로 나타난다. 이를 통해 변동 폭이 클 것으로 예상되는 지표를 예측하는 경우에는 RMSE를 기준으로 모형을 설정하는 것이 더욱 바람직할 수 있음을 추론해 볼 수 있다. 특히, 채권 발행금리와 같이 금융시장에서의 변동성이 크며 미세한 오차가 투자자와 발행자에게 큰 손실을 야기할 수 있는 경우에는 오차의 크기가 클수록 더 많은 패널티가 적용되는 RMSE 기준으로 예측 모형을 설정하는 것이 바람직할 수 있음을 본 연구의 분석결과로 확인해 볼 수 있다.

[그림 1] 모형 별 ESG 채권 발행 금리 예측 성과 비교

Actual은 ESG 채권의 실제 발행금리 스프레드(bp)이며, 그 외 각각의 선 그래프는 모형별로 추정된 ESG 채권의 발행금리 스프레드(bp) 예측값이다. 가로축은 84개 ESG 채권을 발행년월 순으로 정렬하여 붙인 채권의 번호를 의미한다.



V. 결 론

최근 재무·금융 분야의 다양한 연구에서 기업의 주가 예측이나 부도 예측 등에 기계학습을 알고리즘을 이용하여 포괄적인 분석이 이루어지고 있으며, 예측에 있어 전통적인 회귀모형

보다 머신러닝 기법이 더욱 유용하다는 결과가 제시되고 있다. 이에 본 연구는 국내의 채권시장을 대상으로 전통적인 선형회귀모형과 머신러닝 알고리즘을 이용하여 회사채 발행금리 예측력을 상호 비교・분석하였다.

분석을 위한 머신러닝 알고리즘은 선형기반의 LASSO, Ridge 및 Elastic net의 3가지 기법을 활용하였다. 또한, 회사채 발행금리 예측을 위한 변수는 회사채 발행 특성, 거시경제적 요인, 채권 및 주식시장 변수, 채권 발행기업의 재무정보 등을 종합적으로 고려하였다.

본 연구의 주요 분석결과는 다음과 같다. 첫째, 데이터셋의 구성에 따라 회사채 발행금리 예측 모델의 성능이 달라짐을 확인하였다. 특히, 회사채 발행금리 예측에 있어 발행 특성 이외에도 채권 및 주식시장, 거시경제 상황, 기업 재무 변수 등 다양한 변수들의 정보가 중요하게 작용하며, 전통적인 선형회귀모형보다는 머신러닝 알고리즘을 활용하는 것이 예측 성과를 향상시킬 수 있음을 실증적으로 확인하였다. 둘째, 최적의 예측 모형은 모델의 형태나 변수의 개수 그리고 표본의 크기 등에 따라 다양하게 나타남을 알 수 있었다. 특히, 신용등급과 신용스프레드, 장단기금리차 등은 공통적으로 회사채 발행금리 예측에 유의미한 변수로 확인되나, 일부 변수는 모델에 따라서 예측 중요도가 낮거나 유의미하지 않은 것으로 나타났다. 또한, 신용스프레드 기준금리 등과 같이 예측에 중요하게 고려되던 시장 변수는 기업의 총자산 혹은 부채비율과 같이 기업의 현 시점의 재무상태를 나타낼 수 있는 변수와 상호작용하는 경우 예측 기여도가 더 높아지는 것으로 확인되었다. 셋째, 회사채 발행금리 데이터 기반의 모델의 경우 ESG 채권의 발행금리 예측에도 유용하게 사용될 수 있음이 확인되었다. 또한, 신규 채권에 대한 발행금리 예측에는 RMSE 기준으로 모델을 선택하는 것이 MAE 기준보다 더 적절한 것으로 나타났다. 다만, 데이터 수의 부족으로 인해 예측 지표에 따른 예측 성과의 차이를 통계적 검증 모형으로 분석하지 못하였다. 향후 표본 수가 충분히 확보된다면 RMSE, MAE 등 예측 성과 지표 간 예측력 차이를 통계적 유의성을 고려하여 설명할 수 있는 알고리즘을 개발하고 연구를 확장할 가능성이 존재함을 확인하였다.

전체적으로 회사채 발행금리 예측 모형에 있어 데이터 기반의 모델 선택과 최적 파라미터 설정이 중요한 사항인 것으로 판단된다. 한편, 본 연구에서 확인되는 바와 같이 데이터, 모형, 예측 성과 지표 등 예측 알고리즘 개발에 필수적인 요소를 바꿔가며 예측 성과를 비교한 결과 회사채 발행금리 예측을 위한 최적 모형은 다양한 데이터셋 구성과 모형 선택에 따라 달라지는 것이 확인된다. 따라서 회사채 발행금리 예측에 있어 데이터의 종류와 구성 그리고 예측 모형의 유형을 신중하게 선택하는 중요한 사항인 것으로 추론해 볼 수 있다.

추가적으로 본 연구의 결과는 머신러닝 알고리즘을 활용한 모형이 전통적인 선형회귀모형보다 회사채 발행금리 예측 성과가 우수함을 실제 데이터를 통해 확인하였다. 이는

기존의 예측 모델에 새로운 시각을 제공할 수 있으며, 재무금융 분야에서 머신러닝 기법이 유용하게 활용될 수 있음을 시사한다. 즉, 머신러닝 기법을 이용한 추정 모형은 투자자 측면에서 적절한 채권 발행금리를 예측하고 합리적인 투자의사 결정을 내리는 데에 유용한 정보를 제공할 수 있을 것으로 사료된다. 또한, 회사채 발행기업 입장에서도 타당한 자본조달비용을 결정하는 참고가 될 수 있을 것으로 기대된다.

다만, 본 연구는 선형모형을 중심으로 예측 모형을 설정하고 있어 변수 간 비선형 관계에 대해서 제한적인 수준에서 반영할 수밖에 없었다는 한계점이 존재한다. 추후 회사채 및 ESG 채권에 대한 추가적인 자료가 확보되고 비선형 관계를 고려한 예측 모형 추정에 충분한 표본 수가 확보된다면, 다양한 머신러닝 알고리즘을 적용하고 비교하는 연구를 진행할 수 있을 것이라 기대한다.

참 고 문 헌

권혁건, 이동규, 신민수, “RNN(Recurrent Neural Network)을 이용한 기업부도예측모형에서 회계정보의 동적 변화 연구”, 지능정보연구, 제23권 제3호, 2017, 139-153.

김경목, 김선웅, 최흥식, “투자자별 거래정보와 머신러닝을 활용한 투자전략의 성과”, 지능정보연구, 제27권 제1호, 2021, 65-82.

김도완, “안전자산 선호현상이 투기등급 채권발행을 어렵게 하는가?”, 재무관리연구, 제35권 제1호, 2018, 1-25.

김용석, 조성욱, “한국어 텍스트 분석과 적용: 머신러닝을 통한 증권발행신고서의 비정형화된 텍스트 분석”, 한국증권학회지, 제48권 제2호, 2019, 215-235.

배광일, 이순희, “거시경제에 대한 이질적 기대가 채권초과수익률에 미치는 영향”, 재무관리연구, 제37권 제1호, 2020, 1-23.

송민찬, 류두진, “기계학습 기반 기업신용정보 분석을 통한 채무불이행 예측”, 재무연구, 제34권 제4호, 2021, 199-234.

안지영, 임병권, “머신러닝 알고리즘을 이용한 MBS 조기상환율 예측”, 금융연구, 제34권 제2호, 2020, 33-63.

양철원, “한국의 채권과 주식시장 유동성의 상호관계”, 대한경영학회지, 제26권 제2호, 2013, 351-3701.

윤윤석, 김도형, 최찬수, “회사채수익률에 영향을 미치는 회계정보에 관한 연구”, 상업교육연구, 제10권, 2005, 209-226.

이현상, 오세환, “시계열 예측을 위한 LSTM 기반 딥러닝: 기업 신용평점 예측 사례”, 정보시스템연구, 제29권 제1호, 2020, 241-265.

정희준, 이한구, 김종희, “코로나19 전후의 공모회사채 발행가격 형성에 관한 연구”, 금융공학연구, 제20권 제3호, 2021, 27-57.

채병권, 한재현, “수요예측제도 도입과 채권발행가격 결정요인 분석”, 회계와 정책연구, 2020, 제25권 제2호, 303-330.

최보람, 문예영, 구자은, “채권 발행금리와 스프레드에 미치는 영향”, 국제회계연구, 제43집, 2012, 213-240.

황광숙, 이준희, “금융위기 이후 경제정책 불확실성의 회사채스프레드에 대한 영향 분석”, 대한경영학회지, 제35권 제1호, 2022, 35-66.

Athanassakos, G. and P. Carayannopoulos, “An Empirical Analysis of the Relationship

of Bond Yield Spreads and Macro economic Factors," *Applied Financial Economics*, 11(2), (2001), 197-207.

Bernoth, K. and B. Erdogan, "Sovereign Bond Yield Spreads: A Time-Varying Coefficient Approach," *Journal of International Money and Finance*, 31, (2012), 639-656.

Bianchi, D., M. Buchner, and A. Tamoni, "Bond Risk Premiums with Machine Learning," *Review of Financial Studies*, 34(2), (2021), 1046-1089.

Chen, L., D. A. Lesmond, and J. Wei, "Corporate Yield Spreads and Bond Liquidity," *Journal of Finance*, 62(1), (2007), 119-149.

Chen, l., M. Pelger, and J. Zhu, "Deep Learning in Asset Pricing," *Management Science*, forthcoming, (2023).

Favero, C., M. Pagano, and E. Thadden, "How Does Liquidity Affect Government Bond Yields?," *Journal of Financial and Quantitative Analysis*, 45(1), (2010), 107-134.

Golbayani, P., I. Florescu, and R. Chatterjee, "A Comparative Study of Forecasting Corporate Credit Ratings Using Neural Networks, Support Vector Machines, and Decision Trees," *The North American Journal of Economics and Finance*, 54, (2020), 101251.

Gu, S., B. Kelly, and D. Xiu, "Empirical Asset Pricing via Machine Learning," *Review of Financial Studies*, 33(5), (2020), 2223-2273.

Han, S. and X. Zhou, "Informed Bond Trading, Corporate Yield Spreads, and Corporate Default Prediction," *Management Science*, 60(3), (2014), 675-694.

Hastie, T., R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," *Springer Science & Business Media*, 2009.

Hoerl, A. E. and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12(1), (1970), 55-67.

Huang, H., H. Y. Huang, and J. J. Oxman, "Stock Liquidity and Corporate Bond Yield Spreads: Theory and Evidence," *Journal of Financial Research*, 38(1), (2015), 59-91.

Jubinski, D., and A. F. Lipton, "Equity Volatility, Bond Yields, and Yield Spreads," *Journal of Futures Market*, 32(5), (2011), 480-503.

Kim, J. M., D. H. Kim, H. Jung, "Applications of Machine Learning for Corporate Bond Yield Spread Forecasting," *North American Journal of Economics and Finance*, 58, (2021), 101540.

Kim, M., "Adaptive Trading System Integrating Machine Learning and Back-Testing:

Korean Bond Market Case," *Expert Systems with Applications*, 176(15), (2021), 114767.

Li, F., "Annual Report Readability, Current Earnings, and Earnings Persistence," *Journal of Accounting and Economics*, 45(2-3), (2008), 221-247.

Mayberger, M., E. Pana, F. Hoyt, D. Marvin, and D. Mendez-Carbajo, "How Do Bond Specific, Firm Specific and Macroeconomic Factors Influence Corporate Credit Spreads?," *Working Paper*, (2014).

Mishra, S., and S. Padhy, "An Efficient Portfolio Construction Model Using Stock Price Predicted by Support Vector Regression," *The North American Journal of Economics and Finance*, 50(C), (2019), 101027.

Moscatelli, M., F. Parlapiano, S. Narizzano, and G. Viggiano, "Corporate Default Forecasting with Machine Learning," *Expert Systems with Applications*, 161, (2020), 113567.

Wang, J., C. Wu, and F. X. Zhang, "Liquidity, Default, Taxes, and Yields on Municipal Bonds," *Journal of Banking & Finance*, 32, (2008), 1133-1149.

Saltzman, B., and J. Yung, "A Machine Learning Approach to Identifying Different Types of Uncertainty," *Economics Letters*, 171, (2018), 58-62.

Tibshirani, R., "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 1996, 267-288.

Zhang, G., M. Y. Hu, B. E. Patuwo, and D. C. Indro, "Artificial Neural Networks in Bankruptcy Prediction: General Framework and Cross-Validation Analysis," *European Journal of Operational Research*, 116(1), (1999), 16-32.

Zou, H. and T. Hastie, "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society: Series B (statistical methodology)*, 67(2), 2005, 301-320.

THE KOREAN JOURNAL OF FINANCIAL MANAGEMENT
Volume 40, Number 5, October 2023

Forecasting Corporate Bond Yields with Machine Learning*

Jiyoung An** · Byungkwon Lim**

〈Abstract〉

This study analyzes the corporate bond yield spread predictability with machine learning methods (LASSO, Ridge, Elastic net). We consider comprehensive input variables such as bond characteristics, macroeconomic factors, bond and stock market, and financial factors of issuers. The major empirical findings are as follows. First, we find that credit ratings, credit spreads, interest rate spread, base rate, and GDP are critical factors in corporate bond yield spreads. In addition, the optimal prediction method varies depending on the model type, the number of variables, and the sample size. Second, we find that the predictive model with the general corporate bond yield spreads could be helpful in predicting ESG bond initial yield rates.

Overall, our findings show that machine learning for predicting corporate bond yield spread is more valuable than the traditional OLS method. Our evidence provides a new perspective in determining corporate bond rates and practical implications for improving useful information to the bond issuer or investors.

Keywords : Machine Learning, Elastic Net, Corporate Bond, ESG Bond, Yield Spread

* This Research was supported by the Commercialization Promotion Agency for R&D Outcomes (COMPA) funded by the Ministry of Science and ICT (MSIT) (1711195821, Science and Technology Acceleration for Region + Academy (Chungnam National University).

** First Author, Associate Research Fellow, Korea Energy Economics Institute, E-mail: ajy4129@gmail.com
*** Corresponding Author, Professor, Graduate School, Department of Technology Practical Convergence, Chungnam National University, E-mail: bk81.lim@gmail.com