

머신러닝을 이용한 설비투자 예측 및 분석: 뉴스심리지수와 구글 트렌드 데이터를 중심으로*

조두연** · 황보현***

국문초록

설비투자는 국가의 생산능력을 결정하는 중요한 요인 중 하나이기 때문에, 이에 대한 정확한 예측을 통하여 동향을 파악하고 설비투자를 활성화할 수 있는 적절한 정책 시행이 요구된다. 본 연구를 통해 설비투자에 영향을 미치는 다양한 거시경제 변수 외에도 텍스트 데이터(Text Data)인 한국은행 뉴스심리지수(News sentiment index, NSI)와 구글 트렌드 데이터(Google Trends Data)를 이용하여 설비투자에 대한 예측력을 개선할 수 있는지 분석하였다. 설비투자 예측 모형에서 거시경제 관련 설명변수를 이용한 모형의 예측력과 한국은행 뉴스심리지수 및 설비투자 관련 검색어 추세 변화량을 추가한 모형의 예측력을 비교하였다. 설비투자 예측 모형의 예측력을 비교하기 위해 머신러닝(Machine Learning)에 기반한 Adaptive LASSO, Ridge Regression, Random Forest, Neural Network 모형을 추정한 결과, 거시경제 변수를 포함하는 정형데이터만 활용한 모형의 예측력에 비하여 한국은행 뉴스심리지수와 구글 트렌드 데이터를 함께 활용한 모형의 예측력이 개선됨을 보였다. 본 연구의 분석결과에 따르면, 향후 설비투자의 예측에 있어 정형데이터 이외에도 뉴스 데이터 및 설비투자와 연관된 검색어 추세 등 텍스트 데이터를 활용한다면 정확도가 개선될 것이며, 더 나아가 이를 정책 시행에 이용할 수 있을 것으로 보인다.

핵심주제어: 설비투자, 머신러닝, 구글 트렌드 데이터, 뉴스심리지수, 예측

JEL 번호: B030104, B030109

· 논문접수일: 2023년 7월 12일, 심사 및 수정일: 2023년 8월 15일, 게재확정일: 2023년 9월 6일.

* 본 연구의 내용은 전적으로 집필진의 의견이며, 한국은행의 공식적 견해와는 무관함을 밝힌다. 본 논문에 대해 유익한 의견을 주신 두 분의 익명의 심사자에게 감사드린다. 본 논문에 혹시 남아 있을 수 있는 오류는 저자들의 책임임을 밝힌다.

** 성균관대학교 경제대학 경제학과 및 쿼트응용경제학과 부교수. E-mail: dooyeoncho@g.skku.edu.

*** 교신저자. 성균관대학교 경제대학 쿼트응용경제학과 석사, 한국은행 조사역. E-mail: totop2500@g.skku.edu, bhwang95@bok.or.kr

www.kci.go.kr

I. 서론

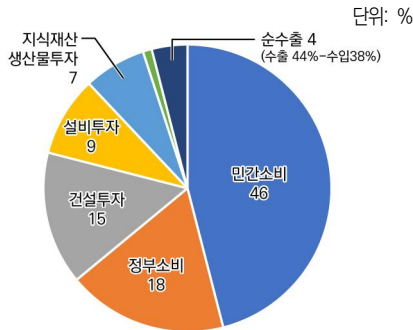
설비투자(Facilities investment)는 기업이 재생산 등을 목적으로 기계장치 및 차량 등의 설비를 도입하는 투자를 의미한다. 민간소비 및 정부소비와 비교하였을 때, 설비투자는 GDP 대비 9% 정도로 상대적으로 비중은 작지만(〈그림 1〉), 경제 전체의 생산능력을 결정하고 고용수준을 결정한다는 점에서 수요와 공급의 성장을 이끄는 중요한 변수 중 하나라고 볼 수 있다. 최근 한국은행의 긴축적 통화정책의 시행으로 인한 고금리 기조가 지속됨에 따라 기업의 투자심리가 위축되었을 뿐 아니라, 미·중 갈등 심화 및 글로벌 IT업황의 둔화로 인해 설비투자의 부진한 흐름이 이어지고 있다. 2023년 1/4분기 GDP 대비 설비투자는 전 분기에 비하여 5.0% 감소하였으며, 설비투자의 선행지표인 설비투자 조정압력도 〈그림 2〉에서와 같이 2022년 하반기 이후 하락을 지속하고 있는 실정이다.¹⁾

설비투자의 감소로 인해 향후 잠재성장률의 저하로 이어질 가능성이 존재한다는 점에서 설비투자는 거시경제와 관련하여 중요한 변수 중 하나이며, 이에 따라 설비투자 동향에 대한 정확한 모니터링이 요구된다. 효과적이고 안정적인 통화정책의 수립 및 집행을 위해서도 설비투자 등을 포함하여 거시경제 전반에 대한 정확한 분석 및 예측이 선행되어야 한다. 하지만 설비투자의 의사결정은 미래에 대한 예측을 기반으로 이루어지고, 설비투자에 대해 가시적인 결과가 나타나는데 적지 않은 기간이 소요된다는 점 때문에 다른 거시 관련 변수 중에서도 상대적으로 불확실성과 변동성이 높다고 할 수 있다. 따라서, 본 연구에서는 머신러닝 기법 및 한국은행의 뉴스심리지수, 구글 트렌드 데이터를 포함하는 텍스트 데이터 등을 활용하여 설비투자 전망에 대한 예측력을 높일 수 있는지를 분석하려고 한다.²⁾

1) 설비투자 조정압력은 제조업 부문의 생산증가율(X)과 생산능력 증가율(Y)의 차이(X-Y)로 계산되며 설비투자에 대한 수요를 의미하는 지표이다. 설비투자 조정압력이 양(+)의 값을 가지며 증가한다는 것은 향후 설비투자 확대에 대한 필요성이 높아짐을 의미하는 반면, 음(-)의 값을 가진다는 것은 향후 설비투자 여력이 상대적으로 낮은 상황임을 나타낸다.

2) 뉴스심리지수(News sentiment index, NSI)는 한국은행에서 개발하여 제공하는 지수이다. 구체적으로, 뉴스기사 중 경제 분야에 해당하는 기사 텍스트를 웹 스크래핑(Web scraping)을 이용하여 수집하고, 이를 자연어 처리(Natural Language Processing, NLP) 기법으로 분석하여, 경제심리를 지수화한 것이다. 한국은행 경제 통계시스템(ECOS)에서 일별 및 월별 자료를 제공하고 있다.

〈그림 1〉 지출 GDP 중 설비투자의 비중



자료 : 한국은행.

〈그림 2〉 설비투자 조정압력



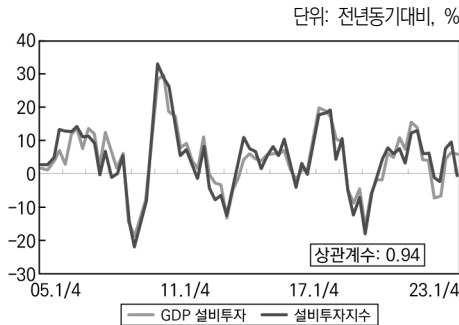
자료 : 통계청.

주: 1) 설비투자 조정압력 = 생산 증가율 - 생산능력 증가율

설비투자의 측정항목으로 한국은행의 GDP 대비 설비투자와 통계청의 설비투자지수가 존재하며, 〈그림 3〉에 제시된 바와 같이 두 지표는 상관관계 0.94로 상당히 높은 것으로 나타나지만, 포괄범위와 기초자료 등의 차이로 인해 완벽하게 일치하지는 않는다. 매 분기마다 발표되는 GDP 자료에 비해 통계청의 설비투자지수는 월별로 발표된다는 점에서 보다 신속하게 설비투자 업황의 흐름을 파악할 수 있다는 장점을 가지고 있지만, 여전히 2개월 정도의 시차가 존재하여 이에 대한 예측이 필요한 실정이다.

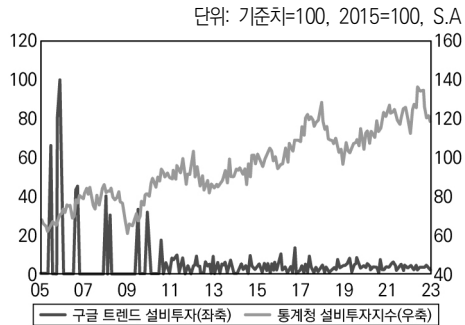
변수 간 비선형 관계를 분석하는 데 유용하다고 알려져 있는 머신러닝 알고리즘 및 기존에 사용되던 정형 데이터 이외에 비정형 데이터인 텍스트 데이터 등을 설명변수에 추가하여 거시 및 금융 관련 지수에 대한 예측력을 높이려는 시도는 활발히 이루어져 왔다. 설비투자에 대한 이전의 연구는 주로 다른 거시변수와 설비투자 간의 상관관계에 초점을 맞추어 분석하거나, 설비투자의 의사결정 과정을 분석하는 방식으로 진행되었다. 예를 들어, 홍정효(2018)는 국내 민간소비, 설비투자, 저축률과 GDP 간의 인과성을 분석하여 설비투자와 소비는 경제성장률과 양(+)의 관계가 있는 반면, 저축률은 경제성장률과 음(-)의 관계가 있음을 보였다. 이용수·현완순(2005)는 불확실성하에서 설비투자에 대한 의사결정이 이루어지는 과정을 분석하였다. 한국은행 보고서(2017)는 경기도를 중심으로 2008년 글로벌 금융위기 이후 설비투자의 결정요인을 분석하였으며, 비용 측 요인보다는 수요 측 요인인 주가지수 및 환율 등이 더 유의하다는 사실을 밝혀내었다. 김병화·

〈그림 3〉 한국은행 GDP 대비 설비투자지수 및
통계청 설비투자지수



자료: 한국은행, 통계청.

〈그림 4〉 구글 트렌드 데이터인 설비투자와
통계청 설비투자지수



자료: 구글(Google), 통계청.

임현준(2002)은 국내 설비투자가 GDP에 가장 큰 영향을 받으며, 상대적으로 국내 설비투자에 금리가 미치는 영향이 크지 않다는 사실을 확인하였다.

머신러닝 기법을 이용하여 기존의 정형 데이터에 텍스트 데이터 등 비정형 데이터를 추가적으로 포함하여 분석함으로써 유가 혹은 수출입에 대한 예측력을 높일 수 있음을 보이는 연구는 활발히 이루어져 왔다. 최근, 김선미·조두연(2022)은 머신러닝 기법을 이용하여 거시변수 외에 추가로 구글 트렌드 데이터를 포함하여 유가 하락기에 유가의 예측력을 향상시킬 수 있음을 보였으며, 최정원·오세경·장재원(2017)은 머신러닝 기법에 기반하여 뉴스 텍스트 데이터를 추가함으로써, 월간으로 기업 부도에 대한 예측력을 높일 수 있음을 보였다. 성노윤·남기환(2017)은 주가를 예측하는 데 있어, 거시경제 지표와 뉴스 텍스트 데이터를 결합할 경우 머신러닝 모형의 예측력이 개선됨을 보였다.

본 연구에서는 위에서 언급한 기존의 연구와 달리, 중요한 거시변수 중 하나인 ‘설비투자’를 예측하는 데 있어 한국은행의 뉴스심리지수, 구글 트렌드 데이터를 포함하는 텍스트 데이터를 활용하여 예측력을 개선할 수 있는지 분석하고자 한다. 즉, 위에서 설명한 바와 같이, 설비투자에 대한 기존의 연구는 주로 다른 거시변수와의 상관관계를 분석하는 데 초점이 맞추어져 있었지만, 본 연구에서는 설비투자 전망에 대한 예측력을 높일 수 있는지를 분석하려고 한다. 본 연구에서는 다양한 머신러닝 기법에 기반한 예측모형을 이용하여, 〈그림 4〉에 제시된 구글 트렌드 데이터의 키워드 검색 추세 및 한국은행 뉴스심리지수를 설명변수로 추

가하여 설비투자액에 대한 예측력을 향상시킬 수 있는지 분석한다.

본 논문의 구성은 다음과 같다. 제Ⅱ장에서는 실증모형과 정형 및 비정형 데이터를 제시하고, 제Ⅲ장에서는 머신러닝 기법에 기반한 Adaptive LASSO, Ridge Regression, Random Forest, Deep Neural Network를 포함하는 다양한 예측모형을 이용하여 분석한 결과를 제시한 뒤, 설비투자액에 대한 예측 결과를 분석한다. 마지막으로, 제Ⅳ장에서는 결론 및 시사점을 제시한다.

Ⅱ. 연구방법 및 데이터

1. 연구방법 및 실증모형

본 연구에서는 머신러닝 기법에 기반하여 거시경제 변수를 예측하고 분석하는데 주로 사용되는 Adaptive LASSO, Ridge Regression, Random Forest 및 Deep Neural Network를 포함하는 다양한 모형을 추정하여 설비투자액 예측력 향상 여부에 대해 비교하고 분석한다.³⁾

가. Adaptive LASSO 모형

LASSO(Least Absolute Shrinkage and Selection Operator)는 최소제곱법의 목적함수에 벌점항(Penalty term)을 추가한 Penalized regression model로, Tibshirani (1996)에 의해 처음으로 제시되었다. 관측치보다 설명변수가 많거나 설명변수들 간 다중공선성(Multicollinearity)이 높아 회귀계수의 과잉추정(Overestimation)이 우려될 경우 사용하는 모형이다. 회귀계수에 제약을 설정하여 추정치의 크기를

3) 본 연구에서는 다수의 설명변수가 사용되었기 때문에 고차원 데이터(High dimensional data)의 차원 축소(Dimension shrinkage)에 유용한 Adaptive LASSO와 Ridge Regression을 활용한다. 한편, Random Forest와 Deep Neural Network 모형은 전통적 회귀분석의 가정이 필요하지 않고, 다양한 변수를 동시에 고려하면서 데이터 간의 관계를 보다 유연하게 포착할 수 있다는 장점을 가지고 있기 때문에 본 연구에서 분석을 위해 활용한다.

축소시키고 변수선택을 가능하게 하며, 어느 정도 편의(Bias)를 허용하는 대신 분산(Variance)을 감소시켜 예측의 정확도를 높인다는 장점이 있다.

$$\hat{\beta}^{lasso} = \arg_{\beta} \min \left\{ \sum_{i=1}^N (y_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j |\beta_j| \right\}, \lambda \geq 0$$

LASSO의 목적함수는 위와 같으며 N은 전체 데이터의 수, i는 데이터의 인덱스를, j는 예측변수의 인덱스를 나타낸다. 첫째항인 $\sum_{i=1}^N (y_i - \sum_j \beta_j x_{ij})^2$ 은 실제 관측값(y_i)과 예측값 간의 차이인 잔차의 제곱합(Residual Sum of Squares, RSS)을 나타내고, 두 번째 항은 수축 패널티(Shrinkage penalty)로 조절 모수(Tuning parameter)인 λ 의 크기에 따라 패널티(Penalty)가 달라지며, λ 가 클수록 많은 패널티가 부여된다. λ 가 0인 경우에 패널티가 사라져 선형회귀(Linear regression, 즉 Ordinary least squares, OLS)와 같은 식이 되고, λ 가 커질수록 패널티가 커져 예측에 불필요한 변수는 0이 된다. 조절 모수 λ 의 값은 예측오차의 평균을 최소화하는 값으로 설정되며, 일반적으로 k-Fold 교차검증(Cross validation, CV)을 통해 결정된다. Adaptive LASSO는 LASSO의 벌점함수(Penalty function)에 가중치(w)를 적용하여 기존의 LASSO를 개선한 모형으로 Zou(2006), Bühlman과 van de Geer(2011)에 의하여 고안되었다. 회귀계수 초기추정치 역수($w = \frac{1}{|\beta|}$)인 가중치 벡터(w)의 크기가 클수록 작게, 크기가 작을수록 많이 축소되기 때문에 초기추정치가 0에 가까울수록 패널티를 더 많이 주게 된다. 이를 통해 기존 LASSO 모형보다 편의(Bias)를 낮추면서 예측성고를 높일 수 있다는 장점을 가지고 있지만 각 계수마다 가중치를 구해야 하므로 추정치를 구하는 데 기존 모형보다 시간이 오래 걸린다는 단점이 있다.

$$\hat{\beta}^{alasso} = \arg_{\beta} \min \left\{ \sum_{i=1}^N (y_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j w_j |\beta_j| \right\}, \lambda \geq 0$$

$$\text{with } w = \frac{1}{|\beta|}$$

해당 연구에서는 가중치(w)는 Medeiros et al.(2021)의 방법론에 따라 $w_i = \frac{1}{|\hat{\beta}^{lasso}| + \frac{1}{\sqrt{N}}}$ 를 사용하여 추정하였으며, 조절 모수 λ 도 BIC 기준을 통해 선정하였다.

나. Ridge Regression

Ridge Regression은 Hoerl and Kennard(1970)에 의해 고안된 기법으로 LASSO와 같이 목적함수(Objective function)에 벌점항(Penalty term)을 추가한 Penalized regression model로 다중공선성(Multicollinearity)이 높아 회귀계수의 과잉추정(Overestimation)이 우려될 경우 사용하는 모형이다. 본 모형은 어느 정도 편의(Bias)를 허용하는 대신 분산(Variance)을 감소시켜 예측의 정확도를 높인다는 장점이 있다.

LASSO는 회귀계수의 절댓값에 벌점을 부과하는 반면, Ridge Regression은 회귀계수의 제곱에 벌점을 부과한다는 차이점이 존재한다. 불필요한 변수를 0으로 만드는 LASSO와 달리 Ridge Regression은 불필요한 변수를 0에 근사하도록 축소시키기 때문에 모든 변수를 고려하지만, 동시에 모든 변수들이 모형에 포함되어 해석에 어려움이 있다는 단점이 있다.

$$\hat{\beta}^{ridge} = \arg_{\beta} \min \left\{ \sum_{i=1}^N (y_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j \beta_j^2 \right\}$$

Ridge Regression의 목적함수는 위와 같으며 N은 전체 데이터의 수, i는 데이터의 인덱스를, j는 예측변수의 인덱스를 나타낸다. 첫째항인 $\sum_{i=1}^N (y_i - \sum_j \beta_j x_{ij})^2$ 은

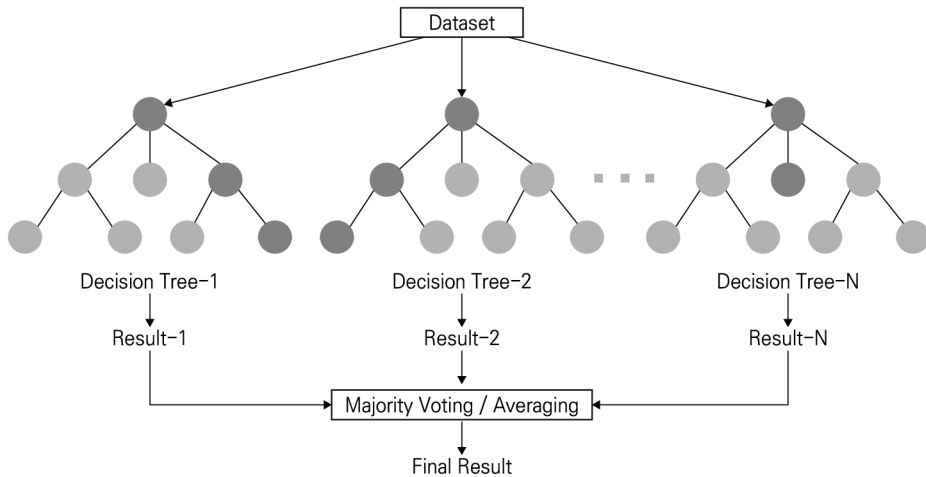
실제 관측값(y_i)과 예측값 간의 차이인 잔차의 제곱합(Residual Sum of Squares, RSS)을 나타내고, 두 번째 항은 수축 패널티(Shrinkage Penalty)로 조절 모수(Tuning parameter)인 λ 의 크기에 따라 패널티가 달라진다. λ 가 0일 경우 선형 회귀(Linear regression, 즉 Ordinary least squares, OLS)와 같은 식이 되며, λ 가 무한대(∞)일 경우 수축 패널티의 영향이 커지게 되며, Ridge Regression 회귀계수의 추정량은 0에 근접하게 된다.

다. Random Forest

Random Forest는 의사결정 나무(Decision tree) 알고리즘을 바탕으로 앙상블(Ensemble) 기법을 적용한 분석방법으로, 여러 개의 의사결정 나무를 생성하고 각 나무(Tree)가 예측한 결과를 종합하여 기존의 의사결정 나무의 단점을 개선한 모형이다. 의사결정 나무(Decision tree)는 복잡한 가정이 필요 없어 직관적이고 처리 속도가 빠르다는 장점을 지니지만 데이터가 조금만 바뀌어도 결과가 바뀌고 과적합(Overfitting)이 발생하기 쉽다는 단점을 지닌다. 이를 개선한 모형인 Random Forest는 여러 개의 모형을 만들고 각 모형의 예측값을 종합하는 방법인 앙상블 기법을 활용하며, 앙상블 기법 중 하나인 배깅(Bootstrap aggregating, Bagging)을 사용한다. 배깅은 부트스트랩(bootstrap)을 통해 집계하는 방식으로, 전체 데이터에서 복원추출을 통해 여러 개의 샘플을 만들고, 만들어진 각각의 샘플에 개별적으로 모형을 학습시킨다. 이렇게 만들어진 의사결정 나무들의 예측을 종합하여 최종 예측이 이루어지며, Random Forest는 각각의 의사결정 나무를 무작위(Randomly)로 선택하며 독립적인 의사결정을 시행하고 여러 나무의 결과가 조합되기 때문에 기존 의사결정 나무의 과적합 문제가 해소될 수 있다. <그림 5>에는 Random Forest 모형의 의사결정 과정이 나타나 있다. Random Forest는 변수 간의 관계를 유연하게 설정할 수 있어 비선형 관계를 예측하는 데 유리하다는 장점을 지니지만, 예측에 오랜 시간이 소요되고 결과가 일정하지 않으며 해석이 어렵다는 단점을 지닌다.⁴⁾

4) 본 연구에서 나무(Tree)의 수는 500개, Split의 수는 배깅(Bagging)의 1/3인 5로 설정하였다.

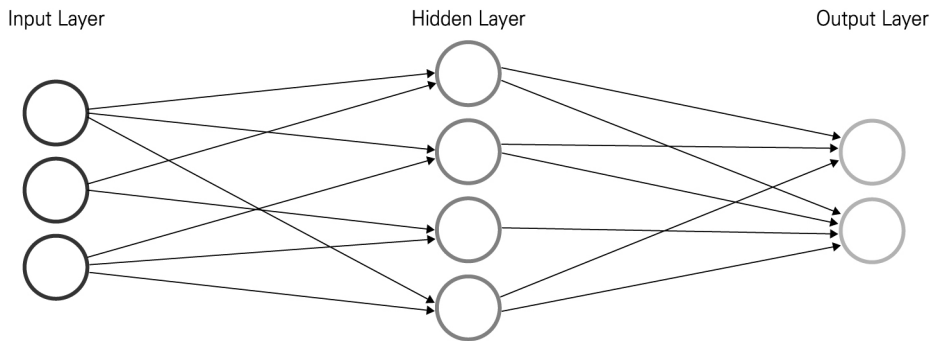
〈그림 5〉 Random Forest 모형



라. Deep Neural Network

인공신경망(Artificial Neural Network, ANN)은 인간의 뉴런구조(Neuron structure)를 모방하여 만든 딥러닝(Deep Learning) 알고리즘으로 다층의 Perceptron이 연결된 형태이다. 〈그림 6〉에 제시된 바와 같이 신경망(Neural Network)은 입력층(Input Layer), 은닉층(Hidden Layer), 출력층(Output Layer)으로 구성되어 있다. 입력된 데이터는 입력층 → 은닉층 → 출력층을 거치면서 예측값으로 변환되는데, 데이터가 입력되면 각 노드(Node)의 입력값과 가중치가 곱해진 값이 합산되고 활성화함수(Activation function)를 통해 최종 결과값이 출력되는 구조이다. 입력층에 자료를 입력하면 은닉층에서는 실제 관측치와 모형 예측치의 차이를 최소화하는 가중치를 골라 계산이 일어나게 되는데, 출력값이 실제값에 근접하도록 도움을 주는 노드 간의 가중치는 강화하고, 유의미한 영향을 주지 못하는 가중치는 약화시킨다. 해당 층에서 일어나는 계산의 결과를 볼 수 없어 은닉층이라 불린다. 출력층에서는 활성화함수를 통해 도출된 값이 나타나게 된다. 인공신경망 모형은 비선형 관계를 예측하는 데 우수하지만 은닉층이 늘어날수록 연산량이 증가하여 시간이 오래 소요되고 과적합(Overfitting) 문제가 발생할 수 있으며, 모형이 제시하는 결과에 대한 변수들의 인과관계를 설명할 수 없고 결과가 일정하지

〈그림 6〉 신경망(Neural Network) 모형



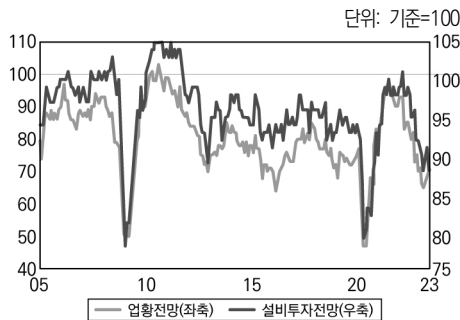
않다는 단점을 지닌다.

해당 연구에서는 인공신경망(ANN)과 같이 입력층과 출력층을 지니지만 은닉층(Hidden layer)의 수가 2개 이상인 심층신경망(Deep Neural Network, DNN) 모형을 추정한다. 은닉층의 수가 늘어날수록 더욱 복잡한 계산이 가능하며, 신경망이 비선형의 구조를 이루게 된다. 심층신경망(DNN) 모형의 활성화함수(Activation function)로 ReLU(Rectified Linear Unit)을 사용하였으며, 은닉층을 3개로 하고 각 은닉층의 퍼셉트론 개수는 각각 32개, 16개, 8개로 구성하였다. 또한, 반복학습(Epochs)을 100으로 설정하였다.

2. 데이터

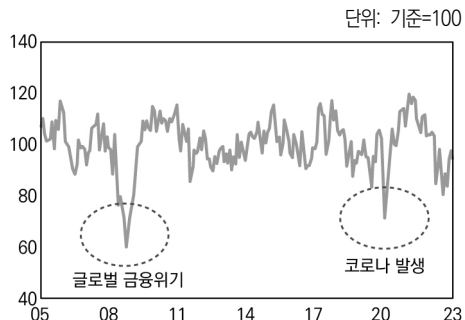
종속변수를 ‘설비투자지수 총지수’로 설정하였으며, 설명변수로는 기계류, 운송장비류 및 정보 통신 기술(Information & Communications Technology, ICT) 설비투자지수, 설비투자의 선행지표인 국내 기계 수주 금액과 자본재 내수출하, 반도체 재고 출하 비율 등을 고려하였으며 이외에도 거시변수인 산업생산, 소매 판매 등을 추가하였다. 수입액은 설비투자에서 주요한 비중을 차지하는 반도체 제조용 기계류, 자동차, 선박, 항공기 등을 선정하였으며, 미래 경기의 흐름을 반영하는 OECD 경기선행지수(Composite Leading Indicator, CLI)도 추가하였다. 기업심리를 반영하는 변수로 한국은행 BSI(Business survey index)를 추가하였다. BSI는 기업경기실사지수로 경제활동과 경기에 대한 기업의 판단, 전망 등을 지수

〈그림 7〉 제조업 업황 및 설비투자 BSI



자료: 한국은행.

〈그림 8〉 한국은행 뉴스심리지수(NSI)



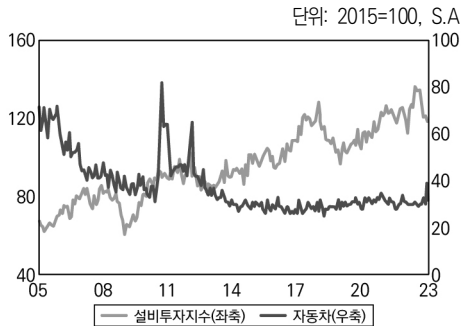
자료: 한국은행.

화한 지표이다. 〈그림 7〉은 업황전망 및 설비투자전망 BSI의 흐름을 나타내고 있으며 2022년 이후 금리 인상으로 기준치(100)를 하회하는 등 투자심리가 악화된 모습을 확인할 수 있다. BSI는 다른 통계자료를 통해서 포착하기 어려운 경제주체들의 심리변화를 측정할 수 있어 기업의 투자계획과 관련성이 크고 설비투자 전망에 유용한 변수라고 할 수 있다. Medeiros et al.(2021)의 방법론에 따라 설명변수에서 추출한 4개의 주성분요소(Principal component factor, PCF), 모든 설명변수 및 종속변수의 4개의 시차변수(Four lags of all variables)도 설명변수로 추가하여 각 모형에 사용하였다.

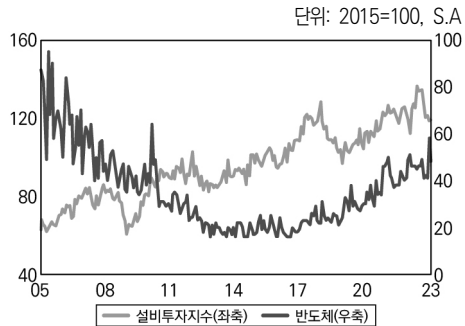
텍스트 데이터로는 한국은행 뉴스심리지수와 구글 트렌드 데이터를 사용하였다. 한국은행 뉴스심리지수는 한국은행에서 경제기사 표본을 뽑아 경기 관련 심리를 정량화한 지표로 100을 기준으로 한다. 〈그림 8〉에 제시된 바와 같이 한국은행 뉴스심리지수는 2008년 글로벌 금융위기나 2020년 코로나19 발생 시기에 급격하게 하락하는 모습을 확인할 수 있다. 구글 트렌드 데이터는 구글 사용자들의 키워드 검색어를 정량화한 데이터로 이를 통해 해당 키워드에 대한 사람들의 관심도를 파악할 수 있다. 설비투자의 구성요소인 기계류와 운송장비류에서 높은 비중을 차지하는 ‘반도체’와 ‘자동차’를 키워드로 설정하여 변수로 활용하였다.⁵⁾ 운송장비의 항목 중 하나인 기타운송장비(선박, 항공기)는 정부의 투자비중(46%)이 상대적으로 높고 항목의 변동성이 커서 포함하지 않았다.⁶⁾ 〈그림 9〉는 구글

5) 설비투자에서 반도체 제조용장비가 포함된 특수산업용 기계가 차지하는 비중은 33%, 자동차의 비중은 15%로 반도체와 자동차 두 항목이 50%가량을 차지한다.

〈그림 9〉 구글 트렌드 데이터: 자동차



〈그림 10〉 구글 트렌드 데이터: 반도체



〈그림 11〉 네이버 뉴스에서 '설비'를 키워드로 추출한 Word Cloud



트렌드에서 '자동차'로 검색한 결과를, 〈그림 10〉은 '반도체'로 검색한 결과를 보여준다.⁷⁾ 〈그림 11〉은 2022년 1월 1일부터 2023년 3월 31일까지 네이버 뉴스에서 '설비'를 검색한 경우에 나타나는 기사 제목에서 80개 이상 나오는 단어를 보여주며, 반도체가 중요한 키워드 중 하나임을 확인할 수 있다.

추정기간의 크기를 고정한 상태에서 시점을 이동하며 예측치를 생성하는 구간 이동(Rolling window) 방법을 사용하였다. 전체 표본의 크기는 2005년 1월에서 2023년 3월까지 219개이며, 추정 기간의 크기(Window size)는 120개월(10년), 예측 기간은 1개월, 3개월, 6개월, 9개월, 12개월로 설정하였다. 정형 데이터만 사용한 경우와 텍스트 데이터를 추가한 경우로 나누어 분석을 시행하였으며, 예측 오차를 계산하기 위해 RMSE(Root Mean Squared Error)를 이용하였다.

6) 2005년 1월부터 2023년 3월까지 각 변수에 대한 표준편차는 설비투자지수 18.1, 기계류 17.8, 운송장비 21.1이며 운송장비 중 자동차와 기타운송장비는 각각 19.4와 35.2이다.

7) '설비'를 키워드로 선정하여 네이버 뉴스에서 검색하였을 때도 '반도체'와 '자동차'와 관련된 키워드가 빈번하게 나오는 것을 확인할 수 있었다.

〈표 1〉 설비투자 예측에 사용된 변수

변수명	변환	변수명	변환
설비투자지수	$\Delta \log$	BSI전망	$\Delta \log$
기계설비류 설비투자지수	$\Delta \log$	반도체 재고/출하비율	$\Delta \log$
운송장비 설비투자지수	$\Delta \log$	반도체 제조용장비 수입액	-
ICT 품목 설비투자지수	$\Delta \log$	승용차 수입액	-
자본재 내수출하지수	$\Delta \log$	항공기 수입액	-
기계수주 국내수요(불변금액)	$\Delta \log$	선박 수입액	-
OECD CLI	log	Google Trends 반도체	$\Delta \log$
소매판매액지수	log	한국은행 뉴스심리지수	log
전산업생산지수	$\Delta \log$	Google Trends 자동차	$\Delta \log$

분석 기간은 2005년 1월부터 2023년 3월까지로 설정하여 OECD CLI 및 한국은행 뉴스심리지수가 이용가능한 시점인 2005년을 시작점으로 설정하였다. ADF(Augmented Dickey-Fuller)와 PP(Phillips-Perron) 단위근 검정 테스트(unit root tests)를 시행하여 I(1)인 변수를 로그(Log) 차분을 통해 안정적인 시계열로 변환하였다. 〈표 1〉은 분석에 사용하기 위해 수집한 변수 및 안정적인 시계열로 만들기 위해 시행한 데이터 변환을 보여준다.

III. 연구결과

1. 예측 결과 및 분석

가. 정형 데이터만 사용한 경우

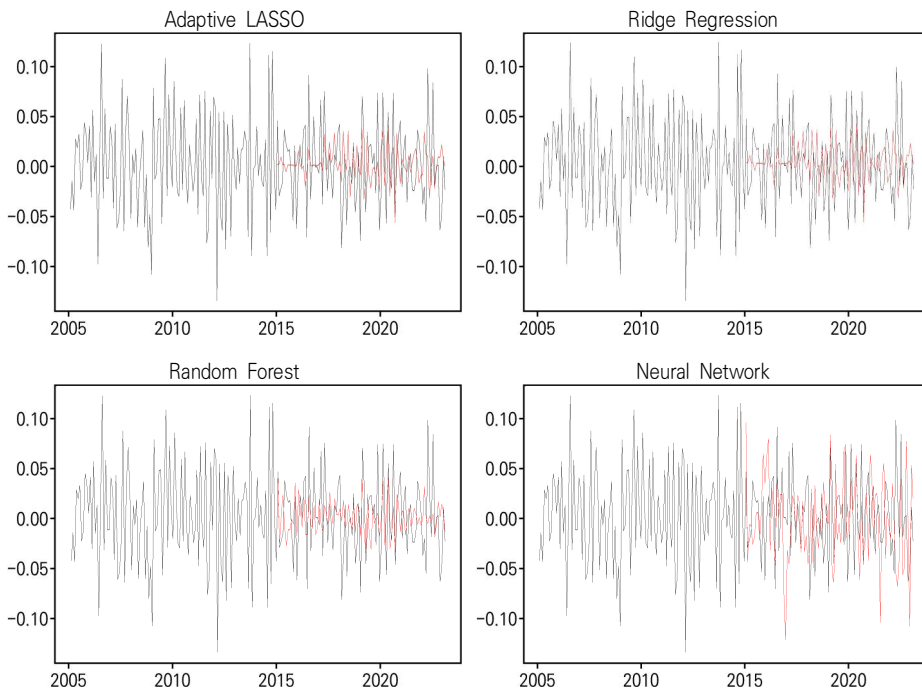
〈표 2〉는 정형 데이터만을 사용한 모형의 예측 결과(RMSE, %)를 각 예측대상 기간별로 제시하고 있다. 각 예측대상 기간마다 가장 낮은 RMSE값을 가지는 모

〈표 2〉 정형 데이터만 사용한 경우 모형의 RMSE(%) 비교

	1개월	3개월	6개월	9개월	12개월
Adaptive LASSO	3.977	3.951	4.087	3.978	3.975
Ridge Regression	4.846	3.978	4.132	4.769	5.203
Random Forest	3.775	3.848	3.972	4.000	4.017
Neural Network	5.203	5.865	5.455	5.283	5.455

주: 각 예측대상 기간마다 가장 낮은 RMSE값을 가지는 모형은 볼드(Bold)체로 표시되어 있으며, 모든 예측대상 기간 내에서 가장 낮은 RMSE값을 가지는 모형은 붉은색 볼드(Bold)체로 표시되어 있다.

〈그림 12〉 모형 추정결과(예측대상 기간 1개월)



형은 볼드(Bold)체로 표시되어 있으며, 모든 예측대상 기간 내에서 가장 낮은 RMSE값을 가지는 모형은 붉은색 볼드(Bold)체로 표시되어 있다. 정형 데이터만을 사용하여 예측을 시행한 경우, 예측대상 기간이 1개월일 경우, Adaptive LASSO의 RMSE(%)는 3.977, Ridge Regression은 4.846으로 추정되었다. 또한, Random Forest 모형의 RMSE(%)는 3.775, Neural Network 모형은 5.203으로 나타났다.

Random Forest, Adaptive LASSO, Ridge Regression, Neural Network 순으로 예측력이 좋았다. 예측대상 기간이 3개월과 6개월인 경우에도 Random Forest가 가장 작은 RMSE를 나타내어 예측력이 가장 좋았으며, 9개월과 12개월인 경우에는 Adaptive LASSO가 예측력이 가장 좋았다. <그림 12>에는 예측대상 기간이 1개월인 경우 정형 데이터만 사용한 모형의 예측결과가 제시되어 있다.

나. 비정형 데이터를 추가한 경우

① 정형 데이터 + 한국은행 뉴스심리지수

<표 3>은 정형 데이터와 한국은행 뉴스심리지수를 함께 사용한 모형의 예측결과(RMSE, %)를 각 예측대상 기간별로 제시하고 있다. 각 예측대상 기간마다 가장 낮은 RMSE값을 가지는 모형은 볼드(Bold)체로 표시되어 있으며, 모든 예측대상 기간 내에서 가장 낮은 RMSE값을 가지는 모형은 붉은색 볼드(Bold)체로 표시되어 있다. 한국은행 뉴스심리지수를 추가하여 예측을 시행한 경우, 예측대상 기간이 1개월일 때 Adaptive LASSO의 RMSE(%)는 3.961, Ridge Regression은 4.865로 추정되었다. 또한, Random Forest는 RMSE(%)가 3.766, Neural Network 모형은 4.841로 나타났다.

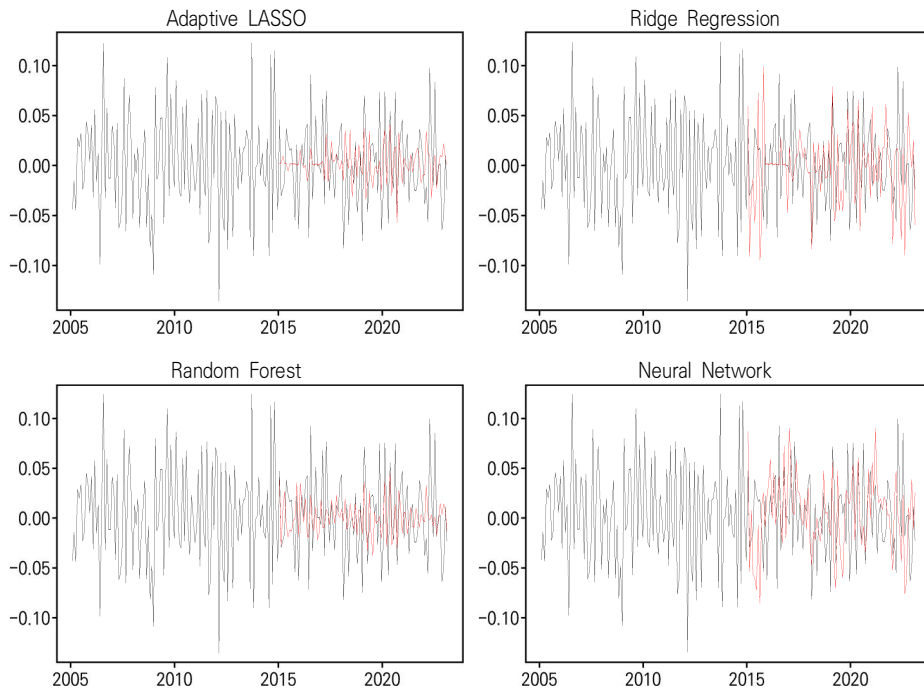
<그림 13>에서 나타난 모형의 예측결과와 같이 Random Forest, Adaptive LASSO, Neural Network, Ridge Regression 모형 순으로 예측력이 좋았다. 또한, Ridge Regression을 제외한 나머지 모형에서 정형 데이터만 사용한 경우에 비해 한국은

<표 3> 정형 데이터와 한국은행 뉴스심리지수를 사용한 경우 모형의 RMSE(%) 비교

	1개월	3개월	6개월	9개월	12개월
Adaptive LASSO	3.961	3.951	4.089	3.978	3.975
Ridge Regression	4.865	3.978	4.206	4.717	5.178
Random Forest	3.766	3.884	3.971	3.959	4.060
Neural Network	4.841	5.594	5.272	5.476	5.660

주: 각 예측대상 기간마다 가장 낮은 RMSE값을 가지는 모형은 볼드(Bold)체로 표시되어 있으며, 모든 예측대상 기간 내에서 가장 낮은 RMSE값을 가지는 모형은 붉은색 볼드(Bold)체로 표시되어 있다.

〈그림 13〉 모형 추정결과(예측대상 기간 1개월)



행 뉴스심리지수를 추가한 경우 RMSE가 낮아져 예측력이 개선됨을 확인할 수 있었다. 예측대상 기간이 3개월, 6개월, 9개월인 경우에도 Random Forest 모형의 예측력이 가장 좋았으며, 12개월인 경우에는 Adaptive LASSO 모형의 예측력이 더 좋음을 확인하였다.

② 정형 데이터 + 구글 트렌드 데이터 ‘자동차’

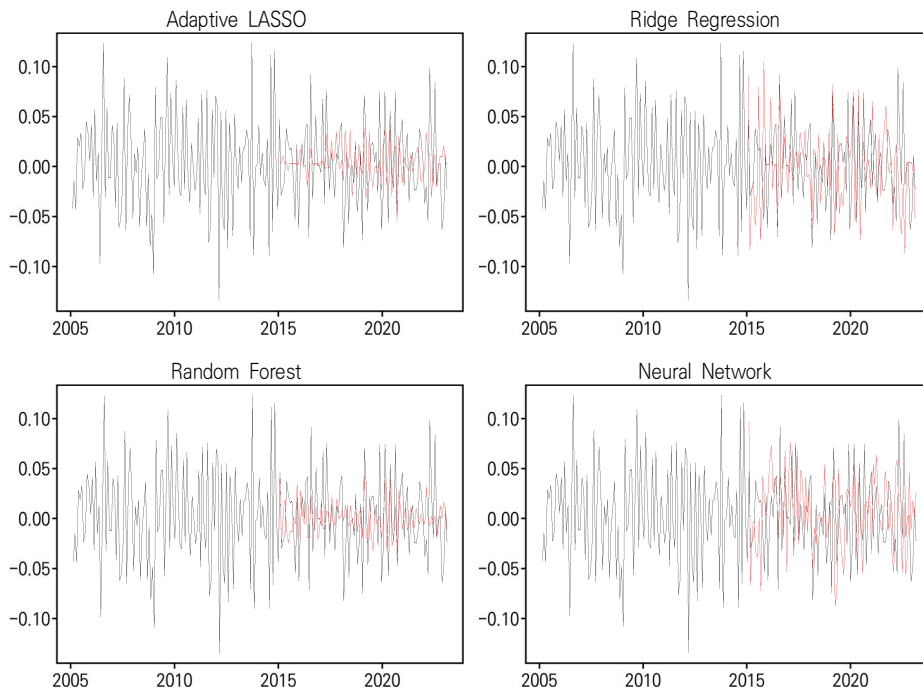
〈표 4〉는 정형 데이터에 추가적으로 구글 트렌드를 통해 ‘자동차’ 검색어 추세를 활용한 모형의 예측결과(RMSE, %)를 각 예측대상 기간별로 제시하고 있다. 각 예측대상 기간마다 가장 낮은 RMSE값을 가지는 모형은 볼드(Bold)체로 표시되어 있으며, 모든 예측대상 기간 내에서 가장 낮은 RMSE값을 가지는 모형은 붉은색 볼드(Bold)체로 표시되어 있다. 예측대상 기간이 1개월일 경우, Adaptive LASSO의 RMSE(%)는 3.969, Ridge Regression은 4.950로 추정되었다. 또한, Random

〈표 4〉 정형 데이터와 구글 트렌드(자동차)를 사용한 경우 예측 모형의 RMSE(%) 비교

	1개월	3개월	6개월	9개월	12개월
Adaptive LASSO	3.969	3.950	4.090	3.980	3.975
Ridge Regression	4.950	3.980	4.660	5.110	5.614
Random Forest	3.756	3.880	3.990	3.980	4.060
Neural Network	4.670	5.760	4.930	5.300	5.624

주: 각 예측대상 기간마다 가장 낮은 RMSE값을 가지는 모형은 볼드(Bold)체로 표시되어 있으며, 모든 예측대상 기간 내에서 가장 낮은 RMSE값을 가지는 모형은 붉은색 볼드(Bold)체로 표시되어 있다.

〈그림 14〉 모형 추정결과(예측대상 기간 1개월)



Forest 모형의 RMSE(%)는 3.756, Neural Network 모형은 4.670로 나타났다. 〈그림 14〉에서 나타난 모형의 예측결과와 같이 Random forest, Adaptive LASSO, Neural Network, Ridge Regression 순으로 예측력이 좋았다. 한국은행 뉴스심리 지수를 사용했을 경우와 마찬가지로, Ridge Regression을 제외한 나머지 모형에서 정형 데이터만을 사용한 경우에 비해 구글 트렌드 데이터 ‘자동차’를 추가한

경우 예측력이 개선되었음을 알 수 있다. 예측대상 기간이 3개월, 6개월, 9개월인 경우에도 Random Forest 모형의 예측력이 가장 좋았다. <그림 14>에는 예측기간이 1개월인 경우 정형 데이터와 추가적으로 구글 트렌드를 통해 ‘자동차’ 검색어 추세를 활용한 모형의 예측결과가 제시되어 있다. 예측기간이 12개월인 경우에는 Adaptive LASSO 모형의 예측력이 더 좋았다.

③ 정형 데이터 + 구글 트렌드 데이터 ‘반도체’

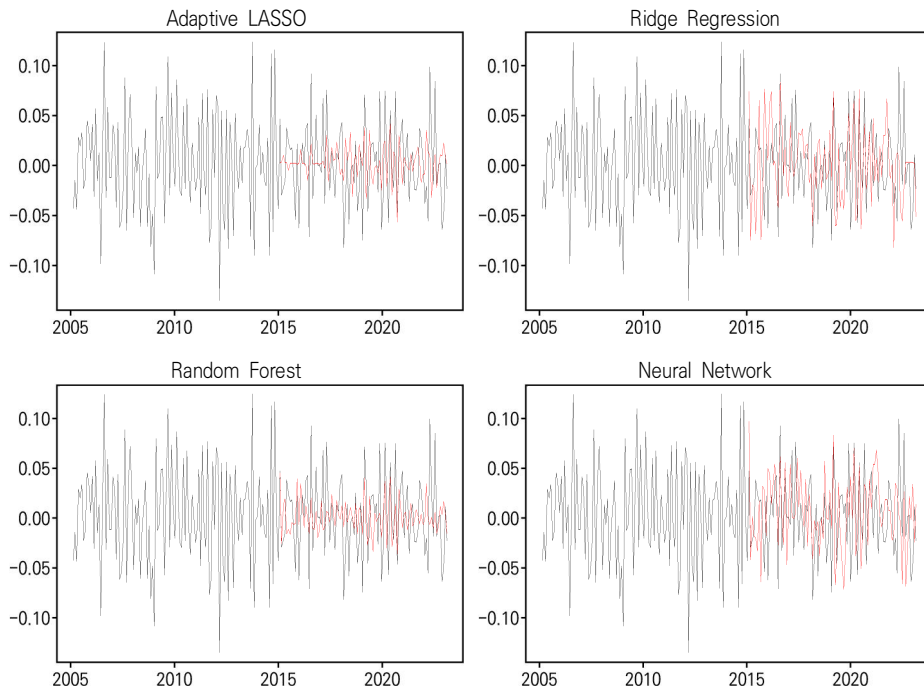
<표 5>는 정형 데이터에 추가적으로 구글 트렌드를 통해 ‘반도체’ 검색어 추세를 활용한 모형의 예측결과(RMSE, %)를 각 예측대상 기간별로 제시하고 있다. 각 예측대상 기간마다 가장 낮은 RMSE값을 가지는 모형은 볼드(Bold)체로 표시되어 있으며, 모든 예측대상 기간 내에서 가장 낮은 RMSE값을 가지는 모형은 붉은색 볼드(Bold)체로 표시되어 있다. 예측대상 기간이 1개월일 경우, Adaptive LASSO의 RMSE(%)는 4.058, Ridge Regression은 4.762로 추정되었다. 또한, Random Forest 모형의 RMSE(%)는 3.767, Neural Network 모형은 4.819로 나타났다. <그림 15>에 제시된 바와 같이, 예측대상이 1개월일 경우, Random Forest, Adaptive LASSO, Ridge Regression, Neural Network 순으로 예측력이 좋았다. 정형 데이터만을 사용한 경우에 비해 구글 트렌드 데이터 ‘반도체’를 추가한 경우 Adaptive LASSO를 제외한 모든 모형에서 예측력이 개선되었음을 알 수 있다. 또한, 예측대상 기간이 3개월, 6개월, 9개월인 경우 Random Forest의 예측력이 가장 좋았고 12개월인 경우에는 Adaptive LASSO가 가장 예측력이 우수하였다.

<표 5> 정형 데이터와 구글 트렌드(반도체)를 사용한 경우 예측 모형의 RMSE(%) 비교

	1개월	3개월	6개월	9개월	12개월
Adaptive LASSO	4.058	3.960	4.090	3.980	3.975
Ridge Regression	4.762	3.980	4.190	4.920	5.262
Random Forest	3.767	3.930	4.000	3.970	4.076
Neural Network	4.819	5.710	5.190	5.310	5.401

주: 각 예측대상 기간마다 가장 낮은 RMSE값을 가지는 모형은 볼드(Bold)체로 표시되어 있으며, 모든 예측대상 기간 내에서 가장 낮은 RMSE값을 가지는 모형은 붉은색 볼드(Bold)체로 표시되어 있다.

〈그림 15〉 모형 추정결과(예측대상 기간 1개월)



④ 정형 데이터 + 한국은행 뉴스심리지수 + 구글 트렌드 데이터 ‘반도체’ & ‘자동차’

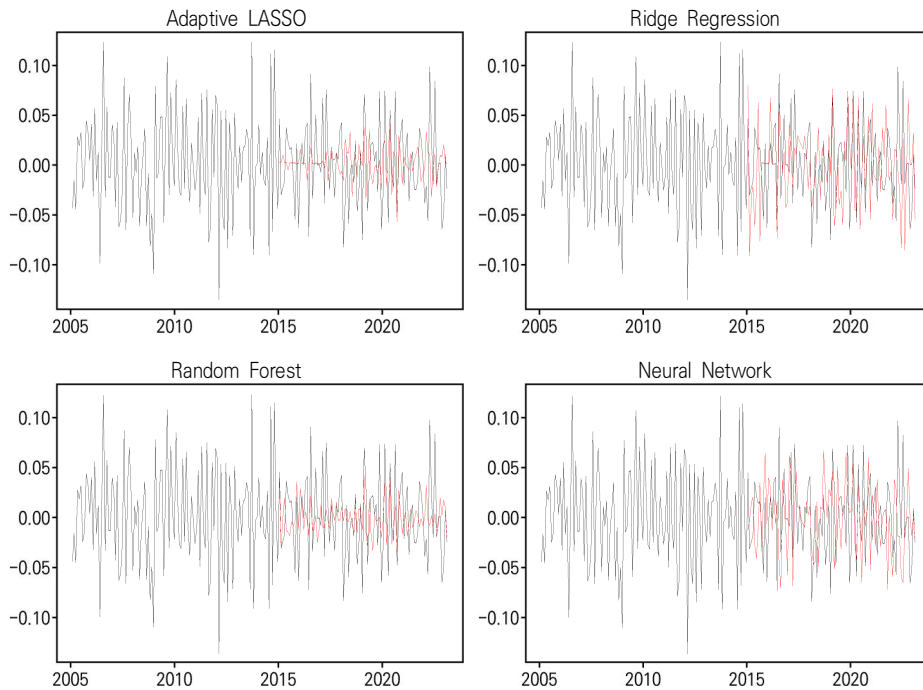
〈표 6〉은 정형 데이터에 한국은행 뉴스심리지수, 구글 트렌드 ‘반도체’와 ‘자동차’ 검색어 추세를 활용한 모형의 예측결과(RMSE, %)를 각 예측대상 기간별로 제시하고 있다. 각 예측대상 기간마다 가장 낮은 RMSE값을 가지는 모형은 볼드(Bold)체로 표시되어 있으며, 모든 예측대상 기간 내에서 가장 낮은 RMSE값을 가지는 모형은 붉은색 볼드(Bold)체로 표시되어 있다. 예측대상 기간이 1개월일 때 Adaptive LASSO의 RMSE(%)는 4.079, Ridge Regression은 4.591로 추정되었다. 또한, Random Forest 모형의 RMSE(%)는 3.733, Neural Network 모형은 5.232로 나타났다. Random Forest, Adaptive LASSO, Ridge Regression, Neural Network 순으로 예측력이 좋았다. 또한, Random Forest의 예측력이 정형 데이터만을 이용한 경우나 다른 비정형 데이터를 추가한 경우보다 RMSE가 가장 낮아져 예측력

〈표 6〉 정형 데이터, 한국은행 뉴스심리지수, 구글 트렌드 데이터(반도체, 자동차)를 사용한 경우 예측모형의 RMSE(%) 비교

	1개월	3개월	6개월	9개월	12개월
Adaptive LASSO	4.079	3.960	4.091	3.978	3.975
Ridge Regression	4.591	3.978	4.561	5.130	6.202
Random Forest	3.733	3.893	4.018	3.968	4.100
Neural Network	5.232	5.498	5.792	5.306	5.455

주: 각 예측대상 기간마다 가장 낮은 RMSE값을 가지는 모형은 볼드(Bold)체로 표시되어 있으며, 모든 예측대상 기간 내에서 가장 낮은 RMSE값을 가지는 모형은 붉은색 볼드(Bold)체로 표시되어 있다.

〈그림 16〉 모형 추정결과(예측대상 기간 1개월)



이 가장 많이 개선되었음을 알 수 있다. 예측대상 기간이 3개월, 6개월, 9개월인 경우 Random Forest 모형의 예측력이 가장 좋았다. 한편 12개월인 경우에는 Adaptive LASSO 모형의 예측력이 가장 우수하였다. 〈그림 16〉에는 예측대상이 1개월인 경우 정형 데이터에 추가적으로 구글 트렌드를 통해 '자동차'와 '반도체' 검색어 추세를 활용한 모형의 예측결과가 제시되어 있다.

따라서, 본 연구에서는 설비투자를 예측할 경우, 거시경제와 관련하여 기존에 사용되던 정형 데이터만 사용한 모형의 예측력에 비해 텍스트 데이터인 한국은행 뉴스심리지수 및 구글 트렌드 검색어 추세('자동차' & '반도체')를 설명변수로 추가하여 분석한 모형의 예측력이 개선된다는 점을 보였다. 또한, 설비투자에 대한 예측대상 기간을 1개월, 3개월, 6개월, 9개월, 12개월로 나누어 분석한 결과, 예측대상 기간이 짧을수록(즉, 1개월) 예측력이 좋다는 점을 발견하였으며, 본 연구에 사용된 머신러닝 기법에 기반한 Adaptive LASSO, Ridge Regression, Random Forest 및 Neural Network 모형 중 텍스트 데이터를 추가하였을 경우, 거의 모든 예측대상 기간에 대하여 Random Forest 모형의 예측력이 가장 개선됨을 보였다.

IV. 결론 및 시사점

민간소비 및 정부소비에 비하여 설비투자는 GDP 대비 9% 정도로 상대적으로 비중은 작지만, 경제 전체의 생산능력을 결정하고 고용수준을 결정한다는 점에서 수요와 공급의 성장을 이끄는 중요한 변수 중 하나이다. 최근 고금리 기조의 지속으로 인해 기업의 투자심리가 위축되어 설비투자의 부진한 흐름이 이어지고 있으며, 설비투자의 감소로 인해 향후 잠재성장률의 저하로 이어질 가능성이 존재한다는 점에서 설비투자에 대한 예측은 중요하다고 할 수 있다. 따라서, 본 연구에서는 머신러닝 기법 및 한국은행의 뉴스심리지수, 구글 트렌드 데이터를 포함하는 텍스트 데이터 등을 활용하여 설비투자 전망에 대한 예측력을 높일 수 있는지를 분석하였다.

본 연구는 2005년 1월부터 2023년 3월까지의 데이터를 이용하여 설비투자지수에 영향을 주는 거시경제 관련 변수 이외에 검색어 추세 등 텍스트 데이터를 추가했을 경우, 예측력이 개선됨을 확인하였다. 다양한 모형의 예측력을 비교하기 위해 머신러닝 기법에 기반한 Adaptive LASSO, Ridge Regression, Random Forest, Neural Network 모형을 추정하여 각 모형의 RMSE(%)를 산출한 결과, 정형 데이터만을 이용한 모형의 예측력보다 비정형 데이터인 한국은행 뉴스심리지

수와 구글 트렌드 데이터를 함께 이용한 모형의 예측력이 개선됨을 보였다.

또한, 설비투자에 대한 예측대상 기간을 1개월, 3개월, 6개월, 9개월, 12개월로 나누어 분석한 결과, 예측대상 기간이 짧을수록(즉, 1개월) 예측력이 좋다는 점을 발견하였으며, 본 연구에 사용된 머신러닝 기법에 기반한 Adaptive LASSO, Ridge Regression, Random Forest 및 Neural Network 모형 중 텍스트 데이터를 추가하였을 경우, 거의 모든 예측대상 기간에 대하여 Random Forest 모형의 예측력이 가장 개선됨을 보였다. 또한, 예측대상 기간이 12개월일 경우, 다른 모형에 비하여 Adaptive LASSO 모형이 가장 예측력이 좋다는 사실을 확인하였다.

향후 설비투자지수의 예측에 있어 텍스트 데이터인 한국은행 뉴스심리지수 및 구글 트렌드 데이터를 활용한다면 정확도가 개선될 것이라고 기대된다. 또한, 설비투자와 연관성이 높은, 보다 다양한 텍스트 키워드(Text keywords)를 선별하여 예측을 시행한다면 예측력을 개선시킬 수 있을 것으로 생각된다. 본 연구에서 다양한 예측 모형을 추정하여 얻은 설비투자 예측 결과를 효과적이고 안정적인 통화정책의 수립 및 집행에 활용할 수 있다는 점에서 중요한 시사점을 제공한다. 마지막으로, 설비투자에 대한 의사결정은 미래에 대한 예측을 기반으로 이루어진다는 점에서, 거시 관련 변수 중에서도 상대적으로 불확실성과 변동성이 높다고 알려져 있는데, 본 연구의 예측결과에 기반하여 설비투자 예측에 대한 정확도가 개선된다면 설비투자와 관련된 불확실성과 변동성을 줄일 수 있는 가능성을 제시했다는 점에서 의미 있는 결과라고 할 수 있다.

참고문헌

- 김병화 · 임현준(2002), “설비투자 결정요인 분석”, 『경제분석』, 8, pp. 32-74.
- 김선미 · 조두연(2022), “Machine Learning과 Google Trends Data를 이용한 유가 예측 및 분석”, 『한국경제학보』, 29, pp. 175-193.
- 김치호 · 김태운 · 박인호 · 안재준(2015), “경제심리지수의 유용성 및 개선방안에 관한 연구”, 『한국데이터정보과학회지』, 26, pp. 1335-1351.
- 박영준 · 엄주영(2017), “글로벌 금융위기 이후의 설비투자 결정요인 변화”, 한국은행 경기본부.
- 박철형(2022), “Lasso 기계학습을 이용한 주요 어종의 소비자가격 예측모형에 관한 연구”, 『해양비즈니스』, pp. 53-71.
- 서범석 · 이영환 · 조형배(2022), “뉴스 텍스트를 이용한 경기 예측: 경제 부문별 텍스트 지표의 작성과 활용”, 한국은행.
- 성노윤 · 남기환(2017), “온라인 뉴스 및 거시경제 변수를 활용한 주가예측”, *Entrue Journal of Information Technology*, 16, pp. 41-54.
- 원지수 · 나종화(2021), “벌점회귀를 통한 서울시 미세먼지 예측”, 『한국데이터정보과학회지』, 32, pp. 631-640.
- 유형주 · 이승오 · 최서혜 · 박문형(2019), “시계열 자료의 예측을 위한 자료 기반 신경망 모델에 관한 연구: 한강대교 수위예측 적용”, 『한국방재안전학회논문집』, 12, pp. 73-82.
- 이용수 · 정수일 · 현완순(2006), “불확실성 하에서의 설비 투자 결정에 대한 연구”, 『대한설비관리학회지』, 11, pp. 35-50.
- 최정원 · 오세경 · 장재원(2017), “빅데이터와 인공지능 기법을 이용한 기업 부도예측 연구”, 2017년 한국재무학회 추계학술대회, pp. 396-435.
- 홍정효(2018), “설비투자, 소비 및 저축률과 경제성장률(GDP)간의 인과성 연구”, 『지역산업연구』, 41, pp. 313-326.
- Bühlmann, P. and S. van de Geer(2011), *Statistics for High-Dimensional Data: Methods, Theory*, Springer-Verlag Berlin Heidelberg.
- Hoerl, A. E., R. W. Kennard(1970), “Ridge regression: Biased estimation for nonorthogonal problems”, *Technometrics*, 12, pp. 55-67.

- Medeiros M. C., Vasconcelos, G. F. R., Veiga, A., E. Zilberman(2021), "Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods", *Journal of Business & Economic Statistics*, 39, pp. 98-119.
- Tibshirani, R.(1996), "Regression shrinkage and selection via the Lasso", *Journal of the Royal Statistical Society, Series B(Methodological)*, 58, pp. 267-288.
- Zou, H.(2006), "The Adaptive Lasso and its oracle properties", *Journal of the American Statistical Association*, 101, pp. 1418-1429.

부 록

부록 I

〈표 A1〉은 2005년 1월부터 2023년 3월까지에 해당하는 기간 동안 예측에 사용된 변수의 평균(Mean), 표준편차(Standard deviation), 중간값(Median), 왜도(Skewness), 첨도(Kurtosis)를 보여주고 있다.

〈표 A1〉 예측 모형에 사용된 변수의 기초통계량

		Mean	S.D.	Median	Skewness	Kurtosis
설비투자지수	$\Delta\log$	0.003	0.047	0.001	0.101	-0.133
기계설비류 설비투자지수	$\Delta\log$	0.003	0.047	0.004	0.138	0.029
운송장비 설비투자지수	$\Delta\log$	0.002	0.108	0.000	0.459	1.444
ICT 품목 설비투자지수	$\Delta\log$	0.002	0.105	-0.002	0.336	0.834
자본재 내수출하지수	$\Delta\log$	0.002	0.040	0.000	-0.213	1.024
기계수주 국내수요	$\Delta\log$	0.004	0.241	-0.004	0.031	1.789
OECD CLI	log	4.606	0.013	4.608	-0.213	0.100
소매판매액지수	log	4.445	0.155	4.441	-0.230	-1.091
전산업생산지수	$\Delta\log$	0.002	0.013	0.002	-0.114	1.539
BSI 전망	$\Delta\log$	0.000	0.020	0.000	-0.709	2.819
반도체 재고/출하비율	$\Delta\log$	0.001	0.122	0.005	0.252	1.810
반도체 제조용장비 수입액	-	9.509	6.228	7.350	1.129	0.431
승용차 수입액	-	2.041	1.562	1.800	0.977	0.477
항공기 수입액	-	2.148	1.052	2.000	1.152	2.119
선박 수입액	-	5.790	3.981	5.300	0.547	-0.608
한국은행 뉴스심리지수	$\Delta\log$	4.599	0.106	4.614	-1.355	3.246
구글 트렌드 '반도체'	log	-0.002	0.174	0.000	0.436	1.469
구글 트렌드 '자동차'	$\Delta\log$	-0.002	0.102	0.000	0.166	1.868

부록 II

본 연구에서는 추정 기간의 크기(Window size)를 120개월(10년)로 설정하였는데 추정 기간의 크기를 변경할 경우, 모형의 예측력이 달라지는지를 검토하기 위하여 강건성 검정(Robustness checks)을 시행하였다. 추정 기간의 크기(Window size)를 84개월(7년)로 설정하였으며 추정 기간의 크기(Window size) 외의 요소는 위에서 설명한 바와 동일하게 설정하였다.⁸⁾

〈표 A2〉~〈표 A6〉은 설비투자지수의 예측에 있어 추정 기간의 크기(Window size)를 84개월(7년)로 설정하였을 때 i) 정형 데이터만 사용한 경우, ii) 정형 데이터에 한국은행 뉴스심리지수를 활용한 경우, iii) 정형 데이터에 구글 트렌드 ‘자동차’ 검색어 추세를 활용한 경우, iv) 정형 데이터에 구글 트렌드 ‘반도체’ 검색어 추세를 활용한 경우, v) 정형 데이터에 한국은행 뉴스심리지수와 구글 트렌드 ‘자동차’, ‘반도체’를 모두 사용한 경우에 대하여, 각 예측기간별(1개월, 3개월, 6개월, 9개월, 12개월)로 모형을 추정하여 얻은 RMSE(%)를 보여준다. 본 연구에서 추정 기간의 크기(Window size)를 120개월(10년)로 설정하였을 때(〈표 2〉~〈표 6〉)와 유사하게 정형 데이터만 사용한 모형의 예측력에 비해 텍스트 데이터인 한국은행 뉴스심리지수 및 구글 트렌드 검색어 추세(‘자동차’ & ‘반도체’)를 설명변수로 추가할 경우 예측력이 개선됨을 확인할 수 있었으며, 예측대상 기간이 짧을수록(즉, 1개월) 예측력이 좋다는 점을 확인하였다.

〈표 A2〉 정형 데이터만 사용한 경우 모형의 RMSE(%) 비교

	1개월	3개월	6개월	9개월	12개월
Adaptive LASSO	4.380	4.531	4.691	4.620	4.565
Ridge Regression	5.482	5.159	4.591	4.589	4.588
Random Forest	4.349	4.653	4.674	4.596	4.817
Neural Network	5.527	6.121	6.100	5.935	6.010

주: 각 예측대상 기간마다 가장 낮은 RMSE값을 가지는 모형은 볼드(Bold)체로 표시되어 있으며, 모든 예측대상 기간 내에서 가장 낮은 RMSE값을 가지는 모형은 붉은색 볼드(Bold)체로 표시되어 있다.

8) 추정 기간의 크기(Window size)를 84개월(즉, 7년)로 설정하여 예측했을 경우에도 120개월(즉, 10년)로 설정하여 예측했을 경우와 큰 차이가 없었음을 확인하였으며, 추정 기간의 크기(Window size)를 다르게 설정하여 예측할 경우에도 본 연구의 주요 결과와 큰 차이가 없음을 확인하였다.

〈표 A3〉 정형 데이터와 한국은행 뉴스심리지수를 사용한 경우 모형의 RMSE(%) 비교

	1개월	3개월	6개월	9개월	12개월
Adaptive LASSO	4.458	4.531	4.691	4.620	4.565
Ridge Regression	5.358	4.586	4.591	4.589	4.588
Random Forest	4.316	4.658	4.613	4.591	4.841
Neural Network	5.203	5.776	5.769	5.744	6.010

주: 각 예측대상 기간마다 가장 낮은 RMSE값을 가지는 모형은 볼드(Bold)체로 표시되어 있으며, 모든 예측대상 기간 내에서 가장 낮은 RMSE값을 가지는 모형은 붉은색 볼드(Bold)체로 표시되어 있다.

〈표 A4〉 정형 데이터와 구글 트렌드(자동차)를 사용한 경우 예측 모형의 RMSE(%) 비교

	1개월	3개월	6개월	9개월	12개월
Adaptive LASSO	4.393	4.541	4.691	4.578	4.490
Ridge Regression	5.835	4.586	4.591	4.589	4.588
Random Forest	4.315	4.617	4.635	4.563	4.761
Neural Network	5.153	5.859	5.634	5.679	6.032

주: 각 예측대상 기간마다 가장 낮은 RMSE값을 가지는 모형은 볼드(Bold)체로 표시되어 있으며, 모든 예측대상 기간 내에서 가장 낮은 RMSE값을 가지는 모형은 붉은색 볼드(Bold)체로 표시되어 있다.

〈표 A5〉 정형 데이터와 구글 트렌드(반도체)를 사용한 경우 예측 모형의 RMSE(%) 비교

	1개월	3개월	6개월	9개월	12개월
Adaptive LASSO	4.383	4.580	4.687	4.620	4.565
Ridge Regression	5.710	4.586	4.591	4.589	4.588
Random Forest	4.285	4.603	4.623	4.588	4.767
Neural Network	5.112	5.793	5.697	5.575	6.017

주: 각 예측대상 기간마다 가장 낮은 RMSE값을 가지는 모형은 볼드(Bold)체로 표시되어 있으며, 모든 예측대상 기간 내에서 가장 낮은 RMSE값을 가지는 모형은 붉은색 볼드(Bold)체로 표시되어 있다.

〈표 A6〉 정형 데이터, 한국은행 뉴스심리지수, 구글 트렌드(반도체, 자동차)를 사용한 경우 예측모형의 RMSE(%) 비교

	1개월	3개월	6개월	9개월	12개월
Adaptive LASSO	4.428	4.583	4.687	4.578	4.611
Ridge Regression	4.599	4.586	4.591	4.589	4.588
Random Forest	4.271	4.636	4.667	4.618	4.754
Neural Network	5.379	6.059	6.419	6.021	5.843

주: 각 예측대상 기간마다 가장 낮은 RMSE값을 가지는 모형은 볼드(Bold)체로 표시되어 있으며, 모든 예측대상 기간 내에서 가장 낮은 RMSE값을 가지는 모형은 붉은색 볼드(Bold)체로 표시되어 있다.

Abstract

Forecasting Facilities Investment using Machine Learning: The Role of the News Sentiment Index and Google Trends Data*

Dooyeon Cho** · Bohyun Hwang***

Since facilities investment is regarded as one of the important factors that determines a country's production capability, it appears to be necessary to implement a proper policy which helps understand patterns and trends and vitalize facilities investment through an accurate prediction. This study analyzes the predictive power using the Bank of Korea's News sentiment index (NSI) and Google Trends Data in addition to a variety of macroeconomic variables. We compare the predictive power of the model for forecasting facilities investment using macroeconomic variables only to that of the model including the Bank of Korea's NSI and Google Trends Data in addition to various macroeconomic variables. We employ Machine Learning-based models such as Adaptive LASSO, Ridge Regression, Random Forest, and Neural Network to compare predictability for facilities investment. The estimation results reveal that predictability is improved with the model

• Received: 12 July 2023, Revised: 15 August 2023, Accepted: 6 September 2023.

* The views expressed in this study are those of the authors and should not be attributed to the Bank of Korea. The authors are grateful to two anonymous reviewers for their helpful comments and suggestions. Any remaining errors are the sole responsibility of the authors.

** Associate Professor, Department of Economics and Department of Quantitative Applied Economics, Sungkyunkwan University. E-mail: dooyeoncho@g.skku.edu

*** Corresponding author. Master in Economics, Department of Quantitative Applied Economics, Sungkyunkwan University, Economist at the Bank of Korea. E-mail: totop2500@g.skku.edu, bhwang95@bok.or.kr

www.kci.go.kr

using the Bank of Korea's NSI and Google Trends Data in addition to various macroeconomic variables compared to the model using macroeconomic variables only. Our study suggests that it would be possible to improve the accuracy of prediction through the use of news-related data and a web-based search trend besides structured data and be useful for policy implementation.

Key words: Facilities investment, Machine Learning, Google Trends Data, News sentiment index, Forecasting
JEL Code: B030104, B0301092