Korea Trade Review (Korea Trade Rev. ; 무역학회지)
Vol. 46, No. 2, April 2021 (pp. 281-299)
https://data.doi.or.kr/10.22659/KTRA.2021.46.2.281

貿易學會誌
ISSN 1226-2765 (Print)

투자와 수출 및 환율의 고용에 대한 의사결정 나무, 랜덤 포레스트와 그래디언트 부스팅 머신러닝 모형 예측

이재득
부산대학교 무역학부 교수

Investment, Export, and Exchange Rate on Prediction of Employment with Decision Tree, Random Forest, and Gradient Boosting Machine Learning Models

Chae-Deug Yi[a]
[a]Department of International Trade, Pusan National University, South Korea

Received 10 April 2021, Revised 23 April 2021, Accepted 28 April 2021

Abstract

This paper analyzes the feasibility of using machine learning methods to forecast the employment. The machine learning methods, such as decision tree, artificial neural network, and ensemble models such as random forest and gradient boosting regression tree were used to forecast the employment in Busan regional economy. The following were the main findings of the comparison of their predictive abilities. First, the forecasting power of machine learning methods can predict the employment well. Second, the forecasting values for the employment by decision tree models appeared somewhat differently according to the depth of decision trees. Third, the predictive power of artificial neural network model, however, does not show the high predictive power. Fourth, the ensemble models such as random forest and gradient boosting regression tree model show the higher predictive power. Thus, since the machine learning method can accurately predict the employment, we need to improve the accuracy of forecasting employment with the use of machine learning methods.

Keywords: Industry Investment, Export, Exchange Rate, Machine Learning, Ensemble Learning
JEL Classifications: F10, F13

[a] E-mail: givethanks@pusan.ac.kr

Ⅰ. 서론

최근 세계는 인공지능(Artificial Intelligence: AI) 시대를 맞이하여 여러 분야에서 이러한 AI 기능을 적용하고 산업과 금융 그리고 물류 분야 등에서도 이것을 응용하여 집중적으로 성장시키려고 하고 있다. 그리하여 우리나라 정부뿐만 아니라 각국 정부들은 자국 경제의 성장과 발전을 위해 AI 산업과 AI 기능을 모든 분야에서 확대하고 발전시키려고 노력하고 있다. AI기능을 장착한 기계가 인간과 같은 지능을 갖추고 자동화와 정보통신 기술의 발달로 많은 수작업들을 대체하고 있다. AI의 산업분야에서의 활용증가로 인하여 고용측면에서 본다면 일반 노동자들의 직업, 작업 등을 대체함으로써 고용을 줄일 수 있는 반면, 다른 한편으로 새로운 AI 관련 업무를 창출하여 새로운 고용을 창출할 수도 있다.

부산시 역시 지역경제의 활성화와 부산지역의 고용을 증가시키고 낙후된 부산지역 산업 및 경제구조의 고도화를 위해 많은 노력을 하고 있다. 그럼에도 불구하고 부산지역 산업들의 상대적인 침체로 인해 고용인구와 고용률의 전국대비 비율은 점점 감소되고 있는 실정이다. 그리하여 부산은 지역경제의 성장과 고용증대, 특히 청년들의 고용을 증대시키기 위하여 신성장산업들을 선정하고 지원하고 육성하기 위해 많은 노력을 하고 있다. 그러나 부산시에서 지원하고 육성하는 이러한 산업들의 성장과 지역고용에 미치는 영향에 대한 좀 더 정확한 진단과 예측이 필요하다. 그러나 기존의 대부분의 연구들은 주로 머신러닝 기법이나 앙상블 러닝 기법이 아닌 주로 전통적인 경제학적 모형이나 시계열 계량기법들을 사용하고 있다.

그러나 전통적 계량경제 모형이나 전통적 구조적 계량모형은 모형의 설정의 오류와 추정방법의 한계상, 예측오차가 많이 생길 수 있다. 왜냐하면 전통적 경제이론에 기반한 모형은 모형설정의 오류를 필연적으로 수반하고 있기 때문에 경제예측에 있어 과잉적합의 문제 등이 표본기간 밖의 예측에 있어 종종 발생하는 문제가 발생한다. 따라서 이들 전통적 모형의 설정에 의한 경제예측의 추정계수들은 일치추정량이 되지 않기 때문에 미래의 경제효과를 예측하는데 있어서 잘못된 추정으로 정책효과를 과장하거나 오도할 수도 있다.

그리하여 최근에는 외국에서는 이러한 전통적 계량경제학 추정을 보완하고 극복하기 위해 인공지능과 머신러닝(Machine Learning) 등의 기법을 도입하여 경제예측에서도 일부 학자들이 도입하여 사용하기 시작하고 있다. 특히 거시경제학적 분석과 지역경제의 예측 등과 같은 분야에서 최근 도입되어 활용되고 있는 실정이며, 향후 경제분석에서 활용영역은 좀 더 정확한 진단과 예측의 정확성을 위해 점점 더 넓어지게 될 것이다. 그러나 우리나라에서도 머신러닝에 의한 경제학적 분석은 거의 없고, 대부분 기존연구들은 머신러닝에 대한 동향이나 개괄적인 소개에 그치고 있을 뿐이다. 물론 단편적인 머신러닝 기법들이 물류, 항만 등에서 일부 사용되고 있으나, 우리나라뿐만 아니라 특히 지역경제분석에 있어 머신러닝 기법은 거의 소개도 안 되어 있는 실정이다. 특히 경제 진단과 예측 측면에서 다양한 머신러닝 기법에 의한 전문적이고 심층적으로 분석한 연구는 거의 없는 실정으로 연구가 극히 미흡한 실정이다.

따라서 본 연구에서는 최근 외국의 연구에서 사용하고 있는 새로운 접근법인 머신러닝 모형과 앙상블 모형(Ensemble Model) 기법들을 가지고 주요 산업들에 대한 투자지수들과 수출 그리고 환율 등을 가지고 부산지역의 고용에 대하여 추정과 예측을 해보고자 한다. 이와 같이 본 연구는 대부분의 기존 경제모형의 설정오류로 인한 과잉적합 문제를 해결하고 오차와 예측성을 높이기 위해서, 의사결정 나무(Decision Tree), 인공 신경망(Artificial Neural Networks ANN), 그리고 랜덤 포레스트(Random Forest)나 그래디언트 부스팅(Gradient Boosting) 모형 등의 앙상블 기법을 이용하여 부산지역의 고용을 중심으로 예측하고자 한다. 본 연구에서는 추정모형을 설정하여 비교하여 분석하기 위해 모형들의 결정계수들과 대표적인 예측 정확도에 대한 검증 방법인 평균제곱근오차(RMSE)를 이용한다.

이와 같이 이러한 머신러닝 기법과 앙상블 러닝 모형에 의한 경제학적 분석은 우리나라는

물론이고 경제분석 차원에서 아직 연구가 극히 미흡하고, 머신러닝과 앙상블 러닝 모형에 의한 새로운 접근법은 모형의 설정오류가 있고 과잉적합을 보이는 기존 구조적 경제모형 혹은 시계량기법에 의한 기존연구들과 차이점이 있다. 또한 본 연구를 통해 주요 산업들의 경제적 효과를 예측할 수 있기 때문에 산업의 지원과 육성을 하는 데도 정책적 함의를 가질 수 있다. 아울러 향후 새로운 접근법인 머신러닝과 앙상블 러닝 모형을 이용한 경제예측 기법 등을 사용하여 분석함으로써 이 분야에서 후속연구들을 유발하고 새로운 해석을 낳을 수도 있을 것이다.

Ⅱ. 문헌연구

부산지역의 고용에 대한 추정과 예측은 부산경제의 활성화를 위해서 매우 중요한 선결과제이다. 물론 전통적 계량방법에 의한 기존의 연구분석은 제법 있었으나 이러한 모형 들은 과적합 문제가 있으므로, 이를 극복하기 위해 좀 더 예측력이 높은 최근 머신러닝과 앙상블 러닝 기법을 도입하여 경제학 분야에서 추정하고 예측한 연구는 그 중요성에도 불구하고 거의 없는 실정이다.

따라서 외국의 머신러닝과 연관된 연구를 보면, Zou and Hastie (2005)는 회귀모형에 있어 과적합 문제 등을 해결하기 위하여 정규화와 변수선택에 대한 연구를 하였다. Chakraborty and Joseph (2017)는 중앙은행에서 머신러닝에 대한 연구를 하였다. Naecker and Peysakhovich (2017)은 리스크의 애매모호한 행동모델을 평가하기 위한 머신러닝에 대한 연구를 하였다. Géron (2017)은 Scikit 학습과 머신러닝 그리고 텐서플로(Tensor Flow)에 대한 연구를 하였다. Kreif and DiazOrdaz (2019)은 정책평가에 있어 과도적합 문제를 극복하기 위해 전통적 계량모형보다는 머신러닝 모형을 도입하여 연구하였다.

그리고 Schapire and Freund (2014)는 부스팅(Boosting)에 대한 알고리즘과 머신러닝 학습에 대한 연구를 하였다. Athey and Wager (2018)은 랜덤 포레스트를 이용한 이질적 효과에 대한 추정에 대해서 연구를 하였다. Athey et al. (2019)는 일반화된 랜덤 포레스트를 연구하였다. Agrawal, Gans and Goldfarb (2018)은 단순한 인공지능 경제학으로 머신의 예측에 대한 연구를 하였다. Athey (2017/2019)는 빅데이트 자료와 인공지능 경제학에 있어 머신러닝의 경제학에 대한 충격에 대한 연구를 하였다. Jean et al. (2016)은 빈곤을 예측하는 데 있어 머신러닝 기법을 사용하였다. Chalfin et al.(2018/2019)은 머신러닝에 의한 생산성과 인적자산에 관한 연구를 하였고, Gu, Kelly and Xiu (2019)은 머신러닝에 의한 자산가격 책정을 연구하였다. Mullainathan and Spiess (2017)은 응용계량적인 분석으로 머신러닝에 대한 연구를 하였다.

그러나 머신러닝 기법들은 경제학 측면에서는 아직 생소하고 초기적인 연구가 주를 이루고 있다. 특히 우리나라에서 비구조적 데이터를 이용한 텍스트 마이닝 기법을 이용한 연구들은 조금 있고, 머신러닝을 이용한 경제 분석에 대한 동향을 간략히 소개한 것은 있다. Kim Soo-Hyon (2020)은 우리나라의 환율이 단기적인 금융시장에 서의 변동에 대한 딥러닝 기법의 적용가능성을 연구하였고, Yi Chae-Deug (2021)은 부산을 중심으로 전략산업의 경제적 효과를 머신러닝을 사용하여 예측하였지만, 주요산업에 대한 앙상블 러닝 기법에 의한 분석은 하지 않았다. 따라서 이들 연구한 것을 제외하고는 머신러닝을 이용한 경제학적인 분석은 드물고, 더구나 우리나라에서 머신러닝과 앙상블 기법을 이용하여 심층적으로 실증적인 자료를 가지고 경제분석을 한 연구는 없는 실정이다.

이와 같이 외국에서는 머신러닝 등을 이용하여 경제학 분야에서 연구가 나오고 있으나 우리나라에서는 머신러닝 기법들과 이들 머신러닝과 딥러닝 기법 등을 융합한 앙상블 러닝 모형을 가지고 거시경제 변수인 고용 등에 대해서 경제예측을 한 연구분석은 당연히 거의 없기 때문에, 향후 이와 같은 머신러닝과 랜덤 포레스트이나 그래디언트 부스팅 모형 등의 앙상블 러닝에 의한 경제 예측분야 연구에 대하여 후발 연구의 디딤돌이 될 것이다. 본 연구에서는

이를 위해 머신러닝과 앙상블 러닝 모형 등에 의해 부산의 주요 산업들과 투자와 수출과 환율 등의 거시경제변수 등이 부산의 지역경제 특히 고용에 어떤 영향을 미치는지 추정하고자 한다.

Ⅲ. 머신러닝과 앙상블 예측

머신러닝은 1950년대 인공 신경망(ANN)이 출현하면서 발전을 시작하였으며, 1980년대 후반 이후에는 상대적으로 연구가 좀 침체하였으나, 2010년대 딥 러닝(deep learning) 방법의 본격적인 출현과 컴퓨터의 발달과 함께 다시 활발히 사용되고 있다. 특히 최근 모든 산업분야에서 적용되고 있는 AI에 힘입어 머신러닝 기법은 분류와 회귀분야에서 아주 좋은 성과를 나타내고 있다.

본 연구에서는 부산지역의 주요 산업들의 발전을 통한 고용에 대한 효과를 살펴보기 위하여 머신러닝 방법인 의사결정 나무(Decision Tree), 인공 신경망(ANN), 랜덤 포레스트(Random Forest)와 그래디언트 부스팅(Gradient Boosting) 앙상블 모형 등을 이용한다. 이하 본 연구에서 사용하는 머신러닝과 앙상블 모형에 의한 추정은 Hastie, Tibshirani and Fridman (2017) 등의 연구를 참조하여 다음과 같이 간단히 요약한다.

1. 의사결정 나무(Decision Tree) 모형

의사결정 나무(Decision Tree)는 데이터에 내재되어 있는 패턴을 계층적 구조로 이루어진 나무형태로 도표화하여 분할하는 모형으로 예측이 어떻게 이루어지는지 비교적 명확하게 알 수 있다. 의사결정 나무는 기본적으로 회귀나무의 경우 MSE값을, 분류는 엔트로피 등의 불순도(Impurity)를 최소로 하는 영역을 찾는 것이다.

불순도의 측정지수는 지니 지수(Gini Index)와 엔트로피 지수(Entrophy Index)를 많이 사용한다. 엔트로피는 얼마만큼 다양하고 불순한 정보가 내포되어 있는가를 측정하며 엔트로피가 클수록 불순도가 높아지고 분류가 힘들다. 지니 지수는 표본을 무작위로 선택할 경우 잘못된 클래스로 분류될 확률을 의미하며 0과 1 사이의 값을 가진다.

그리하여 특성변수와 해당 변수 각각의 분절점(cutpoint)의 모든 가능한 조합을 한꺼번에 고려하는 것은 어렵기 때문에 적극적인 알고리즘을 통하여 어떤 특정한 해당 단계에서만 가장 좋은 분할을 만들어내는 특성변수와 분절점의 조합을 선택한다. 회귀분류의 경우 각 영역 내에서 MSE가 최소가 되도록 이 과정을 반복하고 설명한다. 의사결정 나무는 다음의 앙상블 모형에서도 근간이 되므로 좀 자세히 아래에 설명한다.

1) 회귀나무(Regression Tree)

의사결정 나무 중 회귀나무(Regression Tree)는 먼저 N개의 관측치 각각에 대해 자료(data)가 p개의 투입물(input) x와 1개의 반응(response)물인 y로 구성, 즉 (x_i, y_i), i =1,2,..., N, $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$로 되어 있다고 상정하자. 그리고 변수들과 점들을 자동적으로 분리하고, 어떤 나무의 모양을 결정하는 알고리즘을 필요로 한다고 하자. 그 때, 만약 우리가 한 개의 분할을 M 지역 즉 R1, R2,..., RM, 으로 나눌 때, 각 지역에서 상수 cm을 상정하면 다음과 같은 반응(response)에 대한 식을 모형을 가진다[1].

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m).$$

이 때, 다음 잔차의 자승합(Residual Sum of Squares: RSS)을 최소화시키는 $\hat{c_m}$, 즉 영역 RM에서 y_i의 평균값(ave)를 구할 수 있다.

$$\hat{c_m} = ave(y_i | x_i \in R_m)$$
$$s.t \quad Min \sum_{i=1}^{p} (y_i - f(x_i))^2$$

1) Hastie, Tibshirani and Fridman (2017), pp.307-308.

만약 이항의 분할을 원한다면, 최소자승을 시키는 분할을 찾는 것이 용이하지 않다. 그러나 우리가 분할변수 j와 분할점을 가진다면 다음과 같은 반평면(half-plane)을 정의할 수 있다.

$$R_1(j,s) = X|X_j \le s,$$
$$R_2(j,s) = X|X_j \ge s.$$

이 때 다음 식을 최소화시키는 분할변수 j와 분할점 s를 찾을 수 있다.

$$\min(j,s)[\min(c_1)\sum_{x_i \in R_1(j,s)}(y_i - c_1)^2 + \min(c_1)\sum_{x_i \in R_1(j,s)}(y_i - c_1)^2].$$

그 때, j와 s에 대해, 위의 최소화 문제를 만족시키는 $\hat{c_1}$, $\hat{c_2}$의 값들은 다음과 같이 구할 수 있다.

$$\hat{c_1} = ave(y_i|x_i \in R_1(j,s)), \hat{c_2} = ave(y_i|x_i \in R_2(j,s)).$$

이 때 나무의 크기가 너무 크면 과도 적합성을 가지고 나무의 크기가 너무 작으면 중요한 성질을 놓치는 과소 적합성을 가진다. 따라서 모형의 복잡성을 조정하는 나무의 최적 크기를 결정하는 전략은 최소 노드 크기를 가진 큰 나무(T_0)의 부분집합인 나무 T($T \subset T_0$)를 구하는 것이다. 여기서 터미널 노드를 m, 영역 m을 R_m으로 나타내고, $|T|$를 T에서의 터미널 노드의 개수를 나타내고, 다음과 같이 정의한다.

$$N_m = Number[x_i \in R_m],$$
$$\hat{c_m} = \frac{1}{N_m}\sum_{x_i \in R_m} y_i,$$
$$Q_m(T) = \frac{1}{N_m}\sum_{x_i \in R_m}(y_i - \hat{c_m})^2,$$

여기서 $Q_m(T)$는 자승오차의 노드 불순도(squared-error node impurity)를 측정한다. 그리고 비용 복잡(cost complexity)의 기준을 다음과 같이 설정한다.

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|.$$

여기서 α는 모형의 복잡성을 조절하는 튜닝 모수(tuning Parameter)로서 나무의 크기와 자료의 적합성(goodness of fit) 간의 상충관계(tradeoff)를 조절한다. 그리하여 각 α에 대하여 $C_\alpha(T)$를 최소화시키는 하위 나무(subtree) $T_\alpha(T_\alpha \subseteq T_0)$를 구한다.

2) 분류나무(Classification Tree)

그러나 이러한 의사결정 나무 모형에 의한 지수는 나무의 크기에 따라 민감도가 낮기 때문에 실제에 있어서는 민감도가 높은 불순도의 측정 지수로 지니 지수(Gini Index)와 엔트로피 지수(Entrophy Index)를 사용한다. 먼저 노드 m, 영역 R_m, 그리고 N_m개의 관측치가 있을 때, 노드 m에서 k 클래스의 관측치들의 비율을 $\hat{p}_{mk}$($0 \le \hat{p}_{mk} \le 1$)라고 다음과 같이 설정하면, 지니 계수(G)와 엔트로피 지수(Entrophy Index)는 다음과 같이 정의된다.

$$\hat{p}_{mk} = \frac{1}{N_m}\sum_{x_i \in R_m} I(y_i = k),$$

지니 지수(Gini Index) ;

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

엔트로피 지수((Entrophy Index):

$$D = -\sum_{k=1}^{K} \hat{p}_{mk}\log(\hat{p}_{mk}).$$

여기서 $0 \le \hat{p}_{mk} \le 1$)이므로, 엔트로피 지수(D)≥ 0이다. 그리고 $\hat{p}_{mk}$가 모두 0 혹운

1에 가까이 있으면, 엔트로피 지수는 0에 가까울 것이다. 이와 같이 지니 지수와 엔트로피 지수는 모두 m차 노드에서 순수(pure)하다면 각각 작은 값을 가지게 된다.

2. 인공 신경망(Artificial Neural Network) 모형

인공 신경망(ANN)은 수많은 뉴런(Neuron)과 뉴런을 연결하는 시냅스(synapse)로 구성된 인간의 뇌 신경망을 노드(Node)와 링크(link)로 모형화한 네트워크로서, 뇌의 정보처리 및 전달 프로세스를 생물학적 신경계로 모형화하여 정보처리 및 추론과정을 모방하여 분류 또는 수치 예측을 수행한다.

그리하여 분류예측과 수치예측을 하기 위해 분류예측은 다층 퍼셉트론(Multi-layer Perceptron; MLP) 모형이 구현되는 MLP 분류함수, 수치예측을 위해서는 MLP 회귀함수를 이용한다. MLP 함수에서는 모형의 복잡도의 규제항인 알파(alpha)는 기본값 0.01로 설정하는데 알파를 높이면 규제강화가 일반화되고, 반대로 너무 높이면 과소적합의 문제가 발생할 가능성이 있다.

본 연구에서 사용되는 다층퍼셉트론은 입력층과 출력층 사이에 하나이상의 중간층이 존재하는 구조이다. 신호는 입력층 은닉층 출력층 방향으로 전달되며 은닉층과 출력층에서 입력층으로 직접적인 연결이 존재하지 않는다. 다층 퍼셉트론은 비선형곡선을 매우 정확하게 근사 시킬 수 있다. 신경망 모형은 가중치(weight)라 불리는 알려지지 않은 모수(parameter)를 가지고 있는데 학습데이터를 잘 적합하도록 하기 위해서는 그 값을 찾아내어야 한다.

3. 앙상블 학습 모형 (Ensemble Learning Model)

앙상블 학습모형은 하나의 모형이 아닌 여러 개의 모형을 학습시켜, 그 예측을 결합함으로써 보다 정확한 예측을 도출하는 일련의 머신러닝 기법이다. 이러한 방법을 앙상블 학습(Ensemble Learning)이라 한다. 그리하여 앙상블 학습모형은 의사결정 나무모형, 서포트 벡터 머신 모형, 그리고 인공 신경망 모형 등의 단일 모형의 예측 결과가 떨어질 때, 최근 모형의 정확성과 적합성을 높이기 위하여 이러한 단일 모형을 혼합을 시도하는 모형이다.

그리하여 본 연구에서는 부트스트랩(Bootstrap)과 어그리게이팅(Aggregating)을 결합한 배깅(Bagging) 모형인 랜덤 포레스트(Random Forest) 모형과 성능이 낮은 학습기를 여러 개 결합하여 성능이 강한 학습기를 만드는 기법으로 틀린 예측 데이터에 대해 여러 개의 모형을 순차적으로 학습해나가는 부스팅(Boosting) 모형인 경사하강법을 적용하는 그래디언트 부스팅(Gradient Boosting) 모형을 가지고 추정한다.

Ⅳ. 머신러닝과 앙상블 러닝에 의한 고용 예측

본 장에서는 2000년 1분기부터 2020년 2분기까지 부산의 고용자 혹은 취업자수(empn)가 그 중간값에 해당하는1,650(천명) 이하이면 0으로 두고, 그 취업자수가 만약 1,650(천명)을 초과하면 1로 나누어 영역을 0과 1로 나눈 명목변수를 종속변수로 삼는다. 그리고 독립변수들은 7개의 독립변수를 모두 로그값으로 변환하여 기계류 투자지수(lkm), 정밀기기 투자지수(lkmm), 운송장비 투자지수(lktrans), 부산의 경제활동인구(leactn), 한국의 투자 총지수(lkiv), 한국 원화의 대미 달러 환율(lwus), 그리고 한국의 수출(lex) 등을 선택한다. 이들 독립변수와 종속변수에 대한 모든 자료들은 우리나라 통계청 자료를 구하여 사용하였다.

1. 의사결정 나무(Decision Tree) 모형 예측

1) 의사결정 나무에 의한 분류 예측

본 절에서는 2000년 1분기부터 2020년 2분기까지 부산의 고용자 혹은 취업자수(empn)에

Table 1. The Accuracy of Classification Prediction with a Decision Tree Model

Model Accuracy of Decision Tree Classification	Accuracy of Training Data Set of Decision Tree Model : 982 Accuracy of Test Data Set of Decision Tree Model: 0.840			
Model Performance of Decision Tree Classification		precision	recall	f1-score
	0	0.88	0.88	0.88
	1	0.78	0.78	0.78
	accuracy			0.84
	macro avg	0.83	0.83	0.83
	weighted avg	0.84	0.84	0.84

대한 위의 독립변수들의 영향을 추정하기 위하여 먼저 의사결정 나무를 이용하여 분류예측을 하였다[2]. 이를 위해 머신러닝 모형을 분석하는데 가장 많이 사용하는 파이썬(Python)의 sklearn.tree 모듈과 의사결정분류(Decision Tree Classifier) 함수를 이용하여 의사결정 나무 모형을 생성하였고, 학습용 데이트 세트와 평가용 데이터 세트를 70%와 30%로 나누었고, 과잉적합된 모형을 피하기 위해 의사결정 나무의 사전 가지치기 옵션을 사용하여 나무의 최대 깊이(Max Depth)는 3으로 설정하였다. Yi (2021) 연구에서와 같이 불순도를 측정하기 위해서는 엔트로피 지수도 사용할 수 있겠지만, 본 절에서는 지니 지수를 선택하였다.

그리하여 의사결정 나무 분류 예측 모형을 평가하기 위하여 학습용 데이트 세트와 평가용 데이터 세트에 대한 정확도(accuracy)를 먼저 구하였다. 그 다음 모형의 정밀도(precision)와 재현율(recall), 그리고 정밀도와 재현율을 조화평균한 값인 F1-score를 구하였다. 정밀도(precision)는 모형이 참(true) 혹은 거짓(false)으로 예측한 것 중 실제 참 혹은 거짓이 얼마나 되는가를 판단한다. 재현율(recall)은 실제 참 혹은 거짓 데이터 중에서 참 혹은 거짓으로 분류한 비율을 나타낸다. 그리하여 의사결정 나무를 이용하여 모형의 정확도와 재현율, 그리고 F1-score는 〈Table 1〉에 나타나 있다.

먼저, 분기별 자료를 이용하여 구한 학습용 데이터 세트(Training Data Set)의 정확도는 0.982, 그리고 평가용 데이터 세트(Test Data Set)의 정확도는 0.840으로 나타나, 학습용 데이터 세트의 정확도는 평가용 데이터 세트의 정확도 보다 상대적으로 약간 높게 나와서 아주 약간의 과잉적합이 있을지도 모르지만 그 가능성은 낮다고 보여 진다. 그러나 의사결정 나무 모형의 평가를 나타내는 평가용 데이터 세트의 정확도는 상당히 높게 나아 학습모형은 좋게 나타났다.

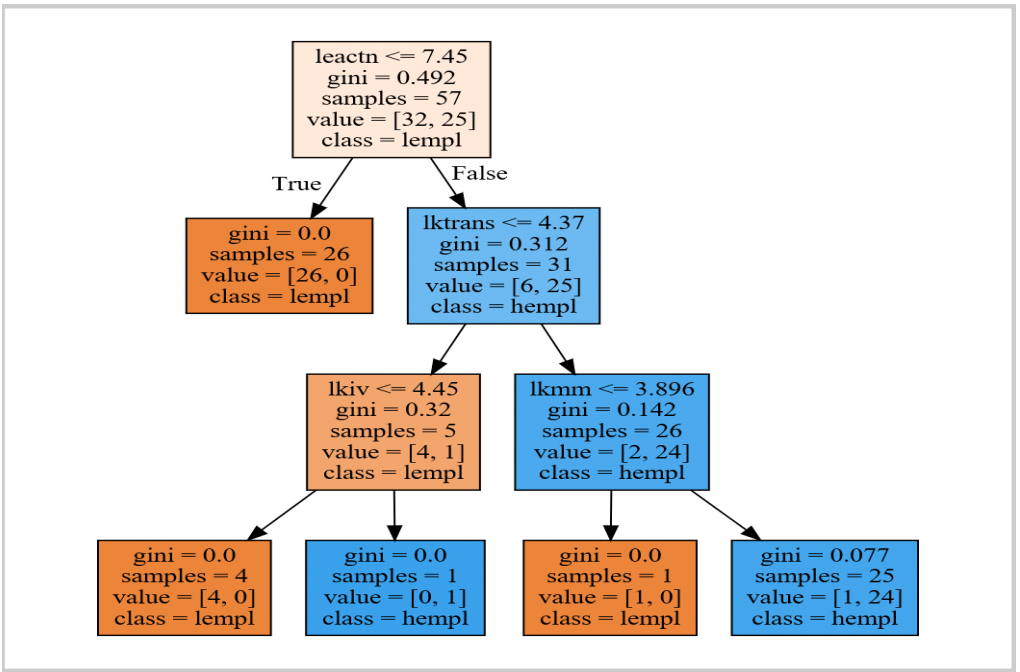
그리고, 의사결정 나무 모형의 성능 평가를 위해 분류 보고서에 있는 정밀도, 재현율, F1 스코어를 살펴보면 〈Table 1〉에서 나타난 것과 같이, 부산의 취업자수가 0인 영역과 1인 영역에서 정확도는 각각 0.88과 0.78로 아주 높게 나왔다. 재현율과 F1 스코어 같은 값으로 0과 1의 영역에서 모두 높게 나왔다. 클래스별 각 성과지표의 단순 평균값과 가중 평균값이 각각 정확도가 0.83과 0.84로 높게 나타나서 의사결정 나무 모형의 예측 성능은 좋은 것으로 나타났다.

이제 의사결정 나무 모형의 시각화를 위하여 파이썬 외부 라이브러리인 Graphviz를 사용하여 나무모양을 그린 것이 〈Fig. 1〉에 나타나 있다.

모든 변수들은 다 로그값으로 표시되어 있다. 이 그림을 해석해보면 먼저 뿌리마디에서 로그로 표시한 부산의 경제활동 인구(leactn)가 7.45보다 작은가 여부를 검사하여, 참(true)이면 취업자가 적은 부류(lempl)인 왼쪽 노드로,

2) Yi, C. and Y. Lee (2020, 참조

Fig. 1. The Classification Prediction with a Decision Tree Model



거짓(false)이면 취업자가 많은 부류(lempl)인 오른쪽 노드로 분류한다.

오른쪽 자식노드에서 또 운송설비투자 지수(lktrans)가 4.37보다 작은지 검사하여 참이면 왼쪽 노드로 거짓이면 오른쪽 노드로 분류한다. 왼쪽 노드에서 다시 한국의 투자지수가 4.45보다 작은지 검사하여 참이면 왼쪽, 거짓이면 오른쪽으로 분류한다. 오른쪽 노트에서 정밀기기 투자지수가 3.896보다 작은지 큰지 조사하여 참이면 왼쪽, 거짓이면 오른 쪽 노드로 분류한다. 이 의사결정 나무에서 끝마디로 갈수록 지니 지수가 0에 가까이 감으로써 불순도가 거의 0에 수렴하며 순수도가 100에 가깝게 나타나는 것을 알 수 있다.

2) 의사결정 나무 모형에 의한 고용 예측

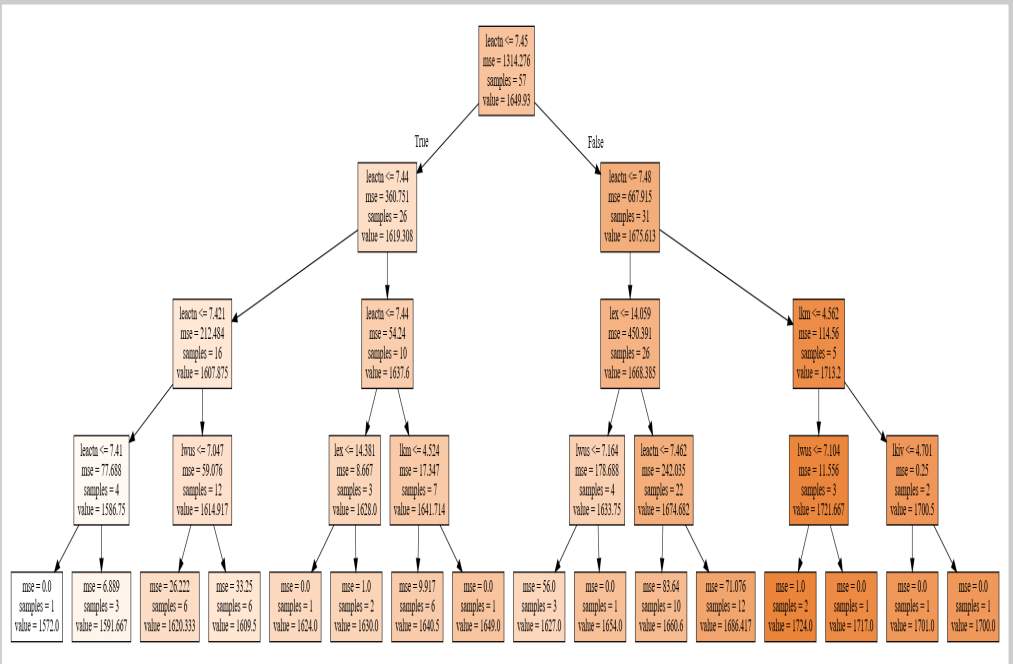
(1) 고용자수 예측

고용자수(empn)을 종속변수로 두고 앞에서와 같이 7개의 독립변수들은 모두 로그값으로 변환해서 부산의 고용자수 혹은 취업자수를 예측하는 의사결정 나무의 회귀모형을 만들어 보면, 다음 〈Fig. 2〉에서와 같이 나타난다.

본 절에서는 의사결정 나무의 사전 가지치기 옵션을 사용하여 나무의 최대 깊이(Max Depth)는 3으로 설정하는 것보다 4로 설정하는 모형이 더 좋아서 4로 설정하였다. 예측용 의사결정 나무 모형을 성능을 평가하기 위하여 결정계수와 MSE를 구한 결과, 학습용 데이터 세트의 결정계수는 0.969, 평가용 데이터 세트의 결정계수는 0.756으로 나타났으며, MSE도 15.799로 0.나타나 비교적 결정계수가 높은 좋은 모형을 나타났다.

한편, 〈Fig. 2〉에서 의사결정 나무는 나무의 깊이가 4로 되어 있는데, 앞에서와 해석과 논리는 같다. 먼저 경제활동 참가자의 로그값이 7.45보다 작으면 참이면 왼쪽으로 거짓이면 오른쪽노드로 분류하고, 오른쪽 노드에서 다시 경제활동인구가 7.48 이하 이면 왼쪽, 7.48보다 크면 오른쪽으로, 그 다음 기계류 투자지수

Fig. 2. The Employment Prediction with a Decision Tree Model



(lkm)가 4.562 이하이면 참으로 분류되어 왼쪽으로 가고, 그 이상이면 오른쪽 마디로 간다.

그 다음 우리나라의 투자지수(lkm)가 4.701 이하로서 참이면 왼쪽으로 가서 목표 예측 변수인 부산의 취업자수는 1,701(천명)으로 예측하게 되고, 4.701 보다 크면 거짓으로 오른쪽 마디로 가서 부산의 취업자수는 1,700(천명)으로 예측하게 된다. 마찬가지로 여러 조건들에 의해 참과 거짓으로 분류되어 분할되며 각 조건에 따라서 16개의 취업자에 대한 예측값이 1,572(천명)에서 1,724(천명)으로 나오게 된다. 그리고 〈Fig. 2〉에서와 나타난 것 같이 끝마디로 갈수록 MSE도 작아지고 있다.

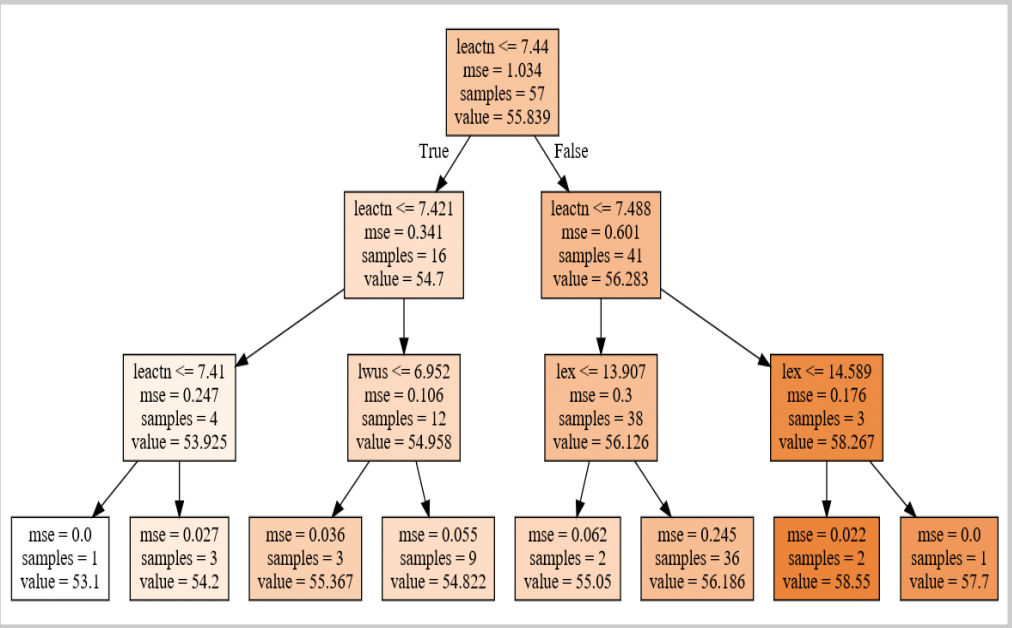
(2) 고용률 회귀 예측

이제 앞에서와 같이 7개의 독립변수들을 사용하여 부산의 고용률을 예측하는 의사결정 나무 모형을 만들어 본다. 역시 의사결정 나무의 사전 가지치기 옵션을 사용하여 나무의 최대 깊이(Max Depth)는 4로 설정하였다. 예측용 의사결정 나무 모형을 성능을 평가하기 위하여 결정계수와 MSE를 구한 결과, 〈Fig. 3〉에서와 같이 학습용 데이터 세트의 결정계수는 0.836, 평가용 데이터 세트의 결정계수는 0.615로 나타났으며, MSE도 0.604로 낮게 나타나 비교적 좋은 모형을 나타내고 있다.

이를 이용한 의사결정 나무가 〈Fig. 3〉에서와 나타나지만 몇 개의 경우만 분석해보면 다음과 같다. 첫째, 만약 경제활동 참가자의 로그값이 7.44보다 높으면 첫째 노드에서 거짓이 되어 오른쪽 노드로 내려간다. 오른쪽 노드에서 다시 경제활동인구가 7.488보다 크면 오른쪽노드로 내려가고, 그 다음 수출(lex)이 14.589 이하이면 참으로 분류되어 왼쪽 노드로 가서 고용률이 58.55%로 예측되고, 수출이 14.589보다 크면 오른쪽 노드로 가서 고용률은 57.7%로 예측된다.

둘째, 만약 제일 위 노드에서 경제활동 참가자의 로그값이 7.44 이하이면 오른쪽 노드로 분류되어 내려가는데, 오른쪽 노드에서 다시 경

Fig. 3. The Employment Rate Prediction with a Decision Tree Model



제활동인구가 7.488 이하이면 보다 크면 위의 경우와 달리 참이 되어 왼쪽 노드로 내려가고, 그 다음 수출(lex)이 13.907 이하이면 참으로 분류되어 왼쪽 노드로 가서 고용률이 58.05%로 예측되고, 수출이 13.907보다 크면 오른쪽 노드로 가서 고용률은 56.186%로 예측된다.

셋째, 만약 첫째 노드에서 경제활동 참가자의 로그값이 7.44보다 작아서 참이면 왼쪽 노드로 내려가고, 그 아래 노드에서 다시 경제활동인구가 7.421 이하 이면 왼쪽으로 가는데, 여기서 만약 경제활동인구가 7.421보다 크면 오른쪽으로 가는데, 그 다음 오른쪽 노드에서 원화의 대미 달러 환율의 로그값(lwus)이 6.952이하이면 고용률이 55.367%로 예측되고, 그 보다 환율이 높으면 오른쪽 노드로 가서 고용률은 54.822%로 예측된다.

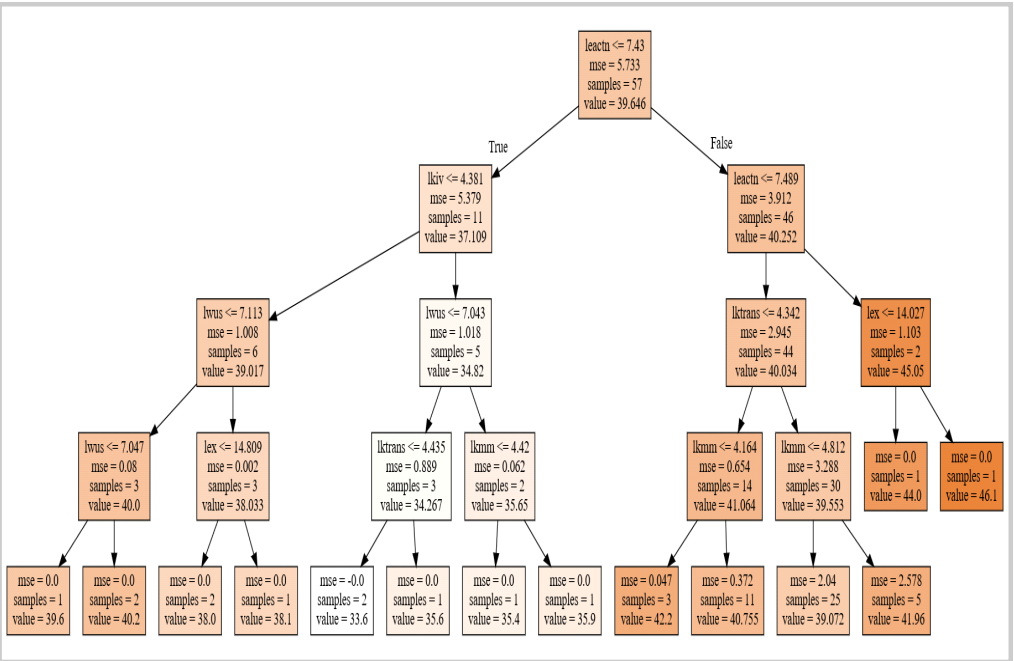
이와 같이 〈Fig. 3〉에서와 나타난 것 같이 의사결정 나무에서는 여러 조건들에 의해 참과 거짓으로 분류되어 분할되며 각 조건에 따라서 노드 16개의 고용률에 대한 예측값이 최하 53.1%에서 최대 58.55% 범위 안에서 다르게 나온다. 한편, 끝노드로 갈수록 MSE도 작아져서 예측오차가 줄어들고 있다.

(3) 청년 고용률 회귀예측

요즘 청년 고용률이 중요한 이슈이니, 청년 고용률(yempg)를 종속변수로 두고 앞에서와 같이 7개의 독립변수들을 사용하여 부산의 청년 고용률을 예측하기 위하여 의사결정 나무의 최대 깊이(Max Depth)가 4인 의사결정 나무 모형을 만들어 본다. 그리하여 예측용 의사결정 나무 모형을 성능을 평가하기 위하여 결정계수와 MSE를 구하였는데, 학습용 데이터 세트의 결정계수는 0.792, 평가용 데이터 세트의 결정계수는 0.496로 나타났으며, MSE는 1.978로 나타났다.

그리고 의사결정 나무의 모형에서 회귀예측의 결과가 다음 〈Fig. 4〉에서와 같이 나타났다. 첫째, 먼저 제일 위 노드에서 경제활동 참가자의 로그값이 7.43보다 크면 거짓이 되어 오른쪽 노드로 분류되어 내려간다. 오른쪽 노드에서 다시 경제활동인구가 7.489보다 크면 오른쪽

Fig. 4. The Youth Employment Rate Prediction with a Decision Tree Model



노드로 다시 내려가서, 그 다음 노드에서 로그 수출액(lex)이 14.027 이하이면 참으로 분류되어 부산의 청년 고용률에 대한 예측값은 44.0%이 되고, 수출액이 그 보다 작으면 오른쪽 마디로 내려가서 부산의 청년 로그 고용률에 대한 예측은 46.1%가 된다.

둘째, 제일 위 노드에서 부산의 로그 경제활동율이 7.43보다 작고 또 다시 두 번째 노드에서 경제활동률이 7.489보다 작을 때, 세 번째 노드로 내려가서 운송장비 투자의 로그값(lktrans)이 4.342 이하이면 왼쪽 노드, 크면 오른쪽 노드로 내려간다. 왼쪽노드에서 다시 정밀기기 투자지수(lkmm)가 4.164 이하이면 왼쪽노드로 내려가서 청년 고용률은 42.2%로 예측되고 정밀투자지수가 4.164보다 크면 오른쪽 노드로 내려가서 청년 고용률은 40.75%가 된다. 만약 위의 세 번째 노드에서 운송장비 투자의 로그값(lktrans)이 4.342보다 크면 오른쪽 노드로 내려가서, 다시 정밀기기 투자지수(lkmm)가 4.812 이하이면 왼쪽노드로 내려가서 청년 고용률은 39.072%로 예측되고 정밀투자지수가 4.812보다 크면 오른쪽 노드로 내려가서 청년 고용률은 41.96%가 된다.

이와 같이 부산의 청년 고용률도 〈Fig. 4〉에서와 나타난 것 같이 의사결정 나무에서는 여러 조건들에 의해 참과 거짓으로 분류되어 분할되며, 각 선행 조건에 따라서 노드 16개의 로그 고용률에 대한 예측값이 최하 33.6%에서 최대 42.2% 범위 안에서 약간씩 다르게 나온다.

2. 인공 신경망(ANN) 모형의 고용 예측

인공 신경망(ANN)을 이용한 모형을 만들어 분류예측과 수치예측을 하기 위해서는 분류예측은 다층 퍼셉트론(Multi-layer Perceptron, MLP) 모형이 구현되는 MLP 분류함수, 수치예측을 위해서는 MLP 회귀함수를 이용한다. MLP 함수에서는 모형의 복잡도의 규제항인 알파(alpha)는 기본값 0.01로 설정하는데 알파를 높이면 규제강화 일반화되고, 반대로 너무 높이면 과소적합의 문제가 발생할 가능성이 있다.

Table 2. Predictions of Employment with Artificial Neutral Network Models

Accuracy of Classification Prediction of ANN Model	MLP Function	alpha=1, hidden layer size of ANN =[20]			
	ANN Model Accuracy	ANN Training Data Set : 0.930 ANN Test Data Set : 0.840			
	Model Performance Classification of Artificial Neutral Network		precision	recall	f1-score
		0	0.88	0.88	0.88
		1	0.78	0.78	0.78
		accuracy(ANN)			0.84
		macro avg(ANN)	0.83	0.83	0.83
		weighted avg(ANN)	0.84	0.84	0.84
ANN Regression Prediction	MLP Function	alpha=1, max_iter=200, hidden layer size of ANN =[30. 30]			
	ANN Model Accuracy	ANN Training Data Set : 0.603 ANN Test Data Set : 0.567 ANN RMSE : 0.471			
ANN Regression Prediction	MLP Function	alpha=1, max_iter=200, hidden layer size of ANN=[30. 30]			
	ANN Model Accuracy	ANN Training Data Set : 0.737 ANN Test Data Set : 0.560 ANN RMSE : 0.319			
ANN Regression Prediction	MLP Function	alpha=1, max_iter=200, hidden layer size of ANN=[20. 20]			
	ANN Model Accuracy	ANN Training Data Set : 0.774 ANN Test Data Set : 0.568 ANN RMSE : 0.315			

은닉층(Hidden Layer)의 크기(size)는 은닉층의 개수(n)과 은닉노드 개수(h)에 의해 정해지는데, 본 절에서 은닉층의 크기(hidden layer size)=[20. 20]으로 은닉층은 2개, 은닉노드는 20개를 상정하였고, 반복회수(max_iter)는 모형이 최적화되는 최대 반복회수로 본 절에서는 1,000으로 상정하고 인공 신경망 추정 결과를 구하였다. 본 절에서는 이러한 MLP 함수의 알파, 반복회수, 은닉층의 개수와 은닉노드 개수를 달리 해서 인공 신경망을 이용하여 인공 신경망 모형에 의한 예측을 하였는데, 그 추정결과는 〈Table 2〉와 같다.

〈Table 2〉에서는 부산의 2000년 1사 분기부터 2020년 2사 분기까지의 분기별 자료를 이용하여 부산의 고용자수의 영역을 2개로 나누고, 부산의 고용자수가 그 중앙값인 1,650(천명) 이하이면 0으로 두고, 그 취업자수가 만약 1,650(천명)을 초과하면 1로 나누어 영역을 0과 1로 나눈 명목변수를 종속변수로 삼는다. 그에 대한 인공 신경망에 의한 분류예측의 모형의 정확도를 살펴보면, 알파(alpha)=1, 은닉층의 크기(hidden layer size)=[20] 일 때, 학습용 데이터 세트 정확도는 0.930, 평가용 데이터 세트 정확도는 0.840으로 나타났으나 역시 모형의 과잉적합 문제 가능성이 조금 있다. 모형의 성능 평가 분류보고서를 보면, 정밀도, 재현율, F1 스코어가 취업자수의 영역이 낮은 0의 영역에서는 모두 0.88, 높은 1의 영역에서는 정밀도, 재현율, F1 스코어가 모두 0.78로 나타났다.

그리고 고용자수에 대한 수치 예측 결과를 보면, 먼저 알파(alpha)=1, 반복회수(max_iter)=200, 은닉층의 크기(hidden layer size)=[30.

30]으로 은닉층은 2개, 은닉노드는 30개를 상정하였을 때, 학습용 데이터 세트 결정계수가 0.603, 평가용 데이터 세트 결정계수는 0.567로 각각 나타났고 RMSE는 0.471로 도출되어 나타났다.

본 연구에서 만약 알파(alpha)=1, 반복회수(max_iter) =200, 은닉층의 크기(hidden layer size)=[20. 20]으로 은닉층은 2개, 은닉노드는 20개를 상정하였을 때, 학습용 데이터 세트 결정계수가 0.737, 평가용 데이터 세트 결정계수는 0.560로 각각 나타났고 RMSE는 0.319로 도출되어 나타났다. 마지막으로, 알파(alpha)=1, 반복회수(max_iter)=300, 은닉층의 크기(hidden layer size)=[20. 20]으로 은닉층은 2개, 은닉노드는 20개를 상정하였을 때, 학습용 데이터 세트 결정계수가 0.774, 평가용 데이터 세트 결정계수는 0.568로 각각 나타났고 RMSE는 0.315로 도출되어 나타났다.

이와 같이 인공 신경망에 의한 회귀예측 모델은 다층 퍼셉트론(MLP) 모형에 따라 모형의 적합성과 정확성이 다르게 나타났지만, 대체로 서포트 벡터 머신 모형에 의한 예측모형의 평가가 낮게 나타난 것을 알 수 있다. 그리하여 인공신경망에 이한 모형과 서포트 벡터 머신 모형에 의한 예측의 정확성 때문에 최근에는 은닉층을 굉장히 증가시킨 딥러닝(Deep Learning) 모형 등을 도입하고 있거나, 의사결정 나무 모형이나 서포트 벡터 머신 모형 등을 혼합한 앙상블 학습모형 등을 도입하여 분석하고 있다.

3. 심층 신경망 모형(DNN)의 고용 예측

딥러닝 모형은 인공 신경망에서 은닉층의 개수를 여러 개 중첩해서 더 많은 계층으로 인공 신경망을 학습하는 모형으로 컴퓨터가 스스로 자가학습을 통해 데이트를 입력한다. 본 절에서는 딥러닝 모형의 심층 신경망(Deep Neural Network : DNN)에 의해서 고용을 예측한다. 그리하여 부산의 2000년 1사 분기부터 2020년 2사 분기까지의 분기별 자료를 이용하여 부산의 고용자 수(empn)가 그 중간값인 1,650천명보다 많으면 1, 적으면 0으로 하는 종속변수로 두고 앞에서와 같이 주요한 7개의 독립변수들로 선택한다.

그리고 학습반복 횟수인 에포크(epochs)와 한 번에 학습되는 데이터의 개수인 배치 크기(batch_size)에 따라 학습용 정확도(train acc)와 검증용 정확도(val acc) 2개의 그래프와 학습용 손실과 검증용 손실을 나타내는 2개의 그래프, 총 4개의 그래프들을 그리기 위해 X축에 에포크, 좌측의 Y축에는 손실(loss), 우측의 Y축에는 정확도(accuracy)를 나타내도록 한다. 그 예측 결과를 에포크와 배치 크기에 따라 학습용과 평가용 세트의 오차와 정확도를 〈Fig. 5〉와 함께 나타내면 다음과 같이 요약된다.

본 절에서는 심층 신경망 모형의 평가를 위해 x 축에 학습 횟수(epoch), Y 축에 오차의 정확도를 갖는 2개의 그래프를 학습용 데이터와 검증용 데이터(validation data)에 각각 적용하여 총 4개의 그래프를 그려 학습과정의 시각화를 가져온다. 그리하여 검증용 데이터는 학습용 데이터에서 20% 비율(validation_split=0.2)을 사용한다.

첫째, 심층신경망 모형 학습을 위해 먼저 에포크=100, 배치 크기=100으로 설정하여 구한 모형의 정확도 그래프를 제시한 후, 모형의 재학습을 통한 모형의 평가하기 위한 학습용 데이터 세트의 오차와 정확도는 각각 0.076, 0.982로 나타났고, 검증용 데이터 세트의 오차와 정확도는 각각 0.260, 0.920로 나타났다.

그러나 〈Fig. 5〉에서 나타난 것과 같이 에포크가 증가할수록 정확도와 오차가 급격히 변화하다가 학습반복 횟수가 20 정도를 넘어가면서 일정해지고 25 정도를 넘어가면서 오히려 학습데이터 세트와 검증용 데이터 세트의 성능이 역전되었다. 따라서 본 모형의 경우 100이라는 학습반복 횟수는 과잉적합 문제를 야기할 수 있으므로 20 전후로 반복횟수를 조정할 필요가 있다.

둘째, 위의 심층신경망 모형이 과잉적합되는 것을 방지하기 위하여 학습 반복 회수를 다시 조정한 모형 재학습을 과정을 통해 에포크=20, 배치 크기=64로 변경하여 구한 모형의 재학습을 통한 모형의 평가하기 위한 학습용 데이터

Fig. 5. Model Accuracy and Loss(epochs=100, batch size =100)

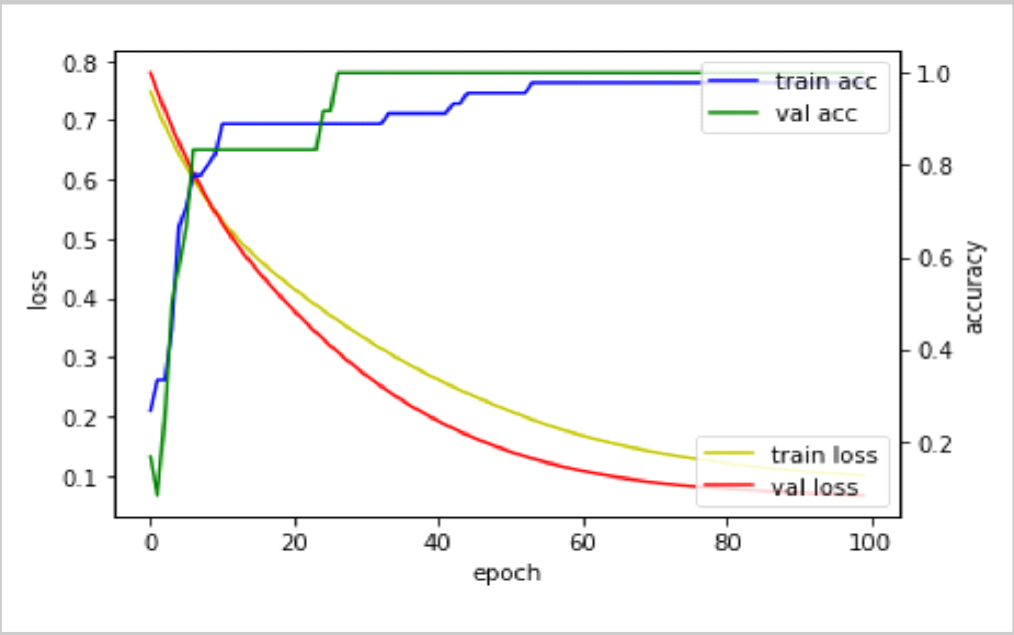
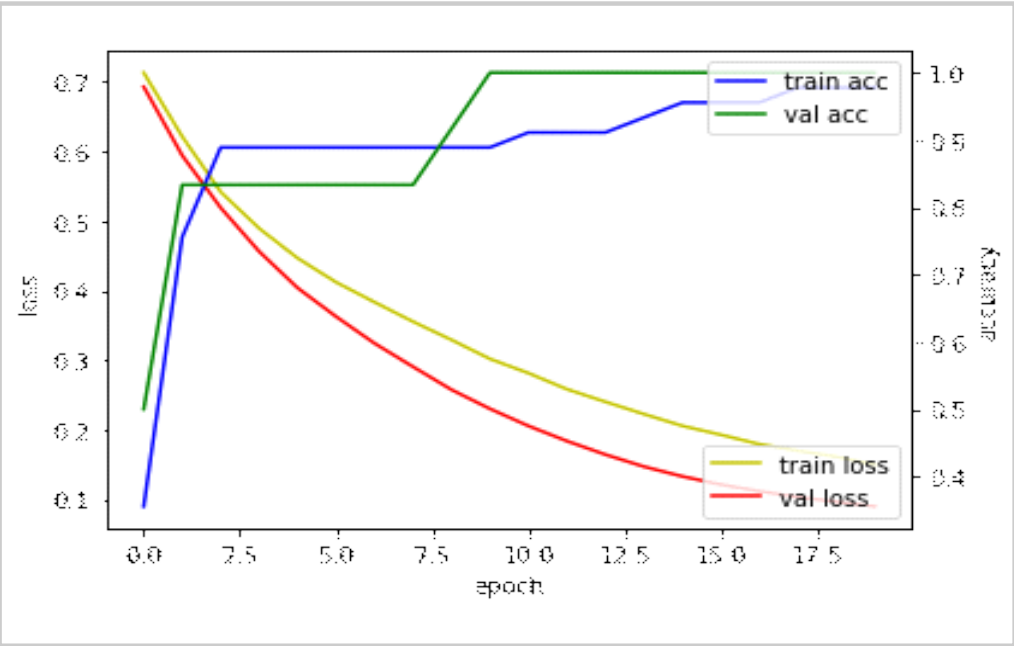


Fig. 6. Model Accuracy and Loss (epochs=20, batch size =64)



세트와 검증용 데이터 세트를 각각 사용하여 구한 오차와 정확도와 이를 이용하여 구한 〈Fig. 6〉은 다음과 같이 나타났다.

그 추정결과, 먼저 심층 신경망 모형에 의한 학습용 데이터 세트의 오차와 정확도는 각각 0.108, 0.982로 나타났고, 검증용 데이터 세트의 오차와 정확도는 각각 0.280, 0.920으로 나타났다.

마지막으로 부산의 고용에 대한 예측하기 위해 부산의 2000년 1사 분기부터 2020년 2사 분기까지의 분기별 자료를 이용하여 앞에서와 같이 고용을 종속변수로 삼고 주요한 기계류 산업들의 투자지수와 경제활동인구, 그리고 수출과 환율 등의 7 개의 변수들을 독립변수로 채택하고, 심층신경망에 의해서 부산의 고용자 수의 로그(lempn)을 종속변수로 두고 회귀 예측을 하였다.

먼저 앞의 모형 재학습과 분류예측에 의해 에포크=20, 배치 크기=64로 설정하였다. 그리고 심층신경망 모형 생성을 위해 1개의 입력층과 64개의 노드를 갖는 5개의 은닉층, 그리고 1개의 출력층을 설정하였다. 가중치는 MSE 손실함수를 기반으로 확률기반 경사하강법을 사용하였다. 그 추정결과, 부산의 고용에 대한 예측오차를 구해보면, 학습용 데이터 세트의 MSE = 2.286, 평가용 데이터 세트 MSE =2.464로 각각 나타났다.

4. 랜덤 포레스트(Random Forest) 앙상블 모형의 고용에 대한 예측

본 절에서는 먼저 데이터의 중복을 허용하는 무작위 샘플링 방식을 통한 부트스트랩과 각 모형이 예측한 값을 결합하는 어그리게이팅 기능을 결합한 배깅(Bagging) 앙상블 모형으로 랜덤 포레스트(Random Forest) 모형을 사용한다. 즉 랜덤 포레스트 모형은 단일 의사결정 나무가 아니라 여러 가지 단일 의사결정 나무들이 모인 숲(Forest)에서 이루어진 모형을 사용하여 예측과 추정을 하는 것으로 최근 점점 더 많이 사용하고 있다.

그리하여 본 절에서는 앞에서와 같이 랜덤 포레스트 모형의 분류 예측을 하기 위하여 먼저 2000년 1분기부터 2020년 2분기까지 부산의 고용자 혹은 취업자수가 그 중앙값인1,650(천명) 이하이면 0으로 두고, 그 취업자수가 만약 1,650(천명)을 초과하면 1로 나누어 영역을 0과 1로 나눈 명목변수를 종속변수로 삼는다.

랜덤 포레스트 회귀모형에 의한 고용에 대한 예측에서는 앞에서와 같이 7개의 독립변수들을 모두 로그값으로 변환하여 기계류 투자지수(lkm), 정밀기기 투자지수(lkmm), 운송장비 투자지수(lktrans), 부산의 경제활동인구(leactn), 한국의 투자 총지수(lkiv), 한국 원화의 대미 달러 환율(lwus), 그리고 한국의 수출(lex) 등을 선택한다. 여기서 개별나무의 개수(estimators)는 300으로 두었고, 나무의 최대 깊이(maximum deaph)는 2로 설정을 하였다. 그리하여 랜덤 포레스트 모형의 분류 예측과 수치예측을 도출한 결과는 다음 〈Table 3〉에 나와 있다.

그리하여 랜덤 포레스트 분류 모형에 대한 추정 결과, 학습용 데이터 세트의 정확도는 0.982, 평가용 데이터 세트 정확도는 0.840 나와 고용에 대한 랜덤 포레스트 모형의 정확성이 상당히 높게 나타났다. 그리고 모형 성능평가 분류 보고를 보면, 분류 모형의 정밀도가 0일 때 0.88, 1일 때 0.78로 나타났으며, 재현율과 F1 스코어 등의 값들도 동일하게 높게 나와서 모형의 성능이 비교적 상당히 좋게 나타난 평가를 보이고 있다.

랜덤 포레스트 회귀모형에 의한 고용에 대한 수치예측을 위해서 본 절에서는 종속변수를 고용자수로 두고 주요 산업에 대한 투자, 경제활동인구, 수출 및 환율 등의 7개의 독립변수들을 가지고 예측하기 위해 랜덤 포레스트 회귀함수를 이용하였다. 본 절에서는 랜덤 포레스트 회귀모형에서 인자는 개별나무의 개수는 100으로 채택하였고, 나무들의 최대 깊이=4로 두고 그 결과를 도출하였다.

그리고 랜덤 포레스트의 회귀예측의 추정결과, 〈Table 4〉에 나와 있듯이 학습용 데이터 세트 결정계수는 0.961, 평가용 데이터 세트 결정계수는 0.851으로 각각 비교적 높게 상당히 높게 나왔다. 한편 랜덤 포레스트 수치 예측 모형의 RMSE는 0.008로 나타났다. 그리하여 부산

Table 3. Predictions of Employment with Random Forest Ensemble Models

Accuracy of Classification Tree of Random Forest Model	Training Data Set Accuracy of Random Forest Model: 0.982 Test Data Set Accuracy of Random Forest Model : 0.840			
Random Forest Model: Number of Trees (Estimators) =300, Maximum Depth of Trees =2		precision	recall	f1-score
	0	0.88	0.88	0.88
	1	0.78	0.78	0.78
	accuracy(RF)			0.84
	macro avg(RF)	0.83	0.83	0.83
	weighted avg(RF)	0.84	0.84	0.84
Regression Tree Prediction of Random Forest Model	Coefficient of the Determination and RMSE of Random Forest Model			
Random Forest Model: Number of Trees (Estimators) = 100, Maximum Depth of Trees =4	Coefficient of the Determination of Training Data Set of Random Forest Model : 0.961, Coefficient of the Determination of Test Data Set of Random Forest Model : 0.851 RMSE : 0.008			

의 고용에 대한 주요 산업의 투자지수들과 수출 및 환율 등 7개의 독립변수들을 사용한 부산의 고용에 대한 예측이 상당히 좋은 모형으로 나타났다.

5. 그래디언트 부스팅(Gradient Boosting) 모형에 의한 예측

마지막으로 부스팅(Boosting) 앙상블 학습모형을 사용하는데, 부스팅 앙상블 학습모형은 여러 모형을 사용하여 순차적으로 앞에서 학습한 모형의 틀린 예측 데이터들을 고쳐나가는 방식을 사용하는데 본 절에서는 의사결정 나무를 사용하고 이전 모형의 예측 오류를 최소화해나가는 경사하강법(Gradient Descent)을 이용하는 그래디언트 부스팅(Gradient Boosting: GB) 방법을 사용한다.

본 절에서는 앞에서와 같이 2000년 1분기부터 2020년 2분기까지 부산의 고용자수를 예측하기 위하여 주요 산업에 대한 투자지수들과 경제활동인구, 수출 및 환율 등 7개의 독립변수들을 가지고 예측하기 위해 그래디언트 부스트 모형을 사용하였다.

먼저 그래디언트 부스팅 앙상블 모형의 분류 예측을 위해서 모형의 인자는 앞에서와 같이 개별나무의 개수는 300으로 두었고, 나무의 최대 깊이=2로 두고, 그 결과를 〈Table 4〉에서와 같이 도출하였다. 그리하여 첫째, 먼저 고용에 대한 그래디언트 부스팅 모형에 의한 분류예측을 보면, 그 추정 결과, 학습용 데이터 세트의 정확도는 거의 1에 가깝게 나타났고, 평가용 데이터 세트 정확도는 0.840 나와 고용에 대한 그래디언트 부스팅 모형의 정확성이 상당히 높게 나타났다.

그리고 그래디언트 부스팅 모형의 성능평가 분류를 보면, 분류 모형의 정밀도가 0일 때 0.93, 1일 때 0.73으로 나타났으며, 재현율도 각각 0.81과 0.89, 그리고 F1 스코어 등의 값들은 0.87과 0.80으로 높게 모형의 분류가 상당히 좋은 것으로 나타났으며 전반적인 정확성과 거시적 정확성, 가중된 정확성도 0.83-0.85로 상당히 높게 나타나 그래디언트 부스팅의 분류모형이 좋은 성능을 보이고 있다.

둘째, 고용에 대한 그래디언트 부스팅 모형에 의한 수치예측을 위하여는 종속변수를 부산

Table 4. Predictions of Employment with Gradient Boosting Ensemble Models

Accuracy of Classification Tree of Gradient Boosting Model	Training Data Set Accuracy of Gradient Boosting Model : 0.999 Test Data Set Accuracy of Gradient Boosting Model : 0.840
Gradient Boosting Model: Number of Trees (Estimators) =300, Maximum Depth of Trees =2	precision recall f1-score 0 0.93 0.81 0.87 1 0.73 0.89 0.80 accuracy(GB) 0.84 macro avg(GB) 0.83 0.85 0.83 weighted avg(GB) 0.86 0.84 0.84
Regression Tree Prediction of Gradient Boosting Model	Coefficient of the Determination and RMSE of Gradient Boosting Model
Gradient Boosting Model: Number of Trees (Estimators) = 100, Maximum Depth of Trees =4 Learning Rate = 0.1	Coefficient of the Determination of Training Data Set of Gradient Boosting Model : 0.994 Coefficient of the Determination of Test Data Set of Gradient Boosting : 0.761 RMSE : 0.010

의 취업자수로 두고 위의 7개의 독립변수 들 즉 부산의 주요 산업들인 기계류 산업들에 대한 투자지수들과 경제활동인구, 환율과 수출 등의 변수들을 가지고 그래디언트 부스팅 회귀함수를 이용하여 예측하였다. 본 절에서 그래디언트 부스팅 모형의 인자는 개별나무의 개수는 100으로 채택하였고, 나무들의 최대 깊이=4로 두고, 학습률은 = 0.1로 두고 그 결과를 도출하였다.

그리하여 그래디언트 부스팅 모형의 회귀예측의 추정결과는 〈Table 4〉에 나타나 있듯이, 학습용 데이터 세트 결정계수는 0.994로 아주 높게 나타났고, 평가용 데이터 세트 결정계수는 0.761로 비교적 높게 나왔고, 랜덤 포레스트 수치 예측 모형의 RMSE는 0.010으로 나타났다. 위에서 평가용 데이트 세트의 결정계수는 약간 낮게 나와 고용에 대한 과잉 적합의 문제가 발생할 가능성이 있지만, 학습용 데이트 세트의 결정계수가 아주 높게 나왔으므로, 이러한 7개의 독립변수들 즉 주요 산업들의 투자와 환율과 수출 등의 경제변수들에 의한 고용수준을 예측하는 데 있어 그래디언트 부스팅 앙상블 모형에 의한 예측력은 아주 높게 나타난 것을 알 수 있다.

Ⅴ. 결언

부산시는 지역경제의 활성화와 지역산업에서의 고용과 청년 일자리를 증가시키기 위해 지역의 전략산업을 비롯한 주요 산업들을 육성하기 위해 많은 노력을 하고 있다. 부산시는 최근 수년간 이러한 지역경제의 활성화 노력과 함께 산업을 지원육성하기 위해 이들 산업들에 금융지원과 투자를 늘이고 있는 형편이다. 본 연구는 부산의 주요 산업들의 부산경제에 대한 효과를 예측하여 분석한다는 것은 중요한 과제이다.

그러나 기존 연구들이 주로 사용한 전통적 계량경제 모형들에 의한 분석에 있어서는 모형설정 등의 오류로 인해 과잉적합 문제가 발생하여 잘못된 추정과 예측을 종종 할 수 있다. 최근에는 외국에서는 머신러닝 기법을 이용한 경제예측을 본격적으로 도입되기 시작하고 있

지만, 우리나라에서는 아직 머신러닝과 특히 앙상블 러닝을 이용한 경제예측과 분석은 거의 없으나 점차 확대되어 갈 것이다.

따라서 본 연구는 이러한 전통적 추정과 예측의 한계점을 보완하고 극복하기 위하여 주요 산업에 대한 투자와 수출 및 환율 등의 고용에 대한 경제적 효과를 예측하기 위하여 머신러닝과 앙상블 러닝 기법 등을 적용하였다. 이를 위해 모든 변수들의 분기별 자료를 사용하여 부산의 고용에 대한 예측을 하였는데, 그 결과를 종합해서 요약해보면 다음과 같다.

첫째, 의사결정 나무의 모형에 의한 주요 산업투자들의 부산의 고용에 대한 분류 예측결과를 보면, 학습용 자료 세트의 정확성이 아주 높게 나타났으며 평가용 자료 세트의 정확성도 비교적 높게 나타났다.

둘째, 의사결정 나무의 회귀예측 모형은 의사결정 나무들의 깊이에 따라서 고용자수와 고용률, 그리고 청년 고용률에 대한 예측값이 조금씩 다르게 나타났으나, 끝마디로 갈수록 불순도도 낮아지고 MSE도 0으로 수렴하는 등 작아지고 있는 것으로 나타났다.

셋째, 인공 신경망 모형에 의한 고용에 대한 예측모형의 결과를 보면, 학습용 데이터 세트와 평가용 데이터 세트 결정계수가 비교적 낮게 나왔으며, 과잉적합의 문제가 발생할 가능성이 높게 나타났다. 그리고 RMSE도 비교적 높아 고용과 소득에 대한 인공 신경망에 의한 모형의 의한 고용 예측력은 높지 않은 것으로 나타났다. 그러나 심층 신경망에 의한 딥러닝 모형으로 예측오차는 작아지고 정확성은 높아지고 있다.

넷째, 랜덤 포레스트 모형이나 그래디언트 부스팅 모델 등의 앙상블 러닝 기법으로 부산의 고용에 대한 추정에서 앙상블 모형이 서포트 벡터 머신 회귀모형 혹은 인공 신경망 회귀모형 등 보다 예측력과 결정계수가 더 높게 나타났고 모형의 정확성도 높게 나타나, 앙상블 러닝 기법에 의한 고용을 예측해볼 필요가 있다.

그리하여 본 연구에서 도입하여 분석한 예측결과를 보면, 의사결정 나무 모형, 그리고 랜덤 포레스트와 그래디언트 부스팅 앙상블 모형 등이 각 모형의 결정계수와 RMSE 등의 기준으로 살펴볼 때 예측력이 좋게 나타났다.

그러나 이러한 머신러닝 모형 혹은 앙상블 모형 기법에 대한 예측력이 모형의 인자에 따라 예측력이 다양하게 나타났다. 그러므로 지역경제 분석에 있어 좀 더 정확한 분석을 위해서 더 엄밀한 머신러닝 모형들과 좀 더 적합한 여러 가지 인자들을 선택하는 것이 중요하다.

그리고 기존의 전통적 계량분석에 의한 경제예측의 과도적합성을 보완하는 측면에서도 경제분석과 예측을 위하여 향후 이러한 머신러닝과 이를 융합한 앙상블 러닝 기법을 도입할 필요가 있다. 그러나 본 연구는 2000년 1분기부터 2020년 2분기까지의 자료를 이용하였으로 표본자료가 충분히 많다고 할 수는 없다. 향후 좀 더 긴 기간과 좀 더 많은 산업과 전국의 각 지역에 대한 더 많은 패널 경제자료들을 수집하여 머신러닝과 앙상블 러닝 기법에 의한 분석을 한다면 좀 더 엄밀한 분석결과와 해석을 할 수 있을 것이지만, 아직 생소한 분야라 후속 연구와 다음 과제로 남긴다.

References

Agrawal, A., J. Gans and A. Goldfarb (2018), *Prediction Machines: The Simple Economics of Artificial Intelligence*, Brighton, MA: Harvard Business Review Press.

Athey, S. (2017), “Beyond Prediction: Using Big Data for Policy Problems”, *Science*, 355(6324), 483-485.

Athey, S. (2019), "*The Impact of Machine Learning on Economics*", The Economics of Artificial Intelligence: An Agenda (1st ed.), Chicago, IL: University of Chicago Press, 507-547.

Athey, S. and S. Wager (2018),"Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests", *Journal of the American Satistical Association*, 113(523), 1228-1242.

Athey, S., J. Tibshirani and S. Wager (2019), "Generalized Random Forests", *Annals of Statistics*, 47(2), 1148-1178.

Chalfin, A., O. Danieli, A. Hillis, Z. Jelveh, M. Luca, J. Ludwig et al. (2016) "Productivity and Selection of Human Capital with Machine Learning", *American Economic Review*, 106(5), 124-27.

Chakraborty, C. and A. Joseph (2017), *Machine Learning at Central Banks* (Bank of England Working Paper, No. 674), London: Bank of England, 1-89.

Géron, Aurélien (2017), *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (1st ed.), CA : O'Reilly Media.

Gu, Shihao, Bryan Kelly and Da-Cheng Xiu (2019), *Empirical Asset Pricing via Machine Learning* (NBER Working Paper No. 25398), Cambridge, MA: National Bureau of Economic Research, 1-80.

Hastie, T., R. Tibshirani and J. Fridman (2017), *The Elements of Statistical Learning* (2nd ed.), Berlin: Springer.

Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell and S. Ermon (2016), "Combining Satellite Imagery and Machine Learning to Predict Poverty", *Science*, 353(6301), 790-794.

Kim, Soo-Hyon (2020), *Macroeconomic and Financial Market Analyses and Predictions through Deep Learning* (BOK Working Paper, No. 2020-18), Seoul: Bank of Korea.

Kreif, Noémi and Karla DiazOrdaz (2019), Machine Learning in Policy Evaluation: New Tools for Causal Inference (Oxford Research Encyclopedia of Economics and Finance), Oxford, England: Oxford University Press.

Mullainathan, S. and Jann Spiess (2017), "Machine Learning: An Applied Econometric Approach", *Journal of Economic Perspectives*, 31(2), 87-106.

Naecker, Jeffrey and Alexander Peysakhovich (2017), "Using Methods from Machine Learning to Evaluate Behavioral Models of Choice under Risk and Ambiguity", *Journal of Economic Behavior & Organization*, 133, 373-384.

Schapire, Robert E. and Yoav Freund (2014), *Boosting: Foundations and Algorithms, Adaptive Computation and Machine Learning Series* (2nd ed.), Cambridge, MA: The MIT Press.

Yi, Chae-Deug (2021), "Machine Learning and Deep Learning Models to Predict Income and Employment with Busan's Strategic Industry and Export", *Korea Trade Review*, 46(1), 169-187.

Yi, Chae-Deug and Y. Lee (2020), *Busan's Economy Prediction and Strategic Industry Using Machine Learning in Artificial Intelligence*, Busan.: Bank of Korea.

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