Korea Trade Review (Korea Trade Rev. ; 무역학회지)
Vol. 46, No. 1, February 2021 (pp. 169-187)
https://data.doi.or.kr/10.22659/KTRA.2021.46.1.169

貿易學會誌
ISSN 1226-2765 (Print)

머신러닝과 딥러닝 기법을 이용한 부산 전략산업과 수출에 의한 고용과 소득 예측

이재득
부산대학교 무역학부 교수

Machine Learning and Deep Learning Models to Predict Income and Employment with Busan's Strategic Industry and Export

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

Received 1 February 2021, Revised 25 February 2021, Accepted 26 February 2021

Abstract

This paper analyzes the feasibility of using machine learning and deep learning methods to forecast the income and employment using the strategic industries as well as investment, export, and exchange rates. The decision tree, artificial neural network, support vector machine, and deep learning models were used to forecast the income and employment in Busan. The following were the main findings of the comparison of their predictive abilities. First, the decision tree models predict the income and employment well. The forecasting values for the income and employment appeared somewhat differently according to the depth of decision trees and several conditions of strategic industries as well as investment, export, and exchange rates. Second, since the artificial neural network models show that the coefficients are somewhat low and RMSE are somewhat high, these models are not good forecasting the income and employment. Third, the support vector machine models show the high predictive power with the high coefficients of determination and low RMSE. Fourth, the deep neural network models show the higher predictive power with appropriate epochs and batch sizes. Thus, since the machine learning and deep learning models can predict the employment well, we need to adopt the machine learning and deep learning models to forecast the income and employment.

Keywords: Strategic Industry, Export, Machine Learning, Deep Learning
JEL Classifications: F10, F13

[a] E-mail: givethanks@pusan.ac.kr
Ⓒ 2021 The Korea Trade Research Institute. All rights reserved.

Ⅰ. 서론

부산시는 1980년대 이래 침체된 지역경제의 활성화와 경제구조의 고도화를 위해 2019년 부산의 지역경제를 성장시키고 지속가능한 미래를 창출하기 위한 제 5차 7대 전략산업을 새롭게 재선정하고 산업 경쟁력을 강화하기 위해 많은 노력을 하고 있다. 그리하여 부산은 현재 4차 산업혁명과 인공지능(AI) 시대에서 전략산업 활성화에 힘쓰고 있다. 부산시의 전략산업 육성의 목표는 결국 부산지역의 산업의 성장을 통해서 부산지역의 경제 즉 고용창출과 지역소득을 증가시키려고 하는 것이다. 이를 위해 여러 가지 전략산업의 육성과 경제정책을 수립하고 효율적으로 성취하려고 노력해오고 있다.

그러므로 부산지역 산업의 경쟁력 향상은 곧 부산지역의 고용과 경제성장으로 연결되며, 부산지역 산업의 경쟁력을 높이기 위한 중요한 요소로는 지역 전략산업 등의 부산지역 경제의 성장과 소득에 대한 예측이 필수적이다. 그러나 부산 지역경제 활성화와 지역산업의 경쟁력을 제고시키기 위해서는 먼저 부산경제와 산업 경쟁력에 대한 진단과 좀 더 정확한 예측과 추정이 필요하다.

물론 몇몇 기존 연구들이 있지만 주로 전통적인 구조적 모형이나 시계열 계량기법들을 사용하고 있다. 부산지역 경제에 대한 기존 연구들은 나름대로 공헌한 점이 있으나 전통적 구조적 모형은 모형설정의 오류가 있을 수 있으며, 전통적 시계열 계량경제인 기법 ARIMA 모델과 VAR 모형 등의 분석에 서는 과잉적합(overfitting) 등의 문제가 종종 생길 수 있다. 특히 최근 같이 세계경기가 불규칙하고 2020-2021년 현재 COVID-19 등의 팬데믹 질병 등으로 변동성이 큰 경우에는 예측력이 많이 떨어져, 중장기 예측에서는 정확도가 많이 낮아지는 한계점을 가진다. 그러므로 이들 전통적 모형에 의한 예측과 추정상 위험이 존재하여 종종 잘못된 추정과 예측을 하기도 한다.

따라서 이러한 전통적 추정과 예측의 한계점을 보완하고 극복하기 위하여 최근 인공지능과 머신러닝(Machine Learning) 등의 기법을 외국 일부분의 학자들이 경제분석에서도 처음으로 도입하여 시도해보고 있다. 우리나라에서도 머신러닝에 의한 경제학적 분석을 시도하고 있지만, 경제학적 분석에서 개괄적인 동향과 소개만 있을 뿐, 특히 경제 진단과 예측 측면에서 다양한 머신러닝 기법에 의한 전문적이고 심층적으로 분석한 연구는 없다. 물론 단편적인 머신러닝 기법들이 물류, 항만 등에서 일부 사용되고 있으나, 부산지역 경제분석에 있어 머신러닝 기법은 거의 소개도 안 되어 있고, 머신러닝에 의한 분석은 전혀 없는 실정이다.

그러나 AI시대 경제학과 지역경제 분석에서는 경제 정책 수요 예측 등과 같은 분야에서 향후 경제분석에서 좀 더 정확한 경제진단과 예측의 정확성을 위해 점점 더 넓어지게 될 것이다. 그러므로 부산지역의 전략산업의 활성화와 정책지원 및 경제효과를 위하서는 먼저 좀 더 엄밀한 추정과 예측이 필요하고, 그리고 이러한 분석을 통해서 지역경제 진단과 추정, 그리고 그에 따른 새로운 해석을 해본다는 것은 매우 중요한 연구과제이다.

따라서 본 연구는 인공지능의 시대에 걸맞게 부산에서 2019년에 선정된 7 개의 전략산업과 22개 세부 유망 산업분야 자료를 중심으로 부산경제를 머신러닝 분석기법들을 활용하여 다음과 같이 분석하고자 한다.

첫째, 본 연구는 여러 가지 머신러닝 모형들을 사용한다. 특히 본 연구는 비선형 머신러닝 기법들인 의사결정나무(Decision Tree), 서포트 벡터 머신(Support Vector Machine), 인공신경망(Artificial Neural Networks, ANN), 그리고 딥러닝(Deep Learning)의 심층 신경망(Deep Neural Networks, DNN) 등을 이용하여 부산경제 즉 부산의 고용 혹은 소득에 대해서 추정하고 예측하고자 한다.

둘째, 머신러닝 기법에 의한 수치 회귀 예측기법 등을 총 망라하여 부산의 전략산업을 중심으로 투자, 수출 및 환율 등에 대한 머신러닝과 딥러닝 기법에 의한 소득과 고용에 대한 예측모형을 통해 모형을 평가하고 실증적으로 분석한다.

셋째, 이러한 머신러닝 기법에 의한 지역경제 분석은 연구가 거의 없는 생소한 분야이다. 기존시계열 계량기법에 의한 개별 독립변수들

의 유효성에 중점을 둔 것과 완전히 달리 머신러닝과 딥러닝 모형 등은 개별 독립변수들의 영향이 아니라, 모든 독립변수들을 포함한 예측결과와 예측오류에 중점을 두고 있으므로, 기존연구들과 분석의 시각과 해석이 완전히 다른 차이점을 보이고 있다.

마지막으로 이러한 머신러닝을 이용한 연구는 해외에서는 최근 도입되고 있으나, 부산은 물론이고 우리나라에서도 머신러닝 기법들을 가지고 소득과 고용 등에 대한 경제예측을 한 분석은 없기 때문에 거의 최초의 연구가 될 것이다. 그리하여 본 연구는 향후 머신러닝과 딥러닝을 이용한 경제예측 기법 등을 사용하려는 후속연구들을 유발하는데 기여할 수 있을 것이다. 아울러 본 연구를 통해 향후 효과적이고 효율적인 전략산업에 대한 정책적인 함의를 제시하는데 도움이 될 것이다.

Ⅱ. 선행연구

인공 신경망을 활용한 기존 선행 연구를 살펴보면, 경제학 측면보다는 금융이나 물류 등의 분석이 많은 편이다. 물류 측면의 연구를 들면, Ding et al.(2019)의 연구에서는 2002년부터 2014년까지 중국 닝보항과 윈저우항의 연간 컨테이너 물동량 자료를 바탕으로 서포트 벡터 머신(Support Vector Machine)과 인공 신경망 모형, 서포트 벡터 머신과 인공 신경망을 결합한 추정 모형을 활용하였다.

외국의 인공지능과 머신러닝과 연관된 금융과 경제 관련 연구를 보면, 경제학적 관련 연구를 보면, Zou and Hastie (2005)는 정규화(Regularization)와 변수선택에 대한 연구를 하였다. Schlkopf and Smola (2001)는 서포트 벡터 머신과 정규화에 대한 커널연구를 하였다. Varian (2014)은 빅데이터와 계량경제에 대한 새로운 기법을 연구하였고, 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. Lopez de Prado (2018)은 머신러닝에 의한 금융을 연구하였다. Kreif and DiazOrdaz (2019)은 정책평가에 있어 머신러닝에 대한 연구를 하였다. Athey et al. (2019)는 일반화된 랜덤포레스트를 연구하였다.

Jean et al. (2016)은 빈곤을 예측하는 데 있어 머신러닝 기법을 사용하였다. Mullainathan and Spiess (2017)은 응용계량적인 분석으로 머신러닝에 대한 연구를 하였다. Gu et al. (2019)은 머신러닝에 의한 자산가격 책정을 연구하였다. Agrawal et al. (2018)은 단순한 인공지능 경제학으로 머신의 예측에 대한 연구를 하였다. Athey (2017, 2019)는 빅데이트 자료와 정책의 문제, 그리고 인공지능 경제학에 있어 머신러닝의 경제학에 대한 충격에 대한 소개를 하였다. Acemoglu and Restrepo (2020)는 미국 노동시장에 있어 로봇과 직업 대체에 관한 연구를 하였다.

그러나 머신러닝 기법들은 공학과 물류 등과는 달리 상대적으로 경제학 측면에서는 아직 생소하고 초기적인 연구가 주를 이루고 있다. 특히 우리나라에서 비구조적 데이터를 이용한 텍스트 마이닝 기법을 이용한 연구들은 조금 있으나, 머신러닝을 이용한 경제학적 분석은 거의 없는 실정이다. Kim Soo-Hyon (2020)은 거시경제와 금융시장 분석을 심층적인 머신러닝의 딥러닝 기법의 적용가능성을 연구하였다.

그러나 우리나라에서 머신러닝의 여러 가지 기법을 이용하여 지역경제 측면에서 분석한 것은 아직 없는 실정이다. 그리하여 본 연구는 2019년 선정된 부산의 7대 전략산업들이 부산의 고용과 소득에 어떤 영향을 미치는지 머신러닝 기법과 딥러닝 기법 들을 도입하여 부산경제에 대한 예측을 실증적으로 분석하고자 한다.

Ⅲ. 머신러닝과 딥러닝 추정 모형

본 연구에서는 부산지역의 전략산업들을 중심으로 하여 투자와 수출과 환율 등이 부산의 소득과 고용에 대한 예측을 분석하기 위하여 머신러닝 대표적 방법들인 의사결정 나무, 서

포트 벡터 머신), 인공 신경망) 모형과 심층 신경망 모형 등을 이용한다. 본 장에서는 이를 위해 머신러닝과 딥러닝 모형 등을 다음과 같이 Hastie et al.(2017) 연구를 참조하여 간략히 요약한다[1].

1. 의사결정 나무(Decision Tree) 모형

의사결정 나무(Decision Tree)는 데이터에 내재되어 있는 패턴을 나무형태로 계층적 구조로 이루어진 것을 도표화하여 예측하는 모형으로 최적의 분할변수와 분할점을 선택한다. 회귀트리의 경우 잔차제곱합(Residual Sum of Squares)을, 분류트리는 엔트로피 등의 불순도(Impurity)를 최소로 하는 영역을 찾는 것이다. 불순도의 측정지수는 지니 지수(Gini Index)와 엔트로피 지수(Entrophy Index)를 많이 사용한다.

그러나 이러한 지수는 나무의 크기에 따라 민감도가 낮기 때문에 실제에 있어서는 민감도가 높은 불순도의 측정 지수로 지니 지수와 엔트로피 지수를 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 지수(G)와 엔트로피 지수(D)는 다음과 같이 정의한다.

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

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 혹은 1에 가까이 있으면, 엔트로피 지수는 0에 가까울 것이다. 이와 같이 지니 지수와 엔트로피 지수는 모두 m차 노드에서 순수하다면 각각 작은 값을 가지게 된다.

그리하여 특성변수와 해당 변수 각각의 분절점(cutpoint)의 모든 가능한 조합 중에서 가장 좋은 분할을 만들어내는 특성변수와 분절점의 조합을 선택한다. 회귀분류에서는 각 영역 내에서 MSE가 최소가 되도록 이 과정을 반복하고 다음과 같이 찾아낸다.

즉 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 설정한다.[2] 그리고 변수들과 점들을 자동적으로 분리하고, 어떤 나무의 모양을 결정하는 알고리즘을 필요로 한다. 그 때, 만약 우리가 한 개의 분할을 M 지역 즉 R1, R2,..., RM으로 나눌 때, 각 지역에서 상수 c_m을 상정하면 다음과 같은 반응에 대한 식을 가진다.

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

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

$$\widehat{c_m} = ave(y_i | x_i \in R_m)$$

$$s.t \quad Min \sum_{i=1}^{p} (y_i - f(x_i))^2$$

1) Hastie et al.(2017) 참조

2) Hastie et al. (2017) 참조.

2. 인공 신경망(Artificial Neural Network) 모형

본 절에서 인공 신경망 모형을 만들어 분류 예측을 하기 위해서 다층 퍼셉트론(Multi-layer Perceptron; MLP) 모형이 구현되는 MLP 분류 함수를 이용하고, 수치예측을 위해서는 MLP 회귀함수를 이용한다. MLP 함수에서는 모형의 복잡도의 규제항인 알파(alpha)를 설정하는데 알파를 높이면 규제강화가 일반화되고, 반대로 너무 높이면 과소적합의 문제가 발생할 가능성이 있다.

본 연구에서 사용되는 다층퍼셉트론은 입력층(input layer)과 출력층(output layer) 사이에 여러 개의 중간층 즉 은닉층(hidden layer)이 존재하는 구조이다. 인공 신경망 모형은 맨위에 k번째의 목표(target) 측정치인 Y_k ($k=1,2,...,K$)가 있을 때, 모형을 설명하는 특성치(features) Z_m은 다음 식과 같이 선형함수로 이루어지고, 그 때 Y_k는 설명변수인 특성치 Z_m의 선형함수로 이루어진다.

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X), m = 1,2,..., M,$$
$$T_k = \beta_{ok} + \beta_k^T Z, \quad k = 1, 2, ..., K,$$
$$f_k(x) = g_k(T), k = 1,2,...,K,$$

여기서 특성치인 $Z = (Z_1, Z_2,..., Z_M)$, $T = (T_1, T_2,..., T_K)$이고, 활성화 함수(activation function)인 $\sigma(v)$는 보통 $\sigma(v) = \frac{1}{1+e^{-v}}$ 지그모이드(sigmoid) 함수를 사용하며 출력 함수는 다음과 같은 softmax 함수를 사용한다.

$$g_k(T) = \frac{e^{T_k}}{\sum_{l=1}^{K} e^{T_l}}$$

3. 서포트 벡터 머신(Support Vector Machine:SVM)

Hastie et al.(2017)의 연구를 따라 서포트 벡터 머신의 원리를 다음과 같이 요약한다. 먼저 N쌍의 학습자료인 $[(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)]$, $x_i \in R^p$, $y_i \in (-1,1)$ 를 상정하면, 그 때 β가 단위 벡터로서 $\| \beta \| = 1$이면, $\{x : f(x) = x^T\beta + \beta_0 = 0\}$ 식으로정의되는 초평면(hyperplane)을 가진다. 이 경우 서포트 벡터 머신 분류는 다음 식에서와 같이 C는 비용 패러미터(cost parameter), $\xi = (\xi_1, \xi_2, ..., \xi_N)$는 슬랙변수(slack variable)라고 할 때 다음과 같은 식으로 구해진다.

$$\min_{(\beta,\beta_0)} [\frac{1}{2} \| \beta \|^2 + C\sum_{i=1}^{N} \xi_i]$$

subject to

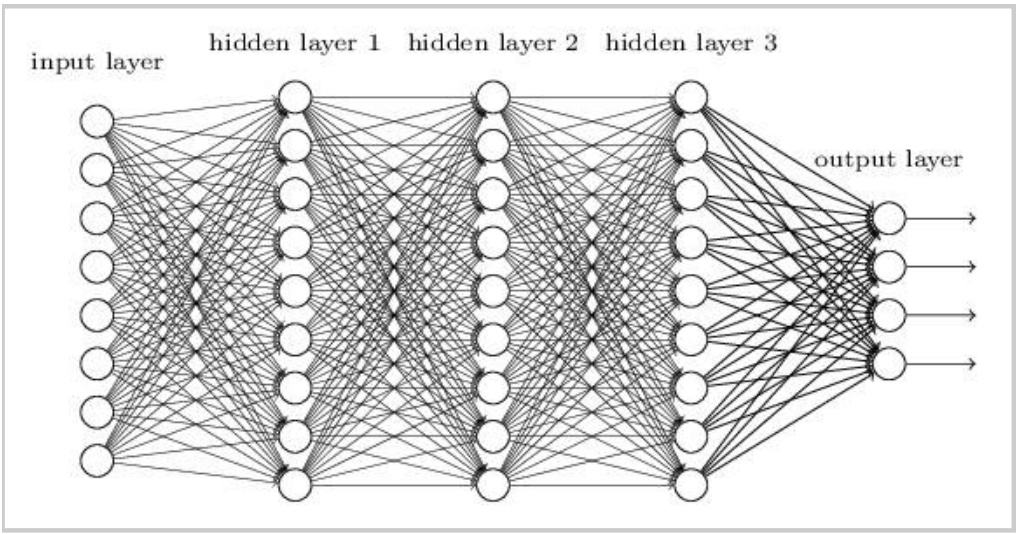
$$\xi_i \geq 0, y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \; \forall \, i$$

그 때, Lagrangian(primal) 함수(L_p)를 다음과 같이 설정하여 β, β_0, ξ_i에 대해 각각 최소화시키는 값을 구할 수 있다.

$$L_P = \frac{1}{2} \| \beta \|^2 + C\sum_{i=1}^{N} \xi_i \quad . \\ - \sum_{i=1}^{N} \alpha_i [y_i(x_i^T\beta + \beta_0) \\ - (1 - \xi_i)] - \sum_{i=1}^{N} \mu_i \xi_i$$

서포트 벡터 머신은 고차원 공간으로의 데이터 이동을 통한 비선형 SVM이 분류 성과를 더 높이기 위해서 보통은 커널 함수(Kernel Function: K)를 이용하여 비선형 SVM을 매핑하여 선형 분리를 가능하도록 다음과 같이 한다. 첫째, 텐서플로 최적화(Tensorflow Optimizer)를 이용하여, 아래의 식에서와 같이

Fig. 1. The Structure of Deep Neural Network



Sources: http://neuralnetworksanddeeplearning.com/chap5.html

SMO(Sequential Minimum Optimization) 알고리즘을 사용하여 다음과 같이 Dual 형식 원리를 이용하여 구한다.

$$L(\alpha) = \sum_{j=1}^{n} \alpha_j - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad ,$$

$$Max\, L(x, \alpha), \;\; 0 \le \alpha \le C, \quad .$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \; \text{for}\, all\; 1 \le i \le n$$

둘째, $\alpha > 0$인 점, 즉 α에 붙은 첨자 j가 동일한 x와 y를 하나 선택해서, 아래와 같은 분류(Classifier)를 만든다.

$$sign\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + y_j - \sum_{i=1}^{n} \alpha_i y_i K(x_i, x_j)\right)$$

4. 딥러닝의 심층 신경망 (Deep neural networks, DNN) 모형

딥러닝(Deep Learning)의 핵심 모델로서, 심층 신경망은 입력층과 출력층 사이에 다수의 은닉층을 가지고 있는 인공 신경망(ANN) 모형의 일종이기 때문에 다양한 비선형적 관계를 학습할 수 있다. Nielsen (2015)에 의하면, 심층 신경망은 인공 신경망의 문제점이었던 과잉적합, 경사도(gradient)의 소멸 등을 ReLU(rectified linear unit), 배치 정규화(batch normalization), 새로운 초기화 방법을 통해 해결하였는데, 심층 신경망의 구조는 〈Fig. 1〉과 같다.[3] 심층 신경망은 피드포워드(Feedforword) 신경망으로 설계되어 있으며, 심층 신경망은 지도 학습(supervised learning)으로 역전파(backpropagation) 알고리즘 통해 각 신경망 노드의 가중치를 갱신하면서 모형을 최적화한다.

3) Nielsen, Michael A.(2015)

Table 1. Busan's 7 Strategic Industries(Selected in 2019)

Classification
Smart Maritime Industry
Intelligent Machinery Industry
Future Tranport Machinery Industry
Global Tourism Industry
Intelligent Information Service Industry
Life Care Industry
Clean Tech Industry

Source: Busan Matropolitan Government(2019).

Ⅳ. 소득과 고용에 대한 머신러닝과 딥러닝 예측

부산시는 지역경제를 성장시키고 부산의 지역산업 경쟁력을 강화하기 위해서, 〈Table 1〉에서 나타나 있듯이 2019년 제 5차 7대 전략산업들과 22개의 세부 전략산업 분야를 새롭게 선정하였다. 그리하여 부산시는 전략산업의 지원 및 육성을 위한 경제정책을 수립하여, 중요한 경제목표들인 부산지역 고용의 증대와 소득을 증가시키려고 하고 있다.

본 장에서는 7대 전략산업들이 부산 경제에 미치는 영향을 분석하기 위하여, 소득 혹은 고용을 종속변수로 두고 7대 전략산업들인 글로벌관광(la), 라이프케어(lb), 미래수송기기(lc), 스마트해양(ld), 지능정보서비스(le), 지능형기계(lf), 클린테크(lg) 산업들을 각각 설명변수로 선택하여 분석한다[4]. 그 다음에는 이러한 7대 전략산업들과 거시경제변수들인 수출 및 투자 그리고 환율 등을 을 독립변수로 두고 소득 혹은 고용을 종속변수로 두고 이에 대한 분석을 해본다. 여기서 종속변수와 설명변수 모두 로그값을 취하여 추정한다. 전략산업 자료와 수출과 환율자료 등은 부산시 통계자료와 통계청의 KOSIS 자료를 이용하였다.

4) Yi and Lee (2020)의 연구 일부분 참조함.

1. 전략산업과 의사결정나무 (Decision Tree) 모형 예측

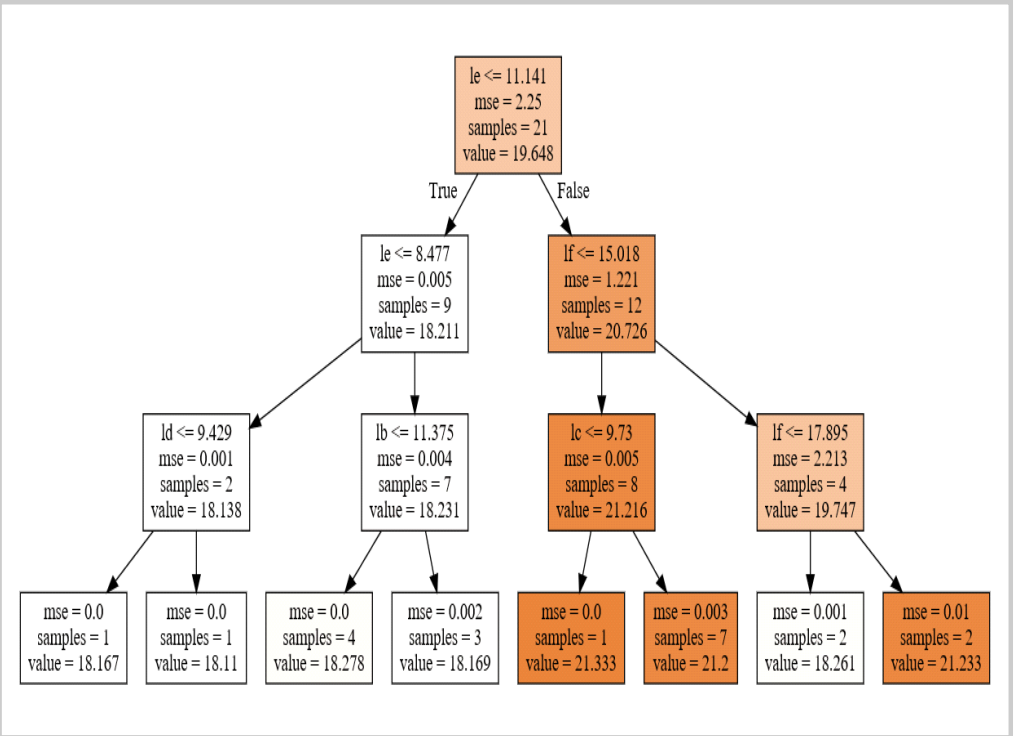
1) 의사결정나무에 의한 전략산업의 소득과 고용에 대한 예측

(1) 부산의 전략산업과 소득 예측

본 절에는 부산의 7대 전략산업들을 독립변수로 사용하여 종속변수인 소득을 예측하는 의사결정나무 모형을 만들어 분석해본다. 이를 위해 먼저 의사결정나무의 사전 가지치기 옵션을 사용하여 나무의 최대 깊이(Max Depth)는 3으로 설정하였다. 그리고 예측용 의사결정나무 모형의 성능을 평가하기 위하여 결정계수와 MSE를 구하였는데, 다음과 같이 학습용 데이터 세트의 결정계수는 0.999, 평가용 데이터 세트의 결정계수는 0.997로 나타났으며, MSE도 0.089로 결정계수가 높은 좋은 모형을 나타내고 있다.

〈Fig. 2〉에서 의사결정나무에서 보면, 지능정보서비스산업의 로그값(le)이 11.141보다 작아서 참(True)이면 왼쪽 노드로 내려가고, 지능정보서비스산업의 로그값이 11.141보다 크면 거짓(False)이므로 오른쪽 노드로 내려간다. 먼저 오른쪽 노드에서 다시 지능형기계산업의 로그값(lf)이 15.018 이하 이면 참이므로 왼쪽 노드, 15.018보다 크면 거짓이므로 오른쪽 노드로 분류된다. 그 다음 지능형기계산업의 로그

Fig. 2. Strategic Industry and Income Prediction



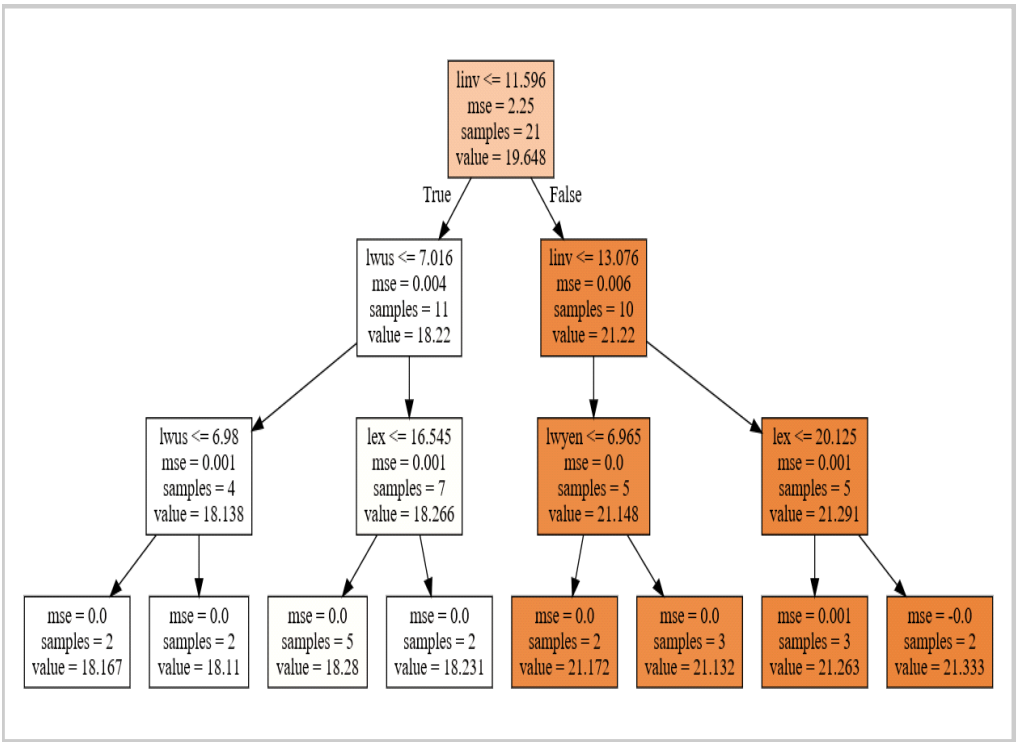
값(lf)이 17.875 이하이면 참으로 분류되어 왼쪽으로 가서 목표 예측 변수인 부산의 로그 소득값은 18.276으로 예측하게 되고, 반면에 지능형기계산업의 로그값(lf)이 17.875 보다 크면 거짓으로 오른쪽 노드로 가서 로그 소득값은 21.333으로 예측하게 된다.

위에서 만약 지능형기계산업의 로그값(lf)이 15.018 이하 이면 참이므로 왼쪽 노드로 내려간다. 여기서 미래수송기기산업의 로그값(lc)가 9.73 보다 작으면 참이 되어 다시 왼쪽 노드로 내려가서 목표 예측 변수인 부산의 로그 소득값은 21.333으로 예측하게 되고, 그 이상이면 거짓이 되어 오른쪽 노드로 내려가서 로그 소득값은 21.2가 된다.

반대로 의사결정나무에서 먼저 지능정보서비스산업의 로그값(le)이 11.141보다 작아서 참(True)이면 왼쪽 노드로 내려가고, 거기서 지능정보서비스의 로그값(le)이 8.477보다 도 작아서 참(True)이면 다시 한 번 더 왼쪽 노드로 내려간다. 거기서 스마트해양산업의 로그값(ld)가 9.429보다 작으면 참이 되어 왼쪽 노드로 내려가서 목표 예측 변수인 부산의 로그 소득값은 18.167으로 예측하게 되고, 반면에 9.429보다 크면 거짓으로 오른쪽 노드로 가서 로그 소득값은 18.11로 예측하게 된다.

마찬가지로 여러 조건들에 의해 참과 거짓으로 분류되어 분할되며 각 조건에 따라서 의사결정나무 깊이가 3일 때 8개의 소득에 대한 예측값의 로그값이 18.11에서 21.333으로 나오게 된다. 한편, 〈Fig. 2〉에서와 나타난 것 같이 끝마디로 갈수록 MSE도 작아지고 있다. 여기서 보면 지능정보서비스산업에 의해 첫마디에서 나누어지므로 상당히 이 산업은 중요한 산업으로 여겨진다. 또한 의사결정나무 모형에 의한 분석에서 각 산업의 선행조건에 의해 부산의 소득이 조금씩 변화하는 것을 알 수 있다.

Fig. 3. Strategic Industry, Investment, Export, Exchange Rate and Income Prediction



(2) 부산의 전략산업, 투자, 수출, 환율과 소득 예측

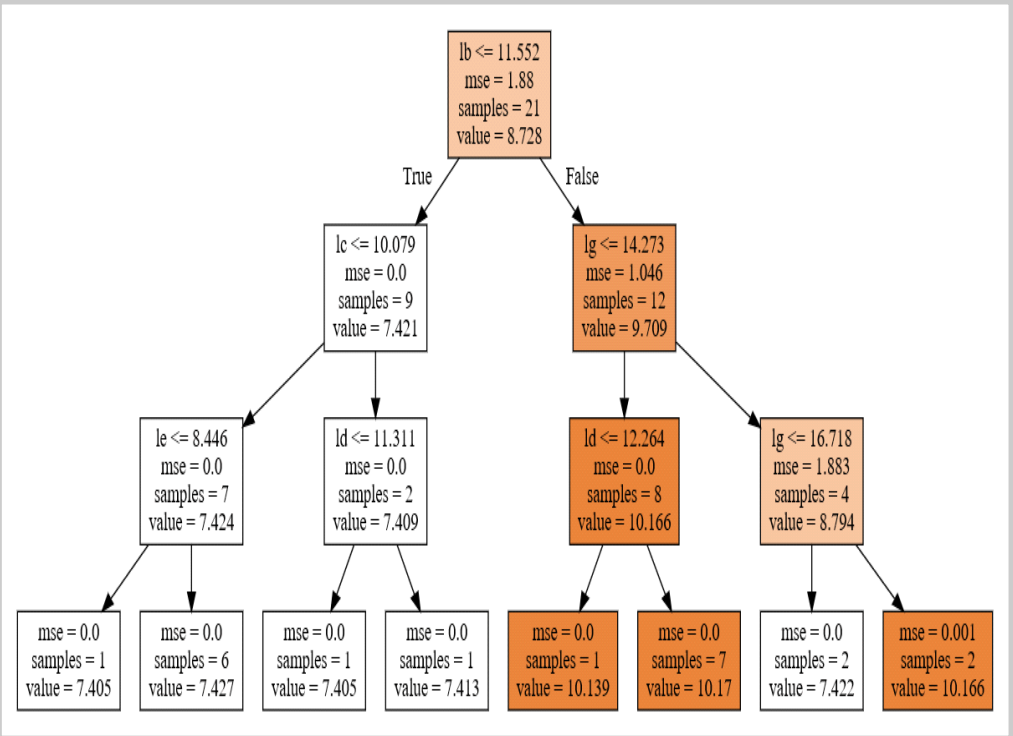
본 절에서는 소득을 종속변수로 두고 앞에서와 같이 7개의 전략산업들인 뿐만 아니라, 중요한 거시경제 변수들인 환율들인 원화의 대미 달러 환율(lwus), 원화의 대 일본 엔화환율(lwyen)을 포함하고 또한 수출(lex)과 투자(linv) 등을 추가로 설명변수로 선택하였고, 종속변수와 설명변수 모두 로그값을 취하여 추정하였다.

이를 위해 의사결정나무의 사전 가지치기 옵션을 사용하여 나무의 최대 깊이(Max Depth)는 3으로 설정하였다. 그리고 예측용 의사결정나무 모형을 성능을 평가하기 위하여 결정계수와 MSE를 구하였는데, 학습용 데이터 세트의 결정계수는 0.999, 평가용 데이터 세트의 결정계수는 0.999로 나타났으며, MSE는 0.017로 나타났다.

먼저 〈Fig. 3〉의 의사결정나무에서 보면, 앞의 의사결정나무와 같이 먼저 로그 투자값(linv)이 이 11.596보다 작아서 참(True)이면 왼쪽 노드로 내려가고, 투자의 로그값이 11.596보다 크면 거짓(False)이므로 오른쪽 노드로 내려간다. 먼저 오른쪽 노드에서 다시 투자의 로그값(lf)이 13.076 이하 이면 참이므로 왼쪽 노드, 13.076보다 크면 거짓이므로 오른쪽 노드로 분류된다. 오른쪽 노드로 내려가서 다시 수출의 로그값(lex)이 20.125 이하이면 참으로 분류되어 왼쪽으로 가서 목표 예측 변수인 부산의 로그 소득값은 21.263으로 예측하게 되고, 반면에 수출의 로그값(lex)가 20.125 보다 크면 거짓으로 오른쪽 노드로 가서 로그 소득값은 21.333으로 예측되어 나타난다.

그리고 〈Fig. 3〉의 두 번째 노드에서 만약 투자의 로그값(linv)이 13.076 이하가 되면 참이므로 왼쪽 노드로 내려간다. 여기서 또 원화의

Fig. 4. Strategic Industry and Employment Prediction



대 일본 엔화환율(lwyen)이 6.965 이하가 되면 참이 되어 다시 왼쪽 노드로 내려가서 목표 예측 변수인 부산의 로그 소득값은 21.172로 예측되고, 그 이상이면 거짓이 되어 오른쪽 노드로 내려가서 로그 소득값은 21.132가 된다.

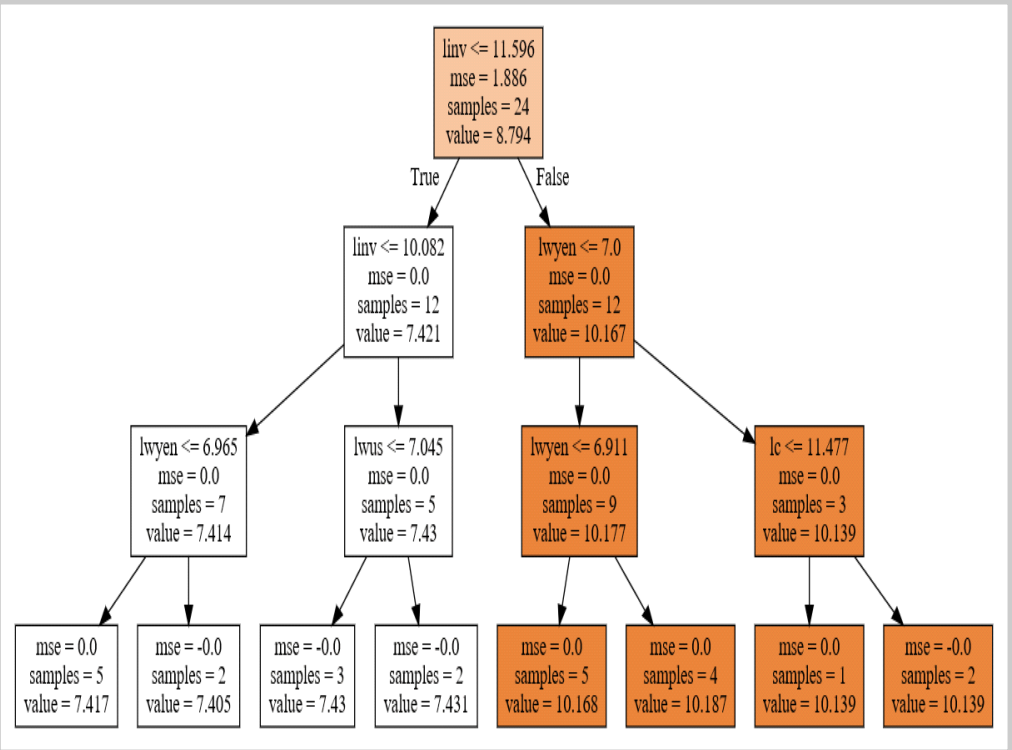
그러나 반대로 의사결정나무에서 먼저 로그 투자값(linv)이 11.596 이하가 되어 참(True)이면 왼쪽 노드로 내려가고, 거기서 한국 원화의 대 미달러 환율의 로그값(lwus)이 7.016 d이하이면 참이 되어 다시 한 번 더 왼쪽 노드로 내려간다. 거기서 대미 달러 환율의 로그값이 6.98 이하가 되면 참이 되어 왼쪽 노드로 내려가서 목표 예측 변수인 부산의 로그 소득값은 18.167으로 예측하게 되고, 반면에 6.98보다 크면 거짓으로 오른쪽 노드로 가서 부산의 로그 소득값은 18.11로 예측하게 된다.

그러므로 〈Fig. 3〉에서 나타난 것과 같이 이러한 의사결정 나무 모형에서는 투자와 환율 그리고 수출이 소득을 결정하는데 중요한 역할을 하는 것으로 나타났다. 이들의 선행 조건들에 의해 참과 거짓으로 분류되는 각 조건에 따라 의사결정나무 깊이가 8개의 소득에 대한 예측값이 나타나게 되고, 끝마디로 갈수록 MSE도 0으로 수렴하여 작아지고 있다.

(3) 부산의 전략산업과 고용 예측

본 절에는 부산의 7대 전략산업들을 독립변수로 사용하여 종속변수인 고용을 예측하는 의사결정나무 모형을 만들었다. 앞에서와 같이 사전 가지치기 옵션을 사용하여 나무의 최대 깊이(Max Depth)는 3으로 설정하였고, 고용자수(lempl)을 종속변수로 두고 앞에서와 같이 7개의 독립변수들을 설명변수로 선택하였고, 종속변수와 설명변수 모두 로그값을 취하여 추정하였다. 그리고 예측용 의사결정나무 모형을 성능을 평가하기 위하여 결정계수와 MSE를 구

Fig. 5. Leading Industry, Investment, Export, Exchange Rate and Employment Prediction



하였는데, 학습용 데이터 세트의 결정계수는 0.999, 평가용 데이터 세트의 결정계수는 0.999로 나타났고, MSE도 0.012로 결정계수가 높은 좋은 모형으로 나타났다.

의사결정나무 모형에 의한 고용에 대한 추정결과를 〈Fig. 4〉에서 보면, 의사결정나무에서 먼저 라이프케어산업의 로그값(lb)이 11.552보다 크면 거짓이 되어 오른쪽 노드로 내려가며, 거기서 클린테크 산업의 로그값(lg)이 14.273보다 크면 다시 오른쪽 노드로 내려간다. 다음에 클린테크산업의 로그값(lg)이 16.718보다 크면, 목표 예측 변수인 부산의 로그 고용자 수는 10.166으로 예측되지만, 그 값이 16.718 이하가 되면 왼쪽 노드로 내려가서 고용자 수의 로그값은 7.422가 된다.

라이프케어산업의 로그값(lb)이 11.552보다 커서 오른쪽 노드로 내려가며, 거기서 클린테크 산업의 로그값(lg)이 14.273 이하가 되면 왼쪽 노드로 내려간다. 거기서 스마트해양산업의 로그값(ld)이 12,264 이하가 되면, 목표 예측 변수인 부산의 로그 고용자 수는 10.139으로 예측되지만, 그 값이 12.264보다 크면 오른쪽 노드로 내려가서 고용자 수의 로그값은 10.17로 예측된다.

한편, 의사결정나무에서 먼저 라이프케어산업의 로그값(lb)이 11.552 이하가 되어 이하가 되어 참(True)이면 왼쪽 노드로 내려가고, 거기서 지능정보서비스 산업의 로그값(le)이 10.079 이하이면 참이 되어 다시 한 번 더 왼쪽 노드로 내려간다. 왼쪽 노드로 내려가서 다시 한 번 지능정보서비스 산업의 로그값(le)이 8.446 이하가 되면 참이 되어 왼쪽 노드로 내려가서 목표 예측 변수인 부산의 고용자 수의 로그값은 7.405으로 예측하게 되고, 반면에 8.446보다 크면 거짓으로 오른쪽 노드로 가서 고용자 수의 로그값 소득값은 7.427로 예측하게 된다.

여기서 고용자 수를 결정하기 위해서는 전략산업 중에서 라이프케어산업을 중심으로 여러 전략산업들의 조건들에 의해 참과 거짓으로 분류되어 분할되며 각 조건에 따라서 의사결정나무 깊이가 8개의 고용자 수에 대한 예측값이 나타나게 되고, 끝마디로 갈수록 MSE도 점점 작아져 0으로 수렴하고 있다.

(4) 주요산업, 투자, 수출, 환율과 고용 예측

본 절에서는 전략산업과 연관성이 높은 7대 주요산업들에 거시경제변수들인 원화의 대미 달러 환율, 원화의 대 일본 엔화환율, 수출과 투자 등을 추가하여 설명변수로 선택하였고, 종속변수와 설명변수 모두 로그값을 취하여 추정하였다.

이를 위해 의사결정나무의 사전 가지치기 옵션을 사용하여 나무의 최대 깊이(Max Depth)는 3으로 설정하였다. 그리고 예측용 의사결정나무 모형을 성능을 평가하기 위하여 결정계수와 MSE를 구하였는데, 학습용 데이터 세트의 결정계수는 0.999, 평가용 데이터 세트의 결정계수는 0.999로 나타났으며, MSE도 0.017로 나타났다.

〈Fig. 5〉에서 나타난 의사결정나무에서 보면, 산업들에 대한 투자값(linv)이 11.596 이하가 되어 참(True)이면 왼쪽 노드로 내려가고, 투자의 로그값이 11.596보다 크면 거짓(False)이므로 오른쪽 노드로 내려간다. 오른쪽 노드에서 원화의 대 일본 엔화환율(lwyen)이 7.0보다 크면 거짓이므로 오른쪽 노드로 내려간다. 거시서 만약 로그 지능정보서비스산업(le)이 11.477보다 크면 거짓이 되어 오른쪽 노드로 내려가서 예측 변수인 로그 고용자 수에 대한 값은 10.139로 예측하게 된다.

반면에 원화의 대 일본 엔화환율(lwyen)이 7.0 이하가 되면 참이 되어 왼쪽 노드로 내려간다. 거기서 다시 한 번 원화의 대 일본 엔화환율(lwyen)이 6.911 이하가 되면 왼쪽 노드로 내려가서 참이 되어 고용자 수의 로그값은 10.168이 되며, 그 값이 6.911보다 크면 고용자 수의 로그값은 10.187로 예측하게 된다.

반대로 의사결정나무에서 맨 먼저 로그 투자값(linv)이 11.596 이하가 되어 참(True)이면 왼쪽 노드로 내려가고, 거기서 만약 투자의 로그값이 10.082 이하가 되면 참이 되어 다시 한 번 왼쪽노드로 내려간다. 다시 원화의 대 일본 엔화환율(lwyen)이 7.0 이하가 되면 참이 되어 왼쪽 노드로 내려간다. 거기서 다시 한 번 원화의 대 일본 엔화환율(lwyen)이 6.965 이하가 되면 왼쪽 노드로 내려가서 참이 되어 고용자 수의 로그값은 7.417로 예측되며, 그 값이 7.414 보다 크면 고용자 수의 로그값은 7.405로 예측된다.

그러나 〈Fig. 5〉에서 나타난 것과 같이 고용을 예측하는 의사결정 나무 모형에서도 수출보다는 투자와 환율 그리고 지능정보서비스산업(le) 등이 부산의 소득을 결정하는데 중요한 역할을 하는 것으로 나타났다. 전략산업에서는 이와 같이 이 경우에 있어서도 여러 조건들에 의해 참과 거짓으로 분류되어 분할되며 각 조건에 따라서 의사결정나무 깊이가 8개의 고용자 수에 대한 예측값이 나타나게 되고, 끝마디로 갈수록 MSE도 점점 작아져 0으로 수렴하고 있다.

2. 인공 신경망(Artificial Neural Network) 모형 예측

1) 전략산업의 소득과 고용에 대한 인공 신경망 예측

인공 신경망에서 은닉층의 크기는 은닉층의 개수와 은닉노드 개수에 의해 정해지는데, 본 절에서는 모형이 최적화되는 최대 반복회수로 1,000으로 상정을 하였다. 은닉층의 크기(hidden layer size)=[20. 20]으로 은닉층은 2개, 은닉노드는 20개를 상정하여 얻은 인공 신경망 추정 결과를 얻는다. 그리하여 본 절에서는 이러한 MLP 함수의 알파, 반복회수, 은닉층의 개수와 은닉노드 개수 달리 해서 인공 신경망을 이용하여 부산의 7대 전략산업의 소득과 고용에 대한 분류예측과 수치예측을 하였는데, 그에 대한 추정결과를 차례대로 〈Table 2〉에서 살펴보면 다음과 같다.

Table 2. Predictions of Income and Employment with Artificial Neural Networks

Classification Prediction (Income: 0,1)	Model Accuracy	Accuracy of Training Data Set: 0.714 Accuracy of Test Data Set: 0.556
	MLP	Alpha=1, max_iter=200, hidden layer size =[20]
	Model Performance Classification Report	precision / recall / f1-score 0: 0.57 / 0.80 / 0.67 1: 0.50 / 0.25 / 0.33 accuracy: 0.56 macro avg: 0.54 / 0.53 / 0.50 weighted avg: 0.54 / 0.56 / 0.52
Regression Prediction (Employmet: lempl)	MLP	Alpha=1, max_iter=200, hidden layer size =[20,20]
	Model Accuracy	Determination Coefficient of Training Data Set: 0.607 Determination Coefficient of Test Data Set: 0.391 RMSE: 0.203
Regression Prediction (Income: ly)	MLP	Alpha=1, max_iter=500, hidden layer size =[30,30]
	Model Accuracy	Determination Coefficient of Training Data Set: 0.468 Accuracy of Test Data Set: 0.401 RMSE:1.176

첫째, 인공 신경망에 의한 고용(lempl)에 대한 분류예측에 대한 모형평가를 보면, 인공 신경망의 인자들인 알파(alpha)=1, 반복회수(max_iter)=200, 은닉층의 크기(hidden layer size)=[20]으로 은닉층은 1개, 은닉노드는 20개를 각각 상정한다. 여기서 소득에 대한 분류를 부산은 1, 전국은 0으로 두고 분류예측 결과를 보면, 모형의 정확도를 나타내는 Accuracy of Training Data Set가 0.714, 평가용 데이터 세트 정확도가 0.556로 각각 나타나, 학습용 정확도가 다소 더 높게 나와 약간의 과잉적합이 있을 가능성이 나타났다.

그리고 모형의 성능 평가분류를 보면, 정밀도와 재현율 등이 그리 높지 않게 나왔다.그리고 f1-스코어 값이 두 영역에서 0.67과 0.33으로 각각 나타났다. 정확도는 약 56%로 정도로 나타났고 평균과 가중평균도 0.50-0.56사이에 있어 높지 않은 것으로 나타나 모형의 분류예측의 정확도는 높게 나타나지 않았다.

둘째, 인공 신경망에 의한 고용(lempl)에 대한 수치예측에 대한 모형평가를 보면, 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 0.607, 평가용 데이터 세트 결정계수가 0.391로 나타나 과잉적합의 문제가 발생할 가능성이 높게 나타났으며, RMSE는 0.203으로 나타나 모형의 적합성은 낮게 나타났다.

셋째, 인공 신경망에 의한 소득에 대한 수치예측에 대한 모형평가 역시, 학습용 데이터 세트 결정계수가 0.468, 평가용 데이터 세트 결정계수가 0.401로 나타났고, 과잉적합의 문제가 조금 있을 수 있으며, RMSE는 1.176으로 도출되어 고용에 대한 인공 신경망에 의한 수치예측 모형의 적합성은 높지 않은 것으로 나타났다.

3. 서포트 벡터 머신(SVM) 모형의 소득과 고용에 대한 예측

보통 파이썬(python)의 Scikit-Learn의 서포트 벡터 머신 함수는 소프트 마진을 기반으로 구현되어 있으며 인자 C(Cost)를 통해 마진 오류의 범위를 결정한다. C를 높게 설정할수록

Table 3. Predictions of Income and Employment with SVM

	SVM RBF Kernel	Determination Coefficient of Training Data Set	Determination Coefficient of Test Data Set	RMSE
Predictions of Income and Employment				
Income	C=10, epsilon=0.5, gamma=0.01	0.906	0.920	0.430
	C=10, epsilon=0.7, gamma=0.01	0.803	0.808	0.665
	C=100, epsilon=1.0, gamma=0.01	0.592	0.529	1.043
	C=1000, epsilon=1.0, gamma=0.01	0.655	0.616	0.939
	C=1000, epsilon=1.0, gamma=0.02	0.615	0.584	0.981
	C=1000, epsilon=1.0, gamma=0.03	0.580	0.555	1.014
Employment : Trinomial Classsifications (0, 1, 2)				
Employment	C=10, epsilon=0.1, gamma=0.01	0.468	0.379	25.202
	C=100, epsilon=0.1, gamma=0.01	0.847	0.837	12.925
	C=1000, epsilon=0.5, gamma=0.01	0.883	0.841	12.755
	C=1000, epsilon=0.5, gamma=0.05	0.956	0.898	10.235
	C=1000, epsilon=1.0, gamma=0.01	0.884	0.846	12.531
	C=1000, epsilon=0.1, gamma=0.1	0.982	0.766	15.476
Employment : Binomial Classsifications (0, 1)				
Employment	C=100, epsilon=0.1 , gamma=0.01	0.799	0.525	0.331
	C=1000, epsilon=1.0, gamma=0.01	0.655	0.618	0.939
	C=1000, epsilon=0.1, gamma=0.05	0.968	0.306	0.400
	C=1000, epsilon=0.1, gamma=0.1	0.967	0.297	0.402

마진폭이 줄어든다. 그리고 서포트 벡터 머신 모형을 최적화하기 위해서는 적용할 커널 함수(kernel function), 오류에 대한 마진을 제어하는 모수인 C, 그리고 데이터의 영향도와 영향력의 범위와 관련된 γ(gamma), 그리고 평가 데이터의 허용 오차율과 관련된 ϵ(epsilon)에 대한 결정이 필요하다.

본 절에서는 커널 함수로 방사기저함수(radial basis function, RBF) 커널을 적용하며, k겹 교차검증에 의해 산출된 검증(validation) 데이터의 RMSE가 최소가 되는 모형을 서포트 벡터 머신 모형으로 결정하였다. 그리하여 본 절에서는 먼저 부산의 소득 혹은 고용을 종속변수로 두고 7대 전략산업들을 독립변수로 삼고 서포트 벡터 머신 회귀모형(SVR) 분석을 한다. 이러한 서포트 벡터 머신 회귀분석과 모형평가에 대한 결과물인 결정계수들과 RMSE가 아래 〈Table 3〉에 나와 있다. 서포트 벡터 머신 회귀분석의 인자로서 오류의 범위와 연관이 있는 모형의 인자인 C를 10, 100, 1000 등으로 변화시켜 보았고, 서포트 벡터 머신 회귀모형의 인자인 epsilon값을 0.1에서 1까지 변화를 시켜보고, 마지막으로 RBF 함수의 커널 계수를 나타내는 gamma 값을 0.01에서 0.1까지 변동을 시켜 보고, 그에 해당하는 결정계수들과 RMSE를 구하였는데 그 추정결과는 〈Table 3〉에 나타나 있다.

첫째, 7대 전략산업들이 부산의 소득에 미치는 영향을 서포트 벡터 머신 회귀 모형으로 예측결과를 〈Table 3〉에서 보면, 서포트 벡터 머신 회귀모형의 인자들이 각각 C=10, epsilon=0.5, gamma=0.01 일 때, 학습용과 평가용의 데이터 세트 모두에서 결정계수가 아주 높은 0.906과

0.920으로 도출되었고, RMSE는 0.430으로 도출되어 예측력이 가장 우수하게 나왔다. 그 다음 모형의 인자들이 각각 C=10, epsilon=0.7, gamma=0.01일 때 두 개의 결정계수들이 각각 0.803과 0.808, RMSE는 0.665 등으로 나와 예측력이 우수하게 나왔다. 그러나 마진과 관계있는 C를 100 혹은 1,000으로 증가 시켰더니 결정계수들이 많이 감소하였다.

둘째, 부산의 고용자 수에 대해서 3개 영역(0,1,2)으로 나누고, 7개의 전략산업들을 독립변수들로 회귀분석을 한 결과, 서포트 벡터 머신 회귀모형의 인자들이 각각 C=1000, epsilon=0.5, gamma=0.05와 C=1000, epsilon=0.1, gamma=0.1일 때 두 결정계수가 각각 매우 높았으며, RMSE도 각각 10.235, 15.476 등으로 각각 나타났다.

셋째, 부산의 고용자 수에 대해서 혹은 2개 영역(0,1)으로 나누어 서포트 벡터 머신 회귀분석을 하였을 때는 모형의 인자들이 각각 C=1000, epsilon=0.1, gamma=0.05와 C=1000, epsilon=0.1, gamma=0.1일 때 모두 학습용 데이터 세트 결정계수는 0.968과 0.967로 아주 높고 RMSE도 대체로 낮았으나, 평가용 데이터 세트 결정계수는 0.306과 0.297로 너무 낮아서 학습용 모형의 다소 과잉적합 문제가 발생할 수 있다.

그러므로 서포트 벡터 머신 모형에 의한 예측은 모형의 인자들의 조합과 선택이 중요하며 그 인자들의 조합에 따라 약간씩 예측값들이 달리 나타났다. 그리하여 위에서 나타난 것과 같이, 먼저 부산의 7대 전략산업들을 가지고 소득을 예측할 때는 서포트 벡터 머신 회귀모형의 인자들이 각각 C=10, epsilon=0.5, gamma=0.01인 경우에 가장 학습용 데이트와 평가용 데이트 세트의 결정계수들이 모두 높아서 좋은 예측 모형으로 나타났다.

그러나 부산의 7대 전략산업들을 가지고 고용자 수를 예측할 때는 고용자 수를 2분류 보다 3분류로 나누어 예측했을 때, 대체로 더 좋은 예측모형으로 나타났다. 서포트 벡터 머신 회귀모형의 인자들이 마진폭을 줄인 C=1000일 때가 마진폭을 넓게한 C=10 혹은 C=100일 때보다 학습용과 평가용 데이트 세트의 결정계수 모두 높은 좋은 모형으로 나타났다.

그리하여 위의 인공 신경망 모형보다는 서포트 벡터 머신 모형에 의해서 부산의 전략산업을 가지고 부산의 고용과 소득에 대한 영향을 예측할 때, 모형의 인자들의 조합과 선택에 의해서 달리 나타나지만, 대체로 결정계수가 더 높게 나타나고 좋은 예측 모형으로 나타난 것을 알 수 있다.

그러나 위의 인공 신경망 모형에 은닉층의 개수를 충분히 10개 이상 증가시킨 딥러닝(Deep Leaning)의 심층신경망(Deep Neural Network) 모형을 채택함으로써 자가학습 수행을 자동화하여 기계가 스스로 데이터에서 주요 특징을 추출하여 예측을 수행하면 더 좋은 예측력을 가진 모형이 될 수도 있다.

4. 딥러닝의 심층 신경망(DNN) 모형의 소득과 고용 예측

딥러닝의 심층 신경망(Deep Neural Network; DNN) 모형을 생성하고 학습 및 평가를 수행하기 위해서는 Keras 모듈의 순차적 함수를 도입하고 밀도(Dense) 함수와 활성화(Activation) 함수를 불러온다. 딥러닝의 중요한 인자에는 전체 데이터에 대한 학습반복 횟수(epochs; 에포크)와 한 번에 학습되는 데이터의 개수(batch_size) 등이 있다. 에포크가 클수록 모형의 성능이 올라갈 수 있으나 학습 속도가 느려져 과잉적합이 발생할 가능성이 있고, 배치 크기는 그 값이 작을수록 가중치 갱신을 자주 해야 하므로, 적정한 값을 찾아 조정해야 한다.

그리하여 딥러닝의 심층 신경망 모형에 의한 부산의 7대 전략산업이 소득(LY)과 고용(LEMPL)에 미치는 영향에 대한 예측결과가 〈Table 4〉에서 나타나 있다. 이를 위해 부산의 7대 전략산업들을 독립변수로 삼고 소득을 종속변수로 두고 추정예측을 하였다.

그리하여 본 절의 이항 분류예측에서는 이항 횡단 엔트로피(Binary Cross-Entrophy)를 설정하고, 경사하강법은 아담모형을 사용하였다. 이를 위해 순차적 함수를 이용하여 순차적 계층모형을 생성하고, 소득에서는 부가(add) 함

Table 4. Predictions of Income and Employment with DNN by Epoch and Batch

Prediction	(Epochs, Batch_Size)	Mean Squared Error(MSE)
Income	(30, 64)	Training Data Set MSE 2.529 Test Data Set MSE 2.544
	(40,64)	Training Data Set MSE 1.672 Test Data Set MSE 1.826
	(50, 64)	Training Data Set Set MSE 1.217 Test Data Set MSE 1.564
	(40,100)	Training Data Set MSE 1.672 Test Data Set MSE 1.826
	(100, 100)	Training Data Set MSE 0.611 Test Data Set MSE 0.693
Employment	(30, 64)	Training Data Set MSE 1.141 Test Data Set MSE 1.810
	(50, 64)	Training Set MSE 0.684 Test Data Set MSE 1.066
	(50, 100)	Training Set MSE 0.232 Test Data Set MSE 0.194
	(100, 100)	Training Set MSE 0.348 Test Data Set MSE 0.528

수를 3개를 이용하였고, 고용에서는 부가함수를 5개를 이용하여 개별 계층을 추가하였다. 각 층의 구조는 밀도(Dense) 함수를 통해 결정되는데, 주요인자로서 활성화(activation) 함수로 은닉층에는 Relu 함수, 출력층에는 Sigmoid 함수를 각각 설정하였다.

또한 회귀를 통한 수치예측에서는 가중치의 업데이트를 위해서, 관성의 방향을 고려한 경사하강법인 모멘텀과 아다그라드(Adagrad)의 보폭 민감도를 보완하는 RMSProp 방식을 결합한 Adam 방식을 사용하였다. Adagrad는 가중치의 업데이트 횟수에 따라 학습률을 조절하는 방법으로 업데이트 하는데, 변화가 잦을수록 학습률을 작게 하는 방법이다. 밀도(Dense) 함수의 활성화는 Relu 함수를 사용하였다.

그리하여 부산의 7대 전략산업들의 고용자 수에 대한 딥러닝에 의한 수치예측 결과들은 〈Table 4〉에서 나타나 있다. 먼저 에포크(반복횟수)와 배치 크기에 따른 MSE를 보면, 에포크(epochs)=100, 배치 크기(batch_size)= 100일 때 구한 학습용 데이터 세트의 MSE = 0.611, 평가용 데이트 세트의 MSE = 0.693로 각각 가장 작게 나타났다. 따라서 다른 에포크와 배치 크기 조합들인 (30,64), (40,64), (50, 64)보다, 에포크와 배치 크기가 (100, 100)일 때 학습용 데이터 세트와 평가용 데이트 세트의 MSE가 가장 작게 나타났기 때문에 가장 좋은 예측 모형으로 나타났다.

다음에는 딥러닝에 의한 2019년 부산이 선정한 7대 전략산업이 고용자 수(LEMPL)에 어떻게 영향을 미치는지에 대한 예측을 추정한 결과를 〈Table 4〉에서 보면, 로그로 표시한 부산의 7대 전략산업들의 고용자 수에 대한 딥러닝에 의한 수치예측 결과를 알 수 있다. 그리하여 에포크(반복횟수)와 배치 크기에 따른 MSE를 보면, 에포크(epochs)=50, 배치 크기(batch_size)= 100일 때 구한 학습용 데이터 세트의 MSE = 0.232, 훈련용 데이트 세트의 MSE = 0.194로 각각 가장 작게 나타났다, 그리하여 다른 에포크와 배치 크기 조합들인 (30,64)과

(50, 64), (100, 100)보다, 에포크와 배치사이즈가 (50, 100)일 때 학습용 데이터 세트와 평가용 데이트 세트의 MSE가 가장 작게 나타나서 예측오차가 가장 작고 가장 좋은 예측 모형으로 나타났다.

Ⅴ. 결론

본 연구는 전통적 구조적 계량모형이나 시계열 계량분석 모형의 과잉적합 문제를 극복하고 좀 더 예측성을 높이기 위해 외국에서 최근 도입하기 시작한 머신러닝과 딥러닝 기법 등을 도입하여 부산의 7대 전략산업들을 중심으로 투자와 수출과 환율 등이 부산의 소득과 고용에 미치는 영향 등을 분실증적으로 분석하였다. 그리하여 본 연구의 결과를 요약하면 다음과 같다.

첫째, 의사결정나무에 의한 전략산업의 소득과 고용에 대한 예측결과를 보면, 전략산업들의 여러 선행조건들에 의해 추정된 예측값들이 참과 거짓으로 분류되어 분할되며, 의사결정나무들의 깊이와 전략산업들과 수출과 환율 등의 선행조건에 따라 소득과 고용에 대한 예측값이 조금씩 다르게 나타났지만, 소득과 고용에 대해 상당히 구체적으로 예측할 수 있었다. 모형에서 끝마디로 갈수록 MSE도 0으로 수렴하는 등 작아지고 있고 좋은 예측모형으로 나타났다.

둘째, 의사결정나무 모형에 전략산업들에 투자와 환율 그리고 수출을 포함하였을 때는 투자와 환율 그리고 수출이 소득을 예측하는데 중요한 것으로 나타난 반면, 고용을 예측하는데 있어서는 투자와 환율 그리고 지능정보서비스산업 등이 부산의 소득을 예측하는데 중요한 기준이 되는 것으로 나타났다.

셋째, 인공 신경망에 의한 고용과 소득에 대한 예측모형의 결과를 보면, 모형의 인자들의 선택과 조합에 따라 약간씩 다르지만, 학습용 데이터 세트와 평가용 데이터 세트 결정계수가 비교적 낮게 나왔으며, 과잉적합의 문제가 발생할 가능성이 높게 나타났다. 그리고 RMSE도 비교적 높아 부산시의 전략산업들의 고용과 소득에 대한 인공 신경망에 의한 모형의 예측력은 그리 높지 않은 것으로 나타났다.

넷째, 서포트 벡터 머신(SVM)에 의한 7대 전략산업들이 먼저 소득에 미치는 영향을 서포트 벡터 머신 회귀 모형으로 예측결과, 학습용과 평가용의 데이터 세트 모두에서 결정계수가 아주 높게 나왔고, RMSE도 낮아 예측력이 우수하게 나왔다. 그 다음에 부산의 고용자 수에 대해서 7개의 독립변수들로 회귀분석을 한 결과, 서포트 벡터 머신 회귀모형 역시 결정계수도 높고 예측력이 높게 나왔다. 그리하여 서포트 벡터 머신 모형에 의한 전략산업들의 부산의 소득과 고용에 대한 예측이 인공 신경망 모형에 의한 예측보다 우월하게 나타났다.

다섯째, 인공 신경망에 은닉층을 중첩한 딥러닝 모형의 심층 신경망 모형에 의해서 부산의 전략산업으로 소득을 예측할 때, 에포크와 배치사이즈가 (100, 100)일 때 학습용 데이터 세트와 평가용 데이트 세트의 MSE가 가장 작게 나타나서 가장 좋은 예측 모형으로 나타났다. 그러나 고용자 수에 미치는 영향에 대한 예측을 살펴보면, 에포크와 배치사이즈가 (50, 100)일 때 학습용 데이터 세트와 평가용 데이트 세트의 MSE가 가장 작게 나타나 MSE가 가장 작은 것으로 나타나 가장 좋은 예측 모형으로 나타났다.

이와 같이 본 연구에서는 주요 머신러닝 기법들에 의한 부산의 전략산업의 소득과 고용에 대한 예측결과를 보면, 의사결정나무 모형, 인공 신경망 모형, 서포트 벡터 머신 모형, 그리고 딥러닝 심층 신경망 모형의 결정계수가 높게 나오고 RMSE 작게 나타났다. 그리고 모형 인자들의 조합과 선택에 따라서 모형의 과잉식별을 줄이고 예측력도 높일 수 있는 것으로 나타났다. 따라서 이러한 머신러닝 모형으로 향후 부산의 전략산업들을 선정하고 전략산업들의 지역경제 즉 지역의 성장과 고용에 미치는 효과를 좀 더 정확히 예측하고 분석할 필요가 있는 것으로 나타났다.

본 연구는 이와 같이 전통적 계량모델의 기법과는 달리 최근 외국에서 도입되기 시작한 머신러닝과 딥러닝 등의 기법을 도입하여 고용과 소득 등에 대해 좀 더 정확하게 예측함으로써 향후 부산시와 유관기관들이 부산의 7대 전

략산업을 효율적이고도 효과적으로 육성하고 지원하기 위한 정책적 시사점을 제공하는데 기여할 수 있을 것이다.

그러나 본 연구는 부산시의 전략산업들을 중심으로 분석하였기 때문에 분석대상과 분석방법, 그리고 좀 많은 자료를 수집하여 사용함으로써 좀 더 엄밀한 분석과 해석을 할 수 있을 것이다. 그럼에도 불구하고 본 연구는 우리나라에서 아직 소개조차 거의 없는 경제분석에 있어 머신러닝과 딥러닝을 이용한 예측 기법 등을 도입함으로써 부산과 우리나라의 경제분석에 대한 후속연구들에 대한 디딤돌이 되는데 기여할 수 있을 것이다.

References

Acemoglu, D. and P. Restrepo (2020), "Robots and Jobs: Evidence from US Labor Markets", *Journal of Political Economy*, 128(6), 2188-2244.

Agrawal, A., J. Gans and A. Goldfarb (2018), *Prediction Machines: The Simple Economics of Artificial Intelligence*, 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, In *The Economics of Artificial Intelligence: An Agenda*, 1 Edition, 507-547, National Bureau of Economic Research Conference Report, University of Chicago Press.

Athey, S., J. Tibshirani and S. Wager (2019), "Generalized Random Forests", *The 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), 124-127.

Chakraborty, C. and A. Joseph (2017), "Machine Learning at Central Banks", *Bank of Eng-Land Staff Working Paper*, No. 674.

De Prado, M. L. (2018), *Advances in Financial Machine Learning*, New York, NY, USA: Wiley.

Ding, M. J., S. Z. Zhang, H. D. Zhong, Y. H. Wu and L. B. Zhang (2019), "A Prediction Model of the Sum of Container Based on Combined BP Neural Network and SVM", *Journal of Information Processing Systems*, 15(2), 305-319.

Géron, A. (2017), *Hands-On Machine Learning with Scikit-Learn and Tensor Flow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 1st Edition, O'Reilly Media.

Gu, S., B. Kelly and D. Xiu (2019), "Empirical Asset Pricing via Machine Learning", *NBER Working Paper* No. 25398.

Hastie, T., R. Tibshirani and J. Fridman (2017), *The Elements of Statistical Learning*, Second Edition, 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", Bank of Korea *Working Paper*, No. 2020-18.

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

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

Nielsen, M. A. (2015), *Neural Networks and Deep Learning*, San Francisco, Determination Press.

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

Schlkopf, B. and A. J. Smola (2001), *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, 1st Edition, Adaptive Computation and Machine Learning Series, The MIT Press.

Varian, H. R. (2014), "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, 28(2), 3-28.

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

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