Korea Trade Review (Korea Trade Rev. ; 무역학회지)
Vol. 47, No. 2, April 2022 (pp. 69-88)
https://data.doi.or.kr/10.22659/KTRA.2022.47.2.69

貿易學會誌
ISSN 1226-2765 (Print)

# 로지스틱 회귀모형과 머신러닝 모형을 활용한 주요산업의 부산 지역총생산 및 고용 효과 예측

이재득
부산대학교 무역학부 교수

# Prediction on Busan's Gross Product and Employment of Major Industry with Logistic Regression and Machine Learning Model

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

Received 27 March 2022, Revised 22 April 2022 Accepted 29 April 2022

## Abstract

This paper aims to predict Busan's regional product and employment using the logistic regression models and machine learning models. The following are the main findings of the empirical analysis. First, the OLS regression model shows that the main industries such as electricity and electronics, machine and transport, and finance and insurance affect the Busan's income positively. Second, the binomial logistic regression models show that the Busan's strategic industries such as the future transport machinery, life-care, and smart marine industries contribute on the Busan's income in large order. Third, the multinomial logistic regression models show that the Korea's main industries such as the precise machinery, transport equipment, and machinery influence the Busan's economy positively. And Korea's exports and the depreciation can affect Busan's economy more positively at the higher employment level. Fourth, the voting ensemble model show the higher predictive power than artificial neural network model and support vector machine models. Furthermore, the gradient boosting model and the random forest show the higher predictive power than the voting model in large order.

**Keywords:** Industry, Export, Gross Production, Employment, Logistic Regression, Machine Learning
**JEL Classifications:** F10, F13

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

# Ⅰ. 서언

2000년대 들어 세계 각국은 인공지능(Artificial Intelligence: AI)과 제 4차 산업혁명 시대를 맞이하여 빅데이터와 사물인터넷, 그리고 사무자동화와 스마트 팩토리 등의 활용을 급속히 진전시키고 있다. 따라서 우리나라도 이러한 세계 경제 환경의 변화에 대응하여 국가와 산업 그리고 기업의 경쟁력을 강화하기 위해서 노력하고 있다.

부산경제는 1970년대와 1980년대를 거치면서 우리나라의 신발, 합판 등의 경공업 중심의 수출주도형 경제성장에 힘입어 상대적으로 지역경제가 활성화하였다. 그러나 1980년대 후반 이후 우리나라의 수출주도형 정책은 경공업에서 중공업으로 바뀜으로 인해 경공업 중심이던 부산지역경제는 수도권에 비해 상대적으로 낙후하기 시작하였고 정체되어 있다.

특히 부산시는 지역경제의 활성화를 위해 낙후된 부산지역 산업구조의 고도화를 통해 2019년 제 5차 7대 전략산업으로 글로벌관광산업, 라이프케어산업, 미래수송기기산업, 스마트해양산업, 지능정보서비스산업, 지능형기기산업, 그리고 클린테크산업 등을 새롭게 재선정하여 부산지역의 고용창출과 지역소득을 증가시키려고 노력하고 있다. 그리하여 부산지역의 전략산업과 주요산업의 영향에 대한 추정과 예측은 부산지역경제의 발전을 위해서 매우 중요한 연구과제이다.

그리하여 부산시의 산업정책에 관련한 기존 연구들은 주로 전통적인 구조적 경제모형이나 시계열 계량경제학 기법들을 사용하고 있다. 그러나 이러한 전통적인 경제모형이나 시계열 경제모형은 정확한 경제모형을 설정하는데 있어 설정오류가 발생하며, 과대적합 등의 심각한 문제가 발생한다. 그러므로 이들 전통적 모형은 정확히 모형을 설정하기 힘들기 때문에 생기는 오류와 추정모형의 적합성 때문에 종종 잘못된 추정과 예측을 초래할 수 있다.

따라서 이러한 과대적합을 야기하는 전통적인 계량방법의 추정과 예측에서 일어날 수 있는 한계점을 보완하기 위하여 외국에서는 AI를 기반으로 하는 머신러닝(Machine Learning) 등의 기법을 경제학적 분석에서도 최근 2010년대부터 Zou and Hastie (2005)는 회귀모형의 변수선택에 대한 정규화 문제를 연구하였고, Varian (2014)은 빅데이터를 사용하는 새로운 계량기법을 도입하는 등 새로운 기법인 머신러닝에 의한 경제학적 연구를 시작하고 있다. 이와 같이 인공지능 시대 머신러닝 기법은 경제학측면에서는 외국 일부분의 학자들이 적극적으로 시도하고 있다.

그러나 우리나라 경제학 분야에서는 개괄적인 머신러닝 분석에 대한 동향소개는 있지만, 머신러닝 기법들을 사용한 심층적인 연구가 거의 없다. 머신러닝 기법에 의한 분석은 통계와 전산 그리고 경제학의 융복합학적인 학문분야에 대한 어려움도 있고, 4-5년 마다 주요산업들의 구분과 이에 따른 자료수집의 한계 등으로 힘들다. 그리하여 우리나라 경제학 분야에서 머신러닝을 도입한 연구가 극히 미약한 실정이다. 그럼에도 불구하고 AI시대 경제정책의 효과 및 수요 예측 등과 같은 분야에서 최근 외국에서 활발히 적용되고 있는 실정으로 볼 때, 향후 우리나라의 경제학 분야와 경제분석에서도 좀 더 정확한 경제 분석과 예측의 정확성과 엄밀함을 위해 점점 더 필요성이 절실해질 것이다.

그러므로 국가 및 지역의 경쟁력을 제고하기 위해서는 우리나라 주요산업투자와 지역의 전략산업 선정 및 육성 등을 통한 효과적인 지원정책이 필요하다. 그러나 먼저 이러한 산업투자 정책과 지원을 위해서는 좀 더 엄밀한 경제진단과 추정과 예측이 필요하다. 이를 위해서는 전통적인 계량경제기법과 아울러 현재 외국의 경제분석에서 도입을 하고 있는 머신러닝에 의한 분석기법들을 활용하여 좀 더 정확한 경제효과의 진단을 위해서는 좀 더 엄밀한 경제효과 예측과 추정이 절실히 요구된다.

따라서 본 연구는 AI 기법인 머신러닝 기법을 사용하여 부산 전략산업 및 우리나라의 주요산업을 중심으로 부산의 지역총생산 혹은 소득과 고용량에 대한 추정과 예측 및 분석을 다음과 같은 주안점을 두고 분석하고자 한다.

첫째, 2019년 선정된 부산의 전략산업과 우리나라 주요산업에 대한 지원과 투자 및 육성, 부산의 지역경제, 특히 부산의 총생산량 혹은

소득과 고용에 미치는 대한 경제효과를 전통적인 통상최소자승 선형모형(OLS)에 종속변수의 1차 시차를 도입하여 분석한다.

둘째, 산업투자의 경제효과를 분석하기 위하여 기존의 OLS 선형회귀분석과 단일 시계열 분석 등 전통적 회귀모형에서 종종 발생하는 설정오류 문제 등을 완화하기 위하여 로지스틱 회귀분석(Logistic regression)과 머신러닝 모형에 의한 비선형 모형의 회귀분석을 한다.

셋째, 좀 더 엄밀한 예측력을 위해 머신러닝 기법에 의한 AI 머신러닝 기법들을 융합한 보팅(Voting Ensemble) 모형, 랜덤 포레스트(Random Forest) 모형, 그리고 그래디언트 부스팅(Gradient Boosting) 모형 등의 비선형 앙상블 모형(Ensemble Model) 등을 이용한다.

넷째, 이러한 머신러닝 기법 등을 활용하여 부산과 우리나라 주요산업투자 외에도 우리나라는 수출 혹은 무역도 중요하기 때문에 우리나라 원화의 대미 달러 환율 및 수출 변수도 포함하여 경제효과 즉 지역총생산 혹은 소득과 고용에 대한 효과를 추정하고 예측을 하고자 한다. 이를 위해 먼저 각 추정모형의 분류예측을 하고 그 다음에 회귀모형을 통한 수치예측 등의 분석을 하고자 한다.

따라서 본 연구는 우리나라는 물론이고 지역경제 차원에서 머신러닝을 활용한 연구는 거의 없는 생소한 분야로서 아직 본격적으로 연구가 수행되고 있지 않다. 그리하여 본 연구는 기존 전통적 연구방법에 의한 예측력을 보완하기 위하여 머신러닝 기법을 도입함으로써 보다 엄밀히 예측하고 분석함으로써 향후 부산의 전략산업 및 우리나라의 주요산업들의 육성효과를 좀 더 엄밀히 예측하고 추정한다. 그리하여 기존 연구들을 보완하여 머신러닝에 의한 경제예측과 새로운 해석을 창출하고, 나아가 우리나라의 산업을 보다 효과적으로 지원하고 육성하는 데 기여함으로써, 향후 좀 더 많은 후발연구를 유발할 수도 있을 것이다.

## Ⅱ. 선행연구

경제학적 분석을 위해 행하여 온 구조적 모형의 추정이나 시계열 모형에 의한 전통적인 계량방법의 추정과 예측은 모형설정의 오류 등으로 인해 특히 변동성이 심한 팬데믹과 미국과 중국 무역 전쟁, 그리고 러시아와 우크라이나 등의 전쟁 등으로 인한 경제 불확실성의 시대에는 더욱 더 모형예측력에 있어 과대평가가 종종 발생할 수 있다.

그리하여 외국에서는 2010년대부터 이러한 전통적 계량모형에서 일어날 수 있는 추정과 예측의 한계점을 보완하기 위하여 머신러닝 기법을 도입하기 시작하고 있다. 경제정책 효과와 예측력의 정확성을 위해 Varian (2014)은 빅데이터와 인공지능을 활용한 머신러닝 기법을 도입하여 새로운 경제분석을 하는 등, 외국 경제학자들이 최근 선구적으로 도입하고 있다.

최근 경제학 분야에서 좀 더 나은 경제효과 분석과 예측을 위하여 머신러닝 기법을 활용한 외국의 대표적 연구를 보면, Chalfin et al.(2016)는 생산성과 인간자본 관계에 대한 분석을 하였고, Jean et al. (2016)은 빈곤에 대한 머신러닝에 의한 예측을 하였다. Mullainathan and Spiess (2017)은 계량경제학에 머신러닝 기법을 융합한 연구를 하였다. Athey (2017, 2019)는 경제정책에 대한 빅데이터 자료의 사용과 인공지능 경제학에 있어 머신러닝의 충격과 이에 대한 연구를 하였고, Athey and Wager (2018)은 머신러닝 모형인 랜덤 포레스트 모형을 사용하여 이질적 경제정책 효과를 추정하였다.

그리고 Chakraborty and Joseph (2017)은 머신러닝의 영국 중앙은행의 금융정책, 금융규제, 그리고 영국 소비자 물가와 경제모델에 대한 예측을 한 결과, 머신러닝 모형에 의한 예측력이 더 좋은 것으로 나타났다. Agrawal et al. (2018)은 인공지능에 의해 경제예측에 대한 분석을 하였다. Gu et al.(2018, 2019)은 머신러닝에 의한 좀 더 엄밀한 자산가격의 산정에 대한 분석을 하였다. Naecker and Peysakhovich (2017)은 머신러닝의 리스크의 행동모델에 대한 연구를 하였고, Acemoglu and Restrepo (2019)는 현재 AI와 4차산업에 있어 중요한 인공지능에 의한 로봇과 노동자들의 직업 대체에 관한 연구를 하였다. Kreif and DiazOrdaz (2019)는 머신러닝에 의한 새로운 인과관계 추

**Table 1.** Prediction and Actual Fact

| Prediction and Actual Fact | | Prediction | |
|---|---|---|---|
| | | True | False |
| Acual Fact | True | True Positive: TP | False Negative: FN |
| | False | False Positive: FP | True Negative: TP |

정에 의한 정책평가를 연구하였다.

그러나 우리나라에서 머신러닝 기법에 의한 경제학적 분석은 그 중요성에도 불구하고 거의 없는 실정이다. Park Ki-Young, and Ko Jeong-Won(2019)은 머신러닝 기법에 기반하는 경제학 분야에서 외국 연구동향에 대해서 간략히 소개하고 있다. Kim Soo-Hyon (2020)은 우리나라의 환율의 단기 금융시장에서 변동성에 대한 딥러닝 기법을 연구하였고, Yi Chae-Deug (2021a, 2021b)은 머신러닝 기법을 이용한 부산지역 전략산업의 경제효과에 대한 연구를 하였지만, 본 연구는 부산의 전략산업뿐만 아니라, 우리나라 주요산업들에 대한 투자와 수출, 그리고 환율 등이 부산지역의 총생산과 고용에 대한 효과를 로짓 회귀모형과 머신러닝 모형을 이용하여 예측하고 있다. 그리하여 우리나라 주요산업의 투자효과 등에 대해 머신러닝을 이용한 경제적 분석은 아직 거의 없는 실정이다.

## Ⅲ. 로지스틱 회귀모형과 머신러닝 모형

### 1. 분류 모형 평가

먼저 모형을 평가하기 위해 분류모형과 수치회귀모형으로 나누어 분석한다. 분류모형의 평가를 위한 정확도(Accuracy)는 모형이 전체 데이터를 얼마나 정확하게 분류하는가를 나타내는 평가지표이다. 모형의 정밀도(precision)는 모형이 참(true) 혹은 거짓(false)으로 먼저 예상한 것 중 실제로 참 혹은 거짓이 얼마나 되는가를 판단한다.

재현율(recall)은 실제 참 혹은 거짓 데이터 중에서 참 혹은 거짓으로 분류한 비율을 나타낸다. 다음과 같이 모형의 정밀도(Precision)와 재현율(Recall), 그리고 정밀도와 재현율을 조화 평균한 값인 F1-score를 다음과 같이 〈Table 1〉에 나타나 있다.

정확도(Accuracy)

$$= \frac{TP+TN}{TP+FN+FP+TN} \text{ ,}$$

정밀도(Precision) $= \frac{TP}{TP+FP}$ ,

F1 Score

$$= 2 \times \frac{\mathrm{Pr}ecision \times F1\,Score}{\mathrm{Pr}ecision + F1\,Score}$$

본 연구에서 머신러닝 회귀모형에 의한 수치예측을 위한 회귀모형을 설정하여 비교분석하는데, 대표적인 회귀모형에 의한 예측 정확도 검증 방법인 평균제곱근오차(RMSE: Root Mean Squared Error)와 독립변수들의 종속변수에 대한 예측력을 나타내는 결정계수($R^2$)를 주로 이용한다.

### 2. 로지스틱 회귀모형

로지스틱 회귀모형 (Logistic Regression은 이항형인 자료를 사용하였을 때 종속 변수(Y)의 결과가 범위[0,1]로 제한하여 독립변수(X)가 주어졌을 때 종속변수(Y)가 1의 범주에 속할 확률 $p(y=1|x)$를 의미한다. 종속변수의 범주의 개수가 두 개인 이항형 로지스틱 회귀(Binomial logistic regression)와 세 개 이상의 범주를 가지는 다항 로지스틱 회귀(Multinomial

logistic regression) 모형이 있다. 본 연구의 실증분석에서는 이항과 다항 분석을 할 것이다.

로지스틱 회귀분석에서 각 독립변수가 종속변수에 미치는 영향은 다음과 같은 오즈비율(Odds Ratio)로 해석된다.

$$\text{Odds} = \frac{p(y=1|x)}{1-p(y=1|x)}$$

오즈 비율에 로그를 취한 함수로서 입력 값의 범위가 [0,1] 일 때 출력 값의 범위를($-\infty, \infty$)로 조정하는 로짓(Logit)변환을 한다.

$$\text{Logit(p)} = \log \frac{p}{1-p}$$

로지스틱 회귀모형은 종속변수와 독립변수 사이의 관계에 있어서 선형 모델과 차이점을 지니고 있다. 로지스틱 회귀모형은 종속변수와 독립변수 사이의 관계에 있어서 종속변수를 이항형인 자료를 사용하며 종속 변수 y의 결과가 범위[0,1]로 제한되며, 종속 변수의 조건부 확률(P(y | x))의 분포가 정규분포 대신 이항 분포를 따른다. 그리하여 일반적인 전통적 선형 회귀모형과 차이를 가지고 있다.

따라서 로지스틱 회귀모형은 로지스틱 함수 또는 시그모이드 함수(Sigmoid Function)들의 로짓변환을 하며, 주로 다음과 같은 로지스틱 함수(Logistic Function)를 가진다.

$$\text{Logistic Function} = \frac{e^{\beta_i X_i}}{1+e^{\beta_i X_i}} .$$

로지스틱 회귀분석에서 각 독립변수가 종속변수에 미치는 영향은 Kim. H.S. (2020)의 연구를 참고하면, 오즈비율(Odds Ratio)로 해석되며, 이것은 다른 독립변수가 일정할 때와 어느 특정한 독립변수($x_i$)가 1 단위가 증가하였을 때 변화하는 종속변수의 효과를 의미하며 이것은 두 오즈의 비율을 다음과 같이 나타낸다.

$$\frac{odds(x_1, x_2, x_3.., x_i+1, ..., x_n)}{odds(x_1, x_2, x_3..., x_n)} = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+ ...\beta_i(x_i+1)...+\beta_n x_n}}{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+ ...+\beta_n x_n}} = e^{\beta_i}$$

따라서 추정된 독립변수의 추정계수($\beta_i$)는 오즈비율인 $e^{\beta_i}$의 값으로 해석된다. 즉 독립변수($x_i$)가 1 단위가 증가하였을 때 목표집단에 속할 확률이 속하지 않을 확률에 비해 $e^{\beta_i}$의 값으로 나타난다. 여기서 추정계수($\beta_i$)값이 양수로 나와 $e^{\beta_i}$의 값이 1보다 크면. 대상 목표집단에 속할 확률이 증가하여 긍정적인 효과가 더 크다는 것을 나타난다. 반면에 추정계수($\beta_i$)값이 음수로 나왔다면 이 $e^{\beta_i}$의 값은 1보다 작게 나오고 이것은 목표집단에 속하지 않고 목표집단에서 이탈될 확률이 더 커서 부정적인 효과가 크다는 것을 의미한다.

## 3. 랜덤 포레스트(Random Forest) 머신러닝 모형

랜덤 포레스트 모형은 나무가 하나 있는 단일 의사결정 나무(Decision Tree)모형을 확장한 것이다. 본 연구에서는 Hastie et al (2017)의 연구에 따라 먼저 의사결정 나무모형은 회귀를 위해 N개의 각각의 관측치에 대해 p개의 자료들을 투입한 독립변수 $x$와 1개의 반응물인 $y$로 구성, 즉 $(x_i, y_i)$, $i$=1,2,..., N, $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$로 되어 있다고 상정한다. 그리고 한 개의 분할을 R1, R2,..,RM의 m 지역으로 나누어 변수들을 분리한 후, 잔차 자승합을 최소화시키는 $\widehat{c_m}$를 구하여 다음 식을 최소화시키는 분할변수 j와 분할점 s를 찾는다.

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

그 다음에 나무의 최적 크기를 결정하기 위해 최소 노드를 가진 큰 나무($T_0$)의 부분집합인 나무 T($T \subset T_0$)를 구하는 것이다. 여기서 터미널 노드 영역을 $R_m$으로, $|T|$를 T에서의 터미널 노드의 개수를 각각 나타내면, 다음과 같다.

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

여기서 $Q_m(T)$는 자승오차의 노드 불순도의 정도를 측정하므로, 비용 복잡성을 나타내는 기준은 다음과 같이 정의된다.

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

위 식에서 $\alpha$는 모형의 복잡성을 조절하는 튜닝 모수로서, 각 $\alpha$에 대하여 $C_\alpha(T)$를 최소화시키는 하위 나무(subtree) $T_\alpha(T_\alpha \subseteq T_0)$를 구한다. 그리하여 랜덤 포레스트 모형은 상관관계가 없는 의사결정 나무들을 모은 후, 그것들의 평균을 구하는 것이다. 그리하여 랜덤 포레스트 회귀분석의 예측치 $\hat{f}_{RF}^{B}(x)$는 다음과 같이 구해진다.

$$\hat{f}_{RF}^{B}(x) = \frac{1}{B} \sum_{b=1}^{B} T(x_i;\Theta_b).$$

## 4. 그레디언트 부스팅(Gradient Boosting) 머신러닝 모형

그래디언트 부스팅 나무모형은 나무의 끝 마디 영역 $R_j$에, 상수 $\gamma_j$는 각 영역에서 다음과 같이 표시된다.

$$f(x) = \gamma_j,\ x \in R_j,$$

그리고 패러미터 $\Theta = \{R_j, \gamma_j\}_1^J$가 주어져 있을 때, 한 개의 나무($T$)에 대해서 그 패러미터는 손실(L)을 최소화 시켜서 구할 수 있다.

$$T(x;\Theta) = \sum_{j=1}^{J} \gamma_j I(x \in R_j)\,,$$
$$\hat{\Theta} = \arg\min_{(\Theta)} \sum_{j=1}^{J} \sum_{x_i \in R_j} L(y_i, \gamma_j).$$

여기서 부스팅된(Booted) 나무모형은 이러한 나무들의 합으로 다음과 같이 이루어진다.

$$f_M = \sum_{m=1}^{M} T(x;\Theta_m)\,.$$

이 때 전방 단계 방식으로 현재 모형이 $f_{m-1}(x)$로 주어져 있다고 상정하면, 그것의 패러미터의 추정량에 대해서 다음과 같이 해를 구한다.

$$\hat{\Theta}_m = \arg\min_{(\Theta_m)} \sum_{i=1}^{N} L(y_i, f(x_i) + T(x_i; f(\epsilon$$

그리고 훈련용 데이터를 사용하여 $y$를 추정하기 위해서 $f(x)$를 이용한 손실함수(L)는 다음과 같이 주어진다고 상정한다. 그 때 $f$의 값은 손실함수 $L(f)$에 대한 최소화 문제를 통하여 $f$의 예측치를 구할 수 있다.

$$L(f) = \sum_{i=1}^{N} L(y_i, f(x_i)),$$
$$\hat{f} = \arg\min_{(f)} [L(f)],$$
$$f = \{f(x_1), f(x_2), ..., f(x_N)\}^T.$$

그리하여 $f$에 대한 최적화 문제는 다음과 같은 벡터를 합함으로써 구해진다. 여기서 초기

**Table 2.** Time-Lag OLS Regression Model

| | | | | | | |
|---|---|---|---|---|---|---|
| Adj. R-squared: | 0.999 | | | | | |
| F-statistic: | 4639. | | | | | |
| Prob (F-statistic): | 2.73e-37 | | | | | |
| Industry | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
| Intercept | 3.7247 | 0.556 | 6.695 | 0.000 | 2.579 | 4.870 |
| YL | 0.2973 | 0.112 | 2.646 | 0.014 | 0.066 | 0.529 |
| Coal and Petroleum Chemistry | 0.1071 | 0.058 | 1.847 | 0.077 | -0.012 | 0.226 |
| Non-Metal and Metal | -0.0080 | 0.040 | -0.198 | 0.844 | -0.091 | 0.075 |
| Electricity and Electronics | 0.1026 | 0.049 | 2.094 | 0.047 | 0.002 | 0.204 |
| Machine and Transport | 0.1869 | 0.069 | 2.726 | 0.012 | 0.046 | 0.328 |
| Information and Telecommunication | 0.0981 | 0.053 | 1.862 | 0.074 | -0.010 | 0.207 |
| Finance and Insurance | 0.1237 | 0.046 | 2.684 | 0.013 | 0.029 | 0.219 |
| **Omnibus:** | | **1.772** | **Durbin-Watson:** | | | **1.858** |
| **Prob(Omnibus):** | | **0.412** | **Jarque-Bera (JB):** | | | **1.098** |
| | | | **Prob(JB):** | | | **0.577** |

추측값은 $f_0 = h_0$ 이고, 이전의 유도된 패러미터 벡터인 $f_{m-1}$이 주어져 있다면, $f_m$ 은 새롭게 도출된 패러미터 벡터이다.

$$f_M = \sum_{m=0}^{M} h_m,\ h_m \in R^N.$$

따라서 최적화 문제는 각 시기의 증분 벡터 $h_m$에 의해 달라지는데, 현재기에 $f_{m-1}(x_i)$이 주어져 있을 때 그래디언트 부스팅은 각 기간에 있어 $\hat{\Theta}$ 를 최소화 시킨다. 최소화를 위해서는 $m$ 번째 반복되는 시기에 있어 나무에 대한 예측치 $t_m$이 음의 그래디언트에 가장 근접하게 나무 $T(x;\Theta_m)$를 유도함으로써 이는 자승오차 기법을 사용하여 다음과 같이 구할 수 있다.

$$\hat{\Theta}_m = \arg\min(\Theta)\sum_{i=1}^{N}(-g_{im} - T(x_i;\Theta))^2$$

## Ⅳ. 로지스틱 회귀모형에 의한 부산 경제효과 분석

### 1. 국내 산업의 지역총생산에 대한 시차 OLS 회귀분석

본 절에서는 6개의 주요산업들인 석탄 및 석유화학, 비금속광물 및 금속, 전기 및 전자, 기계 운송, 정보통신, 금융보험 산업들이 부산의 소득(GRDP)에 각각 어떻게 영향을 미치는지 분석하기 위하여 1985년부터 2018년까지의 우리나라 광공업통계에 나와 있는 연도별 생산액에 대한 자료를 사용하여 추정하였다. 이러한 6개의 주요산업 들은 부산이 선정한 전략산업들이 4-5년 마다 바뀌고 일관성 있는 자료가 없기 때문에 대부분 부산의 전략산업 포함하고 있거나 부산이 선정한 전략산업과 밀접한 연관이 있는 6대 주요산업들을 채택하였다.

이를 추정하기 위해 먼저 최소자승 추정 회귀모형(OLS regression)을 취하였다. 회귀모형에서는 부산의 6대 주요산업들을 설명변수들로 하고 부산의 지역총생산(GRDP) 혹은 소득을 종속변수(Y)로 채택하여 모든 변수에 대해 로그를 취하여 추정하였다. 그리고 년도별 시계열 변수들이 시간의존성을 갖고 있기 때문에

자기 상관성을 줄이기 위해, Arellano and Bond (1991)의 연구를 원용하여 경제변수의 동태적 특성을 감안하고 반영하기 위해 종속변수의 1차 시차변수(YL)를 또한 독립변수로 추가로 도입하여 추정한 결과는 〈Table 2〉에 나타나 있다.

그리하여 그 추정결과는 〈Table 2〉에 나와 있듯이 OLS 모형으로 추정한 결과, 1차 시차변수는 5% 유의수준에서 양의 유의한 값을 가지고 있으며, 부산 산업들의 석탄 및 석유화학 산업과 정보통신산업은 10% 유의수준에서, 전기 및 전자, 기계운송, 금융보험 산업들은 부산의 지역총생산에 5% 유의수준에서 각각 유의한 정의 영향을 미치고 있는 것으로 나타났다.

그러나 비금속광물 및 금속산업은 10% 유의수준에서 부산의 GRDP에 유의적인 영향을 미치지 않는 것으로 나타났다. 이 회귀모형의 결정계수는 아주 높은 것으로 나타났고 자기상관성을 측정하는 더빈-왓슨(Durbin-Watson)통계량은 1.858으로 나타나 자기 상관성 문제는 나타나지 않았다. 한편, 확률(Omnibus)은 0.412, 확률(JB)은 0.577로 각각 나타나 정규분포하고 있다는 귀무가설을 5% 유의수준에서 기각할 수 없는 것으로 나타났다.

## 2. 부산 전략산업의 지역총생산에 대한 로지스틱 회귀분석

전통적 OLS 선형모형은 모형의 설정오류가 발생하므로 본 절에서는 부산의 전략산업들이 부산의 소득에 미치는 영향을 분석하기 위하여 비선형 로지스틱 회귀모형을 설정하여 추정하고자 한다.

따라서 로지스틱 모형에 의해 부산의 전략산업의 소득효과를 예측하기 위하여 부산의 7대 전략산업들인 글로벌관광산업(a), 라이프케어(b), 미래수송기기(c), 스마트해양산업(d), 지능정보서비스(e), 지능형기기(f), 그리고 클린테크(g) 등의 산업들의 생산액을 독립변수들로 삼고 부산의 지역총생산을 종속변수로 각각 채택하였다. 그 후 모든 변수들의 로그를 취하여 부산의 소득에 대한 전략산업들의 효과를 분석하기 위하여 로지스틱 회귀분석 모형을 사용하여 예측하였다.

이를 위해 평가용 데이터 세트의 독립변수의 스케일을 표준화하였고, 로지스틱 회귀분석에서도 규제항의 인자를 나타내는 C값에 의해 규제의 강도를 표시하였다. C값이 너무 크면 규제강도가 작아져 모형의 복합도가 높아져 과대적합이 발생할 확률이 높아진다. 반면에 C값이 작으면 규제강도가 높아져 학습용 데이터가 모형에 제대로 반영되지 않는 것으로 나타나, 적절한 학습용 데이터 모형이 되지 못한다. 그리하여 여러 C값에 의해 구분해본 결과 나온 C값에 다음과 나누어 분석한 결과가 〈Table 3〉에서 나타나 있다.

### 1) C = 0.1

첫째, 로지스틱 모형에서 규제강도 C= 0.1일 때의 모형을 평가하면 다음과 같다. 먼저, 학습용 데이터 세트의 정확도는 0.714, 그리고 평가용 데이터 세트의 정확도는 0.556으로 나타나, 모형의 평가를 나타내는 평가용 데이터 세트의 정확도는 높지 않다. 그리고 학습용 데이터 세트의 정확도는 평가용 데이터 세트의 정확도 보다 높게 나왔으므로 약간의 과대적합이 있을 확률이 있다.

그리고, 로지스틱 모형의 성능 평가를 위해 분류 보고서에 있는 정밀도, 재현율, F1 스코어를 살펴보면 〈Table 3〉에서 나타난 것과 같이, 0의 전국 영역과 높은 1인 부산 영역에서 정확도는 각각 0.57, 0.50으로 나왔다. 재현율은 전국 0.80, 부산 0.25로 나타났다. F1 스코어는 0.67과 0.33으로 각각 나타났다. 그리하여 클래스별 각 성과지표의 단순 평균값과 가중 평균값이 정확도가 0.54로 각각 나타났다. 따라서 모형의 예측 성능은 보통인 것으로 나타났다.

둘째, C= 0.1일 때의 로지스틱 회귀모형의 회귀계수를 보면, 글로벌관광산업(a)= -0.130, 라이프케어산업(b)= -0.104, 미래수송기기산업(c) = -0.065, 스마트해양산업(d)= -0.110, 지능정보서비스산업(e)= -0.202, 지능형기기산업(f)= -0.123, 그리고 클린테크산업(g)= -0.113 등으로 각각 나타났다. 로지스틱 회귀계수들은

**Table 3.** Classification and Regression with a Logistic Regression Model

| | | |
|---|---|---|
| C = 0.1 | Model Accuracy | Accuracy of Training Data Set of Logistic Model: 0.714<br>Accuracy of Test Data Set of Logistic Model : 0.556 |
| | Model Performance of Logistic Classification | precision recall f1-score<br>0 0.57 0.80 0.67<br>1 0.50 0.25 0.33<br>accuracy 0.56<br>macro avg 0.54 0.53 0.50<br>weighted avg 0.54 0.56 0.52 |
| | Regression Coefficicent | a Coefficient: −0.130, b Coefficient: −0.104, c Coefficient: −0.065, d Coefficient: −0.110, e Coefficient: −0.202, f Coefficient: −0.123, g Coefficient: −0.113 |
| C = 0.001 | Model Accuracy | Accuracy of Training Data Set of Logistic Mode: 0.524<br>Accuracy of Test Data Set of Logistic Model : 0.444 |
| | Model Performance of Logistic Classification | precision recall f1-score<br>0 0.44 1.00 0.62<br>1 0.00 0.00 0.00<br>accuracy 0.44<br>macro avg 0.22 0.50 0.31<br>weighted avg 0.20 0.44 0.27 |
| | Model Performance of Logistic Classification | a Coefficient: −0.005, b Coefficient: −0.005, c Coefficient: −0.004, d Coefficient: −0.005, e Coefficient: −0.006, f Coefficient: −0.005, g Coefficient: −0.005 |

각각 exp(-$\beta$i)로 표시되어 있다. 그러므로 exp($\beta$i)값이 클수록 어떤 독립변수가 한 단위 증가할 때 부산지역 소득이 증가한다고 할 수 있다. 그런데 여기서 C= 0.1 일 때 모형의 회귀계수들은 모두 $\beta$값이 음으로 나타났다. 그러므로 -$\beta$의 절대값이 클수록 exp(-$\beta$i)= 1/exp($\beta$i)은 작아진다.

그러므로 미래수송기기산업과 라이프케어산업, 그리고 스마트해양산업의 순으로 부산의 소득에 기여하는 것이 크며, 글로벌관광, 지능형기기산업, 지능형정보서비스산업, 클린테크산업 등이 상대적으로 낮게 나타났다.

### 2) c=0.001

첫째, 규제강도가 C= 0.001 일 때 로지스틱 회귀모형을 평가하면 다음과 같다. 그리하여 〈Table 3〉에서 나타난 것과 같이, 학습용 데이터 세트의 정확도는 0.524, 그리고 평가용 데이터 세트의 정확도는 0.444로 각각 나타나, 모형의 평가를 나타내는 평가용 데이터 세트의 정확도는 다소 낮게 나타났다. 그리고 학습용 데이터 세트의 정확도는 평가용 데이터 세트의 정확도 보다 약간 높게 나왔으므로 아주 약간의 과대적합이 있을 가능성도 조금 있다.

그리고, 로지스틱 모형의 성능 평가를 위해 분류 보고서에 나타나 있는 것을 살펴보면, 정밀도, 재현율, F1 스코어를 살펴보면, 전국 영역인 0의 영역에서 정밀도는 각각 0.44, 0.00으로 낮게 나타났고, 1의 영역에서 정밀도는 0.00, 1의 영역에서는 재현율은 0.20으로 낮게 나타났고, 단순 평균값과 가중 평균값에 대한 F1 스코어는 0.44와 0.31로 각각 다소 낮게 나타났다. 그리하여 모형의 예측 성능은 다소 낮은 것으로 나타났다.

둘째, C= 0.001 일 때의 로지스틱 회귀모형의

회귀계수를 보면, 글로벌관광산업(a)= -0.005, 라이프케어산업(b)=-0.005, 미래수송기기산업(c)=-0.004, 스마트해양산업(d)= -0.005, 지능정보서비스산업(e)=-0.006, 지능형기기산업(f)=-0.005, 그리고 클린테크산업(g)=-0.005 등으로 모두 음수로 비슷하게 나타났다. 그러므로 미래수송기산업이 부산의 소득에 가장 크게 기여하는 것으로 나타났지만, 그 외 다른 산업들은 비슷하게 소득을 증가시키는 것으로 나타났다.

## 3. 국내 산업과 수출 및 환율의 고용에 대한 로지스틱 회귀분석

### 1) 이항 로지스틱 회귀모형 추정

본 절에서는 1999년 3분기부터 2020년 2분기까지 분기별 자료를 사용하여 부산의 고용자수(천명)를 종속변수로 삼고, 독립변수들로서 한국의 기계류 투자지수(a), 정밀기기 투자지수(b), 운송장비 투자지수(c), 부산의 경제활동인구(d), 한국의 투자 총지수(e), 우리나라 원화의 대미 달러 환율(f), 그리고 한국의 수출(g) 등을 삼아, 이들 7개 독립변수들이 부산의 고용에 미치는 영향을 추정하였다.

여기서 모든 독립변수들과 종속변수들에 로그를 취하여 부산의 취업자 수에 대한 먼저 이항 로지스틱 회귀분석으로 추정하였다. 이항 로지스틱 회귀모형에 의한 분석을 하기 위해 본 절에서는 부산의 고용자 혹은 취업자 수가 1,650(천명) 이하이면 0으로 두고, 그 취업자 수가 만약 1,650(천명)을 초과하면 1로 나누어 영역을 0과 1로 나눈 명목변수를 종속변수로 삼는다.

본 절에서는 머신러닝에 의한 이들 독립변수들이 얼마나 잘 예측하는가에 초점을 두고 이들에 대한 모형의 평가를 하기 위해서 주로 학습용 데이터 세트와 평가용 데이터 세트에 대한 정확도(accuracy)를 구하였다. 그리하여 로지스틱 회귀분석에서 규제의 강도를 나타내는 C의 값에 따라서 구한 모형의 정확도와 재현율, 그리고 F1-score는 다음 〈Table 4〉에 나타나 있다.

#### (1) C= 1

첫째, 로지스틱 모형에서 규제강도 C= 1 일 때의 모형을 평가하면 다음과 같이 〈Table 4〉에서 나타나 있는 것과 같이 먼저, 학습용 데이터 세트의 정확도는 0.965, 그리고 평가용 데이터 세트의 정확도는 0.920으로 나타나, 학습용 데이터 세트에서 학습이 잘 되었으며, 평가용 데이터 세트의 정확도는 높게 나왔으므로 모형의 평가가 좋다고 할 수 있다. 그리고 학습용 데이터 세트의 정확도가 평가용 데이터 세트의 정확도보다 조금 높게 나왔으므로 아주 약간의 과대적합이 있을 확률이 있지만 크지 않다.

그리고, 로지스틱 모형의 성능 평가를 위해 분류 보고서(Classification report) 함수를 사용하여 나타낸다. 여기에서 클래스별 각 성과지표의 단순 평균값을 나타내며, 평균 가중치(weighted avg)는 클래스 별 각 성과지표를 표본 수를 바탕으로 가중 평균한 것이다. 분류 보고서에 있는 정밀도, 재현율, F1 스코어를 살펴보면 취업자 수의 영역이 낮은 0의 영역과 높은 1의 영역 모두 정밀도, 재현율, F1 스코어가 모두 높은 것으로 나타나 모형의 예측 성능이 높은 것으로 나타났다.

둘째, 규제강도 C= 1 일 때 로지스틱 모형으로 추정한 회귀계수를 보면, 취업자 수에 대한 한국의 기계류 투자지수(a)=0.040, 정밀기기 투자지수(b)=0.376, 운송장비 투자지수(c)=0.341, 부산의 경제활동인구(d)=2.376, 한국의 투자 총지수(e)=0.059, 한국 원화의 대미 달러 환율(f)=0.659, 그리고 한국의 수출(g)=0.988 등으로 나타났고 각각의 로지스틱 회귀계수들은 각각 $\exp(\beta i)$로 모두 $\beta$값이 양으로 표시되어 있다. 그러므로 $\exp(\beta i)$값이 클수록 어떤 독립변수(i)가 한 단위 증가할 때 취업자 수가 높아질 가능성이 크다고 할 수 있으므로 우리나라의 정밀기기산업, 운송장비 산업, 기계류 산업 등에 대한 투자액이 클수록 부산의 고용이 증가하는 것으로 알 수 있다.

특히 대외부문이 한국의 수출이 증대할수록 부산의 취업자들이 증가하여 고용이 늘어나고 있다. 그리고 한국 원화의 대미 달러 환율이 높아질수록 한국의 원화가 떨어져서 대외 시장에

**Table 4.** Classification and Regression with a Binomial Logistic Regression

| | | |
|---|---|---|
| C = 1 | Model Accuracy | Accuracy of Training Data Set of Logistic Model : 0.965<br>Accuracy of Test Data Set of Logistic Model : 0.920 |
| | Model Performance of Binomial Logistic Classification | precision recall f1-score<br>0 1.00 0.88 0.93<br>1 0.82 1.00 0.90<br>accuracy 0.92<br>macro avg 0.91 0.94 0.92<br>weighted avg 0.93 0.92 0.92 |
| | Binomial Logistic Regression Coefficicent | a Coefficient: 0.040, b Coefficient: 0.376, c Coefficient: 0.341, d Coefficient: 2.376, e Coefficient: 0.059, f Coefficient: 0.659, g Coefficient: 0.988 |
| C=0.1 | Model Accuracy | Accuracy of Training Data Set of Logistic Model : 0.930<br>Accuracy of Test Data Set of Logistic Model : 0.880 |
| | Model Performance of Binomial Logistic Classification | precision recall f1-score<br>**0 0.93 0.88 0.90**<br>**1 0.80 0.89 0.84**<br>**accuracy 0.88**<br>**macro avg 0.87 0.88 0.87**<br>**weighted avg 0.89 0.88 0.88** |
| | Logistic Regression Coefficicent | a Coefficient: 0.161, b Coefficient: 0.212, c Coefficient: 0.393, d Coefficient: 0.878, e Coefficient: 0.210, f Coefficient: 0.237, g Coefficient: 0.257, |
| C=0.01 | Model Accuracy | Accuracy of Training Data Set of Logistic Model : 0.877<br>Accuracy of Test Data Set of Logistic Model : 0.800 |
| | Model Performance of Binomial Logistic Classification | **precision recall f1-score**<br>**0 0.87 0.81 0.84**<br>**1 0.70 0.78 0.74**<br>**accuracy 0.80**<br>**macro avg 0.78 0.80 0.79**<br>**weighted avg 0.81 0.80 0.80** |
| | Binomial Logistic Regression Coefficicent | a Coefficient: 0.090, b Coefficient: 0.092, c Coefficient: 0.138, d Coefficient: 0.184, e Coefficient: 0.104, f Coefficient: 0.043, g Coefficient: 0.087, |

서 경쟁력이 증가하여 수출이 증가하므로, 부산에서도 취업자가 늘어나고 부산의 고용 역시 증가하는 것으로 나타났다.

### (2) C= 0.1

첫째, 회귀모형에서 규제강도 C=0.1 일 때도 학습용 데이터 세트의 정확도는 0.930, 그리고 평가용 데이터 세트의 정확도는 0.880으로 각각 나타나, 평가용 데이터 세트에 대한 모형이 C=1 일 때 보다는 조금 낮지만 여전히 높은 것으로 나타났다. 그러므로 학습용 데이터 세트의 정확도가 평가용 데이터 세트의 정확도보다 아주 약간 높게 나와 과대적합 문제가 있을 가능성은 낮은 것으로 보인다.

한편, 로지스틱 모형의 성능 평가를 위해 모형의 정밀도, 재현율, F1 스코어를 살펴보면 〈Table 4〉에서 나타난 것과 같이, 취업자 수의 영역이 낮은 0의 영역과 높은 1의 영역 모두 정밀도, 재현율, F1 스코어가 모두 0.80보다 높은 것으로 나타나 모형의 예측 성능이 높은 것으로 나타났다.

둘째, 규제강도 C= 0.1 일 때 로지스틱 모형으로 추정한 회귀계수를 보면, 부산의 취업자 수에 대한 한국의 기계류 투자지수(a)=0.161,

정밀기기 투자지수(b)=0.212, 운송장비 투자지수(c)=0.393, 부산의 경제활동인구(d)=0.878, 한국의 투자 총지수(e)=0.210, 한국 원화의 대미 달러 환율(f)=0.237, 그리고 한국의 수출(g)=0.257 등으로 나타났다. 그리하여 로지스틱 회귀계수들은 각각 exp($\beta$i)로 모두 $\beta$값이 양으로 표시되어 있다. 그러므로 우리나라의 운송장비 산업, 정밀기기산업, 기계류 산업 등에 대한 투자액이 클수록 부산의 고용이 증가하는 것으로 알 수 있다.

그리고 특히 대외부문이 한국의 수출이 증대할수록, 그 다음에 한국 원화의 대미 달러 환율이 높아질수록 부산에서 취업자가 늘어나고 부산의 고용 역시 증가하는 것으로 나타났다.

### (3) C= 0.01

마지막으로 규제강도 C=0.01 일 때도 추정계수를 보면, C=1일 때와 같이 취업자 수에 대한 한국의 기계류 투자지수(a), 정밀기기 투자지수(b), 운송장비 투자지수(c), 부산의 경제활동인구(d), 한국의 투자 총지수(e), 한국 원화의 대미 달러 환율(f), 그리고 한국의 수출(g) 등의 영향을 나타내는 로지스틱 모형에 의한 회귀계수들은 대부분 계수의 수치에 있어 약간의 차이는 있었지만, 모두 양수로 나타났다.

그러므로 이들 우리나라의 주요산업 부문들인 기계류 투자, 정밀기기 투자, 운송장비 투자지수(c), 한국의 투자 총 지수(e) 등 투자가 증가할수록 부산의 취업자가 늘어나 고용도 증가하는 것으로 나타났다. 앞에서와 마찬가지고 운송장비 투자지수가 높을수록 부산의 고용 역시 제일 크게 증가하는 것으로 나타났다. 그리고 한국 원화의 대미 달러 환율(f)이 높아지고, 그리고 한국의 수출(g) 등이 늘어날 때 부산의 취업자 수 혹은 부산의 고용자 수 역시 증가시키는 것으로 나타났다.

## 2) 다항 로지스틱 회귀모형 추정

본 절에서는 다항 로지스틱 회귀분석으로 부산의 취업자 수를 바탕으로 3개 그룹으로 나누었는데 취업자 수가 163만 이하이면 0, 163만부터 167.5만 이하이면 1, 그리고 167.5만을 초과하면 2로 두었다. 그리하여 부산의 고용에 영향을 미칠 수 있는 주요 변수들인 7개의 독립변수들인 기계류 투자, 정밀기기 투자, 운송장비 투자지수(c), 부산의 경제활동인구(d), 한국의 투자 총지수(e), 한국 원화의 대미 달러 환율(f), 그리고 한국의 수출(g) 등이 늘어날 때 취업자 수 혹은 부산의 고용에 대한 영향을 규제강도 C에 따라서 추정하였다.

첫째, 다항 로지스틱 회귀분석을 한 결과, 〈표 5〉에 나타난 것과 같이, 첫째, 로지스틱 모형에서 규제강도 C= 1, C=0.1, 그리고 C =0.05일 때의 모형을 평가하면 다음과 같다. 이항 로지스틱 회귀분석 경우보다 3개의 규제강도 경우에 있어서 대체로 학습용 데이터 세트의 정확도와 평가용 데이터 세트의 정확도는 약간 낮아졌지만, 여전히 학습용 데이터 세트의 정확도가 평가용 데이터 세트의 정확도보다 조금 높게 나왔으므로 약간의 과대적합이 있을 수 있지만, 평가용 데이터 세트의 정확도는 각각 0.640, 0.560, 그리고 0.600 등으로 나와서 비교적 양호하게 나타났다.

그리고, 로지스틱 모형의 성능 평가를 위해 분류 보고서를 보면 C= 1, C=0.1, 그리고 C =0.05일 때, 정밀도, 재현율, F1 스코어를 살펴보면 〈Table 5〉에서 나타난 것과 같이, 취업자 수의 영역이 0의 영역, 1의 영역, 2의 영역 모두 정밀도가 0.5보다 높이 나타났다, 재현율, F1 스코어는 C=0.1 일 때 다소 낮게 나타났다.

둘째, 규제강도에 따른 다항 로지스틱 회귀분석의 추정계수를 보면, Python에서는 다항분류를 위한 로지스틱 회귀함수 모형의 복잡도를 나타내는 C와 최적화문제를 해결하는 알고리즘을 설정할 수 있는데, 본 절에서는 newton-cg 알고리즘을 주로 이용하여 다항 로짓분석에서 각 독립변수에 대한 추정계수를 구하였다.

그리하여 기계류 투자지수(a), 정밀기기 투자지수(b), 운송장비 투자지수(c), 부산의 경제활동인구(d), 한국의 투자 총지수(e), 한국 원화의 대미 달러 환율(f), 그리고 한국의 수출(g) 등의 주요산업 부문들에 대한 투자설비 증가와 수출 증가와 환율인상 등의 부산의 고용에 미치는 영향을 나타내는 추정 회귀계수가 〈Table 5〉에 나타나 있다.

**Table 5.** Classification and Regression with a Multinomial Logistic Regression

| | | |
|---|---|---|
| C = 1 | Model Accuracy | Accuracy of Training Data Set of Logistic Model : 0.807<br>Accuracy of Test Data Set of Logistic Model : 0.640 |
| | Model Performance of Multinomial Logistic Classification | precision recall f1-score<br>0 0.62 0.71 0.67<br>1 0.70 0.58 0.64<br>2 0.57 0.67 0.62<br>accuracy 0.64<br>macro avg 0.63 0.65 0.64<br>weighted avg 0.65 0.64 0.64 |
| | Multinomial Logistic Regression Coefficicent | -- Employment 0: Coefficient---<br>a Coefficient: -0.048, b Coefficient: -0.149, c Coefficient: -0.223, d Coefficient: -0.707, e Coefficient: -0.077, f Coefficient: 0.099, g Coefficient: -0.042<br>-- Employment 1: Coefficient---<br>a Coefficient: 0.028, b Coefficient: 0.147, c Coefficient: -0.181, d Coefficient: 0.059, e Coefficient: -0.028, f Coefficient: -0.186, g Coefficient: -0.073,<br>-- Employment 2: Coefficient---<br>a Coefficient: 0.020, b Coefficient: 0.003, c Coefficient: 0.404, d Coefficient: 0.648, e Coefficient: 0.105, f Coefficient: 0.088, g Coefficient: 0.115 |
| C=0.1 | Model Accuracy | Accuracy of Training Data Set of Logistic Model : 0.684<br>Accuracy of Test Data Set of Logistic Model : 0.560 |
| | Model Performance of Multinomial Logistic Classification | precision recall f1-score<br>0 0.50 0.71 0.59<br>1 0.62 0.67 0.64<br>2 0.50 0.17 0.25<br>accuracy 0.56<br>macro avg 0.54 0.52 0.49<br>weighted avg 0.56 0.56 0.53 |
| | Multinomial Logistic Regression Coefficicent | --- Employment 0: Coefficient---<br>a Coefficient: -0.042, b Coefficient: -0.050, c Coefficient: -0.077, d Coefficient: -0.153, e Coefficient: -0.050, f Coefficient: 0.032, g Coefficient: -0.027,<br>--- Employment 1: Coefficient---<br>a Coefficient: 0.005, b Coefficient: 0.018, c Coefficient: -0.031, d Coefficient: 0.016, e Coefficient: -0.004, f Coefficient: -0.057, g Coefficient: -0.009,<br>--- Employment 2: Coefficient---<br>a Coefficient: 0.037, b Coefficient: 0.032, c Coefficient: 0.108, d Coefficient: 0.136, e Coefficient: 0.054, f Coefficient: 0.025, g Coefficient: 0.035, |
| C=0.05 | Model Accuracy | Accuracy of Training Data Set of Logistic Model : 0.789<br>Accuracy of Test Data Set of Logistic Model : 0.600 |
| | Model Performance of Multinomial Logistic Classification | precision recall f1-score<br>0 0.56 0.71 0.63<br>1 0.67 0.50 0.57<br>2 0.57 0.67 0.62<br>accuracy 0.60<br>macro avg 0.60 0.63 0.60<br>weighted avg 0.61 0.60 0.60 |

| | |
|---|---|
| Multinomial Logistic Regression Coefficicent | --- **Employment 0: Coefficient**---<br>a Coefficient: −0.057, b Coefficient: −0.107, c Coefficient: −0.177, d Coefficient: −0.479, e Coefficient: −0.080, f Coefficient: 0.080, g Coefficient: −0. 033,<br>--- **Employment 1: Coefficient**---<br>a Coefficient: 0.017, b Coefficient: 0.082, c Coefficient: −0.113, d Coefficient: 0.046, e Coefficient: −0.016, f Coefficient: −0.152, g Coefficient: −0.041,<br>--- **Employment 2: Coefficient**---<br>a Coefficient: 0.040, b Coefficient: 0.025, c Coefficient: 0.290, d Coefficient: 0.433, e Coefficient: 0.096, f Coefficient: 0.072, g Coefficient: 0.075, |

첫째, C=1인 경우에 있어서는 종사자 수가 비교적 작은 0의 영역의 경우에는 대미달러 환율인상의 경우는 제외한 추정계수들이 모두 음으로 나타났다. 그리하여 환율이 증가할 때 부산의 취업자 수가 작을 때는 환율인상이 부산의 고용증가에 기여하고, 한국의 수출(g) 증가, 기계류 투자지수(a), 한국의 투자 총지수(e) 등이 기여하고 있다.

둘째, C=0.1인 경우와 C=0.05와 인 경우에는 기계류 투자지수(a), 정밀기기 투자지수(b), 그리고 부산의 경제활동인구(d) 등이 1단위 증가하면 취업자 수가 영역 2에 있을 확률이 다소 올라간다. 그 외 다른 독립변수들의 추정 추정계수는 음으로 나타났으므로 고용자 혹은 종사자가 0과 1의 영역에 속할 확률은 낮아진다.

그리고 취업자 수가 많은 2의 영역에 있어서는 기계류 투자지수(a), 정밀기기 투자지수(b), 운송장비 투자지수(c) 주요산업 부문들에 대한 투자설비 증가와 부산의 경제활동인구(d) 증가, 한국의 투자 총지수(e), 한국 원화의 대미 달러 환율(f), 그리고 한국의 수출(g) 등 7개의 모든 독립변수들의 추정계수가 양으로 나타나 고용이 증가하며 2의 영역에 속할 확률이 증가하는 것으로 나타났다.

그리하여 취업자 수가 제일 높은 영역 2를 바탕으로 보면, 추정계수들이 모두가 양으로 나타났고 로지스틱 추정계수들은 각각 $\exp(\beta i)$로 표시되어 있다. 그러므로 $\exp(\beta i)$값이 클수록 어떤 독립변수(i)가 한 단위 증가할 때 취업자 수가 높아질 가능성이 크다고 할 수 있다. 이것은 이들 독립변수들이 증가하면 취업자가 늘어나는 것으로 기계류 투자, 정밀기기 투자, 운송장비 투자지수, 부산의 경제활동인구, 한국의 투자 총지수, 한국 원화의 대미 달러 환율, 그리고 한국의 수출 등이 늘어날 때 취업자 수 혹은 부산의 고용을 증가시키는 것으로 해석된다.

## Ⅴ. 머신러닝 모형에 의한 국내 산업의 부산 지역총생산 효과분석

우리나라에서도 전통적 계량방법에 의한 산업육성과 그 경제효과에 대한 기존의 분석들은 많이 있었지만, 좀 더 예측력이 높은 인공지능을 이용하여 좀 더 예측력이 높은 머신러닝을 활용한 전문적인 경제학적 연구는 그 중요성에도 불구하고 아직 극히 미흡한 실정이다.

그리하여 본 절에서는 우리나라의 주요산업들인 제조업, 석탄 및 석유화학, 비금속광물 및 금속, 전기 및 전자, 기계 운송, 정보통신, 금융보험 등을 포함하고 있는 7개의 독립변수들을 채택하여 먼저 부산의 지역총생산(GRDP) 혹은 지역소득이 1, 그 외 전국 15개 광역시도들의 지역총생산 혹은 지역소득을 0으로 설정하여 여러 앙상블 머신러닝 모형으로 분류 예측을 시도하였다.

본 절에서는 이들 주요산업들이 부산과 우리나라의 소득 즉 지역총생산에 각각 어떻게 영향을 미치는지 분석하기 위하여 1985년부터 2018년까지의 연도별 자료를 사용하여 추정하였다. 그리고 부산과 전국 15개 광역시도들의

**Table 6.** Classification and Performance with a Voting Ensemble

| Classifier | | | | Accuracy of Classification |
|---|---|---|---|---|
| Classifier | Voting(soft) | | | 0.994 |
| | Decision Tree | | | 0.988 |
| | KNeighbors | | | 0.981 |
| Model Performance of Soft Voting Classification | | precision | recall | f1-score |
| | 0 | 1.00 | 0.99 | 1.00 |
| | 1 | 0.91 | 1.00 | 0.95 |
| | accuracy | | | 0.99 |
| | macro avg | 0.95 | 1.00 | 0.90 |
| | weighted avg | 0.99 | 0.99 | 0.99 |

지역총생산 혹은 소득을 종속변수로 두고 회귀를 통한 수치 예측을 한다. 모든 변수들에 대해 로그를 취하여 예측하였다. 본 절에서는 부산시 통계자료와 통계청의 MDIS에서 연도별 패널자료를 사용하였다.

그런데, 앙상블 학습에 사용할 변수 간의 큰 차이가 나는 불균형 데이터 세트를 그대로 학습하면 예측 편의가 발생한다. 따라서 데이터의 균형화 작업을 위하여 Python에서 SMOTE (Synthetic Minority Over-sampling Technique) 함수를 사용하고 데이터의 불균형 문제을 극복하기 위하여 전처리 과정에서 오버 샘플링 방식을 적용하여, 머신러닝 기법에서 사용하는 각 변수들을 표준화하기 위한 함수를 사용하였다.

본 연구의 Python 언어를 사용한 프로그램의 전처리 과정에서 각 변수들 간 수치차이가 컸지만, 표준화를 적용하여 측정단위를 조정하였다. 각 변수들의 표준화를 위해서는 StandardScaler 함수를 사용하여 Scaler 객체를 하나 만들어 핏(fit) 함수를 적용하여 훈련용 데이터 세트의 독립변수와 시험용 데이터 세트의 독립변수의 스케일을 표준화하였다. 이들 모형들의 역시 훈련용 데이터 세트와 시험용 데이터 세트를 70%와 30%로 각각 나누고, 그 평가를 위해 머신러닝 기법의 전처리 단계에서 각 변수들의 표준화 함수를 사용하였다.

## 1. 보팅 앙상블(Voting Emsemble) 모형 분류와 예측

### 1) 분류예측

위와 같이 보팅(Voting) 앙상블에 의해 구한 주요산업의 소득 혹은 지역총생산에 대한 분류예측은 본 절에서도 다른 모형들과 보팅 앙상블 모형 외에도 평가를 아울러 비교하기 위하여 단일 모형으로 의사결정나무 모형과 K-최근접 이웃(K-Nearest Neighbor) 모형을 생성하였다. 그리고 보팅 앙상블 모형의 함수의 개별나무의 개수(estimators) 인자에서는 소프트(soft) 보팅 방식을 선택하였다.

분류 예측 결과, 다음과 같은 〈Table 6〉에서 나타나 있듯이, 의사결정나무 모형의 정확도는 0.988, K-최근접 이웃 모형은 0.981 등으로 높게 나왔다. 보팅 앙상블의 정확도는 이들 보다 다소 높은 0.994로 나왔다. 그리고 보팅 앙상블 분류 모형의 정밀도, 재현율, F1 스코어 등의 모형은 모두 0.90 이상으로 나타나 보팅 앙상블에 의해 구한 주요산업의 소득 혹은 지역총생산에 대한 분류 예측모형은 좋은 것으로 나타났다.

### 2) 수치예측

수치예측을 위하여는 종속변수를 지역총생산으로 두고 부산의 전략산업들과 관계가 있는

**Table 7.** Prediction of SVR. ANN, and Voting Ensemble Model

| Model Accuracy | Support Vector Machine, Artificial Neutral Network, and Voting Model |
|---|---|
| Individual Model Performance | SVR Model's Coefficient of the Determination : 0.996 |
| | MLPRegressor Model's Coefficient of the Determination : 0.995 |
| Voting Ensemble Model Performance | Voting Ensemble model's Coefficient of the Determination: 0.997 |
| | RMSE: 0.096 |

**Table 8.** Classification and Performance with a Random Forest Model

| Model Accuracy | Training Data Set Accuracy of Random Forest Model: 0.936<br>Test Data Set Accuracy of Random Forest Model : 0.938 | | | |
|---|---|---|---|---|
| Model Performance of Random Forest Classification | | precision | recall | f1-score |
| | 0 | 0.94 | 1.00 | 0.97 |
| | 1 | 0.00 | 0.00 | 0.00 |
| | accuracy | | | 0.94 |
| | macro avg | 0.47 | 0.50 | 0.48 |
| | weighted avg | 0.88 | 0.94 | 0.91 |

주요산업들이 어떻게 영향을 미치는지 예측하기 위하여 서포트 벡터 머신 회귀(Support Vector Machine Regression: SVR) 모형과 인공신경망 모형을 결합하여 보팅 앙상블 모형으로 추정한 결과가 〈Table 7〉에 나타나 있다.

그리하여 단일 객체 모형들을 생성하기 위한 SVR 모형은 SVR 함수를 이용하였다. 그리고 인공신경망(Artificial Neutral Network: ANN) 모형의 추정을 위해서는 다층 퍼셉트론(MLP) 회귀모형을 이용하였는데, 그 추정결과가 〈Table 7〉에 나타나 있다.

그리하여 추정결과, 소프트 보팅 앙상블의 추정계수는 0.997, SVR의 결정계수는 0.996, 그리고 인공신경망 모형의 결정계수는 0.995로 모두 비슷하지만 높게 나타났으며, 보팅 앙상블의 결정계수가 SVR과 인공신경망 결정계수보다는 조금 높게 나왔다. 그리고 Voting 앙상블 모형의 RMSE를 구한 결과 0.096으로 상당히 적게 나타나 추정모형의 예측력은 양호하게 나타났다.

## 2. 랜덤 포레스트(Random Forest) 모형의 분류와 예측

### 1) 랜덤 포레스트(Random Forest) 분류 예측

앞의 보팅 앙상블 모형과 마찬가지로 부산 1, 여기에 다른 지역의 지역총생산은 0으로 두고 주요한 산업을 포함하는 독립변수들은 분류예측 모형을 평가하여 모형의 정확도를 도출하고자 한다. 이를 위해 〈Table 8〉에서는 개별나무의 개수(estimators)는 300으로 두었고, 나무의 최대 깊이는 2로 설정을 하였다.

추정 결과, 학습용 데이터 세트의 정확도는 0.936, 평가용 데이터 세트 정확도는 0.938로 나와 학습 모형과 평가 모형의 데이터 세트의 정확도가 높게 나타났다. 그리고 모형 성능평가 분류 보고서에 나타나 있듯이 정확도의 F1 스코어는 0.94로 높게 나타났다. 랜덤 포레스트 모형의 분류 예측의 평균 정밀도, 재현율, F1 스코어 등은 다소 낮게 나왔지만, 이것을 가중평균한 정밀도, 재현율, F1 스코어 등은 각각

**Table 9.** Predictions of Income with a Gradient Boosting Ensemble Model

| (estimators, max depth, learning rate) | Model Accuracy of Gradient Boosting Model |
|---|---|
| (estimators, max depth, learning rate) (100, 2, 0.1) | Training Data Set of Gradient Boosting Model : 1.000<br>Test Data Set of Gradient Boosting Model: 0.999<br>RMSE:0.046 |
| estimators, max depth, learning rate) (100, 2, 0.2) | Training Data Set of Gradient Boosting Model: 0.999<br>Test Data Set of Gradient Boosting Model: 0.998<br>RMSE:0.064 |
| (estimators, max depth, learning rate) (100, 4, 0.1) | Training Data Set of Gradient Boosting Model: 0.999<br>Test Data Set of Gradient Boosting Model: 0.997<br>RMSE:0.074 |
| (estimators, max depth, learning rate) (50, 2, 0.1) | Training Data Set of Gradient Boosting Model: 0.999<br>Test Data Set of Gradient Boosting Model: 0.998<br>RMSE:0.057 |
| (estimators, max depth, learning rate) (30, 2, 0.1) | Training Data Set of Gradient Boosting Model: 0.993<br>Test Data Set of Gradient Boosting Model: 0.991<br>RMSE:0.128 |

0.88, 0.94, 그리고 0.91로 높게 나왔다.

### 2) 랜덤 포레스트 소득에 대한 수치예측

앞에서와 같이 랜덤포레스트 모형에 의한 수치예측을 위하여 종속변수를 지역총생산 혹은 지역소득으로 두고 우리나라의 주요산업들이 부산의 소득에 어떻게 영향을 미치는지 예측하기 위해서 랜덤포레스트 회귀함수를 이용하였다. 인자는 개별나무의 개수=100, 최대 깊이(max_depth)=4로 두고 그 결과를 도출하였다.

랜덤포레스트의 수치예측의 추정결과 학습용 데이터 세트 결정계수는 0.999, 평가용 데이터 세트 결정계수는 0.997으로 각각 상당히 높게 나왔다. 한편 랜덤포레스트 수치 예측 모형의 RMSE는 0.068로 나타나 보팅모형의 0.096보다 더 낮게 나타나 우리나라의 주요산업의 부산의 소득 혹은 지역총생산에 대한 예측력이 더 높게 나타났다.

## 3. 그래디언트 부스팅(Gradient Boosting) 모형의 분류와 예측

### 1) 분류예측

지역총생산에 대한 주요한 산업을 포함하는 7개의 독립변수들의 분류 예측 모형을 평가하기 위하여 모형의 정확도를 도출하고자 한다. 앞에서와 마찬가지로 부산의 지역총생산은 1, 기타 전국 15개 광역시도의 총생산을 0으로 두고 그래디언트 부스팅 모형의 분류 예측 모형을 평가하기 위하여 정확도를 도출한다. 이를 위해 그래디언트 부스팅의 인자들 중 하위 나무의 개수를 지정하는 개별나무의 개수(estimators) = 10, 복잡도를 조정하는 인자를 사용하기 위하여 먼저 하위 개별나무의 최대 깊이(max_depth) = 3으로 설정을 하였고, 학습률(learning rate)는 0.1로 각각 설정하였다.

그리하여 그 추정결과는 학습용 데이터 세트 정확도 = 1.000, 평가용 데이터 세트 정확도=

0.999로 높게 나왔다. 그리고 모형의 성능은 정밀도가 0.97과 1.00, 재현율은 1.00과 0.50, F1 스코어 등 각각 0.98과 0.67로 나타났다. 정확도의 F1 스코어는 0.97로 높게 나타났다. 그래디언트 부스팅 모형의 분류 예측의 평균 정밀도, 재현율, F1 스코어 등도 상당히 높게 나왔고 이것을 가중평균한 정밀도, 재현율, F1 스코어 등은 각각 0.97, 0.97, 그리고 0.96으로 높게 나왔다.

### 2) 그래디언트 부스팅 수치예측

그래디언트 부스팅 모형에 의한 수치예측을 위하여는 종속변수를 지역총생산으로 두고, 8개 독립변수들인 주요산업들이 어떻게 영향을 미치는지 예측하기 위해서 그래디언트 부스팅 회귀함수를 이용하여 구한 추정결과는 〈Table 9〉에서와 같이 나타났다.

그리하여 그래디언트 부스팅 모형의 인자들인 개별나무의 개수, 최대 깊이, 그리고 학습률 등을 여러 가지로 조금 변화시켜 추정한 결과가 〈Table 9〉에서 나타난 것과 같이 그래디언트 부스팅 모형들도 인자값에 따라서 예측모형의 성능도 조금씩 다르게 나타났다. 그러나 전반적으로 그래디언트 부스팅의 수치예측들의 결정계수는 학습률이 0.1, 개별나무의 개수 = 100 혹은 50 이고, 최대 깊이 = 2일 때는 아주 높게 나타났고 그에 연관된 RMSE도 보팅 모형이나 랜덤 포레스트 모형에서 예측하는 것보다 낮게 나타나 지역총생산 혹은 소득 증대에 대한 경제효과의 예측 정확도가 더 높은 것으로 나타났다. 그리하여 그래디언트 부스팅 학습모형의 수치예측 모형은 예측의 정확도가 단일 머신러닝 모형인 의사결정 나무 모형, 서포트 벡터 머신 모형, 혹은 인공신경망 모형에 의한 회귀모형에서의 추정결과보다 더 예측력이 높은 것으로 나타났다.

## Ⅵ. 결언

세계는 현재 4차 산업혁명과 AI 시대를 맞이하고 있고, 한편으로 Covid-19이라는 팬데믹에 빠져 있고 러시아와 우크라이나의 전쟁 등으로 세계경제는 침체되고 있다. 이러한 경제변혁의 시대에는 산업의 육성으로 인한 경제효과 예측을 위해서는 설정오류 등으로 인해 발생하는 과대적합의 한계점을 좀 더 보완하기 위하여 전통적인 최소자승법에 의한 선형 추정모형보다 비선형 기법인 머신러닝 추정모형으로 좀 더 정확한 경제효과에 대한 예측과 분석이 필요하다. 그럼에도 우리나라에서는 머신러닝 기법에 의한 경제분석은 거의 도입되지 않고 있는 실정이다.

그리하여 본 연구는 부산의 산업과 우리나라의 주요산업들에 대한 투자를 중심으로 부산지역경제에 어떤 경제적 효과 즉 부산의 지역총생산 혹은 소득과 고용에 얼마나 영향을 미치는지, 로지스틱 회귀모형과 최근 도입된 머신러닝 모형 등에 의해 그 효과를 추정하고 예측하고자 하였다. 그 추정결과를 요약하면 다음과 같다.

첫째, 전통적인 시차를 도입한 OLS 회귀모형의 추정결과, 전기 및 전자, 기계운송, 금융보험산업 등은 부산의 지역총생산 혹은 지역소득에 5% 유의수준에서 각각 유의한 정의 영향을 미치고 있는 것으로 나타났다. 석탄 및 석유, 정보통신 산업은 10% 유의수준에서 부산지역 소득에 정의 영향을 주고 있다. 비금속광물 및 금속 산업은 부산의 지역 소득에 정의 영향을 미치는 않는 것으로 나타났다.

둘째, 이항 로지스틱 회귀모형에 의한 부산의 전략산업의 부산의 지역총생산에 대한 경제효과 추정결과, 미래수송기기산업과 라이프케어산업, 그리고 스마트해양산업의 순으로 부산의 소득에 기여하는 것이 크게 나타났으며, 글로벌관광, 지능형기기산업, 지능형정보서비스산업, 클린테크산업 등이 상대적으로 낮게 나타났다.

셋째, 다항 로지스틱 회귀모형에 의한 우리나라 주요산업과 수출과 환율 변동이 부산의 고용에 미치는 효과의 추정결과, 우리나라의 정밀기기산업, 운송장비 산업, 기계류 산업 등에 대한 투자가 증가할수록 부산경제에 대한 효과가 긍정적으로 나타나고 부산의 고용이 증가하는 것으로 알 수 있다. 그리고 한국 원화의

대미 달러 환율이 높아지고 수출이 증대할수록 부산의 취업자들이 증가하여 고용이 늘어나는 것으로 나타났다. 그리고 환율 인상과 우리나라 수출의 증대는 부산의 고용이 더 큰 상태에 있는 경우에 고용이 작은 상태에서 보다 더 부산의 경제 즉 고용량에 긍정적인 결과를 가져오는 것으로 나타났다.

넷째, 보팅 앙상블 모형으로 주요산업의 부산의 지역총생산 혹은 소득에 대한 추정결과, 소프트 보팅 앙상블의 추정계수는 비슷하지만 조금 높게 나타났으며, 보팅 앙상블의 결정계수가 서포트 벡터 머신 회귀모형과 인공신경망 결정계수 보다는 조금 높게 나왔다. 그리고 보팅 앙상블 모형의 RMSE를 구한 결과 아주 낮게 나타나 예측력이 이들 단일 머신러닝 모형보다 높게 나타났다.

다섯째, 랜덤포레스트 모형에 의한 우리나라 주요산업의 부산의 지역총생산에 대한 수치예측의 추정결과 결정계수도 높게 나왔고, 랜덤 포레스트 수치 예측 모형의 RMSE는 랜덤 포레스트 회귀모형은 보팅 앙상블 회귀모형보다 더 낮게 나타나 그 예측력이 더 좋은 것으로 나타났다.

여섯째, 그래디언트 부스팅 모형으로 우리나라의 주요산업의 부산의 지역총생산에 대한 추정결과, 전반적으로 그래디언트 부스팅의 수치예측들의 결정계수는 학습률이 0.1, 개별나무의 개수=100 혹은 50 이고 최대 깊이=2일 때는 아주 높게 나타났고 그에 연관된 RMSE도 보팅이나 랜덤 포레스트 모형보다 낮게 나타나 경제효과의 예측 정확도가 더 높은 것으로 나타났다.

그리하여 본 연구에 우리나라의 주요산업들에 대한 투자가 부산의 지역총생산 혹은 소득에 미치는 영향에 대한 머신러닝 예측결과를 종합해 보면 앙상블 머신러닝 모형인 그래디언트 부스팅 모형, 랜덤 포레스트 모형, 보팅 모형 순으로 추정하여 구한 예측력이 더 좋은 것으로 나타났다. 그리고 이러한 앙상블 머신 모형들에 의한 추정 예측력은 단일 머신러닝 모형들인 의사결정나무 모형, 서포트 벡터 머신 모형, 인공신경망 모형 보다 더 정확한 예측력을 보이는 것으로 나타났다.

따라서 부산시의 전략산업과 우리나라의 주요산업 그리고 수출이나 환율 등을 포함하여 부산의 지역경제에 미치는 효과를 분석할 때, 선형모형 설정의 오류와 과대적합 문제를 일으키는 통상적인 최소자승법에 의한 추정과 예측보다는 비선형 로지스

틱 회귀모형에 의한 추정, 그리고 비선형 머신러닝 회귀모형에 의해 좀 더 나은 추정과 예측을 함으로써 보다 향후 우리나라와 지역경제 차원에서도 효과적이고 효율적인 산업정책과 무역정책의 효과를 예측할 수 있을 것이다.

그러나 본 연구는 아직 경제학 분야에서 머신러닝 기법을 도입한 생소한 초기연구이므로 향후 좀 더 엄밀한 후속연구들을 유발할 수 있으며, 그런 점에서 본 연구는 머신러닝 모형을 이용한 예측과 추정에 디딤돌 역할을 할 수 있을 것이다. 특히 본 연구는 더 많은 지역과 기간 그리고 자료를 수집하여 분석할 수 있다면 좀 더 엄밀하고 완성도가 높고 유용한 결과를 얻을 수 있겠지만, 향후 과제로 남긴다.

## 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*, Brighton, MA: Harvard Business Review Press.

Arellano, M., and S. Bond (1991), " Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations," *The review of economic studies*, 58 (2), 277-297.

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.

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.

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, Hyong-Soo (2020), *Step by Step Business Machine Learning in Python*, Predics.

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

Kreif, Noémi, and Karla 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 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.

Park, Ki-Young, and Jeong-Won, Ko (2019), "A Short Guide to Machine Learning for Economists", *Korea Economic Studies*, 26(2), 367-408.

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

Yi, Chae-Deug (2021a), "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 (2021b), "Investment, Export, and Exchange Rate on Prediction of Employment with Decision Tree, Random Forest, and Gradient Boosting Machine Learning Models", *Korea Trade Review*, 46(2), 281-299.

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