https://doi.org/10.7840/kics.2022.47.10.1576

국가종합전자조달시스템 복수예비가격에서 다중선형회귀 및 정규화 모델 기반의 예정가격 예측

엄 상 훈◆, 조 인 휘°

Prediction of Standard Price Based on Multi-linear Regression and Its Regularized Models in Multiple Preliminary Prices of the Korea On-line E-Procurement System (KONEPS)

Sanghoon Eom◆, Inwhee Joe°

요 약

차세대 통합 국가종합전달조달시스템으로 발전하고 있는 나라장터(KONEPS)는 연간 64조원을 거래하는 통합 공공 전자조달시스템으로 조달 업체 등록-입찰-계약-검사-대금 지급 등 조달업무 전 과정을 전자적으로 처리하고 있다. 2021년 기준 등록된 조달업체 수는 668,964개로 낙찰 업체가 되기 위한 치열한 경쟁을 하고 있다. 가격조사를 통해 합리적으로 확정하는 기초금액을 기준해 상한과 하한 범위가 정해져 있어 과당 경쟁으로 터무니없는 낮은 가격이나 높은 가격 등 비합리적인 투찰을 방지하고 있다. 경쟁입찰에서 공정성과 투명성을 강화하는 복수예가 방식은 상한과 하한 범위 내에서 무작위로 정해지는 15개의 예비가격 중 참여하는 조달업체의 무작위 추첨으로 4개의 예비가격을 정해 산술평균으로 예정가격을 정한다. 사실상 무작위로 누구도 알 수 없게 만들어진 예정가격을 사전에 알 수 있다면 낙찰 가능성이 높아진다. 본 논문에서는 예정가격을 예측하는 모델을 구축하기 위해서 상관관계가 분명한 독립변수들을 피어슨 적률상관계수를 적용해 찾고 강한 선형성을 띤 변수들을 다중선형회귀 모델로 분석하였다. 모델의 다중공선성 문제를 해결하기 위해 정규화 회귀 모델을 적용하였고, 모델들의 예측력은 주요 지표로 평가하였다. 제시된 4가지 모델 모두 설명력을 나타내는 R-square가 99% 이상으로 높은 설명력을 보여 주었고, 실제 정답률과 예측 정확률 사이의 오차가 MAE 1%로 우수한 예측력을 보여 주었다. 정규화 모델 중에서는 엘라스틱넷회귀 모델이 설명력과 평가 지표 모두에서 가장 우수한 성능을 보여 주었고, 리지회귀 모델은 독립변수들의 특성을 가장 잘 살리는 것으로 확인되었으며, 라소회귀 모델은 변수 선택 특성이 나타났다. RMSE 지표에서는 라소회귀 모델의 오차가 가장 컸고, MAPE 지표에서는 리지회귀 모델의 오차가 가장 컸다. 4가지 모델 모두 입찰 담당자들에게 예정가격 예측으로 낙찰에 대한 경쟁력을 제공할 수 있을 것으로 보인다.

키워드 : 국가종합전자조달시스템, 복수예가방식, 예정가격, 다중선형회귀, 리지, 라소, 엘라스틱넷
Key Words : KONEPS, Multiple Preliminary Price Method, Standard Price, Multi regression analysis, Ridge, Lasso, Elastic Net

ABSTRACT

Korea Online E-Procurement System (KONEPS) is a developing integrated procurement system that processes 64 trillion won in transactions, annually. As of 2021, roughly 668,964 procurement companies were

◆ First Author : Sunwon Construction Co. Ltd, sheom@sunwon.co.kr, 종신회원
° Corresponding Author : Hanyang University Department of Computer Science, iwjoe@hanyang.ac.kr, 정회원
논문번호 : 202207-150-C-RU, Received July 22, 2022; Revised August 3, 2022; Accepted August 3, 2022

competing for successful bids. A standard price was set to prevent bidding at unreasonably high or low prices. The multiple preliminary method determines the standard price by randomly selecting 15 preliminary prices, with four prices set at random lottery, to ensure transparency in competitive bidding, as knowing the standard price would give a company unfair advantage. This paper discusses the Pearson's correlation coefficient in predicting the standard price, using a multiple linear regression model. A regularized regression model solved the multicollinearity problem, finding predictive power as a major indicator. All four models showed high explanatory power with over 99% of R-square, with an error margin of 1%. The elastic net model best demonstrated both explanatory power and evaluation index. The ridge regression model proved most effective with independent variables, and the Lasso regression model showed variable selection characteristics. The RMSE index indicated greater error in the Lasso model, while the ridge model showed greater error in the MAPE index. All four models are expected to increase competitiveness in predicting the standard price.

Ⅰ. 서 론

국가종합전자조달시스템(KONEPS)인 나라장터(www.g2b.go.kr)의 2021년 연간 발주 규모는 119조 9,929억원이다[1]. 총 63, 798개의 공공기관 중 5,459개 국가기관이 21조 4,800억원(17.9%), 6,997개 지방자치단체가 55조 182억원(45.9%), 12,194개 교육기관이 12조 4,076억원(10.3%), 그 외 3,262개 공기업, 준정부기관, 기타 공공기관 등이 31조 871억원(25.9%)을 발주했다. 발주를 받기 위해 경쟁하는 조달 등록업체 수는 2021년 기준 668,964업체이다. 전체 발주 중에서 공사 부문이 차지하는 비중은 39.9%로 발주액 규모는 47조 8,544억원이고 공사 부문 조달 업체는 113,506개 업체이다[2]. 공사 부문은 건설산업기본법 2조 4항에 따른 건설공사[1] 이외에 전기공사업법에 따른 전기공사, 정보통신공사업법에 따른 정보통신공사, 소방시설공사업법에 따른 소방시설공사를 포함한다. 조달 업체로 낙찰받기 위한 경쟁이 치열할 수밖에 없는 구조이다. 또한 내부업무시스템(ERP)과의 연계를 통한 편리성 및 업무 특수성을 이유로 국가기관, 지자체, 공기업과 그 산하 기관 등이 26개의 자체 조달시스템[2][3]을 운영하고 있어서 국내 전자입찰에서의 공사 규모는 더 크고, 경쟁 또한 치열하다.

공공기관의 입찰 정보를 공고하고 있는 나라장터는 조달업무 전 과정[3]을 전자적으로 처리하고, 1회 등록으로 행정자치부, 금융기관, 관련 협회 등 156개 기관의 조달 입찰에 참여할 수 있다. 투명한 조달 프로세스를 구축할 수 있어서 정보가 실시간으로 공개되고, 대면접촉 감소와 지문인식 전자입찰로 부정부패 차단과 부적격자 입찰 참여 배제 등이 가능해졌으며, 무엇보다 입찰이나 계약 때마다 반복해 제출하던 각종 서류[4]를 연계정보로 전자적으로 해결하면서 공공기관 방문 횟수가 줄고 제출서류 감소로 효율성과 비용을 절감했다.

또한 합리적 기준으로 기초금액[5]을 산정하고 나서 +2%와 -2% 또는 +3%와 -3% 상한과 하한 범위를 지정하여 터무니없는 낮은 금액을 써내지 못하도록 만들어 과당 경쟁으로 의한 저가 수주를 막고 부실 공사가 이루어지는 폐단을 방지하고 있다. 예정가격[6]을 사전에 빼돌려 부정 입찰을 할 수 있었던 때가 있었다. 하지만 복수예가방식으로 참여업체들이 입찰금액을 제시한 후 예정가격을 정하면서 부정 입찰을 사전에 차단할 수 있게 되었다.

복수예가방식은 계약업무 담당자가 예정가격을 확정하는 단일예가방식이나 예정가격을 정하지 않는 비예가방식(협상에 의한 계약)과는 다르다[4]. 2000년대 이전에는 계약업무 담당자가 10개의 예가를 사전에 만들어서 입찰 당일에 무작위로 3개를 선별, 산술평균으로 예정가격을 정하였다. 정보통신기술이 발달하면

1) 토목, 건축, 산업 설비, 조경, 환경시설, 시설물의 설치·유지·보수 및 부지조성, 기계설비나 그 밖 구조물의 설치 및 해체 공사 등.

2) 26개 자체 전자조달시스템 연간 이용금액 72조 9,211억 원(2016년). 운영기관: 방위사업청, 인천국제공항공사, 한국가스공사, 한국국토정보공사, 한국농수산식품유통공사, 한국도로공사, 한국석유공사, 한국수자원공사, 한국전력공사, 한국조폐공사, 한국지역난방공사, 한국철도공사, 한국철도시설공단, 한국토지주택공사, 한국수자력원전, 한국국제협력단, 정보통신산업진흥원, 한국과학기술연구원, 한국과학기술정보연구원, 한국인터넷진흥원, 한국전기연구원, 한국전자통신연구원, 교직원공제회, 강원랜드, 한국마사회, 한전KDN 등.

3) 조달 업체 등록, 입찰, 계약체결, 보증금 수납, 대금 지급 등.

4) 사업자등록증이나 법인등기부 등본, 시·국세 완납증명서, 보증서, 자격심사서류 등

5) 복수예가방식에서 다수의 예가를 산출하기 위해 작성된 기준금액이다. 추정가격을 뽑고 가격조사를 진행하여 확정한다.

6) 낙찰자 및 계약금액의 결정 기준으로 삼기 위하여 미리 작성 비치하여 두는 가액

서 2000년대 이후에는 예가 범위[7]의 변동 폭을 15개 구간으로 나누어서 구간마다 무작위로 하나의 값을 자동으로 추출해 총 15개의 복수 예가를 생성하고 각 예가에 무작위 구간번호를 부여한다. 무작위 구간번호를 가진 15개의 예가 중에 입찰 참여자 각각 2개의 무작위 번호를 선택하면 참여자의 두 배의 구간번호가 복수로 선택되고, 그중에서 가장 많이 선택된 구간번호 4개를 공개하여 예비가격 4개 값을 산술평균한다. 이런 과정으로 정해진 예정가격과 낙찰 하한률[8]을 곱해 그 이상을 써낸 입찰 참여자를 먼저 선별하고 그중에서 가장 최저가격을 입찰가격으로 제시한 참여자를 낙찰자 1순위로 확정한다[9]. 낙찰 하한율이 공개되어 있어서 예정가격만 입찰 이전에 미리 안다면 무조건 낙찰을 받을 수 있어서 입찰에 있어서 예정가격을 예측하는 것이 무엇보다 중요하다[4].

기존의 통계적 방법은 예가사정률(사정률)[10]에 따라 예정가격이 어느 구간에서 많이 만들어지는지를 사정률의 집중화 구간과 분산 구간으로 분석하여 고빈도 낙찰구간 선정, 경쟁률 낮은 구간 선정, 고빈도 낙찰구간과 경쟁률 낮은 구간을 혼합한 산정 방법 등 낙찰 가능성이 큰 입찰금액을 예측하는 다양한 시도를 하고 있다[4,5]. 하지만 소수점 셋째 자리까지의 예가 사정률은 경우의 수가 6,000개나 되어 업체 사정률까지 고려해야 하는 위의 방법은 복잡성이 크고, 무엇보다 난수 체계로 결정되는 복수예가 선정 시스템을 규정화할 수 없어 낙찰가를 추정 하는 자체가 쉽지 않다.

최근에는 데이터 마이닝과 머신러닝, 심층학습 알고리즘을 이용하여 다양한 분야에서 낙찰가격을 예측하는 연구가 이루어지고 있다[5-7]. 공공 IT사업 낙찰 예측이나 부동산 경매 낙찰가 예측, 차압된 자동차의 낙찰 예측 등 데이터 마이닝 기법을 이용한 예측이나 시계열 데이터를 이용한 머신러닝 예측과[5], 다층퍼셉트론(MLP), 신경퍼지추론시스템(ANFIS), 순환 신경망(RNN) 등 딥러닝 알고리즘을 이용한 전자입찰 가격 예측이 이루어지고 있다[5,7]. [5]와 [7]은 2015년부터 2019년까지의 전기공사 낙찰 데이터를 이용하여 4가지 독립변수인 기초금액, 추정금액, 예가 범위, 낙찰하한율로 종속변수인 낙찰금액을 예측하는 딥러닝 모델간 비교를 하고 있다. [5]는 최대최소변환(Min-Max Scaler)으로 데이터를 정규화한 MLP 모델이 ANFIS 모델보다 우수한 성능을 보여 낙찰 확률을 높일 수 있다고 보았고, [7]은 MLP와 RNN과의 비교에서 모델의 평가 지표인 평균절대오차(mean absolute error: MAE)와 평균절대비율오차(mean absolute percentage error: MAPE)에서는 RNN의 오차가 더 적게 나왔지만 평균제곱오차(mean squared error: MSE)와 평균제곱오차제곱근(root mean square error: RMSE)에서는 MLP가 RNN보다 오차가 더 적게 나왔고 특히 RMSE 값이 적을수록 정확도가 높다는 주장과 함께 MLP가 더 우수한 모델이란 평가를 하고 있다. [6]은 k-최근접 이웃 알고리즘을 이용한 입찰가격 예측은 분석 결과 훈련 99.1%, 테스트 99.9%의 정확도가 나왔고, k=9일 때 RMSE가 최소에 도달한 연구 결과를 보여 주었다.

하지만 선행연구 대부분은 전자입찰에서의 예정가격 예측이 아닌 낙찰가격 예측에 집중되어 있다. 또한 독립변수를 선택한 이유나 선형성을 가진 데이터에 비선형 방법으로 예측 모델을 만든 이유도 명확히 설명하고 있지 않다. 또한 수집된 데이터의 기준금액이 설정되어 있지 않고, 전기공사의 성격상 공사 규모가 작아 규모가 큰 건설공사 부문 적용에는 한계가 있을 수 있다. 아직 건설공사 부문 전자입찰 관련 예측 연구는 없는 실정이다.

이런 한계를 해결하는 접근을 위해 본 논문에서는 낙찰가격이 아닌 예정가격을 예측하는 모델을 구성하였고, 독립변수 사이의 상관관계를 조사하였고, 데이터 또한 20억에서 1,000억 이하로 기준을 정해 건설공사에 적합한 자료를 수집하였다. 또한 확보된 데이터의 독립변수와 종속변수 사이의 선형성을 파악하여 비선형 방법이 아닌 선형회귀분석 방법으로 예측 모델을 구성하는 것이 전자입찰에서의 낙찰 가능성을 높일 수 있는 전제로 연구를 진행하였다.

본 논문은 예정가격 모델을 만들기 위해서 조달청에서 발주한 공사 부문의 데이터를 확보하여 독립변수를 예정가격과 직접적 연관성이 있는 기초금액, 추

---

7) 조달청, 한국토지주택공사, 한국전력공사(공사·용역), 한국수자원공사 등은 ±2%, 지방자치단체, 한국도로공사, 서울교통공사 등은 ±3%를 적용한다. 발주기관별로 차이가 있다.

8) 입찰공고문에 "예정가격 이하 87.745% 이상 최저가격을 제시한 자로 결정한다."라고 낙찰 하한율이 제시되어 있다. 조달청은 공사 부문에서 2억~10억 미만 87.745%, 10억~50억 미만 86.745%, 50억~100억 미만 85.495%로 낙찰 하한율이 정해져 있고, 지방자치단체도 나머지 영역은 같고 100억~300억 미만에서 79.995%로 차이가 난다.

9) 투찰 금액이 (예정가격×100%)~(예정가격×낙찰하한율)의 범위 안에 들어가야 낙찰 대상자가 될 수 있다. 예정가격을 초과하면 예정가격 초과로 탈락하고 낙찰하한율 아래이면 낙찰하한가 미달로 탈락한다. 예정가격이 100억이고 투찰률이 87.745%라면 100억원을 초과하면 탈락하고 87억 7,450만원 이하로 입찰하면 탈락한다. 해당 범위 내에서 87억 7,450만원에 가장 가까운 입찰금액을 적은 기업이 1순위 낙찰자가 된다. 원용춘외, 앞의 책.

10) 예가사정률 = $\frac{\text{예정가격}}{\text{기초금액}} \times 100\%$

정금액, 추정가격으로 정하고 종속변수를 예정가격으로 정해서 예정가격 예측을 위한 다중선형회귀분석(multiple regression analysis: MRA)모델을 구성했고, 회귀분석이 가질 수 있는 다중공선성(multicollinearity) 문제를 해결하기 위해 정규화 모델(regularized model)을 적용해 예측 정확도를 높이는 방법을 연구하였다.

Ⅱ. 연구 방법과 분석

2.1 연구 방법

입찰에서 낙찰자를 결정짓는 기준금액은 예정가격이다. 낙찰을 받기 위해서는 예정가격을 잘 예측하는 것이 낙찰의 핵심이다. 하지만 예정가격은 입찰 전에 알 수 없고, 난수 체계를 가진 복수예가방식으로 복잡하게 만들어져서 예정가격을 통계적 방법으로 예측하기가 쉽지 않다. 머신러닝을 이용하면 데이터에서 컴퓨터가 스스로 특징을 추출하여 예측 모델을 만들기 때문에 예정가격을 예측하는데 더 효과적이다.

예정가격을 예측하기 위해 구축한 다중선형회귀 모델은 2개 이상의 독립변수와 한 개의 종속변수 사이의 선형 관계식을 가정한 모델이다. 확보한 데이터로 종속변수를 예측하는 독립변수들의 영향력인 회귀계수를 추정한다. 종속변수 y에 영향을 주는 n개의 독립변수 $X_1, X_2, \ldots, X_n$과 각 독립변수에 해당하는 회귀계수 $\beta_1, \beta_2, \cdots, \beta_n$과 상수항 C로 회귀계수를 추정하는 식을 표현할 수 있는데 식 1에서 살펴볼 수 있다[8]. 다중선형회귀 모델의 경우 노이즈 데이터(noisy data)에 민감해서 불필요한 정보를 가진 데이터가 포함되면 예측 성능이 떨어지는 경향성이 있고[9], 가장 큰 장애물인 다중공선성의 한계를 가지고 있다[10].

$$y = C + \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (1)$$

따라서 다중선형회귀모델 구축에서는 노이즈 데이터를 배제하면서 통계적으로 유의미한 독립변수를 선택하는 것과 함께 독립변수들이 서로 독립되지 않고 의존성이 커지면서 생기는 다중공선성 문제 또한 해결해야 한다. 다중공선성 문제가 발생하면 예측력이 약화할 수 있고 독립변수들 사이의 상호상관관계가 커지면서 회귀계수 추정량의 분산이 커지게 되어 추정된 회귀 결과의 안정성을 해치게 된다[11]. 훈련 데이터에 대한 예측 정확도가 검증 데이터에 대한 예측 정확도 보다 지나치게 높은 과대 적합(overfitting) 문제도 발생해서 일반화(generalization)의 오류가 커져 현장 적용의 한계를 가지게 된다.

다중공선성을 없애는 기본적인 방법은 다른 독립변수에 의존하는 변수를 제거하는 것이다. 하지만 예정가격 예측에 사용되는 독립변수들은 기본적으로 상관관계가 높을 수밖에 없어서 변수를 제거하는 것보다는 예측 정확도와 일반화 가능성을 높이는 조치로 정규화(regularized) 방법을 사용할 수 있다. 본 논문에서는 다중공선성 문제해결을 위해 정규화 방법을 사용하고 있다.

정규화를 위한 일반 패널티 공식은 식 2와 같다. 여기서 첫 번째 항은 최소제곱법으로 예측값과 실제값의 오차를 제곱한 손실함수이다. 두 번째 항은 첫 번째 항을 제어하기 위한 회귀계수에 대한 일반적인 L_q 패널티이다[12]. L_q 패널티에는 정규화 강도(regularization strength) 파라미터인 λ(감마) 값이 있는데 감마를 크게 하면 회귀계수의 값을 작게 만들어 과적합을 개선할 수 있고, 감마를 작게 하면 회귀계수의 값을 증가시켜 데이터의 적합을 개선할 수 있다. 이렇게 감마값으로 패널티를 부여해 회귀계수의 값의 크기를 감소시켜 과적합을 개선하는 것을 정규화라고 한다.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\left(\sum_{j=1}^{p}\beta_j^q\right)^{\frac{1}{q}} \quad (2)$$

대표적인 정규화 방법으로는 q가 2인 L_2 패널티를 가진 리지회귀(ridge regression) 모델과, L_1 패널티를 적용한 회귀분석 방법인 라소회귀(Lasso: least absolute shrinkage and selection operator) 모델, L_1과 L_2 패널티의 볼록 조합(convex combination)으로 L_1 패널티와 L_2 패널티의 두 패널티를 동시에 제약조건으로 가지는 엘라스틱넷회귀(Elastic Net) 모델이 있다.

리지회귀 모델은 선형회귀에서 회귀계수(가중치: β)들의 제곱합에 대한 최소화를 추구하고, 라소회귀는 회귀계수들의 절대값의 합을 최소화하는 패널티를 추가한다. 라소회귀는 특성상 필요하지 않은 독립변수의 회귀계수를 0으로 만들 수 있어서 특성(feature) 선택에 유리하다. 하지만 독립변수 사이에서 상관관계가 강한 경우 임의적으로 독립변수 중 선택된 변수를 제외하고 나머지 변수의 회귀계수 값이 0으로 수렴해 제거되는 문제점이 있다[12].

[10]은 불가피하게 독립변수들 사이에서 높은 상관관계가 있는 경우 리지회귀 모델로 수정하는 것이 다

중회귀분석 결과보다 더 높은 정확도를 제공한다고 주장한다. [11]은 독립변수들 사이에서 강한 상관관계 때문에 생기는 추정량 분산의 확대로 인한 추정 회귀식의 예측 정확도 하락 문제에 대해서는 라소회귀가 리지회귀보다 해석력을 더 높일 수 있다고 주장한다. 회귀계수 축소로 예측 정확도를 높이는 리지 회귀의 정확도와 영향력이 적은 회귀계수 값을 쉽게 0으로 만들 수 있는 변수 선택(variable selection) 기능 때문이다. [12]는 세 가지 정규화 모델의 예측 성능과 관련해서 시뮬레이션 결과 상관관계가 없는 경우에는 리지 회귀의 성능이 높게 나타났지만, 공변량이 상관된 데이터의 경우에는 리지회귀보다는 라소와 엘라스틱넷회귀 모델의 성능이 더 우수하다고 주장한다. 특히 상관관계가 있는 독립변수의 크기가 클수록 리지회귀보다 나머지 두 모델이 성능이 더 우수하게 나타나는 결과를 보여 주고 있다. [12]는 문헌분석 결과로 엘라스틱넷 회귀 모델이 다중공선성 문제해결에서 특히 효과적이며, 공변량이 상관된 데이터의 경우 엘라스틱넷 회귀 모델이 리지와 라소회귀 모델을 능가한 결과를 보여 주는 여러 연구 결과가 있다고 설명하고 있다. 본 연구에서는 다중선형회귀의 다중공선성 문제를 위의 세 가지 주장을 수용하여 3개의 정규화 모델을 적용해 해결하고자 한다.

2.2 분석

2.2.1 분석 대상 데이터

본 연구를 위해 공공데이터 포털(data.go.kr)에서 2008년 1월 1일부터 2021년 12월 31일까지 20억 이상 1,000억원 이하의 복수예가 방법으로 예정가격을 정한 건설공사 관련 데이터 30,100개를 확보하였다. 그중에서 입찰자가 없어 유찰되었거나 복수예가방식이 아닌 다른 방식으로 낙찰자가 정해진 경우에는 예정가격 데이터가 없어 제거하여 29,908개의 공사 관련 입찰데이터를 확보하였다. 확보된 입찰데이터에서 동일한 예가 범위를 가진 조달청 발주 입찰 데이터로 한정하여 총 4,949개 데이터를 분석을 위한 최종 데이터로 확정하였다. 조달청 이외의 기관, 특히 지방자치단체의 경우에는 예가 범위가 달라 이번 분석에서는 제외하였지만, 예가 범위가 다른 데이터에 대한 비교분석은 추후 연구로 진행할 계획이다.

2.2.2 전처리 설계

데이터 전처리를 진행하였다. 데이터의 특성상 선정된 금액 범위가 너무 넓고 커서 특이점(이상치)이 존재할 확률이 높은데, 이런 경우 회귀식의 결정에 큰 영향을 미치게 된다. 하지만 예정가격을 예측해야 하는 실제 상황에서는 이상치가 발생할 확률이 크기 때문에 이상치 또한 중요한 변수가 될 수 있어서 이를 제거하지 않고 그 영향을 줄이기 위해 데이터의 전처리로 로버스트 스케일 변환(robust scale transformation) 방법을 사용하였다. 데이터의 대표적 특성(representativeness)은 그대로 유지하면서 동등화 정확도(equating accuracy)를 향상하는 효과적인 방법으로 이상치 데이터가 존재할 때 훨씬 더 정확한 예측을 하도록 데이터를 변화시켜 주는 방법이다[13]. 확보한 데이터인 종목, 기초금액, 추정금액, 추정가격, 참여업체수, 개찰일 데이터 등을 모델 구성을 위해 전처리하여 독립변수인 입력데이터로는 기초금액(basic amount, BA), 추정금액(presumed amount, PA), 추정가격(presumed price, PP), 종속변수인 출력 데이터는 예정가격(standard price, SP) 데이터로 활용하였다. 학습 데이터 세트는 80%로 3,559개, 테스트 데이터 세트는 20%로 990개로 나누어 다중 선형회귀 모델을 위한 훈련과 검증 데이터로 활용하였다.

2.2.3 상관관계 분석

상관관계 분석은 두 변수 간 선형 관계가 존재하는지와 선형성의 방향과 강도를 측정하여 두 변수 사이의 상관성(밀접한 정도)을 판단하는 기법이다. 두 변수 간 선형적인 관계 정도를 나타내는 상관계수로 선형관계의 방향과 강도를 표현할 수 있다. 상관계수는 상관을 수치화한 공분산을 각 변수의 표준편차로 나누어 표준화한 값이다. 본 연구에서는 피어슨 적률상관계수(Pearson's product moment correlation coefficient)를 적용하여 상관성을 판단하였다[14].

[그림 1]은 기초금액과 추정금액, 추정가격, 예정가격의 상관관계 분석 결과로 3가지 독립변수 모두 종속변수인 예정가격과 양(+)의 강한 선형적인 상관관계를 보여 주고 있다.

[그림 2]는 독립변수들의 상관관계 분석 결과로 3가지 독립변수들 모두 강한 양의 선형적 상관관계가 존재함을 알 수 있다. 독립변수 간 피어슨 상관계수가 0.5 이하이면 다중공선성이 없는 것으로 판단하는데, 예정가격 예측 모델의 경우 상관계수가 모두 0.9 이상으로 강한 상관관계로 인한 변수간 다중공선성 문제가 발생할 수 있어서 정규화 모델 적용의 필요성이 제기된다.

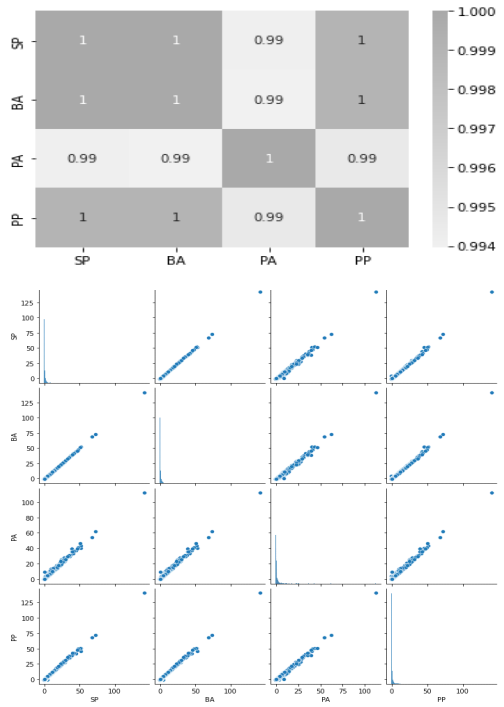

그림 1. 예정가격, 기초금액, 추정금액, 추정가격 상호 간의 상관관계
Fig. 1. Correlation between the standard price(SP), the basic amount(BA), the presumed amount(PA), and the presumed price(PP).

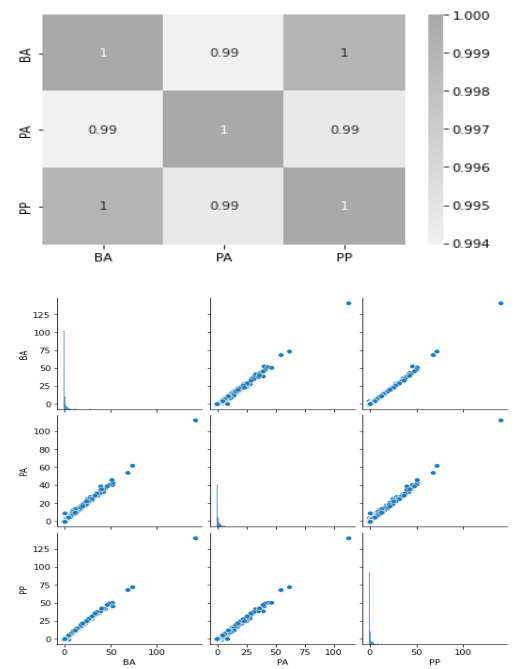

그림 2. 기초금액, 추정금액, 추정가격의 상관관계
Fig. 2. Correlation between the basic amount(BA), the presumed amount(PA), and the presumed price(PP).

Ⅲ. 다중선형회귀모델을 통한 예정가격 예측 및 성능평가

본 논문에서는 독립변수인 기초금액, 추정금액, 추정가격, 종속변수인 예정가격으로 다중회귀분석을 위한 모델을 구성하였다. 최소제곱법(least squares method)에 근거해 회귀계수의 추정량을 구하였다.

[표 1]은 전체 데이터 4,949개로 추정한 회귀식 계수(coefficients)들과 그에 대한 정보들을 보여 준다. Pr(>|t|)열은 t 분포를 사용하여 각 변수가 얼마나 유의미한지를 판단할 수 있는 p-value를 나타낸다[15].

분석 내용을 보면 독립변수 모두와 절편(intercept)의 p-value 값이 유의수준인 0.05보다 작다는 것을 알 수 있다. 따라서 예정가격 예측에서 기초금액, 추정금액, 추정가격의 세 가지 독립변수는 유의미하고, 절편 또한 유의미한 것을 알 수 있다. [표 1]의 회귀식 결과는 [예정가격]={(0.9528×기초금액)+(0.0090×추정금액)+(0.0385×추정가격)-0.0008}로 표현할 수 있다.

예정가격 예측에 관한 회귀 모델의 설명력은 결정계수(R-square)로 평가하였다. 모델이 예측한 예정가격과 실제 예정가격의 분산 비율을 나타내는 R-square 값은 1에 가까울수록 설명력이 좋은 모델로 볼 수 있다. 회귀계수(coefficient)는 독립변수인 각 입력데이터(특징)의 가중치를 보여 주어 예정가격 예측에서 어떤 독립변수(특징)가 주요하게 영향을 미쳤는지 알 수 있게 해 준다.

다중선형회귀와 다중공선성 문제를 해결하는 목적으로 적용한 3가지 정규화 예측 모델의 예측 정확도를 평가하기 위해서 4가지의 성능 평가 지표를 사용하였다. 실제 관측값과 모델에 의한 예측값 사이에서의 오차를 중점적으로 평가하는 MSE, RMSE, MAE와 예측값과 관측값의 에러를 계산하는 대표적인 측정 방법인 MAPE 지표로 예측력을 평가하였다[9]. MSE는 모델로 예측한 예정가격과 관측된 실제 예정가격과의 차이를 제곱해서 평균한 값이다. MSE 값이 적을수록 오차가 작다는 의미로 우수한 성능을 보여

표 1. 추정된 계수 및 정보
Table 1. Estimated coefficients and information.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-0.0008	0.001	-1.518	0.017
Basic Amount	0.9528	0.003	353.855	0.000
Presumed Amount	0.0090	0.001	6.891	0.000
Presumed Price	0.0385	0.003	13.585	0.000

준다. 하지만 과대 적합의 가능성이 있고, 예측된 값이 관측값보다 크게 예측했는지 작게 예측했는지를 파악할 수 없다. RMSE는 MSE값에 루트를 씌운 값으로 오차를 제곱하여 생긴 왜곡을 줄여 준다. 단위와 크기에 영향을 크게 받는 MAE는 두 값의 차이에 대한 절대값의 평균값이다. 본 논문에서는 데이터의 단위 차이로 인한 규모 차이가 없어서 MAE 또한 성능 예측에 적합하다고 할 수 있다. 값이 작을수록 높은 성능을 보인다. MAE와 비슷한 MAPE는 관측값이 0이거나 0에 가까운 매우 작은 값이면 MAPE 값 측정이 어렵거나 매우 커지는 단점이 있는데 연구 모델은 0이나 0에 가까운 값이 없어 성능 평가 지표로 가능하다. MAPE도 값이 작을수록 모델의 성능이 우수하다는 것을 의미한다. 전체 모델은 5-폴드 교차검증(cross validation)을 통해 예측력을 평가하고 비교하였다.

세 가지 정규화 모델은 정규화 과정에서 가장 좋은 성능을 나타내는 정규화 강도(λ) 값을 최적화해야 하는데, 그 검증 결과를 [그림 3, 4, 5]에서 표현하고 있다. R-square, MSE, RMSE로 평가하였고 5-폴드 교차검증으로 최종값을 평가하였다. 제시된 후보 값에서 리지회귀는 1.0, 라소회귀 모델과 엘라스틱넷회귀 모델에서는 0.001 값에서 가장 좋은 성능을 보여 주었다.

```
Alpha:0.0000, R2 0.9999, MSE: 0.0014, RMSE:  0.0369
Alpha:0.0001, R2 0.9999, MSE: 0.0014, RMSE:  0.0369
Alpha:0.0010, R2 0.9999, MSE: 0.0014, RMSE:  0.0369
Alpha:0.0100, R2 0.9999, MSE: 0.0014, RMSE:  0.0369
Alpha:0.1000, R2 0.9999, MSE: 0.0014, RMSE:  0.0369
Alpha:1.0000, R2 0.9999, MSE: 0.0014, RMSE:  0.0369
Alpha:10.0000, R2 0.9999, MSE: 0.0014, RMSE:  0.0379
Alpha:100.0000, R2 0.9998, MSE: 0.0037, RMSE:  0.0610
Alpha:1000.0000, R2 0.9992, MSE: 0.0152, RMSE:  0.1234
```

그림 3. 리지회귀 모델 정규화 강도
Fig. 3. Ridge regression model regularization strength.

```
Alpha:0.0000, R2 0.9999, MSE: 0.0014, RMSE:  0.0378
Alpha:0.0001, R2 0.9999, MSE: 0.0014, RMSE:  0.0378
Alpha:0.0010, R2 0.9999, MSE: 0.0014, RMSE:  0.0378
Alpha:0.0100, R2 0.9999, MSE: 0.0015, RMSE:  0.0381
Alpha:0.1000, R2 0.9999, MSE: 0.0019, RMSE:  0.0435
Alpha:1.0000, R2 0.9974, MSE: 0.0516, RMSE:  0.2273
Alpha:10.0000, R2 0.7551, MSE: 4.9249, RMSE:  2.2192
Alpha:100.0000, R2 0.0000, MSE: 20.1066, RMSE:  4.4840
Alpha:1000.0000, R2 0.0000, MSE: 20.1066, RMSE:  4.4840
```

그림 4. 라소회귀 모델 정규화 강도
Fig. 4. Lasso regression model regularization strength.

```
Alpha: 0.0001, R2:0.9999, MSE: 0.0014, RMSE: 0.0378
Alpha: 0.0010, R2:0.9999, MSE: 0.0014, RMSE: 0.0377
Alpha: 0.0100, R2:0.9999, MSE: 0.0015, RMSE: 0.0393
Alpha: 0.1000, R2:0.9997, MSE: 0.0064, RMSE: 0.0798
Alpha: 0.3000, R2:0.9994, MSE: 0.0125, RMSE: 0.1119
Alpha: 0.5000, R2:0.9991, MSE: 0.0183, RMSE: 0.1354
Alpha: 0.7000, R2:0.9987, MSE: 0.0253, RMSE: 0.1591
Alpha: 1.0000, R2:0.9981, MSE: 0.0386, RMSE: 0.1964
Alpha: 10.0000, R2:0.8968, MSE: 2.0753, RMSE: 1.4406
Alpha: 100.0000, R2:0.0000, MSE: 20.1066, RMSE: 4.4840
```

그림 5. 엘라스틱넷회귀 모델 정규화 강도
Fig. 5. Elastic Net regression model regularization strength.

표 2. 선형회귀 모델별 예측 성능
Table 2. Prediction Performance of 4 Linear Regression Models.

	MSE	RMSE	MAE	MAPE
MLR	0.001394	0.050688	0.011410	22.12
Ridge	0.002017	0.044909	0.011547	52.04
Lasso	0.002029	0.045043	0.011206	47.66
ElasticNet	0.001419	0.037666	0.011346	42.28

[표 2]에서 보면 예측 성능을 비교하였을 때 4가지 모델 모두 MAE가 1%를 조금 상회하기 때문에 평균적으로 실제 정답률과 모델에 의한 예측 정답률 사이의 오차가 1%로 매우 정확한 수준으로 예측함을 알 수 있다. 3가지 정규화 모델 중 라소와 엘라스틱넷 회귀 모델이 리지회귀 모델보다 낮게 나와서 [12]의 연구 결과와 유사하지만, 오히려 라소 회귀가 가장 작은 값을 보여 [12]의 연구 결과와 차이를 보여 주고 있다. 하지만 그 결과의 차이가 0.014%로 연구 결과가 차이가 난다고 보기는 어려워 보인다. MAPE는 3가지 정규화 모델 중 리지회귀 모델이 52.04%로 실제 정답률 대비 오차 비율이 가장 높았다. MSE와 RMSE 평가 지표에서는 라소회귀 모델이 오차 비율이 가장 높았다. MSE에서는 리지회귀 모델이 라소회귀 모델보다 더 작은 평균 MSE를 가지고 있어서 [12]의 연구 결과와 차이가 나지만 근소한 차이라서 [12]의 연구 결과와 배치된다고 보기는 어렵다. 다만 혼합모델인 엘라스틱넷회귀 모델이 3가지 정규화 모델 가운데 전체적으로 가장 낮은 값을 보여 주면서 독립변수의 공변량이 상관되면 엘라스틱넷 회귀가 리지회귀와 라소회귀보다 우수하다는 [12]의 문헌 연구 결과를 지지함을 볼 수 있었다. 4가지 회귀 모델 모두 예정가격 예측에 있어서 높은 정확도를 갖고 있음을 알 수 있다.

[표 3]은 R-square 값과 선정된 정규화 강도 값과 독립변수의 예측 영향도를 보여 주는 회귀계수 값을 나타낸다. 가장 우수한 설명력을 가진 모델은 정규화 모델인 엘라스틱넷회귀 모델이었고, 리지회귀 모델이 라소회귀 모델보다 조금 우수하게 나타났다. 리지회귀 모델은 3가지 독립변수들의 특성을 가장 잘 살려서 예측하지만 변수 선택 능력을 가진 라소회귀 모델의 경우 추정가격(PP)의 값이 0으로 수렴하는 것을 볼 수 있는데 실제 정규화 강도 값이 0.1 이상으로 바뀔 때

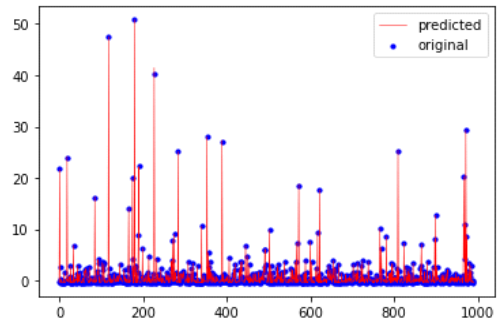

그림 6. 예측값과 실제값의 조합 그래프: 리지회귀 모델
Fig. 6. Combination graph of predicted and actual values: Ridge regression model.

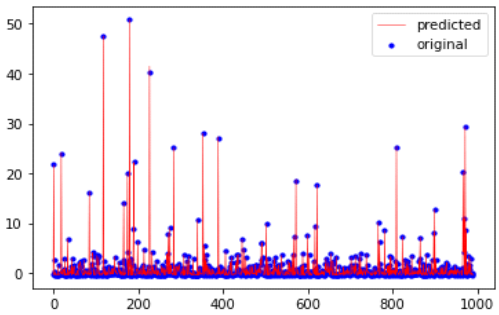

그림 7. 예측값과 실제값의 조합 그래프: 라소회귀 모델
Fig. 7. Combination graph of predicted and actual values: Lasso regression model.

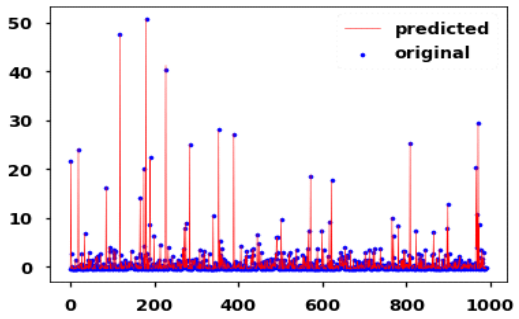

그림 8. 예측값과 실제값의 조합 그래프: 엘라스틱넷회귀 모델
Fig. 8. Combination graph of predicted and actual values: Elastic Net regression model.

추정가격 변수가 제외되는 결과를 볼 수 있었다. 엘라스틱넷회귀 모델에서도 추정가격의 회귀계수 값이 낮게 나왔지만 라소회귀 모델처럼 변수가 제외되지는 않는 특성을 살펴볼 수 있었다. [그림 6, 7, 8]은 예측값과 실제값을 조합한 그래프로 예측값과 실제 값의 오차 확인이 가능하다. 3가지 정규화 모델 모두 빨간색의 예측값이 실제 관측값인 점과 많은 데이터에서 일치하고 있음을 살펴볼 수 있어서 모델들의 예측 정확도가 우수함을 확인해 볼 수 있다.

Ⅳ. 결 론

국가종합전자조달시스템인 나라장터의 규모가 커지고 조달 기업을 선정하는 데 있어서 공정성과 투명성을 높이는 노력이 계속되고 있다. 2025년에는 현재 독립적으로 운영되는 26개 공공기관의 자체 조달시스템도 나라장터로 통합될 예정이라서 통합 전자입찰 시스템의 규모가 더욱 커지고 활발해질 전망이다. 조달 업체로 등록된 기업들은 낙찰을 위한 경쟁을 치열하게 하고 있다. 낙찰자 선정은 예정가격 이하 가격 중에 낙찰 하한율에 가장 가까운 입찰가격을 1순위 낙찰자로 결정한다. 따라서 이미 공고된 낙찰 하한율을 알고 있는 상황에서 예정가격을 안다면 낙찰을 받을 가능성이 커지기 때문에 예정가격을 예측하는 것이 중요해진다. 통계적 방법으로 기존의 다양한 시도들이 있었지만 최근 머신러닝과 딥러닝을 이용한 모델들이 개발되고 있다. 하지만 선행연구 대부분은 전자입찰에서의 예정가격 예측이 아닌 낙찰가격 예측에 집중된 한계가 있고, 선형성이 뚜렷함에도 비선형 모델만이 연구됐다. 이런 한계를 해결하는 접근을 위해 본 논문에서는 낙찰가격이 아닌 예정가격 예측 모델 구성과 상관관계가 분명한 독립변수를 선정하였고 데이터 또한 20억에서 1,000억원 이하로 기준을 정해 건설공사에 적합한 자료를 수집해 건설 부문에서 예정가격을 예측하는 모델을 처음으로 시도해 보았다. 다중선형회귀 분석 모델을 구성하였는데 다중공선성

표 3. 결정계수(R2)와 선형회귀 모델별 회귀계수
Table 3. Coefficient of Determination(R2) and Regression Coefficient of 4 Linear Regression Models.

	R-square	Regularization Strength	Basic Amount coefficient	Presumed Amount coefficient	Presumed Price coefficient
MLR	0.9999062		0.958467	0.011573	0.030545
Ridge	0.9998694	1.0	0.947044	0.007649	0.045824
Lasso	0.9998686	0.001	0.994979	0.005544	0.000006
ElasticNet	0.9999294	0.001	0.991585	0.008879	0.000105

문제를 해결하기 위해 정규화 모델을 적용하였다. 연구 결과 4가지 모델 모두 99%의 높은 정확도로 예측을 하였고, 4가지 주요 평가 지표에서 특히 실제 정답률과 모델에 의한 예측 정확률 사이의 오차가 1%로 모두 우수한 성능을 보여 주었다. 4가지 모델 중 엘라스틱넷회귀 모델이 설명력과 평가 지표 모두에서 4가지 모델 중 가장 우수한 성능을 보여 주었다. 정규화 모델 중에서는 리지회귀 모델이 독립변수의 특성을 가장 잘 살리는 것으로 확인되었고, 라소회귀 모델은 변수 선택 특성이 나타난 것을 살펴볼 수 있었다. MSE와 RMSE에서는 라소회귀 모델이 오차가 가장 컸고, MAPE 지표에서는 리지회귀 모델의 오차가 가장 컸다. 4가지 모델 모두 입찰 담당자들에게 입찰에 앞서 예정가격을 예측하여 공개된 낙찰 하한율에 맞추어 낙찰 1순위에 가까이 접근할 수 있는 경쟁력을 확보할 수 있을 것으로 보인다.

후속 연구로 예가 범위가 달라 이번 연구가 포함되지 않은 데이터에 관한 연구와 기존의 3가지 독립변수 이외에 시뮬레이션을 통한 복수예가 방식을 재현한 데이터를 생산해 냄으로써 일반화 가능성이 더 큰 모델을 추후 연구로 진행할 예정이다.

References

[1] Public Procurement Service, "*Procurement business statistics*," Approval (consultation) no. 131003, pp. 1-27, Dec. 2021.

[2] Public Procurement Service, "*2021 KONEPS Transaction Results*," Reported data from the Public Procurement Service, pp. 1-4, Jan. 2022.

[3] National Assembly Research Service(NARS), "*A study on the operational status of self-procurement system of public institutions and the integration of procurement system*," Field Report, vol. 43, pp. 1-85, 2016.

[4] Y. C. Won and H. G. Kang, *Bidding is Difficult but Winning the bid is Easy*, 2nd Ed. revision, KBID, 2018.

[5] D. H. Hwang and J. Y. Hwang, "A prediction of bid price using deep learning (MLP, ANFIS)," in *Proc. Korea Intell. Inf. Syst. Soc.*, pp. 168-169, 2020.

[6] D. H. Hwang, S. W. Kim, and Y. C. Bae, "A prediction of bid price using k-Nearest neighbhors algorithm," *J. The KIIS,* vol. 29, no. 6, pp. 482-487, 2019. (http://dx.doi.org/10.5391/JKIIS.2019.29.6.482)

[7] D. H. Hwang and Y. C. Bae, "The prediction of bidding price using deep learning in the electronic bidding," *J. KIECS*, vol. 15, no. 1, pp. 147-152, 2020. (https://doi.org/10.13067/JKIECS.2020.15.1.147)

[8] J. H. Chung, M. B. Son, Y. G. Lee, and S. J. Kim, "Estimation of soil moisture using sentinel-1 sar images and multiple linear regression model considering antecedent precipitations," *Korean J. Remote Sensing*, vol. 37, no. 3, pp. 515-530, 2021. (https://doi.org/10.7780/kjrs.2021.37.3.12)

[9] H. J. Park, K. Y. Jang, Y. H. Lee, W. J. Kim, and P. S. Kang, "Prediction of correct answer rate and identification of significant factors for CSAT english test based on data mining techniques," *KIPS Trans. Software and Data Eng.*, vol. 4, no. 11, pp. 509-520, 2015. (https://doi.org/10.3745/KTSDE.2015.4.11.509)

[10] J. Thaithanan, W. Wanishsakpong, T. Panityakul, and D. Prangchumpol, "The efficiency of ridge estimations for multicollinearity multiple linear regression: A monte-carilo simulation-based study," *Int. J. Mathematics and Comput. Sci.*, vol. 16, no. 4, pp. 1721-1727, 2021.

[11] C. Y. Park, "Simple principal componet analysis using Lasso," *J. Korean Data and Inf. Sci. Soc.*, vol. 24, no. 3, pp. 533-541, 2013. (http://dx.doi.org/10.7465/jkdi.2013.24.3.533)

[12] T. Sirimongkolkasem and R. Drikvandi, "On regularisation methods for analysis of high dimenstional data," *Annals of Data Sci.*, vol. 6, no. 4, pp. 737-763, 2019. (https://doi.org/10.1007/s40745-019-00209-4)

[13] Y. He, Z. Cui, and S. J. Osterlind, "New robust scale transformation methods in the presence of outlying common items," *Applied Psychological Measurement*, vol. 39, no. 9, pp. 613-626, 2015. (https://doi.org/10.1177/0146621615587003)

[14] T. H. Lee and H. Y. Jeong, "Spatio-temporal

multinomial regression analysis on the elderly social vulnerability for efficient allocation of welfare resources," *The Stud. in Regional Develop.*, vol. 54, no. 1, pp. 209-239, 2022. (http://dx.doi.org/10.35526/srd.2022.54.1.009)

[15] J. H. Lim, M. B. Lee, H. W. Cho, C. S. Shin, C. W. Park, and Y. Y. Cho, " An analysis study for optimal uptake of nutrient solution based on multiple linear regression model in strawberry hydroponic environments," in *Proc. KIPS*, vol. 26, no. 2, pp. 578-580, Nov. 2019.

엄 상 훈 (Sanghoon Eom)

2019년 9월~현재 : 한양대학교 컴퓨터소프트웨어학과 석박사통합과정

2017년 2월 : 서울벤처대학원대학교 융합산업학과 박사

2004년 5월 : 미국 University of Pennsylvania, Educational Leadership, M.S.

<관심분야> 스마트시티, 머신러닝 및 딥러닝, 강화학습, BEMS, IoT

[ORCID:0000-0002-9701-9669]

조 인 휘 (Inwhee Joe)

현재 : 한양대학교 컴퓨터소프트웨어학과 (정)교수

미국 Georgia Tech, Electrical and Computer Engineering, Ph. D.

미국 University of Arizona, Electrical and Computer Engineering, M.S.

<관심분야> 이동통신, IoT, 딥러닝, XAI, 임베디드 시스템, EV 배터리 및 시뮬레이션

[ORCID:0000-0002-8435-0395]