

건설 공정 지연 예측을 위한 인공지능 기반 경고 시스템 개발

정희찬¹ · 한기웅¹ · 이병훈¹ · 한예지¹ · 허경휘² · 최호창² · 한성원^{1*}

¹고려대학교 산업경영공학과 / ²현대건설 기술연구원

Development of AI-based Warning System for Construction Rate Forecasting

Hoichan Jung¹ · Giwoong Han¹ · Byung Hoon Lee¹ · Ye Ji Han¹ · Kyung Hwi Heo² · Hochang Choi²
Sung Won Han¹

¹Department of Industrial and Management Engineering, Korea University

²Hyundai Engineering and Construction

In the construction industry, the construction process is managed based on the completed construction in comparison to the planned construction. With the current construction management system, it is difficult to respond to construction delays in advance, because the construction process rate cannot be estimated quantitatively. In this paper, we developed an AI-based warning system that proactively predicts the delay in construction based on the completed construction of the past and additional data such as budget, subcontractor, and weather data. We experimented on two types of construction sites and seven machine and deep learning models and evaluated the prediction performance of the models for the construction process rate of the future three months with MAE. We analyze the importance of each feature on the model prediction using SHAP values. Furthermore, we select the adequate target variable for construction process prediction based on the results of prediction and key delay factors.

Keywords: Construction Process Rate Prediction, Time Series Forecasting, Machine Learning, Deep Learning, Feature Importance Extraction

1. 서론

건설업은 2020년 국내총생산의 5.9%를 차지하며 서비스업, 광공업, 제조업에 이어서 4번째로 비중이 큰 산업이다(Korean Statistical Information Service, 2020). 이처럼 건설업은 우리나라 경제에 매우 중요한 역할을 하는 산업이지만, 최근 건설업의 성장률과 투자는 2018년 이후로 지속해서 감소하고 있다(Construction Association of Korea, 2020). 건설업의 낮은 자본투자와 연구 및 개발로 인해 건설 프로젝트의 지연이 발생하며, 세계적 규모의 건설 프로젝트의 90%는 지연으로 인한 예산 초과를 경험한다(Egwim *et al.*, 2021). 건설 프로젝트의 지

연을 줄이기 위한 다양한 노력이 이루어졌지만, 건설 공정 지연은 계속해서 발생하였다. 본 연구에서 머신러닝 기반의 인공지능(Artificial Intelligence) 통해 공정 지연을 예측하여 건설 지연으로 인한 건설업의 피해를 최소화하고자 한다.

현재 건설산업에서 이루어지는 공정관리는 기존에 계획된 공정 진행률과 실제 진행된 공정 진행률을 비교하는 계획 대비 실적으로 이루어진다. 이는 현황 보고 위주로 진행되기 때문에 공정 지연을 정량적으로 예측하는 것이 불가능하고 지연에 대하여 사전 대응이 어렵다. 과거 공정 지연 예측 연구에서는 통계 기반의 회귀분석 모델을 활용하였고, 다양한 머신러닝 기반의 모델 또한 활용하였으며, 나아가 공정 지연 예측에

본 연구는 4단계 두뇌한국21에 의해 지원되었습니다. 본 연구는 (주)현대엔지니어링/현대건설의 지원을 받아 수행되었습니다.

* 연락처 : 한성원 교수, 02841, 서울특별시 성북구 안암로 145 고려대학교 산업경영공학부, Tel : 02-3290-3384, Fax : 02-929-5888,

E-mail : swhan@korea.ac.kr

2022년 1월 28일 접수; 2022년 4월 9일 수정본 접수; 2022년 5월 9일 게재 확정.

큰 영향을 미치는 주요 요인들을 파악하여 모델 설명력을 확보하였다. 최근에는 딥러닝 모델을 적용하여 건설 공정률을 예측하는 연구를 진행하고 있다. 따라서, 본 연구에서는 머신러닝 및 딥러닝 모델을 통해 공정 지연을 정량적으로 예측하고 이들의 성능을 비교한다. 또한, 변수 중요도를 도출하여 건설 지연의 주요 지연 요인을 파악한다. 나아가 종속변수를 구분해서 실험하여 종속변수에 따른 모델의 예측 성능과 주요 지연 요인을 비교한다. 이를 통해 지연에 대해 선제적으로 대응하는 인공지능 기반의 공정 지연 예측 및 경고 시스템을 구축한다.

본 연구의 구성은 다음과 같다. 제2장에서는 과거 건설 공정 지연 예측에 관한 연구를 다루고, 제3장에서는 사용한 방법론에 대해 설명한다. 제4장에서는 본 연구에서 사용한 데이터, 전처리 방법 및 파생 변수를 소개하고, 제5장에서는 평가 지표를 통해 구축된 모델의 성능 및 도출된 변수 중요도를 분석한다. 마지막으로 제6장에서는 본 연구의 결론, 한계점 및 향후 연구에 대하여 언급한다.

2. 선행 연구

과거의 건설 공정률 예측 연구에서는 통계적 기법을 활용하였다. 리모델링 프로젝트의 공정 지연 예측 모델을 구축하는 연구(Min *et al.*, 2005)는 공사 또는 현장 관리자를 대상으로 한 설문지 결과를 수집하여 공정 지연 요인 40개 중 8개를 주요 요인을 독립변수로 설정하고, 프로젝트 공기 수행 지표로 사용되는 공정 수행 편차 수행도(Schedule Deviation Performance Index)를 종속변수로 설정하여 회귀분석을 수행하였다. 공정 지연 요인을 이용하여 초고층 건물 내부공사의 생산성을 예측한 연구(Kim *et al.*, 2011)는 5가지 공정 지연 요인과 관리자의 직책을 독립변수로, 생산성을 종속변수로 설정하여 회귀분석을 수행하였다.

다양한 머신러닝 기법들의 건설 공정 지연 예측 성능을 비교하는 연구 또한 진행되었다. 51개의 건설 현장의 지연을 예측하기 위해 머신러닝 알고리즘을 활용한 연구(Gondia *et al.*, 2020)는 9가지 위험 요인을 활용하여 3가지 범주로 구분된 프로젝트 지연의 정도(Time Overrun)를 분류하였다. 사용한 분류 알고리즘은 Decision Tree(DT)와 Naïve Bayesian이며 작은 데이터의 크기로 인해 10-Fold Cross Validation을 활용하여 학습하였고, Accuracy로 모델 성능을 비교하였다. 고층 건설 공사의 지연 위험 평가를 위해 머신러닝을 활용한 연구(Sanni-Anibire *et al.*, 2020)는 K-Nearest Neighbors(KNN), Artificial Neural Networks(ANN), Support Vector Machines(SVM) 등 머신러닝 모델들의 지연 여부 분류의 성능을 비교하였고, Recursive Feature Elimination(RFE)을 통해 주요 지연 요인을 도출하였다. 건설 공정 예측을 위해 인공지능을 적용한 연구(Egwim *et al.*, 2021)는 Random Forest(RF), Extremely Randomized

Tree(ET), Adaptive Boosting Model(AdaBoost), Gradient Boosting Model(GBM) 등 11가지의 머신러닝 모델의 성능을 비교하였다.

최근에는 딥러닝 모델을 활용하여 건설 공정률을 예측하는 연구가 활발히 진행되고 있다. 터널 건설 현장에서 Shield Tunnel 내부를 구성하는 Soft Soil의 평가 지표인 Tunnel Serviceability Index(TSI)를 예측하는 연구(Zhu *et al.*, 2020)는 유사한 패턴을 보이는 연속된 TSI를 분류하기 위해 Dynamic-Time-Wrapping(DTW) 기반의 K-Means 군집화 알고리즘을 적용하고, 각 군집의 TSI를 Recurrent Neural Networks(RNN) 계열의 딥러닝 모델 중 기울기 소실 문제를 해결한 Long-Short-Term Memory(LSTM) 모델로 예측하였다. 모델에 대한 평가 지표는 Mean Absolute Error(MAE)와 Root Mean Square Error(RMSE)로 설정하였다.

앞선 연구에서는 공정 지연 여부를 분류하거나 다음 미래 시점의 공정률을 예측하는 데 그쳤지만, 본 연구에서는 이전 공정률과 날씨, 현장 및 예산, 하도급 업체 정보를 활용하여 미래 3개월의 공정률 예측을 수행하였다. 또한, 본 연구에서는 선행 연구에서 사용되지 않았던 딥러닝 모델과 변수 중요도 추출 기법에 대하여 분석하고, 종속변수를 변경해보며 예측과 변수 중요도 결과를 비교하였다.

3. 실험 계획

3.1 예측 모델

건설 공정률 예측에 적용할 수 있는 인공지능 모델은 크게 머신러닝과 딥러닝 모델로 구분된다. 개념적으로 딥러닝 또한 머신러닝에 포함되지만, 본 연구에서는 모델의 종류를 구분하기 위해 인공신경망 계열 모델을 딥러닝 모델로 따로 구분하였다. 본 연구에서 실험한 머신러닝 모델은 Extreme Gradient Boosting Model(XGB), Support Vector Regressor(SVR)와 Random Forest(RF)이며, 딥러닝 모델은 Recurrent Neural Networks(RNN), Long-Short Term Memory(LSTM), Sequence-to-Sequence(Seq2Seq)와 Transformer이다. 또한, 건설 공정률 예측에 영향을 미치는 변수의 중요도를 SHapley Additive ex-Planations(SHAP)를 통하여 도출하였다.

XGBoost(XGB)는 Boosting 계열의 모델로 예측에 대한 잔차(Residual)를 다음 모델을 통해 예측하는 과정을 반복하는 Gradient Boosting Machine(GBM)의 원리를 병렬처리가 가능하게 하여 더욱 빠르게 최적화하고, Regularization Term을 추가하여 과적합을 방지한 모델이다(Chen and Guestrin, 2016). <Figure 1>은 XGB의 구체적인 모델 구조를 표현한다. XGB의 Loss Function은 식 (1)과 같으며 l 은 손실함수, Ω 는 Regularization Term, γ 는 Regularization Parameter, T 는 Tree의 최종 Node의 개수 그리고 w 는 최종 Node의 Score를 나타낸다.

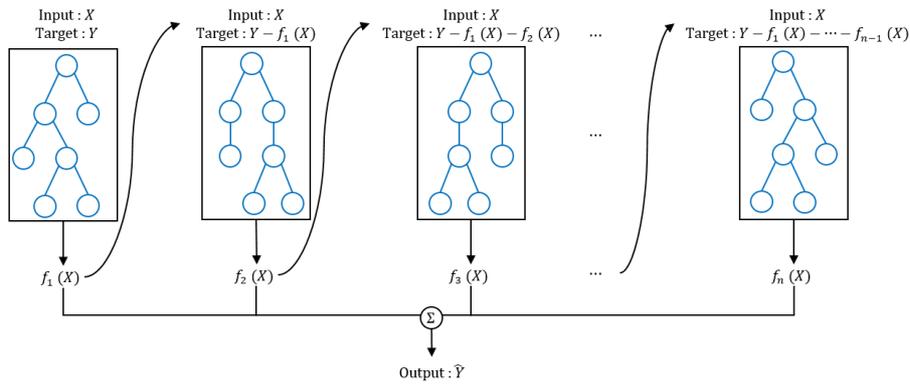


Figure 1. XGBoost

$$loss = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (1)$$

$$\text{where, } \Omega(f) = \gamma T + \frac{1}{2} \|w\|^2$$

Support Vector Machine(SVM)은 기본적으로 d 차원에서 두 범주를 +1 혹은 -1로 분류하는 $d-1$ 차원의 분류 경계면(Hyperplane)을 찾는 선형모델이다(Cortes and Vapnik, 1995). Soft Margin SVM은 Margin을 최대화함과 동시에 Regularization Cost를 통해 선형 분류 경계면 사이에 Noise에 대한 예외를 허용한다. 또한, Kernel Function을 통해 저차원의 데이터를 고차원으로 매핑하여 비선형 분류 경계면으로 기존에 선형 분류 경계면으로 분류할 수 없었던 문제를 해결한다. Support Vector Regressor(SVR)는 Epsilon-tube를 정의하여 $\pm \epsilon$ 사이의 예측에 대한 Loss를 부여하지 않고, 그 이상 예측에만 ξ 만큼 Loss를 부여한다.

Random Forest(RF)는 Bagging 계열의 모델로 데이터를 복원 추출을 허용하여 샘플링하고 변수의 다양성을 확보하기 위해 무작위로 Feature를 선정하여 Decision Tree를 구성한다(Breiman, 2001). Random Forest Regressor는 모든 Decision Tree의 예측값을 평균하여 최종 예측값을 출력한다.<Figure 3>은 RF의 구체적인 모델 구조를 표현한다.

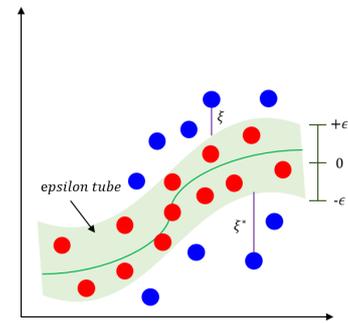


Figure 2. Support Vector Regressor

Recurrent Neural Network(RNN)은 이전 시점의 출력값을 은닉층의 값으로 ‘기억’하여 다음 시점 예측에 활용하는 구조로 Sequence 데이터에 효과적인 인공신경망의 한 종류이다(Rumelhart et al., 1986). 다양한 크기의 입력과 출력을 사용할 수 있는 구조이고, 본 연구에서는 Many-to-Many의 구조를 활용하였다.<Figure 4>는 RNN의 구체적인 모델 구조를 표현한다. RNN의 동작 원리에 대한 식 (2)와 같으며 x_t 는 t 시점의 입력값, h_t 는 t 시점의 은닉층의 값, o_t 는 t 시점의 출력값을 나타낸다. U, V, W 는 각각 x_t, h_t, h_{t-1} 의 가중치(Weight)를 나타내고, b 는 편향(Bias)을 나타낸다.

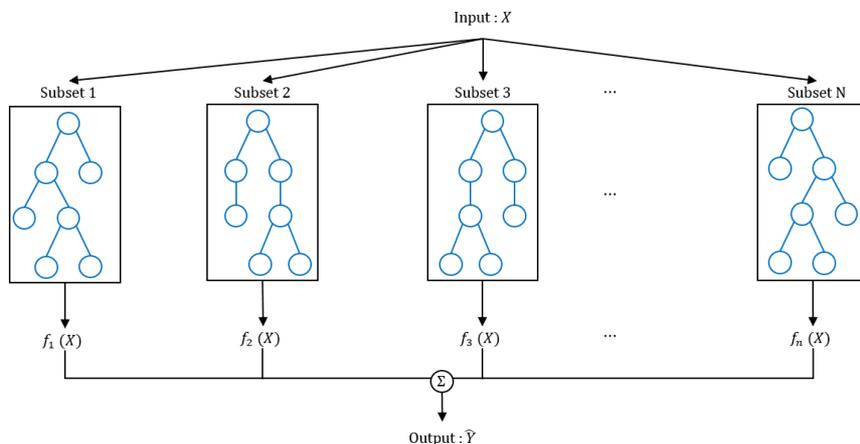


Figure 3. Random Forest

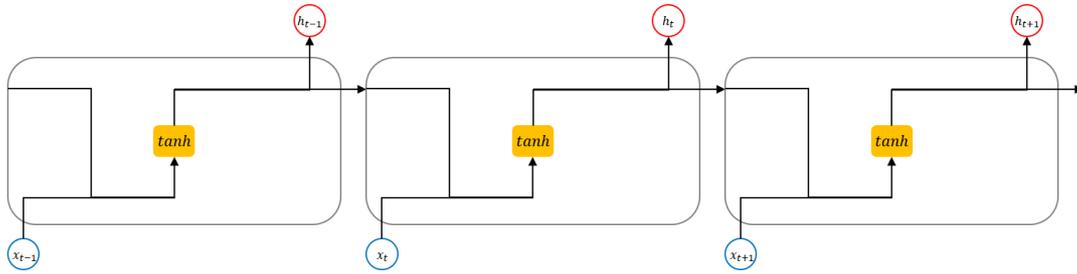


Figure 4. RNN

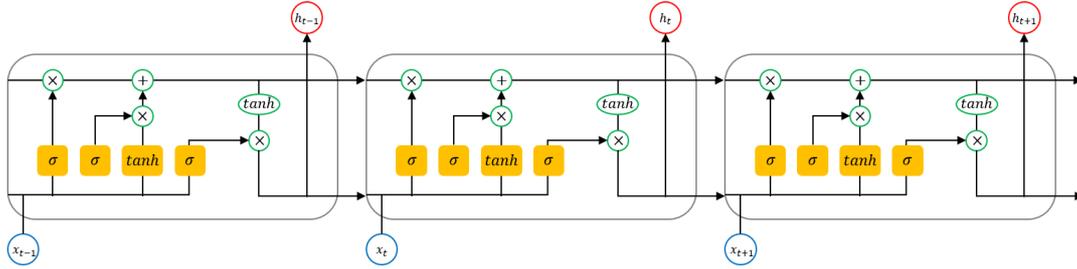


Figure 5. LSTM

$$h_t = \sigma(Ux_t + Wh_{t-1} + b_x) \quad (2)$$

$$o_t = \tanh(Vh_t + b_t)$$

Long-Sort Term Memory(LSTM)는 장기 의존성(Long-Term Dependency)로 인한 기울기 소실(Gradient Vanishing) 문제를 가지고 있는 RNN이 모델 예측이 먼 과거 시점의 정보를 반영하지 못한다는 단점을 해결하기 위해 제안되었다(Hochreiter and Schmidhuber, 1997). LSTM은 기존 RNN 구조에 기억소자(Memory Cell)와 입력 게이트, 망각 게이트, 출력 게이트가 추가된 구조이며, 이를 통해 정보의 흐름을 적절하게 조절한다. <Figure 5>는 LSTM의 구체적인 모델 구조를 표현한다.

LSTM의 동작 원리에 대한 식 (3)과 같다. 먼저, 망각 게이트(f_t)를 통해 과거 정보를 잊어버릴지 결정하기 위해 x_t, h_{t-1}, b^f 의 가중 합(Weighted Sum)에 Sigmoid 함수를 씌워 0과 1 사이의 값으로 계산한다. 1은 해당 위치의 정보를 기억한다는 의미, 0은 반대로 잊어버리겠다는 의미를 나타낸다. 입력 게이트(i_t)는 현재의 새로운 정보를 저장할지 결정하며, 망각 게이트와 유사하게 계산한다. 이때 현재 소자값(\tilde{c}_t)은 x_t, h_{t-1}, b^i 의 가중 합에 tanh 함수를 씌워 -1과 +1 사이의 값으로 계산된다. 최종 소자값(c_t)는 망각 게이트의 출력값과 이전 시점의 기억소자의 값, 입력 게이트의 출력값과 현재 시점의 소자값의 Element-wise

Product를 통해 업데이트된 값이다. 출력 게이트(o_t)는 최종 출력값을 결정하며, 망각과 입력 게이트와 유사하게 계산한다. 최종 출력값이자 은닉층 값(h_t)은 기억할 과거와 현재 정보가 반영된 최종 소자값에 tanh 함수를 씌워 계산한다.

$$f_t = \sigma(U^f x_t + W^f h_{t-1} + b^f) \quad (3)$$

$$i_t = \sigma(U^i x_t + W^i h_{t-1} + b^i)$$

$$\tilde{c}_t = \tanh(U^c x_t + W^c h_{t-1} + b^c)$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t$$

$$o_t = \sigma(U^o x_t + W^o h_{t-1} + b^o) \quad h_t = o_t * \tanh(c_t)$$

Sequence-to-Sequence(Seq2Seq)는 기계 번역(Machine Translation)을 수행하기 위해 여러 층의 RNN으로 구성된 Encoder-Decoder 구조를 가진 모델이다(Sutskever *et al.*, 2014). Encoder는 순서대로 입력되는 입력 Sequence를 입력받아 정보를 하나의 Context Vector로 추약하는 역할을 하고, Decoder는 Encoder가 생성한 Context Vector를 입력받아 출력 Sequence를 생성하는 역할을 한다. 이 모델은 자연어처리뿐만 아니라 시계열 데이터 분석에서도 효과적인 모델이다. 본 연구에서는 LSTM으로 Encoder와 Decoder를 구성하였다. <Figure 6>은 Seq2Seq의 모델 구조를 나타낸다.

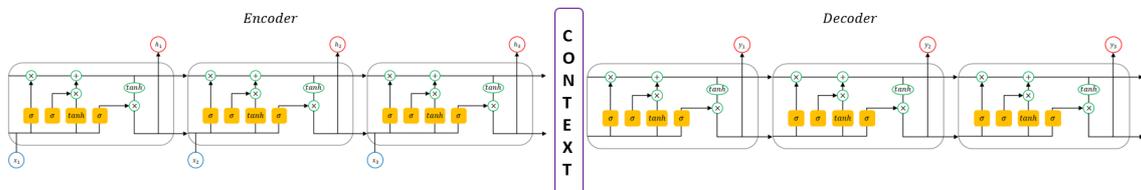


Figure 6. Seq2Seq

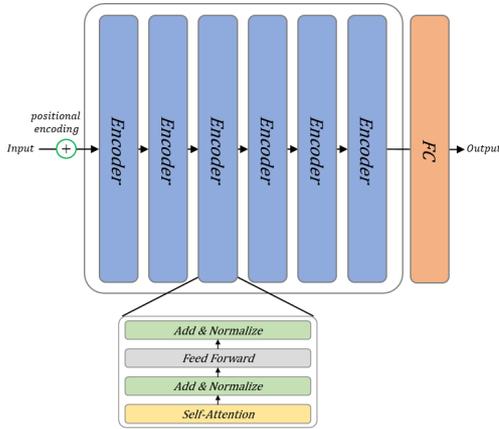


Figure 7. Transformer

Transformer는 Seq2Seq 모델이 입력 Sequence를 하나의 Context Vector로 압축하면서 정보가 손실되는 문제를 해결하기 위해 Self-Attention Mechanism과 Feed Forward Neural Network로만 구성된 모델이다(Vaswani et al., 2017). 이 모델은 자연어처리 뿐만 아니라 시계열 및 이미지 데이터 분석에도 매우 효과적이라고 알려져 있다. Transformer는 6개의 모두 같은 형태의 Layer로 구성된 Encoder-Decoder 구조이며, 세부적으로는 Multi-head Self-attention과 Position-wise Fully Connected Feed-forward Network로 구성된다. Encoder를 통과하기 전, Positional Encoding을 통하여 입력의 상대적인 위치를 반영하도록 하고, 각 Sub-layer마다 Residual Connection과 Layer Normalization을 수행한다. 본 연구에서는 시계열 데이터에 Transformer를 활용하기 위해 Decoder를 Fully-connected Layer로 대체하여 모델을 구축하였다. <Figure 7>은 본 연구에서 사용한 Transformer의 모델 구조를 나타낸다.

3.2 변수중요도 추출 알고리즘

SHapley Additive exPlanations(SHAP) values는 각 변수에 대한 특성을 나타내며, 특정 변수를 조건화했을 때 변경되는 모델 예측의 기댓값으로 변수 중요도를 측정한다(Lundberg and Lee, 2017). <Figure 8>은 변수에 대한 어떠한 정보도 없을 때의 Base value $E[f(z)]$ 에서부터 실제 모델 예측 $f(x)$ 까지 도달하는지를 설명한다. 모델이 비선형적이고 입력 변수가 독립적이지 않을 때 기댓값에 변수가 더해지는 순서가 영향을 미치기 때문에 SHAP values는 가능한 모든 순서에 대한 ϕ_i 의 평균으로 계산된다.

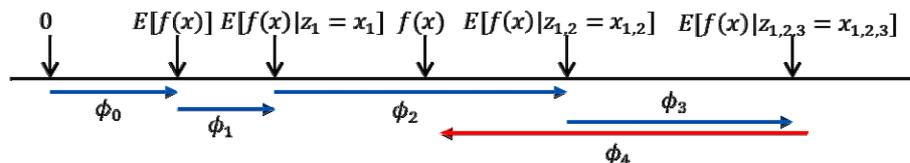


Figure 8. SHAP Values

3.3 데이터 설명 및 전처리

본 연구에서는 2011년 12월부터 2021년 5월까지 수집된 공정보고 공정률과 추가 데이터로 전국 날씨, 현장별 공사예산 금액, 예산 외 품의와 하도급 시행현황 데이터를 사용하였다. 전국 날씨 데이터를 제외한 모든 데이터는 현대건설 측에서 제공하였다. 건설 산업현장에 근무하는 전문가가 경험적으로 판단하였을 때 공정 지연에 영향력 요인들을 반영하여 입력 변수를 선정하였습니다. <Table 1>은 모델 학습에 사용된 최종 데이터의 변수명과 설명을 요약하여 나타낸 표이다.

공정보고 공정률 데이터는 각 현장의 월간 보고된 실적과 계획 공정률을 나타내며, 각 현장의 고유한 현장 코드와 현장의 사업부 구분, 보고기준일 그리고 실적과 계획 공정률로 구성된다. 본 연구의 종속변수인 실적 공정률의 실제로 진행된 공정률을 의미하며, 계획 공정률은 기존에 계획된 공정률을 의미한다. 계획된 전체 공사 기간 대비 진행된 공정 진행률을 파생 변수로 생성하였고, 보고기준일을 기준으로 연차, 월, 계절과 근무일 등의 시간 변수를 파생 변수로 생성하였다.

비정상적인 실적과 계획 공정률 추세를 가지는 현장에 대하여 전처리하였다. 정상적인 현장의 실적과 계획 공정률은 공정 진행에 따라 꾸준히 증가한다. 그러나, 건설현장의 특성상 발생하는 공정 계획의 변경 및 조정으로 인해 실적과 계획을 사람이 임의로 조작하여 이전 시점보다 실적과 계획이 감소하는 현상이 발생한다. 이러한 이상치에 대해 처리하기 위해 실적이 이전 시점보다 감소하는 시점을 기준으로 현장을 구분하여 정상적으로 증가하는 실적 공정률을 예측하도록 전처리하였다.

시계열 특성을 가진 공정보고 공정률 데이터의 이전 시점에 대한 정보를 반영하도록 처리하였다. 딥러닝 모델에 대해서는 Sequence 길이가 다른 각 현장에 대하여 고정된 크기의 Window로 Sequence를 밀어가며(Sliding) 입력 Sequence를 구성하여 실험하였고, Window의 크기는 5로 설정하였다. 머신러닝 모델은 딥러닝의 Sliding Window의 효과를 반영하기 위해 실적과 계획 공정률의 최근 5개월에 대한 Lag 변수를 파생 변수로 생성하였다.

전국 날씨 데이터는 기상청 기상자료개발 포털(<https://data.kma.go.kr/cmmn/main.do>)에서 제공하는 매일의 날씨 현장을 수집한 중관기상관측자료이다. 일간으로 기록된 기상정보를 월간 평균으로 계산하여 활용하였고, 수집된 기상정보에 대한 60개의 변수 중 비 작업 기상 조건과 관련된 6개의 변수를 사용하였다. 또한, 9개의 비 작업 기상 조건에 해당하는 일자를 계산하여 파생 변수로 생성하였다.

현장별 공사예산 금액 데이터는 각 현장에 배정된 세부 공사에 대한 예산을 나타내는 데이터이다. 세부 공사는 건축, 단지 외부, 설비, 전기, 조경, 철거와 토목으로 구분되며, 전체 현장 예산 대비 각 세부 공사의 비율을 파생 변수로 생성하였다.

예산 외 품의 데이터는 각 현장에서 기존에 책정된 예산 외에 필요한 추가 예산에 대해 품의한 정보는 포함한다. 품의가 승인된 일자와 전체 품의 금액과 당사 분의 품의 금액에 대한 데이터이다.

Table 1. Variable Description

Data	Variable	Description
Time Series Construction Data	Report Date	Date of report
	Planned	Planned construction of the report date
	Completed	Completed construction of the report date(Target)
	Progress	Progress compared to total planned construction
	Year	Working year of the report date
	Month	Month of the report date
	Season	Season of the report date
	Working Days	Number of working days of the report date
	Lag	Shifted value of Planned/Completed construction
Weather Forecast Data	Ave. Temp.	Average temperature
	Max Temp.	Maximum temperature
	Min Temp.	Minimum temperature
	Rainfall	Average amount of rainfall
	Wind	Maximum wind speed
	Snowfall	Average amount of snow
	Max Temp. 33	Number of days with Max Temp. over 33℃
	Max Temp. 35	Number of days with Max Temp. over 35℃
	Max Temp. 0	Number of days with Max Temp. under 0℃
	Min Temp. -12	Number of days with Min Temp. over -12℃
	Rainfall 5	Number of days with Rainfall over 5mm
	Rainfall 10	Number of days with Rainfall over 10mm
	Rainfall 20	Number of days with Rainfall over 20mm
	Snowfall 5	Number of days with Snowfall over 5mm
Wind 15	Number of days with Wind Speed over 15m/s	
Construction Budget Data	Construction	Amount of budget allocated for construction
	Outdoor	Amount of budget allocated for outdoor space
	Facility	Amount of budget allocated for facility
	Electricity	Amount of budget allocated for electricity
	Landscaping	Amount of budget allocated for landscaping
	Demolition	Amount of budget allocated for demolition
	Infrastructure	Amount of budget allocated for Infrastructure
	Ratio	Ratio of each construction
Extra-Budgetary Report Data	Total	Total amount of reported extra-budgetary
	Company	Reported extra-budgetary responsible for company
Subcontractor Data	Good	Number of good subcontractors on report date
	Normal	Number of normal subcontractors on report date
	Bad	Number of bad subcontractors on report date
	New	Number of new subcontractors on report date

하도급 시행현황 데이터는 각 현장에 투입된 하도급 업체와 고유한 업체 코드 그리고 업체의 공사 참여 기간을 나타내는 데이터이다. 하도급 업체가 참여한 현장과 참여 기간 중 발생한 지연 비율을 통하여 하도급 업체를 구분하였다. 지연 비율 30% 이하는 ‘ 좋음’, 30% 초과 60% 미만은 ‘ 보통’, 60% 이상은 ‘ 나쁨’으로 구분하였고, 구분되지 않은 그 외 업체는 ‘New’로 구분하였다. 구분된 하도급 업체를 현장과 참여 기간을 기준으로 매핑하여 카테고리별 하도급 업체의 수를 계산하여 파생 변수로 생성하였다.

4. 실험 결과

4.1 하이퍼파라미터 탐색

머신러닝 모델은 파이썬의 Scikit-learn 라이브러리를 이용하여 구현하였다. <Table 2>는 격자 탐색(Grid Search)을 통하여 탐색한 각 모델의 파라미터 범위를 나타낸 표이다.

딥러닝 모델은 파이썬의 Pytorch 라이브러리를 이용하여 구현하였다. 미니 배치의 크기는 64, 학습률은 0.001, Optimizer는 Adaptive Moment Estimation(Adam)(King and Ba, 2014), Epochs는 200으로 설정하였다. Scheduler는 ReduceLROnPlateau 사용했으며 Factor와 Patience를 0.1과 10으로 설정하여 Validation Set에 대한 Loss가 10 Epochs 내에 줄어들지 않았을 때 Learning Rate를 0.1배만큼 줄이도록 설정하였다. <Table 3>은 각 모델의 세부 파라미터를 나타낸 표이다.

4.2 평가 지표

4.2.1 Mean Absolute Error(MAE)

MAE는 실제값과 예측값의 오차에 대한 절댓값의 평균을 나타내는 값으로 계산식은 식 (4)와 같다.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4)$$

y_i 는 실제값을, \hat{y}_i 는 예측값을, n 는 표본의 개수를 나타낸다. 오차의 절댓값을 취하기 때문에 가장 직관적이며 이상치에 대하여 강건한 평가지표고, 본 연구의 주요 평가지표다.

4.2.2 Accuracy(ACC)

Accuracy는 올바르게 분류한 데이터의 수를 전체 데이터의 수로 나눈 값으로 계산식은 식 (5)와 같다.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

여기서 지연은 각 현장의 공정 완료에 대한 지연이 아니라 각 보고기준일에 목표 공정률 대비 실제 공정률의 늦어짐을 의미한다. 즉, 실적 공정률이 계획 공정률에 비해 낮으면 지연으로 구분하며, 예측한 실적 공정률 또한 같은 기준을 적용한다. 가장 직관적인 분류 평가 지표이며, 본 연구의 보조 평가 지표이다.

4.2.3 F1 Score

F1 Score는 정밀도(Precision)와 재현율(Recall)의 조화평균으로 계산되는 값이다. 정밀도와 재현율을 함께 반영하는 종합적인 분류 평가지표며, 본 연구의 보조 평가지표다. 정밀도는 모델이 지연으로 예측한 결과 중 실제로 지연인 비율을 나타내고, 재현율은 실제 지연인 결과 중 모델이 지연으로 예측한 결과의 비율을 나타낸다. 정밀도, 재현율 및 F1 Score에 대한 계산식은 식 (5)~(7)과 같다.

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

4.3 실험 결과

본 연구에서는 총 2개의 사업부(주택, 토목)에 소속된 218개

Table 2. Parameters of Machine Learning Models

XGB		SVR		RF	
n_estimators	[5, 10, 20, 50, 100, 200]	C	[0.1, 1, 10, 100]	n_estimators	[5, 10, 20, 50, 100, 200]
reg_alpha	[0.01, 0.1, 1, 10, 100]	gamma	['auto', 'scale']	max_features	['auto', 'sqrt', 'log2']
learning_rate	[0.001, 0.01, 0.1, 0.2, 0.3]	kernel	['rbf', 'poly', 'sigmoid']	max_depth	[5, 10, 25, 50, 100]
max_depth	[5, 10, 25, 50, 100]	degree	[1, 2, 3]	-	-

Table 3. Parameters of Deep Learning Models

RNN/LSTM		Seq2Seq		Transformer	
Number of Layers	[2, 4, 8, 16]	Number of Layers	[2, 4, 8, 16]	Number of Layers	[2, 4, 8, 16]
Hidden Size	[32, 64, 128]	Hidden Size	[32, 64, 128]	Number of Attention Heads	[1, 2, 3]
Dropout Rate	[0.0, 0.1]	Dropout Rate	[0.0, 0.1]	Dropout Rate	[0.0, 0.1]

의 현장에 대하여 실험하였다. 주택은 주택 및 아파트 등 주거용 건축물을 건설하는 현장을 포함하며, 토목은 도로, 항만, 교량, 철도 등의 토지를 조성 및 개량하는 현장을 포함한다. 토목 현장의 공사는 주택에 비해 규모가 크기 때문에 공사 기간이 2배가량 더 길고, 지연 비율 또한 6%가량 높다.

Train과 Test Set은 특정 시점을 기준으로 Set을 구분하는 것이 아닌 현대건설 측에서 평가하기를 원하는 현장을 기준으로 구분하였다. Train Set에 대한 5-Fold Cross Validation을 통해 최적의 하이퍼파라미터 탐색하였습니다. Train Set을 5개의 Fold로 균등하게 나눈 뒤, 4개의 Fold를 Train Set 나머지 1개의 Fold를 Validation Set으로 설정하여 각 Fold에 대해 예측 결과의 평균하였습니다. 5개 Fold 예측 결과의 평균값이 가장 우수하였던 하이퍼파라미터 조합을 선정하여 Train Set 전체를 재학습한 후 Test Set에 대하여 추론하였습니다. <Table 4>는 각 사업부의 Train과 Test 현장의 개수와 총 데이터의 크기를 나타낸 표이다.

본 연구에서는 과거 5개월의 공정보고 공정률 데이터와 추가 데이터를 활용하여 미래 3개월 공정률을 예측하였다.

본 연구에서는 종속변수를 실적 공정률(y_t)과 실적 공정률의 증가분(Δy_t)으로 구분해서 실험하였다. 단순히 실적 공정률을 종속변수로 설정하였을 때 모델이 이전 시점의 실적과 계획 공정률의 경향에 영향을 가장 많이 받으면서 시점마다 변화하는 시간, 기상 및 하도급업체 변수에 대한 정보를 잘 반영하지 못한다. 따라서, 종속변수에 따른 모델의 예측 성능과 도출되는 변수 중요도를 비교하여 공정보고 공정률 데이터와 추가 데이터를 활용한 공정률 예측에 적합한 종속변수를 선택한다. 실적 공정률의 증가분은 이전 시점의 실적 공정률에 예

측 실적 공정률 증가분을 더하여 예측 실적 공정률로 변환하여 실적 공정률과 공정하게 평가하였다. 평가 지표의 결과는 소수점 5자리에서 반올림하였다.

$$\Delta y_t = y_t - y_{t-1} \quad (8)$$

실험 1은 종속변수 실적 공정률(y_t)에 대한 각 모델의 예측 결과를 평가하였다. <Table 5>는 사업부별 모델의 1, 2, 3개월 후의 실적 공정률에 대한 예측 결과를 나타낸다.

실험 1은 사업부별 미래 3개월에 대한 실적 공정률 예측 결과의 MAE를 평균하여 모델의 성능을 평가하였다. 주택과 토목 사업부에 실적 공정률 예측 성능이 가장 우수한 모델은 각각 RF 모델과 XGB 모델이며, 최적의 하이퍼파라미터는 <Table 6>에 표기하였다. 모델 종류 간 성능을 비교하면 딥러닝 모델보다 머신러닝 모델의 성능이 더 우수하였는데, 이는 시계열 특성이 있지 않은 변수가 포함되어있기 때문에 순환신경망 기반의 딥러닝 모델이 효과적이지 못했다. 사업부별 성능을 비교하면 토목보다 주택 현장의 예측 성능이 우수했으며, 이는 공사 기간이 길어지고 지연 비율이 높을수록 공정률과 지연을 예측하기 어려워짐을 의미한다. 예측 시점별 성능을 비교하면 모든 모델이 가까운 미래일수록 즉 1, 2, 3개월 순으로 예측 성능이 우수하였다.

실험 1의 사업부별 예측 성능이 가장 우수한 모델의 SHAP value를 통해 변수 중요도를 추출하였다. <Figure 9>와 <Figure 10>은 각 모델에 대한 SHAP value 절댓값의 평균으로 변수 중요도를 계산하여 중요도가 높은 상위 10개의 변수를 시각화하였다.

Table 4. Number of Sites and of Data for Each Department

Department	Housing		Infrastructure	
	Train	Test	Train	Test
Number of Sites	106	22	42	18
Number of Samples	2442	498	2057	886

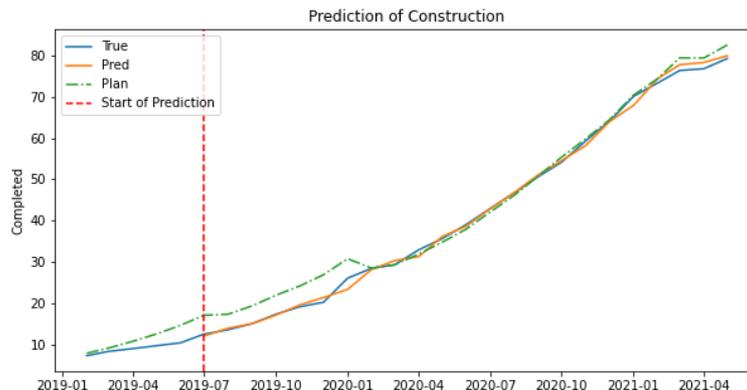


Figure 9. Overview of Model Prediction

Table 5. Prediction Results of Experiment 1

Department	Housing											
Month	1			2			3			Mean		
Evaluation Metrics	MAE	ACC	F1	MAE	ACC	F1	MAE	ACC	F1	MAE	ACC	F1
XGB	0.8111	0.8424	0.7458	1.3486	0.7709	0.5840	1.5336	0.7523	0.5368	1.2311	0.7885	0.6222
SVR	0.9271	0.8172	0.6904	1.3884	0.7379	0.5576	1.7610	0.7083	0.4474	1.3588	0.7545	0.5651
RF	0.7768	0.8382	0.7372	1.2305	0.7907	0.6468	1.5131	0.7269	0.5042	1.1735	0.7853	0.6294
RNN	0.8771	0.8299	0.7593	1.3457	0.8198	0.6900	1.9297	0.7733	0.6061	1.3842	0.8077	0.6851
LSTM	1.4092	0.7680	0.5632	2.0041	0.7558	0.4940	2.6704	0.7122	0.4142	2.0279	0.7453	0.4905
Seq2Seq	1.2362	0.7835	0.6111	1.7114	0.7703	0.5537	2.1845	0.7384	0.4828	1.7107	0.7641	0.5492
Transformer	1.7839	0.7423	0.4910	2.2899	0.7238	0.4551	2.7616	0.7035	0.3951	1.7089	0.7232	0.4471

Department	Infrastructure											
Month	1			2			3			Mean		
Evaluation Metrics	MAE	ACC	F1	MAE	ACC	F1	MAE	ACC	F1	MAE	ACC	F1
XGB	1.1713	0.7933	0.7538	1.8603	0.7259	0.6938	2.0169	0.7338	0.6917	1.6828	0.7510	0.7131
SVR	1.1645	0.7275	0.7157	1.8036	0.6612	0.6588	6.8385	0.5624	0.5499	3.2689	0.6504	0.6415
RF	1.0770	0.7771	0.7381	1.8077	0.7506	0.7196	2.2101	0.6799	0.6491	1.6983	0.7359	0.7023
RNN	1.1304	0.7344	0.6923	1.8682	0.6714	0.6172	2.4842	0.6377	0.5830	1.8276	0.6812	0.6308
LSTM	1.4290	0.7095	0.6746	2.0059	0.6571	0.6140	2.6623	0.6351	0.5862	2.0324	0.6672	0.6249
Seq2Seq	1.4294	0.7294	0.6862	1.9405	0.6857	0.6333	2.5032	0.6701	0.6163	1.9577	0.6951	0.6453
Transformer	2.9668	0.6322	0.6455	3.2031	0.6208	0.6452	3.4746	0.6091	0.6441	3.2148	0.6207	0.6449

Table 6. Hyperparameters for Experiment 1

Dept.	Housing			Dept.	Infrastructure		
Model	RF			Model	XGB		
Parameter	1	2	3	Parameter	1	2	3
n_estimators	200	50	20	n_estimators	100	100	200
max_features	auto	auto	auto	reg_alpha	0.1	0.01	0.01
max_depth	10	10	10	learning_rate	0.2	0.2	0.1
-	-	-	-	max_depth	5	25	5

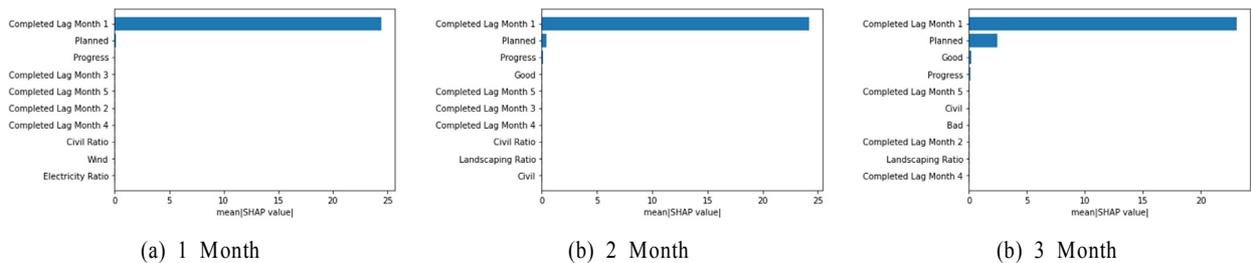


Figure 10. SHAP value of Experiment 1(Housing)

주택의 변수별 SHAP value는 Lag Feature인 1개월 전 시점의 실적 공정률이 가장 높았으며, 그다음으로 계획 공정률이 높았다. 특히, 계획 공정률은 예측 시점이 멀어질수록 중요도가 높아졌으며, 이는 모델이 예측 시점이 멀어질수록 실적 공정률을 예측하는데 계획 공정률을 더 많이 활용한다는 것을 의미한다. 그 외 공정 진행률, “Good” 하도급 업체의 수, 그리고 실적 공정률의 Lag Feature 순으로 중요도가 높았지만, SHAP

value는 가장 중요한 변수인 1개월 전 시점의 실적 공정률에 비해 낮았다. 이는 1개월 전 시점의 실적 공정률 외에 다른 변수는 모델 예측 결과에 대한 기여도가 매우 낮음을 나타낸다.

토목의 변수별 SHAP value는 1개월 예측 모델은 1개월 전 시점의 실적 공정률이 가장 높았으나 2개월과 3개월 예측 모델은 예측 시점의 계획이 가장 높았고, 그다음으로 1개월 전 시점의 실적 공정률이 높았다. 이는 주택과 마찬가지로 모델

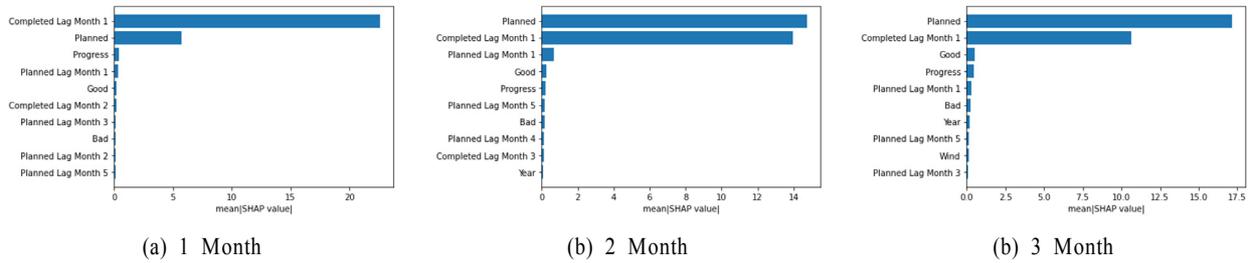


Figure 11. SHAP value of Experiment 1 (Infrastructure)

이 예측 시점이 멀어질수록 실적 공정률을 예측하는데 계획 공정률을 더 많이 활용한다는 것을 의미하지만, 토목 현장에서는 실적 공정률을 예측하는데 실적 공정률보다 계획 공정률의 기여도가 더 높음을 의미한다. 그 외 공정 진행률, “Good” 하도급 업체의 수, 그리고 계획과 실적 공정률의 Lag Feature의 중요도가 높았지만, 주택과 유사하게 상위 2개의 변수의 SHAP value만 극단적으로 높았다.

실험 2는 종속변수 실적 공정률의 증가분(Δy_t)에 대한 각 모델의 예측 결과를 평가하였다. <Table 7>은 사업부별 모델의 1, 2, 3개월 후의 실적 공정률의 증가분(Δy_t)에 대한 예측 결과를 나타낸다.

실험 2는 사업부별 미래 3개월에 대한 실적 공정률의 증가분 예측 결과의 MAE를 평균하여 모델의 성능을 평가하였다. 주택과 토목 사업부에 실적 예측 성능이 가장 우수한 모델은

XGB 모델이며, 최적의 하이퍼파라미터는 <Table 8>에 표기하였다. 실험 1과 유사하게 모델 종류 간 성능을 비교하면 딥러닝 모델보다 머신러닝 모델의 성능이 더 우수하였으며, 사업부별 성능을 비교하면 토목보다 주택 현장의 예측 성능이 우수했다. 예측 시점별 성능을 비교하면 대부분 모델이 가까운 미래일수록 즉 1, 2, 3개월 순으로 예측 성능이 우수하였다. 실험 1과 2를 비교하였을 때, 예측 결과의 대부분이 실험 2에서 MAE가 더 낮았다. 이는 실적 공정률의 증가분을 종속변수로 설정하였을 때 모델의 예측 성능이 더 우수하다는 것을 나타낸다.

실험 2의 사업부별 예측 성능이 가장 우수한 모델의 SHAP value를 통해 변수 중요도를 추출하였다. <Figure 12>와 <Figure 13>은 각 모델에 대한 SHAP value 절댓값의 평균으로 변수 중요도를 계산하여 중요도가 높은 상위 10개의 변수를 시각화하였다.

Table 7. Prediction Results of Experiment 2

Department	Housing											
Month	1			2			3			Mean		
Evaluation Metrics	MAE	ACC	F1	MAE	ACC	F1	MAE	ACC	F1	MAE	ACC	F1
XGB	0.5967	0.8538	0.7752	0.7255	0.8565	0.7746	0.8881	0.8214	0.7331	0.7368	0.8439	0.7610
SVR	0.6689	0.8411	0.7573	0.7801	0.8318	0.7458	0.9014	0.8190	0.7361	0.7835	0.8306	0.7464
RF	0.6172	0.8665	0.7974	0.7684	0.8587	0.7789	0.8718	0.8476	0.7681	0.7525	0.8576	0.7815
RNN	0.9872	0.7392	0.6941	1.0810	0.7338	0.6696	1.1788	0.7427	0.6601	1.0823	0.7386	0.6746
LSTM	0.9774	0.7923	0.6634	1.0107	0.7907	0.6636	1.0585	0.7950	0.6733	1.0155	0.7927	0.6668
Seq2Seq	0.9756	0.7896	0.6667	0.9685	0.7936	0.6603	1.0148	0.7919	0.6492	0.9863	0.7917	0.6587
Transformer	1.1522	0.7978	0.5925	1.1820	0.8081	0.5805	1.1777	0.8043	0.5743	1.1706	0.8034	0.5824

Department	Infrastructure											
Month	1			2			3			Mean		
Evaluation Metrics	MAE	ACC	F1	MAE	ACC	F1	MAE	ACC	F1	MAE	ACC	F1
XGB	0.6393	0.8151	0.7843	0.7786	0.8048	0.7747	0.9021	0.7864	0.7581	0.7733	0.8021	0.7724
SVR	0.9631	0.7515	0.7376	0.9929	0.7487	0.7404	1.1025	0.7524	0.7347	1.0195	0.7509	0.7376
RF	0.6526	0.8163	0.7805	0.7968	0.8125	0.7783	0.9283	0.7837	0.7488	0.7926	0.8042	0.7692
RNN	3.0054	0.7392	0.6503	2.8156	0.7338	0.6555	2.8088	0.7427	0.6608	2.8766	0.7386	0.6555
LSTM	2.419	0.8232	0.7542	2.656	0.8026	0.7143	2.495	0.7984	0.7088	2.5233	0.8081	0.7258
Seq2Seq	2.2504	0.7774	0.6988	2.0576	0.8000	0.7363	2.1092	0.7931	0.7329	2.1391	0.7902	0.7227
Transformer	2.1599	0.8346	0.7427	2.1217	0.8364	0.7439	2.1100	0.8435	0.7500	2.1305	0.8382	0.7455

Table 8. Hyperparameters for Experiment 2

Dept.	Housing			Dept.	Infrastructure		
Model	XGB			Model	XGB		
Parameter	1	2	3	Parameter	1	2	3
n_estimators	100	50	20	n_estimators	20	50	10
reg_alpha	1	1	0.1	reg_alpha	0.1	0.1	1
learning_rate	0.1	0.1	0.2	learning_rate	0.2	0.2	0.2
max_depth	5	5	5	max_depth	5	5	5

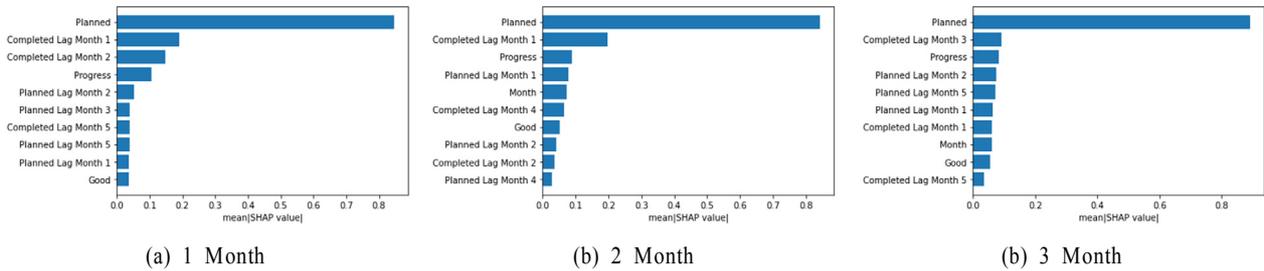


Figure 12. SHAP value of Experiment 2(Housing)

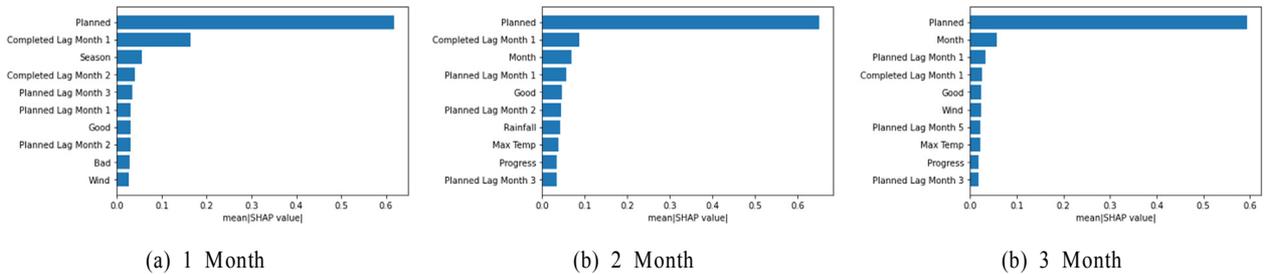


Figure 13. SHAP value of Experiment 2 (Infrastructure)

주택의 변수별 SHAP value는 계획 공정률이 가장 높았으며, 그다음으로 1개월 혹은 3개월 전 시점의 실적 공정률이 높았다. 그 외 실험 1과 유사하게 계획된 전체 공사 기간 대비 진행된 공정 진행률, 그리고 실적과 계획 공정률의 Lag Feature의 중요도가 높았으며, 그리고 새롭게 시간 변수인 보고기준일의 월을 나타내는 변수의 중요도가 높았다. 또한, 실험 1과 비교하였을 때 여러 변수의 SHAP value가 증가하였으며, 이는 실험 2가 실험 1보다 다양한 변수의 상관관계를 고려하여 모델이 예측을 수행했음을 알 수 있다.

토목의 변수별 SHAP value는 계획 공정률이 가장 높았으며, 그다음으로 1개월 전 시점의 실적 공정률 혹은 보고기준일의 월이 높았다. 그 외 계절 변수, 실적과 계획 공정률의 Lag Feature, “Good” 하도급 업체의 수, 그리고 강수량, 최고기온과 풍속 등 기상 변수의 중요도가 높았다. 이전 실험에서는 나타나지 않았던 계절과 기상 변수의 중요도가 높은 것으로 보아, 토목 현장의 공사는 계절과 기상의 영향을 많이 받는 것으로 해석된다.

5. 결론 및 향후 연구

본 연구에서는 주택과 토목 2개의 건설 사업부의 미래 3개월의 실적 공정률을 예측하는 문제에 대하여 7가지 인공지능 모델의 성능을 비교하고 변수 중요도를 추출하였다. 특히, 중속 변수를 실적과 실적 증가분으로 구분해서 실험했으며, 이에 따른 예측성과 변수 중요도를 비교하였다. 최종적으로 XGB 모델을 활용하여 실적 공정률의 증가분을 중속변수로 설정하였을 때 미래 실적 공정률을 더 정교하게 예측하였다. 그럴 뿐만 아니라 SHAP를 통해 변수 중요도를 분석하였을 때 다양한 변수들의 상관관계가 모델 예측에 반영된 것으로 나타났다. 이를 통해 공정보고 공정률과 추가 데이터를 활용하는 건설업에서 실적 공정률의 증가분을 예측하는 XGB 모델을 활용하여 미래 3개월의 실적을 정확하게 예측하고, 주요 지연 요인을 파악할 수 있음을 나타낸다.

향후 과제 및 연구 방향은 아래와 같다. 본 연구의 데이터는 시계열 특성을 가진 공정보고 공정률, 전국 날씨, 하도급 시행

현황 데이터와 시계열 특성을 가지지 않는 현장별 공사예산 금액과 예산 외 품의 데이터로 구성된다. 시계열 데이터 분석에 활용 가능한 딥러닝 모델이 다양한 크기의 변수를 입력받을 수 있지만, 시계열 특성이 없는 변수가 포함되었기 때문에 성능이 저조하였다. 시계열 변수와 그렇지 않은 변수가 함께 입력 변수로 구성되었기 때문에 딥러닝 시계열 예측 모델보다 의사결정나무 기반의 머신러닝 모델의 성능이 우수하였다. 따라서, 데이터의 특성에 맞게 모델을 학습하여 양상불하는 Multi-Model Stacking 모델을 통해 더 높은 성능향상을 기대할 수 있다. 또한, 주택과 토목 2개의 사업부 전체에 월간 보고된 실적 공정률은 총 5,883건이다. 이는 인공지능 모델 특히 딥러닝 모델 구축을 위해 충분한 데이터 크기는 아니며, 매월 보고된 실적 공정률은 보고 사이에 시간이 길뿐만 아니라 사람에 의해 임의로 조정되는 이상치 또한 많이 존재한다. 이 때문에 더욱 정확한 공정 지연 예측 모델 구축을 위해서는 시계열 특성을 가진 데이터를 정확하게 추가 수집하는 것이 필요하다.

참고문헌

- Breiman, L. (2001), Random forest, *Machine Learning*, **45**, 5-32.
- Chen, T. and Guestrin, C. (2016), XGBoost: A scalable tree boosting system, *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.
- Construction Association of Korea (2020), http://www.cak.or.kr/board/boardView.do?menuId=86&cms_site_id=&sel_tab=&searchctg1=&searchCondition=all&searchKeyword=&sidhp=&subhp=&boardId=statistic_build&dataId=37495&pageIndex=1.
- Cortes, C. and Vapnik, V. (1995), Support-vector networks, *Machine Learning*, **20**(3), 273-297.
- Egwim, C. N., Alaka, H., Toriola-Coker, L. O., Balogun, H., and Sunmola, F. (2021), Applied artificial intelligence for predicting construction projects delay, *Machine Learning with Applications*, **6**, 100166.
- Gondia, A., Sian, A., El-Dakhakhni, W., and Nassar, A. H. (2020), Machine learning algorithms for construction projects delay risk prediction, *Journal of Construction Engineering and Management*, **146**(1), 04019085.
- Hochreiter, S. and Schmidhuber, J. (1997), Long short-term memory, *Neural Computation*, **9**(8), 1735-1780.
- Kim, H.-M., Kim, T.-H., Shin, Y.-K., Kim, Y.-S., and Han, S. (2011), Regression technique-based productivity estimation conducting construction delay factor analysis on interior works in high-rise building construction, *The Korea Institute of Building Construction*, **11**(1), 323-324.
- Korean Statistical Information Service.(2020), https://kosis.kr/statHtml/statHtml.do?orgId=301&tblId=DT_111Y002&vw_cd=&list_id=&scrId=&seqNo=&lang_mode=ko&obj_var_id=&itm_id=&conn_path=E1&docId=0193134234&markType=S&itmNm=%EC%A0%84%EA%B5%AD.
- Lundberg, S. M. and Lee, S.-I. (2017), A unified approach to interpreting model predictions, *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 4768-4777.
- Min, H.-S., Eom, S.-J., and Hong, P.-J. (2005), Developing predictive model for delay of building remodeling projects, *Architectural Institute of Korea*, **25**(1), 543-546.
- Rumelhart, D., Hinton, G., and Williams, R. (1986), Learning representations by back-propagating errors, *Nature*, **323**, 533-536.
- Sanni-Anibire, M. O., Zin, R. M., and Olatunji, S. O. (2020), Machine learning model for delay risk assessment in tall building projects, *International Journal of Construction Management*.
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014), Sequence to sequence learning with neural networks, *NIPS '14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, **2**, 3104-3112.
- Vaswan, A., Shazee, N., Parma, N., Uszkorei, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017), Attention is all you need, *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 6000-6010.
- Zhu, H., Wang, X., Chen, X., and Zhang, L. (2020), Similarity search and performance prediction of shield tunnels in operation through time series data mining, *Automation in Construction*, **114**, 103178.

저자소개

정희찬: 동국대학교 경영학과에서 2020년 학사를 취득하고, 고려대학교에서 산업경영공학과 석사과정에 재학 중이다. 연구 분야는 Machine Learning, Image Denoising, Time Series Forecasting 이다.

한기웅: 가톨릭대학교 수학과에서 2020년 학사학위를 취득하고, 고려대학교에서 산업경영공학과 석사과정에 재학 중이다. 연구 분야는 Deep Learning in Non-Euclidean spaces, Variational Inference, Representation Learning이다.

이병훈: 서울과학기술대학교 산업정보시스템학과에서 2021년 학사학위를 취득하고, 고려대학교에서 산업경영공학과 석·박사과정에 재학 중이다. 연구 분야는 Machine Learning, Self-supervised Learning, Multimodal Deep Learning이다.

한예지: 고려대학교 세종캠퍼스 응용수리과학부에서 2021년 학사학위를 취득하고, 고려대학교에서 산업경영공학과 석·박사과정에 재학 중이다. 연구 분야는 Machine Learning, Multimodal Deep Learning이다.

허경휘: 연세대학교 경제학부에서 2015년 학사학위, 성균관대학교 정보통신대학원에서 2021년 석사학위를 취득하였다. 현재 현대건설 기술연구원 연구원으로 재직 중이다. 연구 분야는 Machine Learning, Text Mining, Image Classification 등이다.

최호창: 국민대학교 경영학부에서 2017년 학사학위, 비즈니스IT

전문대학원에서 2019년 석사학위를 취득하였다. 현재 현대건설 기술연구원 연구원으로 재직 중이다. 연구 분야는 Data Mining, Text Mining, Image Mining, Social Network Analysis 등이다.

한성원 : 고려대학교 산업 시스템 정보 공학과에서 2003년 학사학위를 취득하였다. Georgia Institute of Technology에서 2006년 Operation Research 석사학위, 2007년 Statistics 석사학위, 2010년 Mathematics 석사학위를 취득하고 2010년 Industrial Engineering 과 Statistics 박사학위를 취득하였다. University of Pennsylvania, Department of Biostatistics and Epidemiology에서 Post-doctoral

Researcher(2010.07-2012.06), Hoffmann-La Roche Inc., Department of Non-clinical Safety에서 Post-doctoral Fellow(2012.07-2013.08), New York University, Department of Population Health에서 Research Scientist(2013.08-2015.12), New York University, Department of Population Health에서 Senior Research Scientist (2016.01-2016.02)를 역임하고, 2016년부터 고려대학교 산업경영공학부 교수로 재직하고 있다. 연구분야 중 방법론 분야는 probabilistic graphical model, network analysis, deep learning 등이 있으며, 응용 분야로는 바이오 의료, 소재 정보학, 품질 모니터링, 텍스트 마이닝이 있다.