JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 27 No. 5, pp. 29-36, May 2022
https://doi.org/10.9708/jksci.2022.27.05.029

A Design and Implement of Efficient Agricultural Product Price Prediction Model

Jung-Ju Im*, Tae-Wan Kim*, Ji-Seoup Lim*, Jun-Ho Kim**, Tae-Yong Yoo**, Won Joo Lee**

*Graduate Student, Dept. of Applied Artificial Intelligence, Hanyang University, Ansan, Korea
*Graduate Student, Dept. of Applied Artificial Intelligence, Hanyang University, Ansan, Korea
*Graduate Student, Dept. of Applied Artificial Intelligence, Hanyang University, Ansan, Korea
**Student, Dept. of Computer Science & Engineering, Inha Technical College, Incheon Korea
**Student, Dept. of Computer Science & Engineering, Inha Technical College, Incheon Korea
**Professor, Dept. of Computer Science & Engineering, Inha Technical College, Incheon Korea

[Abstract]

In this paper, we propose an efficient agricultural products price prediction model based on dataset which provided in DACON. This model is XGBoost and CatBoost, and as an algorithm of the Gradient Boosting series, the average accuracy and execution time are superior to the existing Logistic Regression and Random Forest. Based on these advantages, we design a machine learning model that predicts prices 1 week, 2 weeks, and 4 weeks from the previous prices of agricultural products. The XGBoost model can derive the best performance by adjusting hyperparameters using the XGBoost Regressor library, which is a regression model. The implemented model is verified using the API provided by DACON, and performance evaluation is performed for each model. Because XGBoost conducts its own overfitting regulation, it derives excellent performance despite a small dataset, but it was found that the performance was lower than LGBM in terms of temporal performance such as learning time and prediction time.

▸ **Key words**: Agricultural Product Price forecasting, Machine Learning, Gradient Boosting Algorithm, DACON

[요 약]

본 논문에서는 DACON에서 제공하는 데이터셋을 기반으로 한 효과적인 농산물 가격 예측 모델을 제안한다. 이 모델은 XGBoost와 CatBoost 이며 Gradient Boosting 계열의 알고리즘으로써 기존의 Logistic Regression과 Random Forest보다 평균정확도 및 수행시간이 우수하다. 이러한 장점들을 기반으로 농산물의 이전 가격들을 기반으로 1주, 2주, 4주뒤 가격을 예측하는 머신러닝 모델을 설계한다. XGBoost 모델은 회귀 방식의 모델링인 XGBoost Regressor 라이브러리를 사용하여 하이퍼 파라미터를 조정함으로써 가장 우수한 성능을 도출할 수 있다. CatBoost 모델은 CatBoost Regressor를 사용하여 모델을 구현한다. 구현한 모델은 DACON에서 제공하는 API를 이용하여 검증하고, 모델 별 성능평가를 실시한다. XGBoost는 자체적인 과적합 규제를 진행하기 때문에 적은 데이터셋에도 불구하고 우수한 성능을 도출하지만, 학습시간, 예측시간 등 시간적인 성능 면에서는 LGBM보다 성능이 낮다는 것을 알 수 있었다.

▸ **주제어**: 농산물 가격예측, 머신러닝, Gradient Boosting Algorithm, DACON

• First Author: Jung-Ju Im, Corresponding Author: Won Joo Lee
*Jung-Ju Im (dlawjdwn12@naver.com), Dept. of Applied Artificial Intelligence, Hanyang University
*Tae-Wan Kim (zmzmdlfjs@gmail.com), Dept. of Applied Artificial Intelligence, Hanyang University
*Ji-Seoup Lim (crobat126@gmail.com), Dept. of Applied Artificial Intelligence, Hanyang University
**Jun-Ho Kim (rlawnsgh8395@naver.com), Dept. of Computer Science & Engineering, Inha Technical College
**Tae-Yong Yoo (dbxodyd9162@naver.com), Dept. of Computer Science & Engineering, Inha Technical College
**Won Joo Lee (wonjoo2@gmail.com), Dept. of Computer Science & Engineering, Inha Technical College
• Received: 2022. 01. 18, Revised: 2022. 02. 17, Accepted: 2022. 02. 17.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

자연 재난은 태풍, 홍수, 호우, 강풍, 풍랑, 해일, 대설, 한파, 낙뢰, 가뭄, 폭염, 지진, 황사, 조류 대발생, 조수, 화산활동, 소행성·유성체 등 자연 우주물체의 추락·충돌 등의 자연현상으로 인하여 발생하는 재해이다[1], 행정안전부 재해 연보에 따르면 2020년도에는 총 27회의 크고 작은 재난이 발생하여 72명이 사망하고 13,182억 원의 재산피해가 발생하였다[1]. 이러한 자연재해는 직•간접적으로 큰 인명피해와 농산물 재배에도 큰 피해를 준다. 농림축산식품부의 농작물재해보험의 최근 6년간 보험금 지급액 또한 2015년도를 기준으로 2020년도까지 꾸준히 증가하고 있다[2], 이러한 피해는 사전에 예측하여 대응책을 마련한다면 그 피해를 줄일 수 있으며, 농부들에게 필수적인 정보가 될 것이다[3]. 태풍은 많은 비와 바람을 동반하여 농가에 피해를 주며 공급이 부족한 농산물의 가격 상승에 영향을 준다. 이러한 직접적인 피해와 다르게 간접적으로 소비자 물가 상승에도 영향을 미치기 때문에 관심을 가져야 한다.

농산물의 가격 상승이 자연재해에 따른 것만은 아니지만 기후변화에 따라 농산물의 가격이 변동한다. 계절별로 생산량이 부족해 공급이 부족한 농산물은 가격이 오를 것이고, 생산량이 많아 공급이 기존 재고보다 많아진다면 농산물의 가격은 낮아질 것이다. 하지만 생산량에 따른 가격 변동만 있는 것이 아닌 유통망, 수요, 기후변화, 트렌드 등 수많은 요인이 농산물의 가격변동에 영향을 준다. 하나의 예로 추석이나 설날 등 명절 음식 재료들은 일정 기간에 수요가 증가하며 가격이 인상된다[4]. 이 기간에 태풍과 같은 자연재해나 기타 외부적인 요인들이 발생한다면 그 가격의 인상 폭은 커지게 된다. 가격에 영향을 주는 요인들을 과거의 데이터를 통해 미리 파악하는 예측하는 모델을 만든다면 농산물 공급과 수입량 등 가격변동에 따른 공급량을 사전에 조절한다면 농산물 가격을 안정시킬 수 있을 것이다. 따라서 본 논문에서는 농산물 가격 예측 모델을 개발하고, 여러 알고리즘을 통하여 성능을 평가하고자 한다. 또한, 개발한 농산물 가격 예측 모델로 Kaggle ML competition에 참여한다.

Kaggle ML competition은 데이터 과학 경진대회에서 가장 잘 알려진 기계 학습 경진대회 플랫폼이다[5]. 이 경진대회에는 많은 참가자가 팀 또는 개인 자격으로 참가한다. 최근 Kaggle ML Competition에서 Boosting 알고리즘들이 우승하며 성능의 우수성을 입증하고 있다. 특히, XGBoost는 검색엔진의 추천 시스템에도 적용되며 일종의 브랜드로 자리잡고 있다[6]. 또한 XGB(XGBoost)와 LGB(LightGBM)는 기존의 LR(Logistic Regression)과 RF(Random Forest) 알고리즘보다 평균 정확도가 높고, 훈련 시간도 RF보다 1~1.5배 정도 빠르다. 하지만 데이터 크기가 작은 데이터셋의 경우 LGB의 평균 정확도는 떨어진다[7]. 이러한 빠른 훈련 시간(Training Time)과 정확도(Accuracy)를 갖춘 부스팅 알고리즘들의 특성을 이용하여 DACON 농산물 가격 예측 모델을 개발하고, 성능을 비교 분석한다.

본 논문의 구성은 다음과 같다. 2장에서는 DACON Baseline 1에서 사용된 Seq2Seq와 대표적인 Boosting 알고리즘 LGBM, XGBoost, CatBoost를 설명한다. 3장에서는 2장에서 설명한 모델 중 과적합 규제가 내장된 XGBoost와 빠른 수행시간을 가진 CatBoost를 직접 적용해 모델링한 농산물 가격 예측 모델을 제안한다. 4장에서는 제안한 농산물 가격 예측 모델의 성능을 평가하고, 5장에서 결론을 맺는다.

II. Preliminaries

1. Related works

본 논문에서 적용하고자 하는 XGBoost와 CatBoost 뿐만 아니라 DACON에서 제공하는 Baseline에서 사용되는 Seq2Seq와 LGBM의 알고리즘의 특성은 표 1과 같다.

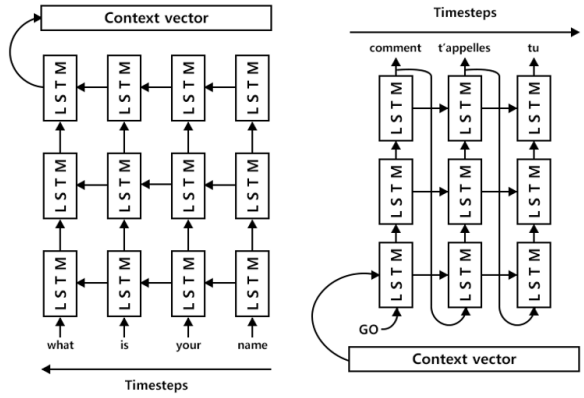
Table 1. Analysis of existing algorithms

Algorithm	Structure	Advantages	Disadvantages
Seq2Seq	Encoder, Decoder	Sequence different length	The longer the sentence, the more difficult it is to acquire information.
LGBM	leaf oriented tree split	Faster learning and execution time compared to GBoost	Increased chance of overfitting when using small datasets
CatBoost	Symmetrical tree	Categorical variable automatic preprocessing	Not suitable for error-rich datasets
XGBoost	Decision tree	Faster execution time compared to GBM, over-conformity regulation	When the dataset is large, the training time is slow.

1.1 Seq2Seq(Sequence to Sequence)

Seq2Seq는 두 개의 순환신경망(RNN)으로 이루어진 종단 간 학습 모델이다. Seq2Seq를 구성하는 두 개의 순환

신경망은 입력값을 받아 고정된 크기의 Context Vector로 변환하는 Encoder와 Context Vector를 Seed로써 사용해 출력값을 만드는 Decoder로 구성된다[8].



(a) Encoder (b) Decoder

Fig. 1. Structure of Seq2Seq

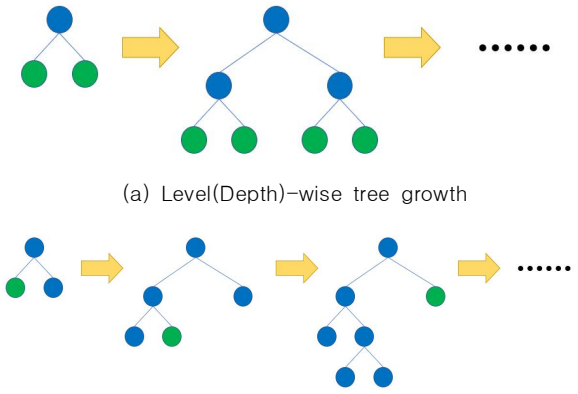
그림 1의 (a) Encoder는 시계열 데이터를 입력받아 하나의 벡터로 정보를 압축하는 기능을 제공하며 입력 시퀀스를 역전된 순서로 처리하는 특징을 가진다. (b) Decoder는 압축된 데이터를 다른 시계열 데이터로 변환하는 기능을 제공하며 stacked LSTM 구조로 이루어져 있다. Decoder는 Encoder와 달리 hidden state가 Encoder에서 생성한 Context Vector로 initializing 된다.

1.2 Gradient Boosting

Gradient Boost는 Boosting 계열의 앙상블 알고리즘이다[9]. GBM(Gradient Boosting Machine)은 모델 X를 통해 A를 예측하고 남은 잔차(residual)를 통해 다시 Y라는 모델을 만든 후 X+Y 모델을 통해 A를 예측하여 모델을 만들게 된다. 손실함수(loss function)를 제곱 오차(squared error)로 설정했을 때 잔차는 negative gradient를 가지게 된다. 따라서 residual fitting을 통해 모델을 만드는 것은 negative gradient를 이용해 다음 모델을 순차적으로 생성한다. 이 방법을 통해 잔차를 줄여나가며 예측 모형을 만들게 되며, 경사하강을 통해 다음 모델을 생성하기 때문에 Gradient Boosting Algorithm이라 한다. 또한 예측 모형에서 발생하는 과적합(Over-Fitting)을 예방하기 위해 Regularization을 사용한다.

1.3 LightGBM

LightGBM은 트리 기반 학습 알고리즘을 사용하는 그래디언트 부스팅 프레임워크이다[10]. GBDT(Gradient Boosting Decision Tree)는 정보를 획득할 수 있는 지점의 모든 데이터 인스턴스를 탐색하여 많은 시간을 소비하는 문제를 해결하기 위해 두 가지 방법을 사용한다. 첫 번째로 GOSS(Gradient-Based One-Side Sampling)과 EFB(Exclusive Feature Bundling)이다. GOSS는 작은 경사 값을 가지는 데이터의 많은 부분을 제외하고 나머지 부분을 통하여 정보를 얻는다. EFB는 데이터의 Feature 개수를 감소시키기 위해 상호 배타적인 Feature들을 하나로 묶어, 효과적으로 Feature들의 수를 감소시킨다. 또한 LGBM은 히스토그램 기반 알고리즘을 사용하여 연속적인 feature 값을 개별적인 bins로 묶기 때문에, 훈련 속도가 빠르고 메모리 사용량이 줄어든다.



(a) Level(Depth)-wise tree growth

(b) Leaf-wise (Best-Frist) tree growth

Fig. 2. Tree growth method

결정 트리 학습 알고리즘은 그림 2(a)와 같이 level (depth)-wise 방식으로 트리를 성장시키는 반면, LGBM은 그림 2(b)와 같은 leaf-wise 방식으로 트리를 성장시킨다. 또한, leaf-wise 방식은 데이터가 작을 때 과적합을 유발할 수 있으므로, LightGBM에는 트리의 최대 깊이를 제한하는 매개 변수를 포함한다. 또한, LightGBM은 데이터를 수직으로 나누지 않고, 전체 데이터를 보유하기 때문에, 데이터 분할 결과에 대해 통신할 필요가 없고 데이터가 더 커지지 않기 때문에 분산, 병렬 처리에도 합리적이다. 그 외에도 GPU 학습을 지원한다는 이점이 있으며, 효율적으로 분산되도록 설계된 프레임워크이다.

1.4 CatBoost

GBM의 과적합 문제를 해결하면서 기존의 GBM 계열 알고리즘(XGBoost, LGBM)보다 학습속도를 개선하는 장점을 앞세워 개발되었다[11]. Gradient Boosting 알고리즘의 구현에 존재하는 Target Leakage로 인해 발생하는

Prediction Shift에 대응하기 위해 순서형 부스팅과 범주형 Feature를 처리하기 위해 제안되었다. 또한, CatBoost는 기존의 GBM 계열 알고리즘의 트리 구조와는 다르게 대칭 트리 형성 구조를 갖는다. 이는 기존 GBM 계열 알고리즘과 비교하여 CatBoost만의 장점이다. 또한, CatBoost는 트리 구조에 대한 Ordered Mode와 Plain Mode 두 가지의 모드를 갖는다. Ordered Mode는 불규칙한 훈련 데이터 셋에 관해 사용되며, 상대적으로 작은(40k 미만 훈련) Dataset 에서 유용한 걸 확인할 수 있다. Plain Mode는 표준 GBM 알고리즘에 사용된다.

1.5 XGBoost

부스팅 알고리즘 중 하나인 XGBoost는 분산 환경에 최적화된 성능과 자원 효율이 우수한 그래디언트 부스팅 라이브러이다[12]. XGBoost는 병렬 처리로 학습하여 분류 속도가 빨라 GBM 대비 빠른 수행시간을 가지며, 그림 3과 같이 병렬 트리 부스팅을 제공한다. 또한, 표준 GBM의 경우 과적합 규제기능이 없으나, XGBoost는 자체 과적합규제 기능과 CART(Classification and regression tree) 앙상블 모델을 사용함으로 분류와 회귀영역에서 뛰어난 예측 성능을 발휘한다. XGBoost는 Early Stopping, Cross Validation을 지원하며 신경망과 비교하면 시각화가 쉽다. 하지만 영상 인식, 컴퓨터 비전, 자연어 처리와 문제 이해 등 같은 경우에는 오히려 딥러닝보다 효율이 떨어진다.

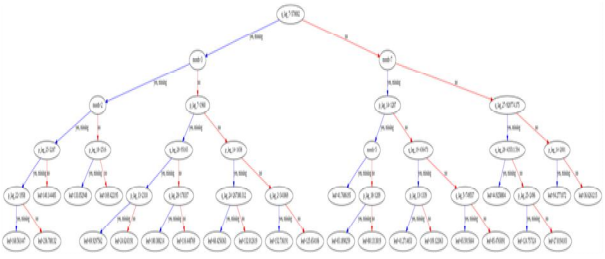


Fig. 3. Decision Tree of XGBoost

III. The Propose of Agricultural Product Price Prediction Model

1. Agricultural Product Price Prediction Model Design

본 논문에서는 DACON에서 제공하는 Baseline인 Seq2Seq, LGBM와 XGBoost, CatBoost를 사용하여 모델을 설계하고 구현한다. 본 논문에서 사용한 데이터셋은 수치형 데이터셋으로 CatBoost의 범주형 변수와는 다르지만, CatBoost의 결정 트리 구조인 대칭 트리 형성을 통한 성능 향상을 기대한다. 표 2와 같이 DACON에서 제공하는 Public Dataset의 Training, Testing 데이터셋을 사용하여 모델링 한다. Train 및 Test에 사용되는 데이터셋은 각 품목, 거래량, 가격을 feature로 사용한다.

Table 2. Dataset provided by DACON

Dataset	Name
Training	train.csv
Validation	train.csv
Testing	test_year-month-day.csv

표 2의 Training 데이터셋은 2016년부터 2020년까지의 농산물의 가격 데이터이다. Testing 데이터셋은 2020년 9월부터 11월까지의 데이터이다.

또한, DACON에서 제공하는 Baseline 1은 어텐션이 적용된 Seq2Seq를 적용하여 평가점수를 측정하고, Baseline 2는 LightGBM을 적용하여 평가점수를 측정한다. 트리 구조로 학습을 진행하여 모델을 생성하는 Boosting 방식의 알고리즘을 적용하여 Baseline에서 적용되는 시계열 학습 알고리즘뿐만 아니라 Boosting 방식의 모델들을 비교하고자 한다.

2. XGBoost Learning Model

Gradient Boosting 기반 XGBoost, LightGBM CatBoost 중 Baseline 2에서 사용된 LGBM을 제외한 XGBoost와 CatBoost에 관한 성능 실험을 진행하며, Baseline 2의 LightGBM 모델 대신 XGBoost를 적용해 실험을 진행한다. Baseline 2는 품목, 품종별 개별 모델을 학습시키는 방식을 사용하기 때문에 분류 모델인 XGBoost Classification 대신 XGboost Regression 회귀 모델을 사용하기 위해 XGBoost의 Scikit-Learn Wrapper API인 XGBRegressor를 사용한다. 모델의 학습 데이터 전처리는 Baseline 2의 전처리 과정을 사용하여 개별 모델의 학습을 진행하고, 학습의 지표가 되는 evaluation metric은 Baseline 2의 custom metric을 사용하여 XGBoost의 학습을 진행한다. 학습 데이터는 train.csv 를 사용한다.

XGBoost의 하이퍼 파라미터는 XGBoost의 공식문서를 참조하여 튜닝을 진행한다[13]. XGBoost의 트리 구조 알고리즘 hist, 과적합을 방지하기 위한 스텝 사이즈 eta, 무작위로 트리 학습 중 샘플링 정도를 정하는 subsample과 colsample_by_tree, 부스팅 반복 정도인 num_boost_ro

und는 같게 실험을 진행한다. 그리고 트리 생성 개수인 n_estimators와 결정 트리의 깊이인 max_depth를 조정해 나가며 3번의 실험을 진행한다. 또한, 더 이상의 성능 향상이 되지 않으면 학습을 종료하는 early_stopping_rounds도 조정한다. submission_score는 예측이 끝난 submission.csv를 DACON에 제출하여 얻은 예측 결괏값이다. 3개 모델의 학습을 진행한 후 submission_score를 산정하여 모델의 성능을 검증한다.

Table 3. XGBoost Regressor Hyper Parameter

Hyper Parameter	1st	2nd	3rd
tree_method	hist	hist	hist
n_estimators	1000	2000	500
max_depth	7	5	10
eta	0.1	0.1	0.1
subsample	0.7	0.7	0.7
colsample_bytree	0.7	0.7	0.7
num_boost_round	10	10	10
early_stopping_rounds	none	200	100
submission_score	0.21936	0.23674	0.21311

표 3과 같이 2차 학습 모델은 1차 학습 모델보다 트리의 개수는 늘리고, 트리의 깊이는 줄인 결과, 1차 학습 모델보다 낮은 score를 확인할 수 있다. 이를 참조하여 트리의 개수는 줄이고 트리의 깊이를 올린 결과로 3차 학습 모델에서 가장 좋은 score를 확인할 수 있다.

XGBoost는 결정 트리를 생성할 때 CART(Classification And Regression Trees)라 불리는 앙상블 모델을 사용한다. 이후 Tree 부스팅을 사용하여, Leaf 노드 하나의 결정 값을 가지는 결정 트리와 달리 CART는 모든 Leaf 노드가 최종 스코어와 연관되어 있다.

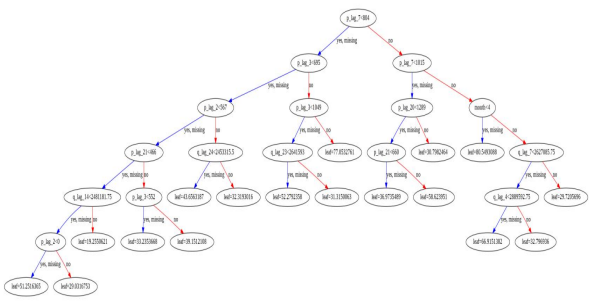


Fig. 4. XGBoost Onion Week-1 Model Tree

그림 4와 같이 XGBoost의 결정 트리는 전처리 과정에서 개별 모델링을 진행하기 위해 넣어준 x일전 가격인 p_lag_x, x일전 거래량인 q_lag_x의 값에 따라 Yes, No의 트리를 반복하여 마지막 Leaf 노드에 도달하여 값을 추정하게 된다.

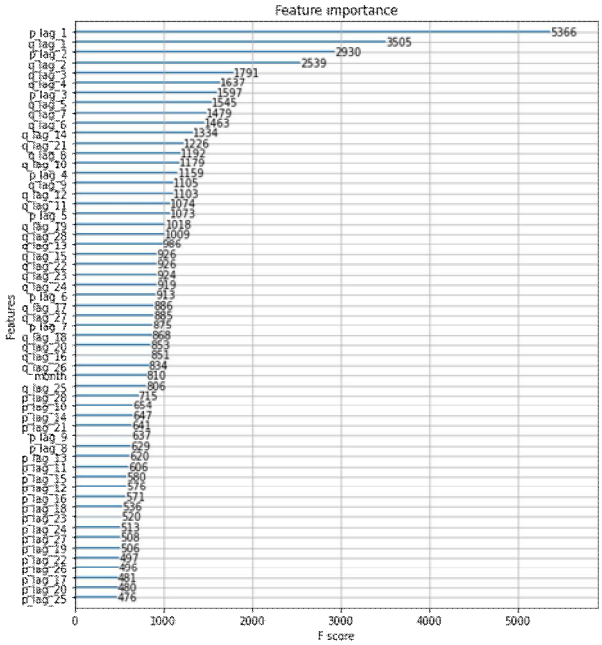


Fig. 5. XGBoost Onion Week-1 Model Feature Importance

XGBoost의 Plot_Importance를 사용하여 학습 시 결정 트리에 영향을 미치는 정도를 알아본 결과, 양파의 1주 후 가격 모델 결정 트리인 그림 5에서는 p_lag_1과 q_lag_1의 값이 결정 트리를 형성하는데 가장 큰 중요도를 띄고 있고 하위 5개인 p_lag_22, 26, 17, 20, 25는 결정 트리 형성 시에 중요도가 가장 낮다.

3. CatBoost Learning Model

XGBoost를 학습시킨 동일한 환경에서 실험을 진행하였다. Baseline 2에서 사용된 eval_metric으로 사용된 at_name을 사용하지 않고 CatBoost Regressor의 eval_set만을 사용하여 학습을 진행하였다. eval_metric으로는 추정값 또는 모델이 예측한 값과 실제 환경에서 관찰되는 값의 차이를 다룰 때 사용되는 RMSE(Root Mean Square Error)를 사용하였다. XGBoost와 동일하게 n_estimators와 max_depth의 수치를 변경하며 3번의 실험을 진행하였으며, gradient step을 줄이는데 사용하는 learning_rate를 조절한다. CatBoost의 공식문서를 참조하여 튜닝을 진행한다[14]. 이전 XGBoost 실험과 동일하게 submission_score를 산정한다.

Table 4. CatBoost Regressor Hyper Parameter

Hyper Parameter	1st	2nd	3rd
n_estimators	1000	500	100
max_depth	6	10	5
eval_metric	RMSE	RMSE	RMSE
subsample	0.8	0.8	0.8
learning_rate	0.05	0.1	0.1
submission_score	0.22994	0.24141	0.23432

표 4를 살펴보면 1차 실험 결과의 submission_score는 0.22994로 XGBoost의 1차 실험 결과보다 높다. 2차 실험에서는 n_estimators를 500으로 줄이고 max_depth를 10으로 설정한 후, learning_rate를 0.1로 설정한다. 그 결과 1차 실험 결과보다 높은 submission_score를 보였다. 3차 실험에서는 n_estimators를 100까지 줄이고 max_depth를 실험 중 가장 낮은 수치인 5로 줄인 상태로 실험을 진행한다. 그 결과, 1차 실험보다는 높지만 2차 실험보다는 낮은 submission_score를 도출하였다.

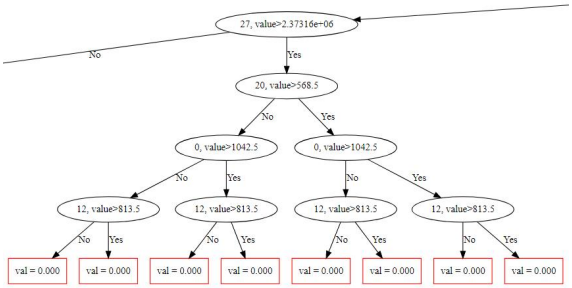


Fig. 6. CatBoost Onion Week_1 Model Fragment Tree

또한, 양파의 1주 후 가격 예측 모델 결정 트리는 그림 6에서 보는 바와 같이 트리 대칭 분할 방식을 사용한다. CatBoost는 결정 트리 생성 시 좌우대칭 트리로 학습 모델을 생성한다. 본 논문에서 사용된 데이터셋은 CatBoost가 사용되는 범주형 Feature 데이터셋이 아닌 수치형 데이터셋을 사용하여 CatBoost가 가지는 장점을 활용한 성능평가를 하지 못하여 CatBoost Regressor를 활용한다.

IV. The Performance Evaluation of Agricultural Product Price Prediction Model

1. Experiment environment

본 논문의 실험 환경은 표 5와 같고, 사용한 파이썬 라이브러리 버전은 표 6과 같다.

Table 5. System Environment

Item	Value
Virtual Environment	Google Colabratory
Memory	12.69GB
Runtime	Python 3 Google Compute Engine

Table 6. Python Library Version

Module	Version
python	3.7.12
xgboost	0.90
CatBoost	1.0.3
numpy	1.19.5
pandas	1.1.5
matplotlib	3.2.2

2. Performance analysis

DACON에서 산정한 Submission score는 그림 7과 같다. Submission score는 오차를 점수로 나타낸 지표이며 낮을수록 성능이 좋다.

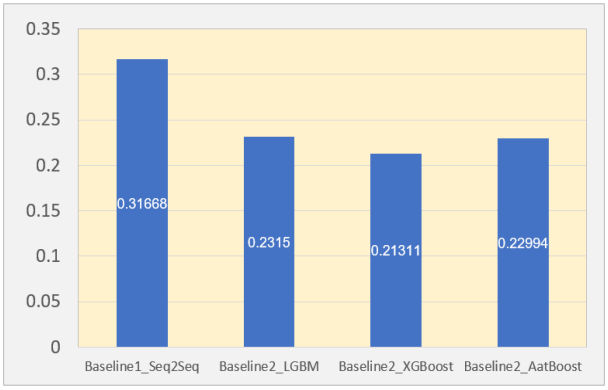


Fig. 7. Submission score of prediction model

그림 7을 살펴보면 Baseline 1의 Attention이 적용된 Seq2Seq submssion score는 0.31668이며, LGBM을 적용한 Baseline2_LGBM의 submssion_score는 0.23150이다. XGBoost를 적용한 모델 중 가장 성능이 좋은 Baseline2_XGBoost의 submssion score는 0.21311이며, CatBoost 적용 모델 중 가장 성능이 좋은 Baseline2_CatBoost의 submssion score는 0.22994이다. 이 결과로 미루어 보아 Kaggle ML competition에서 상위권을 차지하는 GBM 계열 알고리즘들의 성능이 우수한 것을 확인할 수 있다.

LGBM은 leaf-wise 방식의 트리 분할을 채택하여 소요시간 및 메모리를 감소시킬 수 있기 때문에 XGBoost보다 더

빠른 학습시간을 가진다. 하지만 실험에 사용된 데이터셋의 양이 많지 않아 과적합 문제를 피하지 못해 XGBoost보다 낮은 점수 결과를 받게 되었다. LGBM은 학습, 예측시간에서 강한 장점을 보여주기 때문에 많은 양의 데이터셋을 통한 농산물 가격 예측 모델링에 적합하다는 것을 알 수 있다.

XGBoost는 자체적인 과적합 규제를 진행하기 때문에 적은 데이터셋에도 불구하고 우수한 성능을 도출하지만, 학습시간, 예측시간 등 시간적인 성능 면에서는 LGBM보다 성능이 낮다는 것을 알 수 있었다. 또한, 가장 먼저 출시되어 모델의 정보나 하이퍼 파라미터 튜닝 등에 관한 다양한 정보들을 접할 수 있어, 처음 머신러닝 모델링을 한다면 XGBoost를 사용할 것을 추천한다.

CatBoost는 범주형 변수처리에 특화된 모델로 실험에 사용된 Baseline 2의 전처리 방식이 아닌 전체 모델에 따른 분류를 통한 방식의 모델 학습이 이뤄지면 회귀 방식의 모델링보다 좋은 성능을 보여줄 것으로 기대한다.

V. Conclusions

본 논문에서는 DACON의 2021 농산물 가격 예측 경진대회에서 제공하는 데이터셋을 기반으로 Baseline의 Attention이 적용된 Seq2Seq, LightGBM과 기존의 예측 모델의 성능을 높인 XGBoost, CatBoost를 적용한 농산물 가격 예측 모델을 제안하였다. 본 논문에서 사용한 LGBM, XGBoost, CatBoost 모두 트리 방식의 학습 알고리즘을 사용함으로써 큰 성능의 차이는 보여주지 못하였지만, 하이퍼 파라미터의 설정을 통해 같은 데이터셋이지만 예측 성능의 차이가 생길 수 있다는 결론을 도출하였다. 이러한 결과를 통하여 농산물 가격 예측 모델을 학습시키기 위한 데이터셋의 양과 데이터셋의 형식 등에 따라 모델 선정 기준이 달라진다. 또한 학습을 시키기 위한 전처리 과정에서 각 feature별 모델링인지, 하나의 모델에 관한 모델링인지에 따라 어떤 알고리즘을 선택해야 하는지 고려해야 한다. 본 논문의 실험은 GPU의 사용 없이 Colab의 CPU만 사용하였기에 학습 및 예측에 오랜 시간이 소요되어 다양한 결과를 도출하지 못하였다. 향후 실험에서는 GPU의 사용과 좋은 성능의 리소스를 활용한 추가적인 실험으로 성능을 비교하고자 한다.

REFERENCES

[1] http://www.index.go.kr/potal/main/EachDtlPageDetail.do?idx_cd=1628

[2] https://www.mafra.go.kr/mafra/1336/subview.do

[3] M. Fafchamps and B. Minten, “Impact of SMS-Based Agricultural Information on Indian Farmers,” In the World Bank Economic Review, Vol. 26, Issue.3, pp. 383-414, Nov. 2012 https://doi.org/10.1093/wber/lhr056

[4] https://biz.newdaily.co.kr/site/data/html/2021/09/05/2021090500031.html.

[5] https://www.kaggle.com/competitions

[6] “Topic 10. Gradient Boosting,” Kaggle, last modified Jul 01, 2020, accessed Oct. 09, 2021, www.kaggle.com/kashnitsky/topic-10-gradient-boosting

[7] Jae Byung Lee, "A Study on Recent Boosting Methods," Thesis, Konkuk University, 2020

[8] Sutskever, I., Vinyals, O., & Le, Q. V., “Sequence to Sequence Learning with Neural Networks,” NIPS. Dec. 2014, https://arxiv.org/abs/1409.3215

[9] Friedman, J. H. (2002). “Stochastic gradient boosting,” Computational statistics & data analysis, Vol. 38, No. 4, pp. 367-378, Feb. 2002 https://doi.org/10.1016/S0167-9473(01)00065-2

[10] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T., “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)

[11] Dorogush, A.V., Gulin, A., Gusev, G., Kazeev, N., Ostroumova, L., & Vorobev, A., “Fighting biases with dynamic boosting,” June 2017, https://arxiv.org/pdf/1706.09516v1.pdf

[12] Chen, T., & Guestrin, C., “XGBoost: A Scalable Tree Boosting System,” Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, June 2016, https://arxiv.org/abs/1603.02754

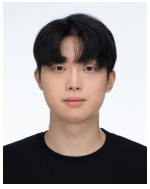
[13] “XGBoost Parameters,” https://xgboost.readthedocs.io/en/stable/parameter.html

[14] “CatBoost Training Parameters,” https://CatBoost.ai/en/docs/references/training-parameters/

Authors

Jung-Ju Im received B.S. degree in 2022 from the Department of Computer Science Inha Technical College, Incheon Korea. He is currently a student in Department of Applied Artificial Intelligence, Hanyang Graduate School. His research interests include Mobile Computing, Image Recognition and Maching Learning Network.

Tae-Wan Kim received B.S. degree in 2022 from the Department of Computer Science Inha Technical College, Incheon Korea. He is currently a student in Department of Applied Artificial Intelligence, Hanyang Graduate School. His research interests include Machine Learning, Health Care and CNN.



Ji-Seoup Lim received B.S. degree in 2022 from the Department of Computer Science Inha Technical College, Incheon Korea. He is currently a student in Department of Applied Artificial Intelligence, Hanyang Graduate School. His research interests include Operating System, Real-time System and Machine Learning.

Jun-Ho Kim currently a student in Inha Technical College. He will have received the B.S. degree from the Department of Computer Science and Engineering Inha Technical Colleage, Incheon Korea in 2022. His research interests include Web Programming, Data Analysis and Machine Learning.

Tae-Yong Yoo received B.S. degree in 2021 from the Department of Computer Science Inha Technical College, Incheon Korea. He is currently a student in Inha Technical College. He will have received the B.S degree from the Department of Computer Science Inha Technical College, Incheon Korea in 2022. His research interests include Machine Learing.

Won Joo Lee received the B.S., M.S. and Ph.D. degrees in Computer Science and Engineering from Hanyang University, Korea, in 1989, 1991 and 2004, respectively. Dr. Lee joined the faculty of the Department of Computer Science at Inha Technical College, Incheon, Korea, in 2008, where he has served as the Director of the Department of Computer Science. He is currently a Professor in the Department of Computer Science, Inha Technical College. He has also served as the Vice-president of The Korean Society of Computer Information. He is interested in parallel computing, internet and mobile computing, and cloud computing, data science, artificial intelligence.