트리 기반 인공지능 모델을 활용한 원자재 가격 예측 자동화 파이프라인 구축

이은지 [1], 유헌창 [2]
[1] 고려대학교 SW·AI 융합대학원 석사과정
[2] 고려대학교 정보대학 컴퓨터학과 교수

2023511038@korea.ac.kr, yuhc@korea.ac.kr

Automated Pipeline for Raw Materials Price Forecast Using Tree-Based Artificial Intelligence

Eun-Ji Lee[1], Heon-Chang Yu[2]
[1]Dept. of Applied Artificial Intelligence, Graduate School of SW·AI Convergence, Korea University
[2]Dept. of Computer Science & Engineering, Korea University

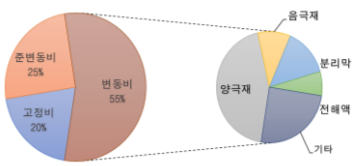
요 약

본 연구는 이차전지 산업의 원가 경쟁력 확보를 위한 원자재 가격 예측 자동화 파이프라인을 제안한다. XGBoost 모델을 기반으로 한 이 파이프라인은 데이터 전처리, 모델 학습 및 예측, 결과 시각화까지의 전 과정을 자동화하여 사용자의 개입을 최소화하고 실시간 의사결정을 지원한다. 본 연구에서는 단변량 및 다변량 예측을 동시에 수행하였으며, 이를 위해 니켈 원자재 가격 데이터와 유가, CPI, PPI, CRB 등 9 개 변수를 활용하였다. 획득한 데이터를 개발한 파이프라인에 적용하여 차분 횟수, 최적 변수 조합, 하이퍼파라미터 결정을 자동 수행하였고, 최종적으로 평균 절대 백분율 오차(MAPE) 0.2%, 결정계수(R^2) 0.99 라는 높은 성능을 달성했다.

1. 서론

유럽연합의 2035 년 내연기관차 완전 판매 금지와 같은 규제는 전 세계적으로 전기차 생산에 관한 관심을 가속하고 있으며, 대한민국도 이러한 글로벌 추세에 발맞추어 이차전지 시장에서의 경쟁력 강화를 위해 지속적인 노력을 기울이고 있다.

이차전지 산업은 다른 기간 산업과 달리 원가의 80% 이상이 변동비로 구성되어 있다는 특징을 가지고 있다. 특히, 이차전지의 원가 구조에서 양극재는 전체 변동비의 약 50%를 차지하며, 그중에서도 니켈은 가장 중요한 원자재로 간주한다. 이러한 산업 특성으로 인해 원자재 가격의 변동은 이차전지 산업의 경쟁력에 직접적인 영향을 미치게 된다.



(그림 1) 리튬 이온 전지 총원가 기준 개념별 비용

본 연구의 목적은 대한민국 이차전지 산업의 원가 경쟁력 확보를 위한 인공지능 모델 기반의 원자재 가격 예측 자동화 파이프라인을 구축하는 것이다. 이차전지 원자재 중 가장 중요한 것으로 꼽히는 니켈을 중심으로 연구하여 원자재 가격의 정확한 예측, 예측 모델의 성능 최적화 및 자동화, 그리고 실시간 예측 결과 제공을 통한 산업 현장의 의사결정 지원에 초점을 맞추고 있다.

2. 관련 연구 및 이론

2.1 시계열 예측 관련 이전 연구

시계열 예측 모델은 시간에 따라 변화하는 데이터를 바탕으로 미래 시점의 값을 예측하는 데 활용된다. 기존의 가격 예측 관련 연구들은 전통적인 통계적 방법인 ARIMA 모형과[1] 기계학습 기반의 LSTM 모델[2]이 주로 사용되고 있다. 과거 관측 값과의 오차를 통해 미래 값을 예측하는 ARIMA 모형은 시계열 데이터의 선형적 관계를 모델링하는 데 유용하지만, 비선형성을 효과적으로 처리하는 데 한계를 가진다. 또

한 LSTM 모델은 장기 의존성을 학습할 수 있는 능력이 뛰어나 복잡한 시계열 패턴 예측에 활용되지만, 과적합의 위험과 해석의 어려움이 존재한다. 이러한 기존 연구의 한계는 더 나은 예측 성능과 해석 가능성을 위한 새로운 접근법의 필요성을 제기하고 있다.

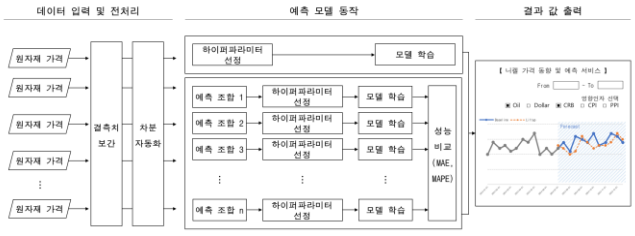
2.2 트리 기반 지도 학습 모델

본 연구에서는 XGBoost 를 학습 모델로 채택하였다. XGBoost 는 CART(Classification And Regression Tree)를 기반으로 하는 의사결정나무 앙상블 모델로, 이전 모델의 오차를 순차적으로 개선하는 부스팅 방식을 적용하여 높은 예측 성능을 발휘한다[3]. XGBoost 는 시계열 데이터의 특성을 효과적으로 반영할 수 있는 능력이 뛰어나며, 시간에 따른 패턴과 변동성을 잘 포착하는 데 강점을 가진다. 또한, 특성 중요도를 통해 모델의 예측 결과를 해석할 수 있어 실무적 활용도가 높다. 이러한 이유로 본 연구에서는 시장 상황에 따라 변동성이 큰 원자재 가격 예측에 XGBoost 가 매우 적합하다고 판단하였다.

3. 원자재 가격 예측 자동화 파이프라인 구축

3.1 파이프라인 구조

본 연구에서 제안하는 전체 파이프라인은 (그림 2)와 같이 세 가지 주요 단계로 구성된다.



(그림 2) 제안하는 원자재 가격 예측 파이프라인

첫 번째, 데이터 입력 및 전처리 단계는 예측에 사용될 데이터를 전처리하는 단계로 결측치 보간과 차분 작업을 포함한다. 이때 결측치는 타겟 데이터인 원자재 가격의 시계열에 맞춰 보간되며, 차분은 자기상관 분석을 반복적으로 수행하며 자동으로 결정된다.

두 번째, 예측 모델 동작 단계에서는 예측 수행 단계가 포함된다. 원자재 가격의 단변량 데이터를 학습시켜 모델의 기본 결과 값으로 제공하며, 동시에 여러 영향 인자를 바탕으로 최적의 예측 조합을 찾아 다변량 모델을 구축하여 단변량 모델의 설명력을 보강한다.

마지막으로 결과 값 출력 단계는 예측 모델의 결과 값을 바탕으로 현장에서의 활용에 중점을 두고 설계된 사용자 중심의 실시간 대시보드이다. 이 대시보드는 실시간으로 예측 결과를 시각화하여 제공하며, 사용자가 영향 인자의 조합을 자유롭게 변경할 수 있도록 하여 더욱 효과적인 의사결정 과정을 지원한다.

이와 같은 세 단계의 파이프라인 구축을 통해 본 연구는 원자재 가격 예측의 실용성을 극대화하고, 사용자에게 실질적인 가치를 제공하는 것을 목표로 한다.

3.2 데이터 입력 및 전처리

3.2.1 데이터 출처 및 구성

원자재 가격은 원유, 곡물 등 다양한 요소에 영향을 받는다. 특히, 코로나 19 확산 이후 원자재 가격의 상승에는 글로벌 유동성 확대, 경기 회복, 공급 병목 현상 등 복합적인 요인들이 작용하고 있는 것으로 추정된다[4]. 이러한 복잡한 요인을 효과적으로 예측하기 위해서는 적절한 입력 변수를 선택하는 것이 매우 중요하다. 따라서 본 연구에서는 니켈 원자재 가격 데이터와 함께 원자재 가격에 영향력을 미치는 것으로 알려진 8 개의 변수를 <표 1>과 같이 수집하였다.

<표 1> 입력 변수 목록

변수 구분	출처	설명
니켈 원자재 가격	네이버 증권 크롤링	일 단위 니켈 원자재 종가
Dubai 유가 가격	네이버 증권 크롤링	일 단위 국제 유가 종가
Brent 유가 가격	네이버 증권 크롤링	일 단위 국제 유가 종가
WTI 유가 가격	네이버 증권 크롤링	일 단위 국제 유가 종가
USD/KRW	네이버 증권 크롤링	일단위 환율(USD/KRW) 종가
금 가격	네이버 증권 크롤링	일 단위 금 가격 종가
CPI 지수	Quant API	월 단위
PPI 지수	Quant API	월 단위
CRB (Commodity Research Bureau) 지수	Stooq Excel 다운	국제 원자재 및 선물 조사회사 CRB 社 지표, 19개 원자재 선물 가격 평균, 일 단위

3.2.2 전처리

(1) 결측치 보간

결측 값이 있는 데이터를 제거하고 분석을 진행할 경우, 모델이 편향적으로 학습될 위험이 있어 일반화된 모델 설계에 어려움이 발생할 수 있다. 따라서 본 연구에서는 결측치를 보간하는 방법을 선택하였다.

2007 년 04 월 05 일부터 2024 년 08 월 02 일까지 일 단위로 수집된 니켈 원자재 가격 데이터는 휴일 및 공휴일에 대한 정보가 결여되어 있다. 이에 따라 (그림 3)과 같이 날짜를 기준으로 데이터 보간을 수행하였다.

(그림 3) 일자 기준 데이터 보간 수행 방법

니켈 원자재 가격 데이터의 영향 인자로 수집한 8 개의 변수에 또한 니켈 원자재 가격 데이터에 맞추어 보간을 실시하였으며, 특히 CPI 와 PPI 의 경우 월 단위 데이터로 수집되어 일 단위로 변환하기 위해 추가적인 보간이 필요하였다.

(2) 자기상관 분석 및 차분 자동화

시계열 데이터 예측에서는 데이터의 정상성을 확인하고, 비정상적인 데이터를 정상 시계열로 변환하는 과정이 필수적이다. 본 연구에서는 Algorithm 1 과 같이 주어진 입력 변수의 정상성을 검증하고, 차분 횟수를 스스로 결정하는 자동화 모듈을 구축하였다.

Algorithm 1 checkTimeSeriesStationarity

```
for each column in dataframe.columns do
  if dataframe[column].dtype is float or int then
    current_value ← dataframe[column].copy
    diffrence_count ← 0
    while diffrence_count < max_diffrence_count do
      adf, p_value ← adfuller(current_value)
      if p_value > 0.05 then
        current_value ← current_value.diff().dropna()
        diffrence_count ← diffrence_count + 1
      else
        break
      end if
    end while
  end if
end for
```

자동화 모듈을 통해 결정된 차분 차수와 차분 결과는 <표 2>에 제시되어 있으며, P-value 값이 0 에 가까워짐으로써 데이터가 정상 시계열로 변환되었음을 확인할 수 있다.

<표 2> 변수 별 차분 횟수와 ADF 및 P-value 값

변수 구분	차분 횟수	차분 전		차분 후	
		ADF	P-Value	ADF	P-Value
니켈 원자재 가격	0	-5.01	0.00002	-5.01	0.00002
Dubai 유가 가격	1	-2.72	0.07126	-10.60	0.00000
Brent 유가 가격	1	-2.15	0.22615	-12.22	0.00000
WTI 유가 가격	1	-2.35	0.15548	-25.19	0.00000
USD/KRW	1	-2.46	0.12562	-15.38	0.00000
금 가격	1	-0.42	0.90718	-17.71	0.00000
CPI 지수	1	-0.33	0.92142	-6.67	0.00000
PPI 지수	1	-1.58	0.49609	-5.27	0.00001
CRB 지수	1	-1.71	0.42627	-25.41	0.00000

3.3 예측 모델 동작

3.3.1 예측 조합 생성 및 최적 조합 선정

본 연구는 다변량 예측 결과만을 제공하는 기존의 시계열 예측 연구들과 달리, 단변량 및 다변량 예측 결과를 동시에 제시함으로써 차별성을 둔다.

단변량 예측은 니켈 원자재 가격 데이터의 단일 시계열 값에 기반하여 수행되었으며, 다변량 예측을 수행하기 위해서는 니켈 원자재 가격 데이터를 포함한 9 개 입력 변수를 기반으로 총 512 개의 조합을 생성하였다. 이는 각 변수의 포함 여부에 따른 모든 가능한 조합을 고려한 것이다. 각 조합에 대해 XGBoost 모델을 학습한 후, 평균 절대 오차(MAE)와 평균 절대 백분율 오차(MAPE)를 계산하여 모델 성능을 평가하였다. <표 3>에서는 상위 3 개 조합과 하위 3 개 조합의 성능 결과를 제시한다.

<표 3> Best/Worst 3 예측 조합 및 성능 결과 값

	변수 조합 구분 (개수)	MAE	MAPE
Best 3	니켈 원자재 가격, PPI 지수 (2)	48.33	0.21
	니켈 원자재 가격, USD/KRW, PPI 지수 (3)	50.32	0.22
	니켈 원자재 가격, Brent 유가 가격, USD/KRW, CPI 지수, PPI 지수 (5)	56.93	0.24
Worst 3	Dubai 유가 가격 (1)	5211.93	31.19
	금 가격 (1)	5222.21	31.17
	WTI 유가 가격 (1)	5276.00	31.93

3.3.2 하이퍼파라미터 결정 자동화

인공지능 모델의 성능은 모델의 하이퍼파라미터에 크게 의존한다. 하이퍼파라미터란 기계학습 모델의 구조나 알고리즘을 구체적으로 결정하는 변수로, 이를 최적화하기 위한 적절한 파라미터 탐색이 필수적이다. 본 연구에서는 그리드서치(Grid Search) 방법을 활용하여 하이퍼파라미터 결정 과정을 자동화하였다.

그리드서치는 하이퍼파라미터의 탐색 범위를 설정한 후, 모든 가능한 조합을 일정한 간격으로 실험하는 방법이다[5]. 이 방법은 시간과 계산량 측면에서 비효율적일 수 있으나, 본 연구에서는 현장에서의 실효성을 중시하여 가능한 많은 조합을 조사함으로써 예측 정확도를 향상하는 데 집중하였다. <표 4>와 같이 그리드서치의 사전 탐색 범위를 설정하였으며, 단변량 모델(니켈 원자재 가격 데이터만 사용) 및 3.3.1 의 연구 결과로 도출한 상위 예측 조합 모델의 하이퍼파라미터 결과는 <표 5>에서 제시된다.

<표 4> 하이퍼파라미터 목록 및 그리드서치 범위

파라미터 구분	탐색 범위	설명
max_depth	[3, 5, 7]	결정트리 모델의 최대 깊이 깊이가 깊어질수록 모델의 복잡도 증가 및 과적합 위험
learning_rate	[0.01, 0.05, 0.1, 0.2, 0.3]	가중치 업데이트 비율 낮은 학습률은 더 많은 반복을 수행하나 과적합 방지 높은 학습률은 더 적은 반복을 수행하나 과적합 위험
min_child_weight	[1, 2, 3, 4, 5, 6]	자식 노드 분할을 위한 최소 샘플 가중치 합 값이 작으면 모델의 복잡도 증가 값이 크면 분할 어려움 증가
n_estimators	[100, 200, 300, 500]	생성할 트리 개수 트리 개수가 많을수록 모델의 성능은 증가하나 과적합 위험
subsample	[0.6, 0.7, 0.8, 0.9, 1.0]	각 트리의 학습 샘플 비율 1.0이면 전체 데이터 Set 사용, 0.5이면 절반의 데이터 Set 사용
colsample_bytree	[0.6, 0.7, 0.8, 0.9, 1.0]	트리 생성시 사용할 변수 비율 1.0이면 모든 변수 사용, 0.5이면 절반의 변수 사용

<표 5> 예측 모델별 결정된 하이퍼파라미터 결과

모델 구분	파라미터					
	max_depth	learning_rate	min_child_weight	n_estimators	subsample	Colsample_bytree
단변량 예측 모델 (니켈 원자재 가격)	7	0.3	5	300	0.6	0.6
다변량 예측 모델 (니켈 원자재 가격, PPI 지수)	7	0.1	6	300	0.6	1.0

4. 자동화 파이프라인 성능 평가

본 연구에서 구축한 자동화 파이프라인은 3.2 의 전처리 과정을 거쳐 3.3.1 에서 도출한 최적의 변수 조합과 3.3.2 에서 결정한 하이퍼파라미터를 적용하여 최종 예측을 수행한다. 예측 결과는 사용자의 즉각적인 이해와 의사결정을 지원하기 위해 (그림 4)와 같이 실시간 시각화 대시보드로 제공된다.

(그림 4) 니켈 원자재 가격 예측 결과 시각화

<표 6>은 자동화 파이프라인을 통해 도출된 예측 모델의 평가 지표를 보여준다.

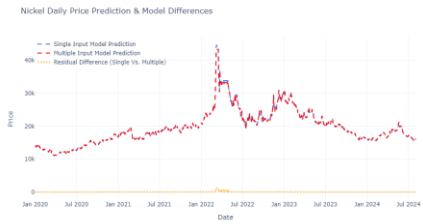
<표 6> 예측 모델별 최종 예측 결과 평가 지표

모델 구분	예측 성능 평가			
	MAE	MAPE	MES	R^2
단변량 예측 모델 (니켈 원자재 가격)	64.54	0.24	81944.75	0.9971
다변량 예측 모델 (니켈 원자재 가격, PPI 지수)	72.89	0.26	134078.60	0.9953

평가 지표 중 특히 주목할 만한 것은 시계열 예측의 정확성을 평가하는 지표인 평균 절대 백분율 오차(MAPE) 값이 0.2 로 높은 성능을 달성하였다는 점이다. 이는 예측 값이 평균적으로 실제 값과 0.2% 정도의 오차만을 보인다는 것을 의미하며, 원자재 가격과 같이 변동성이 큰 시계열 데이터 예측에서는 매우 높은 정확도로 평가할 수 있다.

또한 결정계수(R^2) 값이 0.99 로 1 에 근사한 결과를 보여, 본 모델이 니켈 가격 변동의 대부분을 설명할 수 있음을 통계적으로 입증하고 있다. 이는 본 연구에서 제안한 예측 모델의 신뢰성과 정확성을 강력하게 뒷받침한다.

한편, (그림 5)의 그래프 하단과 같이 단변량 모델(니켈 원자재 가격 데이터만 사용)과 다변량 모델(최적 변수 조합 사용) 간의 성능 차이가 미미한 점은 주목할 만하다.



(그림 5) 모델별 상세 예측 결과 값과 잔차

이는 니켈 원자재 가격 자체의 시계열 패턴이 예측에 있어 매우 중요한 요소임을 시사하며, 동시에 현재 선정된 외부 변수들이 예측 정확도 향상에 제한적인 영향을 미쳤음을 의미한다.

5. 결론 및 고찰

본 연구는 학술적 측면에서 트리 기반 모델의 시계열 예측 능력을 입증하였으며, 실용적 측면에서는 산업 현장에서 즉시 활용할 수 있는 자동화 파이프라인을 제시하였다. 구축된 파이프라인은 평균 절대 백분율 오차(MAPE) 0.2%, 결정계수(R^2) 0.99 라는 높은 성능을 보였다.

이는 변동성이 큰 원자재 가격 예측에서 매우 우수한 결과로, 본 연구에서 제안한 방법론의 유효성을 입증한다. 더 나아가 본 연구에서 제안한 방법론은 니켈 외 다른 원자재 가격 예측, 나아가 주식 시장 동향 예측 등 다양한 경제 지표 예측에도 응용될 수 있는 잠재력을 지니고 있다. 이는 데이터 기반의 의사결정이 있어야 하는 다양한 산업 분야에서 활용될 수 있을 것으로 기대된다.

다만, 단변량 모델과 다변량 모델 간의 성능 차이가 미미한 점은 현재 선정된 외부 변수들의 예측력 향상 기여도가 제한적임을 시사한다. 이에 따라 외부 변수의 영향력 개선과 더욱 다양한 시나리오에 대한 검증 등이 다음 연구 과제로 남아있다. 향후 연구에서는 본 연구의 한계점을 보완하고, 더욱 다양한 상황에서의 모델 검증을 통해 그 적용 범위를 확장해 나갈 계획이다.

참고문헌

[1] Dong-Hwan Kim, Jin Kim, "Price Prediction Analysis in Seoul APT Market Using Time Series Model", Korea Real Estate Society, Vol.71, 193-209, 2024

[2] Su-A Kim, Mi-Ju Kwon, Hyon-Hee Kim, "Sentiment Analysis of News Based on Generative AI and Real Estate Price Prediction: Application of LSTM and VAR Models", The Transactions of the Korea Information Processing Society, Vol.13, No.5, 209-216, 2024

[3] Seon-Ku Lee, Seon-Jong Yoo, "Predicting Real Estate Fractional Investment Prices with the XGBOOST Model", Vol.26, No.1, 1-22, 2024

[4] Chan-Woo Kim, Jeong-Hyeok Lee, "Analysis of Price Impact by Factors of Raw Material Price Variation", Monthly Bulletin of Korea Bank, Vol.76, No.4, 2022

[5] Jae-Eun Lee, Young-Bong Kim, Jong-Nam Kim, "Hyperparameter Optimization for Image Classification in Convolutional Neural Network", The Journal of Korea Institute of Convergence Signal Processing, Vol.21, No.3, 148-153, 2020