

팜유 가격 예측을 위한 딥러닝 기반 단기 시계열 예측 모델링

Deep Learning-Based Short-Term Time Series Forecasting Modeling for Palm Oil Price Prediction

배 성 호 (Sungho Bae) 광주과학기술원 전기전자컴퓨터공학부 학부생
김 명 선 (Myungsun Kim) LGCNS Generative AI 사업팀, 선임연구원
정 우 혁 (Woohyuk Jung) CJ올리브네트웍스 AI연구소 연구원
우 지 환 (Jihwan Woo) CJ올리브네트웍스 AI연구소 연구소장, 고려대학교 기술경영전문대학원 겸임교수, 교신저자

요 약

본 연구에서는 딥러닝 기반의 팜유(Crude Palm Oil: CPO) 가격 예측 방법론을 개발하였다. 팜유는 그 생산 수율과 경제적 효율성으로 인해 다양한 산업에서 중요한 자원으로 활용되고 있으며, 이로 인해 팜유 가격 변동성에 대한 산업계의 관심이 증가하고 있다. 따라서, 팜유 가격 예측을 위한 연구가 활발히 진행되고 있으나, 많은 연구가 시계열 예측 기반으로 정확도에 한계점을 가지고 있다. 본 연구는 기존 방법론의 주요 한계인 정상성 부재 문제를 해결하기 위해 현재 가격 대비 미래 가격의 비율을 종속변수로 사용하는 새로운 모델을 제시한다. 이 접근법은 주식 가격 예측에서의 수익(return) 모델링에 착안하여 개발되었으며, 단순 가격 예측보다 더 높은 성능을 나타낸다. 또한, 다변량 시계열 예측에서 중요한 요소인 독립변수의 지연 값(lag)을 고려하여, 불필요한 잡음을 제거하고 예측 모델의 안정성을 높이는 방법론을 채택했다. 이 연구는 팜유 가격 예측의 정확도를 향상시키는데 중요한 기여를 하며, 시계열 데이터가 중요한 다른 경제적 예측 문제에도 적용 가능한 접근법을 제시한다는 점에서 산업계에 큰 의미가 있다.

키워드 : 팜유 가격 예측, 단기 시계열 예측, 다변량 시계열 분석

I. 서 론

팜유(crude palm oil: CPO)는 기름야자 열매의 과육에서 얻어지는 유지이다. 팜유는 단위 면적당 생산 수율이 다른 석유 작물에 비해 현저히 높기 때문에 그 가격이 저렴하고 생산 효율이 뛰어나 중요한 경제 작물로 여겨진다. 이러한 특성으로

인해 팜유는 생필품, 화장품, 제조업, 요식업 등 여러 분야에서 널리 사용되고 있다. 팜유 생산이 글로벌 경제에 미치는 영향력이 커짐에 따라 팜유 가격 변동에 대한 관심 또한 증가하고 있다. 시장에서의 리스크를 최소화하고 안정적인 재고를 확보하기 위해서는 팜유 가격의 변동을 정확하게 예측하는 것이 중요한 실정이다(Murphy et al., 2021).

팜유 가격 예측의 중요성이 커짐에 따라 많은 연구에서 시계열 예측 방법론을 활용한 팜유 가격 예측이 시도되었다(Karia *et al.*, 2013). 가격 변동성은 랜덤 워크와 유사한 패턴의 양상을 보이기 때문에 매우 도전적인 과제로 여겨진다(Patalay and Bandlamudi, 2020). 대표적인 시계열 예측 방법론으로는 전통적인 통계적 모델링 방법론과 인공지능 기반 모델링 방법론 2가지로 나눌 수 있다(김명선 등, 2023). 통계적 모델링의 경우 시계열의 선형적인 패턴을 모델링하는 데 특화된 반면 인공지능 기반 모델링의 경우 복잡한 비선형적 패턴까지 모델링이 가능하기 때문에, 최근 연구들에서는 팜유 가격 변동에 영향력 있는 여러 요소를 결합한 인공지능 기반 모델링 방법론들이 좋은 성능을 거두었다(Karia *et al.*, 2013; Yoo *et al.*, 2011).

팜유 가격 예측을 위한 기존 방법들의 경우 1달에서 6개월 뒤의 팜유 가격을 활용한 모델링이 주를 이룬다. 하지만 가격 데이터의 특성상 시간에 따른 변동성이 크기 때문에 정상성을 만족하지 않는다. 정상성은 통계학에서 데이터가 정규분포를 따르는 성질을 말한다. 많은 통계적 방법들이 데이터가 정규분포를 따른다는 가정 하에 최적으로 작동하기 때문에, 단순히 가격을 활용한 예측 방법론은 실효성이 크지 않다. 따라서 본 연구에서는 정상 시계열 상태를 충족하기 위해 가격 대신 현재 가격과 미래 가격의 비율(ratio)을 종속변수로 사용한 모델링 방법론을 제시한다. 주식 가격 예측에서 주식의 수익(return)이 예측 모델링시 종속변수로 사용되는 점에 착안하였는데, 수익의 경우 비율과 비교했을 때 편차함에 대한 학습이 추가적으로 필요하기 때문에 비율 값을 종속변수로 채택하였다. 초기 분석에서 원시 가격 데이터가 많은 통계 모델과 방법론의 전제 조건인 정상성을 만족하지 않는 것으로 나타났다. 원시 가격 분포의 비대칭성과 치우침은 이후의 통계적 분석을 위해 정상성 조건을 만족시키기 위한 변환을 필요로 하였다.

이 문제를 해결하기 위해, 가격을 일련의 비율로 변환하는 방법을 사용하였다. 변환된 데이터셋은 원래 가격 데이터에 대한 연속적인 수정을 통해 파생된 “M+1”, “M+2”, “M+3” 비율 분포로 명명되었다(식 (1.1)). 이 방법론적 접근은 가격 절대값 대신 가격 비율이 정규 분포와 더 일관된 특성을 나타낼 것이라는 전제에 기초하고 있었다.

$$R_{M+1} = \frac{V_{M+1}}{V_M}, R_{M+2} = \frac{V_{M+2}}{V_M}, \quad (1.1)$$

$$R_{M+3} = \frac{V_{M+3}}{V_M}$$

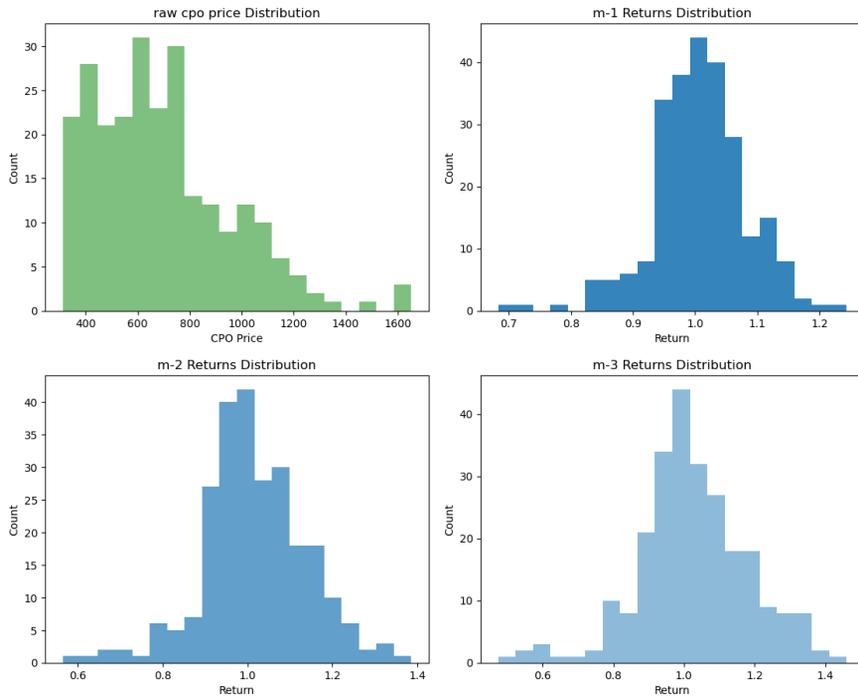
V_T : T step에서의 가격값

M: month

R_T : T step에서의 비율

변환의 효과는 변환된 <그림 1>의 데이터 히스토그램에서 명백하게 나타났으며, 이는 정규분포의 전형적인 종 모양 곡선을 밀접하게 닮았다. 원시 가격 데이터에서는 이러한 정상성이 나타나지 않았음을 보여, 우리의 비율 변환 기법이 분산을 정상화하고 분포를 정규화 하는 데 성공했다는 것을 나타냈다.

본 연구에서는 다변량 시계열 예측을 위한 모델링 과정에서 독립변수의 지연 값(lag)의 중요성을 인식하고 있다. 팜유 가격과 연관성이 낮은 변수나 부적절한 지연 시점의 데이터를 모델에 포함시키면 예측에 잡음(noise)이 발생하여 모델의 정확성이 저하될 위험이 있다. 이에 본 연구는 각 독립변수의 지연 시점별로 중요도를 철저히 분석하고, 오직 중요한 특징들만을 예측 모델에 반영하여 예측의 정확성을 높이는 방법을 채택하였다. 이를 통해 모델은 일관된 결과를 도출할 수 있고, 불필요한 변동성을 최소화하여 모델의 예측 정확도를 극대화하는 데 기여하였다.



〈그림 1〉 시간별 raw 가격과 ratio 가격

II. 문헌연구

팜유 가격 예측을 위해 다양한 시계열예측 방법론들이 제안되었다. 본 연구에서는 이러한 방법론들을 크게 3가지로 분류하여 검토한다.

2.1 지수 평활법을 통한 예측 방법론

지수 평활법은 간단하면서도 효과적인 시계열 예측 방법으로, 최신 정보에 높은 가중치를 부여하고 계산 효율적이기에 실시간 예측 및 의사 결정에 유용하다. 이 방법을 통해 시계열 데이터의 특성에 따라 적응적으로 예측 모델을 조절할 수 있는 장점을 가지고 있어, 관련 연구가 활발히 이루어지고 있다. Siregar *et al.*(2017)에서는 삼중 지수 평활법(Triple Exponential Smoothing)과 이중 이동 평균(Double Moving Average)을 사용하여 미래 팜유 생산량을 예측하였다. Sandika *et al.*(2023)에

서도 마찬가지로 단일 지수 평활법, 이중 지수 평활법, 삼중 지수 평활법 등의 다양한 방법을 사용하여 팜유 생산량을 예측하였다. 두 논문 모두 삼중 지수 평활법 식 (2.4)을 사용한 모델이 가장 우수한 성능을 보였다. Muainuddin(2021)에서는 코로나 팬데믹 기간의 팜유 가격을 예측하였는데 이 경우에는 특이하게도 단일 지수 평활법, 이중 지수 평활법, 삼중 지수 평활법 중 단일 지수 평활법을 사용한 모델이 가장 좋은 결과가 나왔다. 코로나 팬데믹의 특수성을 반영하기 위해서는 최근 데이터에 가중치를 더 주어야 하고 따라서 계절성과 추세성을 반영하지 않는 단일 평활 모델이 다른 모델들에 비해 좋은 성능을 보였다고 말한다.

레벨 평활(Level Smoothing)은 시계열 데이터의 레벨(또는 평균)을 평활화하고, 최신 데이터에 더 큰 중요성을 부여하면서도 과거의 평활화된 추세를 고려한다. 이 과정은 다음 수식에 의해 정의된다.

〈표 1〉 선행 연구 조사

저자	방식	내용
Siregar <i>et al.</i> (2017)	지수 평활법 (단일, 이중, 삼중)	단일 지수 평활법, 이중 지수 평활법, 삼중 지수 평활법을 비교하여 팜유 생산량을 예측하였다. 상수를 $\alpha = 0.6$, $\beta = 0.02$, $\gamma = 0.02$ 로 설정한 삼중 지수 평활법이 RMSE = 0.09로 가장 결과가 좋았다. 계절적 요소를 고려할 때 삼중 지수 평활법이 우수한 예측 결과를 제공한다는 것을 보여준다.
Suppalakpanya <i>et al.</i> (2019)	이중 지수 평활법, 삼중 지수평활법(덧셈, 곱셈), 개선된 삼중 지수 평활법	2005년부터 2017년까지의 팜유 가격 데이터를 기반으로 하여 여러 가지 지수 평활법을 사용하여 팜유 가격을 예측하였다. 기존의 이중 지수 평활법과 삼중 지수 평활법(Additive Holt-Winters: AHW) 보다 더 좋은 성능의 확장된 삼중 지수 평활법(Extended Additive Holt-Winters: EAHW)을 제시하였다.
Muainuddin and Abu(2021)	지수 평활법 (단일, 이중, 삼중)	코로나 팬데믹 기간 말레이시아에서의 팜유 가격을 예측하기 위한 다양한 지수 평활 모델을 적용하였다. 단일, 이중, 삼중 지수 평활 모델을 비교하였을 때, 단일 지수 평활 모델이 MAPE, MAD가 가장 작게 나왔다. 코로나 팬데믹 때의 특수성을 반영하기 위해서는 최근 데이터에 더 가중치를 많이 두는 단일 지수 평활 모델이 가장 적합하다는 결과를 보여준다.
Sandika <i>et al.</i> (2023)	이중 이동 평균, 삼중 지수 평활법	2013년 1월부터 2022년 12월까지의 팜유 데이터를 기반으로 하여 두 가지 예측 방법을 비교 분석했다. 상수를 $\alpha = 0.2$, $\beta = 0$, $\gamma = 0.2$ 로 설정한 삼중 지수 평활법이 MAPE = 9.48%로 결과가 가장 좋았고, 이중 이동 평균 방법은 그에 조금 못 미치는 MAPE = 11.2%의 결과가 나왔다.
Ariff <i>et al.</i> (2015)	자기회귀 누적 이동평균(ARIMA)	2010년부터 2012년까지의 팜유 일일가격을 ARIMA 모델을 사용하여 모델링하고 예측하였다. ARIMA(1, 2, 1) 모델이 AIC, BIC 값이 가장 낮게 나오는 것을 확인하였다.
Khalid <i>et al.</i> (2018)	ARDL, ARIMA, ARIMAX	자기회귀 시차 분포모형(Auto Regressive Distributed Lag: ARDL), ARIMA, 외생변수를 포함한 자기회귀 이동 누적 평균모형(Auto Regressive Integrated Moving Average Model with exogenous inputs model)을 비교하여 팜유 가격을 예측한다. 콩기름 가격, 팜유 재고, 원유 가격, 환율 등을 외생변수로 추가한 ARIMAX(2, 1, 2) 모델이 가장 결과가 좋았다.
Kanchymalay <i>et al.</i> (2017)	다층 퍼셉트론(MLP), 서포트 벡터 회귀(SVR), 삼중 지수 평활법	1987년 1월부터 2017년 2월까지의 월별 팜유 가격 데이터를 사용하여 팜유 미래 가격을 예측하였다. MLP, SVR, 삼중 지수 평활법의 3가지 모델을 4가지 평가지표를 토대로 비교해본 결과, DA는 MLP 모델이 가장 높았고, MAPE, MAE는 SVM 모델에서 높게 나왔다.

$$\ell_t = \alpha(y_t - S_{t-L}) + (1-\alpha)(\ell_{t-1} + b_{t-1}) \quad (2.1)$$

α : 레벨 평활 상수, 최신 관측치에 줄 가중치를 결정

ℓ_t : 시간 t 에서의 예측된 수준

y_t : 시간 t 에서의 실제 관측값

s_t : 시간 t 에서의 계절적 요소

b_t : 시간 t 에서의 추세 요소

L : 계절성 주기의 길이

추세 평활(Trend Smoothing)은 이전 레벨의 변

화를 기반으로 하여 추세를 조정하는 과정이다. 이는 다음과 같이 계산된다.

$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1} \quad (2.2)$$

β : 추세 평활 상수, 추세 변화에 줄 가중치를 결정

계절성(Seasonality) 평활은 계절성 패턴을 적용적으로 갱신하여 시계열 데이터에서의 주기적 변동을 반영한다. 계절성 평활은 다음과 같은 수식

으로 표현된다.

$$s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1 - \gamma)s_{t-1} \quad (2.3)$$

γ : 계절성 평활 상수, 계절성 패턴에 줄 가중치를 결정

여기서 γ 는 계절성 평활 상수로 최근 관측치가 계절성 패턴에 얼마나 영향을 줄 것인지를 결정한다.

삼중 지수 평활법(Triple Exponential Smoothing Additive holt-winters method)은 위에서 설명한 레벨, 추세, 그리고 계절성 평활을 결합하여 미래의 값을 예측한다.

$$\hat{y}_{t+m} = (\ell_t + mb_t) + s_{t-L+1+(m-1)\text{mod}L} \quad (2.4)$$

m : 예측하고자 하는 미래 시점까지의 기간 수

2.2 자기 회귀 모델을 통한 예측 방법론

Box와 Jenkins에 의해 개발된 자기회귀 누적 이동평균(autoregressive integrated moving average: ARIMA)은 자기회귀(Autoregressive: AR), 누적(Integrated: I), 이동평균(moving average: MA)의 세 가지 구성 요소를 결합한 통계 모델이다(Box *et al.*, 1970). ARIMA는 ARIMA(p, d, q)로 표시되며, 여기서 p, d 및 q는 매개 변수이다. 매개 변수 p와 q는 자기회귀 및 이동평균 구성 요소의 순서를 나타낸다. 매개 변수 d는 데이터를 정상 상태로 만들기 위해 필요한 차이의 정도를 나타낸다. 차이화는 일련의 시간에 따른 선형 및 지수적 추세를 제거하는 데 일반적으로 사용된다. 예측을 위해 ARIMA는 과거 기간 값에 상수를 더하여 간접적으로 특정 시간 간격 내의 평균 변화를 추정하는데 도움을 준다. ARIMA 모델에서 예측 변수는 이전 관측치의 선형 조합으로 나타낸다(Jain and Mallick, 2017). Khalid and Hakimah(2018)에서는 ARIMA에 팜유 예측에 도움이 될 만한 외생 변수를 추가한 자기회귀 통합 이동평균(autoregressive

integrated moving average with exogenous variable: ARIMAX)모델을 사용하여 예측 정확도를 더 향상시켰다. 추가로 팜유 현물 가격이 팜유 재고량, 원유(crude oil) 가격 및 대두유(soybean oil) 가격에 높은 영향을 받음을 확인할 수 있다. 아래는 ARIMA 모델과 ARIMAX 모델은 아래와 같이 표현된다.

$$ARIMA(p, d, q) : \Phi(B)(1 - B)^d Y_t = \Theta(B)\epsilon_t \quad (2.5)$$

Y_t : t 에서의 관측 값

B : 백 시프트(back shift) 연산자

$\Phi(B)$: AB 항을 나타내는 다항식

$$\Phi(B) = 1 - \Phi_1 B - \Phi_2 B^2 - \dots - \Phi_p B^p$$

$\Theta(B)$ 항을 나타내는 다항식

$$\Theta(B) = 1 + \Theta_1 B + \Theta_2 B^2 + \dots + \Theta_q B^q$$

ϵ_t : 오차 항

$$ARIMAX(p, d, q) : \Phi(B)(1 - B)^d Y_t = \Theta(B)\epsilon_t + \sum_{i=1}^k \beta_i X_{i,t} \quad (2.6)$$

$$= \Theta(B)\epsilon_t + \sum_{i=1}^k \beta_i X_{i,t}$$

$X_{i,t}$: 외생 변수

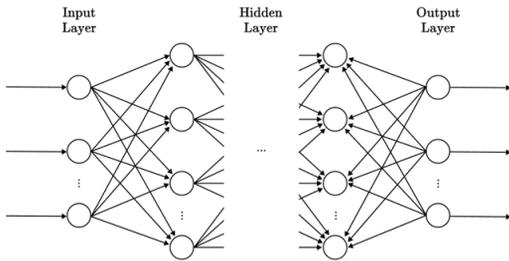
β_i : 각 외생 변수에 대한 계수

k : 외생 변수의 수

2.3 딥러닝 기반 예측 방법론

Kanchymalay(2017)에서는 다층 퍼셉트론(Multi-Layer Perceptron: MLP)과 서포트 벡터 회귀(Support Vector Regression: SVR)와 삼중 지수 평활법(Holt-Winters) 기법을 활용하여 팜유 가격 예측 모델을 만들고 성능을 비교한다. 공통으로 팜유 가격과 관련성이 높다고 알려진 몇 가지 요소(식용유 가격, 원유 가격 및 월별 환율 데이터)를 활용한다. 다변량 시계열 모델링을 수행한다. 제안하는 방법론을 평가하기 위해 평균 제곱근 오차(root mean squared error: RMSE), 평균 절대 오차(mean abso-

lute error: MAE), 평균 절대 백분율 오차(mean absolute percentage error: MAPE) 및 방향 정확도(direction of accuracy: DA)의 4가지 지표를 활용한다. 결과의 성능 비교에 따르면, MLP는 DA에 대해 71.02%의 가장 높은 성능을 보였으며, SVR는 MAE와 MAPE에서 각각 59.15와 7.8073으로 가장 낮은 오차 값을 기록하며 높은 성능을 나타냈다.

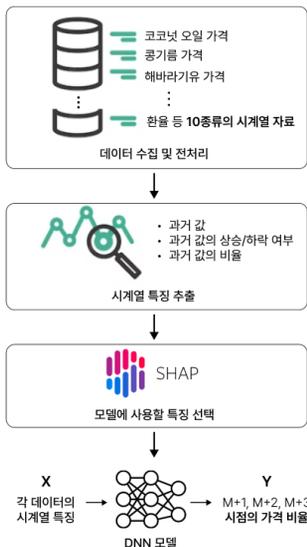


<그림 2> Multi-Layer Perceptron Model 구조

III. 제안 모델

3.1 개략도

제안하는 방법의 개략도는 <그림 3>과 같다.



<그림 3> 개략도

3.2 데이터 수집 및 전처리

본 연구에서는 10개의 시계열 데이터 자료를 사용한다(Kanchymalay *et al.*, 2017). 코코넛 오일(Coconut oil) 가격, 콩기름(Soybean oil) 가격, 해바라기유(Sunflower oil) 가격, 유채 기름(Rapeseed oil) 가격, 올리브유(Olive oil) 가격, 땅콩기름(Peanut oil) 가격의 월별 증가를 Indexmundi¹⁾에서, 원유(Crude oil)가격, 현물 원유(Spot crude oil) 가격, 팜유(Crude palm oil) 가격, 미국 달러와 말레이시아 링깃화의 환율을 FRED Economic Research²⁾에서 찾아 사용한다. 기간은 모든 데이터가 공통적으로 있는 구간인 2002년 2월부터 2023년 5월까지로 정하고, 2002년 2월부터 2019년 12월까지를 학습(Train)구간으로 그 이후를 테스트(Test)구간으로 잡고 실험을 진행한다.

먼저 수집된 자료의 결측치에 대하여 자료의 전처리를 수행한 뒤, 관련 변수들 간의 패턴과 경향성을 파악하기 위하여 통계 분석을 수행한다. 이후 팜유 가격과 다른 식물성 기름 가격들 간의 상관관계를 조사하기 위해 스피어만 상관계수를 사용하였다. 스피어만 상관계수(식 (3.1))는 변수들 간의 관계가 선형적이지 않거나, 데이터가 정규 분포를 따르지 않을 때 유용한 비모수적 방법이다.

$$\rho_{xy} = \frac{\sum_{i=1}^n \left(\text{rank}(X_i) - \frac{n(n+1)}{2} \right) \left(\text{rank}(Y_i) - \frac{n(n+1)}{2} \right)}{\sqrt{\sum_{i=1}^n \left(\text{rank}(X_i) - \frac{n(n+1)}{2} \right)^2} \sqrt{\sum_{i=1}^n \left(\text{rank}(Y_i) - \frac{n(n+1)}{2} \right)^2}} \quad (3.1)$$

X, Y: 사용된 두 변수

$\text{rank}(X_i), \text{rank}(Y_i)$: i번째 데이터 포인트의 순위

n: 데이터 셋에 있는 관측치의 개수

이는 순위 정보를 기반으로 상관관계를 측정하므로, 이상치의 영향을 덜 받고 변수 간의 단조관계를 강조할 수 있다. 피어슨 상관계수가 변수

1) <http://www.indexmundi.com>.

2) <https://fred.stlouisfed.org/>.

들 간의 선형적 관계를 측정하는 데 주로 사용되는 반면, 이 연구에서는 팜유 가격과 다른 기름 가격들 간의 비선형적인 관계나 순위의 변화를 더 잘 반영할 수 있는 스피어만 상관계수를 선택하였다. 분석 초기 단계에서 원본 가격 자료를 바탕으로 각 변수의 상관성을 평가함으로써, 모델이 실질적으로 의미 있는 데이터를 반영하도록 하였다.

스피어만 상관성 검정 결과 팜유 가격과 콩기름 가격의 상관계수 값이 0.96으로 가장 높았고, 유채기름 0.93, 코코넛 기름 0.81, 땅콩기름 0.77 순으로 나타났다. 그러나 올리브유는 -0.17로 팜유에 대해 매우 낮은 상관성을 보였고, 팜유와 단기 관계가 없다고 판단하여 예측에서 제외하였다.

〈표 2〉 데이터 별 스피어만 상관계수

데이터	스피어만 상관계수	p-value
코코넛 오일 가격	0.81	0.00
콩기름 가격	0.96	0.00
해바라기유 가격	0.71	0.00
유채기름 가격	0.93	0.00
올리브유 가격	-0.17	0.01
땅콩기름 가격	0.77	0.00
원유 가격	0.76	0.00
현물 원유 가격	0.73	0.00
미국-말레이시아 환율	-0.24	0.00

〈표 3〉 데이터 별 통계치

데이터	평균	표준편차	25%	50%	75%
팜유 가격	693.45	263.33	485.97	653.48	836.38
코코넛 오일 가격	1,054.53	431.93	704.75	992.83	1,393.40
콩기름 가격	917.69	318.05	720.06	838.40	1,127.07
해바라기유 가격	3,984.15	879.48	3,278.48	3,950.78	4,545.34
유채기름 가격	1,546.06	424.84	1,313.33	1,433.00	1,877.12
올리브유 가격	973.20	326.54	780.98	874.87	1,167.09
땅콩기름 가격	1,122.95	408.49	840.83	1,035.77	1,434.10
원유 가격	66.16	24.56	47.67	63.18	85.76
현물 원유 가격	67.67	26.35	47.03	63.77	89.93
미국-말레이시아 환율	3.74	0.44	3.34	3.80	4.13

3.3 모델링 단계

향후 3달의 팜유 가격이 상승/하락을 예측하는 것이 목표이기에 평가지표 중 평균 방향 정확도 (Directional Accuracy: DA)의 수치가 높은 선행연구의 모델을 사용한다. Kanchymalay(2017)에서 ANN 기반의 다층 퍼셉트론(Multi-Layer Perceptron: MLP)을 사용한 모델이 DA가 가장 좋게 나온 것을 확인했고 월별 CPO 가격 예측 모델을 MLP로 선정하였다.

MLP는 각각 64개, 32개의 은닉 유닛을 가지는 2개의 은닉층으로 구성되며, 옵티마이저로 Adam과 SGD를 교차 사용하였다. 스케줄러로는 MultiStepLR을 적용하였고, 학습률을 0.2, 0.01, 0.005, 0.0001로 변화시키며 실험하였을 때, <표 4>에 제시된 설정인 학습률 0.01, 에포크 3000회, Early Stopping Patience 100회, MultiStepLR 스케줄러(milestones = [50], gamma = 0.2)가 모델의 학습과정에서 loss가 안정적으로 감소하였고, 수렴 속도 측면에서도 가장 우수한 결과를 가져왔다. 이 설정은 다양한 하이퍼파라미터 조합을 테스트한 결과로, 모델이 데이터에 대해 더 잘 일반화하고, 더 높은 정확도를 달성하는 데 기여하였음을 확인할 수 있다. 모델에 사용된 하이퍼 파라미터는 <표 4>와 같다.

모델에서 각 변수들의 공헌도를 확인하기 위해 샤프 지수(SHapley Additive exPlanations: SHAP)를 사용한다. SHAP는 게임 이론의 샤프리 값에 기반하여 각 변수가 모델 예측에 미치는 영향을 정량

적으로 측정한다. 이는 각 변수가 학습에 포함될 때 모델 성능에 미치는 마진 기여도를 평균적으로 측정함으로써 중요도를 부여한다. SHAP 값은 식 (3.2)에 의해 정의된다(Lundberg and Lee, 2017).

〈표 4〉 사용한 하이퍼 파라미터

하이퍼 파라미터	값
에폭(Epoch)	3000
Early Stopping Patience	100
학습률(Learning Rate)	0.01
스케줄러(Scheduler)	MultiStepLR(milestones = [50], gamma = 0.2)
옵티마이저(Optimizer)	Adam

$$SHAP_i(v) = \sum_{S \subseteq N(i)} \frac{|S|!(n-|S|-1)!}{n!} (v(S \cup i) - v(S)) \quad (3.2)$$

v : 특성들의 집합에 대한 가치함수(모델의 예측 결과)

S : 특성 i 를 제외한 모든 특성의 부분집합

SHAP는 XAI(설명 가능한 인공지능)에 필수적인 요소로 복잡한 모델에서 의사결정 과정을 투명하게 만들어준다. 이 공식은 하나의 특성이 주어진 모든 가능한 특성 조합에 얼마나 기여하는지를 평균 내어 계산한다. 이를 통해 해당 특성이 모델 예측에 미치는 영향의 크기와 방향을 정량화할 수 있다. 이 내용을 통해 모델의 각 변수들이 어떻게 최종 예측에 기여하는지를 이해하고, 이를 기반으로 더 나은 특성 선택과 모델의 해석 가능성을 높일 수 있다.

원시 가격 데이터(raw prices)가 비정상 시계열(non-stationary time series)로 나열되어 있으면 모델이 값을 효과적으로 예측하지 못한다. 이 문제를 해결하기 위해 원시 값 대신 가격의 상승/하락 여부(up/down value)와 비율(ratio)을 종속 변수로 사용하여 시계열을 정상 시계열(stationary time

series)로 변환한다. 이 방법은 시계열 데이터를 분석할 때 일반적으로 사용되며, $\frac{V_{T+1}}{V_T}$ 형태를 비율을 적용함으로써 등분산(homoscedasticity) 특성을 갖는 정상화된 데이터로 형식이 변하게 된다. 상승/하락 여부는 식 (3.3)과 같고, 비율은 식 (3.4)와 같다.

$$U_{T+1} = 1 \text{ for } V_{T+1} > V_T \text{ else } 0 \quad (3.3)$$

U_T : T step에서의 상승, 하락 여부

V_T : T step에서의 가격값

$$R_{T+1} = \frac{V_{T+1}}{V_T} \quad (3.4)$$

R_T : T step에서의 비율

비율을 사용하는 것은 차분을 적용하는 원리와 유사하며, 이는 특히 시계열 데이터가 비정상적인 분포를 보일 때 효과적이다. 차분은 원 데이터의 절대적인 값 대신 시간 간의 차이를 분석함으로써 데이터의 정상성을 개선하지만, 이 연구에서 적용된 비율 방식은 가격 데이터의 비율 변화를 분석함으로써 유사한 효과를 낸다. 비율 변환은 데이터의 스케일을 줄이고, 추세와 계절성을 제거하여 등분산성을 획득하는 데 도움을 준다. 추가적으로, 이 방식은 이상치의 영향을 줄이고, 모델이 더 정상적인 분산을 가진 데이터에 기반하여 학습하도록 함으로써, 더 나은 예측 성능을 달성할 수 있게 한다. 이러한 이유로 본 연구에서 비율 데이터는 등분산성 확보를 위한 타당한 전략으로 활용되었다.

IV. 실험 및 결과

본 연구에서는 다변량 시계열 분석을 통해 пам유 가격 예측의 정확성을 향상시키기 위한 방법론을

모색하였다. 기존의 연구를 기준 삼아 변수를 변경해가며 팜유 가격 예측 모델의 성능을 비교 분석하였다. 각 변수와 지연 값이 모델 예측력에 미치는 영향을 객관적으로 평가할 수 있게끔 모든 실험에 사용된 모델의 하이퍼 파라미터는 동일한 설정을 유지하였다. 변수 별 모델 학습 결과는 <표 5>와 같다. 초기 모델 구축 시 Kanchymalay et al. (2017)을 기준으로 삼아, 가격 데이터에 어떠한 변환도 적용하지 않은 원본 가격 값을 독립변수와 종속변수로 사용하여 성능을 평가하였다. 이후에 상승/하락 여부를 독립변수만 변경하였을 때는 이 데이터에서 유의미한 패턴을 학습하지 못하는 것으로 확인되었다. 이에 대한 해결책으로, 가격 데이터를 비율로 변환하여 종속변수의 정상화를 시도하였고, 이는 DA 및 평균 절대 백분율 오차(MAPE)에서 눈에 띄는 성능 향상을 보여주었다.

추가적으로 독립변수도 비율로 변경하여 모델에 제공함으로써, 예측 모델의 정밀도를 향상시켜 보다 정교한 예측이 가능하게 하였다. 이러한 변환을 통해 모델이 상세한 가격 변동을 반영할 수 있게 되었으며, 평균 절대 백분율 오차(MAPE)의 감소를 통해 예측값과 실제 가격 사이의 오차가 줄어들었음을 확인할 수 있었다.

특히, DA 지표에서는 상승/하락 값을 변수로 사용했을 때, MAPE 지표에서는 비율 값을 변수로 사용했을 때 모델의 성능이 눈에 띄게 개선되었다. 이는 상승/하락 값이 미래 가격의 방향성을 예측하는 데 효과적이지만, 비율 값이 예측의 정확성을 높이는 데 기여한다는 것을 시사한다. 이러한 발견을 바탕으로, 두 변수의 장점을 결합하여 새로운 모델을 구성하였다.

상승/하락 데이터와 비율 데이터를 각각 독립

<표 5> 독립 변수, 종속 변수 별 결과

독립 변수, 종속 변수 (X, y)	DA M+1 (%)	DA M+2 (%)	DA M+3 (%)	MAPE M+1 (%)	MAPE M+2 (%)	MAPE M+3 (%)
가격, 가격(기준)	71.79	64.10	56.41	19.90	27.86	35.82
상승/하락, 가격	30.77	38.46	43.59	16.81	24.25	32.33
상승/하락, 비율	94.87	82.05	74.36	5.49	10.11	15.69
비율, 비율	84.62	74.36	74.36	4.42	9.72	15.29
상승/하락+비율, 비율	97.44	79.49	74.36	4.89	9.96	15.33

<표 6> 미 선택 데이터 별 결과

	DA M+1 (%)	DA M+2(%) 미 선택 데이터	DA M+3 (%)	MAPE M+1 (%)	MAPE M+2 (%)	MAPE M+3 (%)
X	87.18	82.05	71.79	5.47	10.43	15.66
팜유 가격	66.66	64.10	58.97	6.48	11.70	16.41
코코넛 오일 가격	92.31	79.49	71.79	5.47	9.73	15.16
콩기름 가격	92.31	82.05	79.49	5.60	10.22	15.46
땅콩기름 가격	89.74	82.05	69.23	5.43	10.25	15.84
해바라기유 가격	94.87	79.49	74.36	5.02	10.02	15.68
유채유 가격	89.74	82.05	69.23	5.41	10.02	15.51
현물 원유 가격	94.87	79.49	74.36	5.02	10.06	15.60
원유 가격	92.31	82.05	74.36	5.38	10.08	15.49
환율	92.31	82.05	74.36	5.43	10.22	15.46

변수로 설정한 두 모델에 대해 지연 값의 최적화 실험을 수행하였다. 지연 값의 범위를 1에서 7까지 조절해가며 각 모델의 예측 성능에 미치는 영향을 평가하였고, 각 데이터에 대한 최적의 지연 값을 결정하였다.

두 모델에 사용한 변수들을 합치고, 최적의 지연값을 설정해준 뒤에 SHAP값과 특징 선택을 통해 모델의 성능을 저하시키지 않는 변수만 선택했다. 일부 특성을 제외하였을 때 결과는 <표 6>과 같다. 해바라기유와 현물 원유 특성을 제외하였을 때 모델의 성능향상을 확인했고 최종적으로 지연 값 1, 2의 상승 하락 데이터와 지연 값 1, 2, 3의 비율 데이터에 해바라기유와 현물 원유 가격 변수를 제외한 것을 독립변수로 사용하고, 비율 데이터를 종속변수로 사용하였을 때 가장 우수한 성능을 보였다.

V. 결 론

본 연구에서는 여러 산업 분야에서 중요한 경제 작물로 여겨지는 팜유의 가격을 예측하는 방법론을 제안한다. 팜유 가격 예측 모델링을 위해 팜유 가격과 관련성이 높다고 알려진 원유 가격, 환율 등을 활용하여 시계열 예측 모델링을 진행하였으며 딥러닝 모델 중 다층 퍼셉트론 모델을 활용한다. 가격 데이터는 정상성을 충족하지 못한다는 점에 착안하여 모델의 종속변수는 가격이 아닌 가격의 비율로 선정한다. 모델의 특징으로는 코코넛 오일, 콩기름 등 9개의 데이터셋을 활용한 지연 값, 상승 및 하락 여부, 과거 가격과 현재 가격의 비율 등을 활용한다. 이 때, 종속변수 예측에 영향력이 없는 특징을 제거하기 위해 SHAP 점수를 활용하여 특징 중요도를 분석한다. 이를 통해 최종적으로 총 45개의 특징 중 35개를 설정한다. 실험 결과 DA값이 1달 뒤, 2달 뒤, 3달 뒤 예측에 대해 각각 기존 모델 대비 68.38%, 53.11%, 41.40%가량 향상된 성능을 보였고, MAPE값은 70.91%, 58.93%, 52.58%가량 향상되었다.

LSTM과 RNN를 기반으로 제작된 모델들은 시계열 데이터 분석에 있어 뛰어난 성능을 자랑한다. 반면, MLP는 LSTM이나 RNN과 같이 재귀적인 구조를 가지고 있지 않으며, 전통적인 전방향 신경망 아키텍처를 기반으로 한다. MLP는 각 레이어가 다음 레이어로 정보를 단방향으로 전달하며, 시계열 데이터의 시간적 의존성을 직접 모델링하는 능력이 LSTM이나 RNN에 비해 제한적이다. 그러나 MLP는 구조적 단순성과 계산 효율성 때문에, 빠른 학습과 더 적은 계산 자원을 요구한다는 장점이 있다. 따라서 경우에 따라 MLP만으로도 충분한 예측 성능을 달성할 수 있다.

본 연구에서 수행한 방법론의 핵심은 MLP 모델의 기본 구조를 유지하면서 데이터의 정상성을 달성하는 방법을 모색한 데 있다. 이를 통해 시계열 예측에서 종종 발생할 수 있는 비정상성 문제를 해결하고자 했다. 데이터를 정상 시계열로 변환하는 과정은 예측 모델이 데이터의 본질적인 특성과 숨겨진 패턴을 더 명확하게 포착하고 학습할 수 있게 하여, 예측의 정밀도를 향상시키는 데 결정적인 역할을 한다.

따라서 본 연구는 최신 모델의 성능을 단순히 추구하기보다는, 데이터의 기본적인 특성을 변환함으로써 MLP 모델의 잠재력을 극대화하는 데에 그 의의를 두고 있다. 이러한 방법론적 접근은 MLP 모델이 가진 기존 한계를 극복하고, 시계열 데이터의 예측에 있어서 새로운 가능성을 탐색하는 데 중요한 발걸음을 내디뎠다고 할 수 있다.

참 고 문 헌

- [1] 김명선, 정우혁, 우지환, “인공지능 기반 한국어 시 생성 시스템 개발 연구”, *경영정보학연구*, 제25권, 제3호, 2023, pp. 43-57.
- [2] Ariff, N. M., N. H. Zamhawari, and M. A. A. Bakar, “Time series ARIMA models for daily price of palm oil”, *American Institute of Physics*,

- Vol.1643, No.1, 2015. pp. 283-288.
- [3] Box, G. E. P., G. M. Jenkins, and G. Reinsel, *Time series analysis: Forecasting and control Holden-day San Francisco*, BoxTime Series Analysis: Forecasting and Control Holden Day, 1970, 1970.
- [4] Jain, G. and B. Mallick, "A study of time series models ARIMA and ETS", Available at SSRN 2898968, 2017.
- [5] Kanchymalay, K., N. Salim, A. Sukprasert, R. Krishnan, and U. R. Hashim, "Multivariate time series forecasting of crude palm oil price using machine learning techniques", *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, 2017, p. 012117.
- [6] Karia, A. A., I. Bujang, and I. Ahmad, "Forecasting on crude palm oil prices using artificial intelligence approaches", *American Journal of Operations Research*, Vol.3 No.2, 2013, Article ID: 29230.
- [7] Khalid, N., H. N. A. Hamidi, S. Thinagar, and N. F. Marwan, "Crude palm oil price forecasting in Malaysia: An econometric approach", *Jurnal Ekonomi Malaysia*, Vol.52, No.3, 2018, pp. 263-278.
- [8] Lundberg, S. M. and S. I. Lee, "A unified approach to interpreting model predictions", *Advances in Neural Information Processing Systems*, 2017, p. 30.
- [9] Muainuddin, M. M. A. M. and N. Abu, "Forecasting the Malaysian crude palm oil price during COVID-19 pandemic", *Global Business and Management Research: An International Journal*, Vol.13, No.4, 2021, pp. 564-574.
- [10] Murphy, D. J., K. Goggin, R. Paterson, and M. Russell, "Oil palm in the 2020s and beyond: Challenges and solutions", *CABI Agriculture and Bioscience*, 2021, pp. 1-22.
- [11] Patalay, S. and M. R. Bandlamudi, "Stock price prediction and portfolio selection using artificial intelligence", *APJIS*, Vol.30, No.1, 2020, pp. 31-52.
- [12] Sandika, R. A., S. K. Gusti, L. H. and S. Ramadhani, "Implementasi triple exponential smoothing dan double moving average untuk peramalan produksi kernel kelapa sawit", *Journal of Information System Research (JOSH)*, Vol.4, No.3, 2023, pp. 883-893.
- [13] Siregar, B., I. A. Butar-Butar, R. F. Rahmat, U. Andayani, and F. Fahmi, "Comparison of exponential smoothing methods in forecasting palm oil real production", In: *Journal of Physics: Conference Series*, IOP Publishing, 2017. p. 012004.
- [14] Suppalakpanya, K., R. Nikhom, T. Booranawong, and A. Booranawong, "Forecasting oil palm and crude palm oil data in Thailand using exponential time-series methods", *Engineering and Applied Science Research*, Vol.46, No.1, 2019, p. 44-55.
- [15] Yoo, H. J., H. K. Park, and B. H. Song, "Investigating the correlation between cognition and emotion characteristics and judgmental time-series forecasting using a self-organizing neural network", *APJIS*, Vol.11, No.4, Dec 2011.

Deep Learning-Based Short-Term Time Series Forecasting Modeling for Palm Oil Price Prediction

Sungho Bae^{*} · Myungsun Kim^{**} · Woo-Hyuk Jung^{***} · Jihwan Woo^{****}

Abstract

This study develops a deep learning-based methodology for predicting Crude Palm Oil (CPO) prices. Palm oil is an essential resource across various industries due to its yield and economic efficiency, leading to increased industrial interest in its price volatility. While numerous studies have been conducted on palm oil price prediction, most rely on time series forecasting, which has inherent accuracy limitations. To address the main limitation of traditional methods—the absence of stationarity—this research introduces a novel model that uses the ratio of future prices to current prices as the dependent variable. This approach, inspired by return modeling in stock price predictions, demonstrates superior performance over simple price prediction. Additionally, the methodology incorporates the consideration of lag values of independent variables, a critical factor in multivariate time series forecasting, to eliminate unnecessary noise and enhance the stability of the prediction model. This research not only significantly improves the accuracy of palm oil price prediction but also offers an applicable approach for other economic forecasting issues where time series data is crucial, providing substantial value to the industry.

Keywords: *Palm oil Price Prediction, Short-term Time Series Forecasting, Multivariate Time Series Analysis*

* Undergraduate Student, School of EECS, Gwangju Institute of Science and Technology

** Researcher, LGCNS, Generative AI Business Team

*** Researcher, CJ OliveNetworks AI Research

**** Corresponding Author, Head, CJ OliveNetworks AI Research, Adjunct Professor, School of Management of Technology, Korea University

○ 저 자 소개 ○



배 성 호 (boat1234@gm.gist.ac.kr)

광주과학기술원 전기전자컴퓨터 전공의 4학년 학부생이다. 지능 정보 시스템 연구실에서 일사량을 활용한 전력 사용량 데이터 보간 연구와 지역 간 전이 학습을 통해 가구원 정보 예측 등의 연구를 진행하고 있다. 이러한 활동을 통해, 실시간 데이터 분석 및 예측 모델 개발에 기여하고 있다.



김 명 선 (ms.kim83@cj.net)

광주과학기술원에서 에너지 데이터 분석을 전공하고, 전지전자컴퓨터공학 석사 학위를 취득하였다. 지능 정보 시스템 연구실에서 데이터 분석 관련 연구들을 진행하였으며, 현재 LGCNS에 재직 중이다. 주요 연구 분야는 시계열 데이터 분석이다.



정 우 혁 (wh.jung2@cj.net)

현재 CJ올리브네트웍스 AI연구소에서 Natural language processing 및 데이터 분석 연구를 수행 중이다. 주요 연구 분야로는 NLP, Data Analytics, HCI, Bio signal processing 등이다.



우 지 환 (jihwan_woo@korea.ac.kr)

삼성전자 삼성 리서치, 미국 카네기멜론대학 로봇연구소 등을 거쳐 현재 CJ 올리브네트웍스 AI연구소에서 연구소장으로 재직 중이며, 고려대학교 기술경영 대학원의 겸임교수이다. 컴퓨터 비전 기반 로봇틱스 및 AI 연구 개발 경력이 있으며, 이와 함께 기술 전략 수립 및 가치 평가 등에 관심이 있다.

논문접수일 : 2023년 12월 01일

게재확정일 : 2024년 02월 21일

1차 수정일 : 2024년 01월 12일

2차 수정일 : 2024년 02월 16일