

# Random Forest Model & 계절 데이터를 활용하여 향후 커피 가격 예측

## 1. 사용한 데이터

`data/final/train_weather.csv`을 사용하여 모델을 학습시켰습니다.

- 데이터 컬럼에 대한 설명은 [이곳](#) 4. Final Train Weather Data에서 확인할 수 있습니다.

## 2. 학습에 사용한 모델

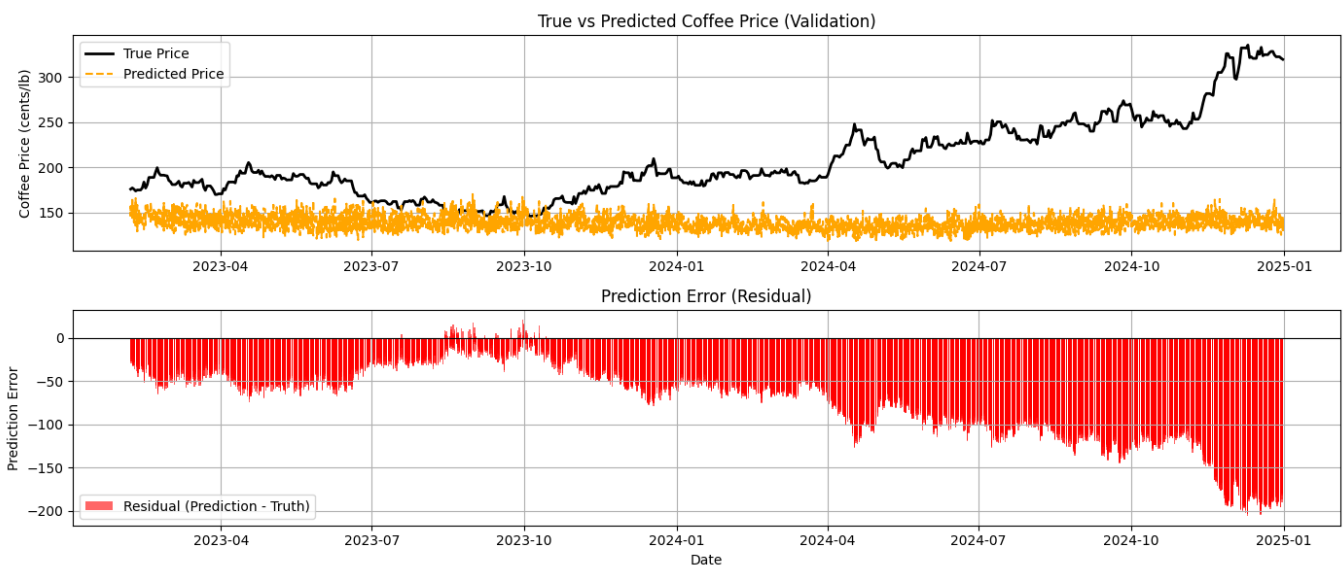
```
RandomForestRegressor(  
    random_state=42,  
    max_features='sqrt',  
    n_estimators=300,  
    n_jobs=-1  
)
```

현재는 예측 가능성을 확인하기 위해서 **Random Forest**를 사용하였으며, 하이퍼 파라미터는 위와 같이 설정하였습니다.

`train_data`는 **2015/01/01 - 2024/12/31**이며 이 기간동안의 기후 데이터를 학습하여, **2025/01/01 ~ 2025/04/01**까지 총 4개월을 예측하도록 코드를 작성하였습니다.

예측 코드는 [이곳](#)에서 확인할 수 있습니다.

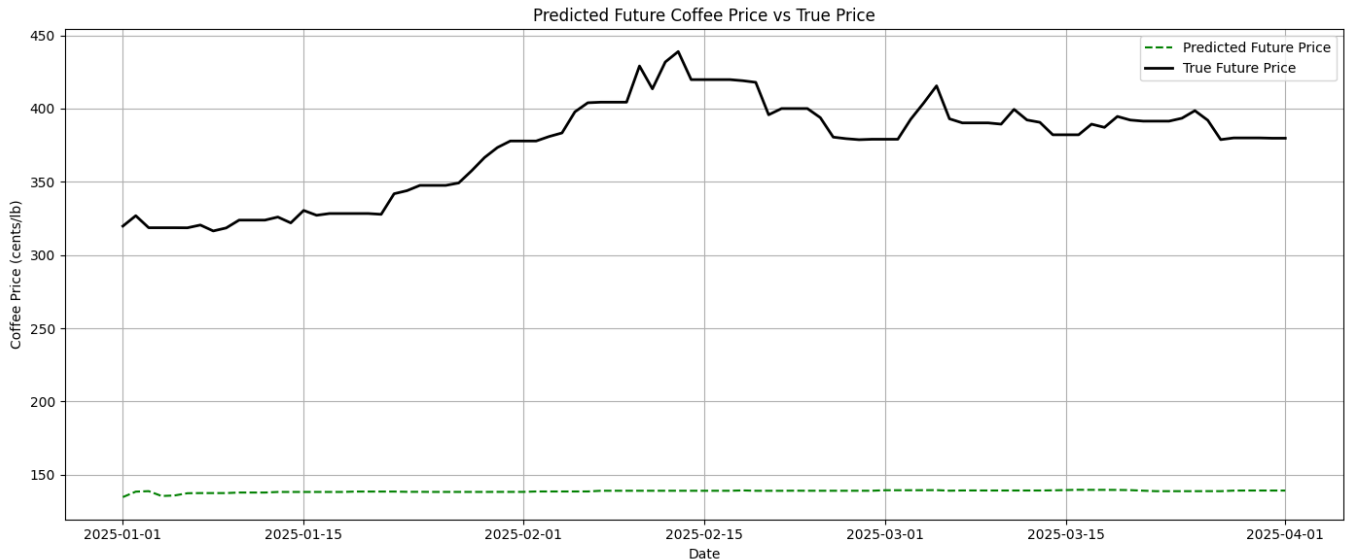
## 3. Coffee Price 예측 결과



[이미지 링크](#)

예측 Target은 당연히 커피의 가격이며, **2025/01/01 ~ 2025/04/01** 범위를 예측하기 전 train data의 최근 20퍼센트를 검증 데이터로 활용하였습니다. 검증 기간동안의 가격 예측 결과는 위와 같습니다.

위 그래프를 보면 예측 결과와 실제 값의 오차가 매우 큰 것을 알 수 있습니다. 2023년 전에는 낮은 가격에서 상승 및 횡보를 하였고, 2023년 이후에는 큰 폭으로 가격이 상승하였기에 모델은 낮은 가격의 범위에서 예측 결과를 도출하게 되어 현재 가격대에 전혀 대응을 하지 못하고 있습니다.



#### 이미지 링크

위 이미지는 테스트 기간인 **2025/01/01 ~ 2025/04/01** 동안의 커피 가격을 예측한 결과입니다. 그래프를 보면, 모델이 여전히 낮은 가격대에서 예측하려는 경향을 보이며, 이로 인해 **실제값 - 예측값**의 차이가 크게 발생하고 있습니다.

이러한 문제를 해결하기 위한 방법으로는, **가격에 로그 변환을 적용**하여 전체 값을 비슷한 범위로 스케일링함으로써 고가와 저가를 균형 있게 학습하도록 유도할 수 있습니다. 하지만 현재처럼 오차가 큰 상황에서는, 로그 변환만으로는 충분한 개선을 기대하기 어렵다고 판단하였습니다.

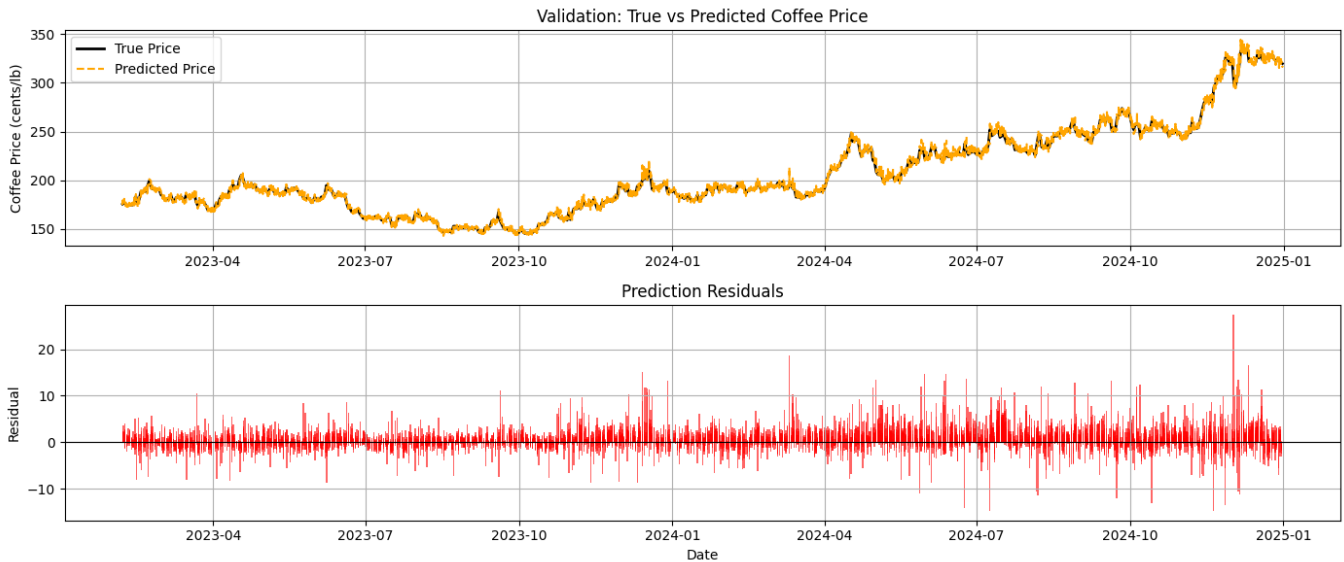
따라서 보다 직접적인 접근 방식으로, **커피 가격의 변화율(Coffee Price Return)**을 예측 대상으로 설정하여, 가격 자체보다 변동폭에 집중하는 방식으로 전환하였습니다. 이로써 모델은 절대적인 가격 수준이 아닌 상대적인 움직임에 초점을 맞추어 더 안정적인 예측 성능을 기대할 수 있게 됩니다.

---

## 4. Coffee Price Return 예측 결과

### 4-1. 수치형 피쳐 스케일링 후 진행한 결과

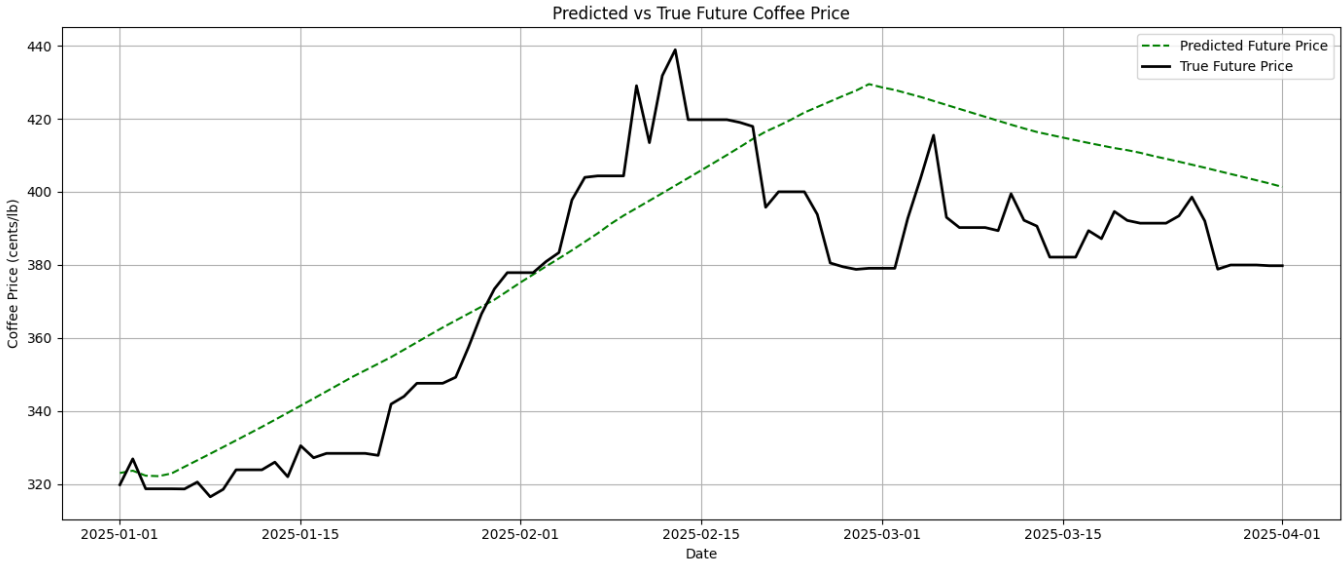
예측 코드는 [이곳](#)에서 확인할 수 있습니다.



예측 Target은 `Coffee_Return` 컬럼이며, 검증 구간(최근 20%)에서의 예측 결과는 위와 같습니다. 모델은 변화율을 예측하고, 이를 직전 커피 가격에 적용하는 방식으로 최종 가격을 계산합니다.

이처럼 커피 가격 자체를 직접 예측하는 것보다 변화율(Return)을 예측하는 방식이 훨씬 더 정교한 결과를 만들어냅니다. 실제 관측값이 존재하는 검증 구간에서는 급격한 기후 변화나 패턴 등을 모델이 잘 포착해내기 때문에 정확도 높은 예측이 가능합니다.

하지만 아래와 같이 미래를 예측할 경우, 실제 기후 데이터를 알 수 없으므로 오직 과거 관측값을 활용한 lag feature들만 사용할 수밖에 없습니다. 이로 인해 입력 변수들의 정보가 제한되고, 결과적으로 만족스럽지 못한 예측 결과를 초래합니다.

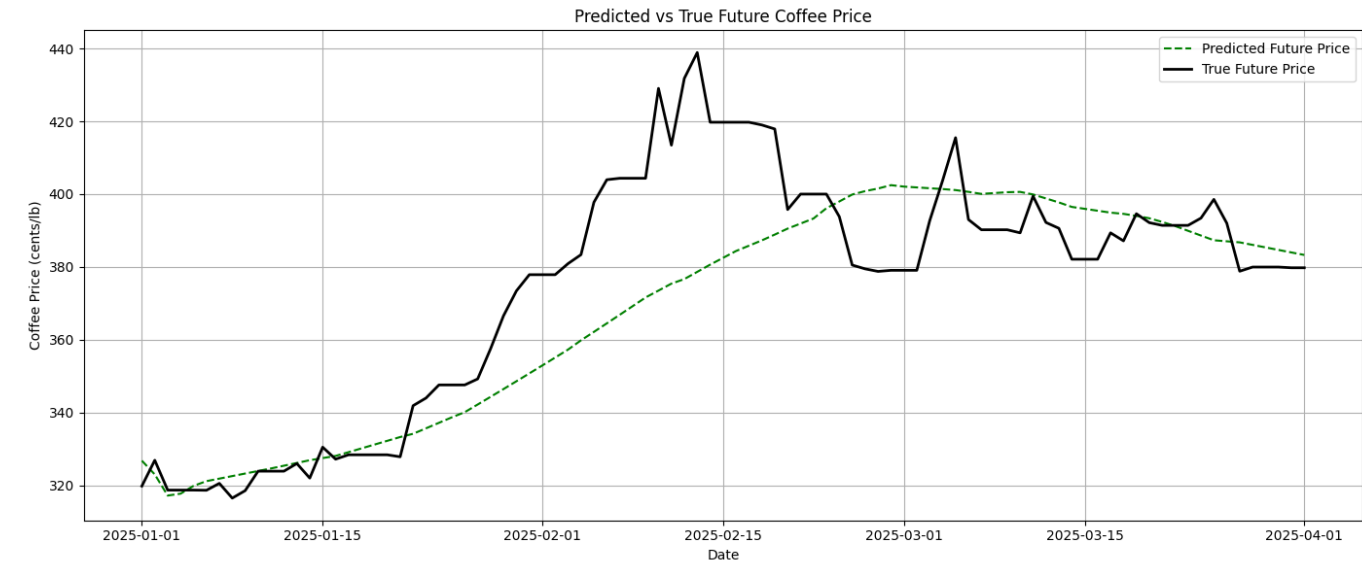


그럼에도 불구하고, 전반적인 상승 혹은 하락의 추세는 어느 정도 예측이 가능하다는 점에서 의미 있는 방향성을 보여줍니다.

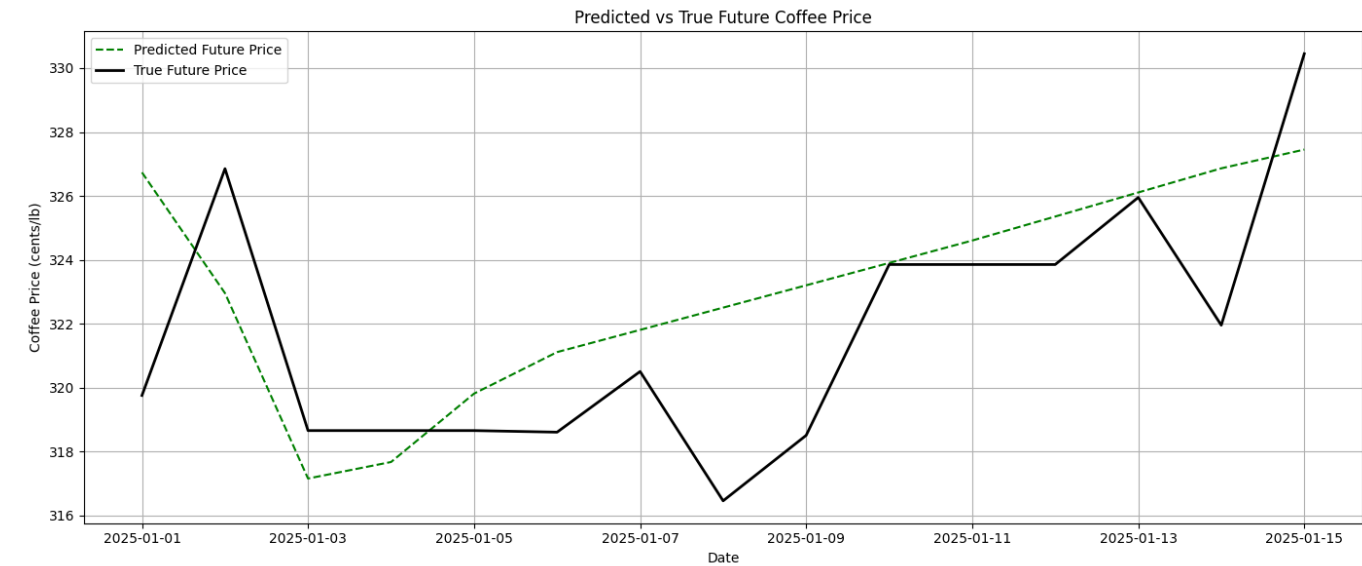
4-2. 수치형 피쳐 스케일링 없이 진행한 결과

앞서 4-1에서 수치형 피쳐에 스케일링을 적용한 채 모델을 학습했을 때, Random Forest 특성상 보수적으로 학습되며 큰 변화율을 예측하지 못하고, 비교적 완만한 상승 또는 하락만을 보였습니다.

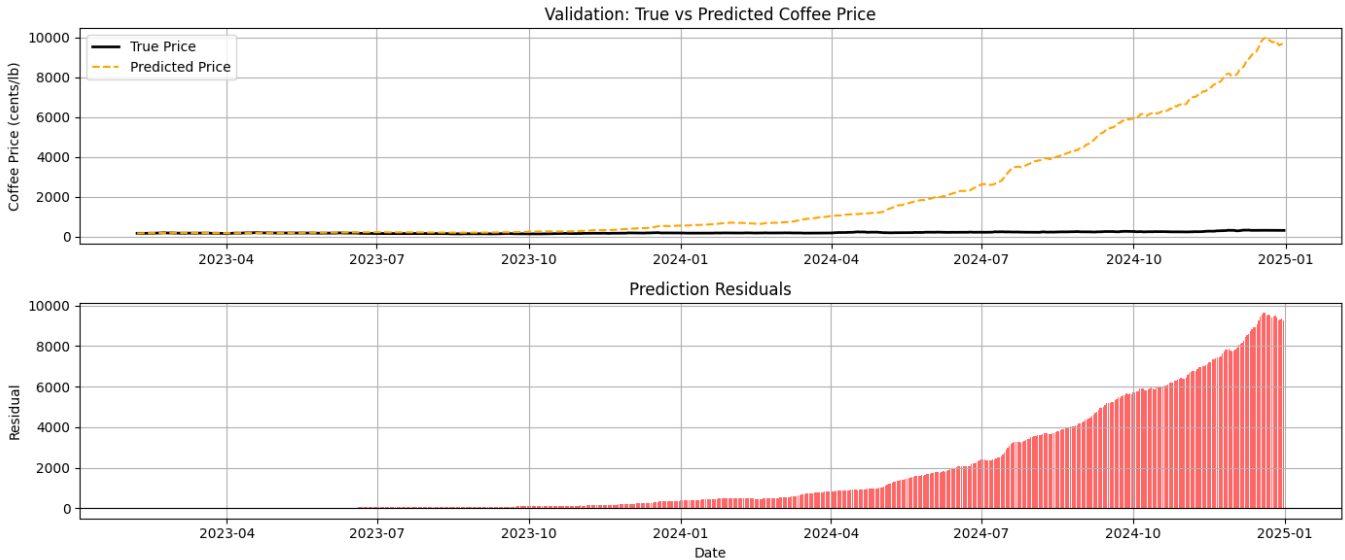
이에 따라 스케일링 없이 모델을 학습한 결과는 아래와 같습니다.



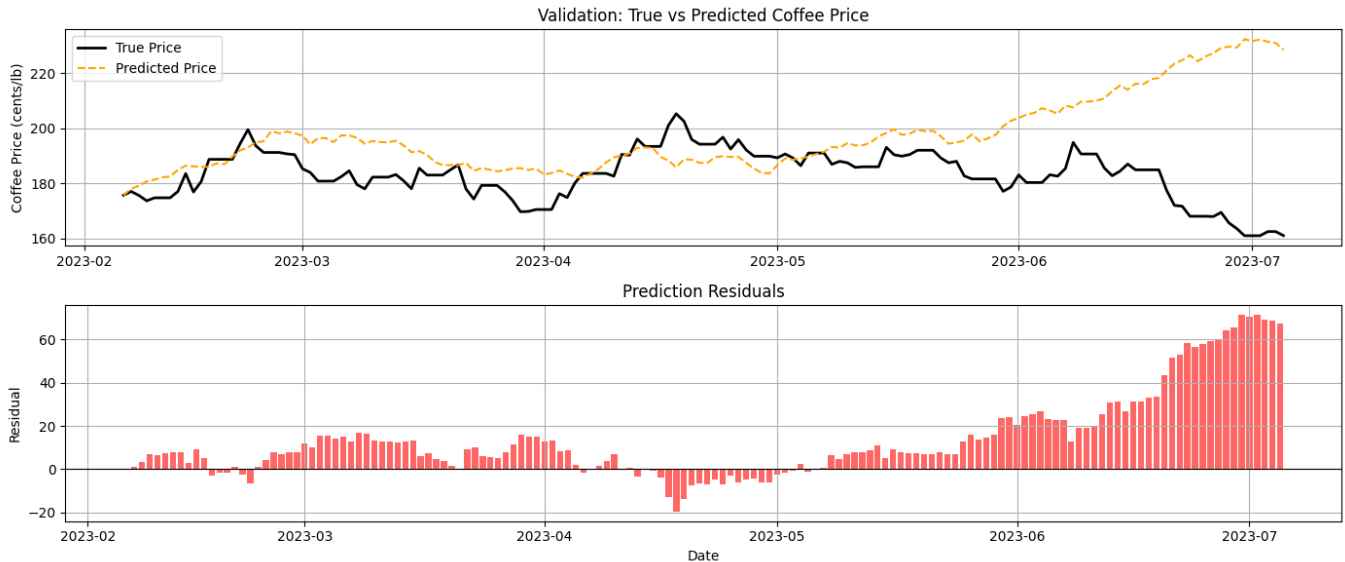
이번에는 예측이 보다 과감하게 이루어진 것처럼 보이지만, 전체적인 정확도 측면에서는 여전히 만족스럽다고 보기는 어렵습니다. 다만, 단기 예측(2주) 성능은 꽤 괜찮은 결과를 보였습니다.




아래 이미지는 validation 구간에서의 예측 결과입니다. 전반적으로 과감한 예측 경향으로 인해, 특정 시점 이후 부터 예측값이 급격히 상승하는 모습을 확인할 수 있습니다.



예측의 신뢰 구간이 어느 정도까지 유효한지 확인하기 위해, 앞쪽 5개월만 잘라 시각화한 결과는 다음과 같습니다.



이를 바탕으로 판단해보면, 최대 약 3개월 이내의 예측이 한계선이라고 생각합니다.

 rf\_pred\_result\_v2\_first3m\_plot

## 5. 정확도 향상을 위한 개선 방안

현재 모델은 Random Forest 기반의 회귀 모델로, 과거 기후 및 계절 데이터를 바탕으로 커피 가격의 변화율을 예측하는 구조입니다. 다만, 미래 예측 구간에서는 실제 관측값이 없어 lag feature에 의존해야 하므로 예측의 한계가 존재합니다.

이러한 한계를 극복하고 예측 성능을 향상시키기 위해 다음과 같은 방법들을 고려할 수 있습니다:

- 시계열 특화 모델 도입:
  - **LSTM (Long Short-Term Memory), GRU, Transformer** 기반의 시계열 모델 등을 활용할 수 있습니다.

- 이들은 순차적인 데이터를 효과적으로 학습할 수 있으며, 시점 간의 연속성을 고려할 수 있어 미래 예측에 더 적합할 수 있습니다.
- 특히 LSTM은 과거의 흐름을 기억하고, 그 패턴을 기반으로 미래 값을 예측하는 데 강점을 지니고 있습니다.
- **외부 데이터 활용:**
  - 현재는 이미 기록된 기후 데이터를 활용하였지만, **예보 데이터**를 통해 예측된 기후 데이터를 활용 (온도, 예측 강수량 등)한다면 실제 관측값에 가까운 기후 데이터를 feature를 사용할 수 있습니다.
  - 환율, 국제 유가 등 커피 가격과 관련된 외부 데이터를 활용한다면, 더 높은 정확도를 기대할 수 있습니다.
  - **거시경제 지표**도 커피 가격에 영향을 줄 수 있으므로, 이를 feature로 추가한다면 성능 향상을 기대할 수 있습니다.
- **모델 앙상블 또는 하이브리드 방식 활용:**
  - 서로 다른 알고리즘을 조합한 앙상블 모델(Random Forest + LSTM 등) 을 구성하면, 단일 모델의 한계를 보완할 수 있습니다.
  - 예를 들어, 단기 예측은 시계열 모델로, 중장기 예측은 기후 변수 기반 모델로 분리하여 예측하는 하이브리드 전략도 고려할 수 있습니다.