

2025년도 커피 가격 예측 프로젝트
AI개발 수행내역서

과제명	AI기반 예측모델 개발 및 시각화
담당자	최두석

2025년 12월 29일

사업과제 : ML 기반 커피 가격 상승/하락 예측 모델 개발 및 시각화

* 순서

- ① 프로젝트 개요
- ② 데이터 수집 및 전처리
- ③ 탐색적 데이터 분석(EDA)
- ④ 모델 선정 및 성능 평가
- ⑤ Streamlit 상호작용 시각화
- ⑥ 결론 및 향후 계획

1. 개요

1.1 추진배경

- 2024년 기준 **국제 커피 가격**은 약 **47년 만에 최고치**를 기록하며, 전 세계 원자재 시장에서 높은 변동성을 보이고 있다.
- **한국**은 1인당 연간 약 367잔을 소비하는 **세계 2위 커피 소비국**으로, 국제 커피 가격 변동은 국내 프랜차이즈 가격, 원가 구조, 소비자 물가에 직접적인 영향을 미친다.
- 최근 브라질의 가뭄, 베트남의 생산 차질, 엘니뇨·라니냐 등 기후 변화는 커피 생산량 및 공급망 불안정을 초래하며 가격 변동성을 더욱 확대시키고 있다.

1.2 추진 목적

- 본 과제는 글로벌 원자재 시장에서 변동성이 큰 커피 가격을 대상으로, 과거 가격, 환율, 유가, 기후 요인 등을 활용하여 **머신러닝 기반 가격 상승/하락 예측 모델을 개발**하고, 예측 결과를 시각화 및 사용자 입력 기반 시스템(Streamlit)으로 구현하는 것을 목표로 한다.
- 단순 시계열 예측을 넘어, 기후 변수(기온, 강수량, 엘니뇨 지수), 거시경제 변수(환율, 유가) 등을 독립 변수로 활용하여 가격 변동의 구조적 원인을 분석한다.
- 최종적으로는 예측 결과를 시각화하여 커피 프랜차이즈, 유통업체 등에서 가격 전략 수립 및 리스크 관리에 실무적으로 활용 가능한 의사결정 지원 도구를 구축하는 것을 목표로 한다.

1.3 과제 범위

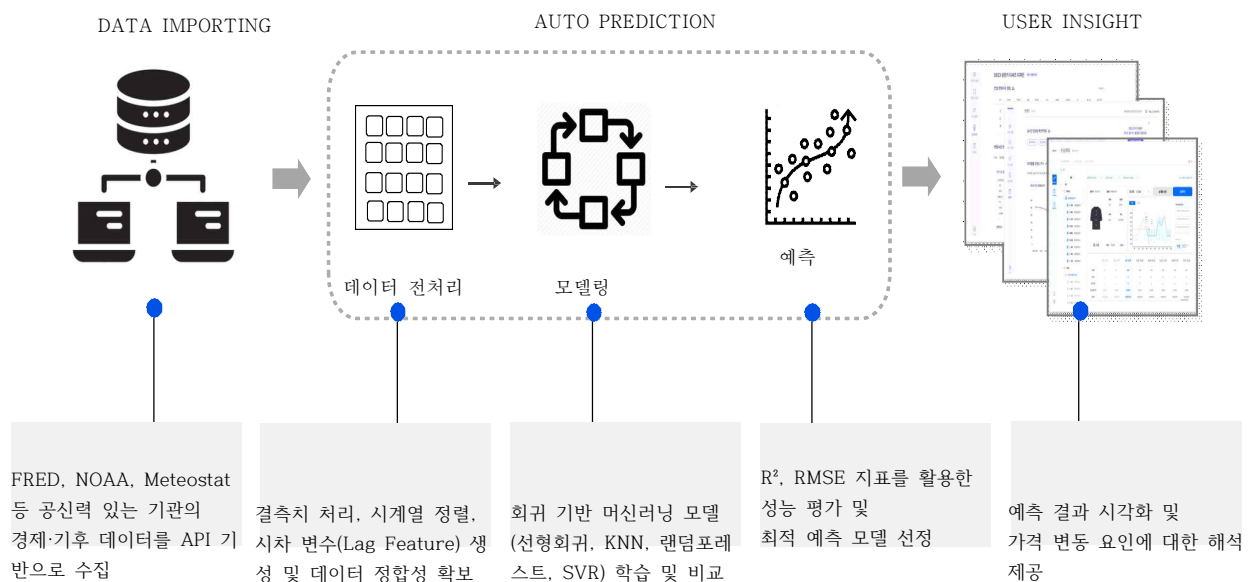
과제구분		내용
AI	AI 기반 예측모델 구현	원시 데이터 수집 및 데이터셋 구축
		데이터 전처리, 표준화, 상관관계 분석 (EDA도구 등 활용)
		예측모델 선정 및 학습
		RMSE, MAE 등 평가지표를 활용한 모델 성능 평가
		STREAMLIT 및 프로토타입 구축
		예측모델 웹기반 시스템 구축
		테스트

1.4 과제 추진 방법

1) AI 예측 분석모델 적용 대상

구분	수집 데이터	예측모델인자(독립변수)	AI예측 분석 대상
	국제 커피 가격, 환율, 유가, 기후 데이터	USD/BRL 환율, WTI 유가, 기온·강수량, 엘니뇨 지수, 시차 변수(Lag)	국제 커피 가격(아라비카) 예측

2) AI 분석모델 구축 프로세스



2. 데이터 수집 및 전처리 · 탐색적 분석

2.1 데이터 수집

- 본 연구에서는 공공 신뢰도가 높은 국제 기관(FRED, NOAA, Meteostat)에서 제공하는 데이터를 활용하였다.

커피 가격은 국제 원자재 가격(FRED)을 기준으로 하였으며, 브라질 생산 환경을 반영하기 위해 환율, 유가, 엘니뇨 지수, 강수량, 폭염 데이터를 함께 수집하였다.

- 커피 가격: FRED (PCOFFOTMUSDM)
- 환율(USD/BRL): FRED
- 유가(WTI): FRED
- 엘니뇨 지수(ONI): NOAA CPC
- 기후 데이터(강수량, 폭염): Meteostat (브라질 주요 산지)

뉴스

-(<https://www.hankyung.com/article/2025112825641>)

-(https://www.beanbrothers.co.kr/info/bb_contents_detail?goodsNo=1000001085&cateNm=%EC%A0%80%EB%84%90)

- 커피 가격 사상 최고가의 원인으로 브라질 기상문제(강수량, 건조, 가뭄, 한파), 환율 문제를 언급하여 이들을 feature로 선정하였다.

- 커피 최대 생산국이 브라질, 베트남이기 때문에, 운송비를 반영하기 위해서 feature에 유가를 반영하였다.

- 커피는 1년에 최대 1~2회 수확을 한다는 점을 고려하여, 3개월 전 강수량, 3개월 전 평균 온도, 3개월전 엘니뇨 지수를 feature로 선정하였다.

- 유가와 달러를 제외하고는 커피 최대 생산국인 브라질의 아라비카 커피 가격, 브라질 기후를 사용하였다.

2.2 분석 기간

2015년 1월 ~ 2025년 12월 (월 단위 데이터)

2.3 Feature(독립 변수) 구성 및 선정 기준

- Feature(독립 변수) 구성 및 선정 기준

커피 가격 예측에 사용된 Feature는 커피 가격에 영향을 미치는 것으로 알려진 기후, 거시경제, 시계열 요인을 중심으로 구성하였다. 각 변수는 선행 연구 및 산업 보고서를 참고하여 선정하였다.

시차 변수	커피 가격: 1개월, 3개월 전 폭염 발생 여부: 3개월 전 강수량: 1개월 전 엘니뇨 지수: 1개월 전	가격 및 기후 요인의 즉각적인 영향뿐 아니라 지연 효과(delay effect)를 반영하기 위해 1~3개월 시차 변수를 생성하였다. 실제 농산물 시장에서 기후 영향이 즉각적으로 반영되지 않는 특성 이 있기 때문이다.
변화율 / 기술적 지표	전월 대비 가격 변화율, 환율 및 유가 변화율, 3개월 이동평균 대비 가격 비율	단순 가격 수준뿐 아니라 변동성 및 추세 정보를 반영 하기 위해 전월 대비 변화율과 이동평균 대비 편차 지표를 추가하였다.
타겟 변수	다음 달 커피 가격이 상승 → 1 하락 또는 동일 → 0	본 연구에서는 가격 자체 예측이 아닌, “다음 달 가격 상승 여부”를 이진 분류 문제 로 정의하였다. 이는 실무적 의사결정 (구매 시점, 재고 전략)에 보다 직접적인 활용이 가능하기 때문이다.

coffee_lag1	1개월전 커피 가격	단기 추세 파악
coffee_lag3	3개월전 커피 가격	중기 추세 파악, 계절적 패턴
coffee_pct_1m	커피 전월 대비 가격 변화율 (%)	급등/급락 감지
usd_brl_pct_1m	USD/BRL 환율 변화율	브라질은 세계 1위 커피 생산국
wti_pct_1m	WTI 유가 변화율	물류비 (배송비) 영향
price_vs_ma3	3개월 이동평균 대비 현재가	평균 회귀 성향 파악
heatwave_lag3	3개월 전 폭염 발생 여부 (브라질 주요 산지에서 30°C 이상 연속 2일 발생)	커피나무는 고온 스트레스에 약함 3개월 시차: 생육 피해가 가격에 반영되는 시간
precip_lag1	1개월 전 강수량	브라질 주요 산지의 월 강수량 너무 많으면 → 병충해 발생 너무 적으면 → 가뭄 스트레스
elnino_lag1	엘니뇨 지수 (ONI)	적도 태평양 해수면 온도 편차 ONI > +1.0 → 강한 엘니뇨 → 가격 상승 압력 ONI < -1.0 → 강한 라니냐 → 가격 하락 압력
month	예측 대상 월	브라질 수확기: 5~9월 → 공급 ↑ → 가격 ↓ 비수확기: 11~3월 → 공급 ↓ → 가격 ↑

3. 탐색적 데이터 분석(EDA)

3.1 결측치 확인

```
-----  
#      Column      Non-Null Count  Dtype  
-----  
0      date         126 non-null    datetime64[ns]  
1      coffee_price  126 non-null    float64  
2      usd_brl       82 non-null     float64  
3      wti_price     82 non-null     float64  
4      el_nino_index  126 non-null    float64  
5      br_precip     48 non-null     float64  
6      heatwave_2d_month 108 non-null    float64  
dtypes: Float64(1), datetime64[ns](1), float64(5)  
memory usage: 7.1 KB
```

usd_brl(34.92%), wti_price(34.92%), br_precip(61.9%),
heatwave_2d_month(14.29%)의 결측치가 발생하였다.

3.2 결측치 처리 방법

단순 결측치 제거 시 데이터 수가 크게 감소하여 시계열 모델 학습에 불리하다고 판단하였다. 이에 따라 환율 및 유가와 같은 연속적인 시계열 변수의 특성을 고려하여, 직전 및 이후 시점의 값을 활용한 보간(Forward Fill, Backward Fill) 방식을 적용하였다.

반면, 커피 시차(Lag) 변수 생성 후 발생한 초기 결측치는(3개월(월단위 자료이므로 3개)의 결측) 데이터 손실을 최소화하기 위해 인접 시점 값을 활용하여 보간 처리하였다.

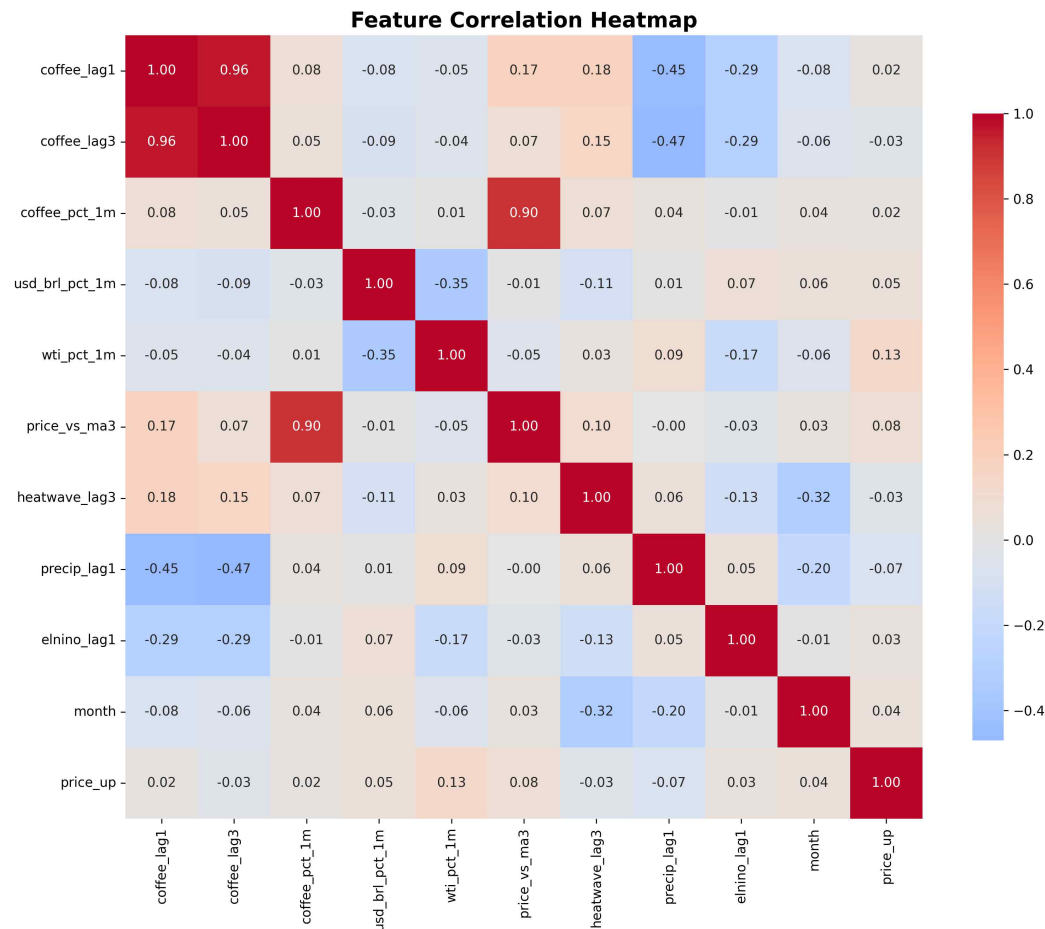
3.3 데이터 분할 전략

시계열 데이터의 특성을 고려하여 임의 섞기(shuffle)를 사용하지 않았다.

학습 데이터: 전체의 80%, 테스트 데이터: 이후 20%

이를 통해 미래 예측 상황을 현실적으로 모사하였다.

3.4 다중공선성 분석 (Heatmap)



- 변수 간 상관관계 분석 결과

위 그림은 주요 입력 변수 및 타겟 변수(price_up) 간의 피어슨 상관계수를 시각화한 결과이다. 상관관계 분석 결과, **과거 커피 가격 변수(coffee_lag1, coffee_lag3)** 간에는 매우 높은 양의 상관관계(약 0.96)가 확인되었다. 이는 시계열 데이터 특성상 인접 시점의 가격이 유사한 값을 가지는 구조적 특성에 기인한 것으로, 자연스러운 현상으로 해석된다.

- 가격 파생 변수 간 관계

또한 ****커피 가격 변화율(coffee_pct_1m)****과 **이동평균 대비 가격 비율(price_vs_ma3)** 간에는 높은 양의 상관관계(약 0.90)가 관찰되었다. 이는 두 변수가 모두 단기 가격 추세 및 모멘텀을 반영하는 파생 변수이기 때문에 동일한 시장 움직임을 부분적으로 공유하는 결과로 판단된다.

- 기후 및 외생 변수의 특성

환율(usd_brl_pct_1m), 유가(wti_pct_1m), 강수량(precip_lag1), 폭염(heatwave_lag3), 엘니뇨 지수(elnino_lag1) 등 외생 변수들은 가격 변수들과 비교적 **낮은 상관계수**를 보였다. 이는 가격 변동이 단일 요인에 의해 결정되지 않고, 복합적인 요인의 상호작용에 의해 발생함을 시사한다.

- 타겟 변수(price_up)와의 관계

타겟 변수인 ****price_up(가격 상승 여부)****는 개별 변수와의 단순 상관계수가 전반적으로 낮게 나타났다. 이는 다음 달 가격 상승 여부가 단일 변수보다는 여러 요인의 **비선형적 결합**에 의해 결정된다는 점을 의미하며, 머신러닝 기반 분류 모델 적용의 타당성을 뒷받침한다.

- 다중공선성 발생 가능성
상관관계 분석 결과 일부 가격 기반 변수 간에는 높은 상관관계가 존재하여 다중공선성 발생 가능성이 확인되었다.
coffee_lag1 ↔ coffee_lag3
coffee_pct_1m ↔ price_vs_ma3

- 다중공선성에 대한 대응 전략

가. 모델 특성 기반 대응

Random Forest, Gradient Boosting 등 트리 기반 모델은 다중공선성에 비교적 강건함

Logistic Regression 등 선형 모델은 성능 비교 및 해석 목적 중심으로 활용

나. 성능 기반 Feature 유지

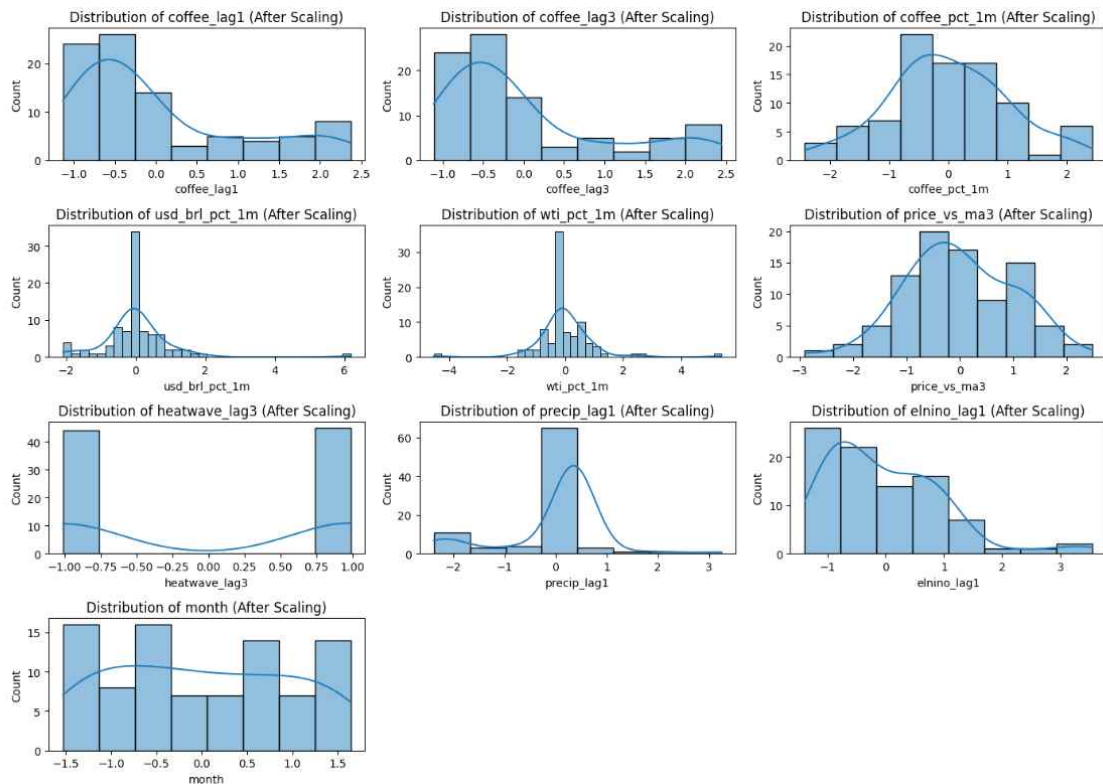
단순 상관관계가 높더라도 모델 성능(ROC-AUC) 및 Feature Importance 분석 결과 예측 기여도가 확인된 변수는 제거하지 않고 유지

다. 예측 목적 중심 설계

본 과제는 계수 해석 중심의 인과 분석이 아닌, 예측 정확도 및 판별력 향상을 목적으로 함

3.5 수치형 변수 스케일링

- **StandardScaler** 수행



Logistic Regression / SVM / KNN을 쓰는 경우 → 반드시 스케일링 필요
 수치형 변수 분포를 시각화한 결과, 각 변수는 서로 다른 범위와 분포 형태를 가지며
 일부 변수는 치우친 분포를 보였다. 이러한 특성을 고려하여 스케일링을 수행하였다.

3.6 온도를 feature에 넣는게 맞는가?

커피 나무에 대해 조사를 하니, *커피는 1년에 1~2회 수확을 한다. *커피 나무는 30도 이상 폭염은 연속 2~3일 이상 겪으면, 생장 둔화, 광합성 저하, 개화 지연 등의 영향을 미친다는 것을 파악하였다.
 처음 단순 월 온도를 feature 변수에서 제거하고, 1달중 30도 이상 2일 이상 사건이 있었는가를 feature 변수에 넣었다.

3.7 이상치(outlier) 처리

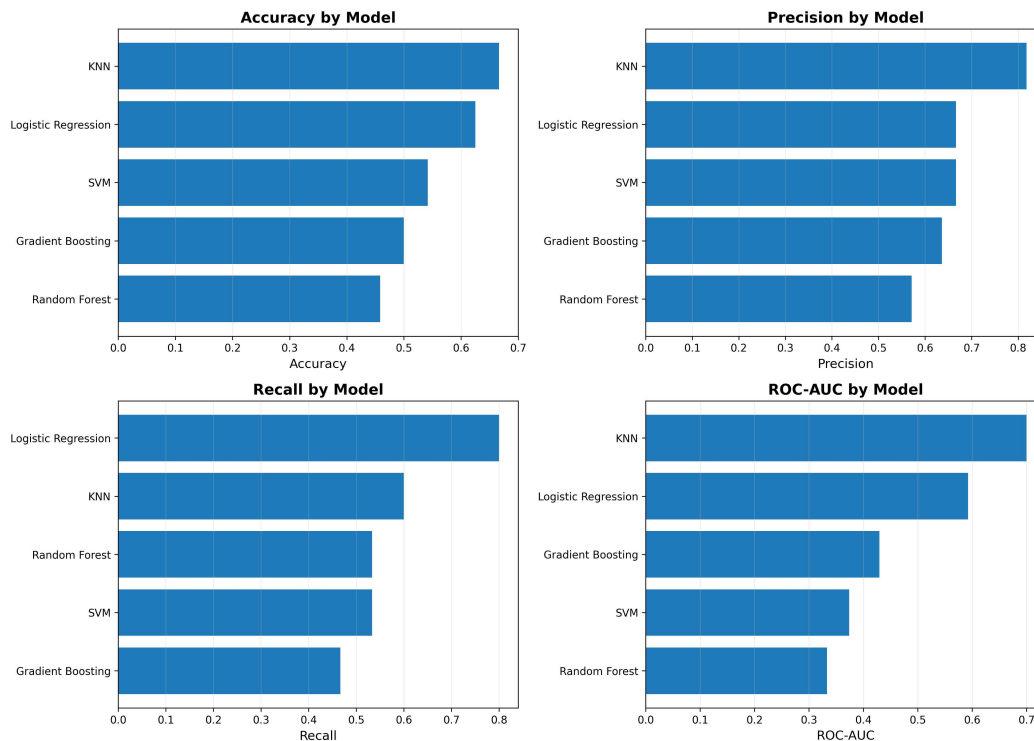
본 연구에서는 커피 가격 급등락을 이상치가 아닌 시장 이벤트로 판단하여 별도의 이상치 제거를 수행하지 않았다.

4. 모델 선정 및 성능 평가

4.1 모델 비교 및 선정

모델 성능 비교:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
KNN	0.739130	0.909091	0.666667	0.769231	0.745833
Logistic Regression	0.652174	0.666667	0.933333	0.777778	0.483333
Random Forest	0.478261	0.588235	0.666667	0.625000	0.475000
Gradient Boosting	0.478261	0.600000	0.600000	0.600000	0.391667
SVM	0.608696	0.750000	0.600000	0.666667	0.350000



KNN 모델

Accuracy: 0.739 (가장 높음)

ROC-AUC: 0.746 (가장 높음)

KNN 모델은 전체 정확도와 ROC-AUC에서 가장 우수한 성능을 보였다.

이는 유사한 과거 패턴을 기반으로 한 예측 방식이 커피 가격 데이터의 국소적 패턴을 효과적으로 포착했기 때문으로 해석된다.

Logistic Regression

Recall: 0.933 (매우 높음)

Precision은 상대적으로 낮음

Logistic Regression 모델은 Recall이 매우 높아 가격 상승을 실제로 놓치지 않는 데 강점이 있다. 이는 **가격 상승 신호를 민감하게 탐지**하는 특성에 기인한다.

다만 Precision이 상대적으로 낮아 상승이 아닌 경우를 상승으로 예측하는 경향이 존재한다.

Random Forest

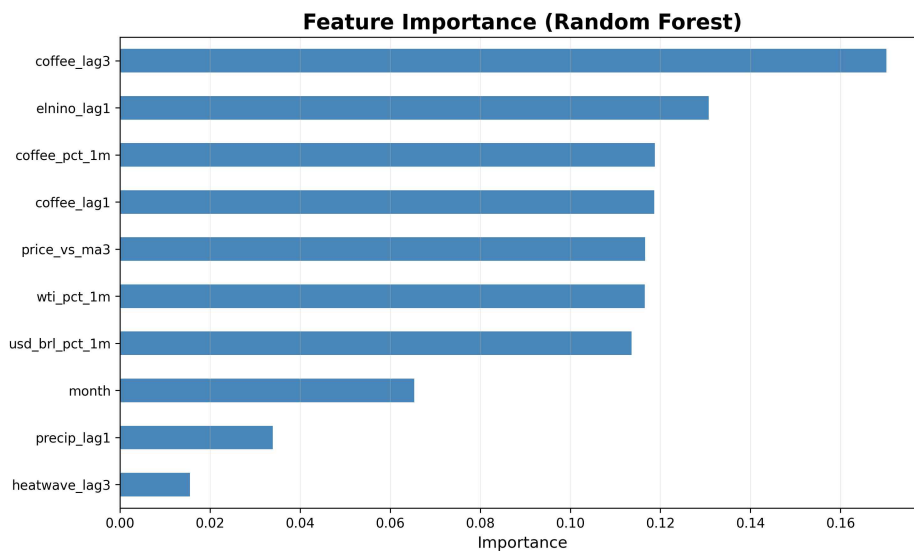
전반적으로 **중간 수준의 안정적인 성능**

Random Forest 모델은 모든 지표에서 중간 수준의 성능을 보였으며, 특정 지표에 치우치지 않은 비교적 안정적인 결과를 나타냈다. 이는 트리 기반 앙상블 모델의 일반적인 특성으로 판단된다.

본 과제에서는 단순 Accuracy만으로 모델을 평가하지 않았다. 이는 가격 상승/하락 데이터가 불균형할 가능성이 있으며, Accuracy만으로는 실제 판별 성능을 과대평가할 수 있기 때문이다. 이에 따라 ROC-AUC와 Recall, Precision을 함께 고려하여 모델을 해석하였다.

4.2 feature importance 설명

Random Forest 기반 중요도 분석 결과, 커피 가격 예측에는 과거 가격 정보와 가격 추세 변수의 영향이 가장 큰 것으로 나타났으며, 기후 요인과 거시 변수는 보조적인 역할을 수행하였다.



분석 결과, **coffee_lag3(3개월 전 가격)** 변수가 가장 높은 중요도를 보였으며, 이는 커피 가격이 단기적인 급변보다는 **중기적인 가격 흐름의 영향을 크게 받는다는 점**을 시사한다.

또한 coffee_lag1(1개월 전 가격), coffee_pct_1m(전월 대비 가격 변화율), price_vs_ma3(이동평균 대비 가격 위치) 등 가격 추세 및 모멘텀 관련 변수들이 상대적으로 높은 중요도를 차지하였다.

기후 요인 중에서는 ****elnino_lag1(엘니뇨 지수)****가 폭염(heatwave_lag3)이나 강수량(precip_lag1)보다 상대적으로 높은 중요도를 보였다.

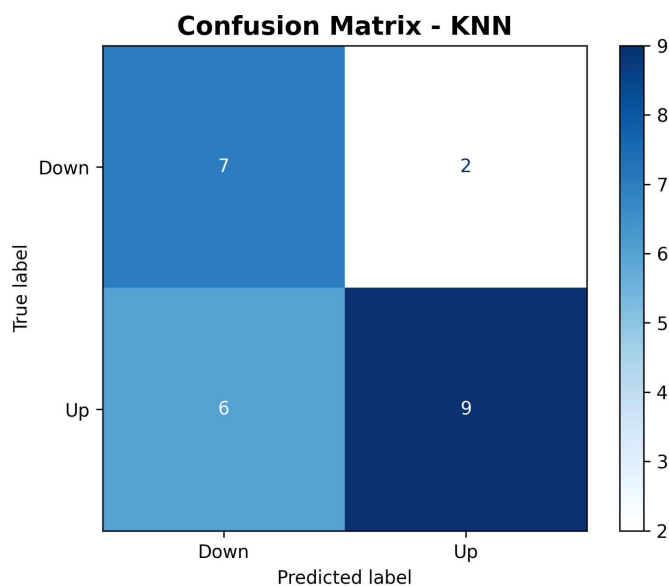
이는 단기 기상 이벤트보다 **대규모 기후 패턴이 커피 생산 및 가격에 더 구조적인 영향을 미친다는 점**을 시사한다.

환율(usd_brl_pct_1m)과 유가(wti_pct_1m)는 중간 수준의 중요도를 보였으며, 이는 생산·물류 비용을 통해 가격에 간접적인 영향을 미치는 요인으로 해석할 수 있다.

폭염 발생 여부(heatwave_lag3)와 강수량(precip_lag1)은 뉴스에서 언급된 것과 달리 상대적으로 낮은 중요도를 보였다.

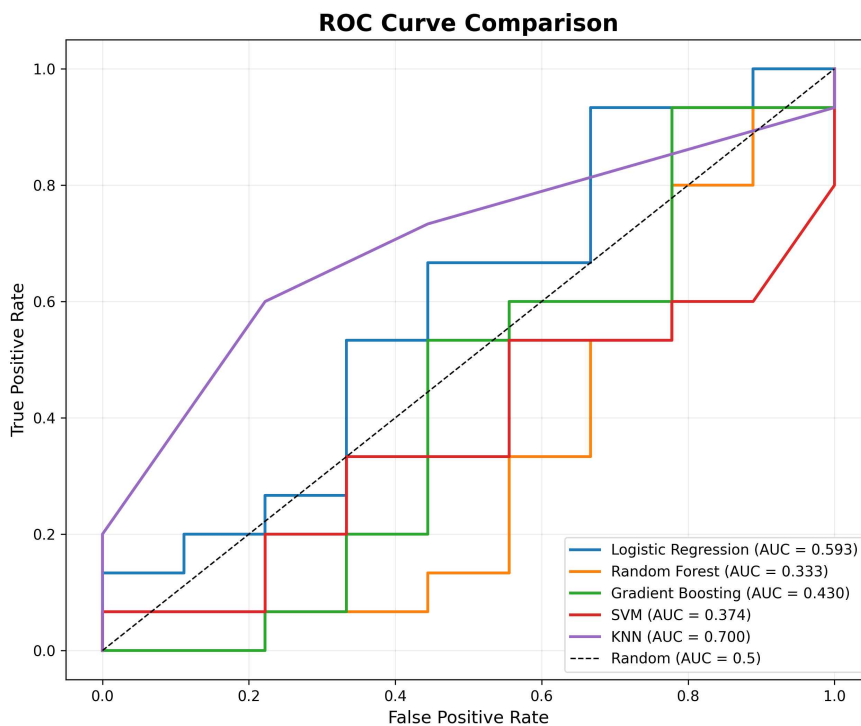
월 단위 데이터로 집계되면서 단기 기후 충격이 완화되었을 것으로 해석하였다.

4.3 confusion matrix



KNN 모델은 **가격 상승(Up)에 대해서는 비교적 잘 예측하였으나**, 하락(Down)을 상승으로 잘못 예측하는 사례가 일부 존재하였다. 단독 사용보다는 다른 지표와 병행하는 것이 바람직할 것으로 보인다.

4.4 ROC Curve



모델별 ROC Curve 및 AUC를 비교한 결과, **KNN 모델이 AUC 0.700으로 가장 우수한 성능을 보였다.** 이는 다양한 임계값 구간에서 가격 상승과 하락을 비교적 안정적으로 구분할 수 있음을 의미한다.

Logistic Regression 모델은 AUC 0.593으로 무작위 예측(AUC = 0.5) 대비 일정 수준의 판별력을 보였으나, KNN 모델에 비해서는 상대적으로 낮은 성능을 나타냈다. 반면 Random Forest, Gradient Boosting, SVM 모델은 AUC가 0.5 이하로 나타나, 본 데이터셋에서는 가격 방향성 구분에 한계가 있음을 확인하였다.

KNN 모델은 가격 변동 패턴이 유사한 과거 사례를 기반으로 예측을 수행하기 때문에

본 데이터셋과 같이 패턴 반복성이 존재하는 경우 상대적으로 우수한 판별 성능을 보인 것으로 해석된다.

반면 트리 기반 모델은 데이터 규모 및 Feature 구조상 충분한 분기 정보를 확보하지 못해

ROC-AUC 성능이 제한적으로 나타난 것으로 판단된다.

ROC-AUC는 특정 임계값에 의존하지 않고 모델의 전반적인 분류 순위 능력을 평가

하는 지표이다. 따라서 Accuracy가 상대적으로 높더라도 ROC-AUC가 낮게 나타날 수 있으며, 본 과제에서는 가격 상승/하락 예측은 특정 임계값에 따라 의사결정이 달라질 수 있으므로, 임계값에 독립적인 ROC-AUC 지표가 적합하다고 판단했습니다.

4.5 최종 모델 : KNN

5. Streamlit 상호작용 시각화

5.1 feature 설명 시각화

☕ 커피 가격 상승/하락 예측 시스템

[Feature 설명](#) [모델 정보](#) [사용 가이드](#)

Feature 설명

1 과거 가격 정보

- `coffee_lag1`: 1개월 전 커피 가격 - 단기 추세 파악
- `coffee_lag3`: 3개월 전 커피 가격 - 중기 추세 파악

2 변화율

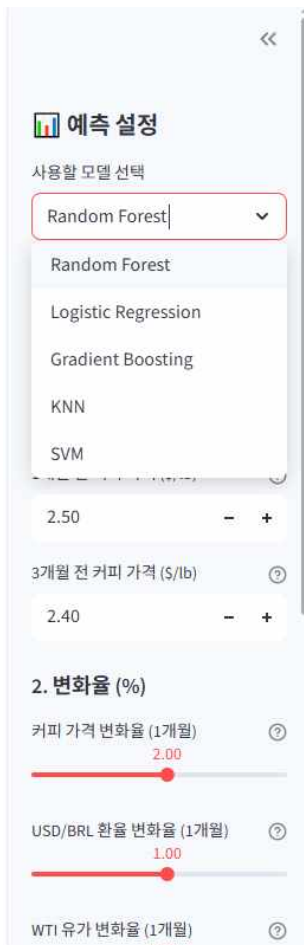
- `coffee_pct_1m`: 커피 가격 변화율 - 모멘텀 지표
- `usd_brl_pct_1m`: 환율 변화율 - 브라질 수출 경쟁력 영향
- `wti_pct_1m`: 유가 변화율 - 물류비 및 생산비 영향

3 기술적 지표

- `price_vs_ma3`: 3개월 이동평균 대비 - 과매수/과매도 판단

feature 설명 페이지를 넣어서, 어떤 요인이 커피 가격에 영향을 줄수 있는지 설명하였습니다.

5.2 모델 선택 시각화



☺ 커피 가격 상승/하락 예측 시스템

☐ Feature 설명 ☒ 모델 정보 ☐ 사용 가이드

Feature 설명

1 과거 가격 정보

- coffee_lag1: 1개월 전 커피 가격 - 단기 추세 파악
- coffee_lag3: 3개월 전 커피 가격 - 중기 추세 파악

2 변화율

- coffee_pct_1m: 커피 가격 변화율 - 모멘텀 지표
- usd_brl_pct_1m: 환율 변화율 - 브라질 수출 경쟁력 영향
- wti_pct_1m: 유가 변화율 - 물류비 및 생산비 영향

3 기술적 지표

각 모델의 특성이 다름을 활용할 수 있도록, 모델을 선택하여, 예측을 할 수 있도록 하였습니다.

6. 결론 및 향후 계획

6.1 모델 개별 결과 요약

모델 비교 결과, 단일 지표 기준의 우열 판단은 적절하지 않았으며 예측 목적에 따라 적합한 모델이 상이함을 확인하였다.

본 과제에서는 전반적인 판별력이 가장 우수한 KNN 모델을 최종 분석 중심 모델로 활용하였다.

6.2 실무 적용 방안

커피 수입·유통 기업은 다음 달 가격 상승 가능성이 높은 경우 선구매 또는 재고 확보 전략을 사전에 검토할 수 있다.

공공기관 관점에서는 본 모델을 활용하여 원자재 가격 변동이 소비자 물가에 미치는 영향을

사전에 모니터링할 수 있다.

6.3 한계점 및 개선방향

가. 기후 피해 현상의 정교한 반영 한계

본 과제에서는 폭염, 강수량, 엘니뇨 지수 등 대표적인 기후 변수를 활용하였으나, 커피 생산에 치명적인 특정 기후 피해 현상을 충분히 반영하지 못한 한계가 있다.

예를 들어, 브라질 고산지대에서 발생하는 이른바 *검은 서리(Black Frost)*는 단순히 특정 임계 온도(예: -2°C 이하)만으로 발생 여부를 판단하기 어렵고, 습도, 바람, 지형 조건 등이 복합적으로 작용하는 현상이다.

그러나 본 과제에서는 일 단위 기온 및 간단한 폭염 지표만 활용하였기 때문에, 이와 같은 복합 기상 피해를 직접적으로 설명하는 Feature를 충분히 설계하지 못하였다.

-> 복합 기상 위험 지표를 설계 (습도 \times 기온 조합 지표)

나. 주요 생산국의 지역별 영향 미반영

본 모델은 브라질을 중심으로 한 기후 변수에 초점을 두었으나, **베트남(세계 커피 생산량 2위 국가)**을 포함한 다른 주요 생산국의 생산 환경은 반영하지 못하였다.

특히 베트남은

*로부스타 품종 중심

*강수 패턴 및 기후 조건이 브라질과 상이

함에도 불구하고, 해당 지역의 생산량 변화나 기후 리스크를 모델에 포함하지 못한 점은 글로벌 공급 측면에서의 설명력을 제한하는 요소로 작용하였다.

-> 베트남, 콜롬비아, 인도네시아 등 주요 생산국의 기후 및 생산량 데이터 추가

다. 월 단위 데이터 사용에 따른 정보 손실

본 연구는 월 단위 데이터 기반 분석으로, 일 단위 또는 주 단위에서 발생하는 급격한 기후 이벤트나 시장 충격을 충분히 반영하지 못하였다.

이로 인해 단기적인 가격 급변 국면에서는 모델 예측 성능이 제한될 가능성이 있다.

-> 일 단위 또는 주 단위 데이터로 확장

라 .비정형·정성적 요인의 미반영

커피 가격은 기후 요인 외에도 다음과 같은 비정형 요인의 영향을 크게 받는다.

*수출입 정책 변화

*지정학적 리스크

*물류 차질

*시장 심리 및 투기 수요

그러나 본 과제에서는 정량화 가능한 수치 데이터 중심으로 분석을 수행하여, 이러한 정성적 요인을 직접적으로 반영하지 못하였다.

->뉴스, 리포트, SNS 등 텍스트 데이터를 활용한 감성 분석(Sentiment Analysis)을 결합