S&P 500 지수 방향성 예측을 위한 경제지표 기반 분류 모델링

2팀(김도현, 김정운, 오현민, 이영섭, 장다향)

1. 프로젝트 개요

(1) 프로젝트 배경 및 목적

미국 S&P 500 지수는 글로벌 증시의 방향성과 투자 심리를 반영하는 핵심 지표이다. 본 프로젝트는 다양한 금융 자산과 거시경제 지표 데이터를 활용하여 S&P 500의 단기적 상승/하락 여부를 예측하는 분류 모델을 구축하는 것을 목표로 한다. 이를 통해 시장 변화에 대한 조기 인지와 데이터 기반의 투자 판단을 지원하고자 한다. 또한 경제 지표와 자산 간 상관관계를 분석함으로써 시장 구조에 대한 통찰을 도출할 수 있다.

(2) 예측 대상 정의

- 예측 대상
  - 미국 S&P 500 지수
- 예측 방식
  - 미래 일정 기간의 S&P 500 종가가 이전 종가 대비 상승/하락할지 여부를 예측
- 입력 변수
  - 10년(2015.01.01 ~ 2024.12.31) 간의 S&P500 지수
  - 금융 자산 가격: US 10년물 국채 수익률, 원유, 구리, 금, 천연가스, 옥수수, VIX 지수
  - 거시경제 지표: 소비자물가지수, 기준금리, 실업률, 소매판매, 기대 인플레이션, 개인소비지출
  - 기준 데이터 소스: yfinance, FRED API

(3) 기대 효과 및 활용 방안

S&P 500의 방향성이 어떤 변수들과 상관관계를 가지는지를 분석함으로써 경제지표와 시장 간의 상호작용에 대한 이해도를 증진시킬 수 있다. 또한 정형 데이터를 기반으로 한 예측 결과를 활용하여 투자 전략 수립 시 참고 가능한 근거 자료를 제공할 수 있다.

2. 데이터 수집 및 전처리

본 프로젝트에서는 S&P500 지수의 단기 상승 여부를 예측하기 위해 다양한 경제 지표와 시장 데이터를 통합한 시계열 데이터셋을 사용하였다. yfinance는 Python에서 직접 주식 및 금융 데이터를 불러올 수 있는 API 기반 라이브러리로, 주가, 금리, 원자재 등 실시간 금융시장 데이터를 손쉽게 조회할 수 있다. FRED는 미국 연방준비은행에서 제공하는 공식 경제지표 데이터 플랫폼으로, 신뢰도 높은 경제 통계를 제공한다. 수집한 모든 데이터는 날짜(Date)를 기준으로 병합하여 시계열 데이터프레임으로 구성했다. 해당 데이터는 다음과 같은 정보를 포함한다

금융 시장 데이터 (yfinance)

- S&P500 :미국의 주식시장에 상장된 500대 기업의 주가를 종합하여 산출하는 주가지수
- VIX :옵션시장 기반의 공포지수
- NASDAQ : 기술주 중심의 전자식 미국 주식시장
- DXY :미국 달러 인덱스
- US10Y :미국 10년 만기 국채 수익률
- US2Y: 미국 2년 만기 국채 수익
- WTI_Oil : 원유
- Copper : 구리
- Gold : 금
- NatGas : 천연가스
- Corn : 옥수수

거시경제 지표 (FRED API)

- CPI : 소비자물가지수
- Fed Funds Rate : 기준금리
- Unemployment Rate : 실업률
- Retail Sales : 소매판매
- Inflation Expectation 10Y : 10년 기대 인플레이션
- PCE : 개인소비지출

가격 정보는 주로 종가(Close) 기준으로 사용하며, 필요 시 변동률(Return), 이동평균, 변동성(표준편차) 등 기술 지표로 확장할 계획이다.

(1) 전처리 전략

- 날짜 변환: 해당 데이터는 .xlsx 포맷의 엑셀 파일로 수집되었으며, 날짜는 Excel 기준 일 수로 저장되어 있어 pandas.to_datetime() 함수를 이용해 표준 날짜 형식으로 변환하였다.
- 타깃 변수 정의: 예측 목표는 다음 거래일의 S&P500 지수가 0.2% 이상 상승하는지 여부다. 이를 위해 다음과 같은 방식으로 타깃 변수를 생성했다.
  df["Return"] = df["S&P500"].pct_change().shift(-1) # 다음날 수익률 계산
  df["Target"] = (df["Return"] > 0.002).astype(int) # 0.2% 초과 상승이면 1, 아니면 0
- 파생 변수 생성: 모델의 성능을 높이기 위해 아래와 같은 파생 변수를 추가로 생성하였다.

- 경제 지표 기반 변화율: 외부 경제 환경의 변화를 수치로 반영하여 시장 영향을 파악

변수명	설명
Inflation_change	기대 인플레이션의 전일 대비 변화율
Retail_change	소매판매 지표의 변화율
WTI_change	WTI 원유 가격의 변화율
PCE_change	개인소비지출 변화율
Gold_change	금값의 변화율
CPI_change	소비자물가지수 변화율
Unemp_change	실업률 변화율
Retail_change_lag1	소매판매 변화율의 하루 지연값
Unemp_change_lag1	실업률 변화율의 하루 지연값

- 기술적 분석 지표: S&P500 자체의 움직임과 가격 패턴을 수치화

변수명	설명
MA10	10일간 S&P500의 단순 이동평균
Above_MA10	현재 주가가 MA10 위인지 여부 (0/1)
Return_3d_avg	최근 3일간 평균 수익률
Momentum_5d	최근 5일간 누적 수익률
Volatility_5d	최근 5일간 수익률의 표준편차 (변동성)
MACD	12일 EMA - 26일 EMA (추세를 보여주는 기술적 지표)

- 파생 수익률 및 변동성 관련 변수: 수익률 변화나 변동성의 흐름 자체를 파악하기 위한 변수

변수명	설명
CumulativeReturn_3d	3일간 누적 수익률
Return_change_1d	수익률의 하루 간 변화량
Volatility_change	5일 변동성의 하루 간 변화량

- 상호작용 및 시장 심리 관련 변수: 자산 간 관계나 투자자 심리를 반영하는 논리적 변수

변수명	설명
SP_Gold_interaction	S&P500 수익률 × 금 변화율 (위험자산과 안전자산의 동시 반응)
Gold_up_while_SP_down	금 상승 + S&P500 하락 여부 (위험 회피 심리 지표)

- 결측치 처리: 파생 변수 생성 과정에서 발생한 NaN 값은 모두 제거하고, 인덱스를 재설정하였다.
- 정규화 및 샘플링: 입력 변수들은 StandardScaler를 이용해 정규화(표준화)하였고, 클래스 불균형을 해소하기 위해 SMOTE 오버샘플링 기법을 적용하였다.

3. 예측 대상 및 목표 변수 정의

본 프로젝트의 예측 대상은 미국 주식시장의 대표적인 지수인 S&P 500 지수의 단기적인 방향성(상승 또는 하락)이다. S&P 500은 시장 전반의 흐름을 반영하는 핵심 지표이기 때문에, 이를 예측하는 모델은 다양한 자산 운용 및 리스크 관리 전략에 실질적인 활용 가치를 가질 수 있다. 모델의 목표는 S&P 500 지수가 다음 거래일에 상승할지, 하락할지를 이진 분류 형태로 예측하는 것이다. 이를 위해, 본 프로젝트에서는 S&P 500의 일일 종가 데이터를 기반으로 하며, 이에 영향을 미칠 수 있는 국채금리(예: 10년물 수익률), 변동성 지수(VIX), 국제유가(WTI), 금, 구리 등의 원자재 가격, 그리고 CPI, 소비지출(PCE), 연방기금금리(Fed Funds Rate), 소매판매(Retail Sales) 등 주요 거시경제 지표를 함께 수집했다. 예측을 위한 목표 변수는 S&P 500의 향후 1일 수익률을 기준으로 정의된다. 현재 시점에서 다음 날 종가가 0.2%를 초과하여 상승했는지 여부를 기준으로 1(0.2% 초과 상승), 0(0.2% 이하 상승 또는 하락)으로 라벨링하여 이진 분류 모델의 타겟 변수로 사용한다.

4. 모델링 전략

(1) 사용 알고리즘 개요 (Logistic Regression, Random Forest)

이진 분류 문제 해결을 위해 로지스틱 회귀(Logistic Regression)와 랜덤 포레스트(Random Forest) 알고리즘을 주요 후보로 채택하였다.

- 로지스틱 회귀(Logistic Regression)는 독립 변수와 종속 변수 간의 선형적인 관계를 기반으로 하며, 시그모이드 함수를 활용해 결과를 0과 1 사이의 확률값으로 변환한 뒤 특정 기준에 따라 이진 분류를 수행한다.
    - 모델 구조가 단순하고 각 변수의 계수를 통해 해석이 가능하다는 장점을 지니며, 변수 간의 관계를 직관적으로 파악할 수 있어 설명 가능성이 중요한 상황에 적합하다.
    - 입력 변수와 결과 간의 관계가 비선형이거나 변수 간 상호작용이 복잡할 경우 성능이 저하될 수 있으며, 전처리와 변수 선택의 영향이 크다는 한계를 갖는다.

- 랜덤 포레스트(Random Forest)는 다수의 결정 트리를 기반으로 하는 앙상블 학습 알고리즘으로, 각 트리에 대해 무작위로 샘플링된 데이터와 선택된 피처를 사용하여 학습하고, 이를 종합하여 최종 예측을 도출한다.
    - 모델의 분산을 줄이고 과적합을 방지하는 데 효과적이며, 비선형 관계나 변수 간 복잡한 상호작용을 잘 포착할 수 있어 다양한 데이터 환경에서 높은 예측 성능을 보인다.
    - 변수 중요도를 제공함으로써 피처 선택과 해석에도 활용할 수 있지만, 트리 구조가 많아질수록 모델의 복잡도가 증가하고 해석력이 떨어질 수 있다.

(2) 변수**(feature)** 선택 및 엔지니어링 전략

변수 선택에 있어서는 모델의 성능을 높이기 위해 아래와 같이 파생 변수를 추가로 생성하였다.

- 금융 자산 및 경제 지표 기반 변화율: 외부 경제 환경의 변화를 수치로 반영하여 시장 영향을 파악

변수명	설명
Inflation_change	기대 인플레이션의 전일 대비 변화율
Retail_change	소매판매 지표의 변화율
WTI_change	WTI 원유 가격의 변화율
PCE_change	개인소비지출 변화율
Gold_change	금값의 변화율
CPI_change	소비자물가지수 변화율
Unemp_change	실업률 변화율
Retail_change_lag1	소매판매 변화율의 하루 지연값
Unemp_change_lag1	실업률 변화율의 하루 지연값

- 기술적 분석 지표: S&P500 자체의 움직임과 가격 패턴을 수치화

변수명	설명
MA10	10일간 S&P500의 단순 이동평균
Above_MA10	현재 주가가 MA10 위인지 여부 (0/1)
Return_3d_avg	최근 3일간 평균 수익률
Momentum_5d	최근 5일간 누적 수익률
Volatility_5d	최근 5일간 수익률의 표준편차 (변동성)
MACD	12일 EMA - 26일 EMA (추세를 보여주는 기술적 지표)

- 파생 수익률 및 변동성 관련 변수: 수익률 변화나 변동성의 흐름 자체를 파악하기 위한 변수

변수명	설명
CumulativeReturn_3d	3일간 누적 수익률
Return_change_1d	수익률의 하루 간 변화량
Volatility_change	5일 변동성의 하루 간 변화량

- 상호작용 및 시장 심리 관련 변수: 자산 간 관계나 투자자 심리를 반영하는 논리적 변수

변수명	설명
SP_Gold_interaction	S&P500 수익률 × 금 변화율 (위험자산과 안전자산의 동시 반응)
Gold_up_while_SP_down	금 상승 + S&P500 하락 여부 (위험 회피 심리 지표)

엔지니어링 전략으로는 다양한 하이퍼파라미터의 조합을 실험해보는 한편, 함수를 사용해 최적의 파라미터를 찾는 작업도 진행하였다. class_weight='balanced' 를 사용하여 모델이 상승, 하락 중 하나의 경우만 잘 예측하는 것을 방지하고자 하였다. 로지스틱 리그레션에서는 변수들간의 공선성이 있는 경우 모델 성능이 하락하므로 일정 상관계수 이상의 변수를 제거하거나, 중요도가 높은 변수만 선별하여 학습을 진행하는 등 다양한 방식을 취해보았다. 랜덤포레스트에서도 다양한 하이퍼파라미터의 조합을 실험해보는 한편, Feature Importance를 측정하여 중요도가 높은 변수를 선별하여 학습을 진행했다.

(3) 학습 데이터 **/** 테스트 데이터 분리 기준

학습 데이터와 테스트 데이터는 8:2 비율로 나누었다. 시계열 데이터임을 감안, shuffle=False 로 지정하여 시계열 구조를 유지하며 데이터를 분리했다. 학습과 테스트에 2015년 1월 1일부터 2024년 12월 31일까지 총 10년의 데이터를 사용하여, 학습에는 2015년부터 2022년까지 약 8년, 테스트에는 2023년부터 2024년까지 약 2년의 데이터를 사용하였다.

5. 모델 성능 평가

(1) 평가 지표

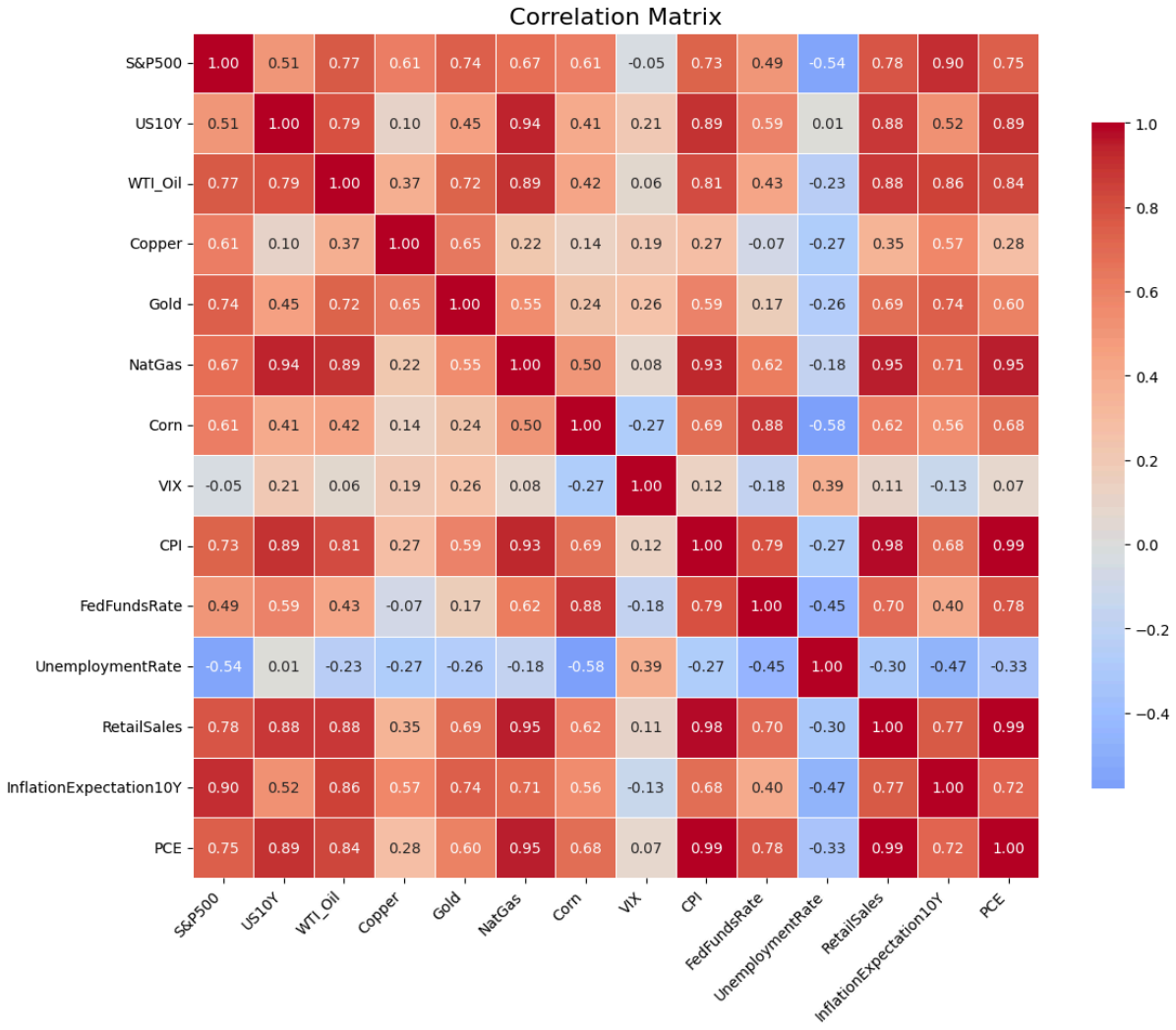
분류 모델의 성능을 정량적으로 평가하기 위해 **Accuracy, Precision, Recall, F1 Score** 네 가지 주요 지표를 사용하였다.

- **Accuracy** (정확도): 전체 데이터 중 모델이 맞춘 정답의 비율
- **Precision** (정밀도): 양성(주가 상승)이라고 예측한 것 중 실제로 양성인 비율
- **Recall** (재현율): 실제 양성(주가 상승) 중 모델이 양성이라고 예측한 비율
- **F1 Score**: Precision과 Recall의 균형을 고려한 종합적인 성능 지표

이 네 가지 지표 중, 전반적인 분류 성능을 평가하는 데 유용한 **F1 Score**를 향상시키는 데에 초점을 맞추어 모델 빌딩을 진행하였다.

(2) 모델 빌딩 - **Logistic Regression**

다양한 하이퍼파라미터의 조합을 실험해보는 한편, 함수를 사용해 최적의 파라미터를 찾는 작업도 진행하였다. 파라미터로 class_weight='balanced' 를 사용하여 모델이 상승, 하락 중 하나의 경우만 잘 예측하는 것을 방지하고자 하였다. 이러한 전략에도 불구, 초기 모델은 정확도 0.51 수준의 낮은 성능을 보였다. 로지스틱 리그레션에서는 변수들간의 공선성이 있는 경우 모델 성능이 하락하는데 아래 상관관계 도표에서 보이는 것과 같이 파생 변수를 생성하기 전의 변수들은 대부분이 서로 상관관계가 있는 것으로 나타났다. 이를 보완하기 위해 일정 상관계수 이상의 변수를 제거하거나, 중요도가 높은 변수만 선별하여 학습을 진행하는 등 다양한 방식을 취해보았으나 모델의 성능은 나아지지 않았다.



낮은 모델 성능을 피처 정의 단계에서의 문제로 파악하고 앞서 '변수 선택' 부분에서 기술한 것과 같이 파생변수를 생성하여 새로운 피처로 정의했다. 수정 전의 학습 데이터에서는 원자재 가격의 절대적인 수치를 변수로 사용했는데, 모델의 예측 목표인 S&P500 지수의 상승과 하락은 원자재 가격의 수치보다는 전일 대비 원자재 가격의 등락 여부가 더 영향을 미칠 것으로 판단하였다. 이에 원자재 가격의 전일 대비 상승, 하락 비율을 새로운 변수로 지정하였다. 또한 다양한 하이퍼파라미터 조합을 실험하여 아래와 같은 성능을 기록하였다.

항목	모델 1	모델 2
penalty	l1	l1
class_weight	balanced	balanced
solver	liblinear	liblinear
Feature	상관계수 0.9 이상 제거 후 8개만 사용	수정한 변수 중 상위 10개 사용
Accuracy	0.5090	0.7163
Precision	0.5100	0.7179
Recall	1.0000	0.6914
F1 Score	0.6700	0.7044

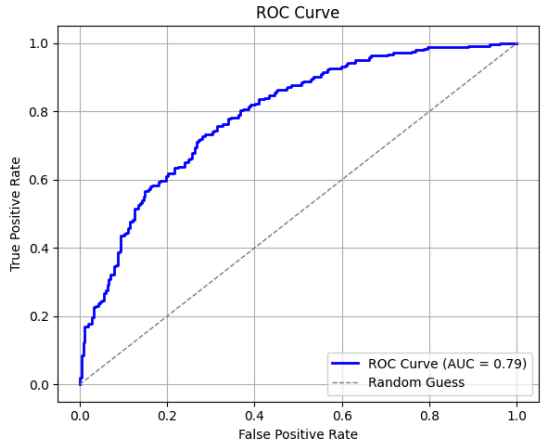
로지스틱 리그레션에서 가장 높은 성능을 보인 모델은 다음과 같은 하이퍼파라미터를 사용했다.

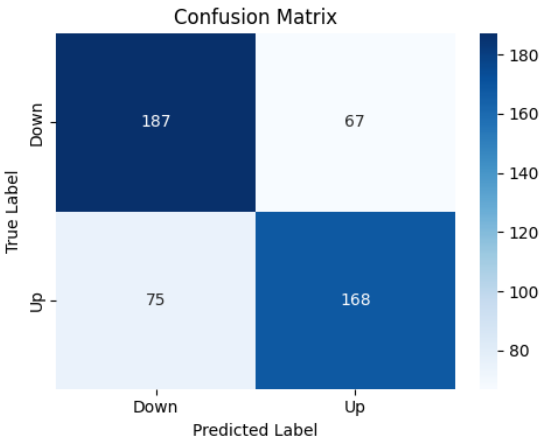
```
# 로지스틱 모델 정의
log_reg = LogisticRegression(penalty='l1', solver='liblinear', class_weight='balanced')
```

또한 RFE로 상위 10개의 변수만 선택해 학습에 사용했으며, 사용한 변수는 아래와 같다.

- 선택된 피처: ['Inflation_change', 'WTI_change', 'Gold_change', 'CPI_change', 'Unemp_change', 'Above_MA10', 'Return_3d_avg', 'Volatility_5d', 'Momentum_5d', 'Retail_change_lag1']

다음으로 ROC Curve와 Confusion Matrix는 아래와 같다. AUC(Area Under the Curve)는 0.79로 예측력이 어느 정도는 있는 것으로 판단된다.





다음으로 회귀계수의 값을 계산하였을 때, 각 변수의 회귀 계수는 아래와 같이 나타났다. 표에서 확인할 수 있는 것처럼 가장 높은 의미를 보인 계수는 'S&P 3일 평균 수익률'과 'S&P

5일 변동성’이다. 반면 많은 영향을 미칠 것으로 예측했던 전일 대비 국제 유가 변화율이나 금 가격 변화율은 S&P 관련 파생 변수에 비해 적은 영향을 미치는 것으로 나타났다.

변수	의미	회귀 계수	절대값
Return_3d_avg	S&P 3일 평균 수익률	3.18376	3.18376
Volatility_5d	S&P 5일 변동성	1.350204	1.350204
Momentum_5d	S&P 5일 누적 수익률	0.643461	0.643461
Above_MA10	이동평균 위인지 여부	-0.358155	0.358155
Unemp_change	전일 대비 실업률 변화율	0.326022	0.326022
Retail_change_lag1	소매판매 변화율의 1일 지연 값	0.258327	0.258327
Inflation_change	전일 대비 인플레이션 변화율	-0.173904	0.173904
CPI_change	전일 대비 소비자 물가 지수 변화율	0.142382	0.142382
Gold_change	전일 대비 국제 유가 변화율	-0.131342	0.131342
WTI_change	전일 대비 금 가격 변화율	-0.077152	0.077152

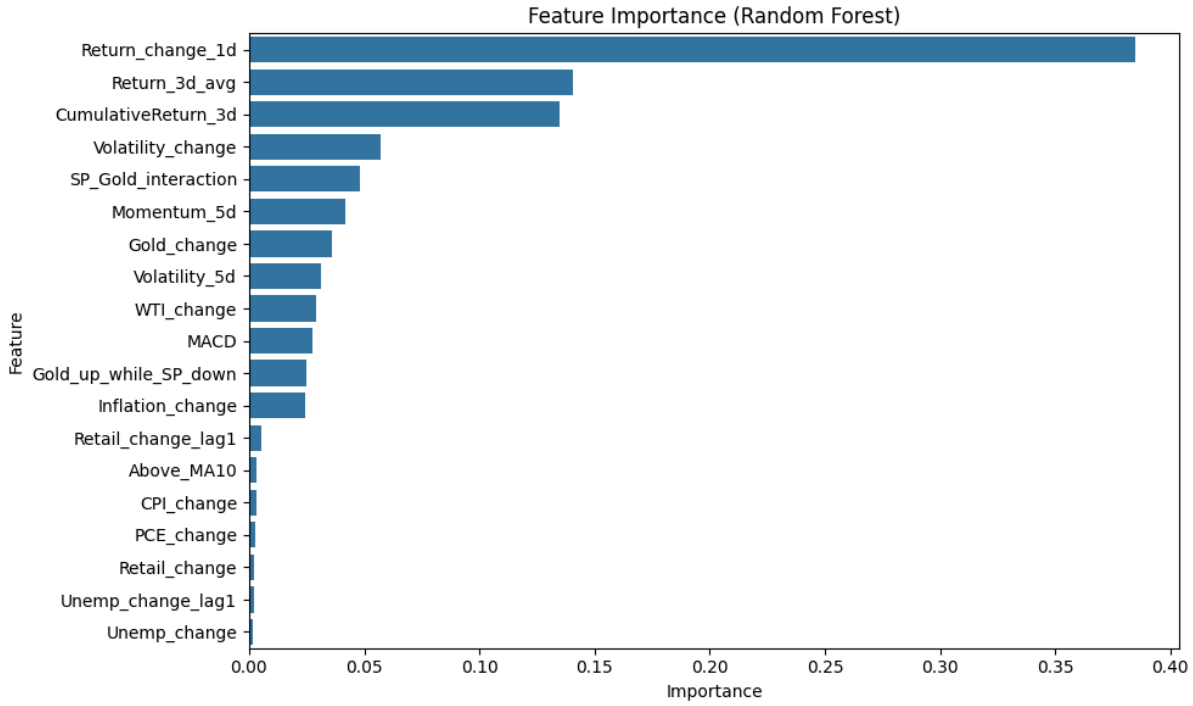
(3) 모델 빌딩 **- Random Forest**

본 프로젝트에서는 트리 기반 앙상블 모델인 Random Forest Classifier를 사용하여 S&P500 지수의 단기 상승 여부를 예측하였다. 초기에는 원본 피처만을 사용하여 모델을 학습시켰으며, class_weight='balanced' 파라미터를 적용해 클래스 불균형으로 인한 편향을 방지하였다. 또한, RandomizedSearchCV를 활용하여 주요 하이퍼파라미터(n_estimators, max_depth 등)를 최적화하였다. 초기 모델은 정확도 F1 Score 0.59 수준의 제한적인 성능을 보였다. 이는 단순 경제 지표만으로는 상승/하락의 복잡한 패턴을 충분히 설명하기 어렵기 때문으로 해석되었다. 이에 따라 시장 구조 및 심리, 가격 추세를 반영한 다양한 파생 변수를 추가하였다. 이동평균(MA), 변동성(Volatility), 모멘텀(Momentum), 수익률 변화(Return Change), 그리고 금 가격이나 안전자산과의 상호작용 지표 등을 포함한 새로운 피처들은 예측 정보량을 크게 증가시켰다. 그 결과, 모델 성능은 F1 Score 기준 약 0.87 이상으로 크게 향상되었으며, 이는 파생 변수의 도입이 예측 정확도 개선에 핵심적인 역할을 했음을 보여준다.

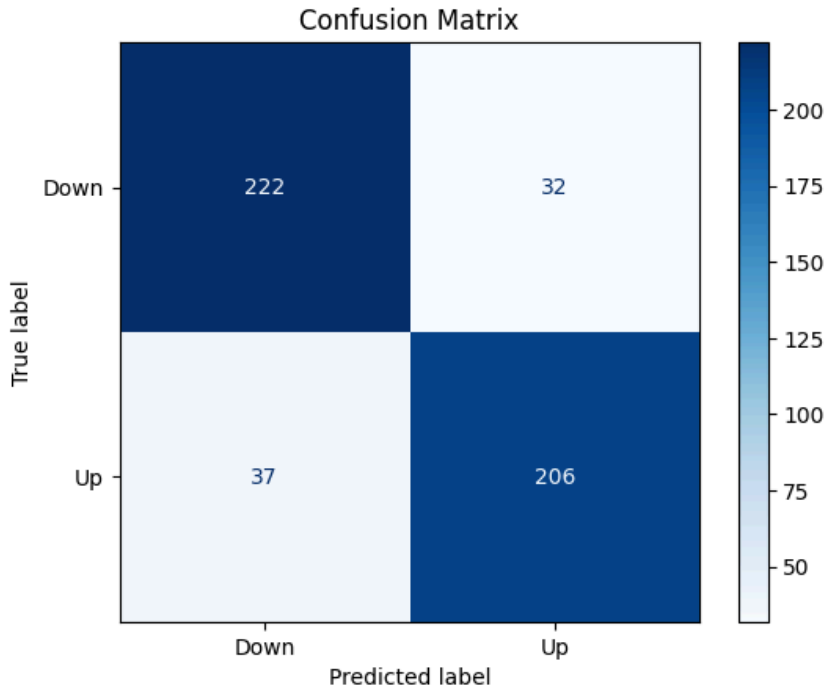
항목	기본 모델	확장 모델
n_estimators	500	2000
max_depth	10	20
random_state	42	42
min_samples_split	5	2
min_samples_leaf	5	1

max_features	—	None
class_weight	—	balanced
Accuracy	0.53	0.8753
Precision	0.53	0.8787
Recall	0.67	0.8642
F1 Score	0.59	0.8714

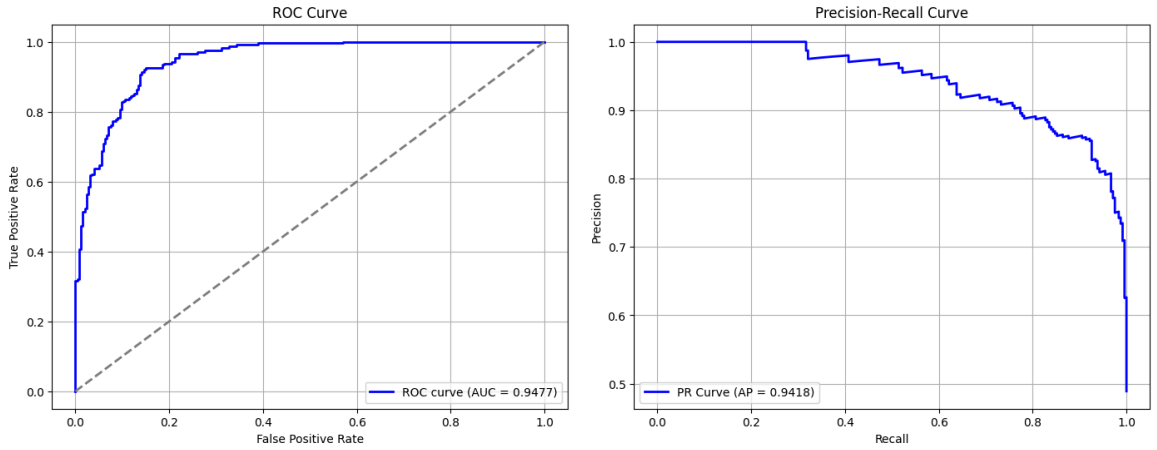
피처 중요도**(Feature Importance)**



혼동 행렬**(Confusion Matrix)**



ROC Curve / Precision-Recall Curve



→ ROC 곡선과 PR 곡선에서 모두 **0.94** 이상의 아주 높은 성능을 보인다.

6. 리스크 및 한계

(1) 데이터 품질 및 외부 요인(정책, 뉴스 등)의 영향

본 프로젝트는 공개 API(yfinance, FRED)를 통해 수집한 금융 자산 및 거시경제 데이터를 기반으로 모델을 학습하였다. 그러나 해당 데이터는 발표 시점의 시간차, 측정 방식의 변화, 지연 보정(revision) 등으로 인해 실제 시장 상황을 완벽하게 반영하지 못할 수 있다. 또한 금융시장은 금리 인상, 지정학적 위기, 정부 정책 발표 등과 같은 예기치 못한 외부 요인에 민감하게 반응하는데, 본 모델은 이러한 질적 데이터(정책 발표, 시장 뉴스 등)를 고려하지 않으므로 예측력에 한계가 존재한다.

(2) 과적합 가능성 및 일반화 이슈

Random Forest는 비선형 데이터에 강건하며 예측 성능이 우수한 장점이 있으나, 트리 기반 알고리즘 특성상 학습 데이터에 과적합(overfitting)될 위험이 있다. 특히 수많은 파생 변수와 시차(lag) 변수를 생성함에 따라 변수 간 상관관계가 복잡해지고, 이는 테스트셋에서는 좋은 성능을 보이더라도 실제 신규 데이터에서는 일반화 성능이 저하될 수 있다. 이를 방지하기 위해 교차 검증과 하이퍼파라미터 튜닝을 적용하였으나, 여전히 일반화 성능에 대한 지속적인 검증이 필요하다.

(3) 예측 불확실성에 대한 설명 방안

본 프로젝트는 분류(classification) 결과만을 출력하므로, 투자 판단 시 중요한 예측의 확신도(confidence) 또는 불확실성 수준에 대한 정보를 직접적으로 제공하지 않는다. 향후 확률 기반 출력(probability output)을 도입하여 모델의 판단 신뢰도를 함께 제공하거나, 예측 결과에 대한 Shapley value 또는 Feature Importance 기반 해석 모델을 병행함으로써 사용자에게 예측의 타당성과 해석 가능성을 높일 수 있는 보완 방안이 필요하다.