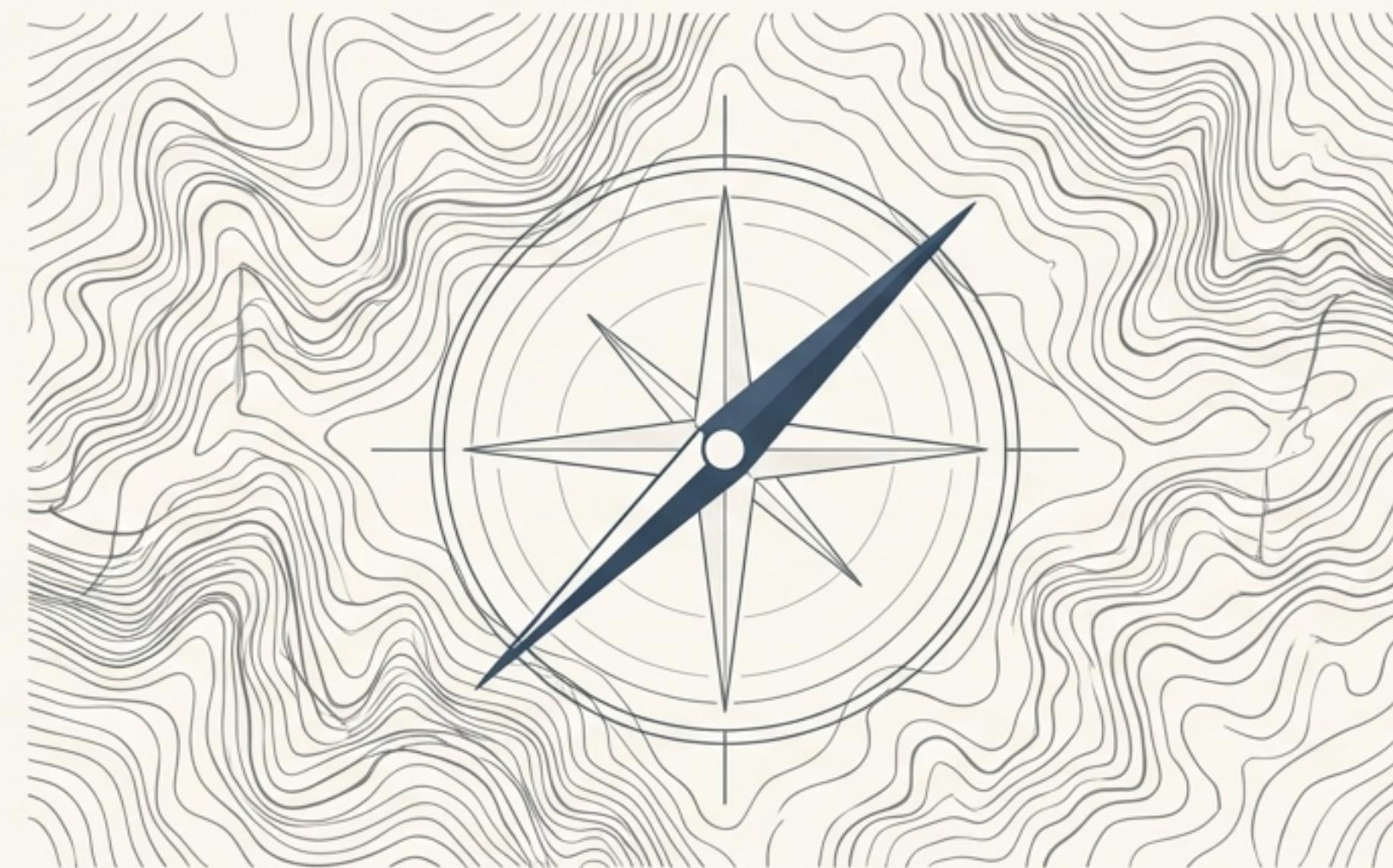


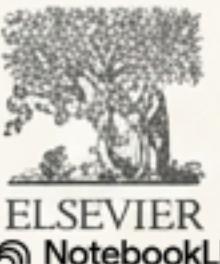
새로운 나침반: 머신러닝은 주식 시장의 방향성을 예측할 수 있는가?

Giovanni Campisi et al. (2024)의 연구 '휘발성 지수를 기반으로 한 미국 주식 시장 방향 예측을 위한 머신러닝 방법 비교' 핵심 요약 및 분석



International Journal
of Forecasting

f



시장 예측은 금융 분야의 오래된 난제입니다

- * 주식 시장의 방향을 예측하는 것은 자산 가격 책정 및 리스크 관리에 있어 매우 중요한 문제입니다.
- * 전통적으로 이 문제는 선형 회귀 분석과 같은 통계적 모델에 의존해 왔습니다.
- * 하지만 시장의 복잡성과 비선형적 특성은 이러한 전통적 방법론의 예측력을 제한합니다.

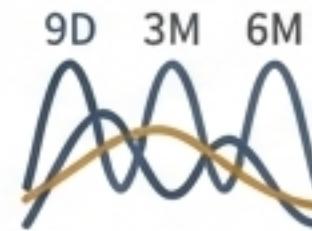
핵심 질문: 대량의 데이터를 처리하고 변수 간의 복잡한 관계를 학습할 수 있는 머신러닝(ML) 방법론이 더 우수한 예측 성과를 보일 수 있을까요?



예측을 위한 11가지의 휘발성 및 리스크 지표를 활용합니다

**VIX**

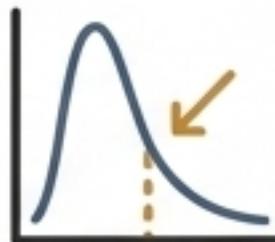
S&P 500의 30일 예상
변동성 ("시장의 공포 지수")

**VIX9D, VIX3M, VIX6M**

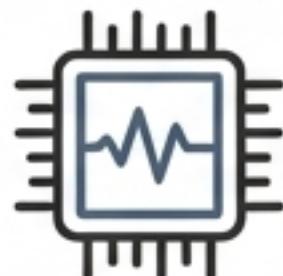
단기(9일), 중기(3개월),
장기(6개월) 예상 변동성

**VVIX**

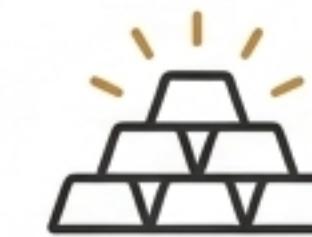
VIX 자체의 변동성
("변동성의 변동성")

**SKEW**

S&P 500의 30일 수익률
왜도 ("테일 리스크" 지표)

**VXN**

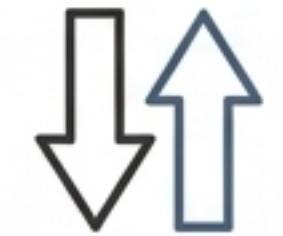
NASDAQ-100의 30일
예상 변동성

**GVZ**

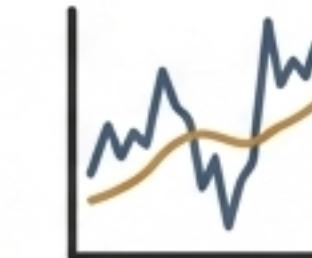
금(Gold ETF) 가격의
30일 예상 변동성

**OVX**

원유(Crude Oil ETF)
가격의 30일 예상 변동성

**PUTCALL**

S&P 500 지수 풋/콜
옵션 비율

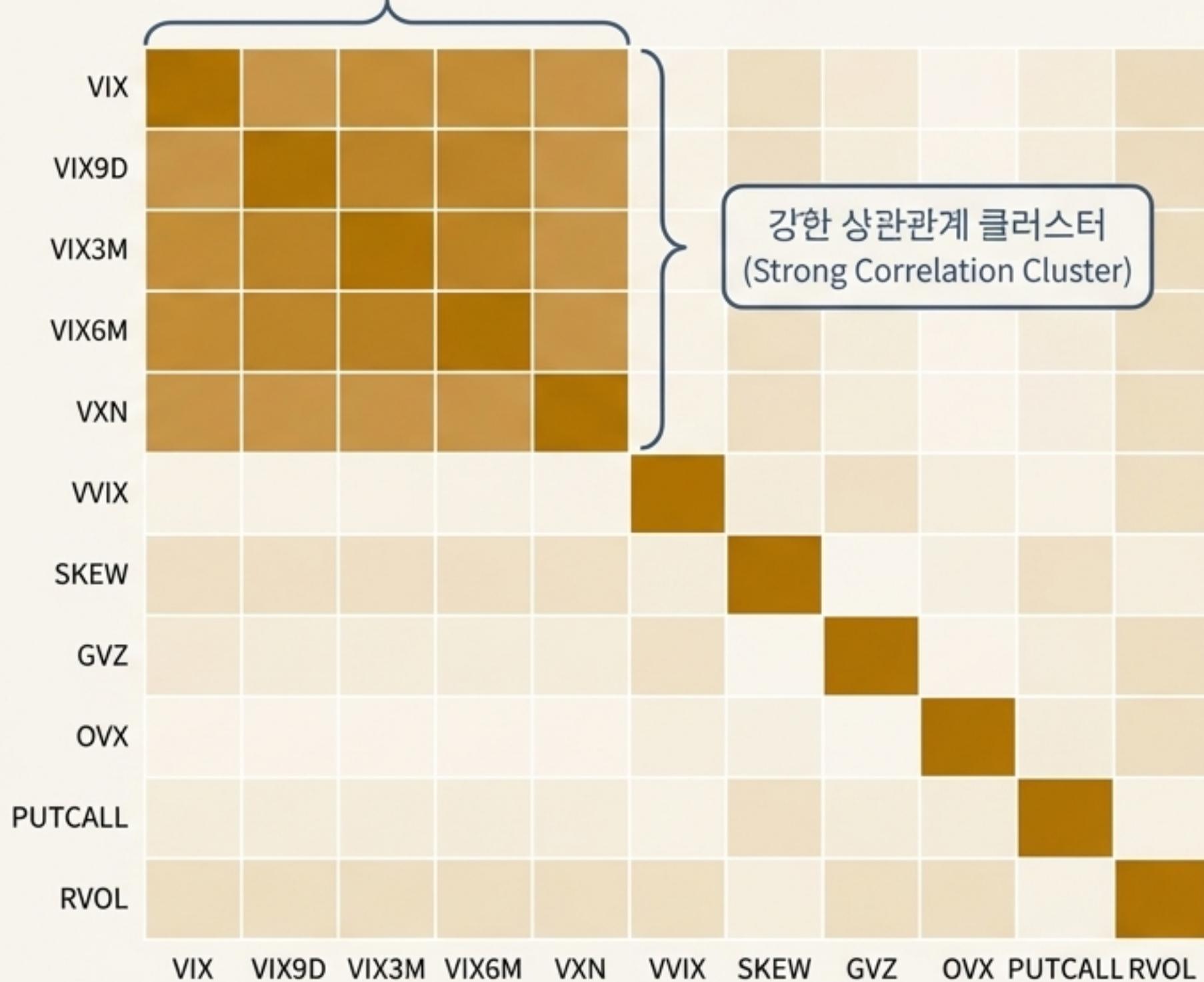
**RVOL**

과거 30일간의 실현 변동성

Source Reference: 데이터는 2011년 1월부터 2022년 7월까지의 3,040개 일일 관측치를 포함합니다.

다수의 예측 변수 간에는 높은 상관관계가 존재합니다

- * 분석에 사용된 많은 휘발성 지수들은 서로 강한 양의 상관관계를 보입니다. 특히 VIX, VIX9D, VIX3M, VIX6M, VXN 등 기간이 다른 변동성 지수들 간의 상관관계가 매우 높습니다.
- * 이러한 다중공선성(Multicollinearity)은 모델의 안정성을 저해하고 개별 변수의 영향력을 해석하기 어렵게 만듭니다. 이는 정교한 변수 선택 과정의 필요성을 시사합니다.



새로운 도전자들: 머신러닝 모델 라인업

본 연구는 전통적인 선형 회귀 분석을 벤치마크로 삼아, 다양한 머신러닝 모델들의 예측력을 비교 분석합니다. 모델은 크게 세 가지 범주로 나뉩니다.

분류 (Classification)

시장의 '방향'(상승/하락)을 직접 예측

- 로지스틱 회귀 (Logistic Regression)
- 선형 판별 분석 (Linear Discriminant Analysis - LDA)

회귀 (Regression)

시장의 '수익률'을 예측한 후 방향으로 변환

- 라쏘 회귀 (Lasso Regression)
- 릿지 회귀 (Ridge Regression)

앙상블 (Ensemble)

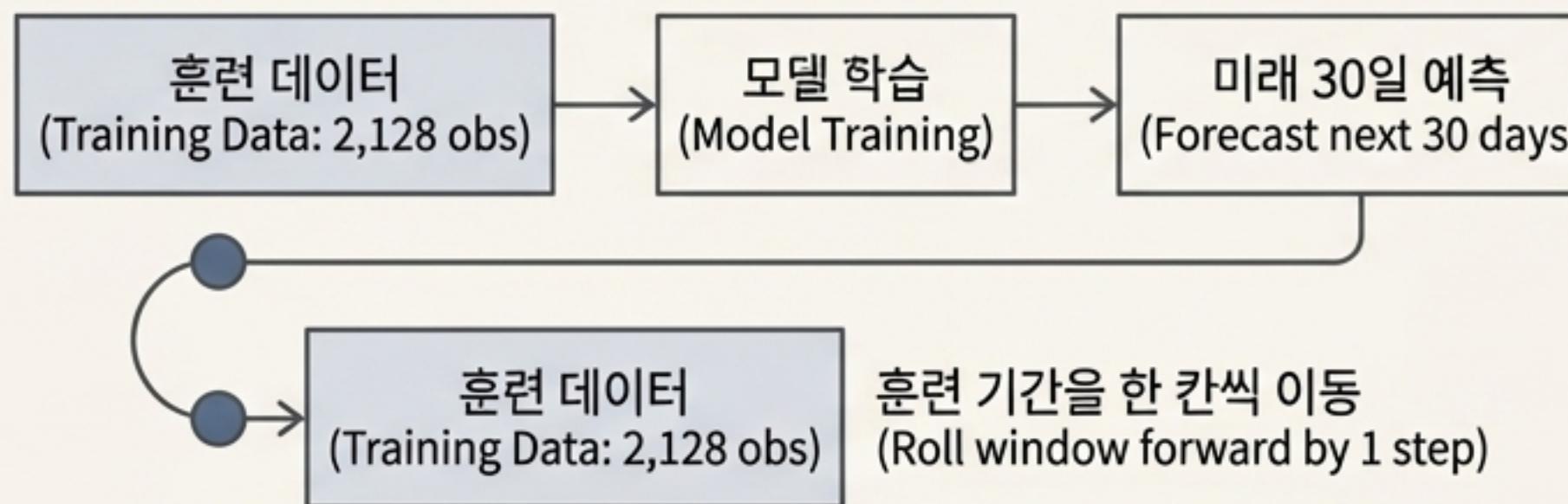
여러 개의 약한 모델을 결합하여 강력한 예측 모델 생성

- 배깅 (Bagging)
- 랜덤 포레스트 (Random Forest)
- 그래디언트 부스팅 (Gradient Boosting)

공정한 비교를 위한 엄격한 테스트 프레임워크를 설정합니다

모델의 예측 성과를 객관적으로 평가하기 위해 시계열 데이터에 적합한 검증 방법을 사용합니다.

검증 방법: Walk-Forward Validation



이 과정을 반복하여 총 883개의 예측치를 생성합니다.

(Repeat to generate 883 total forecasts)

이 방법은 미래 데이터를 사용하여 과거를 예측하는 '미래예측오류(Forward Looking)'를 방지합니다.

성과 측정 지표

정확도 (Accuracy)

전체 예측 중 방향을 올바르게 맞춘 비율

AUC (Area Under the Curve)

모델이 양성(상승)과 음성(하락)을 얼마나 잘 구별하는지에 대한 척도

F-measure

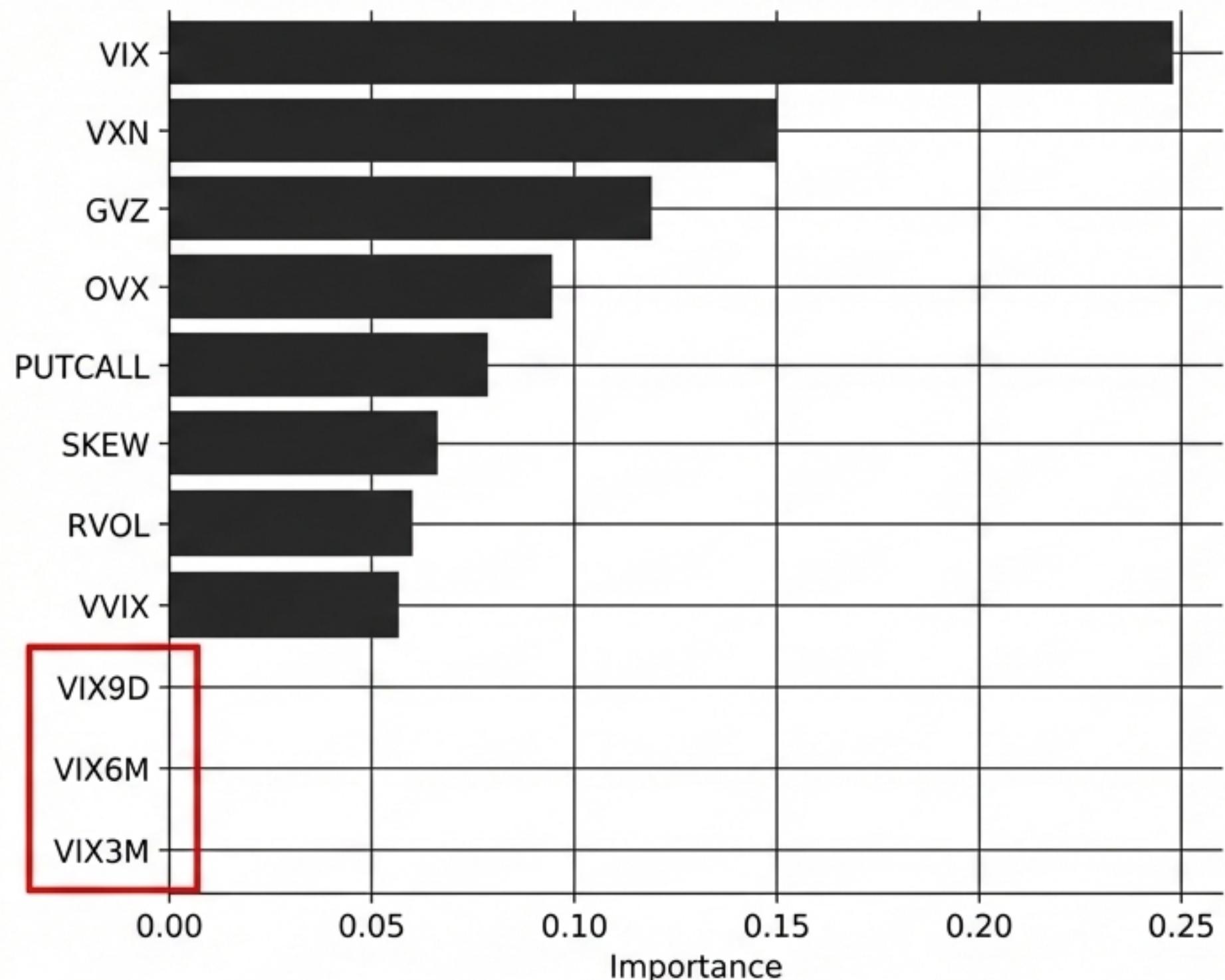
정밀도(Precision)와 재현율(Recall)의 조화 평균

핵심 신호와 노이즈를 분리합니다: 라쏘 회귀를 통한 변수 선택

모델의 성능을 개선하고 다중공선성 문제를 해결하기 위해,
라쏘(Lasso) 회귀를 사용하여 변수 선택을 수행했습니다.

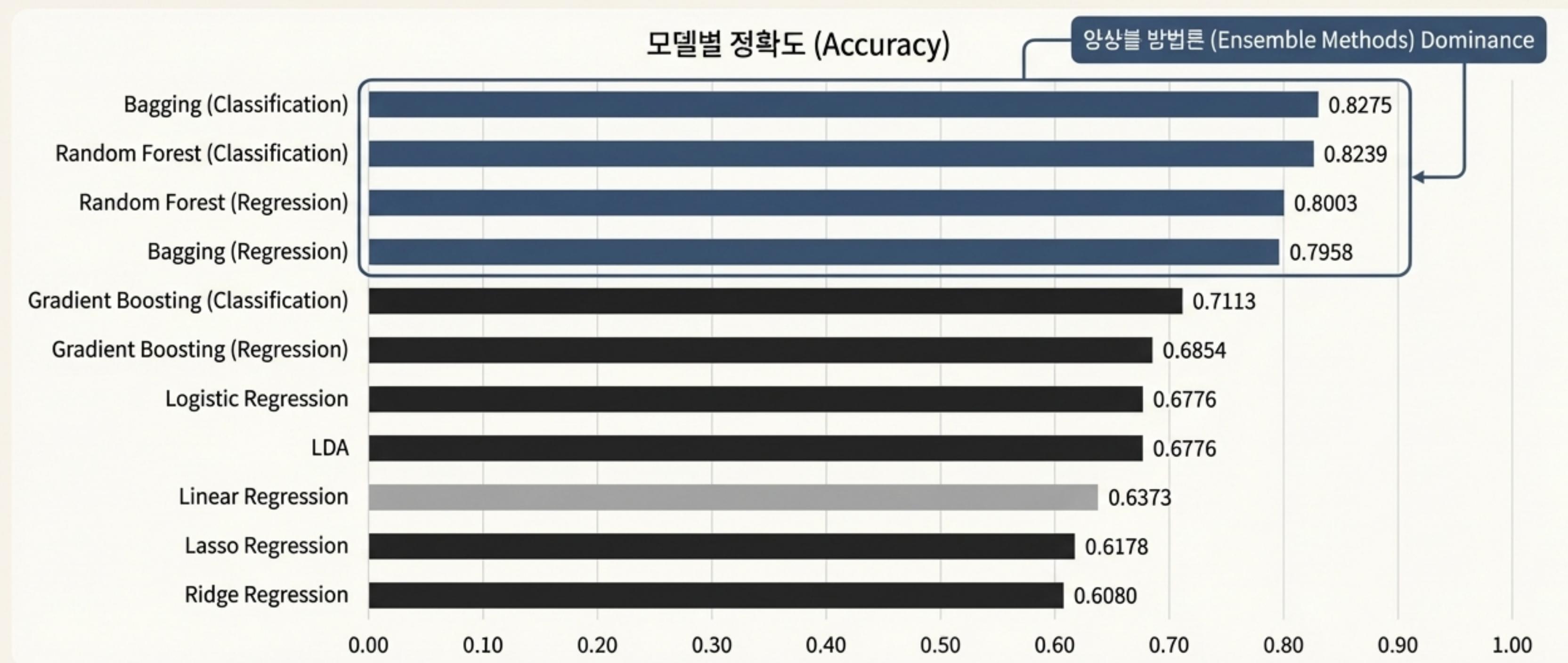
라쏘 회귀는 중요도가 낮은 변수의 계수를 정확히 0으로
만들어, 가장 영향력 있는 예측 변수들만 식별해냅니다.

분석 결과, **VIX9D**, **VIX3M**, **VIX6M** 세 가지 변수는 모델의
예측력에 거의 기여하지 못하는 것으로 나타나 최종 분석
에서 제외되었습니다.



최종 결과: 앙상블 모델이 전통적 방법을 압도합니다

변수 선택 후, 모든 모델의 예측 성과를 비교한 결과, 앙상블 방법론, 특히 랜덤 포레스트와 배깅이 가장 뛰어난 성과를 보였습니다. 이는 머신러닝 모델이 전통적인 선형 회귀 분석보다 S&P 500의 방향성을 예측하는 데 더 효과적임을 명확히 보여줍니다.



세부 분석: ‘방향’ 예측(분류)이 ‘수익률’ 예측(회귀)보다 효과적입니다

전반적으로 동일한 양상을 기법(랜덤 포레스트, 배깅) 내에서는 수익률을 예측한 후 방향으로 변환하는 회귀 모델보다, 직접적으로 상승/하락 방향을 예측하는 분류 모델이 더 높은 성과(정확도 및 F-measure)를 기록했습니다.

이는 시장의 정확한 수익률 수치를 예측하는 것보다 방향성을 맞추는 것이 더 실용적이고 달성 가능한 목표일 수 있음을 시사합니다.

분류 모델 성과 (Classification Model Performance)

모델	ACC	F-measure
Logistic	0.6776	0.8076
LDA	0.6776	0.8076
RF	0.8239	0.8828
Bagging	0.8275	0.8845
GB	0.7113	0.8215

회귀 모델 성과 (Regression Model Performance)

모델	ACC	F-measure
Linear	0.6373	0.7614
RF	0.8003	0.8592
Bagging	0.7958	0.8500
GB	0.6854	0.7694
Ridge	0.6080	0.7212
Lasso	0.6178	0.7465

판정: 모델 간의 성과 차이는 통계적으로 유의미합니다

모델 간의 예측 정확도 차이가 통계적으로 유의미한지 확인하기 위해 Diebold-Mariano(DM) 테스트를 실시했습니다.

[랜덤 포레스트 ≈ 배깅]



[그래디언트 부스팅]



[선형 회귀]



[라쏘 회귀 ≈ 릿지 회귀]

핵심 결과:

- 랜덤 포레스트(RF)와 배깅(Bagging)은 서로 통계적으로 구별할 수 없는 최상위 성능을 보였습니다.
- 이 두 모델은 선형 회귀, 라쏘 회귀, 릿지 회귀 등 다른 모든 모델보다 통계적으로 유의미하게 우수한 예측력을 가졌습니다.
- 선형 회귀는 라쏘 및 릿지 회귀보다 우수한 성과를 보였습니다.

결론: 시장의 불확실성을 항해하는 더 나은 도구

본 연구는 휘발성 지수를 활용한 주식 시장 방향 예측에 있어 머신러닝의 우수성을 명확히 입증합니다.

1. 양상을 모델의 명백한 우위

랜덤 포레스트와 배깅은 전통적인 선형 회귀 분석을 포함한 다른 모든 모델을 통계적으로 유의미하게 능가하는 예측 정확도를 보였습니다.

2. 변수 선택의 중요성

라쏘 회귀를 통한 체계적인 변수 선택은 다중공선성 문제를 해결하고 모델 성능을 향상시키는데 결정적인 역할을 했습니다.

3. 방향성 예측의 실효성

수익률 자체를 예측하기보다 상승/하락의 '방향'을 직접 예측하는 분류 모델이 더 높은 성과를 보이는 경향이 있었습니다. 이는 시장 타이밍 전략에 있어 중요한 시사점을 제공합니다.

부록 1: 전체 예측 변수 상세 설명

VIX	CBOE 변동성 지수, S&P 500의 30일 예상 변동성
VIX9D	CBOE S&P 500 9일 변동성 지수
VIX3M	CBOE 3개월 변동성 지수
VIX6M	CBOE S&P 500 6개월 변동성 지수
VVIX	VIX의 30일 선도 가격에 대한 예상 변동성
SKEW	S&P 500 수익률의 30일 왜도 추정치
RVOL	과거 30일간의 일별 S&P 500 로그 수익률 표준편차
VXN	CBOE NASDAQ-100 변동성 지수
GVZ	CBOE 금 ETF 변동성 지수
OVX	CBOE 원유 ETF 변동성 지수
PUTCALL	S&P 500 지수의 풋/콜 옵션 비율

부록 2: 모델 튜닝에 사용된 하이퍼파라미터

모델 (Model)	R 패키지 (R-package)	튜닝 파라미터 (Tuning parameters)	파라미터 값 (Values)
Lasso Regression	glmnet	λ ($\alpha=1$)	$\lambda \in [10^{-4}, 10^2]$
Ridge Regression	glmnet	λ ($\alpha=0$)	$\lambda \in [10^{-4}, 10^2]$
Random Forest	randomForest	mtry, ntree	$mtry=\sqrt{6}$, $ntree=500$
Bagging	randomForest	mtry, ntree	$mtry=6$, $ntree=500$
Gradient Boosting	caret	ntree, shrinkage, etc.	$ntree \in [50, 150]$, $shrinkage \in [0.01, 0.1], \dots$

*(Note: Logistic Regression and LDA require no tuning.)