

Data-Driven Equity

Data Science Approaches to Stock Trend Prediction

김도현 서인선 양동영 윤진영

프로젝트 개요

01

목적



주식의 추세를 예측하는
예측모델 개발

02

분석대상



KOSPI INDEX
(2014.01.01~2024.05.21)

03

수집 데이터



가격, 거래량
주체별 거래량
거시경제지표
뉴스심리지수

04

데이터 라벨링



Trend-Scanning Labeling /
(로페즈 데 프라도, 2018)
Variant Trend-Scanning Labeling

프로젝트 개요

05

EDA



데이터 종류 별
탐색적 데이터 분석 수행

06

Modeling



Baseline Model : LSTM
성능개선 Model : LightGBM, XGBoost

07

모델의 활용성과 한계



모델의 특성과 한계
추후 요구되는 과제

1

문제 정의 & 데이터 구성



1. 문제 정의와 데이터 구성

1. 문제 정의

코스피 지수의 추세를 예측하는 머신러닝 모델 개발

2. 대상 데이터

2014년 1월 ~ 2024년 5월 / 전체 KOSPI INDEX 가격 & 거래량 데이터

3. 학습 방법

머신러닝 지도학습 방법론

4. 예측 변수

기술적 지표, 거래량 지표, 주체별 거래량, 주체별 거래량 파생변수,
미국시장 지표, 거시경제 지표, 뉴스지수, 심리지수

4. 예측 변수

기술적 지표 상세

거래 추세 지표 (Trend Index) (1)

선정 이유: 가격의 장단기 추세를 파악하여 시장 방향성 예측

- MACD: 단기 및 장기 이동평균 차이로 추세 판단
- ADX: 가격 추세의 강도 측정
- TRIX: 단기와 장기 추세 동시 확인
- Mass Index(MI): 추세 반전 시점 포착



거래량 지표 (Volume Index)

선정 이유: 거래량과 주가의 상관관계를 분석하여 투자자 행동 예측

- CMF: 자금의 유입/유출을 파악 ($CMF > 0$: 자금 유입, $CMF < 0$: 자금 유출)
- FI: 주가 변화율과 거래량의 상호작용 측정
- MFI: 가격과 거래량을 결합한 모멘텀 지표
($MFI \geq 80$: 과매수, $MFI \leq 20$: 과매도)
- VPT: 가격 변화와 거래량 변화의 상관관계 측정
(VPT 증가: 매수 압력, VPT 감소: 매도 압력)
- EMV: 가격 변동성과 거래량 고려하여 주가 움직임 용이성 측정
($EMV > 0$: 주가 상승 용이, $EMV < 0$: 주가 하락 용이)

변동성 지표 (Volatility Index)

선정 이유: 시장 변동성 평가로 투자 위험성 분석

- ATR: 주가 변동폭 측정
- UI: 가격 하락 위험 측정
- BB: 주가 변동성 측정하여 상한선, 중심선, 하한선 표시
(상한선에 가까워지면 과매수, 하한선에 가까워지면 과매도)
- KC: 주가 변동성 측정하여 이동평균 중심으로
상한과 하한선 표시(ATR 사용)
- DC: 주가 최고가, 최저가 기준 상한과 하한 표시
(상한선 돌파: 매수 신호, 하한선 돌파: 매도 신호)

4. 예측 변수

기술적 지표 상세

추가 거래 추세 지표 (Trend Index) (2)

선정 이유: 가격의 장단기 추세를 파악하여 시장 방향성 예측

- Aroon : 주가 최고가, 최저가 도달 빈도 측정하여 교차점에서 매도, 매수 신호 확인
- Parabolic SAR: 주가 추세 따라 추세 전환시점 (< 주가: 매수, > 주가: 매도)
- Ichimoku Cloud:
주가 추세, 지지선 및 저항선 분석 지표로 종합적 시장상태 제공

거래량 변화율

- 5일, 10일, 20일 거래량 변화율
(vol_change_5, vol_change_10, vol_change_20):
주어진 기간 동안 거래량의 퍼센트 변화
- 5일, 10일, 20일 수익률
(ret_5, ret_10, ret_20): 주어진 기간 동안 종가의 퍼센트 변화

추가 변동성 지표 (2)

- 30일 종가 표준편차 (std_30): 30일 동안의 종가 변동성
- 30일 거래량 표준편차 (vol_std_30): 30일 동안의 거래량 변동성

모멘텀 지표 (Momentum Index)

선정 이유: 가격 상승/하락 모멘텀 측정으로 주가의 힘 평가

- RSI
가격의 상대강도 측정
- Williams %R
최근 가격 위치를 백분율로 표시 (과매수/과매도 식별)
- Stochastic Oscillator
주가가 특정 기간 동안 최고가와 최저가 범위 내에서 현재 위치 표시
- TSI
주가 모멘텀 측정하여 추세 방향과 강도
(+ → - : 매도 신호, - → + : 매수 신호)
- ROC
일정기간 동안 주가 변화율 측정
- AO
주가의 중기와 장기 이동평균선 차이 이용하여 시장 강도 측정
- CCI
주가 평균 가격으로부터 차이 측정
($CCI > 100$: 과매수, $CCI < -100$: 과매도)
- CMO
주가 상승, 하락 비교하여 모멘텀 측정
($CMO > 0$: 상승 모멘텀, $CMO < 0$: 하락 모멘텀)

4. 예측 변수

투자자별 거래량 데이터

- 투자자별 거래량**
- 데이터 수집: PyKRX 활용(출처: 한국거래소 정보시스템) 일 단위 주체별 거래량 데이터 약 10년치(2014.~2024.) 크롤링
 - 투자 주체 거래 실적과 그 변동 패턴을 함께 고려하여, 하나의 거래량 지표로 볼 때보다 시장의 전반적 움직임 더 정확히 파악
 - 주요 투자 주체: 개인 투자자, 기관 투자자(금융투자, 보험, 투신, 은행, 기타 금융, 연기금 등, 기타 법인), 외국인 투자자

투자자별 거래량

- 일중강도, 기관/개인 매수 비율, 매도 비율, 외국인 체결 강도, 매수 주요 그룹 합계, 매도 주요그룹 합계
- 대규모 자금을 운용하는 주체에서 시장 주요 변동 요인 및 코스피 시장에서 외국인 투자 동향 및 시장 참여 정도와 영향 포착

파생 변수

변수명	수식	변수명	수식
금융투자_매수/매도 비율	매수 금융투자 / 매도 금융투자	외국인 체결 강도	매수 외국인 / 매도 외국인
보험_매수/매도 비율	매수 보험 / 매도 보험	순매수 주요 그룹 합계	순매수 금융투자 + 순매수 보험 + 순매수 투신 + 순매수 사모 + 순매수 연기금 등
개인 활동성	(매도 개인 + 매수 개인) / 2	매수 주요 그룹 합계	매수 금융투자 + 매수 보험 + 매수 투신 + 매수 사모 + 매수 연기금 등
기관/개인 매수 비율	매수 기관합계 / 매수 개인	매도 주요 그룹 합계	매도 금융투자 + 매도 보험 + 매도 투신 + 매도 사모 + 매도 연기금 등
기관/개인 매도 비율	매도 기관합계 / 매도 개인		

4. 예측 변수

거시 경제 지표

미국시장 지표 (DJI, S&P500, NASDAQ)

- 1997년 외환위기 이후, 국내 주식시장이 외국인에게 전면 개방됨에 따라 국내 주가와 미국 주가의 연동이 강해짐
- 2000년 이후 외국인투자가 급격히 늘면서 **양의 상관관계**를 가지며, 상관계수가 0.6 이상으로 높아짐

뉴스심리지수 (NSI, News Sentiment Index | 기준 : 100)

- 한국은행에서 2022년부터 신규로 발표하고 있는 지표로, 경제기사 텍스트를 웹크롤링 기법으로 수집하고 자연어 처리로 분석
- 월별 뉴스심리지수는 주요 경제심리지표 및 실물경기지표에 1~2개월 선행하며 높은 **양의 상관관계**를 보임

무역수지

- 국경에서 통관하는 물품의 수출액과 수입액의 차이를 집계하는 지표로, 무역수지와 주식시장은 **양의 상관관계**를 가짐
- 한국의 경제구조가 제조업을 기반 대외의존적인 모습을 띠고 있으므로, 무역수지 데이터는 매우 중요한 지표

시중통화량 (M2)

- 중앙은행은 물가안정, 경기 부양 등을 위해 화폐량을 조절하며, 통화량이 많으면 주식시장에 자금이 유입되어 주가 상승
- 10년치 데이터 (14~24년) 기준, 코스피 인덱스와 0.67 수준의 **양의 상관관계**를 가짐

* M2(광의통화, 평잔) : 경제 내 유통되는 화폐 양을 측정하는 대표적인 척도로, 현금, 예금취급기관의 결제성예금, 저축성예금, 시장형/실적배당형 금융상품, 금융채 등 포함

환율, 금리

- 일반적으로 원·달러 환율과 금리 vs 주가는 **음의 상관관계**를 가지고, 외화부채 평잔, 외국인투자자 수급 등 양방향에 영향을 미침
- 환율 상승 시, 수출기업의 수출 경쟁력 강화 / 환율 하락 시, 외국인 투자자금이 쉽게 빠져나가지 못해 주가에 호재로 작용
- 금리 인상 시, 자금조달 비용 증가, 외화부채 평잔 증가, 시중 예금금리 인상으로 인한 주식 매력도 하락 등의 이유로 주가에 악재
- 단순 금리/ 장단기금리차는 코스피 주가와 연관성이 생각보다 미미하였으며, 시장 변동성과 유동성에 대한 참고사항으로만 활용

1. 문제 정의와 데이터 구성

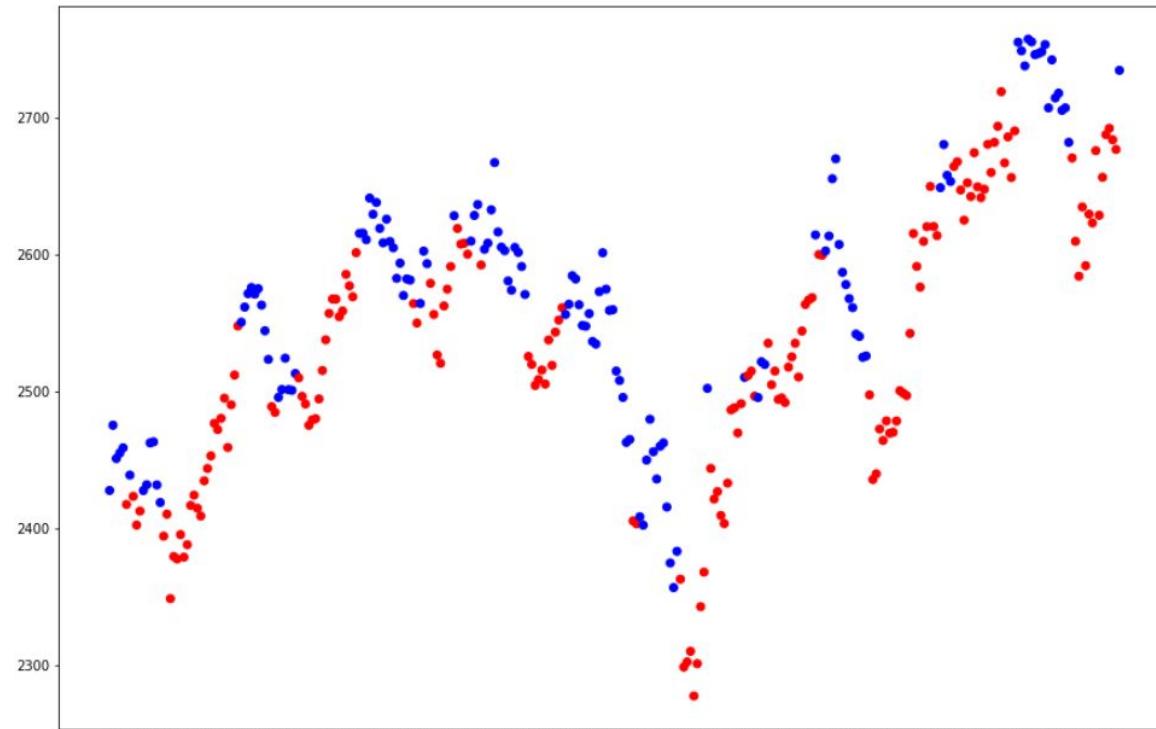
5. 반응 변수

지도학습에 필요한 반응변수를 정의하기 위해 Trend-Scanning 기법 활용

- 1) 'Advances in Financial Machine Learning(로페즈 데 프라도, 2018)'에서 소개
- 2) 주식 가격 시계열에서 현재 시점으로부터 미래 n일에 해당하는 데이터 포인트에 대해 선형회귀를 사용하여 기울기(베타)를 구함
- 3) 각 베타값 중 절댓값이 가장 큰 t-value의 부호에 따라 현재 시점의 라벨을 1 또는 -1로 정의

1. 문제 정의와 데이터 구성

Trend-Scanning을 활용한 코스피 지수 라벨링 결과 (최근 300일)



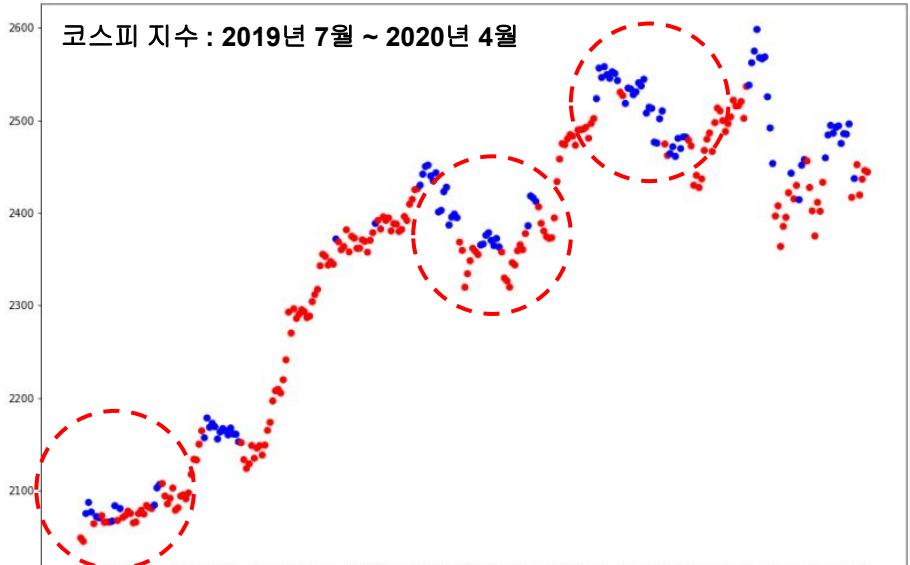
1. 문제 정의와 데이터 구성

6. 추가적 시도

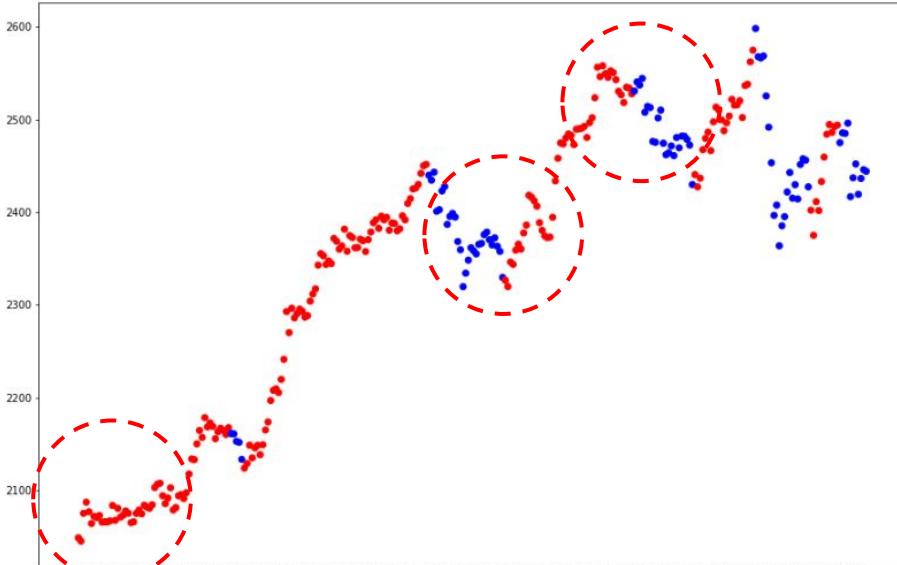
1. 주식 가격 시계열에 Trend-Scanning 방법을 있는 그대로 적용하면 라벨링 결과에 노이즈가 반영됨
2. 이는 예측변수와 반응변수 사이의 연관성을 왜곡할 수 있고, 모델 성능에도 부정적인 영향을 끼치게 됨
3. 주식 가격 시계열에 **Kalman Filter**를 적용하여 가격 시계열에 대한 이동평균 추정값에 대해 Trend-Scanning 수행
4. 결과적으로 **노이즈가 일부 제거된 라벨링 결과**를 얻을 수 있음

1. 문제 정의와 데이터 구성

기존 Trend Scanning 라벨링 결과

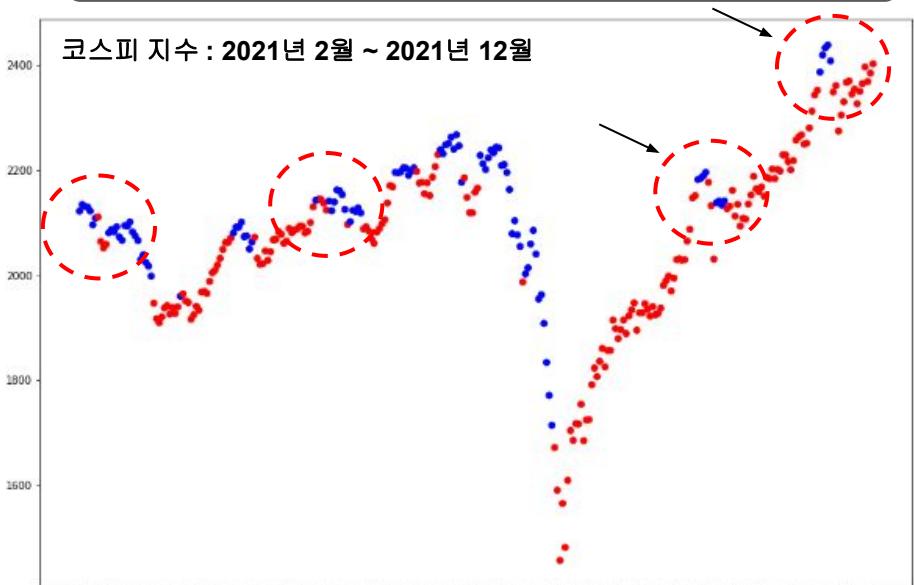


Kalman Filter를 적용한 라벨링 결과

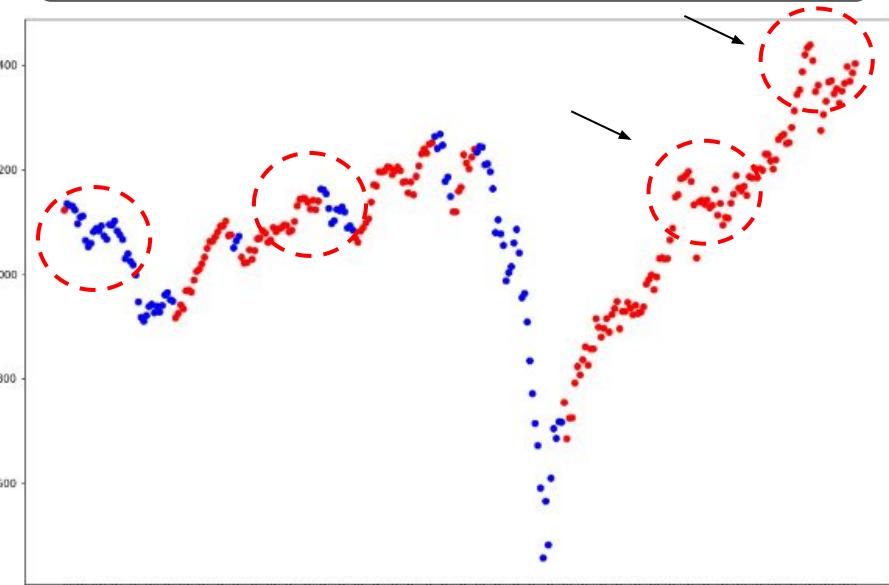


1. 문제 정의와 데이터 구성

기존 Trend Scanning 라벨링 결과



Kalman Filter를 적용한 라벨링 결과



1. 문제 정의와 데이터 구성

7. 효과

1. Kalman Filter를 적용하여 라벨링을 했을 때, EDA 과정에서 예측변수와 반응변수 사이의 상관관계와 예측변수가 반응변수에 가지는 예측력에 긍정적인 변화 발생한 것을 확인
2. 결론적으로, 금융 데이터를 어떻게 **라벨링**하느냐에 따라 예측모델의 성능이 좌우되는 것을 확인함

2

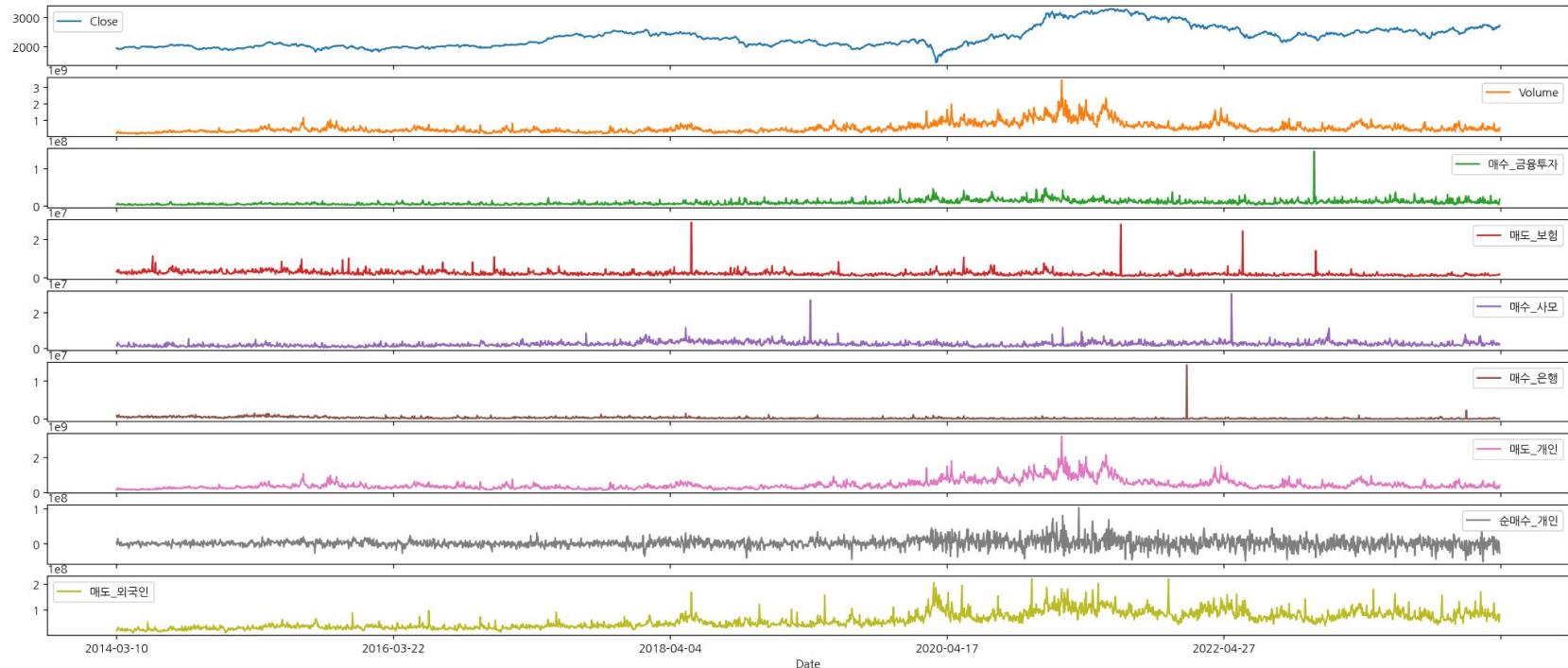
탐색적 데이터 분석 (EDA)



2. 탐색적 데이터 분석

1. Motivation

코스피 시계열 데이터의 분포를 시각화하는 것은
데이터 특성을 직관적으로 이해하는데 큰 도움이 되지 않음



2. 탐색적 데이터 분석

2. 방향성

: 효율적인 탐색적 데이터 분석을 위해

- 1) 각각의 예측변수가 반응변수에 대해 얼마나 예측력이 있는지 파악하는데 초점
- 2) Kalman Filter를 적용한 라벨링 전후 차이를 비교분석하는 것을 중심으로 탐색적 데이터 분석 수행

2. 탐색적 데이터 분석

3. 개요

: 데이터 스누핑, Test Set 과적합을 방지하기 위해 Train Set에 대해서만 탐색적 분석을 수행

- 1) 예측변수와 반응변수 사이의 상관관계 분석
- 2) 예측변수의 구간에 따른 타겟값의 비율 차이 분석
- 3) T-test를 통해 타겟값에 따른 예측변수값의 평균의 차이가 통계적으로 유의미한지를 검정

2. 탐색적 데이터 분석 - 상관관계 분석 | train-target

Train
데이터 분할

1 ~ 5 lagged
예측변수 생성

현재 시점의
반응변수와 비교

상관관계 계산

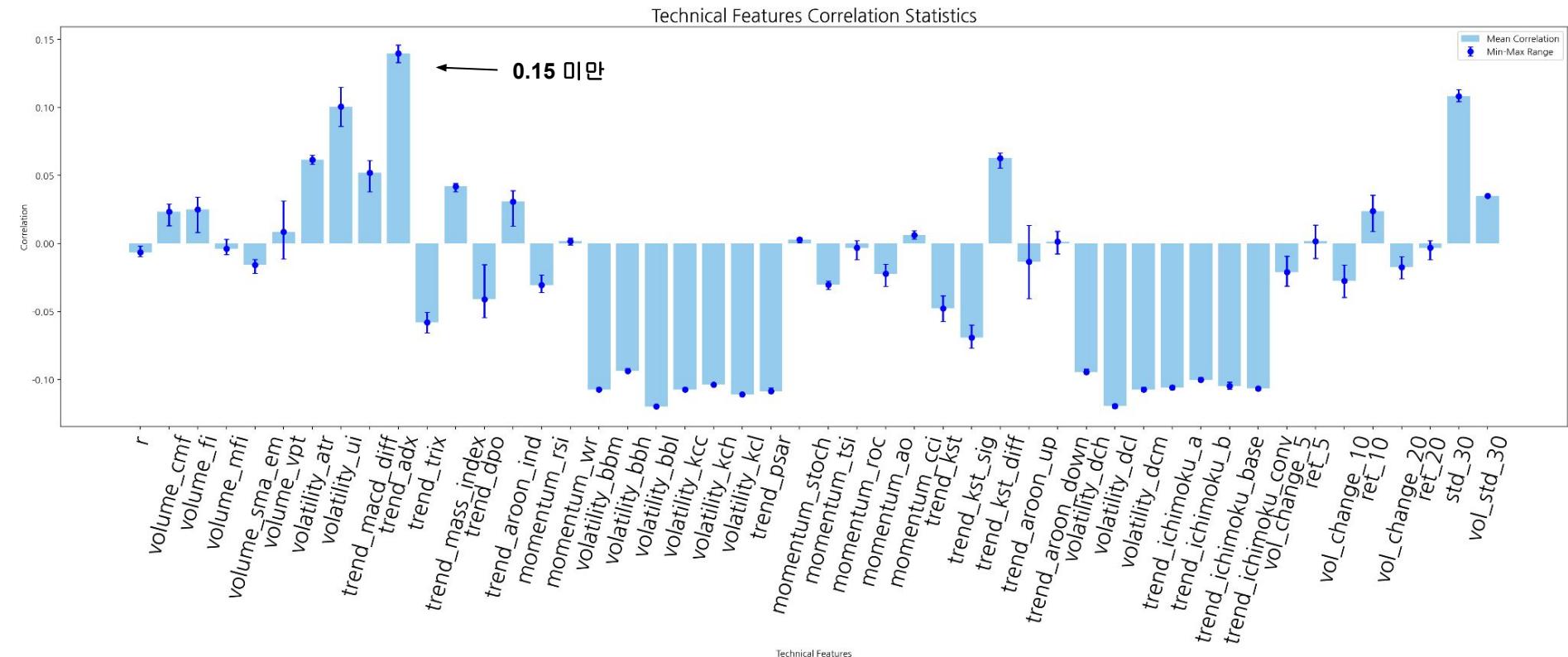
요약 통계
& 시각화

1. 예측변수 별로 각각 1 ~ 5일 지연시킨 값을 가지고 현재 시점의 반응변수와 상관계수를 구한다.
2. 하나의 예측변수 당 5개의 피어슨 계수 값을 얻게 된다.
3. 예측변수 별로 5개 상관계수의 평균값과 범위(Range)를 시각화한다.
4. 라벨링에 노이즈가 제거되기 전후 Target에 대한 예측변수의 상관계수의 변화를 비교

- **Technical Indicators:** Volume, volatility, trend, and momentum indicators (e.g., volume_cmf, volatility_atr, trend_macd_diff, momentum_rsi, trend_ichimoku, volatility_dch). Additionally, kalman_ma_log with kalman filter.
- **Macroeconomics Indicators:** Economic indices and financial metrics (e.g., dowjones, nasdaq, s&p500, 뉴스심리지수, USDKRW, 경제심리지수, 경상수지, 상품수지, 무역수지).
- **Investors Metrics:** Buy, sell, net buy/sell activities by various investor types (e.g., 금융투자, 보험, 투신, 사모, 은행, 연기금 등, 기타법인, 개인, 외국인, 기타외국인).
- **Investors Metrics2:** Investor ratios and activity indices (e.g., 매수/매도 비율, 주요그룹_합계, 체결강도, 활동성).

2. 탐색적 데이터 분석 - 상관관계 분석 | train-target

1. 노이즈 제거 전 (기술적 지표)

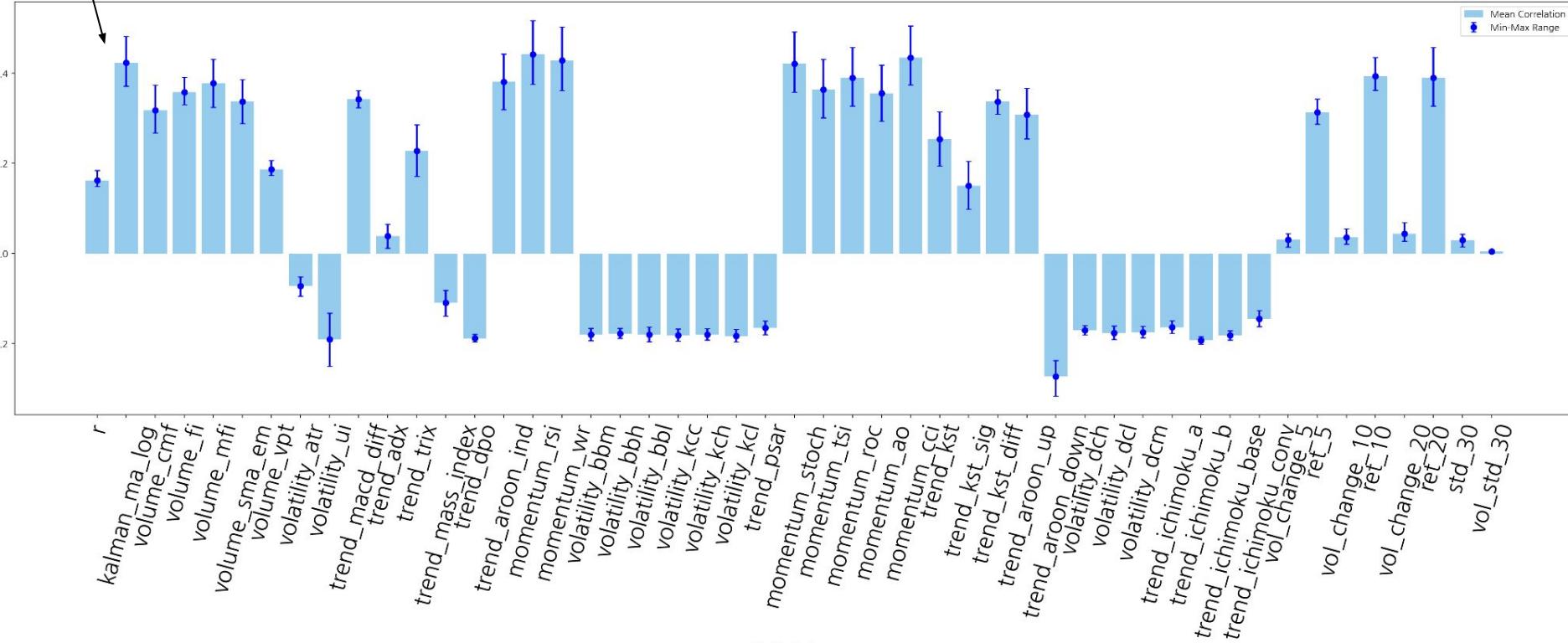


2. 탐색적 데이터 분석 - 상관관계 분석 | train-target

0.4 초과

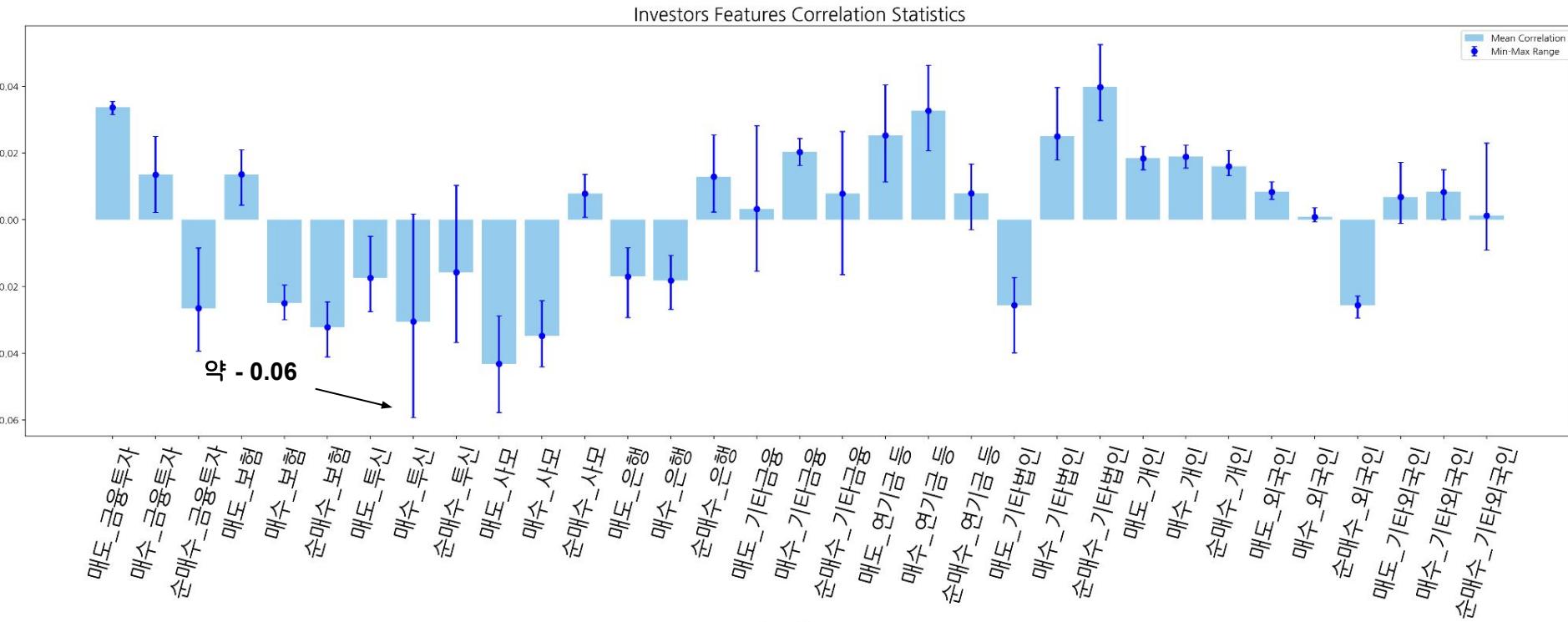
1. 노이즈 제거 후 (기술적 지표)

Technical Features Correlation Statistics



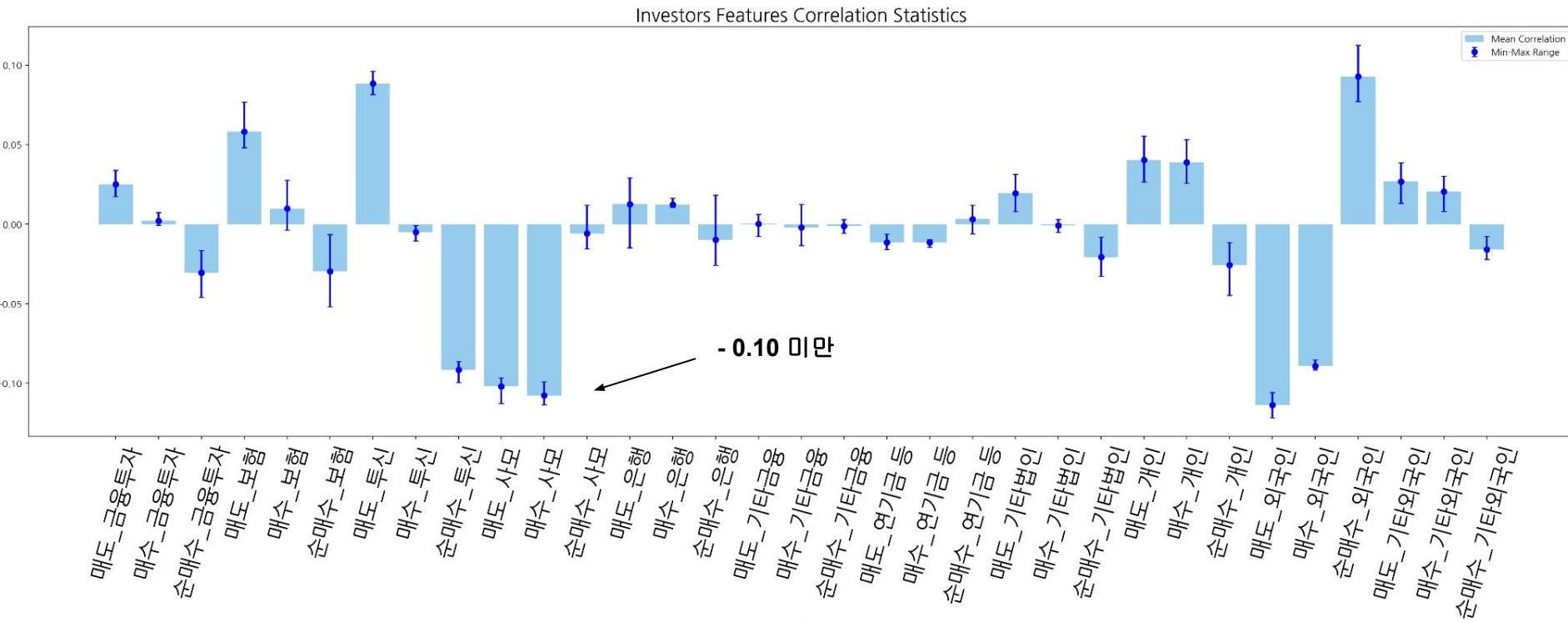
2. 탐색적 데이터 분석 - 상관관계 분석 | train-target

2. 노이즈 제거 전 (주체별 거래량)



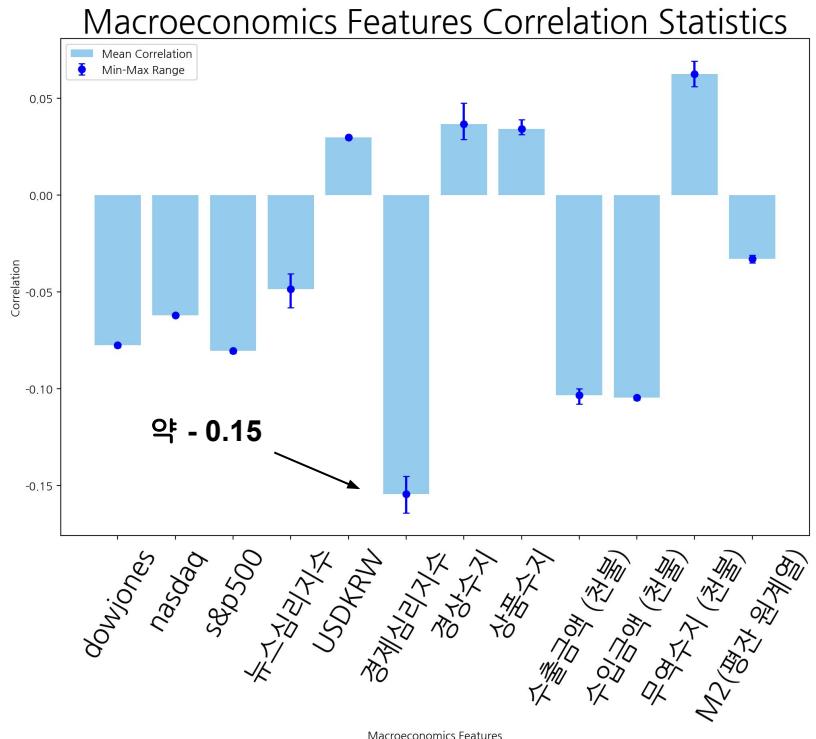
2. 탐색적 데이터 분석 - 상관관계 분석 | train-target

2. 노이즈 제거 후 (주체별 거래량)

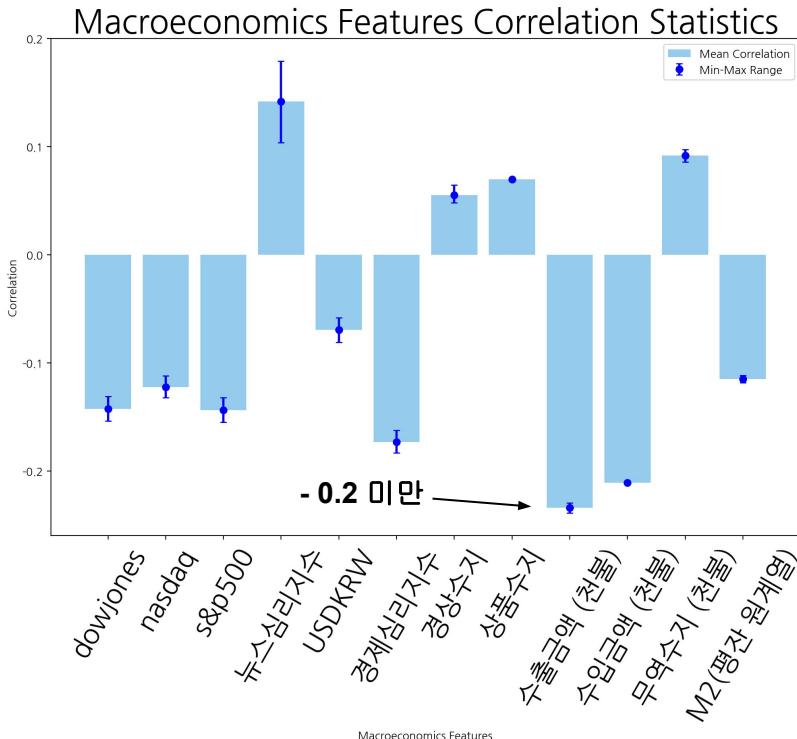


2. 탐색적 데이터 분석 - 상관관계 분석 | train-target

3. 노이즈 제거 전 (거시경제 지표)

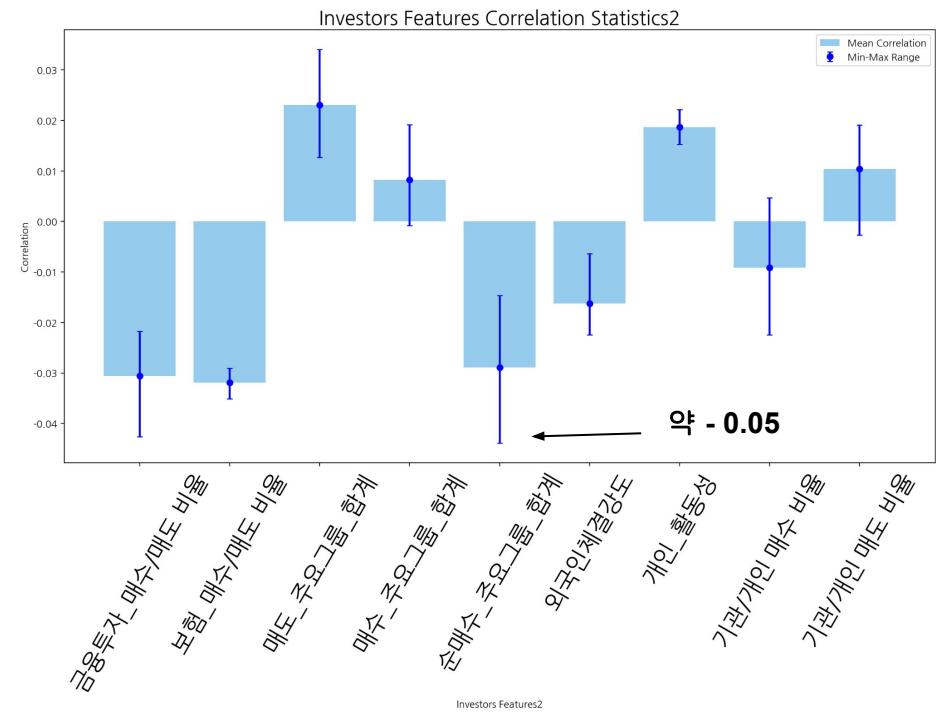


3. 노이즈 제거 후 (거시경제 지표)

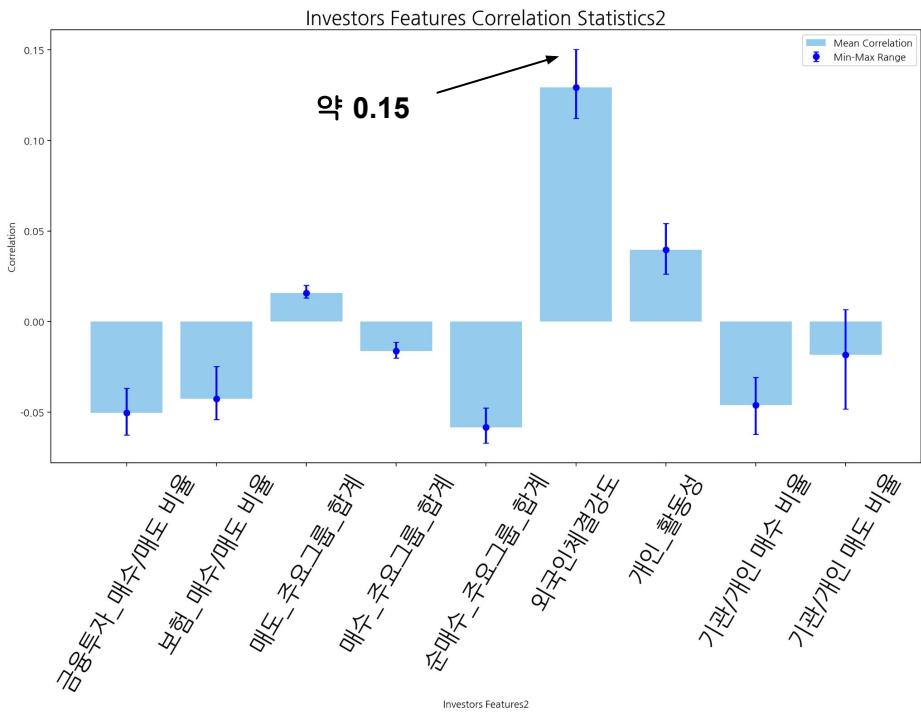


2. 탐색적 데이터 분석 - 상관관계 분석 | train-target

4. 노이즈 제거 전 (투자자별 거래량 파생변수)



4. 노이즈 제거 후 (투자자별 거래량 파생변수)

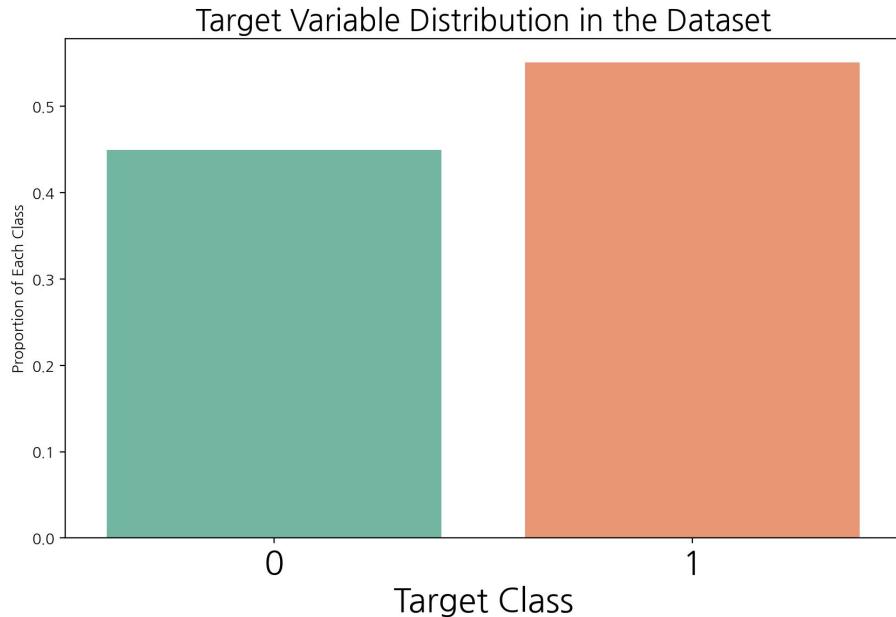


2. EDA - 예측변수 값의 구간에 따른 타겟값의 비율 분석

Train Data 의 타겟 비율

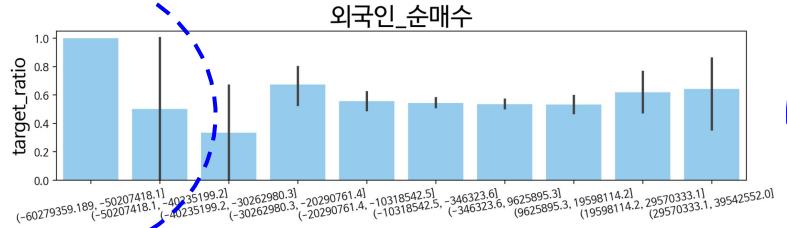
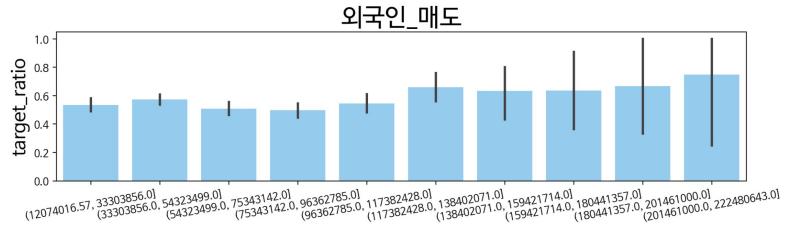
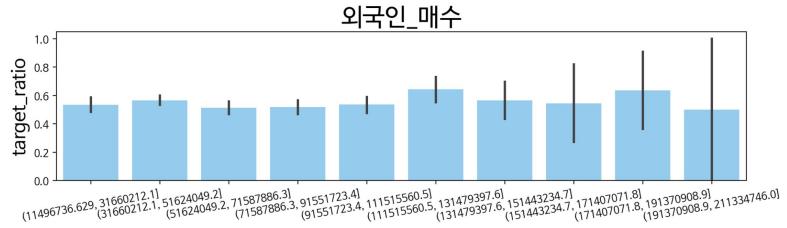
```
# 타겟값의 비율  
df_train_scaled['target'].value_counts(1)
```

1	0.5511
0	0.4489

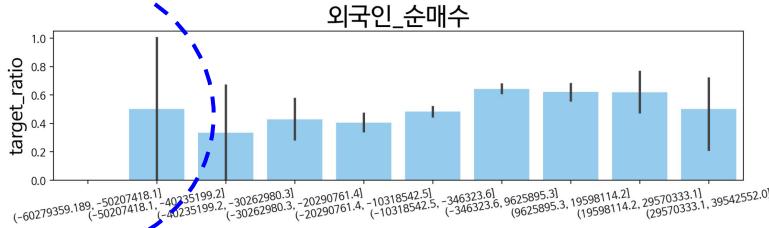
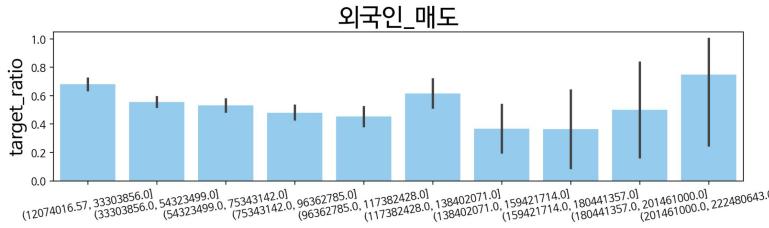
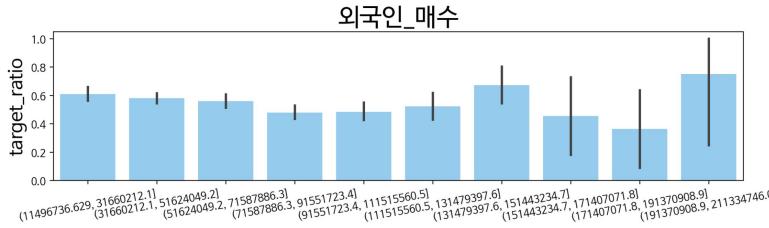


2. EDA - 예측변수 값의 구간에 따른 타겟값의 비율 - 외국인 거래량

노이즈 제거 전

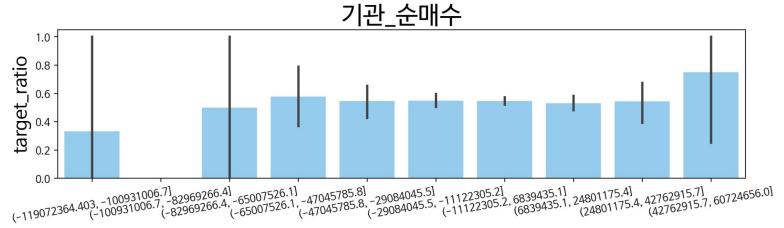
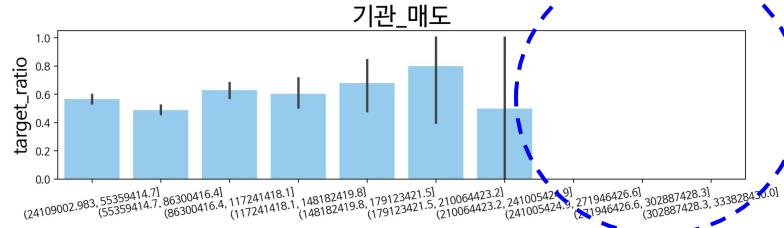
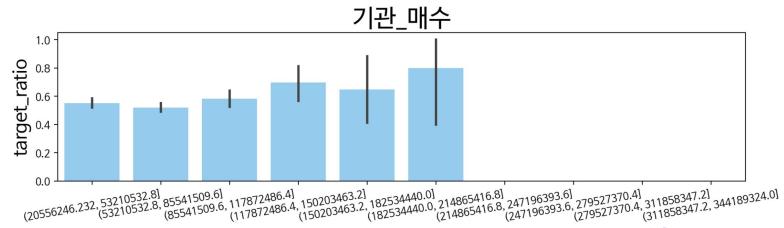


노이즈 제거 후

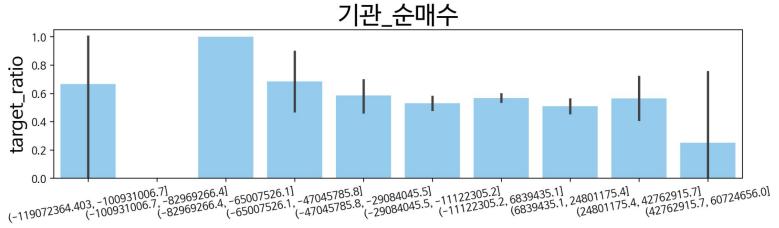
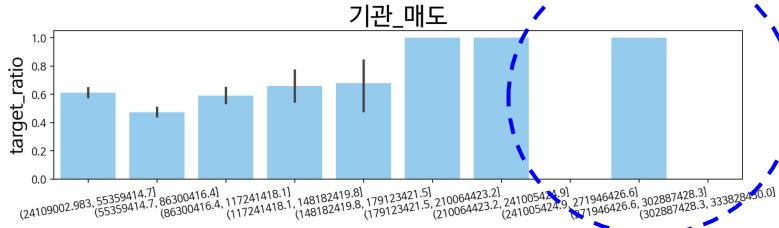
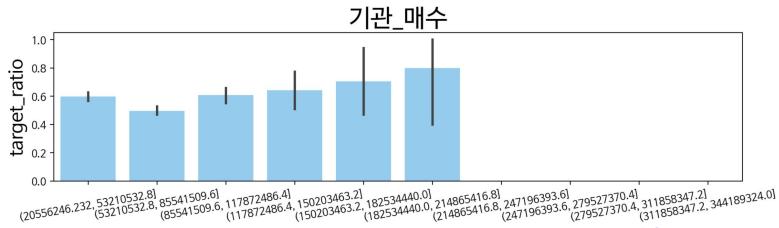


2. EDA - 예측변수 값의 구간에 따른 타겟값의 비율 - 기관 거래량

노이즈 제거 전

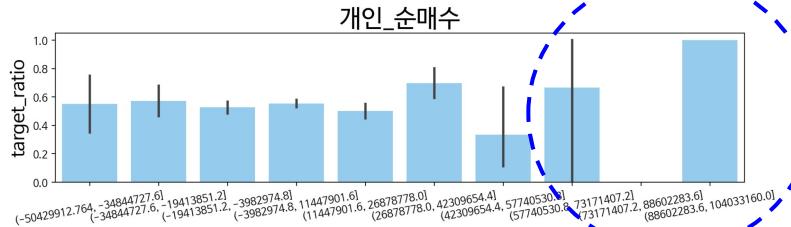
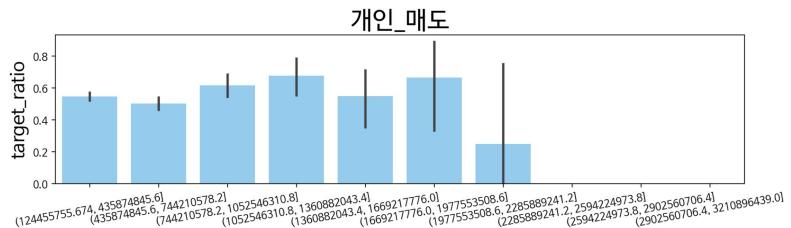
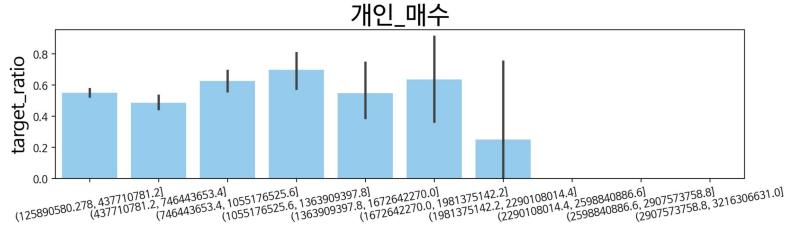


노이즈 제거 후

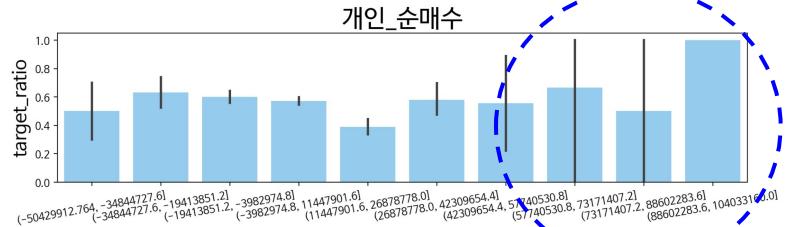
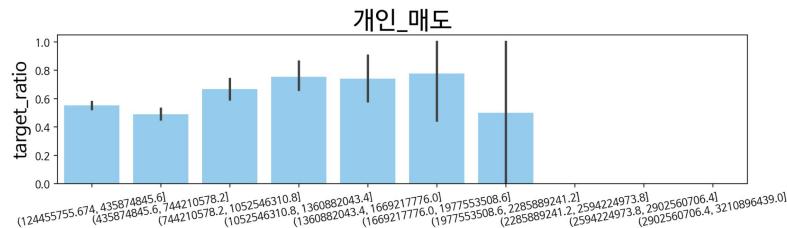
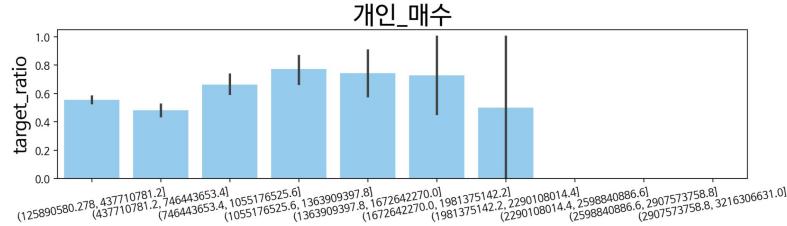


2. EDA - 예측변수 값의 구간에 따른 타겟값의 비율 - 개인 거래량

노이즈 제거 전

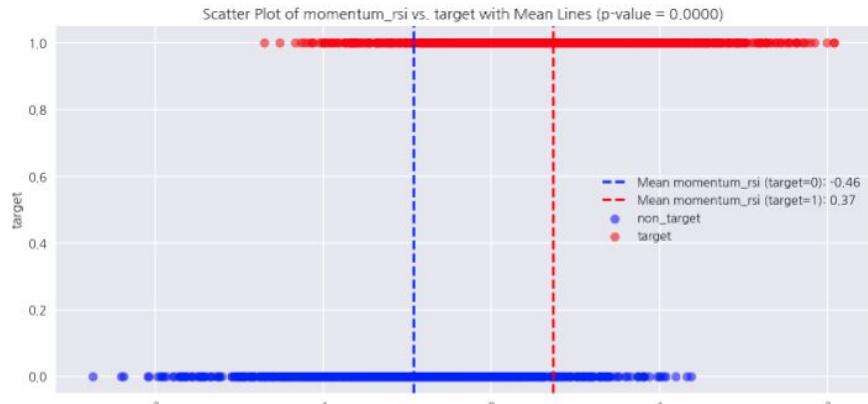


노이즈 제거 후



2. 탐색적 데이터 분석 - T-test

- 모든 예측변수 별로 반응변수에 대해 T-test 실행
- 총 104개 예측변수 개수 만큼 104번의 검정 수행
- 다중 테스트에 따른 알파 인플레이션을 고려하여 본페로니 교정(Bonferroni Correction) 적용
- 유의수준 = $0.05 / 104 = 0.000054$ 로 설정



2. 탐색적 데이터 분석 - T-test

노이즈 제거 전

Selected features based on p-values :

총 23개

'kalman_ma', 'volatility_ui', 'trend_adx', 'volatility_bbm', 'volatility_bbh',
'volatility_bbl', 'volatility_kcc', 'volatility_kch', 'volatility_kcl', 'trend_psar',
'volatility_dch', 'volatility_dcl', 'volatility_dcm', 'trend_ichimoku_a',
'trend_ichimoku_b', 'trend_ichimoku_base', 'trend_ichimoku_conv',
'std_30', 'dowjones', 's&p500', '경제심리지수', '수출금액 (천불)',
'수입금액 (천불)'

노이즈 제거 후

Selected features based on p-values :

총 66개

'매도_투신', '순매수_투신', '매도_사모', '매수_사모', '매도_외국인',
'매수_외국인', '순매수_외국인', 'kalman_ma', 'volume_cmf', 'volume_fi',
'volume_mfii', 'volume_sma_em', 'volume_vpt', 'volatility_atr',
'volatility_ui', 'trend_macd_diff', 'trend_trix', 'trend_mass_index',
'trend_dpo', 'trend_aroon_ind', 'momentum_rsi', 'momentum_wr',
'volatility_bbm', 'volatility_bbh', 'volatility_bbl', 'volatility_kcc',
'volatility_kch', 'volatility_kcl', 'trend_psar', 'momentum_stoch',
'momentum_tsi', 'momentum_roc', 'momentum_ao', 'momentum_cci',
'trend_kst', 'trend_kst_sig', 'trend_kst_diff', 'trend_aroon_up',
'trend_aroon_down', 'volatility_dch', 'volatility_dcl', 'volatility_dcm',
'trend_ichimoku_a', 'trend_ichimoku_b', 'trend_ichimoku_base',
'trend_ichimoku_conv', 'ret_5', 'vol_change_10', 'ret_10',
'vol_change_20', 'ret_20', 'dowjones', 'nasdaq', 's&p500',
'외국인체결강도', '기관/개인 매수 비율', '뉴스심리지수', 'USDKRW',
'경제심리지수', '상품수지', '수출금액 (천불)', '수입금액 (천불)',
'무역수지 (천불)', 'M2(평잔 원계열)', 'r', 'kalman_ma_log'

3

모델링 (Modeling)

3. 모델링

1. 기본 가정

1. 머신러닝 관점에서의 문제 정의 : 과거 20일치 데이터로 다음날 추세 예측

Look-Back Period (past 20 days)

$t - 19, t - 18, \dots, t - 1, t$ (현재시점)

예측 시점

$t + 1$ (다음날)

2. 예측결과를 바탕으로 장후 시간외 거래 및 다음날 장전 시간외 거래를 통해 매매
3. Backtesting 과정에는 증권사 수수료, 거래세, 그리고 고정 슬리피지 반영

3. 모델링

2. Train/Test 데이터 구성



시계열 자기상관으로 인해 Train set 마지막 부분 데이터가 Test set 시작 부분에 있는 예측변수의 지연값 패턴을 학습하는 것을 제한
: Data leakage 방지

3. 모델링 - Baseline Model

3. 베이스라인 모델

LSTM

1. 기본 구조

Instance Normalization,
Batch Normalization,
Weight Initialization,
Dropout 적용

2. LSTM layer = 2

3. Hidden Unit Dim = [64, 32]

4. Optimizer : Adam

5. 활성화 함수 : ReLU / Sigmoid

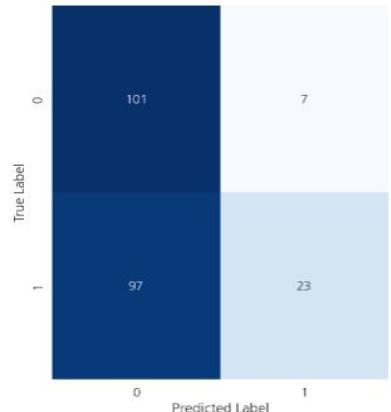
6. 교차검증

: Time Series Cross Validation, n = 4

노이즈 제거 전 성능

AUC: 0.5634

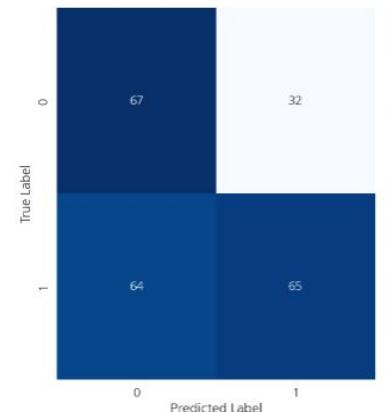
	precision	recall	f1-score	support
0	0.51	0.94	0.66	108
1	0.77	0.19	0.31	120
accuracy			0.54	228
macro avg	0.64	0.56	0.48	228
weighted avg	0.65	0.54	0.47	228



노이즈 제거 후 성능

AUC: 0.5903

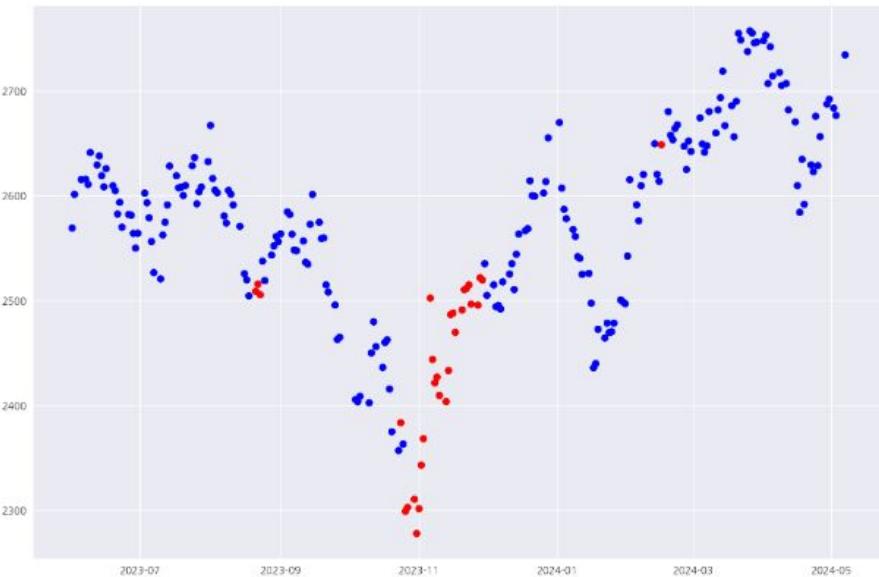
	precision	recall	f1-score	support
0	0.51	0.68	0.58	99
1	0.67	0.50	0.58	129
accuracy			0.58	228
macro avg	0.59	0.59	0.58	228
weighted avg	0.60	0.58	0.58	228



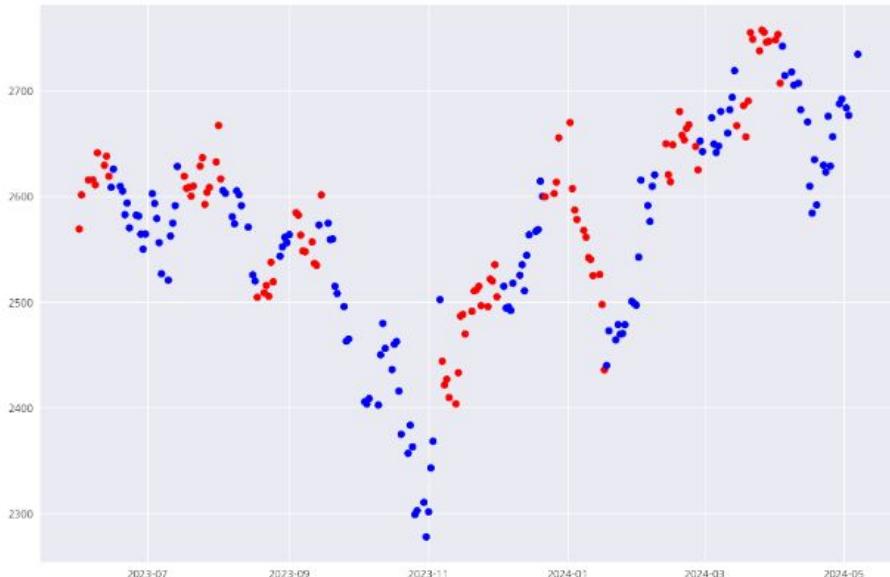
3. 모델링 - Baseline Model

4. 예측 결과

(노이즈 제거 전) 모델의 추세 예측



(노이즈 제거 후) 모델의 추세 예측



3. 모델링 - Baseline Model

5. Backtesting

- 모델 예측값이 1에서 0으로 바뀌는 시점 - Short Position, 반대는 Long Position.
- 거래비용 반영 : 증권사 수수료 - 0.015%, 거래세 - 0.18%, 고정슬리피지 - 0.05%

노이즈 제거 전 성능



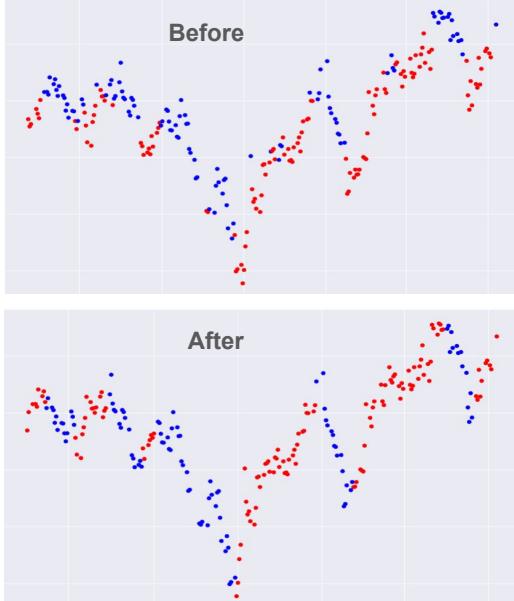
노이즈 제거 후 성능



3. 모델링 - 성능 개선 : LightGBM, XGBoost

1. Boosting Models

Kalman Filter 적용 Labeling

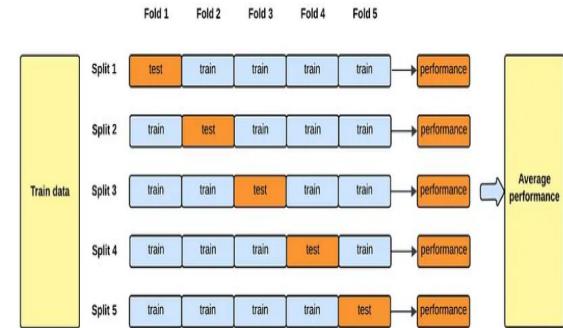


특성 중요도 분석 및 Feature Selection

1. 모든 변수 (104개)
2. T-test
: p-value가 가장 낮은 15개 변수 선택
3. Correlation
: 피어슨 계수의 절댓값이 큰 상위 15개 변수 선택
4. Mean Decrease Accuracy, MDA
: 교차검증 성능에 영향을 끼친 상위 15개 변수 선택
5. Recursive Feature Elimination, RFE
: 교차검증 시 모든 변수를 포함한 상태에서 기여도에 따라 변수를 삭제. 남은 13개 변수 선택
6. Sequential Forward Selection, SFS
: 순차적으로 변수를 추가하거나 제거하여 발견한 최적의 조합 15개 변수 선택

최적화 및 일반화 능력 개선

1. 하이퍼파라미터 튜닝
: 베이지안 최적화 적용
2. 과적합 대처 및 일반화 능력 향상
: OOP 예측 (Out-of-fold Prediction) 활용



3. 모델링 - Feature Selection

모든 변수 104개

T-Test 상위 15개

Corr 상위 15개

MDA 상위 15개

RFE 조합 13개

SFS 조합 15개

- ['매도_금융투자', '매수_금융투자', '순매수_금융투자', '매도_보험', '매수_보험', '순매수_보험', '매도_투신', '매수_투신', '순매수_투신', '매도_사모', ...]
- '경제심리지수', '경상수지', '상품수지', '수출금액(천불)', '수입금액(천불)', '무역수지(천불)', 'M2(평잔원계열)', 'r', 'kalman_ma_log']

- ['momentum_wr', 'momentum_rsi', 'momentum_cci', 'momentum_stoch', 'kalman_ma_log', 'ret_20', 'momentum_roc', 'ret_10', 'trend_aroon_ind', 'volume_mfi', 'momentum_tsi', 'volume_fi', 'momentum_ao', 'volume_sma_em', 'trend_macd_diff']

- ['momentum_rsi', 'momentum_cci', 'momentum_wr', 'momentum_stoch', 'kalman_ma_log', 'momentum_ao', 'momentum_roc', 'ret_20', 'trend_aroon_ind', 'ret_10', 'volume_mfi', 'momentum_tsi', 'momentum_ao', 'volume_fi', 'volume_sma_em', 'volume_cmf']

- ['trend_kst', 'volume_sma_em', 'kalman_ma_log', 'momentum_ao', 'momentum_wr', 'momentum_rsi', 'momentum_cci', '경제심리지수', 'volume_fi', 'ret_5', 'momentum_stoch', 'std_30', '매수_보험', '수입금액(천불)', 'trend_adx']

- ['volatility_atr', 'volatility_ui', 'trend_adx', 'trend_mass_index', 'volatility_bbm', 'trend_kst_sig', 'trend_kst_diff', 'volatility_dcl', 'std_30', 'vol_std_30', '뉴스심리지수', '수출금액(천불)', '무역수지(천불)']

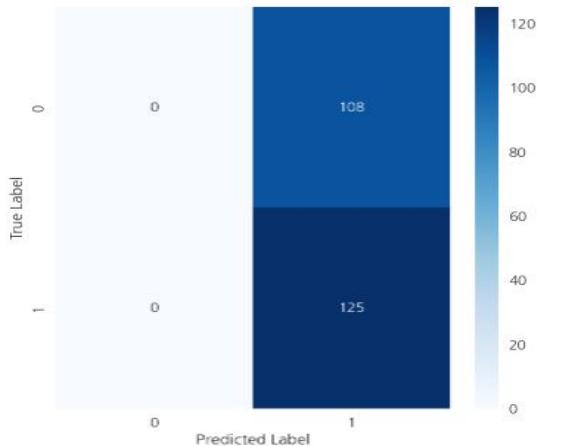
- ['매수_금융투자', '매도_투신', '매수_투신', '순매수_투신', '매도_사모', '매수_은행', 'momentum_rsi', 'momentum_roc', 'ret_5', 'ret_10', 'ret_20', '매수_주요그룹_합계', '외국인체결강도', '기관/개인 매수 비율', 'kalman_ma_log']

3. 모델링 - 성능 개선 : LightGBM, XGBoost

2. Boosting 기본 모델

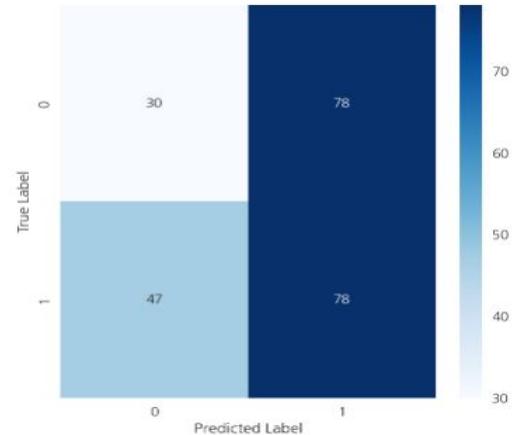
(노이즈 제거 전) LightGBM 성능

AUC: 0.4911				
	precision	recall	f1-score	support
0	0.00	0.00	0.00	108
1	0.54	1.00	0.70	125
accuracy			0.54	233
macro avg	0.27	0.50	0.35	233
weighted avg	0.29	0.54	0.37	233



(노이즈 제거 전) XGBoost 성능

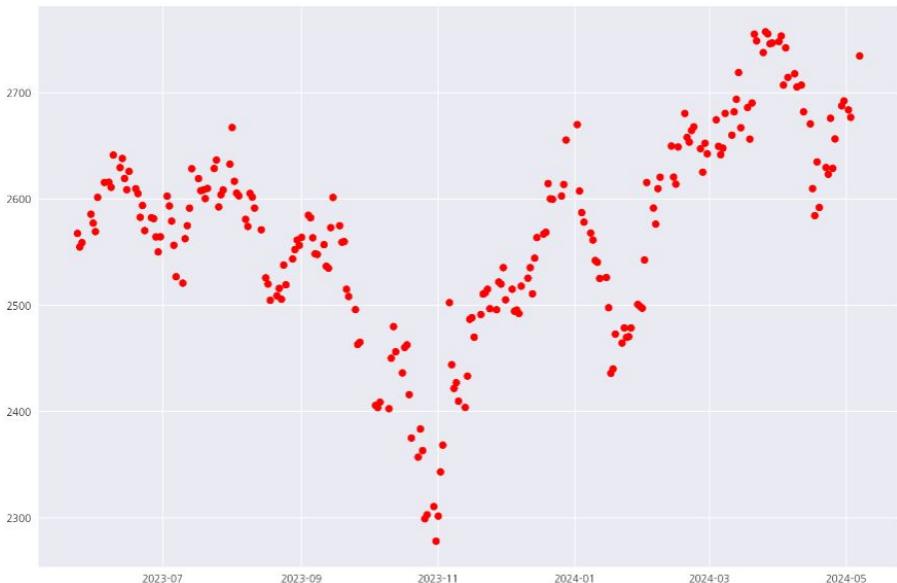
AUC: 0.4581				
	precision	recall	f1-score	support
0	0.39	0.28	0.32	108
1	0.50	0.62	0.56	125
accuracy			0.46	233
macro avg	0.44	0.45	0.44	233
weighted avg	0.45	0.46	0.45	233



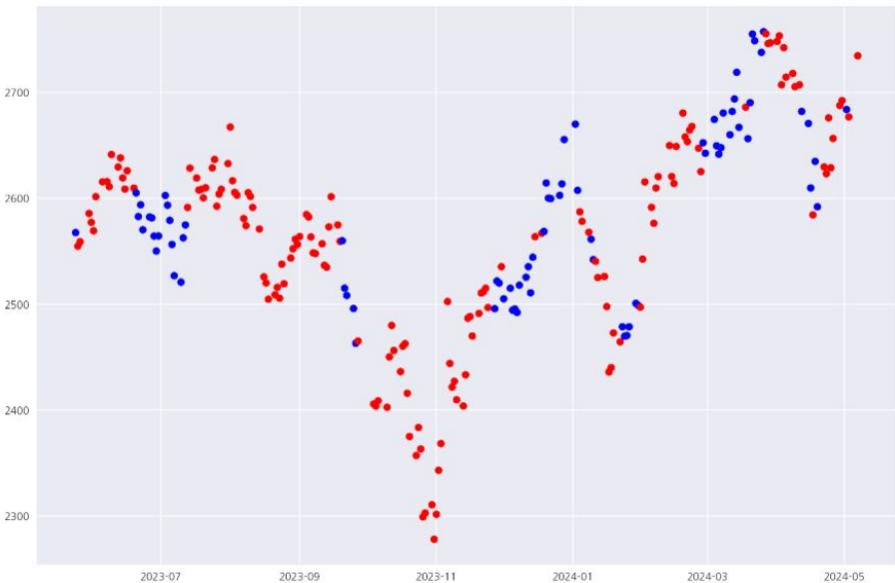
3. 모델링 - 성능 개선 : LightGBM, XGBoost

2. Boosting 기본 모델

(노이즈 제거 전) LightGBM 예측 결과



(노이즈 제거 전) XGBoost 예측 결과



3. 모델링 - 성능 개선 모델 : 실험 결과

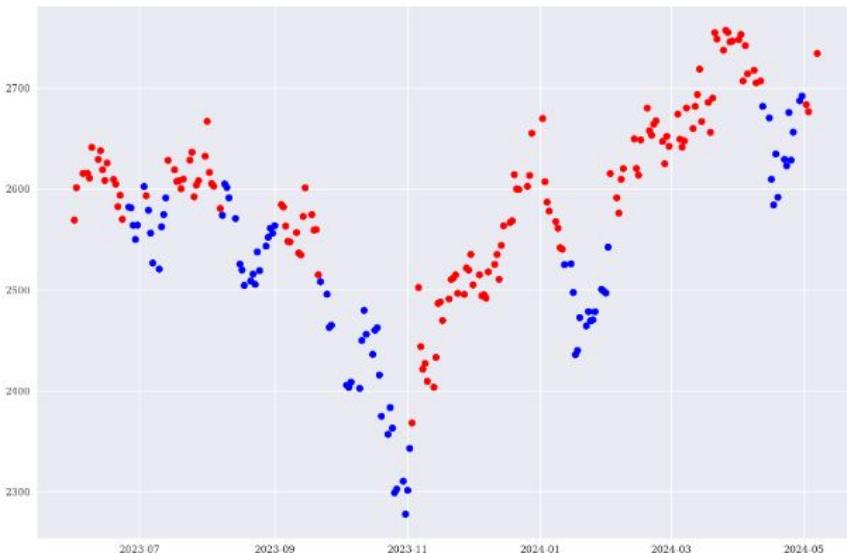
		모든 변수(104개)		T-Test(15개)		Corr(15개)		MDA(15개)		RFE(13개)		SFS(15개)	
Look-back period	Model Metric	L-GBM	XGB	L-GBM	XGB	L-GBM	XGB	L-GBM	XGB	L-GBM	XGB	L-GBM	XGB
Lag 1	AUC	0.7630	0.7469	0.7497	0.7487	0.7510	0.7438	0.7664	0.7673	0.6157	0.5879	0.7685	0.7587
	Net Profit(%)	-3.1	-0.68	1.12	-5.65	-5.22	2.47	2.95	-1.3	7.85	-26.47	-2.83	2.11
Lag 5	AUC	0.7375	0.7513	0.7523	0.7438	0.7438	0.7456	0.7904	0.7539	0.5286	0.5671	0.7684	0.7414
	Net Profit(%)	5.48	-12.59	5.32	4.34	8.47	4.27	8.47	6.23	-22.96	-24.61	1.94	-9.58
Lag 10	AUC	0.7574	0.7339	0.7389	0.7443	0.7558	0.7315	0.7600	0.7511	0.5327	0.5938	0.7669	0.7404
	Net Profit(%)	-8.13	-6.38	7.34	-0.03	0.05	2.78	9.77	4.63	-23.80	-8.1	9.77	-3.17
Lag 15	AUC	0.7445	0.7459	0.7615	0.7354	0.7085	0.7363	0.7944	0.7798	0.6055	0.6208	0.7542	0.7434
	Net Profit(%)	-6.62	0.13	11.65	1.52	6.04	1	-2.26	2.49	0.21	-4.13	1.91	-6.99
Lag 20	AUC	0.7500	0.7452	0.7431	0.7520	0.7281	0.7465	0.7281	0.7465	0.5643	0.5950	0.7483	0.7462
	Net Profit(%)	-0.14	-3.46	17.23	12.92	-1.66	18.43	-1.66	18.43	-16.32	5.65	-3.26	-5.66

3. 모델링 - 성능 개선 : LightGBM, XGBoost

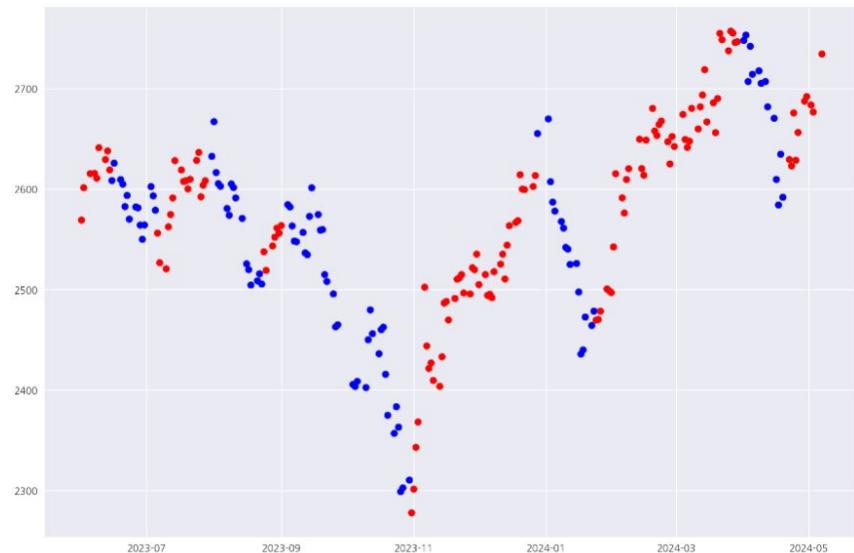
3. 최고 성능 모델

Corr 15, MDA 15

(노이즈 제거 후) LightGBM, XGBoost 예측 결과



(노이즈 제거 후) Ground Truth

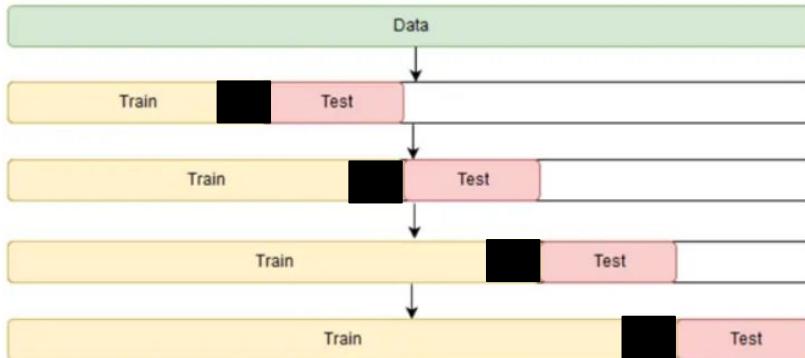


3. 모델링 - 성능 개선 : LightGBM, XGBoost

4. 교차검증

Corr 15, MDA 15

Time Series Cross Validation



Validation 성능

```

##### 풀드 1 / 풀드 4
[0]      valid-error:0.27002
[100]    valid-error:0.21968
[200]    valid-error:0.20366
풀드 1 지니계수 : 0.5851973858987302
  
```

정규화 지니계수

CV 평균 : 0.558

```

##### 풀드 2 / 풀드 4
[0]      valid-error:0.63158
[100]    valid-error:0.44165
[200]    valid-error:0.43249
풀드 2 지니계수 : 0.45017553335133675
  
```

테스트 : 0.452

OOO 예측성능 : 0.49

```

##### 풀드 3 / 풀드 4
[0]      valid-error:0.25172
[100]    valid-error:0.21739
[199]    valid-error:0.21510
풀드 3 지니계수 : 0.5403391715318321
  
```

```

##### 풀드 4 / 풀드 4
[0]      valid-error:0.64302
[100]    valid-error:0.31808
[200]    valid-error:0.31350
풀드 4 지니계수 : 0.6599598503513094
  
```

3. 모델링 - 성능 개선 : LightGBM, XGBoost

5. Backtesting

(노이즈 제거 후) LightGBM, XGBoost 결과



(노이즈 제거 전) XGBoost 결과

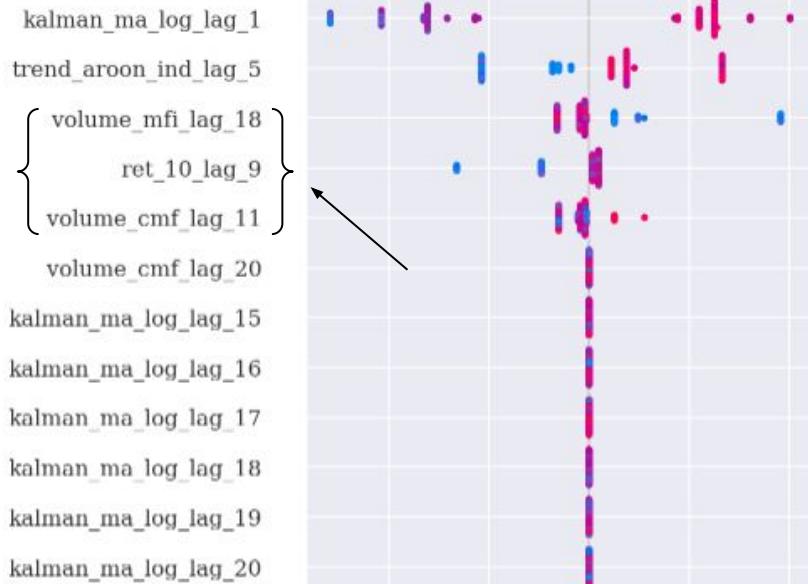


3. 모델링 - 성능 개선 : LightGBM, XGBoost

6. 사후 분석 (SHAP)

- 학습된 모델에서 Train set 과 Test set 에서 나온 SHAP 결과가 비슷한 것을 확인
- 모델의 성능 (AUC : 0.75, Accuracy : 0.67, F1 : 0.73) - 개선 가능성 O
- 세부 분석(일반화 능력, 일관성, 신뢰성, 안정성)은 보류하고, 성능을 개선하는데 비중을 둠

SHAP - Train Set



SHAP - Test Set



3. 모델링 - 성능 개선 : LightGBM, XGBoost

6. 사후 분석 (SHAP)

모델의 일반화 능력

1. 모델이 학습 데이터에서 학습한 패턴이 테스트 데이터에서도 잘 적용되고 있음을 의미함
2. 모델이 과적합되지 않았고 일반화 능력이 좋다는 것을 시사

특성 중요도의 일관성

1. 학습 데이터와 테스트 데이터의 SHAP 값이 비슷하다는 것은 두 데이터셋에 대해 모델이 특성의 중요도를 일관되게 평가하고 있음을 의미함

모델의 안정성과 신뢰성

1. SHAP 값이 일관되면 모델의 예측 결과를 해석하고, 신뢰할 수 있는 설명을 제공하는데 긍정적인 요인으로 생각할 수 있음

4

모델 의의 &
추후 연구 과제

4. 모델 특성, 한계, 그리고 추후 연구 과제

모델의 특성

- 1) 긴 상승, 하락 구간이 있는 경우
양호한 예측 성능과 경제적 성능이
나타남
- 2) 반면에, 짧은 상승, 하락, 횡보가
있는 구간에 대해서는 예측 신호의
정확도는 양호하나 경제적 성능이
하락하는 것으로 나타남

모델의 한계

- 1) 정확도가 높게 나와도 경제적
성능이 안 좋게 나오는 경우가 많음
- 2) 경제적 성능이 좋게 나온 결과는
많은 실험의 결과로 발견된 1종 오류
(거짓 양성)일 가능성이 있음

추후 연구 과제

- 1) 입력변수의 노이즈를 제거하는
작업을 통해 모델의 성능향상 도모
- 2) 예측변수 별로 지연값을
군집화하여 군집 MDA를
수행함으로 예측변수의 기여도를
상세히 분석

References

- 유주현, *Finance Time Series 데이터 활용하기*. 아이펠 캠퍼스.
- Marcos Lopez de Prado**, *Advances in Financial Machine Learning*. Wiley, 2018.
- Marcos Lopez de Prado**, *자산운용을 위한 금융 머신러닝*. 에이콘, 2021.
- New York Institution of Finance**, *Machine Learning for Trading*.
- New York University**, *Machine Learning and Reinforcement Learning in Finance*.
- 박유성, *시계열 예측과 분석*. 자유아카데미, 2024.
- 신백균, *머신러닝 딥러닝 문제해결 전략*. 골든레빗, 2022.
- 대신경제연구소, "통화량과 주가변동, 상관관계 높다," *매일경제*, 1988년 10월 25일.
- US FRED**, "10-Year Treasury Constant Maturity Minus 2-Year Treasury Constant Maturity," accessed June 3, 2024,
<https://fred.stlouisfed.org/series/T10Y2YM>.
- 오현우, "금리차 역전되면 경기 침체 오는데...'기묘한' 美 경제," *한국경제*, 2023년 6월 23일.
- 권재희, "원달러환율과 외국인 수급의 관계," *아시아경제*, 2022년 8월 7일.
- 천일영, "미국 주가 변동이 금융시장에 미치는 영향," *현대경제연구원*, 2000년 1월 6일.
- 윤서연, "한은 '뉴스심리지수, 경제지수보다 1~2개월 선행'," *인포스탁데일리*, 2022년 2월 9일.
- 이도훈, "무역수지와 주가지수의 상관관계," *매일경제*, 1996년 10월 25일.
- 한국경제연구원, "무역수지 적자나면 외국인 주식 순매도 확률 28.3% 증가," *한국경제연구원*, 2022년 9월 21일.
- 통계청, *e-나라지표*. Accessed June 3, 2024.
- 한국은행, *한국은행 API*. Accessed June 3, 2024.

Data-Driven Equity

감사합니다.

Data-Driven Equity

Q & A

Data Science Approaches to Stock Trend Prediction

김도현 서인선 양동영 윤진영