김영민, 박진철, 설위준, 송동민, 차수환

시계열 및 ML/DL기반 주가예측 모델링

< 美 Index주식(S&P 500 및 나스닥100) **종가 예측 모델 구현 >**

Contents

다양한 독립변수를 통한
미래주가 예측 모델링 구현

1. 주제선정 배경

점점 대중들에게 익숙해지고 있는 주식
하지만, 역대급 변동성으로 인해 주가 예측은 갈수록 어려워지고 있음

남녀노소 주식 광풍(狂風)

국제 정세의 역대급 변동성

이데일리 · www.edaily.co.kr › News

트럼프가 코스피도 흔들까...美 대선 변동성 경고

2024.07.12. 오는 11월 치러지는 미국 대선이 하반기 국내 증시의 주요 변수로 떠올랐다. 도널드 **트럼프** 전 대통령이 우세를 보이면서다. 증권사 리서치 센터장들은 **트럼프** 전 대통령이 대선에서 승기를 잡으면 시장의...

조선일보 · 7시간 전 · 네이버뉴스

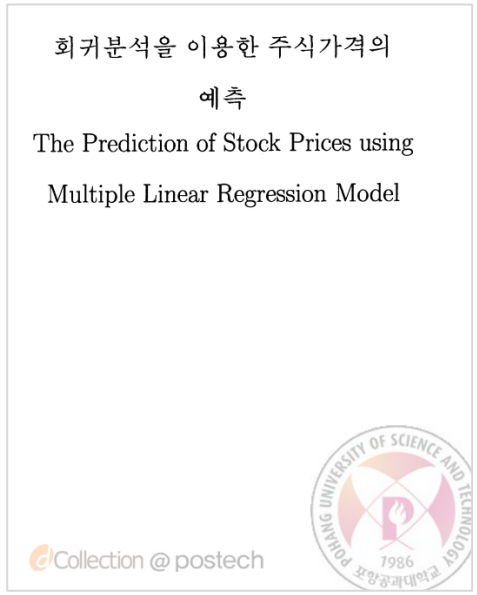
다시 불붙는 미·중 전쟁에서 대한민국의 최선책은?

서태평양 자유 진영의 최전선 대한민국과 중화민국(대만)이 미·중 **전쟁**의 **영향**을 가장 크게 받는다. 긴박한 미·중 대결의 한가운데서 한국과 대만이... 러시아-**우크라이나 전쟁**을 24시간 안에 종식할 수 있다는 트럼프의...

2. 선행논문 참고

기존 논문을 통해 미래 주가 예측을 위한 '분석 방법론' 참고

[회귀분석 활용]



회귀분석을 이용한 주식가격의

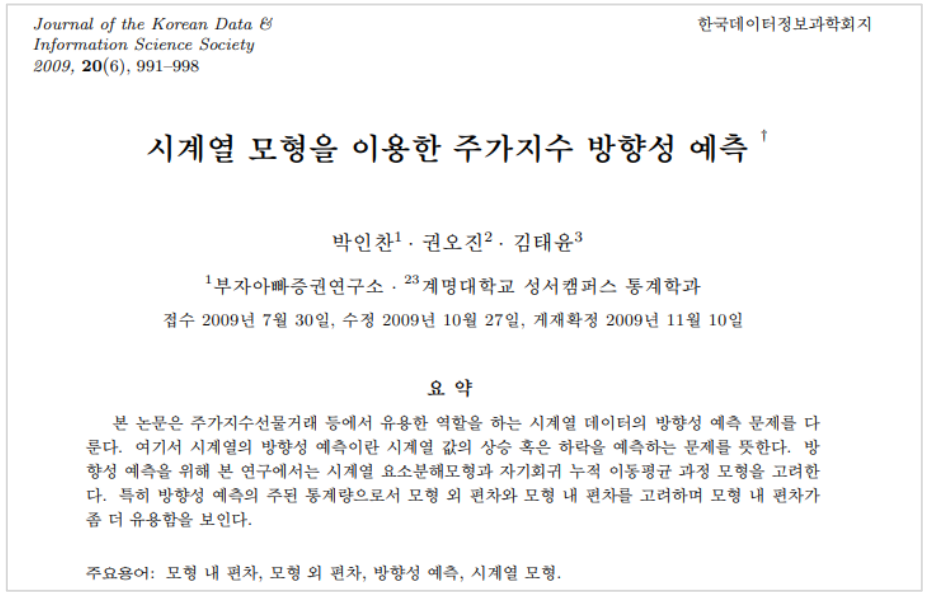
예측

The Prediction of Stock Prices using

Multiple Linear Regression Model

Collection @ postech

[시계열 모형 활용]



Journal of the Korean Data & Information Science Society
2009, **20**(6), 991–998

한국데이터정보과학회지

시계열 모형을 이용한 주가지수 방향성 예측 †

박인찬[1] · 권오진[2] · 김태윤[3]

[1]부자아빠증권연구소 · [23]계명대학교 성서캠퍼스 통계학과

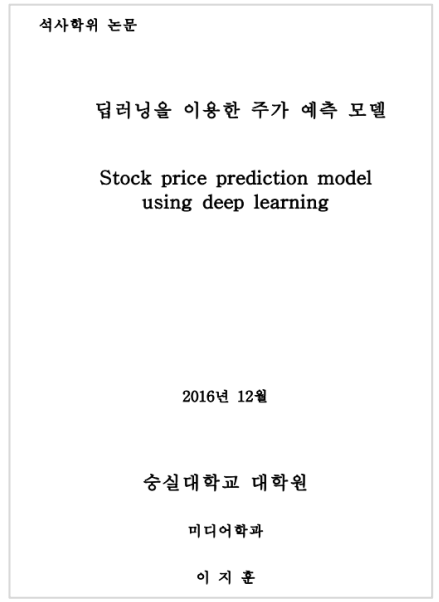
접수 2009년 7월 30일, 수정 2009년 10월 27일, 게재확정 2009년 11월 10일

요 약

본 논문은 주가지수선물거래 등에서 유용한 역할을 하는 시계열 데이터의 방향성 예측 문제를 다룬다. 여기서 시계열의 방향성 예측이란 시계열 값의 상승 혹은 하락을 예측하는 문제를 뜻한다. 방향성 예측을 위해 본 연구에서는 시계열 요소분해모형과 자기회귀 누적 이동평균 과정 모형을 고려한다. 특히 방향성 예측의 주된 통계량으로서 모형 외 편차와 모형 내 편차를 고려하며 모형 내 편차가 좀 더 유용함을 보인다.

주요용어: 모형 내 편차, 모형 외 편차, 방향성 예측, 시계열 모형.

[딥러닝 활용]



석사학위 논문

딥러닝을 이용한 주가 예측 모델

Stock price prediction model
using deep learning

2016년 12월

숭실대학교 대학원

미디어학과

이 지 훈

3. Data 수집

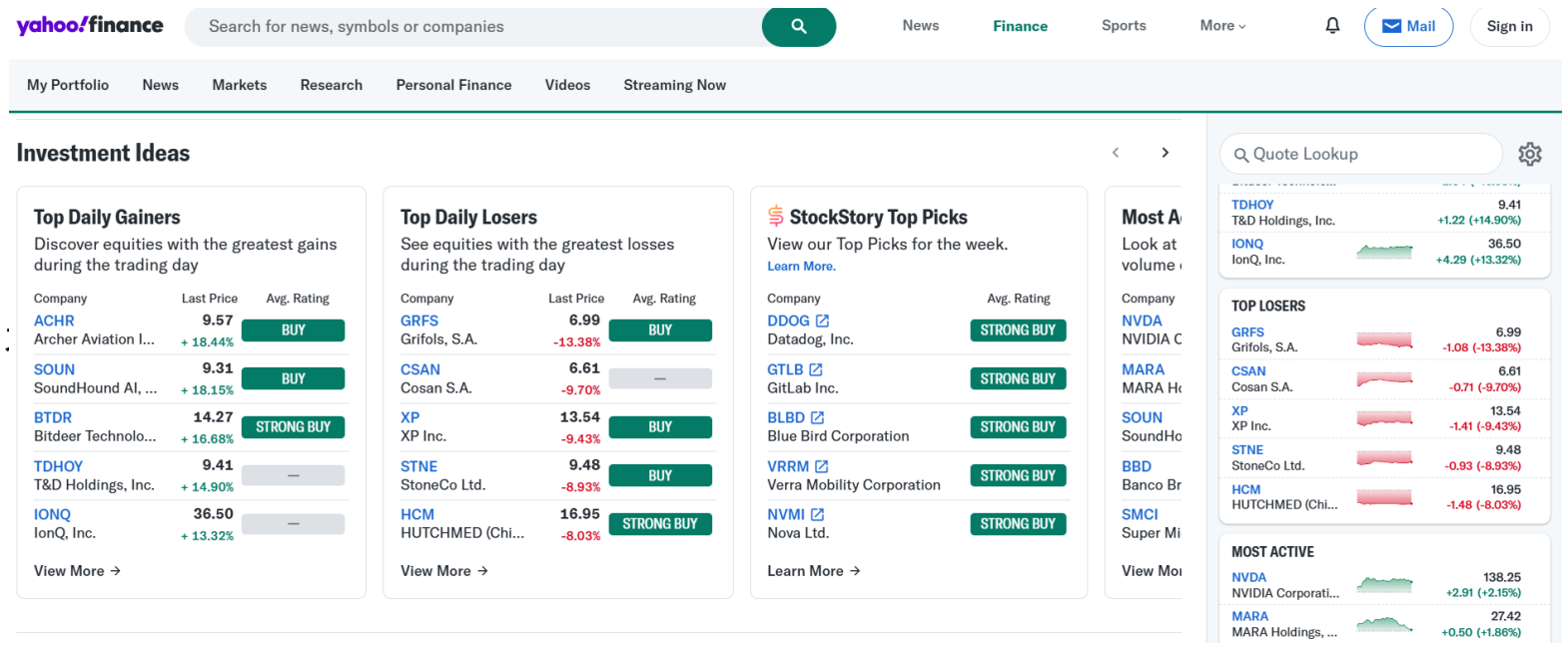
미국 대표 기술주 선정

- 마이크로소프트
- 구글
- 엔비디아
- 메타
- 애플
- 테슬라
- 아마존

※ 수집기간 : 2019년 ~ 2024년 (5년간)

3. Data 수집

'Yahoo Finance' Stock Data 웹크롤링



Stock Data

- 주가
- 배당금
- 재무정보

대외 환경변수

- 금리, 환율
- 석유, 금값
- 실업률, 비트코인 가격 등

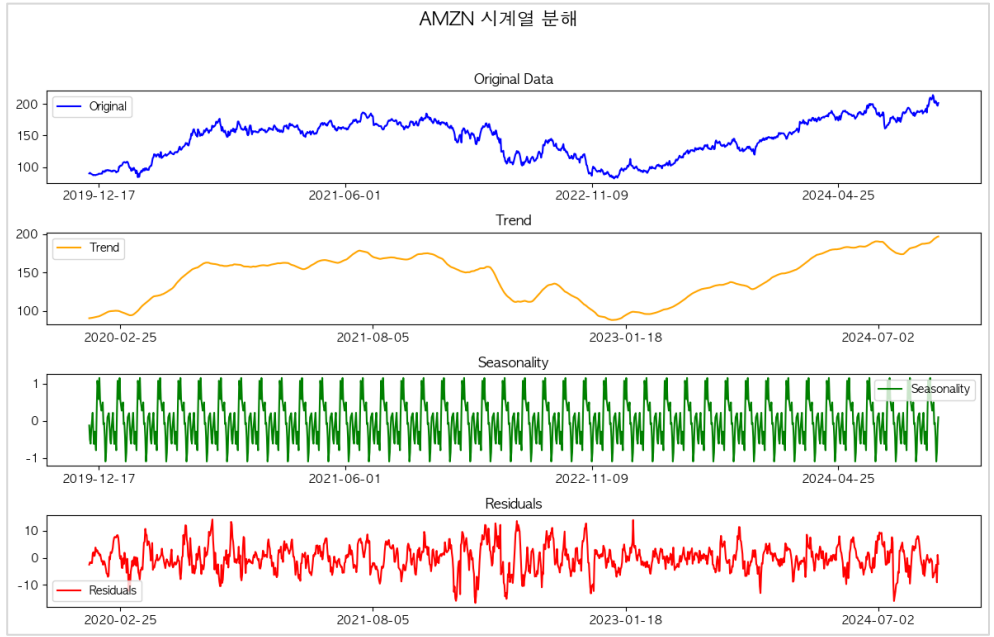
4. 분석/모델링

시계열 기반 분석 기법을 통한 주가 예측 시도

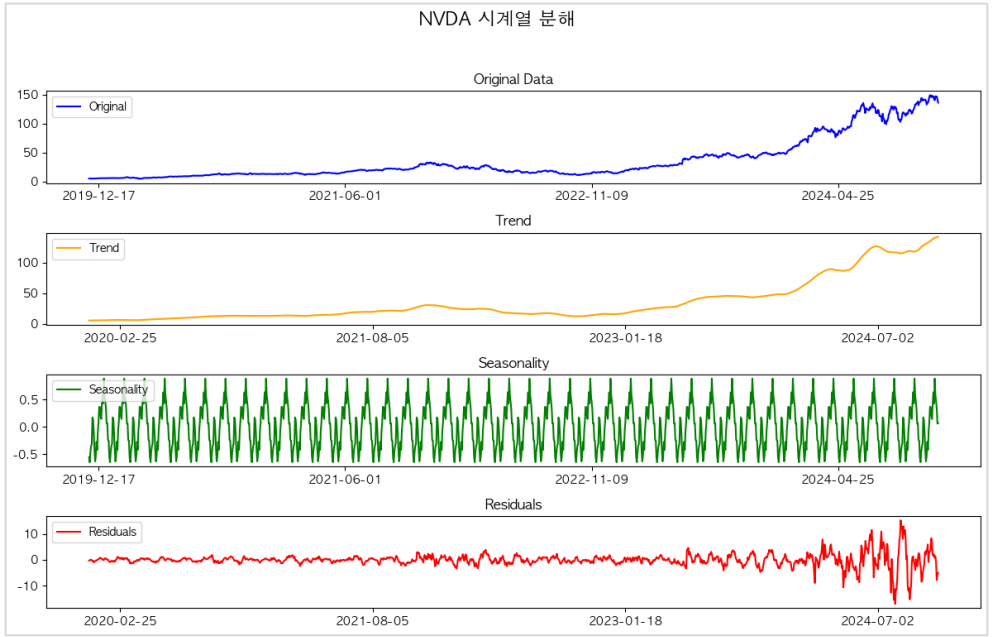
기법	장점	단점	사용 사례
Decomposition	데이터 구조 이해가 쉬움.	예측 능력이 부족함.	데이터 분석, 계절성 및 추세 분리.
ARIMA	간단하고 효율적.	계절성 데이터에 취약, 복잡한 패턴을 학습하기 어려움.	비계절성 시계열 데이터 (예: 주식, 매출).
SARIMA	계절성 데이터 처리 가능.	복잡한 계절성 및 다변량 데이터를 처리하기 어려움.	계절성 데이터 (예: 월별 매출, 전력 소비).
LSTM	비선형적이고 복잡한 패턴을 학습 가능.	계산 비용이 높고, 많은 데이터 필요.	복잡하고 다중 변수의 데이터 (예: 금융, 날씨).

4. 분석/모델링

총 7개 주식의 시계열 분해 작업 진행



[(예시) 아마존 최근 5년간 종가 시계열 분해]



[(예시) 엔비디아 최근 5년간 종가 시계열 분해]

4. 분석/모델링

1) Decompostion기반 예측 모델링 (계절성 값 10으로 선정)

```
#Decomposition모델 예측

period = 10 #period = 10

for stock in stock_columns:

    # 데이터 정렬
    train_stock_data = train_data[['Date', stock]].dropna().sort_values('Date')
    train_stock_data.set_index('Date', inplace=True)

    test_stock_data = test_data[['Date', stock]].dropna().sort_values('Date')
    test_stock_data.set_index('Date', inplace=True)

    # 시계열 분해 (추세, 계절성, 잔차)
    decomposition = seasonal_decompose(train_stock_data[stock], model='additive', period=period)

    # 추세 예측 (Linear Regression 사용)
    trend = decomposition.trend.dropna()
    trend_index = np.arange(len(trend)).reshape(-1, 1)  # 인덱스를 Feature로 사용
    trend_model = LinearRegression()
    trend_model.fit(trend_index, trend.values)  # 선형 회귀 모델 학습

    # 추세 예측 연장
    future_index = np.arange(len(trend), len(trend) + test_size).reshape(-1, 1)
    future_trend = trend_model.predict(future_index)

    # 계절성 예측
    seasonality = decomposition.seasonal
    future_seasonality = np.tile(seasonality[-period:], int(np.ceil(test_size / period)))[:test_size]

    # 예측값 합산
    predictions = future_trend + future_seasonality
```

- 학습데이터와 테스트데이터를 분리

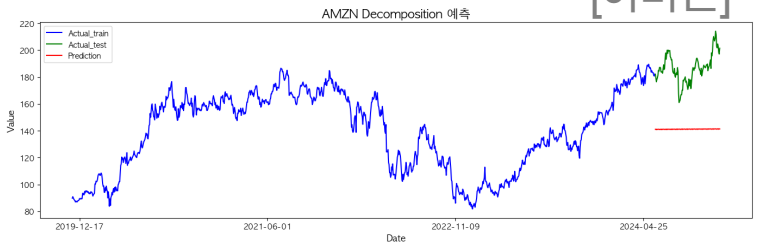
- seasonal_decompose를 사용하여 시계열 데이터를 분해

- 계절 주기를 10으로 설정

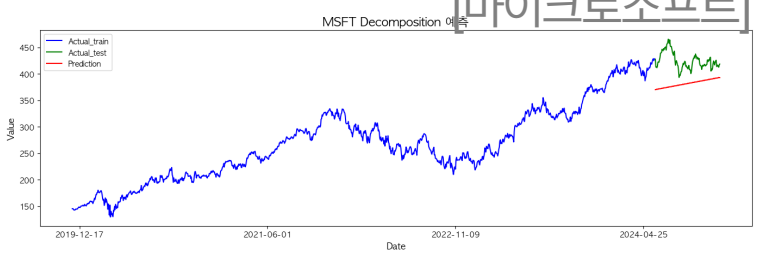
4. 분석/모델링

1) Decomposition 기반 예측 값

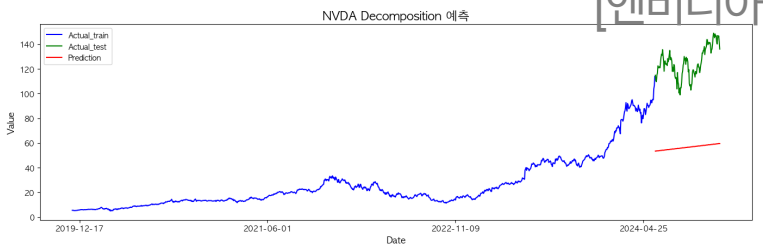
[아마존]



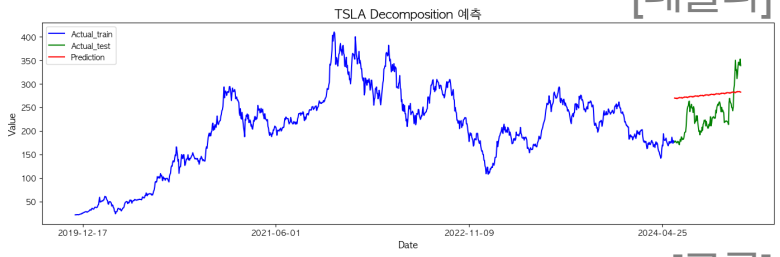
[마이크로소프트]



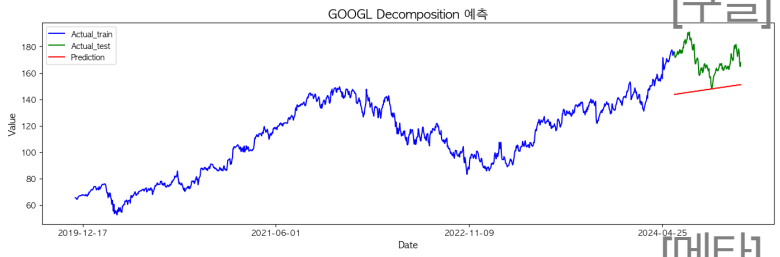
[엔비디아]



[테슬라]

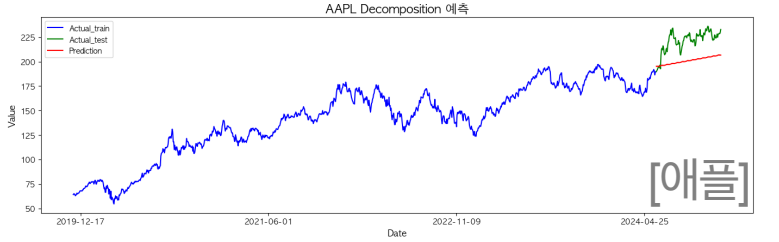


[구글]



[메타]

[애플]



4. 분석/모델링

2) Arima 기반 예측

```
# auto_arima를 사용해 최적의 ARIMA 모델 찾기

for stock in stock_columns:

    # 데이터 정렬
    train_stock_data = train_data[['Date', stock]].dropna().sort_values('Date')
    train_stock_data.set_index('Date', inplace=True)

    test_stock_data = test_data[['Date', stock]].dropna().sort_values('Date')
    test_stock_data.set_index('Date', inplace=True)

    model = auto_arima(train_stock_data, seasonal=False,trace=True, suppress_warnings=True)

    # 예측값 얻기
    forecast = model.predict(n_periods=test_size) #test_size만큼 예측

    # 시각화
    plt.figure(figsize=(12, 8))
    plt.plot(train_stock_data[stock], label='Actual_train', color='blue')
    plt.plot(test_stock_data[stock], label='Actual_test', color='green')
    plt.plot(test_stock_data.index,forecast, label='Prediction', color='red')
    plt.title(f"{stock} ARIMA 예측", fontsize=16)
    plt.gca().xaxis.set_major_locator(mdates.DayLocator(interval=365))
    plt.xlabel("Date", fontsize=12)
    plt.ylabel("Value", fontsize=12)
    plt.legend()
    plt.tight_layout(rect=[0, 0.03, 1, 0.95])
    plt.show()
```

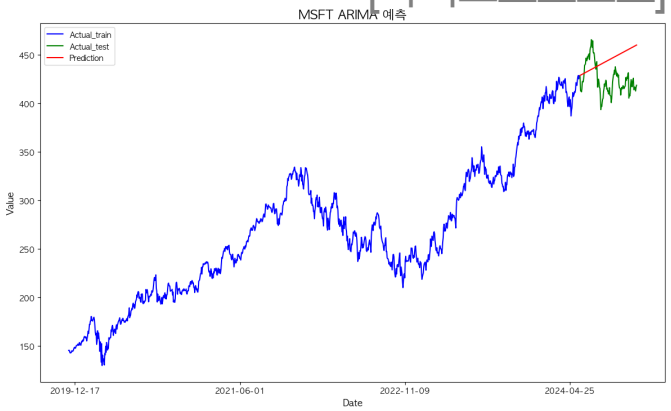
- Auto Arima 라이브러리 활용 시행

4. 분석/모델링

2) Arima 기반 예측

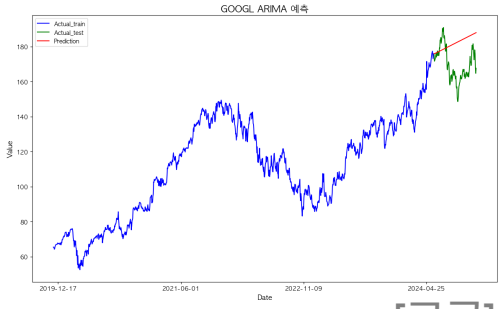
[아마존]

[마이크로소프트]



[엔비디아]

[테슬라]



[구글]

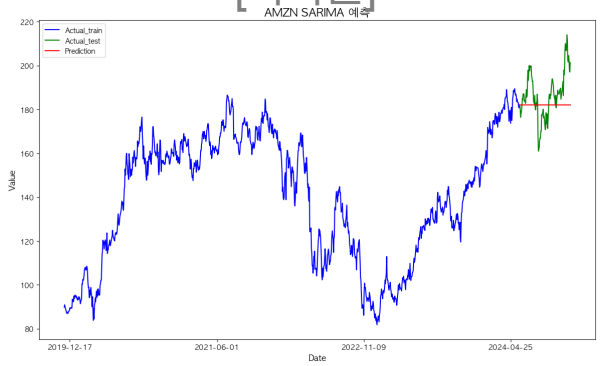
[메타]

[애플]

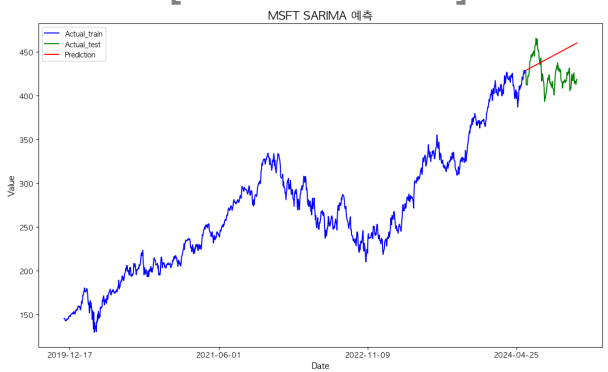
4. 분석/모델링

3) Sarima 기반 예측 (계절성 값 10)

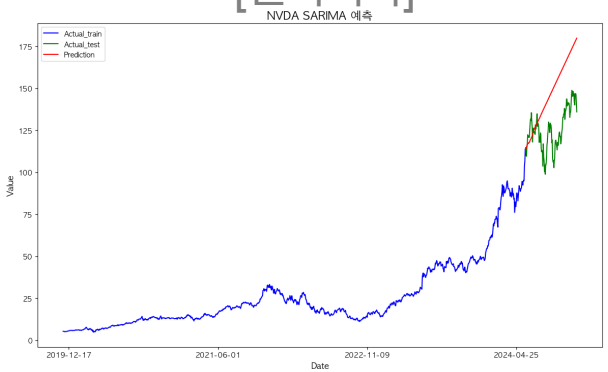
[아마존]



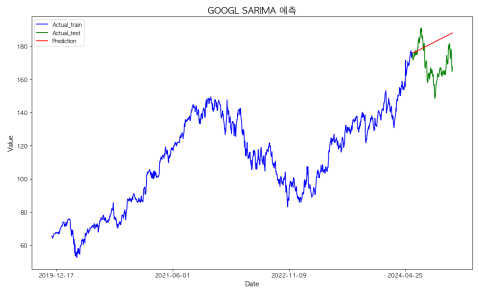
[마이크로소프트]



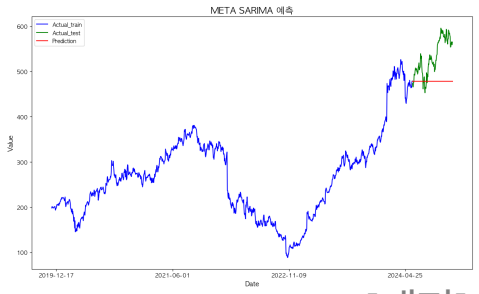
[엔비디아]



[테슬라]



[구글]



[메타]

[애플]

4. 분석/모델링

4) LSTM 기반 예측 모델링

```
# 학습/테스트 데이터 분리
split = int(len(X) * 0.9)
X_train, X_test = X[:split], X[split:]
y_train, y_test = y[:split], y[split:]

model = Sequential()

# 첫 번째 LSTM 레이어
model.add(LSTM(units=50, return_sequences=True, input_shape=(X_train.shape[1], 1)))
model.add(Dropout(0.2))

# 두 번째 LSTM 레이어
model.add(LSTM(units=50, return_sequences=True))
model.add(Dropout(0.2))

# 세 번째 LSTM 레이어
model.add(LSTM(units=50))
model.add(Dropout(0.2))

# 출력 레이어
model.add(Dense(units=1))

# 모델 컴파일
model.compile(optimizer='adam', loss='mean_squared_error')

model.fit(X_train, y_train, epochs=50, batch_size=32)

predicted_price = model.predict(X_test)
predicted_price = scaler.inverse_transform(predicted_price.reshape(-1, 1))  # 역정규화
```

- 3개의 LSTM 레이어와 1개의 Dense(출력) 레이어로 구성

- 과적합 방지를 위해 20%의 뉴런을 무작위로 비활성화

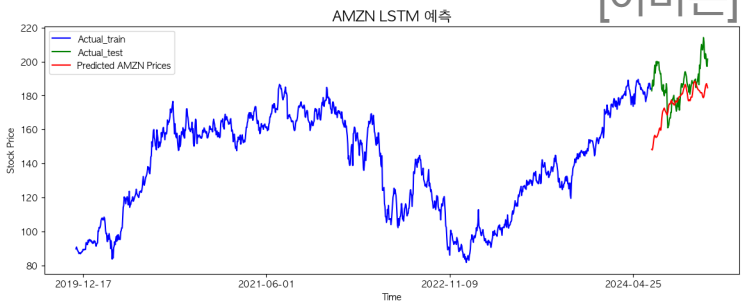
- 최적화 기법: adam (효율적이고 빠른 학습)

- 학습 반복 (Epochs): 50회

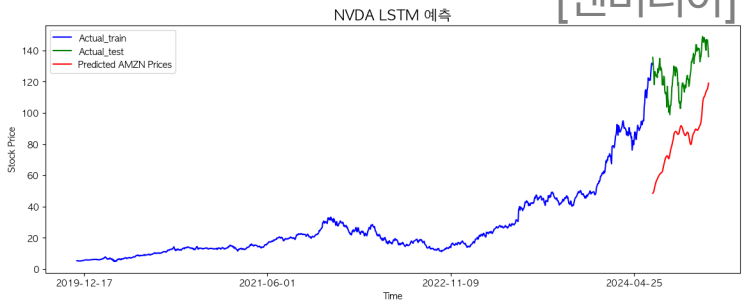
4. 분석/모델링

4) LSTM 기반 예측 모델링

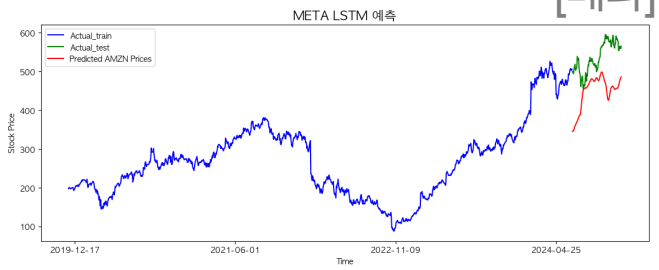
[아마존]

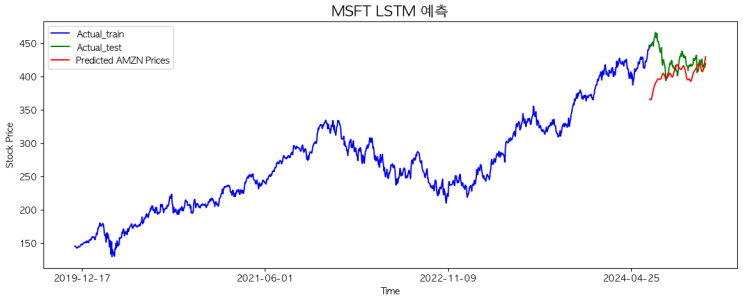


[엔비디아]

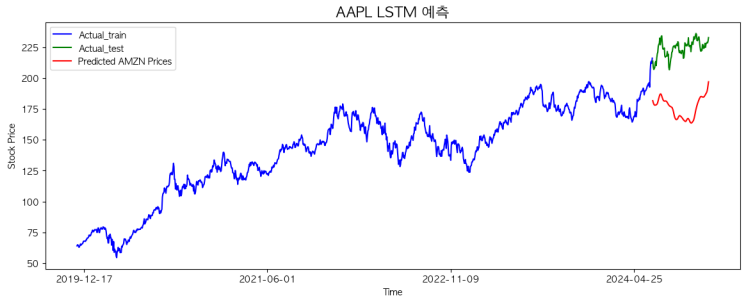


[메타]





[마이크로소프트]

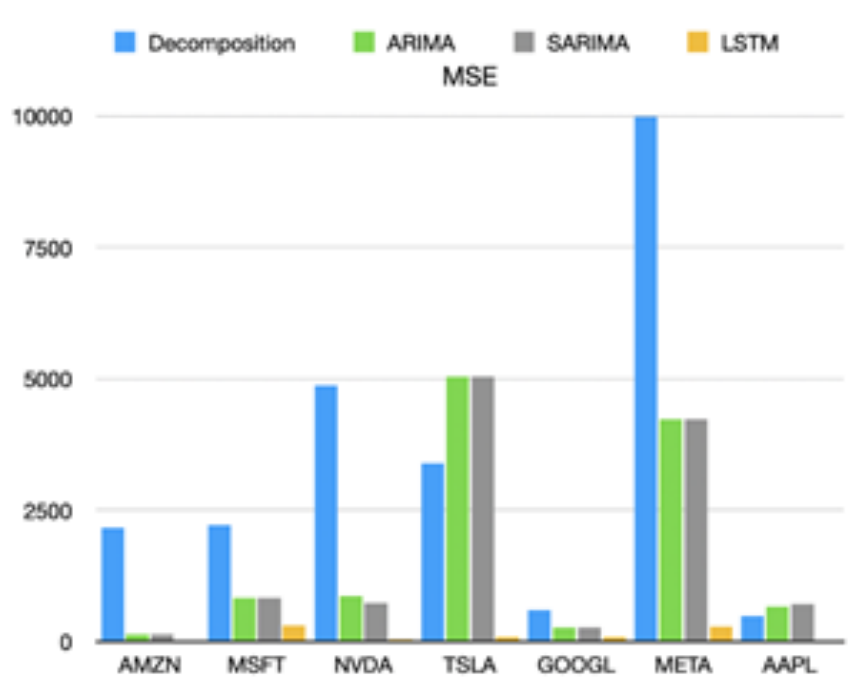


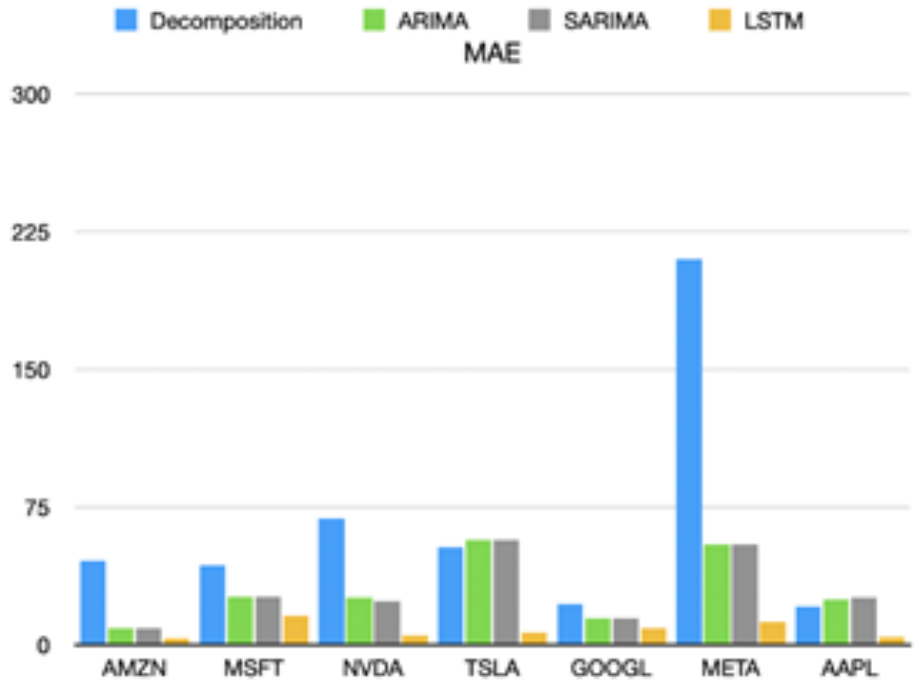
[테슬라]

[구글]

4. 분석/모델링

시계열 모델 MSE/MAE





4. 분석/모델링

5) 회귀분석 – 주식가격을 종속변수로, 주가에 영향을 미치는 요인을 분석(67개)

비트코인 가격	주식 정보	이자 수익	이자 비용	세금 비용	보고된 총 영업 이익	조정된 매출원가
매출원가	영업 수익	기본 평균 주식수	희석 평균 주식수	비영업 이자 비용	희석 주당 순이익 (주주 귀속)	보통주 주주 귀속 순이익
비영업 이자 수익	판매 일반관리비	계산용 세율	총비용	거래량	나스닥 종가	천연가스 가격
계속 사업 순이익	비지배지분 포함 순이익	비정상 항목 세금 효과	S&P500 종가	원유 가격	소비자 물가지수 (CPI)	희석 주당순이익
정상화된 순이익	기본 주당순이익	금리	순이익	계속 사업 및 비지배지분 포함 순이익	계속 사업 및 중단 사업 순이익	총 수익
금 가격	영업이익 전 이자 및 세금 (EBIT)	정상화된 EBITDA	EBITDA	조정된 감가상각비	주식 분할	브렌트유 가격
은 가격	구리 가격	S&P500 거래량	나스닥 최고가	요일	배당금	S&P500 최고가
나스닥 주식 분할	S&P500 최저가	나스닥 최저가	비영업 순이자 비용	순이자 수익	총 이익	나스닥 거래량
영업이익	S&P500 배당금	개인소비지출 (PCE)	나스닥 시가	상수	기타 수익 비용	나스닥 배당금
S&P500 시가	S&P500 주식 분할	영업비용	세전 수익			

4. 분석/모델링

5) 회귀분석 – 주식가격을 종속변수로, 주가에 영향을 미치는 요인을 분석(67개)

비트코인 가격	주식 정보	이자 수익	이자 비용	세금 비용	보고된 총 영업 이익	조정된 매출원가
매출원가	영업 수익	기본 평균 주식수	희석 평균 주식수	비영업 이자 비용	희석 주당 순이익 (주주 귀속)	보통주 주주 귀속 순이익
비영업 이자 수익	판매 일반관리비	계산용 세율	총비용	거래량	나스닥 종가	천연가스 가격
계속 사업 순이익	비지배지분 포함 순이익	비정상 항목 세금 효과	S&P500 종가	원유 가격	소비자 물가지수 (CPI)	희석 주당순이익
정상화된 순이익	기본 주당순이익	금리	순이익	계속 사업 및 비지배지분 포함 순이익	계속 사업 및 중단 사업 순이익	총 수익
금 가격	영업이익 전 이자 및 세금 (EBIT)	정상화된 EBITDA	EBITDA	조정된 감가상각비	주식 분할	브렌트유 가격
은 가격	구리 가격	S&P500 거래량	나스닥 최고가	요일	배당금	S&P500 최고가
나스닥 주식 분할	S&P500 최저가	나스닥 최저가	비영업 순이자 비용	순이자 수익	총 이익	나스닥 거래량
영업이익	S&P500 배당금	개인소비지출 (PCE)	나스닥 시가	상수	기타 수익 비용	나스닥 배당금
S&P500 시가	S&P500 주식 분할	영업비용	세전 수익			

구 분	값
Adj. R-squared:	0.986

→98.6%의 설명력을 가짐

4. 분석/모델링

5) 회귀분석 – 주식가격을 종속변수로, 주가에 영향을 미치는 요인을 분석(67개)

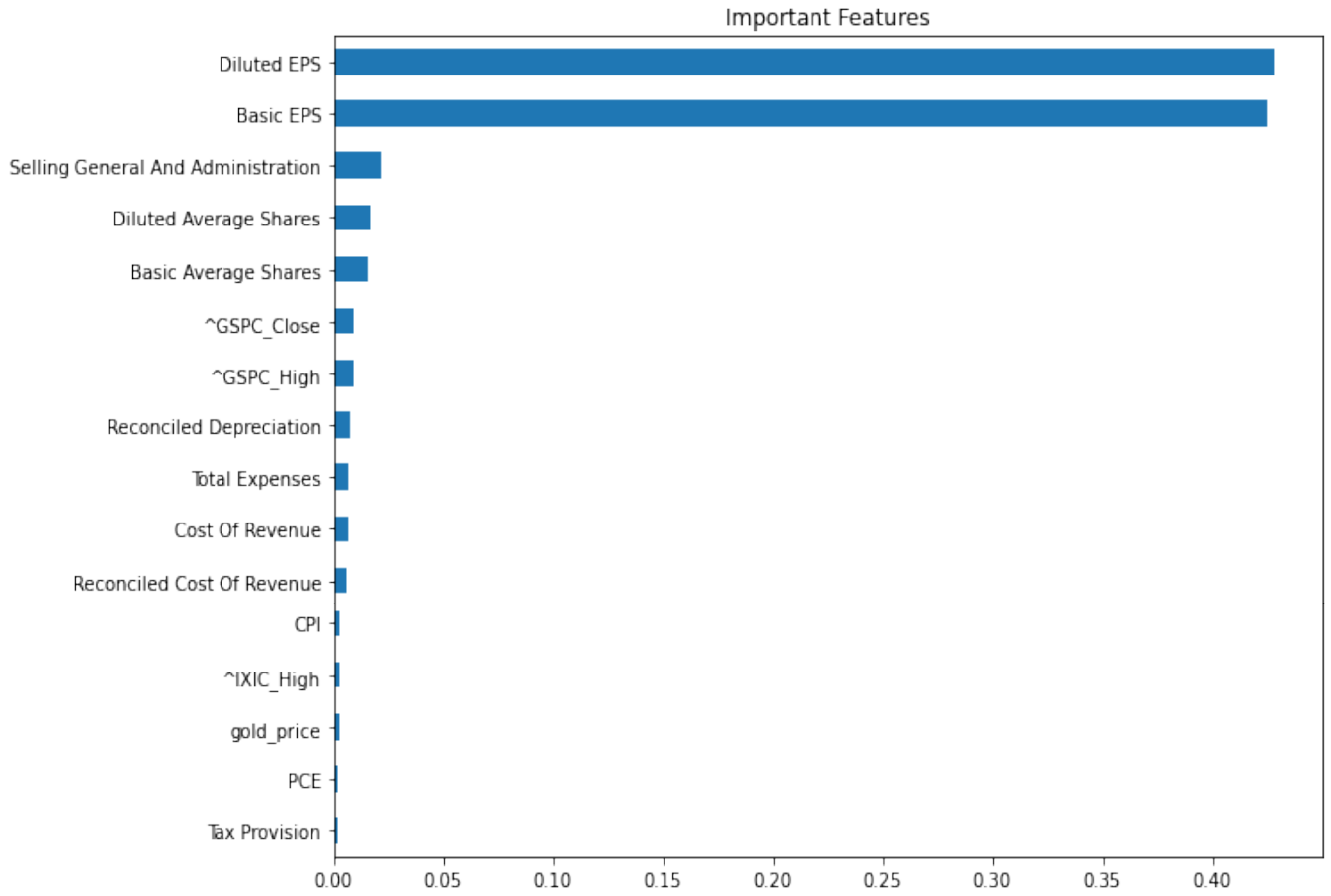
변수	coef	std err	t	P>|t|	절대값 coef
Operating Revenue	(93110)	6.65E+03	-14.01	0	93110
Net Income Common Stockholders	(62190)	9.08E+03	-6.847	0	62190
Net Income Continuous Operations	61010	2.76E+04	2.208	0.027	61010
Net Income Including Noncontrolling Interests	61010	2.76E+04	2.208	0.027	61010
Diluted Average Shares	(24290)	2053.999	-11.826	0	24290
Basic Average Shares	24070	2034.293	11.833	0	24070
Reconciled Cost Of Revenue	21340	3348.233	6.375	0	21340
Cost Of Revenue	21340	3348.233	6.375	0	21340
Total Expenses	19180	6634.912	2.891	0.004	19180
Diluted NI Availto Com Stockholders	18280	1494.896	12.23	0	18280
Tax Provision	(16010)	1.96E+03	-8.182	0	16010
Total Operating Income As Reported	3715	809.652	4.589	0	3715
Interest Expense	445	66.703	6.677	0	445
Interest Expense Non Operating	445	66.703	6.677	0	445
Selling General And Administration	(419)	8.87E+01	-4.727	0	419
Interest Income	(363)	59.881	-6.063	0	363
Interest Income Non Operating	(363)	59.881	-6.063	0	363
Tax Effect Of Unusual Items	323	1.58E+02	2.039	0.042	323
^IXIC_Close	311	122.318	2.54	0.011	311
stock(주식 정보{APPL, AMZN 등등})	75	6.276	11.941	0	75
Tax Rate For Calcs	(41)	12.58	-3.274	0.001	41
Volume	25	9.11E+00	2.783	0.005	25
bitcoin_price	(14)	3.88	-3.574	0	14
naturalgas_price	(8)	3.735	-2.256	0.024	8

→ 총 24개의 독립변수가 유의미
(P-value가 0.05미만)

→ 학습된 회귀식으로 예측된 성능
MAE : 43.28 , MSE : 2674.45

4. 분석/모델링

랜덤포레스트 기반 변수 중요도 분석 추가 시행



→ Diluted EPS(희석 주당 순이익)

→ Basic EPS(주당 순이익)

2가지 요인이 유의미한 변수로 측정

→ 학습된 값으로 예측된 성능
MAE : 13.39 , MSE : 283.06

5. 분석결과 정리

<< 보완점 >>

- 특정 기술주 위주의 선정으로, 좀 더 보편화된 예측 모델링 보완 필요
  (현재 7개 → 나스닥 100 편입 종목 100개로 확대 등)

- 5년간의 데이터를 활용했으나, 기간이 짧아 과적합 이슈 발생 가능
  (기간 내 전쟁 등 특정 이벤트 등이 과도한 영향을 미칠 가능성이 있음)

- 실제 유의미한 값이 맞는지, 향후 주가 추세를 보고 추가적인 검증이 필요함

End of Document