Newbie Investors! Should You Sell or Hold?

Buy/Sell Recommendation For Stock Market Beginners

2023-1 Team timeseries
Kim Min/ Kim Dong-hwan/ Seo Yoo-jin/ Lee Soo-rin/ Jang Da-yeon

1 Introduction

Background of Topic Selection

Stock Market

The stock market quickly worsens, leading to significant losses for many investors.

Novice investors

Novice investors, lacking investment knowledge, face substantial losses

Let's create a service to provide easily accessible investment insights for novice investors!

1 Introduction

Background of Topic Selection

Team Time Series Topic Analysis

Buy/Sell Recommendation Service For Stock Market Beginners

Input information affecting stock price fluctuations into the model

→ the model learns to recommend buy/sell decisions

Providing investors with recommendations to buy or sell based on the current situation!

▶ Offering a indicator that provides insights/information for investment decisions

2 Data overview

Collected Data

SK Hynix

Shinhan Financial Group

Hyundai Motor Company

Selected as analysis targets for

building a robust model applicable to various stocks

2 Data overview

Collected Data

Individual indicators

- ✓ Stock price trend data
- ✓ Investor transaction performance data
- ✓ Foreign ownership data
- ✓ Short selling data
- ✓ Domestic news data
- ✓ English news data
- ✓ Naver Stock Discussion Forum data
- ✓ Naver search volume data

Common indicators

- ✓ KOSPI data
- ✓ Bitcoin trading data
- ✓ Economic sentiment index
- ✓ News sentiment index
- ✓ Industrial production index
- ✓ Consumer price index
- ✓ Consumer confidence index
- ✓ Consumer sentiment index
- ✓ Unemployment rate
- ✓ Bank of Korea base rate
- ✓ Exchange rate

2 Data overview

Collected Data

Individual indicators

- ✓ Stock price trend data
- ✓ Investor transaction performance data
- ✓ Foreign ownership data
- ✓ Short selling data
- ✓ Domestic news data
- ✓ English news data
- ✓ Naver Stock Discussion Forum data
- ✓ Naver search volume data

Common indicators

- ✓ KOSPI data
- ✓ Bitcoin trading data
- ✓ News sentiment index
- ✓ Consumer price index
- ✓ Consumer confidence index

Utilizing data that reflects public opinion to grasp fluctuations based on sentiment and investor psychology!

2 Data overview

Collected Data

Individual indicators

- ✓ Stock price trend data
- ✓ Investor transaction performance data
- ✓ Foreign ownership data
- ✓ Short selling data
- ✓ Domestic news data
- ✓ English news data
- ✓ Naver Stock Discussion Forum data
- ✓ Naver search volume data

Common indicators

- ✓ ata
- ✓ Bitcoin trading data
- ✓ Economic sentiment index
- ✓ News sentiment index
- ✓ Industrial production index
- ✓ Consumer price index
- ✓ Consumer confidence index
- ✓ Consumer sentiment index
- ✓ Unemployment rate
- ✓ Bank of Korea base rate
- ✓ Exchange rate

Data collection period: 2017/06/08 ~ 2023/03/31

The data collection period is aligned with the launch date of the Naver Stock Discussion Forum service, which started on June 8, 2017.

3 Data Preprocessing

Sentiment analysis of Korean news articles

Data New articles Headlines from 2017-06-08 ~ to 2023-03-31.

Date	Headline	label
2017-06-08	[fnRASSI] 장마감, 거래소 하락 종목 (신한 -8.4%)	-1
2017-06-08	'오늘의 증시 메모 [6월 8일]	0
⋮	⋮	
2023-03-31	신한금융 "데이터센터 전력 재생에너지로 조달 "	1
2023-03-31	신한금융, 데이터센터 전력 100% 재생 에너지 추진	1

Date	Sentiment score
2017-06-08	0.148148
2017-06-09	0.142857
⋮	⋮
2023-03-30	0.166667
2023-03-31	0.384615

Labeling through the trained model

3 Data Preprocessing

Sentiment analysis of Foreign news articles

Data

The data comprises a total of 9,995 rows, crawled from CNN from July 2017 to March 2023

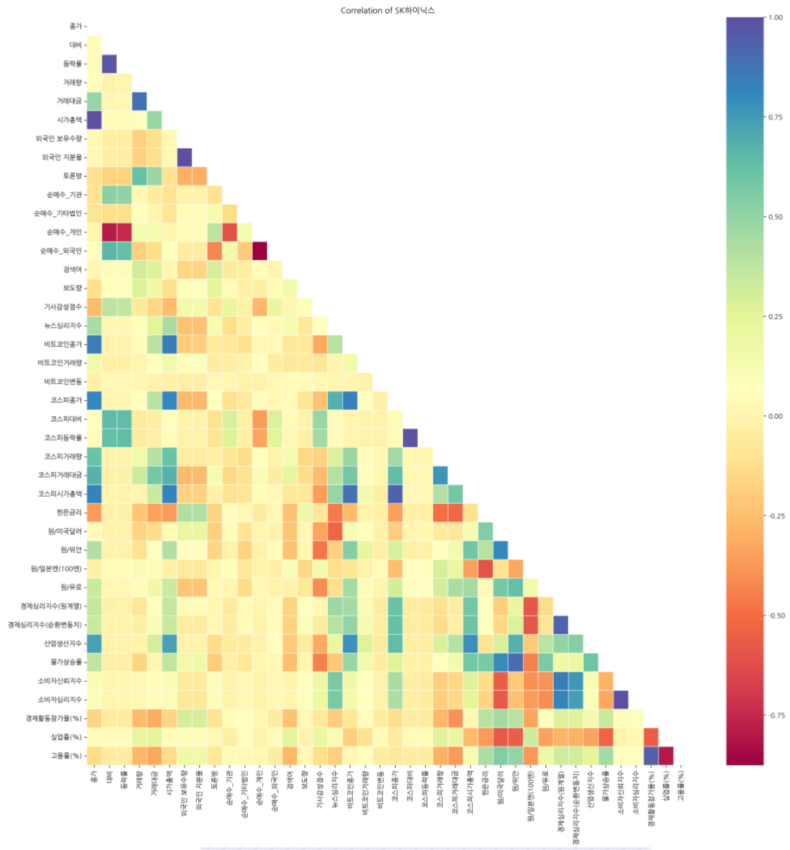
Date	Headline	score
2017-07-19	US general warns of … control killer robots	-0.6908
2017-07-21	Pompeo signals want for N. Korea regime change	0.0772
⋮	⋮	
2023-03-25	How AI turned the ancient sport of Go upside down	0
2023-03-29	US & South Korea stage joint military drills	0

Date	Sentiment score
2017-07-14	0.365775
2017-07-15	0.261127
⋮	⋮
2023-03-30	0.120400
2023-03-31	0.079550

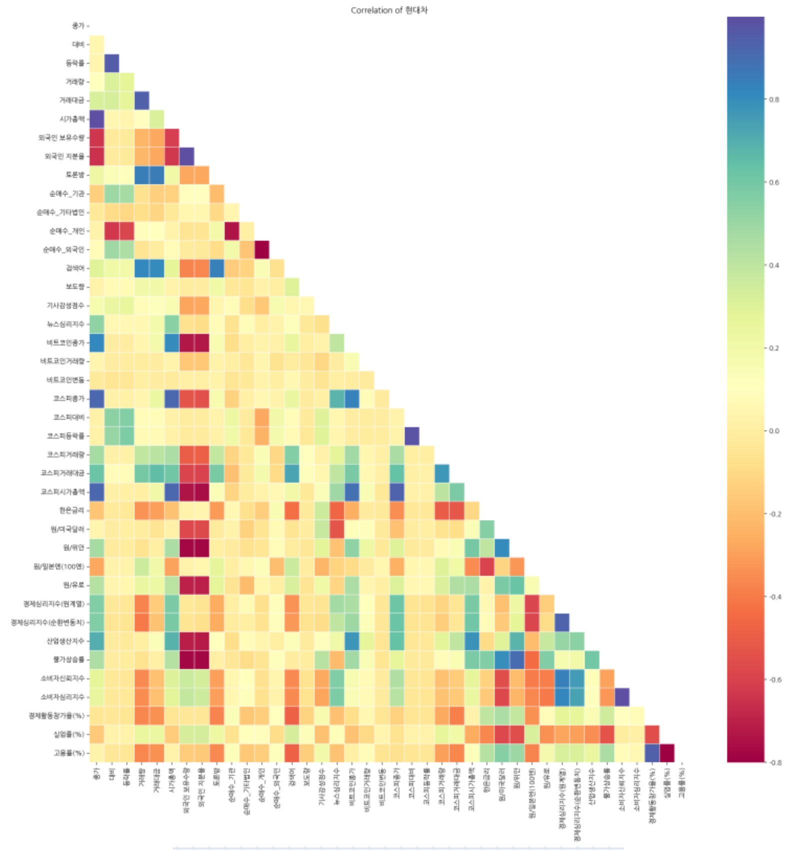
Get a score using nltk.sentiment.vader

4 EDA

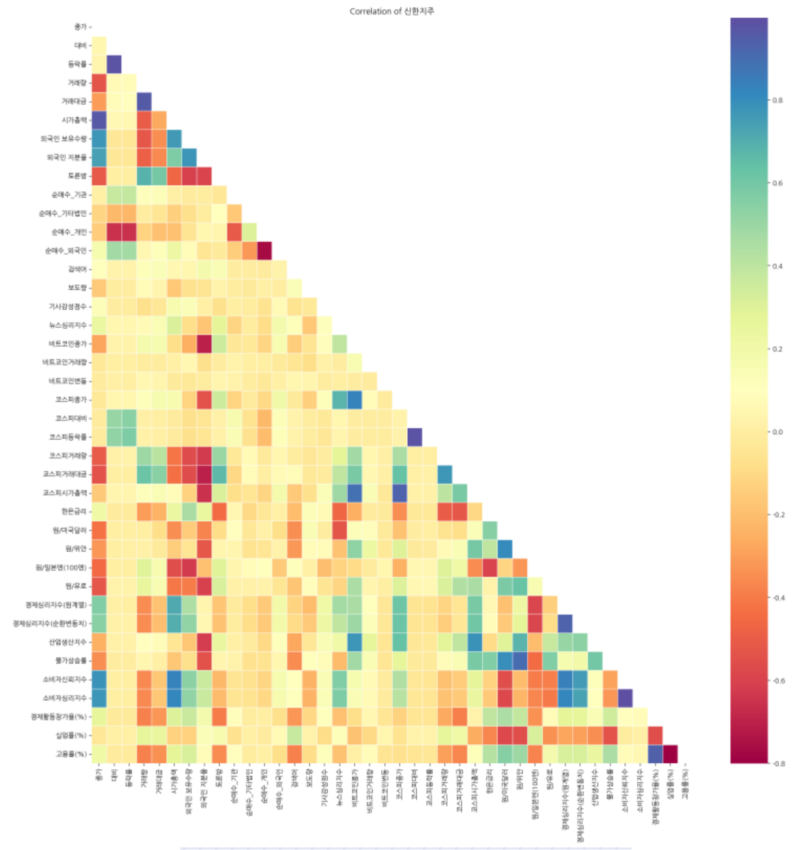
Correlation with Fluctuation rate



SKHynix



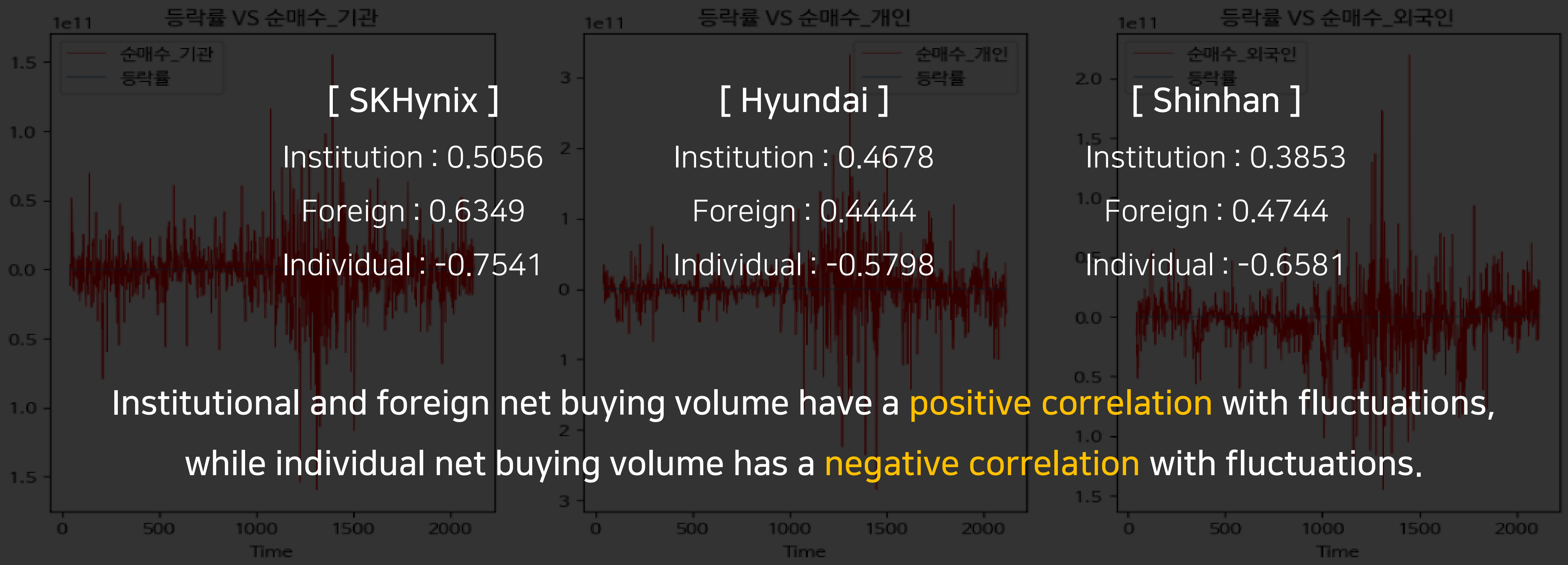
Hyundai



Shinhan

4 EDA

Net purchases by institutions, individuals, foreign



[SKHynix]

Institution : 0.5056

Foreign : 0.6349

Individual : -0.7541

[Hyundai]

Institution : 0.4678

Foreign : 0.4444

Individual : -0.5798

[Shinhan]

Institution : 0.3853

Foreign : 0.4744

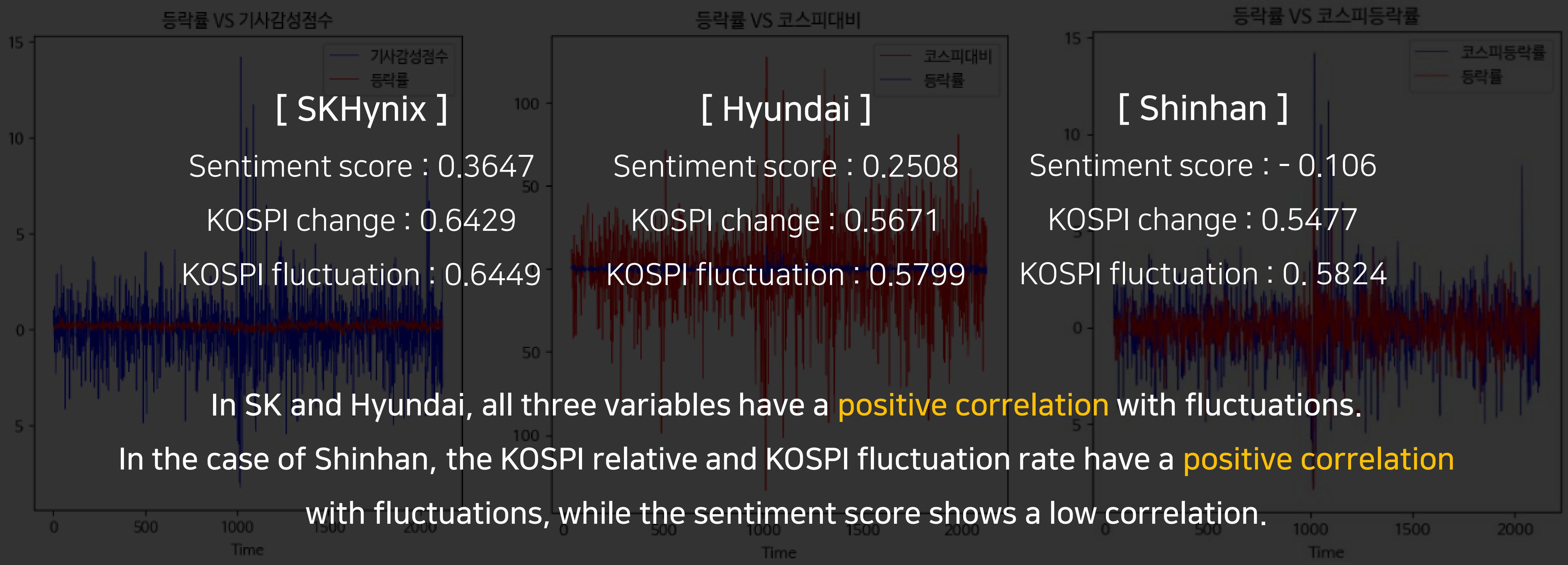
Individual : -0.6581

Institutional and foreign net buying volume have a positive correlation with fluctuations, while individual net buying volume has a negative correlation with fluctuations.

Shinhan

4 EDA

KOSPI change, KOSPI fluctuation rate



[SKHynix]

Sentiment score : 0.3647

KOSPI change : 0.6429

KOSPI fluctuation : 0.6449

[Hyundai]

Sentiment score : 0.2508

KOSPI change : 0.5671

KOSPI fluctuation : 0.5799

[Shinhan]

Sentiment score : - 0.106

KOSPI change : 0.5477

KOSPI fluctuation : 0. 5824

In SK and Hyundai, all three variables have a positive correlation with fluctuations.

In the case of Shinhan, the KOSPI relative and KOSPI fluctuation rate have a positive correlation with fluctuations, while the sentiment score shows a low correlation.

Shinhan

1 Summary of week 1

Labeling for Y variable

Fluctuation(%)

Correlation between X and Y

Polyserial correlation

Correlation between X and labeled Y

Label based on the point where the
two correlations become maximally similar!

1-day FR threshold: 3%

3-day FR threshold: 5%

FR : Fluctuation Rate

1-day(tomorrow) FR <= -3% : SELL

1-day(tomorrow) FR >= 3% : BUY

Otherwise : HOLD (maintain)

4 EDA

Labeling for Y variable;Compare Correlation of Y and labeled Y

SK Hynix (daily fluctuation rate)			
net_purchase_insti	0.5056	Sentiment score	0.3647
net_purchase_foreign	0.6349	KOSPI change	0.6429
net_purchase_indi	-0.7541	KOSPI fluctuation	0.6449

Hyundai (daily fluctuation rate)			
net_purchase_insti	0.4678	Sentiment score	0.2508
net_purchase_foreign	0.4444	KOSPI change	0.5671
net_purchase_indi	-0.5798	KOSPI fluctuation	0.5799

Shinhan (daily fluctuation rate)			
net_purchase_insti	0.3853	Sentiment score	-
net_purchase_foreign	0.4744	KOSPI change	0.5477
net_purchase_indi	-0.6581	KOSPI fluctuation	0.5824

Interpretation

Variables with positive correlation indicates that as the variable increases, the chances of BUY increases, and the SELL decreases

4 EDA

Labeling for Y variable;Compare Correlation of Y and labeled Y

SK Hynix (daily fluctuation rate)			
net_purchase_insti	0.5056	Sentiment score	0.3647
net_purchase_foreign	0.6349	KOSPI change	0.6429
net_purchase_indi	-0.7541	KOSPI fluctuation	0.6449

Hyundai (daily fluctuation rate)			
net_purchase_insti	0.4678	Sentiment score	0.2508
net_purchase_foreign	0.4444	KOSPI change	0.5671
net_purchase_indi	-0.5798	KOSPI fluctuation	0.5799

Shinhan (daily fluctuation rate)			
net_purchase_insti	0.3853	Sentiment score	-
net_purchase_foreign	0.4744	KOSPI change	0.5477
net_purchase_indi	-0.6581	KOSPI fluctuation	0.5824

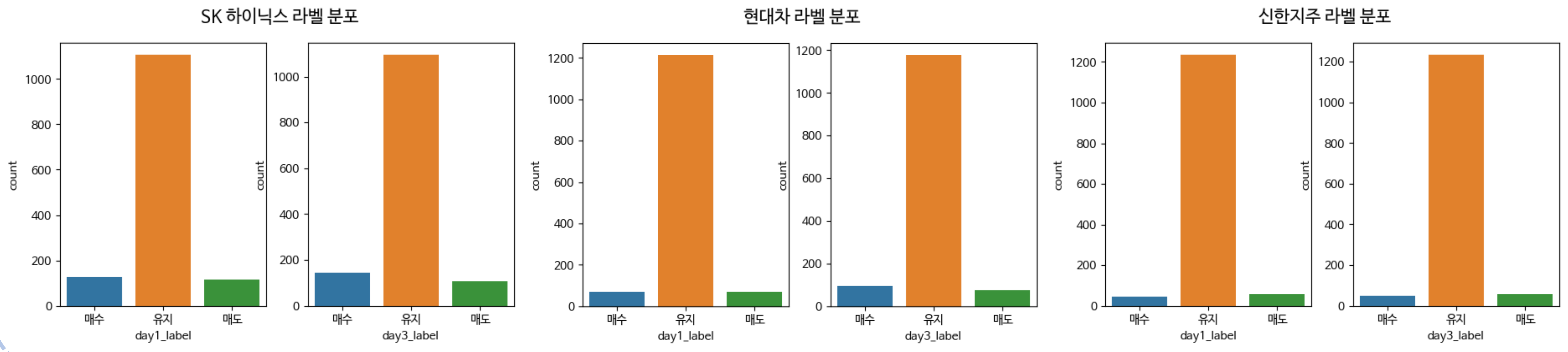
Interpretation

Variables with negative correlation indicates that as the variable increases, the chances of BUY decreases, and the SELL increases

4 EDA

Class imbalance in labeled Y

Each stock's distribution of labeled Y



The overall accuracy is high, but the model predictions are biased towards 'hold', resulting in issues with properly predicting 'buy' and 'sell'

5 Modeling process

Variable Selection

1

[Causality Test]

2

[VIF]

3

[Feature Importance]

4

[KS test]

5

[Full Model]

But for almost every case,
Full model have the best performance…

5 Modeling process

Modeling overview

Input X variables at the current time(minmax scaled /full model)

Output Recommendations based on Stock price fluctuation Predictions for the next 5 days

- Recommend Buy if the rate of change is expected to rise by more than 5% over the 5 days from t+1 to t+6.
- Recommend Sell if the rate of change is expected to fall by more than 5% over the 5 days from t+1 to t+6.
- Recommend Hold if the absolute rate of change is expected to be within 5% over the 5 days from t+1 to t+6.

Fit the model to SK Hynix data, which has the least class imbalance, and then apply the same model to all three stocks.

Model Selection Criteria: How well does it predict buy/sell? (Among models with high accuracy in buy/sell, select the one with the highest f1-score.)

5 Modeling process

Customize optuna score

01 Mean accuracy of BUY, SELL, HOLD

```
cm=confusion_matrix(y_test, y_pred)
bacc=cm[0,0]/sum(cm[0]) # BUY accuracy
sacc=cm[2,2]/sum(cm[2]) # SELL accuracy
macc=cm[1,1]/sum(cm[1]) # HOLD accuracy
rst=np.mean([bacc,sacc,macc])
```

02 Mean accuracy of BUY, SELL

```
cm=confusion_matrix(y_test, y_pred)
bacc=cm[0,0]/sum(cm[0]) # BUY accuracy
sacc=cm[2,2]/sum(cm[2]) # SELL accuracy
rst=np.mean([bacc,sacc])
```

03 Mean F1 score and accuracy of BUY,SELL

```
cm=confusion_matrix(y_test, y_pred)
bacc=cm[0,0]/sum(cm[0]) # BUY accuracy
sacc=cm[2,2]/sum(cm[2]) # SELL accuracy
f1=sum(scores)/len(scores)
rst=np.mean([bacc,sacc,f1])
```

04 Mean accuracy and precision of BUY,SELL

```
cm=confusion_matrix(y_test, y_pred)
bacc=cm[0,0]/sum(cm[0]) # BUY accuracy
sacc=cm[2,2]/sum(cm[2]) # SELL accuracy
bpre=cm[0,0]/np.sum(cm, axis=0)[0] #BUY Precision
spre= cm[2,2]/np.sum(cm, axis=0)[2] #SELL Precision
rst=np.mean([bacc,sacc,bpre,spre])
```

5 Modeling process

List of models attempted

Models

- LSTM
- CNN
- SVM
- Logistic regression
- Naïve Bayes
- XGB
- LGBM
- LGBM regressor
- LGBM-CNN regressor

When tasks pile up in front of me,
It become even less motivated to do them

Selecting the best model using Optuna

based on 4 scores

6 Final model

Final model introduction

XGB classifier

Used Data : SK Hynix

variables : Full Model

evaluation : custom optuna score

classification : 다중 분류(매수, 매도, 유지)

LSTM regressor

Used Data : SK Hynix

variables : VIF

highlight : predict labeled Y

by regression and classify

using Threshold function afterward

6 Final model

XGB Classifier

1. Variable selection

Data : SK Hynix

variables : Full Model

Using SK Hynix data, which exhibits the least class imbalance, for hyperparameter tuning. Afterwards, apply tuned model to remaining stocks

Variables
X: 'Closing Price', 'Price Change', 'Fluctuation Rate', 'Volume', 'Transaction Amount', 'Market Cap', 'Foreign Ownership Quantity', 'Foreign Ownership Ratio', 'Discussion Forum', 'Net Purchases by Institutions', 'Net Purchases by Other Corporations', 'Net Purchases by Individuals', 'Net Purchases by Foreigners', 'Search Volume', 'News Coverage', 'Article Sentiment Score', 'Sentiment Index', 'Bitcoin Closing Price', 'Bitcoin Volume', 'Bitcoin Fluctuation', 'KOSPI Closing Price', 'KOSPI Fluctuation Rate', 'KOSPI Volume', 'KOSPI Transaction Amount', 'KOSPI Market Cap', 'Bank of Korea Interest Rate', 'KRW/USD ', 'KRW/CNY ', 'KRW/JPY ', 'KRW/EUR ', 'Economic Sentiment Index (Original Series)', 'Economic Sentiment Index (Cyclically Adjusted)', 'Industrial Production Index', 'Inflation Rate', 'Consumer Confidence Index', 'Consumer Sentiment Index', 'Labor Force Participation Rate (%)', 'Unemployment Rate (%)', 'Employment Rate (%)', 'KOSPI Comparison'
Y : 'day5_label'

6 Final model

XGB Classifier

2. Label encoding

Perform label encoding with target label (day5_label)

Buy	0
Hold	1
Sell	2

3. MinMax Scaling

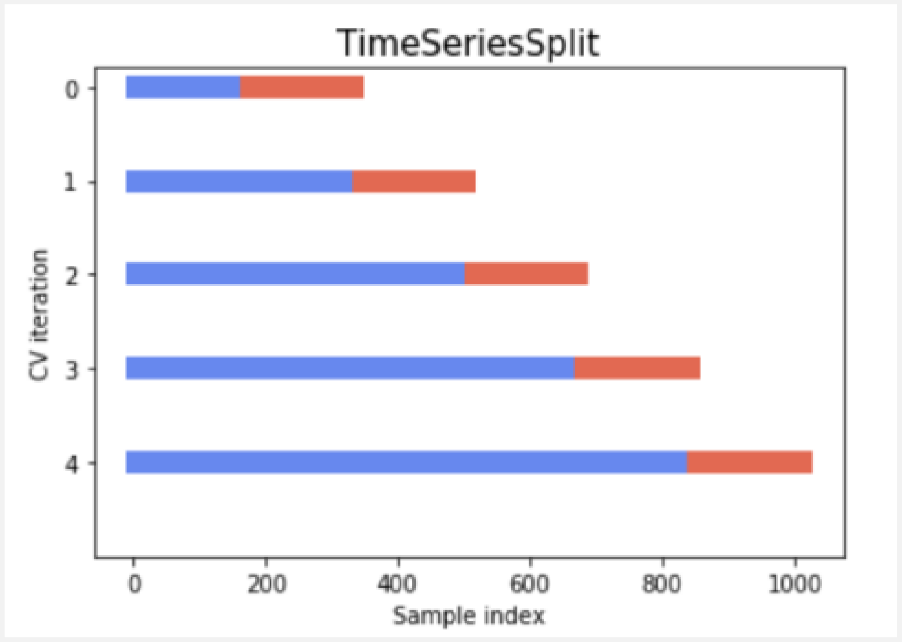
$$\frac{x - Min(X)}{Max(X) - Min(X)}$$

- Apply MinMax scaling to every continuous X variables
- Normalization scaling (range: [0, 1])
- To reduce the scale difference between variables to fitting into the same hyperparameters

6 Final model

XGB Classifier

3. Expanding Window CV



Utilized Expanding Window CV with n_splits = 4

Since Split increases size of validation set go decreases,

which can lead to severe class imbalance issues within a single validation set.

6 Final model

XGB Classifier

4. Class weights

Use the inverse of the proportion of
each class as the sample weight for that class!

```
class_weights = class_weight.compute_sample_weight( class_weight='balanced', y=y_train )
```

▶ Function that calculates the sample weights for each class for the imbalanced training data

```
xgb_model=xgb.XGBClassifier(**params, random_state = 42)
xgb_model.fit(x_train, y_train, sample_weight=classes_weights)
```

You can utilize it with the fit function in this way!

6 Final model

XGB Classifier

5. Optuna hyperparameter tuning

XGBoost Classifier hyperparameters

- max_depth: Maximum depth of the tree; deeper trees are more complex
- learning_rate : the step size at each iteration while moving toward a minimum of a loss function
- n_estimators : number of trees
- min_child_weight : Minimum Hessian weight needed for a split
- gamma : Minimum loss reduction required for a split
- subsample : Data sampling ratio for each tree.
- colsample_bytree : Feature sampling ratio for each tree
- reg_alpha :   L1 regularization weight
- reg_lambda : L2 regularization weight

6 Final model

XGB Classifier

5. Optuna hyperparameter tuning

To create a model that predicts 'buy' and 'sell' well, which can directly impact trading profits.

exclude the high-proportion 'maintain' class and include the accuracy of 'buy' and 'sell' in evaluation metrics

Accuracy

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Precision

$$\frac{TP}{TP + FP}$$

To prevent the model from excessively predicting only ‘buy’ and ‘sell’, and to maintain predictive power for ‘hold’, include precision of ‘buy’ and ‘sell’ in the evaluation metrics

Optuna evaluation metircs

Average Buy accuracy, Buy precision,
Sell accuracy, and Sell precision

6 Final model

XGB Classifier

5. Optuna hyperparameter tuning

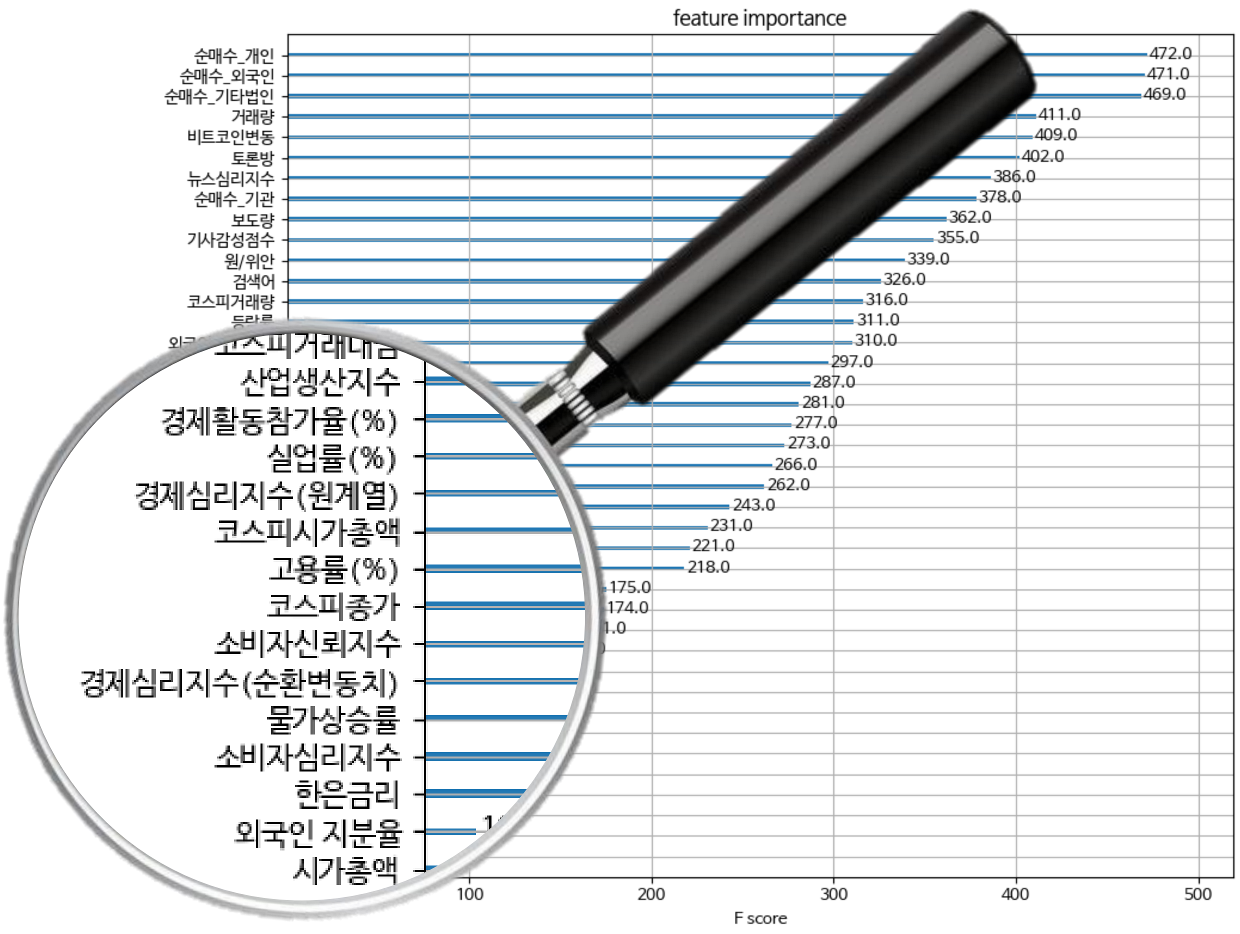
net purchase(institution/foreign/individual/other),Bitcoin volatility, Discussion forum post count, News sentiment index, Sentiment score, Article coverage volume, Search volume

Net purchase data and public opinion and investor sentiment related data appears as important variables

6 Final model

XGB Classifier

5. Optuna hyperparameter tuning



Industrial Production Index, Labor Force Participation Rate, Unemployment Rate, KOSPI, Economic Sentiment Index, Employment Rate, Inflation Rate, Bank of Korea Interest Rate …

On the other hand,

macroeconomic-related data appears

as relatively less important variables

6 Final model

XGB Classifier

6. Prediction

- with test set

[SK Hynix]

```
========== SK 하이닉스 ==========
[[27  7  3]
 [45 90 63]
 [ 6  7 33]]

전체 정확도 :  0.5338078291814946
전체 f1-score :  0.5563170430999059

매수 정확도 :  0.7297297297297297
매도 정확도 :  0.717391304347826
유지 정확도 :  0.45454545454545453
```

F1 score : 0.56

Buy accuracy: 0.73

Sell accuracy : 0.72

[Hyundai motor]

```
========== 현대차 ==========
[[  2   1   0]
 [ 24 117  11]
 [  0   3   3]]

전체 정확도 :  0.7577639751552795
전체 f1-score :  0.8229783067649851

매수 정확도 :  0.6666666666666666
매도 정확도 :  0.5
유지 정확도 :  0.7697368421052632
```

F1 score : 0.82

Buy accuracy : 0.66

Sell accuracy : 0.5

[Shinhan Financial]

```
========== 신한지주 ==========
[[ 10   8   1]
 [ 40 163  37]
 [  0  13  13]]

전체 정확도 :  0.6526315789473685
전체 f1-score :  0.6975956808520171

매수 정확도 :  0.5263157894736842
매도 정확도 :  0.5
유지 정확도 :  0.6791666666666667
```

F1 score : 0.69

Buy accuracy : 0.52

Sell accuracy : 0.5

6 Final model

LSTM Regressor

1. Variable selection

Data : SK Hynix

variables : selected based on VIF index

Variables
X : 'Economic Sentiment Index (Cyclically Adjusted)', 'Market Cap', 'Bitcoin Closing Price', 'KRW/USD ', 'Consumer Confidence Index', 'KOSPI Transaction Amount', 'Net Purchases by Individuals', 'Net Purchases by Foreigners', 'Industrial Production Index', 'KOSPI Volume', 'News Sentiment Index', 'KRW/EUR ', 'Discussion Forum', 'Unemployment Rate (%)', 'KRW/JPY', 'Volume', 'Article Sentiment Score', 'Foreign Ownership Quantity', 'KOSPI Fluctuation Rate', 'Labor Force Participation Rate (%)', 'Search Volume', 'News Coverage', 'Net Purchases by Institutions', 'Bitcoin Volume', 'Bitcoin Fluctuation', '5-Day Fluctuation Rate' Y : 'day5_label'

3 Final model

LSTM Regressor

2. Label encoding

Perform label encoding with target label (day5_label)

buy	0
maintain	1
Sell	2

3. MinMax Scaling

$$\frac{x - Min(X)}{Max(X) - Min(X)}$$

- Apply MinMax scaling to every continuous X variables
- Normalization scaling (range: [0, 1])
- More suitable for a regression model than a classification model

6 Final model

LSTM Regressor

3. Create Window dataset

EXAMPLE) window size = 3

sliding window

Date	Bitcoin Closing price	umeployment	Trading volume	Search term volume	Press volume	…	day5_label
2017-07-11	2324.3	3.4	3187332	8.10396	58	…	1
2017-07-12	2403.1	3.4	3462150	8.16834	65	…	1
2017-07-13	2362.4	3.4	5432312	11.22361	90	…	1
2017-07-14	2234.2	3.4	2931832	9.64898	72	…	0
2017-07-17	2233.4	3.4	2804598	9.12856	50	…	0
2017-07-18	2320.2	3.4	2066194	7.92513	76	…	1
2017-07-19	2282.6	3.4	2009799	7.69511	42	…	1
2017-07-20	2866.0	3.4	1647153	7.71154	31	…	1

X_train[0]

y_train[0]

6 Final model

LSTM Regressor

3. Create Window dataset

Create window dataset with Window size = 10

Apply window sliding

original data size : (1409, 27)

data size afterwards : (1399, 10, 27)

window size

4. train, validation, test split

data size : (1399, 10, 27)

Train set : (1120, 10, 27)

Validation set : (140, 10, 27)

Test set : (139, 10, 27)

5. LSTM Regressor

- with train set

hidden_size	2
num_layers	1
learning_rate	0.0001
loss function	MSE loss
optimizer	Adam
epoch	8000

Labeled Y is categorical variable consisting of 0, 1, and 2

But conduct prediction through regression

→ The optimal LSTM model is saved as a checkpoint

6 Final model

LSTM Regressor

5. LSTM Regressor

- with validation set

	매도 정확도	매수 정확도	유지 정확도	매도 정밀도	매수 정밀도	f1 score	평균
조합 1	O	O	X	O	O	X	
조합 2	O	O	O	X	X	X	★BEST★
조합 3	O	O	X	X	X	O	
조합 4	O	O	O	X	X	O	

Role of Threshold Function

1. The threshold is determined based on the combination that maximizes the average of buy accuracy, sell accuracy, and hold accuracy from validation set
2. Automate buy/sell/hold predictions based on the determined threshold

6 Final model

LSTM Regressor

6. Prediction

- with test set

[Shinhan Financial]

```
========== 신한지주 ==========
[[ 23  15   0]
 [ 24 529  62]
 [  1  15  39]]

전체 정확도 :  0.8347457627118644
전체 f1-score :  0.8503657789754228

매수 정확도 :  0.6052631578947368
매도 정확도 :  0.7090909090909091
유지 정확도 :  0.8601626016260162
```

F1 score : 0.85

Buy accuracy : 0.61

Sell accuracy : 0.71

[SK Hynix]

```
========== SK하이닉스 ==========
[[104  18   0]
 [113 313  45]
 [  3  27  76]]

전체 정확도 :  0.7052932761087267
전체 f1-score :  0.7165113879684445

매수 정확도 :  0.8524590163934426
매도 정확도 :  0.7169811320754716
유지 정확도 :  0.6645435244161358
```

F1 score : 0.72

Buy accuracy : 0.85

Sell accuracy : 0.67

[Hyundai motor]

```
========== 현대차 ==========
[[  9   9   3]
 [ 47 248  57]
 [  4   7  14]]

전체 정확도 :  0.6809045226130653
전체 f1-score :  0.7416229778038823

매수 정확도 :  0.4285714285714285
매도 정확도 :  0.56
유지 정확도 :  0.7045454545454546
```

F1 score : 0.74

Buy accuracy : 0.43

Sell accuracy : 0.56

Final Model

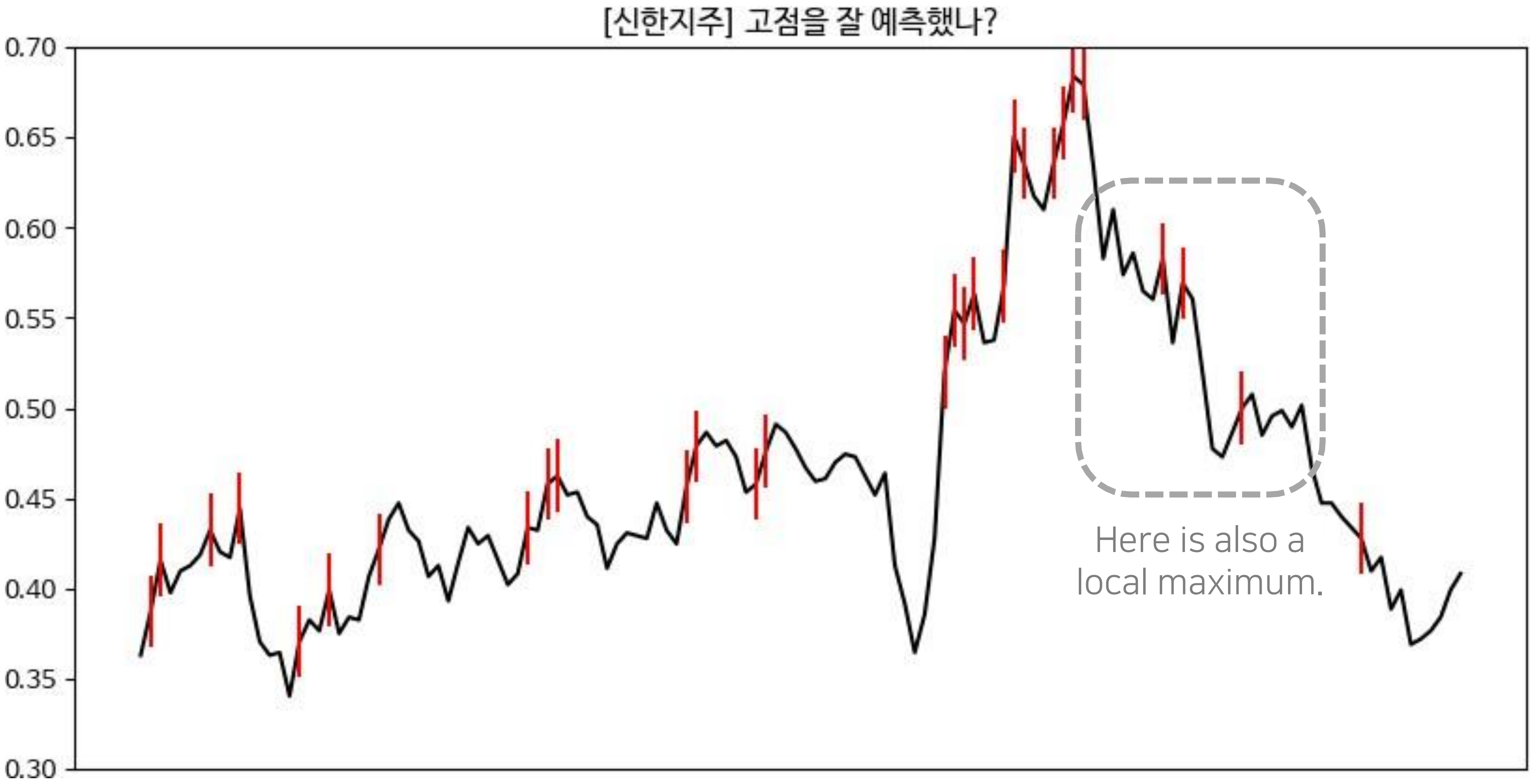
Visualization of prediction results

Expectation of a price increase of more than 5% within the next 5 days based on the red point.

▶ Buy

Final Model

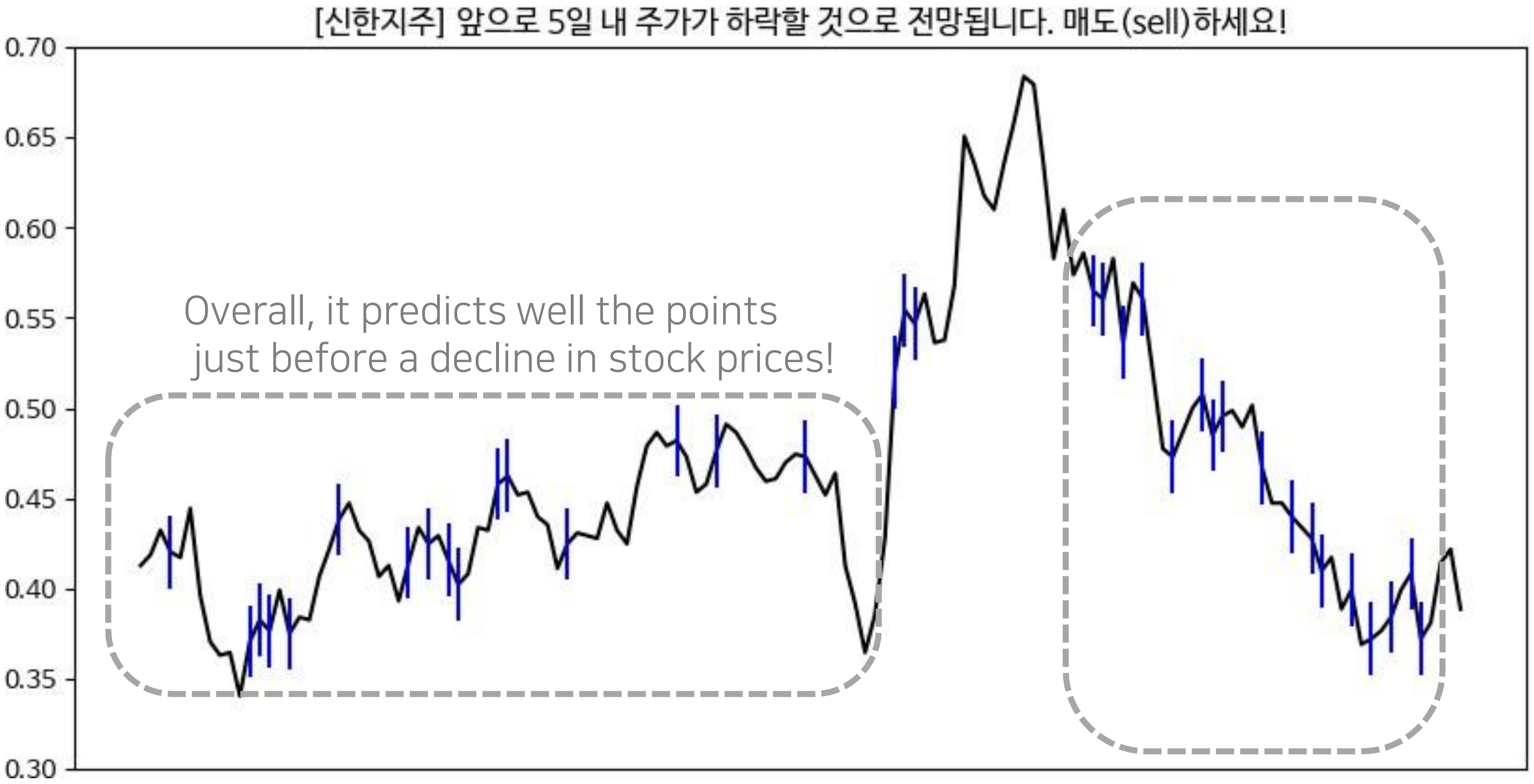
Visualization of prediction results



As a result of moving the red point to 5 days later, it generally matches well with points where the stock price records local peaks!

Final Model

Visualization of prediction results

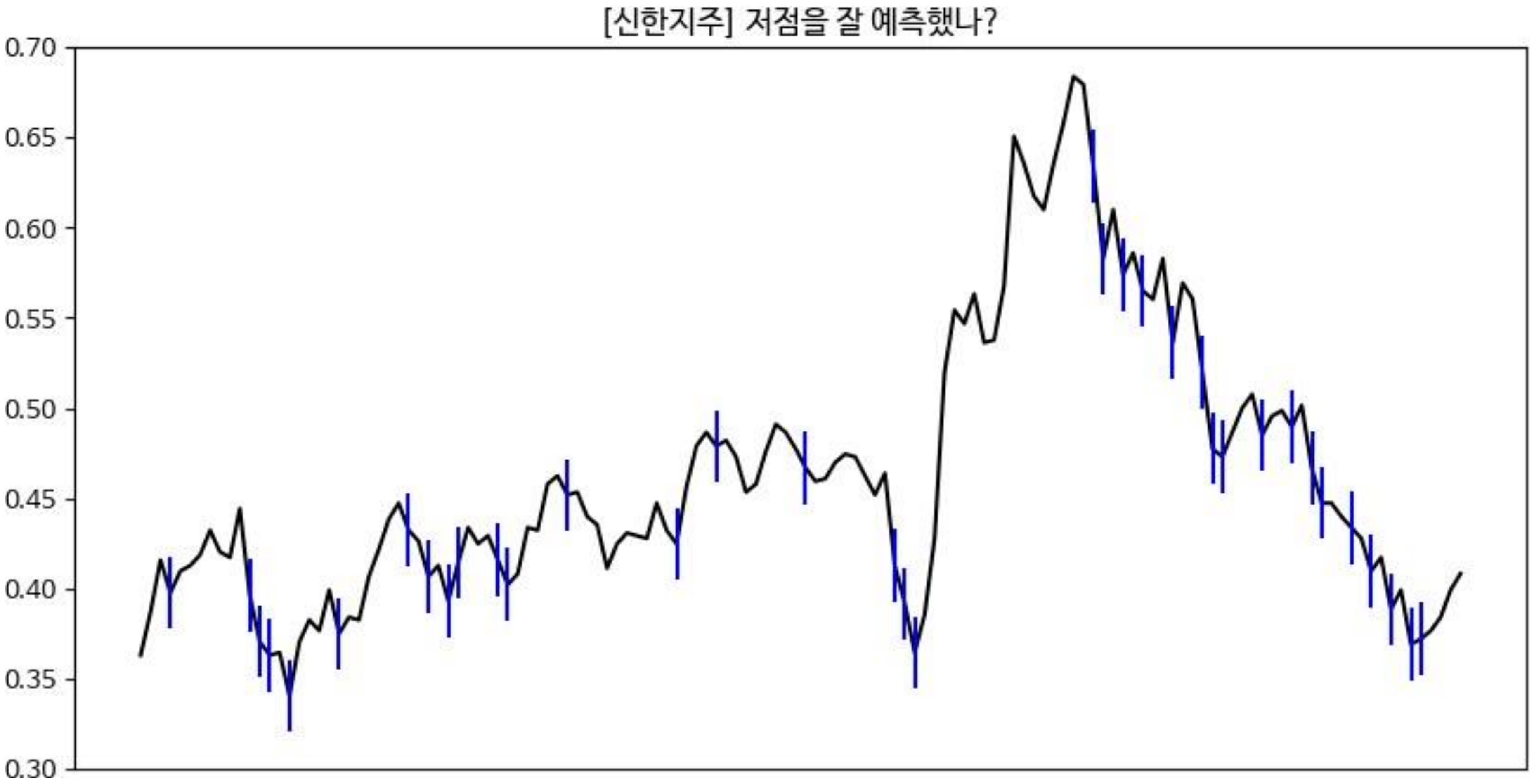


Based on the blue point, the stock price is expected to decrease by more than 5% within the next 5 days.

▶ Sell

Final Model

Visualization of prediction results



As a result of moving the blue point to 5 days later, it generally matches well with points where the stock price records local lows!

4 Conclusion

Significance of the project & Limitations

Significance of the project

- Using structured and unstructured data as well as various datasets to predict fluctuations, deriving significant results
- Analyzing stock data considering its characteristics (imbalanced data, time series data)
- Developing a robust model demonstrating consistent accuracy unaffected by domain-specific influences

Limitations

- To apply it in real-life scenarios, automation of data collection is necessary
- Whether the model can be applied to a wider range of stocks beyond the three stocks used as the dataset has not been tested

Thank you!!!