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

INDEX

1 Summary of week1

2 Modeling process

3 Final model

4 Conclusion

1

Summary of week1

1 Summary of week 1

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 Summary of week 1

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

Team Time Series Topic Analysis

Buy/Sell Recommendation Service For Stock Market Beginners

1 Summary of week 1

Utilized data

신한금융지주회사

SK Hynix

Shinhan Financial Group

Hyundai motor Company

8 individual indicators

11 common indicators

1 Summary of week 1

Utilized 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

1 Summary of week 1

Utilized 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

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

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 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

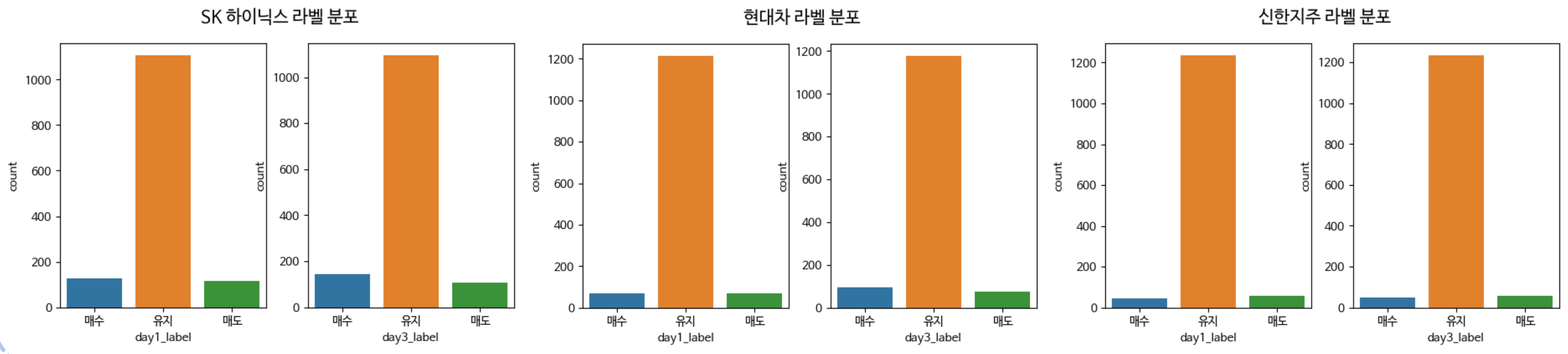
1-day(tomorrow) FR >= 3% : BUY

Otherwise : HOLD (maintain)

1 Summary of week 1

Labeling for Y variable

Each stock's distribution of labeled Y

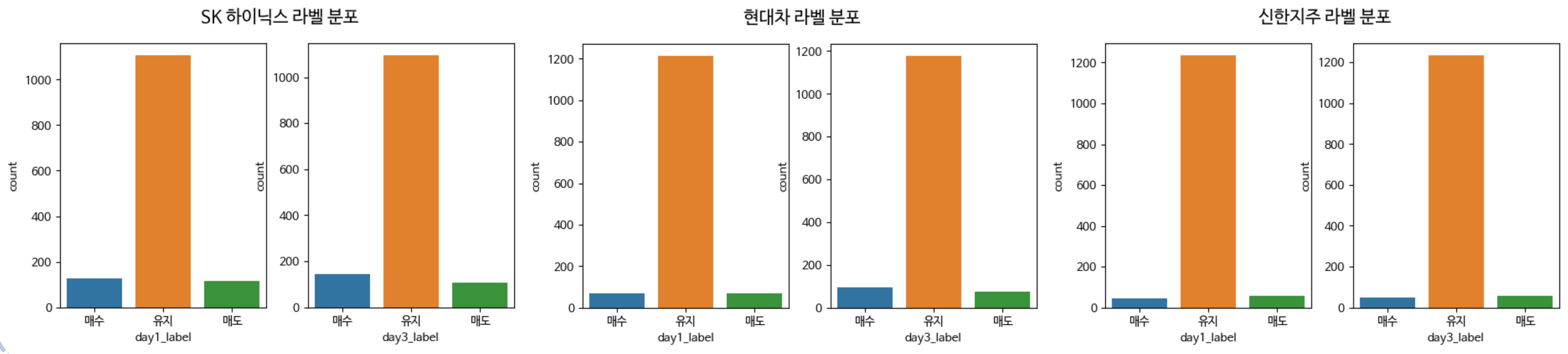


In all stocks, the class imbalance between 'buy'/'sell' versus 'hold' is significant

1 Summary of week 1

Labeling for Y variable

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'

1 Summary of week 1

Labeling for Y variable

Difficulties in augmenting timeseries data

- Traditional data augmentation techniques do not effectively leverage the time-dependent nature of time-series data
- Failure to consider temporal dependence, irregularity, and complexity of patterns can lead to overfitting

2

Modeling process

2 Modeling process

5-day fluctuation labeling

SK Hynix

Shinhan Financial Group

Hyundai motor Company

Team [deep learning] leader :

After applying the model, there are performance issues in predicting the 1-day fluctuations

2 Modeling process

5-day fluctuation labeling

1-day FR predictions

LGBM

	BUY	HOLD	SELL
BUY	0	22	0
HOLD	0	201	0
SELL	0	15	0

XGB

	BUY	HOLD	SELL
BUY	0	22	0
HOLD	9	186	6
SELL	1	14	0

Logistic

	BUY	HOLD	SELL
BUY	1	2	19
HOLD	4	38	159
SELL	0	2	13

FR : Fluctuation Rate

Model performance is bad for 1-day FR prediction

(The deep learning models were also not that great)

2 Modeling process

5-day fluctuation labeling

3-day FR predictions

LGBM

	BUY	HOLD	SELL
BUY	3	24	0
HOLD	3	186	3
SELL	0	15	4

XGB

	BUY	HOLD	SELL
BUY	11	16	0
HOLD	5	186	1
SELL	0	15	4

Logistic

	BUY	HOLD	SELL
BUY	8	24	0
HOLD	3	222	1
SELL	0	26	1

FR : Fluctuation Rate

A certain level of performance in
predicting the 3-day FR was observed

2 Modeling process

5-day fluctuation labeling

Why the 1-day FR predictions are worse

than 3-day FR predictions…?

☝ Estimated cause

The dependency among data become larger when predicting 3-day FR

Compare to that, 1-day FR less depend on X variables and have attribute that is closer to random …

✌ Estimated cause

Class imbalance is more seriously on 1-day FR than 3-day FR and it might lead poorer performance for the 1-day FR compared to the 3-day FR

2 Modeling process

5-day fluctuation labeling

3-day FR predictions

LGBM

	BUY	HOLD	SELL
BUY	3	24	0
HOLD	3	186	3
SELL	0	15	4

XGB

	BUY	HOLD	SELL
BUY	11	16	0
HOLD	5	186	1
SELL	0	15	4

Logistic

	BUY	HOLD	SELL
BUY	8	24	0
HOLD	3	222	1
[illegible]	0	26	1

FR : Fluctuation

A certain level of performance in predicting the 3-day FR was observed

If prediction performance is better for the 3-day FR than the 1-day FR, would it not be meaningful to predict longer that 3-day?

Decided to try predicting the

5-day fluctuation rate

2 Modeling process

5-day fluctuation labeling

```
SK[‘day5_label'] = SK[‘5일 등락률'].apply(lambda x: 'maintain' if abs(x) < 5 else 'buy' if x >= 5 else 'sell')
```

Same as the second week, calculate the polyserial correlation and determine the labeling threshold by comparing it with the original one

Like the 3-day fluctuation rate,
the threshold is determined to be 5%

2 Modeling process

Variable Selection

1

[Causality Test]

2

[VIF]

3

[Feature Importance]

4

[KS test]

5

[Full Model]

2 Modeling process

Variable Selection

Way1. Causality test & Correlation analysis

Causality Test

For two datasets with the same time range, if linear regression can be performed on one dataset against the other and it is significant, it suggests the existence of a Granger Causality

2 Modeling process

Variable Selection

Way1. Causality test & Correlation analysis

Causality Test

For two datasets with the same time range, if linear regression can be performed on one dataset against the other and it is significant, it suggests the existence of a Granger Causality

Perform variable selection

based on correlation analysis from the previous week, and Causality test

2 Modeling process

Variable Selection

Way1. Causality test & Correlation analysis

Causality Test

For two datasets with the same time range, if linear regression can be performed on one dataset against the other and it is significant, it suggests the existence of a Granger Causality

Variable selection result
Fluctuation rate, Closing price, Trading amount, Trading volume, Market capitalization, net institutional buying, net individual buying, net foreign buying, Sentiment score

2 Modeling process

Variable Selection

Way2. VIF

Variance Inflation Factor

In multiple polynomial regression analysis, it is commonly considered that independent variables exhibit multicollinearity when their VIF exceeds 10

$$VIF_i > 10 \iff \frac{1}{1-r_i} > 10$$
$$1 > 10 - 10r_i$$
$$r_i > 0.9$$

If the i-th independent variable is removed, the remaining variables still explain over 90% of the response variable.

2 Modeling process

Variable Selection

Way2. VIF

Variance Inflation Factor

In multiple polynomial regression analysis, it is commonly considered that independent variables exhibit multicollinearity when their VIF exceeds 10

Remove the variables with VIF values exceeding 10 from the final datasets

2 Modeling process

Variable Selection

Way2. VIF

Variance Inflation Factor

In multiple polynomial regression analysis, it is commonly considered that independent variables exhibit multicollinearity when their VIF exceeds 10

Variable selection result
Economic Sentiment Index (Cyclical Fluctuations), Market Capitalization, Bitcoin Closing Price, USD/KRW, Consumer Sentiment Index, KOSPI Trading Volume, Net individual buying, Net foreigner buying, Industrial Production Index, KOSPI Transaction Volume, News Sentiment Index, EUR/KRW, Discussion Forums, Unemployment Rate, JPY/KRW, Transaction Volume, Sentiment Score, Foreign Held Quantity, KOSPI Fluctuation Rate, Labor Force Participation Rate, Search Terms, Media Coverage Volume, Net institutional buying, Bitcoin Transaction Volume, Bitcoin Fluctuations

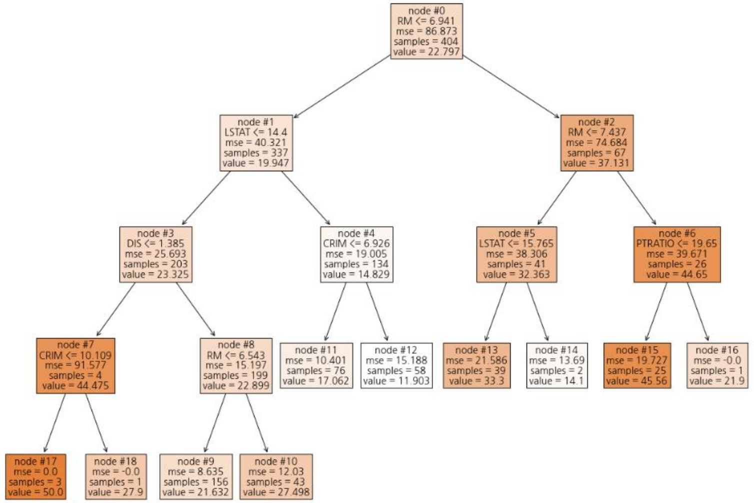
2 Modeling process

Variable Selection

Way3. Feature Importance

feature importance

When classification is conducted using a tree-based model, this value represents the ranking of how frequently and importantly the variable is utilized at each split.



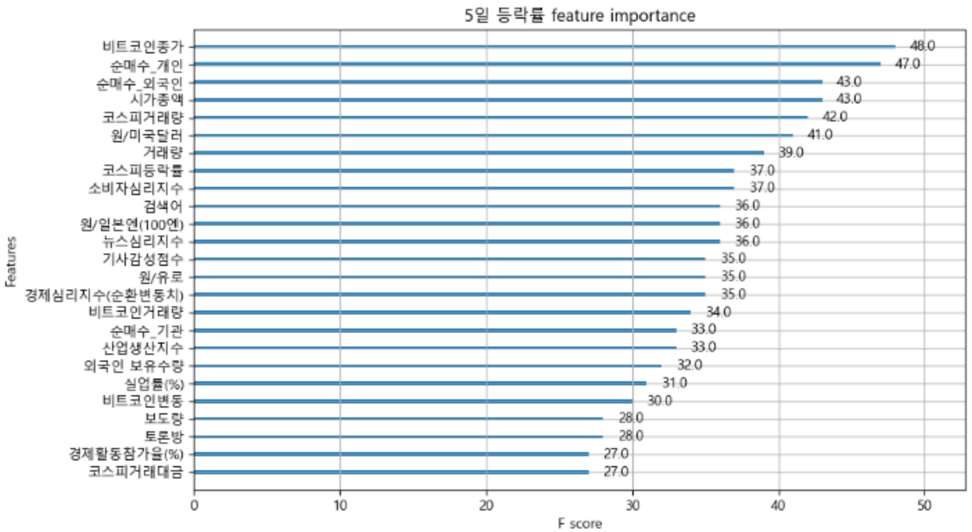
2 Modeling process

Variable Selection

Way3. Feature Importance

feature importance

When classification is conducted using a tree-based model, this value represents the ranking of how frequently and importantly the variable is utilized at each split.



Since there are no significant differences in feature importance, simply removed variables with multicollinearity and selected all remaining variables.

If you perform variable selection based on FI, you must handle the multicollinearity issue first

2 Modeling process

Variable Selection

Way3. Feature Importance

feature importance

When classification is conducted using a tree-based model, this value represents the ranking of how frequently and importantly the variable is utilized at each split.

Variable selection result
Net individual buying, Net foreigners buying, Net other Corporations buying, Trading Volume, Bitcoin Fluctuations, Discussion Forums, News Sentiment Index, Net institutional buying, Media Coverage Volume, Sentiment Score, USD/CNY, Search Terms, KOSPI Transaction Volume, Fluctuation Rate, Foreign held Quantity, Transaction Amount, Change, Bitcoin Closing Price, EUR/KRW, Closing Price, Bitcoin Transaction Volume, KOSPI Fluctuation Rate, USD/KRW, JPY/KRW, KOSPI Change, KOSPI Transaction Amount

2 Modeling process

Variable Selection

Way4. KS test

Kolmogorov Smironov Test

The non-parametric methods determine the rejection region

by comparing the differences in empirical distribution between two distributions

Frequently used techniques in credit scoring models for classifying good/bad customers

The objective of credit scoring models is similar to that of the topic analysis modeling,

and could it be possible to overcome the limitations of variable selection based only

on linear relationships, given its non-parametric approach?!

2 Modeling process

Variable Selection

Way4. KS test

Kolmogorov Smironov Test

The non-parametric methods determine the rejection region by comparing the differences in empirical distribution between two distributions

Frequently used techniques in credit scoring models for classifying good/bad customers

The objective of credit scoring models is similar to that of the topic analysis modeling, and could it be possible to overcome the limitations of variable selection based only on linear relationships, given its non-parametric approach?!

What is Empirical distribution function?

$$F_n(x) = \frac{\Sigma_{i=1}^{n} 1(X_i \le x)}{n} \quad (\textit{Where n is the total number of observations})$$

The definition of Population X's CDF is $F_n(x) = P(X \le x)$ which has high similarity to EDF

$$P(\lim_{n\to\infty} |F(x) - F_n(x)| = 0) = 1$$

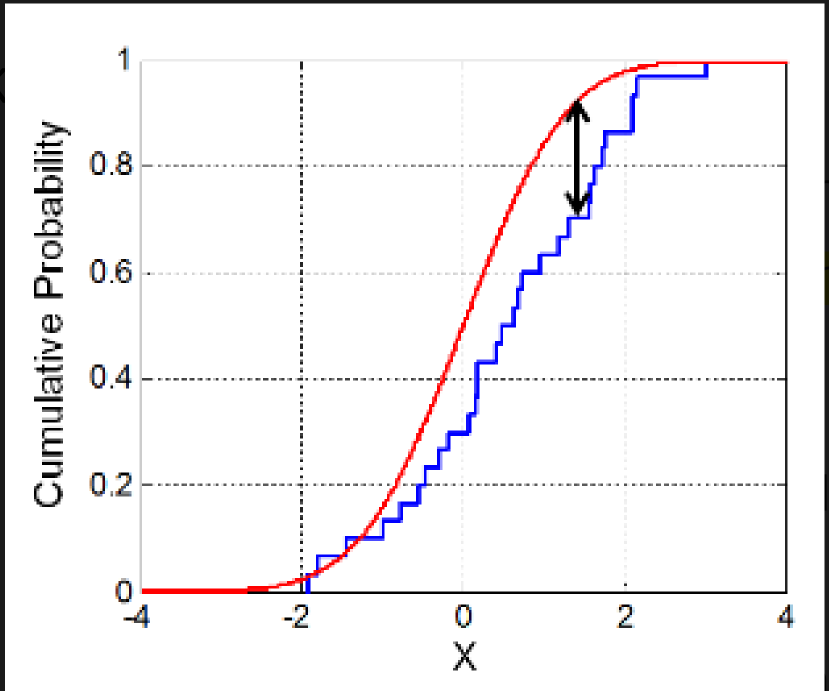
According to the Klimenko-Catelli theorem, EDF converges in probability to the CDF.

In other words, the EDF can be used to define statistics corresponding to the population's CDF

2 Modeling process

Variable Selection

Way4. KS test



The graph shows both an EDF and a CDF

black arrow between them represents KS Statistics

methods determine the rejection region

$$D_{n,m} = sup(|F_{1,n} - F_{2,m}(x)|$$

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}}$$

it scoring models for classifying good/bad customers

Calculate the test statistics by using the formula

if the test statistic is greater than critical value, reject the Null hypothesis (i.e., They are from same distribution)

odels is similar to that of the topic analysis modeling,

and could it be possible to overcome the limitations of variable selection based only on linear relationships, given its non-parametric approach?!

2 Modeling process

Variable Selection

Way4. KS test

Kolmogorov Smironov Test

The non-parametric methods determine the rejection region

by comparing the differences in empirical distribution between two distributions

The number of variables that significantly differed

in the distribution of sell/hold, hold/buy, and sell/buy across all three stocks was small

Among the nine combinations of distributions(S/H,H/B,S/B - 3 X 3 –Shinhan, Sk, Hyundai),

select variables that significantly differ in distribution six or more times

2 Modeling process

Variable Selection

Way4. KS test

Kolmogorov Smironov Test

The non-parametric methods determine the rejection region

by comparing the differences in empirical distribution between two distributions

Variable selection result
Closing price, Fluctuation rate, Trading volume, Trading value, Market capitalization, Discussion Forum, Net institutional buying, Net other Corporations buying, Net individual buying, Net foreigners buying, Search terms, media coverage volume, Sentiment score, KOSPI Fluctuation rate

2 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…

2 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…

2 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.

2 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.)

2 Modeling process

Customize optuna score

Confusion matrix

		PREDICTIVE VALUES	
		POSITIVE (1)	NEGATIVE (0)
ACTUAL VALUES	POSITIVE (1)	TP	FN
	NEGATIVE (0)	FP	TN

Accuracy

how often a classification ML model is correct overall

(TP+TN)/(TP+TN+FP+FN)

Precision

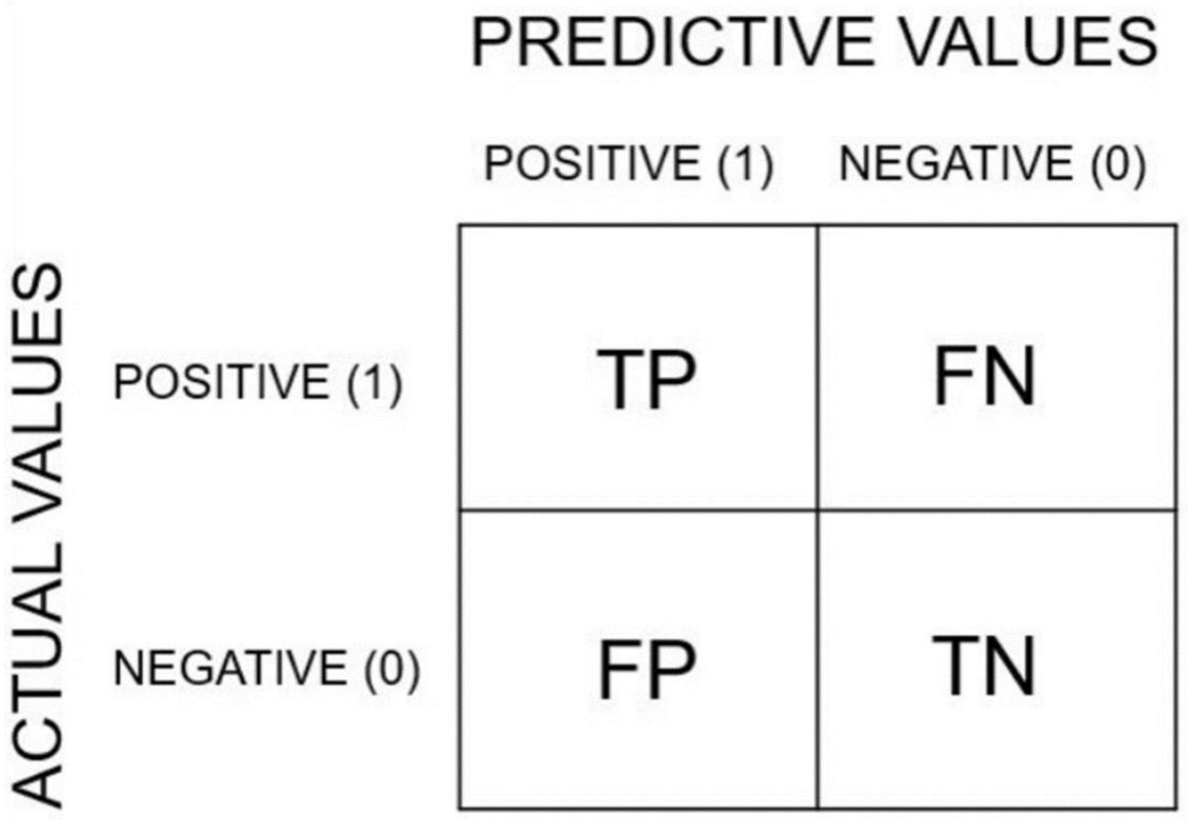
how often an ML model is correct when predicting the target class.

TP/(TP+FP)

2 Modeling process

Customize optuna score

Confusion matrix



Recall

shows whether an ML model

can find all objects of the target class

TP/(TP+FN)

F1 score

Harmonic mean of

Precision and Recall

2(Precision*Recall)/(Precision+Recall)

2 Modeling process

Customize optuna score

Optuna

Explore the hyperparameter spce to find out the
composition of parameters which maximize or minimize the objective function

2 Modeling process

Customize optuna score

Optuna

Explore the hyperparameter space to find out the
composition of parameters which maximize or minimize the objective function

Optuna score custom

Defining and Providing evaluation metrics for the objective function to be optimized

Function returning results
for each hyperparameter combination in Optuna

2 Modeling process

Customize optuna score

Optuna

Explore the hyperparameter space to find out the
composition of parameters which maximize or minimize the objective function

Optuna score custom

Defining and Providing evaluation metrics for the objective function to be optimized

Process Optuna with F1score
and 3 additional custom scores

2 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])
```

2 Modeling process

List of models attempted

Models

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

2 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

3

Final model

3 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

3 Final model

Final model introduction

XGB classifier

Used Data : SK Hynix

variables : Full Model

evaluation : custom optuna score

Type: multiclass classification

LSTM regressor

Used Data : SK Hynix

variables : VIF

highlight : predict labeled Y

by regression and classify

using Threshold function afterward

3 Final model

Final model introduction

XGB classifier

Used Data : SK Hynix

variables : Full Model

evaluation : custom optuna score

Type: multiclass classification

LSTM regressor

Used Data : SK Hynix

variables : VIF

highlight : predict labeled Y

by regression and classify

using Threshold function afterward

3 Final model

XGB Classifier

1. Variable selection

Data : SK Hynix

variables : Full Model

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'

3 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'
Categorical variable based on the predicted fluctuation rate of 5 days after the given date Have 3 categories : buy, sell, maintain
Y : 'day5_label'

3 Final model

XGB Classifier

2. Label encoding

Perform label encoding with target label (day5_label)

Buy	0
Hold	1
Sell	2

3. MinMax Scaling

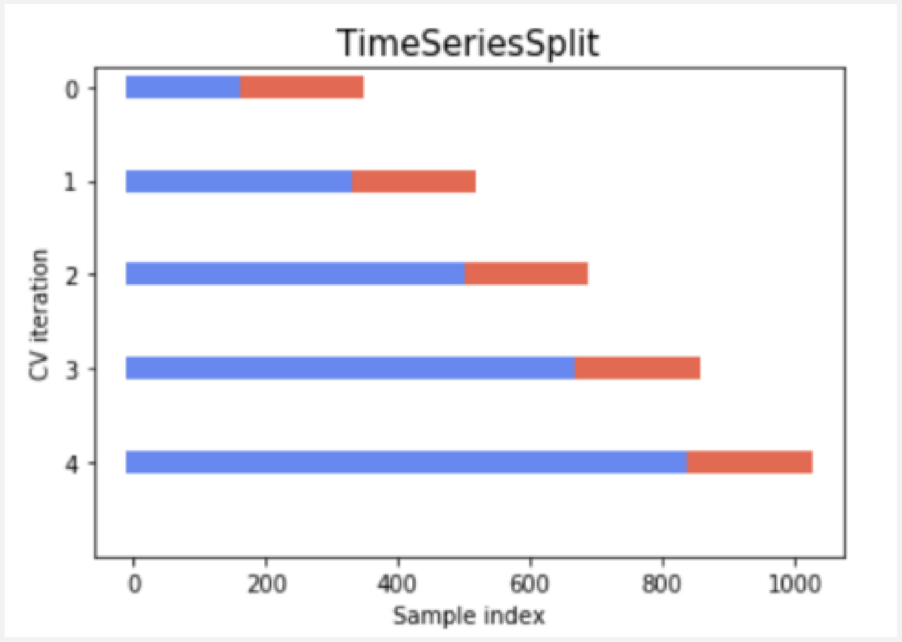
$$\frac{x - Min(X)}{Max(X) - Min(X)}$$

- 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

3 Final model

XGB Classifier

3. Expanding Window CV



time series cross-validation technique

where a window of the same size accumulates and moves incrementally.

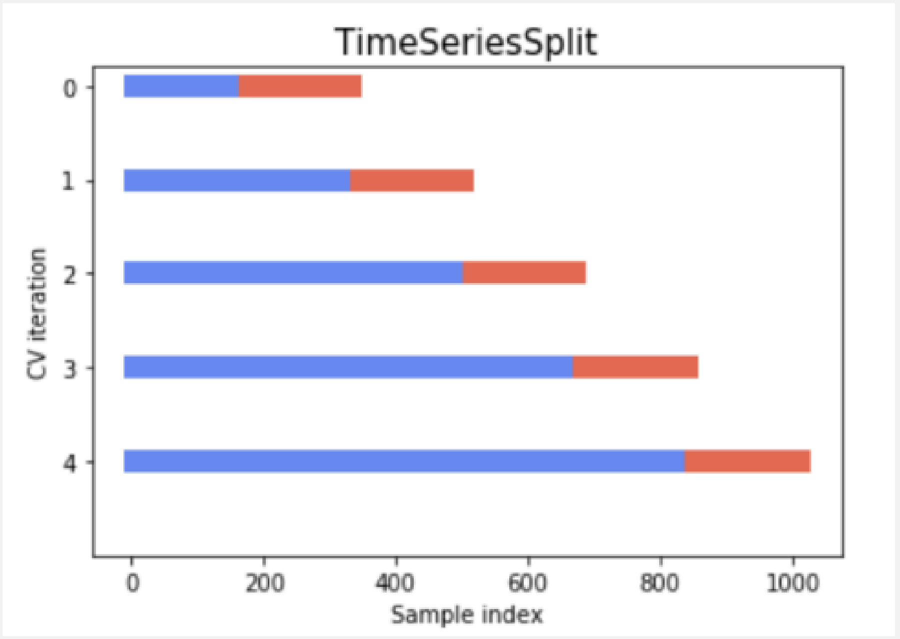
In each step, the training set and validation set from the

previous stage are utilized as the training set for the current stage

3 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.

3 Final model

XGB Classifier

4. Class weights

How to deal with the class imbalance issue?

If there is class imbalance in the data,

can you simply use the scale_pos_weight parameter?

3 Final model

XGB Classifier

4. Class weights

How to deal with the class imbalance issue?

~~If there is class imbalance in the data,~~
~~can you simply use the scale_pos_weight parameter?~~

it can only be used with binary classification…

3 Final model

XGB Classifier

4. Class weights

How to deal with the class imbalance issue?

~~If there is class imbalance in the data,~~

~~can you simply use the scale_pos_weight parameter?~~

In the case of Multiclass classification,

sample_weight parameter can be used with the fit function !!

3 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!

3 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

3 Final model

XGB Classifier

5. Optuna hyperparameter tuning

Accuracy

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

Precision

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

Optuna evaluation metircs

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

3 Final model

XGB Classifier

5. Optuna hyperparameter tuning

Accuracy

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

Precision

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

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

Optuna evaluation metircs

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

3 Final model

XGB Classifier

5. Optuna hyperparameter tuning

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

3 Final model

XGB Classifier

5. Optuna hyperparameter tuning

Best Parameters

- max_depth : 14
- learning_rate : 0.01843
- n_estimators : 604
- min_child_weight : 6
- gamma : 0.13475
- subsample : 0.14467
- colsample_bytree : 0.96275
- reg_alpha : 8.16062e-05
- reg_lambda : 2.07248e-08

3 Final model

XGB Classifier

5. Optuna hyperparameter tuning

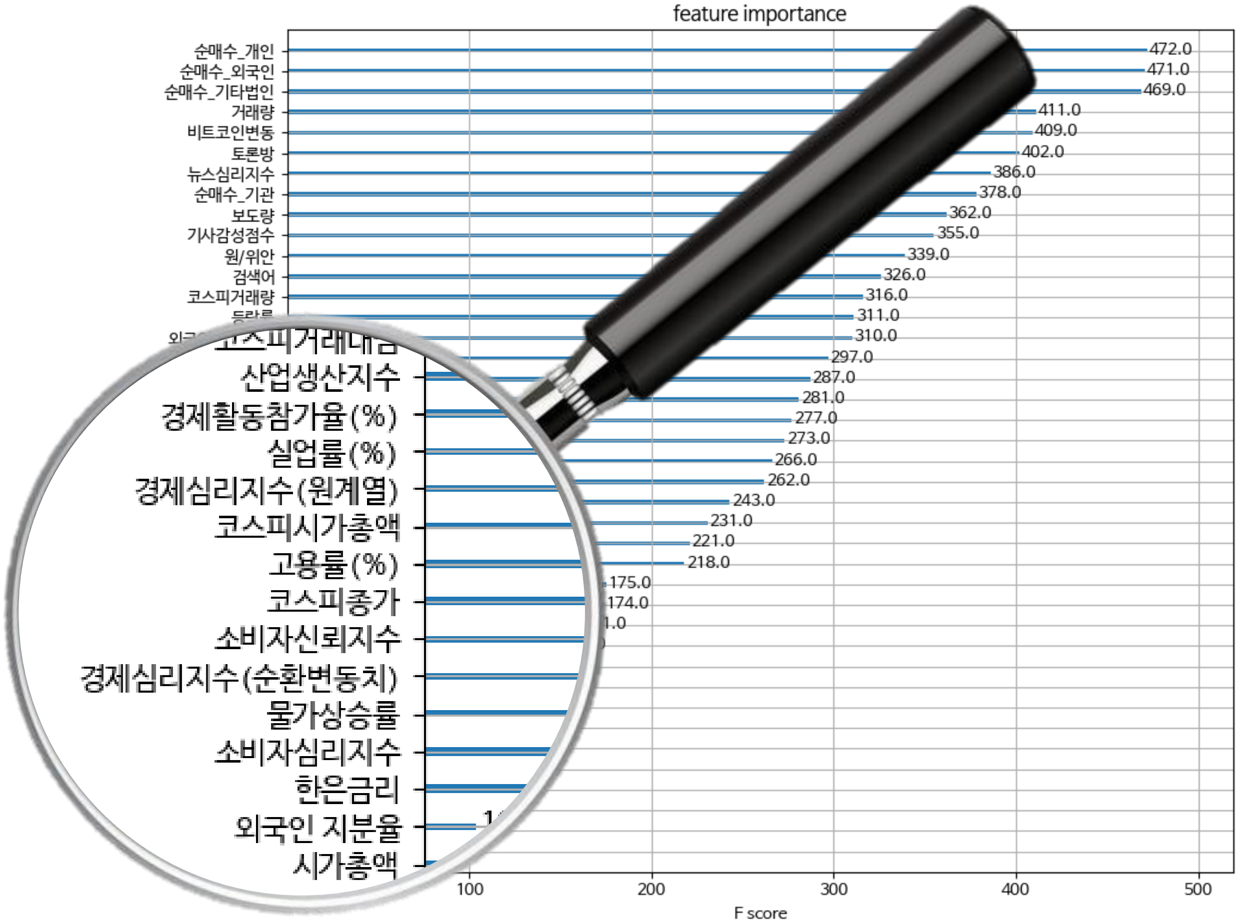
net purchase(institution/foreign/individual/other),Bitcoin volatility, Discussion forum post count, News sentiment index, Sentiment score, Article coverage volume, Search volume

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

3 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

3 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

3 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

3 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

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'

Categorical variable based on the predicted fluctuation rate of 5 days after the given date

Have 3 categories : buy, sell, maintain

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

3 Final model

LSTM Regressor

3. Create Window dataset

General data

the evaluation is conducted by randomly splitting the dataset into train and test datasets

Time series data

randomly splitting the dataset may not reflect the temporal characteristics

3 Final model

LSTM Regressor

3. Create Window dataset

General data

the evaluation is conducted by randomly splitting the dataset into train and test datasets

Time series data

randomly splitting the dataset may not reflect the temporal characteristics

window datasets are created using a sliding window approach

sliding window

utilizes the previously established window size to incorporate past time steps into the training process, enabling predictions for the subsequent time steps

3 Final model

LSTM Regressor

3. Create Window dataset

General data

the evaluation is conducted by randomly splitting the dataset into train ar

Time series data

randomly splitting the dataset may not reflect the temporal ...istics

window size

The number of previous time steps used for predicting a single point

window datasets are created using a sliding window approach

sliding window

utilizes the previously established window size to incorporate past time steps into the training process, enabling predictions for the subsequent time steps

3 Final model

LSTM Regressor

3. Create Window dataset

EXAMPLE) window size = 3

sliding window

Date	Bitcoin Closing price	umeplo yment	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

3. Create Window dataset

EXAMPLE) window size = 3

sliding window

Date	Bitcoin Closing price	umeplo yment	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]

3 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[1]

y_train[1]

3 Final model

LSTM Regressor

3. Create Window dataset

EXAMPLE) window size = 3

sliding window

Date	Bitcoin Closing price	umeplo yment	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[2]

y_train[2]

3 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[3]

y_train[3]

3. Create Window dataset

EXAMPLE) window size = 3

sliding window

Date	Bitcoin Closing price	umeplo yment	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[4]

y_train[4]

3 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)

Since one time step is predicted

using 10 preceding time steps,

the first 10 time steps are excluded from the data

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

3 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)

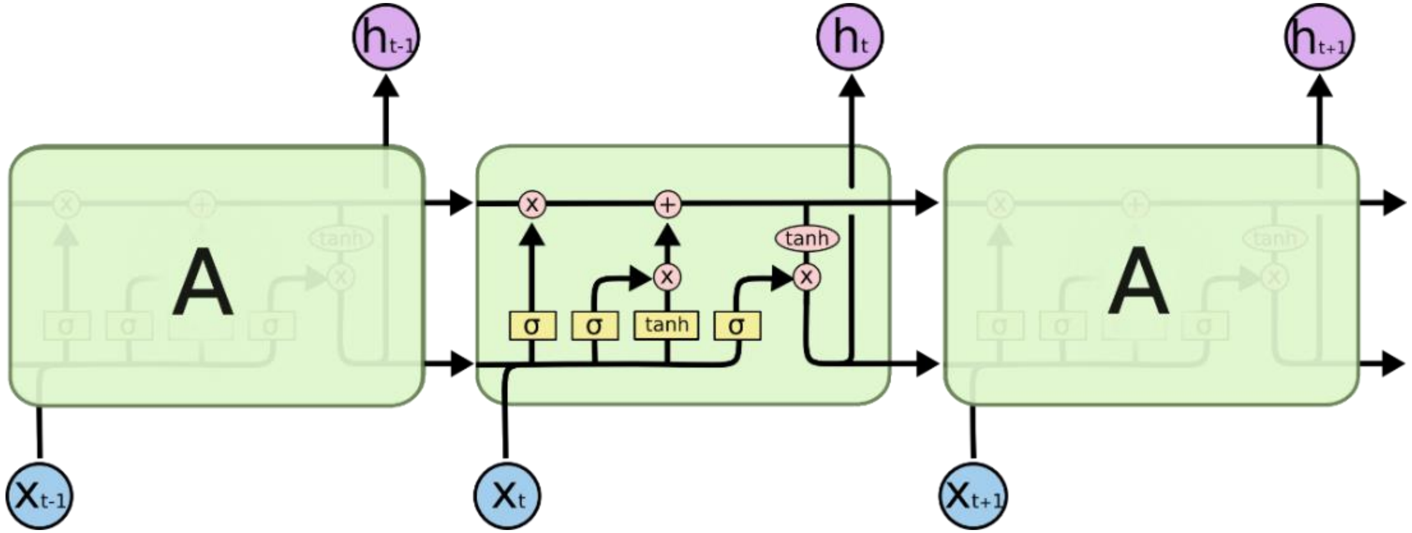
3 Final model

LSTM Regressor

Refer to Deep Learning Team Cleanup Week 3
for more detailed explanations about LSTM

5. LSTM Regressor

- with train set



A structure introduced to address the long-term
dependency issue in RNNs, composed of input gate, forget gate,
and output gate, resulting in a model with excellent long short-term memory

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

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	

Threshold is needed to convert each numerically predicted value into categorical value

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

But conduct prediction through regression

→ Prediction values are numeric rather than categoric

3 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

3 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 :  0.42857142857142855
매도 정확도 :  0.56
유지 정확도 :  0.7045454545454546
```

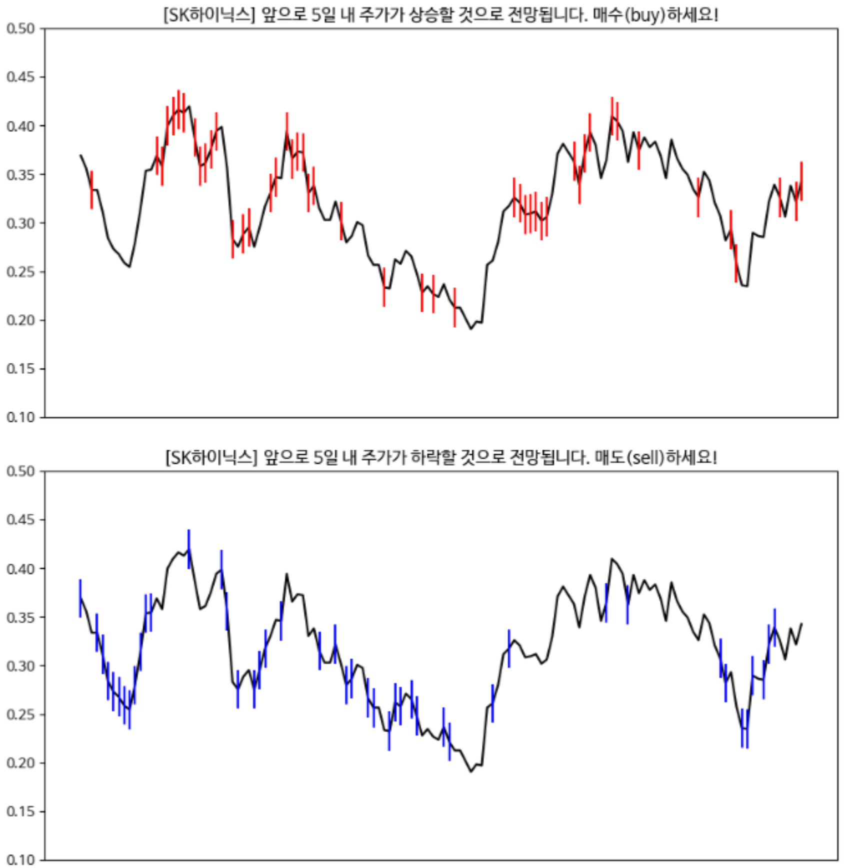
F1 score : 0.74

Buy accuracy : 0.43

Sell accuracy : 0.56

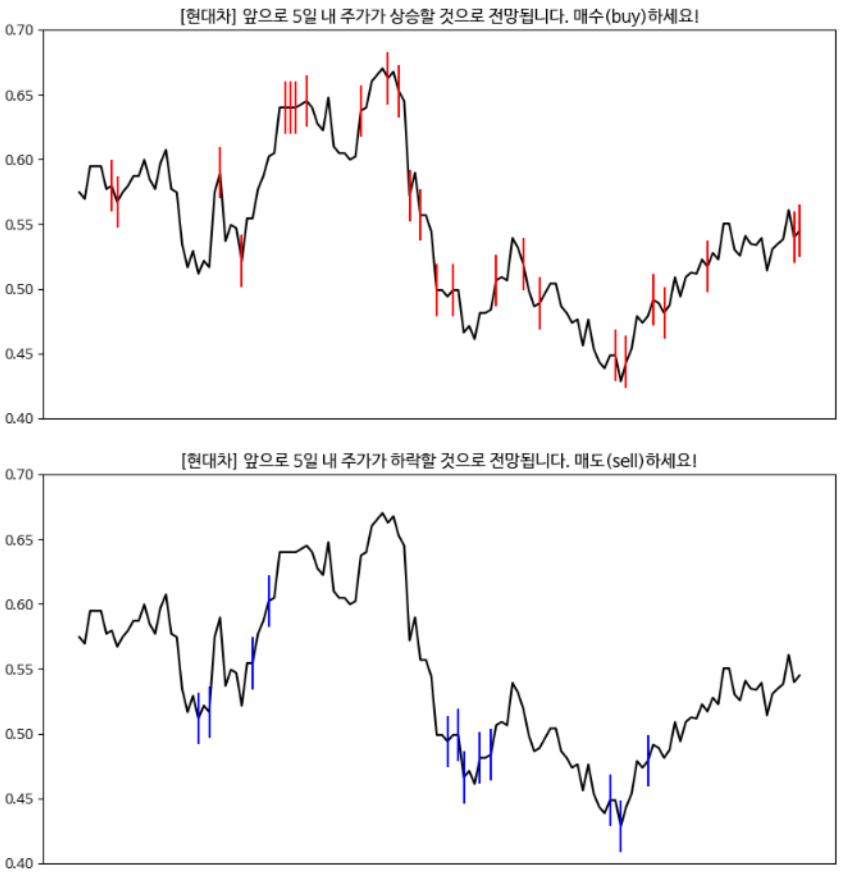
3 Final model

Visualization of prediction result



[SK Hynix]

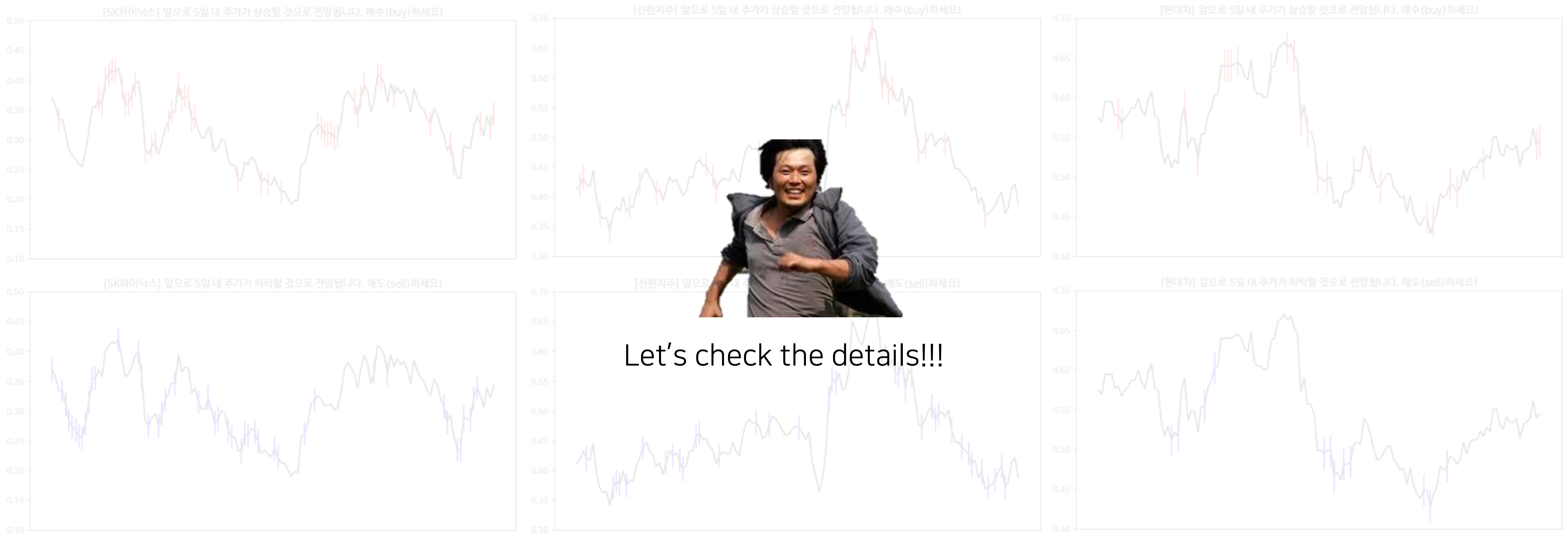
[Shinhan Financial]



[Hyundai motor]

3 Final model

Visualization of prediction result



Let’s check the details!!!

[SK하이닉스] [신한지주] [현대차]

3 Final model

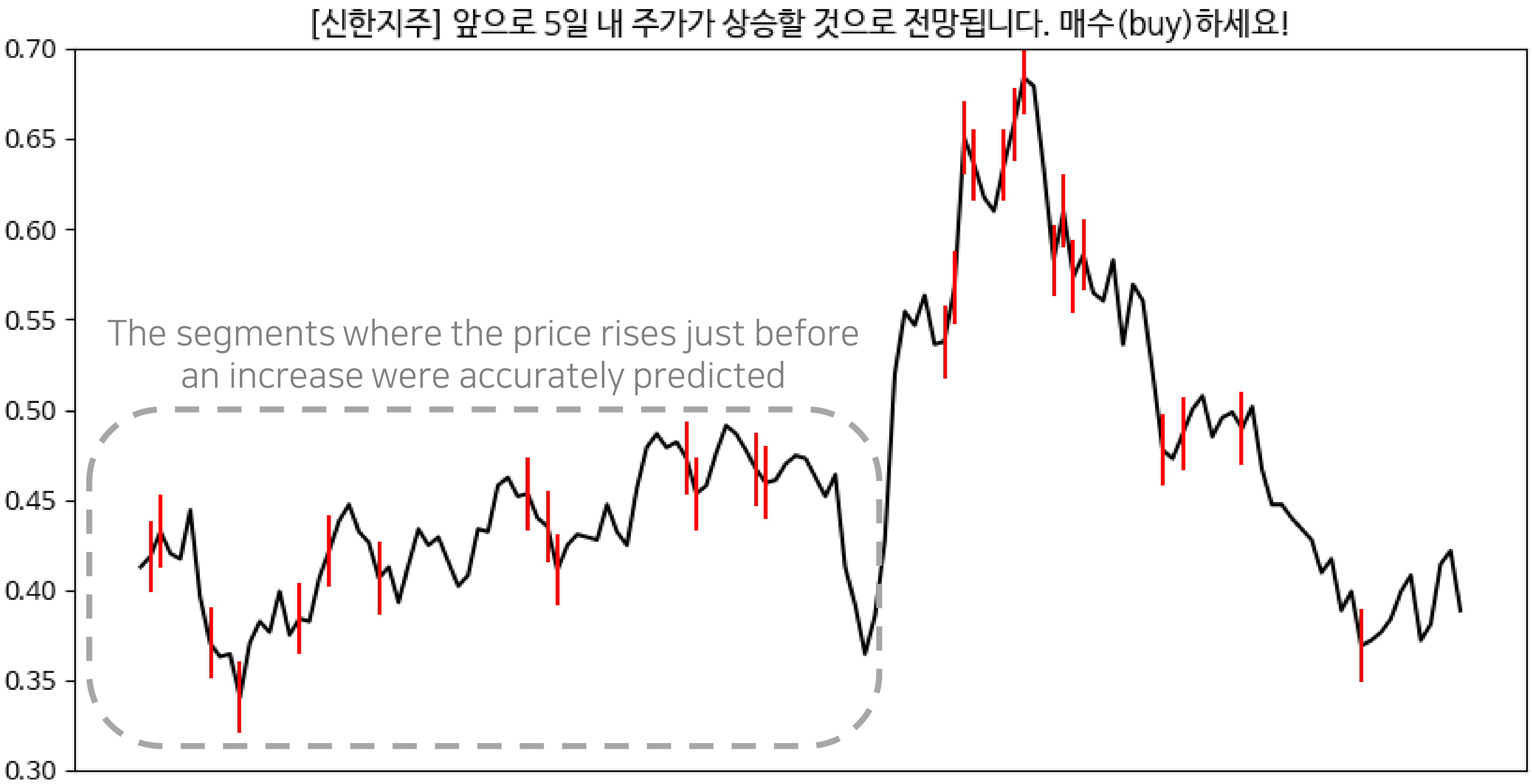
Visualization of prediction result

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

▶ Buy

3 Final model

Visualization of prediction result



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

▶ Buy

3 Final model

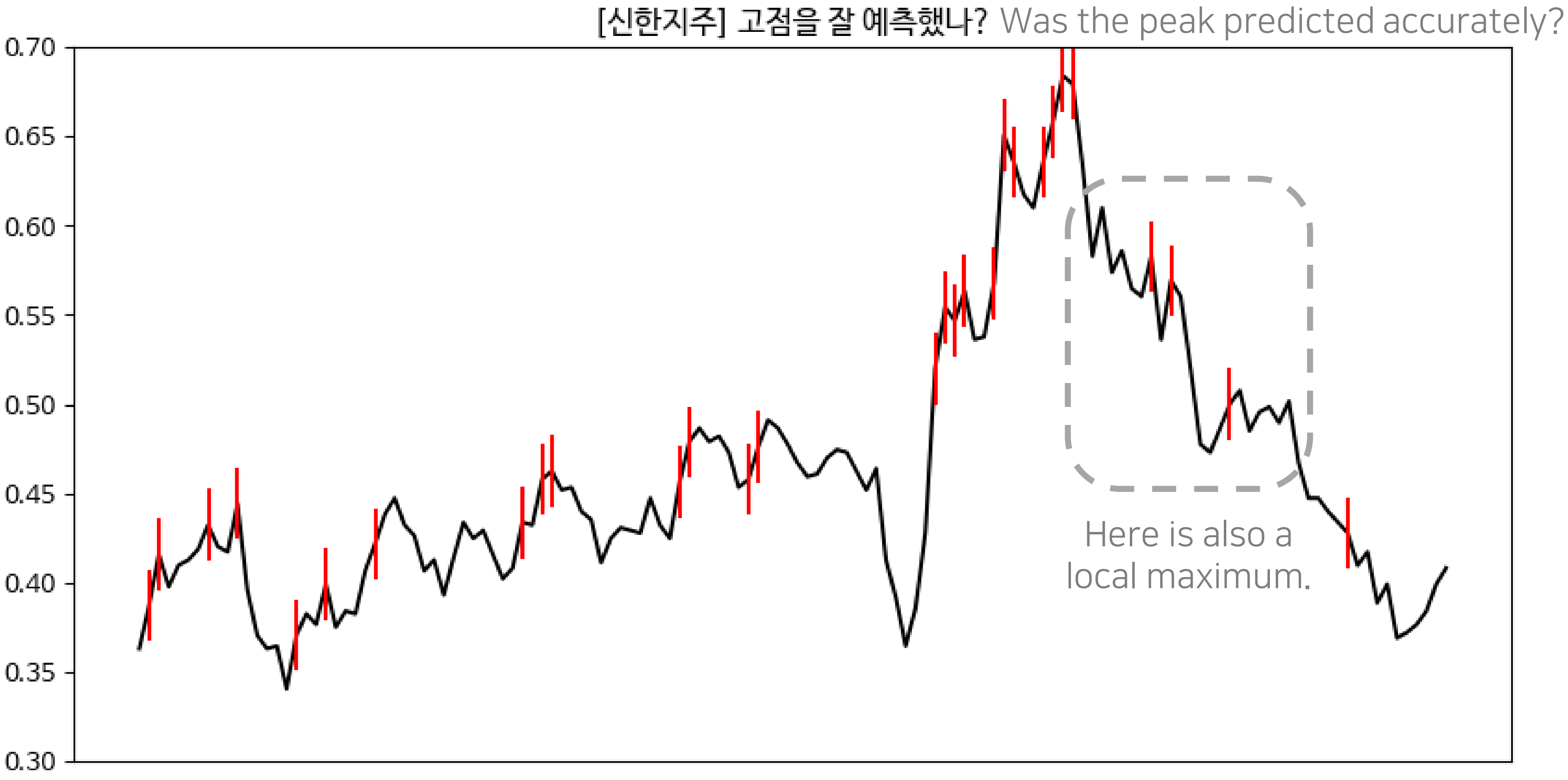
Visualization of prediction result

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

▶ Buy

3 Final model

Visualization of prediction result



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!

3 Final model

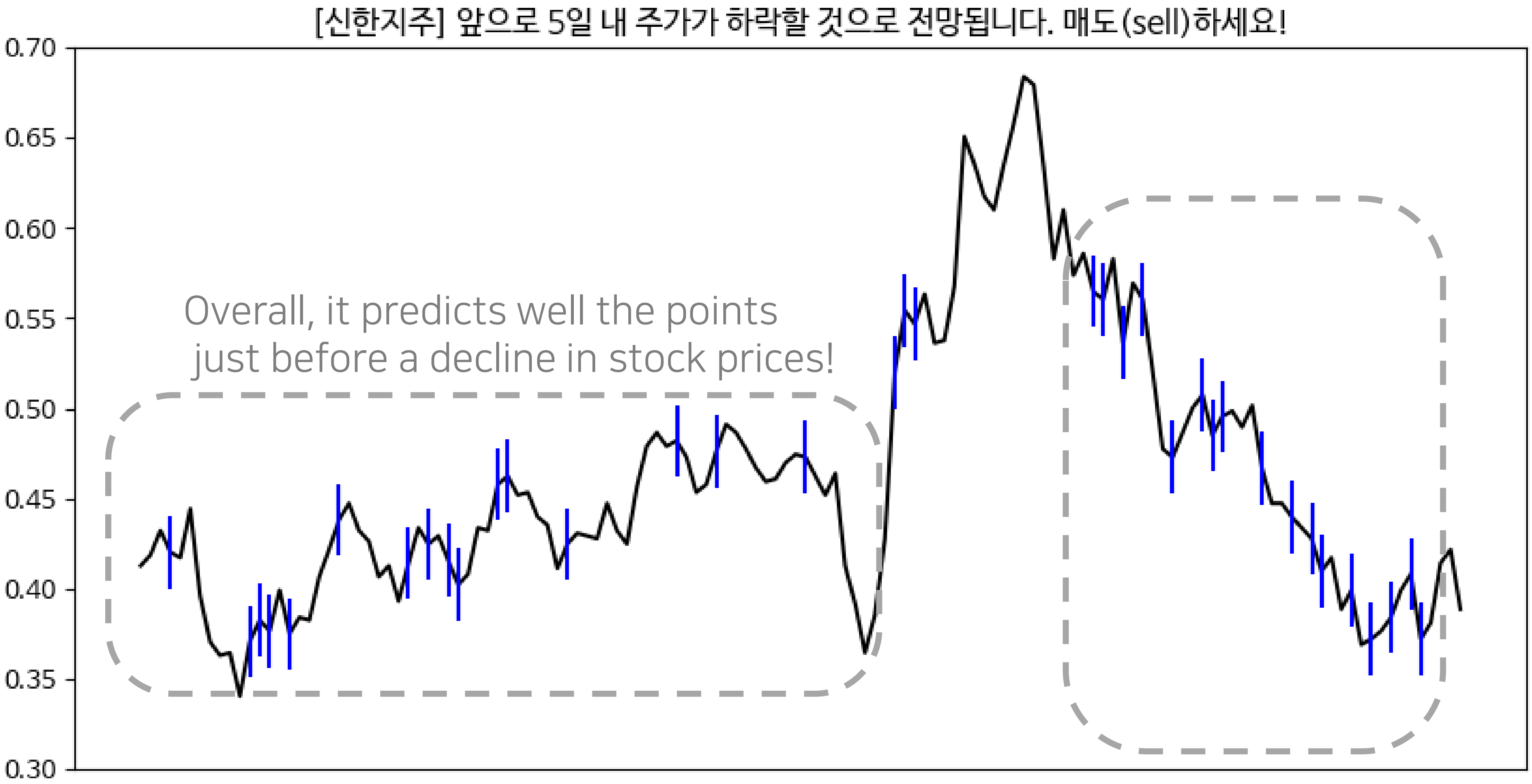
Visualization of prediction result

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

▶ Sell

3 Final model

Visualization of prediction result

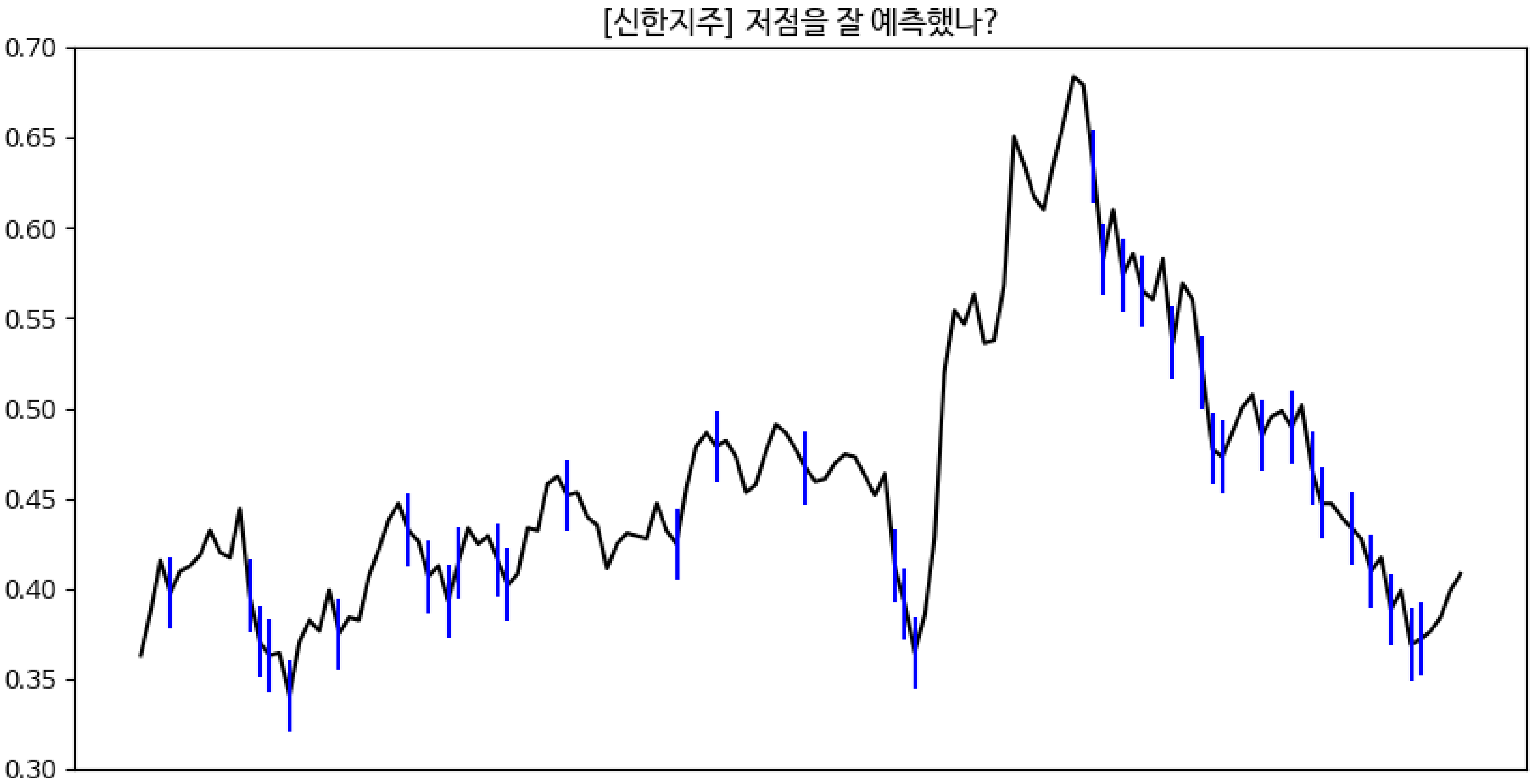


Based on the blue point, the stock price is expected to decrease by more than 5% within the next 5 days.

▶ Sell

3 Final model

Visualization of prediction result



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

4 Conclusion

Topic analysis concept

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!

▶ providing simple,easy and accessible investment insights for everyone

Offering simple and straightforward

investment indicators for novice stock investors

4 Conclusion

Expected impact & Expandability

Expected impact

- If the service is developed into an app, it could attract customers in their 20s who are just starting to invest in stocks
- Most young adults and novice investors tend to continue using the platform they initially signed up for, making it possible to secure loyal customers

Expandability

- By accepting user-defined thresholds for labeling fluctuation rates(3-day,5-day..), it's possible to offer personalized buy/sell recommendations tailored to individual preferences
- It's possible to offer customized buy/sell recommendation service based on investment preferences

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!!!