Forecasting Costs of Fund Index of Korea Utilizing Various Machine Learning Models

Jaeyoung Baak

Department of Economics Sungkyunkwan University

I. Introduction

Uncertainty in financial markets has been rising. During the COVID-19 pandemic, inflation and real estate prices surged, while household debt in Korea reached record-high levels. Adding to this uncertainty, the ongoing Russian-Ukraine war has disrupted global stability. To address these challenges, the Bank of Korea has repeatedly raised the base interest rate, causing a corresponding rise in related interest rates. As a result, the financial burden of household loans has significantly increased.

The Costs of Fund Index (COFIX) is a weighted average of funding costs from eight domestic banks in Korea. It reflects changes in interest rates across products such as deposits and bank debts and serves as a benchmark for floating interest rates in the banking sector. As of October 2022, the COFIX reached 3.98%, its highest level since its introduction in 2010, with a record monthly increase of 0.58 percentage points. Several factors contributed to this sharp rise. Successive base rate hikes by the Bank of Korea drove up interest rates on deposits and installment savings. Additionally, the *Lego Land Incident* in Korea led to a surge in bond yields, further elevating the COFIX. In response, banks swiftly adjusted their loan interest rates to reflect the higher COFIX. Consequently, the upper range of floating interest rates for mortgage loans and "*Jeonse*" loans at Korea's four major commercial banks exceeded 7%, significantly increasing household financial burdens.

The COFIX responds to changes in the base rate with a time lag and serves as a leading indicator for various floating-rate financial products. Its impact is particularly evident in household loans, including those issued by the Korean Housing Finance Corporation. As a predictor of household loan trends, the COFIX is instrumental in determining rates for unsecured loans (Kim, 2016). Therefore, forecasting COFIX movements is essential for understanding and managing fluctuations in floating interest rates.

Accurate COFIX predictions benefit both banks and households. For banks, forecasting the index enables proactive adjustments to internal decisions, rather than reacting to COFIX announcements with delayed interest rate changes. For households, reliable predictions support more informed decisions regarding consumption, savings, and loan sizes. In this context, understanding and predicting COFIX behavior is critical for fostering strategic financial decision-making across the economy.

II. Data and Forecasting Procedure

The target variable, COFIX, has been reported monthly since January 2010. Consequently, the sample period extends from January 2010 to August 2022, covering 152 monthly observations. For explanatory variables, the dataset includes a range of macroeconomic and financial indicators specific to Korea, such as loan rates and employment rates. Additionally, it incorporates the Korean version of the categorical Economic Policy Uncertainty (EPU) index (Cho and Kim, 2020) and the KOSPI volatility index.

Given Korea's status as a small open economy with strong economic ties to the United States, the dataset also includes selected variables from the FRED-MD database. This database is a comprehensive monthly macroeconomic dataset designed for empirical research in data-rich environments. Overall, the dataset contains 102 explanatory variables over the same sample period as the dependent variable, with four lagged terms considered for each variable. This setup results in a total of 408 potential predictors, including autoregressive terms.

To ensure stationarity, Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests are performed on the entire dataset, including the target variable. Appropriate transformations are applied to achieve stationarity, as detailed in Appendix Table A1. For the target variable, COFIX, both tests indicated an I(1) process, leading to the use of its first difference in the analysis. Table 1 presents descriptive statistics for COFIX and the explanatory variables.

Figure 1. The first difference of COFIX (Jan, 2010 – Aug, 2022)

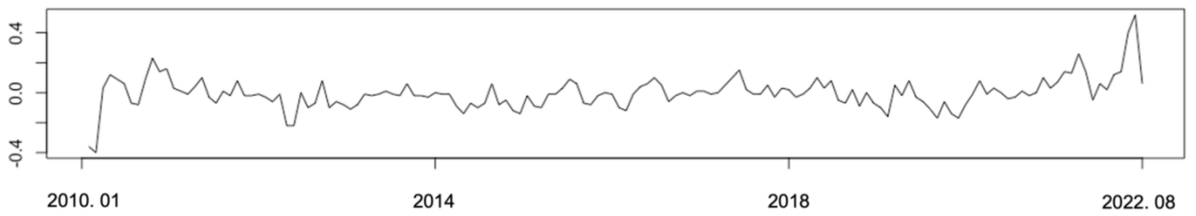

Table 1. Descriptive Statistics for the first difference of COFIX
January 2010 – August 2022

Mean	0.0000
Median	-0.0100
Minimum	-0.4000
Maximum	0.5200
Standard Deviation	0.1100
Skewness	0.6600
Kurtosis	5.8300
ADF	-6.9890 ***
KPSS	0.0980 ***

The forecast horizons are set at 1, 3, and 6 months. As previously mentioned, the forecast period spans from January 2010 to August 2022, covering a total of 152 observations. For each model, predictions of the target variable are made using a rolling window forecasting scheme. The predictive performance of each model is evaluated by calculating the root mean squared error (RMSE) and mean absolute error (MAE) for each horizon. The model demonstrating superior predictive accuracy for each horizon is selected. To further compare predictive performance, the Giacomini-White (GW) test is used to assess the accuracy of predictions across models, and the Model Confidence Set (MCS) test identifies models with better performance at each forecast horizon.

Interpreting results from machine learning models poses challenges, particularly in determining which variables contribute most to the prediction of COFIX. If the entire dataset is used, it becomes difficult to pinpoint the variables with the strongest explanatory power. To address this, the Boruta Algorithm is employed to identify key predictors. First, the algorithm ranks the covariates in the dataset based on their explanatory power. New explanatory variable sets are then created, consisting only of variables identified as significant for each forecast horizon. Finally, the prediction results using the full dataset are compared with those using the reduced variable sets for the Random Forest (RF) and Extreme Gradient Boosting (XGBoost) models. This approach provides clearer insights into which variables are most effective in explaining the target variable at different horizons.

III. Models

In this study, I employ various forecasting models to predict COFIX. The Random Walk (RW) model and the autoregressive (AR) model serve as benchmarks to evaluate the performance of more advanced methods. To enhance predictive accuracy and address potential overfitting, I apply several shrinkage methods, including Ridge Regression (RR), Least Absolute Shrinkage and Selection Operator (LASSO), Adaptive LASSO (adaLASSO), and Elastic Net (ElNet). These regularization techniques manage multicollinearity and effectively select relevant variables by penalizing less important coefficients.

I also utilize factor models to reduce dimensionality and extract key information from the dataset. Specifically, the Target Factors (TFact) approach captures underlying patterns that influence COFIX. Additionally, I implement ensemble and boosting methods, such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost), to model complex nonlinear relationships between predictors and the target variable. These methods are particularly suitable for high-dimensional datasets with potential interactions among variables.

For RF and XGBoost, I test datasets selected for each forecast horizon using the Boruta Algorithm. Although the Boruta Algorithm is based on RF and may not align perfectly with XGBoost, incorporating the selected variables significantly improves XGBoost's forecasting performance. This highlights that the variables identified by the Boruta Algorithm are robust and effectively enhance predictions across different modeling techniques.

IV. Results

The dataset expands to 408 explanatory variables by incorporating four lagged terms for each of the 102 original variables. The Boruta Algorithm identifies a total of 14 "important" variables for the 1- and 3-month horizons and 16 for the 6-month horizon. The rankings, presented in Table 2, indicate that while the variables selected differ across horizons, they exhibit consistency within each horizon.

Table 2. Variable Rankings determined by Boruta Algorithm

	1 month ahead	3 months ahead	6 months ahead
1	YT_1 (1)	USCPI (4)	IPI_B (4)
2	YT_3 (1)	BSI_P (4)	USCPI (1)
3	PI_CA (1)	SPI_I (1)	TradeP (4)
4	Unemployed (6-12) (2)	BSI_HR (1)	HousePrice_S (4)
5	YT_5 (1)	BSI_HR (2)	BSI_P (1)
6	BaseRate (1)	BSI_HR (4)	BSI_HR (1)
7	CB_3_AA (1)	TB_spread (4)	TradeP (3)
8	SPI_RI (3)	BSI_BC (4)	BSI_BC (1)
9	BSI_BC (2)	USCPI (3)	LaggingCI (2)
10	SPI_I (3)	TB_spread (2)	HousePrice (4)
11	BSI_BC (1)	BSI_HR (3)	SPI_RI (1)
12	BSI_SG (1)	LaggingCI (1)	SPI_I (4)
13	BSI_SG (2)	BSI_SG (1)	Unemployed_3mts (3)
14	BSI_HR (3)	SPI_F (1)	SPI_I (1)
15			Employed_total (2)
16			HousePrice_S (3)

For the 1-month horizon, the key variables include the first lag of Treasury bond yields of various maturities, the base rate, Business Survey Indices (BSI), and Supply Price Indices (SPI). These variables directly influence COFIX over the short term, reflecting the immediate effects of monetary policy and market conditions.

For the 3-month horizon, additional variables emerge, such as the third and fourth lags of the U.S. Consumer Price Index (CPI) and Treasury bond spreads. These variables reflect the broader influence of the U.S. economic environment, indirectly impacting COFIX by shaping the Bank of Korea's base rate decisions. This highlights the role of international economic conditions in medium-term forecasts.

For the 6-month horizon, alongside BSI and SPI, macroeconomic variables such as house prices, employment indicators, and Korea's Trade Policy Uncertainty Index gain prominence. These variables capture the influence of broader macroeconomic trends on COFIX, emphasizing their relevance for long-term forecasts.

Using cross-validation, I determine the optimal number of variables for each horizon to be 12 for the 1- and 6-month horizons and 8 for the 3-month horizon. This ensures the selection of a concise, yet informative subset of variables tailored to the specific forecasting horizon.

Figure 2. The Number of Variables and In-Sample OOB MSE

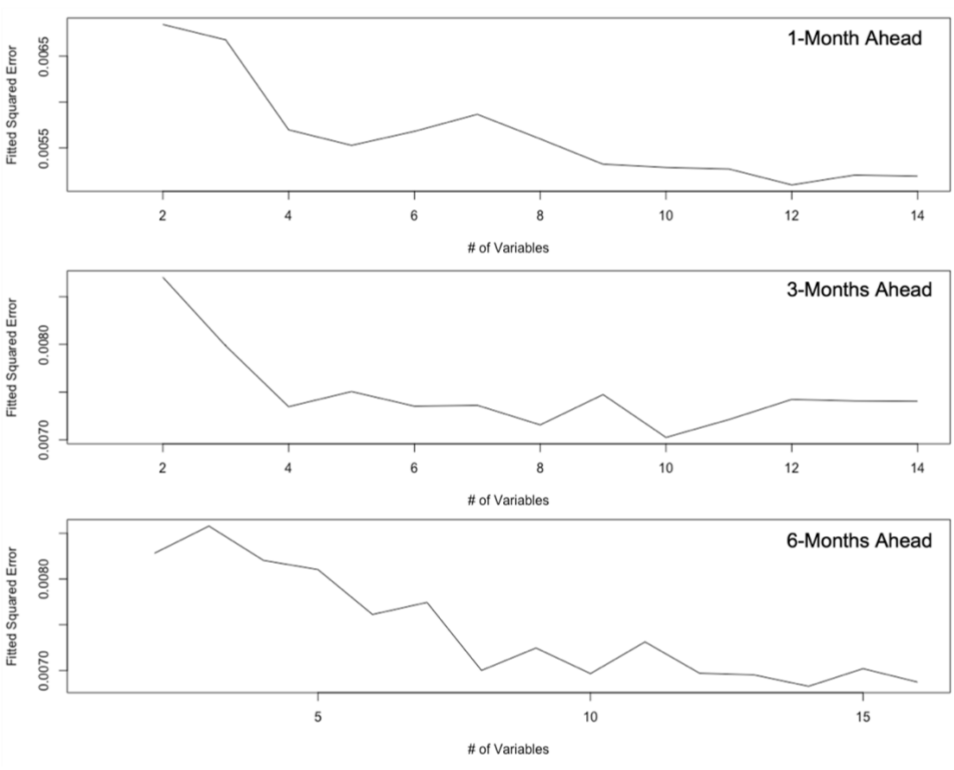

Table 3 presents the Root Mean Squared Errors (RMSE) and Mean Absolute Errors (MAE) for each model and forecast horizon. For the 1-month horizon, most models exhibit similar levels of prediction error. However, the models RF_selected and XGB_selected, which use variables selected by the Boruta Algorithm, demonstrate notably lower errors compared to others. This highlights the advantage of variable selection in short-term forecasting.

For the 3- and 6-month horizons, prediction errors generally increase with the length of the forecast horizon, reflecting the greater difficulty in predicting over longer periods. Despite this, RF_selected and XGB_selected consistently outperform other models, maintaining comparatively better predictive accuracy. These results underscore the robustness of the Boruta Algorithm in identifying explanatory variables that enhance model performance over extended timeframes.

While the Boruta Algorithm is traditionally associated with RF, its application to XGBoost as a robustness check yields substantial improvements in XGBoost's performance. Across all forecast horizons, XGB_selected consistently achieves the smallest forecast errors, demonstrating its effectiveness as a predictive model when paired with a carefully curated set of variables.

The Giacomini-White (GW) test is employed to assess the comparative performance of the forecasting models. As shown in Table 4, the results indicate no significant superiority among the selected models for the 1- and 3-month horizons. This suggests that the predictive accuracy of the models is relatively similar over shorter timeframes. However, for the 6-month horizon, the XGB_selected model demonstrates a clear advantage, outperforming many other models. Its forecast errors remain relatively low despite the general increase in prediction errors across longer horizons, highlighting its robustness and effectiveness for longer-term forecasts.

Table 3. Model Performances Measured by RMSE and MAE for Each Horizon

	RMSE	MAE	RMSE	MAE	RMSE	MAE
	One Month Ahead		Three Months Ahead		Six Months Ahead	
RW	0.124619421	0.089	0.14320149	0.102	0.165378354	0.119
AR	0.127395222	0.088467947	0.151918581	0.102178073	0.151961691	0.106729511
LASSO	0.119712577	0.080207118	0.154827965	0.120279306	0.159777278	0.130753998
adaLASSO	0.128031957	0.087133686	0.14301554	0.097141541	0.143394396	0.092158403
ElNet	0.114978742	0.075423615	0.137145771	0.096920728	0.148256201	0.109470299
adaElNet	0.123409858	0.090268848	0.148121589	0.102468672	0.158484205	0.108747564
Ridge	0.110254062	0.070515304	0.130988827	0.091699571	0.139926382	0.103642627
T. Factor	0.177352007	0.131113114	0.170186245	0.125790382	0.177760041	0.126888964
RF	0.121195711	0.084123522	0.130207451	0.092448244	0.130130973	0.093095733
XGBoost	0.116139923	0.08120311	0.118255116	0.084890094	0.125062087	0.090219024
RF_selected	0.098830009	0.064738911	0.111307425	0.081238585	0.109846165	0.075682678
XGB_selected	0.090905809	0.067335299	0.100677933	0.073206475	0.102820074	0.076389901

Table 4. Giacomini-White Test for Predictive Ability

Panel A.	RF_selected			Panel B.	XGB_selected		
	1 month	3 month	6 month		1 month	3 month	6 month
RW	0.5058	0.1449	0.0055	RW	0.2103	0.1556	0.0000
AR	0.1861	0.1309	0.0751	AR	0.1146	0.1563	0.0079
LASSO	0.9378	0.0162	0.0086	LASSO	0.3656	0.0231	0.0000
adaLASSO	0.5763	0.1794	0.8245	adaLASSO	0.1421	0.2134	0.2093
ElNet	0.6636	0.2517	0.0249	ElNet	0.5324	0.1995	0.0008
adaElNet	0.3094	0.1467	0.1747	adaElNet	0.0411	0.2009	0.0525
Ridge	0.3211	0.4803	0.1122	Ridge	0.7365	0.3281	0.0003
T. Factor	0.0026	0.0003	0.0031	T. Factor	0.0001	0.0104	0.0003
RF	0.6937	0.2907	0.5544	RF	0.0991	0.2907	0.0346
XGBoost	0.1111	0.6458	0.0006	XGBoost	0.6630	0.4256	0.8831

Table 5. MCS Test Results

Model Confidence Set (α = 0.5)			
	1-Month	3-Months	6-Months
RW	5		
AR			
LASSO	6		
adaLASSO			
ElNet	4		
adaElNet			
Ridge	3		
T. Factor			
RF			
XGBoost	7		
RF_selected	2		
XGB_selected	1	1	1

The MCS test is used to identify models with superior predictive power for each forecast horizon. As shown in Table 5, multiple models are included in the confidence set for the 1-month horizon, reflecting similar levels of predictive accuracy among the top-performing models over short-term forecasts. However, for the 3- and 6-month horizons, the XGB_selected model consistently appears in the confidence set, distinguishing itself from other models. This highlights the strong performance of XGB_selected in multi-step forecasts, particularly for longer horizons, where its predictive accuracy remains robust.

V. Conclusions

This study employs various machine learning methods to predict COFIX, the target variable. While the variables selected by the Boruta Algorithm differ significantly across horizons—making direct cross-horizon comparisons challenging—a notable consistency is observed within each horizon. This consistency enables meaningful interpretation, particularly highlighting the importance of macroeconomic variables in long-term forecasts.

Among the models tested, RF and XGBoost demonstrate superior predictive performance, especially when combined with the Boruta Algorithm for variable selection. The GW and MCS tests further confirm that these models perform better in multi-step forecasts than in 1-month forecasts, underscoring their effectiveness in capturing longer-term trends. Notably, applying the Boruta Algorithm's selected variables to XGBoost, an unconventional approach, significantly enhances its predictive accuracy. This finding suggests that the Boruta Algorithm is both robust and broadly applicable, potentially benefiting other models beyond RF.

Research on COFIX prediction remains limited, with existing studies predominantly focusing on its relationship with housing mortgage loan interest rates using traditional econometric methods. By introducing machine learning techniques, this study provides a practical tool that can assist stakeholders in making more informed and rational decisions. Future research could explore additional models and methodologies not addressed here, further advancing the understanding and forecasting of COFIX.

VI. References

Bregman, L. (2001), Random Forests, *Machine Learning*, 45, 5–32.

Christensen, K., Siggaard, M., & Veliyev, B. (2021). A machine learning approach to volatility forecasting. *Available at SSRN*.

Kim, B. Y., & Han, H. (2022). Multi-Step-Ahead Forecasting of the CBOE Volatility Index in a Data-Rich Environment: Application of Random Forest with Boruta Algorithm. *Korean Economic Review*, *38*(3), 541-569.

Kim, W. S., & Han, K. S. (2017). Do COFIX rates lead housing mortgage loan interest rates? *Korean Journal of Business Administration*, 30(12), 2127-2145.

Kursa, M. B., and W. R. Rudnicki (2010), "Feature Selection with the Boruta Package," *Journal of Statistical Software*, 36, 1–13.

Medeiros, M. C., G. F. Vasconcelos, A. Veiga, and E. Zilberman (2019), "Forecasting Inflation in a Data-rich Environment: The Benefits of Machine Learning Methods," *Journal of Business & Economic Statistics*, 1–45.

VII. Appendix

Table A1. Explanatory Dataset with Appropriate Transformation Codes

	Code	Description	tcode
1	PI_CA	Capital goods Production index(seasonally adjusted)	5
2	SI_CA	Capital goods Shipment index(seasonally adjusted)	5
3	II_CA	Capital goods Inventory index(seasonally adjusted)	5
4	PI_I	Intermediate goods Production index(seasonally adjusted)	5
5	SI_I	Intermediate goods Shipment index(seasonally adjusted)	5
6	II_I	Intermediate goods Inventory index(seasonally adjusted)	5
7	PI_CO	Consumers' goods Production index(seasonally adjusted)	5
8	SI_CO	Consumers' goods Shipment index(seasonally adjusted)	5
9	II_CO	Consumers' goods Inventory index(seasonally adjusted)	5
10	PI_W	Whole country Production index(seasonally adjusted)	5
11	SI_W	Whole country Shipment index(seasonally adjusted)	5
12	II_W	Whole country Inventory index(seasonally adjusted)	5
13	PI_MM	Monthly survey of Mining and Manufacturing Production index(seasonally adjusted)	5
14	SI_MM	Monthly survey of Mining and Manufacturing Shipment index(seasonally adjusted)	5
15	II_MM	Monthly survey of Mining and Manufacturing Inventory index(seasonally adjusted)	5
16	PI_MQ	Monthly survey of Mining and Quarrying Production index(seasonally adjusted)	5
17	SI_MQ	Monthly survey of Mining and Quarrying Shipment index(seasonally adjusted)	5
18	II_MQ	Monthly survey of Manufacturing Production index(seasonally adjusted)	5
19	PI_MF	Monthly survey of Manufacturing Shipment index(seasonally adjusted)	5
20	SI_MF	Monthly survey of Manufacturing Inventory index(seasonally adjusted)	5
21	II_MF	Monthly survey of Electricity, gas and steam supply Production index(seasonally adjusted)	5
22	SI_E	Monthly survey of Electricity, gas and steam supply Shipment index(seasonally adjusted)	5
23	Unemployment	Unemployment rate	2
24	Act_pop_total	Total Economically active population (Thousand Person)	5
25	Employed_total	Total Employed persons (Thousand Person)	5
26	Unemployed_3mts	Unemployed persons by duration of seeking for work Less than 3 months	5
27	Unemployed_3_6mts	Unemployed persons by duration of seeking for work 3 Months and over & less than 6 months	5
28	Unemployed_6_12mts	Unemployed persons by duration of seeking for work 6 Months and over & less than 12 months	5
29	Unemployed_ov6	Unemployed persons by duration of seeking for work 6 months and over	5
30	Unemployed_ov12	Unemployed persons by duration of seeking for work 12 Months and over	2

31	HousePrice	House price index national	5
32	HousePrice_S	House price index - Seoul capitial area	5
33	HousePrice_P	House price index - provincial area	5
34	BSI_BC	National Future Tendency Business Condition BSI	1
35	BSI_SG	National Future Tendency Sales Growth BSI	1
36	BSI_P	National Future Tendency Profitability BSI	1
37	BSI_FS	National Future Tendency Financial Situation BSI	1
38	BSI_HR	National Future Tendency Human Resources BSI	1
39	PI_P	Index of all industry production Industrial production	5
40	PI_C	Index of all industry production Construction	5
41	PI_SI	Index of all industry production Service Industry	5
42	PI_PA	Index of all industry production Public administration	5
43	MB	Monetary Base(Average, SA)	5
44	Currency_cir	Currency in circulation	5
45	CBLIAB_DC	Central bank liabilities to depository corporations	5
46	M1	M1	5
47	M2	M2	5
48	CD	CD	2
49	CP	CP	2
50	YT_1	Yields of Treasury Bonds(1-year)	2
51	YT_3	Yields of Treasury Bonds(3-year)	2
52	YT_5	Yields of Treasury Bonds(5-year)	2
53	YT_10	Yields of Treasury Bonds(10-year)	2
54	CB_3_AA	Yields of Corporate Bonds (AA-)	2
55	BaseRate	Bank of Korea Base Rate	2
56	CB_3_BBB	Yields of Corporate Bonds (BBB-)	2
57	FX_US	Won per United States Dollar(Basic Exchange Rate)	5
58	FX_JP	Won per Japanese Yen(100Yen)	5
59	FX_EU	Won per Euro	5
60	FX_BR	Won per United Kingdom Pound	5
61	FX_CA	Won per Canadian Dollar	5
62	FX_SW	Won per Swiss Franc	5
63	CPI	Finished Consumer Goods	5
64	CPI_F	Consumer Price indices Food and non-alcoholic beverages	5
65	CPI_C	Consumer Price indices Clothing and footwear	5
66	CPI_H	Consumer Price indices Furnishings, household equipment and rountine household maintenance	5
67	CPI_HE	Consumer Price indices Health	5
68	CPI_T	Consumer Price indices Transport	5
69	CPI_E	Consumer Price indices Education	5
70	CPI_M	Consumer Price indices Miscellaneous goods and services	5
71	SPI_R	Domestic Supply Price indices Raw materials	5

72	SPI_I	Domestic Supply Price indices Intermediate goods and services	5
73	SPI_F	Domestic Supply Price indices Final goods and services	5
74	SPI_RI	Domestic Supply Price indices Raw & intermediate materials	5
75	IPI_A	Import Price Indices(Basic Groups) Agricultural, forestry & marine products	5
76	IPI_B	Import Price Indices(Basic Groups) Bitnminous coal	5
77	IPI_C	Import Price Indices(Basic Groups) Crude oil	5
78	IPI_L	Import Price Indices(Basic Groups) Liquified natural gas	5
79	IPI_M	Import Price Indices(Basic Groups) Manufacturing products	5
80	EPI_A	Export Price Indices(Basic Groups) Agricultural, forestry & marine products	5
81	EPI_M	Export Price Indices(Basic Groups) Manufacturing products	5
82	KOSPI	KOSPI index	5
83	KOSDAQ	KOSDAQ index	5
84	Expected_inf	Expected Inflation	2
85	Oilprice	Oilpricex (Fred-MD)	5
86	TB_spread	Treasury Bonds spread(10year-1year) (Fred-MD)	2
87	USCPI	US CPI index CPIAUCSL (Fred-MD)	5
88	IR_bigfirm	Interest Rates Loans to Big-corporates	2
89	IR_smallfirm	Interest Rates Loans to Small-corporates	2
90	IR_firm	Interest Rates Loans to corporate	2
91	IR_House	Interest Rates Loans to house	2
92	IR_saving	Interest Rates Loans to Deposits	2
93	IR_gen_credit	Interest Rates Loans to Credit	2
94	USB_10	US 10-year bond yield	2
95	LeadingCI	Composite Index of Business Indicators Leading Composite Index	5
96	LaggingCI	Composite Index of Business Indicators Lagging Composite Index	5
97	vkospi	Volatility index of KOSPI200	1
98	Korean Overall EPU	Korean Overall Economic Policy Uncertainty index	4
99	MonetaryP	Korean Monetary Policy Uncertainty index	4
100	FiscalP	Korean Fiscal Policy Uncertainty index	4
101	TradeP	Korean Trade Policy Uncertainty index	4
102	FXP	Korean Foreign Exchange Market Policy Uncertainty index	4