

랜덤 포레스트(Random Forest)의 시계열 적용에 관한 연구: 한국 물가상승률 예측 사례 분석

한 희 준*

논문 초록 본고는 한국과 미국의 물가상승률 예측에 랜덤 포레스트 모델을 적용할 때, Stationary Bootstrap이나 Moving Block Bootstrap 등 Block Bootstrap을 사용하는 것이 통상적인 독립 부트스트랩(Independent Bootstrap)을 사용하는 것에 비해 통계적으로 유의한 수준으로 예측력을 개선하지는 못한다는 것을 보인다. 그리고 FRED-MD를 참고한 총 93개의 관련 국내외 거시경제/금융 변수들을 사용하고, XGBoost, LSTM 등 다양한 머신러닝 방법을 활용하여 한국의 물가상승률을 예측하고 분석한다. 2004년 9월에서 2022년 3월까지의 표본을 이용하였고, 1개월에서 12개월의 예측 대상기간(Forecast Horizon)을 고려하였다. 총 13개의 모형 중 대부분의 예측 대상기간에 있어 예측력이 우수한 모형이 존재하는 것으로 나타났는데, 이는 보루타 알고리즘(Boruta Algorism)을 통해 중요한 변수로 분류된 변수들만을 랜덤 포레스트에 적용하는 모형이다. Giacomini and White(2006)와 Hansen et al.(2009)의 검정을 통해 대부분의 예측 대상기간에서 통계적으로 유의하게 예측력이 우수함을 확인하였는데, 특히 경제활동인구 및 취업자 수의 증가율 등 고용시장 관련 변수, 기업경기실사지수, 주택가격 변화율 등이 물가상승률 예측에 중요한 변수로 선택되는 것으로 나타났다.

핵심 주제어: 랜덤 포레스트, 부트스트랩, 물가상승률, 예측, 머신러닝

경제학문헌목록 주제분류: C32, C50, E31

투고 일자: 2023. 2. 9. 심사 및 수정 일자: 2023. 4. 26. 게재 확정 일자: 2023. 8. 25

* 성균관대학교 경제학과/퀀트응용경제학과 교수, e-mail: heejoonhan@skku.edu

I. 서 론

Stock and Watson (2010) 은 미국 물가상승률 예측에서 임의보행 등 단순한 모형들보다 체계적으로 우월한 모형을 찾기가 매우 어렵다고 강조하였고, 이러한 결론은 여러 논문에 의해 확인되었다(Faust and Wright, 2013 참조). 그러나 이러한 기존 연구들은 최근 활발히 적용되는 머신러닝 기법들이 고려되지 않은 것이었다. Medeiros et al. (2021) 은 FRED-MD¹⁾ 데이터 전체를 설명변수로 이용하고 다양한 시계열 모형 및 머신러닝 기법들을 적용하여 미국 물가상승률을 예측하고 분석하였는데, 머신러닝 기법들이 임의보행, 자기회귀모형 등의 벤치마크(Benchmark) 모형보다 우월한 예측치를 제공한다는 것을 보였다. 특히 Medeiros et al. (2021) 은 예측 대상기간이 길수록 Breiman (2001) 이 제안한 랜덤 포레스트 모형의 예측력이 우수함을 보였으며, 이는 랜덤 포레스트 모형이 변수 간의 비선형 관계를 모형화하고 주요 설명변수 선택에 뛰어난 점들이 중요한 역할을 하기 때문이라 것을 강조하였다.

랜덤 포레스트²⁾ 기법에는 원래의 데이터에서 부트스트랩 샘플을 추출하는 절차가 포함되어 있는데, 횡단면 데이터에는 통상적인 독립 부트스트랩을 사용한다. 반면 Medeiros et al. (2021) 은 시계열을 다루기 때문에 랜덤 포레스트를 적용할 때 Block Bootstrap을 사용한다고 논문의 4.5장에 명시하였다. 이는 정상시계열(Stationary Time Series)의 부트스트랩 샘플을 추출할 때는 Moving Block Bootstrap 또는 Stationary Bootstrap 등의 Block Bootstrap을 적용하는 것이 시계열의 의존성을 반영하기에 더욱 적절한 것으로 알려져 있기 때문이다. 그러나 저자들이 공개한 R 코드³⁾는 랜덤 포레스트에서 통상적인 독립 부트스트랩을 사용하고 있으며, 코드를 공개한 저자에게 문의한 결과 논문에서 명시한 것과 달리 실제로는 독립 부트스트랩을 랜덤 포레스트에 사용하였음을 확인받았다. 본 연구의 첫 번째

1) 미국의 거시경제 연구를 위한 월별 데이터 세트인 St. Louis 연방준비은행에서 업데이트하며, 다음 웹사이트에서 이용 가능하다. <https://research.stlouisfed.org/econ/mccracken/fred-databases/>

2) 랜덤 포레스트는 배깅(Bootstrap Aggregating, Bagging)을 기반으로 하면서 임의 노드 분할(Random Node Split)을 통해 부트스트랩 표본에서 추정된 결정 트리(Decision Tree)들의 상관관계를 제거(Decorrelate)하여 결과적으로 분산을 낮추는 방법이다.

3) <https://github.com/gabrielrvsc/>

동기는 실제 정상시계열 예측에 랜덤 포레스트를 적용할 때 독립 부트스트랩 대신 시계열 의존성을 반영하는 Block Bootstrap을 사용하는 것이 모형의 예측력에 유의한 차이를 야기하는지를 살펴보는 것이다. 다만 본 연구는 실제 우리나라와 미국의 물가상승률 예측과 관련된 특정 데이터만을 이용하여 이를 살펴보고 있으며, 정상시계열 예측의 일반적인 상황에 적용되는 연구를 수행하는 것은 아니라는 점을 명확히 하고자 한다.

본 연구의 두 번째 동기는 우리나라 물가상승률 예측과 관련된 기존 연구들은 고려한 설명변수의 수 또는 적용한 머신러닝 기법이 제한되어 있으며, FRED-MD와 같이 방대한 거시경제/금융 데이터를 활용⁴⁾하면서 Medeiros et al. (2021) 처럼 다양한 머신러닝 기법들을 적용한 연구가 현재까지 없다는 점이다. 강규호 외 (2021), 김종화·이중식(2005), 김현학(2015), 이진희·김덕과(2014), 최인·황승진(2012) 등은 우리나라 물가상승률 예측을 위해 시계열 모형들과 예측조합 등 다양한 계량경제모형들을 고려하였으나, 랜덤 포레스트, XGBoost (Extreme Gradient Boosting), LSTM (Long Short-Term Memory) 등 최근 실무에서 활발히 예측에 사용되는 머신러닝 기법들은 고려하지 않고 있다. 참고로 Medeiros et al. (2021) 또한 XGBoost 모형이나 LSTM 모형은 고려하지 않고 있다. 추가로 본 연구에서는 모형별 예측오차 비교에 그치는 것이 아니라 관련 연구 분야에서 표준적으로 사용하는 Giacomini-White (2006)의 검정과 Hansen et al. (2011)의 Model Confidence Set (MCS) 검정을 시행하여 모형의 예측력 우수성이 통계적으로 유의한 지도 살펴보고자 한다.

본 논문은 2004년 9월부터 2022년 3월까지의 월별 소비자 물가상승률을 종속변수로 사용하였고, 이의 예측을 위해 FRED-MD와 기존 연구를 참고하여 총 93개의 국내외 거시경제/금융 데이터를 수집/활용하였다. 정상시계열로 변환한 후 각 변수와 변수들에서 추출한 4개의 주성분요소(Principal Component Factor)의 1차에서 4차까지의 시차변수들을 개별 설명변수로 취급한 결과, 각 머신러닝 모형에 사용하는 설명변수의 수는 총 388개가 되었다. 표준적인 예측 절차인 구간이동(Rolling Window) 방법을 채택하여 1개월에서 12개월 앞 표본 외(Out-of-sample) 예측치들

4) Medeiros et al. (2021)은 FRED-MD의 123개(2022년 2월 vintage 기준) 변수와 이 변수들로부터 추출한 4개의 주성분요소(Principal Component Factor)들의 과거값들(1차에서 4차까지의 시차변수)을 포함하여 총 508개의 설명변수를 사용하였다.

을 생성하였다.

본 논문의 주요 결과는 다음과 같다. 첫째, 우리나라와 미국의 물가상승률 예측 관련 데이터를 이용해서 살펴본 결과, Moving Block Bootstrap 또는 Stationary Bootstrap 등 Block Bootstrap을 이용하는 것이 통상적인 독립 부트스트랩을 이용하는 것에 비해 랜덤 포레스트 모형의 예측력을 통계적으로 유의한 수준으로 개선하지는 않는 것으로 나타났다. 둘째, XGBoost 모형이나 LSTM 모형 등을 포함한 총 13개의 시계열 모형 및 머신러닝 기법들을 우리나라 물가상승률 예측에 적용한 결과, 대부분의 예측 대상기간에서 가장 낮은 예측오차를 지니는 예측력이 우월한 모형이 존재하는 것으로 나타났다. 이 모형은 보루타 알고리즘을 통해 중요한 변수로 분류된 변수들만을 랜덤 포레스트에 적용하는 것인데, 이는 Kim and Han (2022)이 제안한 모형에서 교차검증 부분을 제외하여 더욱 간소화한 것으로 본 논문에서 제안하고 처음으로 예측력을 평가해 본 모형이다. Giacomini-White 검정과 MCS 검정 결과를 살펴보면, 대다수 경쟁 모형들에 비해 통계적으로 유의한 수준으로 이 모형의 예측력이 우월한 것으로 나타났다.

마지막으로 우리나라 물가상승률 예측에 중요한 역할을 하는 설명변수가 무엇인지를 살펴보았는데, 보루타 알고리즘을 통해 중요한 변수로 분류되는 변수들과 Adaptive LASSO 모형에 의해 선택된 변수들을 비교하였다. 보루타 알고리즘은 중요한 변수들을 변수 중요도의 순위와 함께 제시하는데, 가장 큰 특징은 대부분의 예측 대상기간에 걸쳐 경제활동인구 및 취업자 수의 증가율 등 고용시장 관련 변수들이 변수 중요도의 높은 순위를 차지한다는 점이다. 이외 식료품, 보건, 교육 분야 물가지수와 유가 등 물가지수와 직접적으로 관련된 변수들뿐만 아니라 기업경기실사지수와 주택매매가격지수 등도 여러 예측 대상기간에 걸쳐 포함되었다. 반면 Adaptive LASSO 모형에 의해 선택된 변수들을 살펴보면, 물가지수들의 과거값, 고용시장 관련 변수, 기업경기실사지수 등 보루타 알고리즘이 중요한 변수로 분류한 것들과 겹치는 경우가 있었으나 예측 대상기간이 길어질수록 선택된 변수의 수가 크게 줄어들어 유용한 정보를 제공하는 데에는 커다란 한계가 있는 것으로 나타났다.

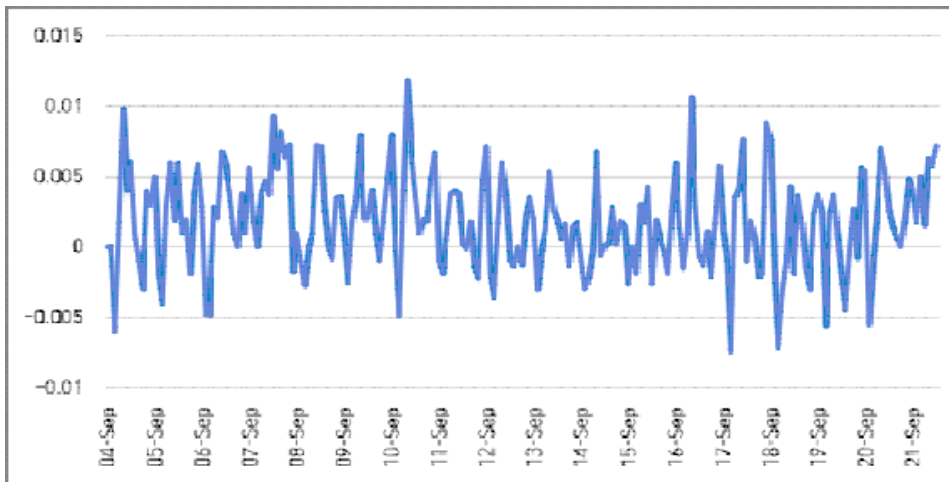
본 논문의 구성은 다음과 같다. 제Ⅱ장에서는 우리나라 물가상승률 예측을 위해 사용하는 거시경제/금융 데이터 개요와 각 데이터를 정상시계열로 변환하는 과정을 설명한다. 제Ⅲ장에서는 우리나라와 미국의 데이터를 활용하여 랜덤 포레스트에 각

각 다른 부트스트랩 방법을 사용할 때의 예측력을 비교한다. 제Ⅳ장에서는 우리나라 물가상승률 예측에 고려한 시계열 모형 및 머신러닝 방법들과 연구에 사용한 프로그램과 코드를 설명한다. 제Ⅴ장에서는 우리나라 물가상승률 예측 결과, 예측력 검정 결과 및 예측에 중요한 역할을 하는 변수 등을 분석한다. 제Ⅵ장에서 연구의 결론을 제시한다.

Ⅱ. 데이터 및 정상시계열로의 변환

본 연구에서는 2004년 9월부터 2022년 3월까지의 월별 소비자 물가상승률(소비자 물가지수의 전월 대비 로그차분)을 종속변수로 사용하였다. 세인트루이스 연방준비은행(St. Louis Fed)의 FRED-MD를 구성하는 변수들과 유사한 우리나라의 거시경제 데이터를 수집하고 이용하였는데, 주택매매가격지수들(전국, 수도권, 지방)이 2004년 8월부터 이용 가능하므로 로그차분 후 표본 기간이 2004년 9월부터 시작하게 된다. 〈Figure 1〉은 표본 기간의 물가상승률을 보이고 있는데, 확률적 추세(Stochastic Trend)가 없어 보이는 것처럼 단위근 검정 결과 안정적인 시계열인 것으로 나타났다.

〈Figure 1〉 Inflation Rate (CPI) in South Korea



Notes: The figure provides the log difference of CPI index (month-on-month growth rate) in South Korea from September 2004 to March 2022.

FRED-MD는 127개 (2022년 2월 vintage 기준)의 거시/금융 월별 데이터로 구성되어 있으며, 산출/소득, 노동시장, 주택 착공/허가, 수주/재고, 통화/신용, 금리/환율, 가격, 주식시장 등 총 8개 범주의 광범위한 데이터를 포괄한다. 본 연구를 위해 최대한 FRED-MD의 구성 변수들에 대응하는 우리나라의 거시/금융 자료를 수집하였다. 이에 더해 우리나라 물가상승률 예측 관련 기존 연구에서 사용한 변수들을 추가하여 데이터 세트(Dataset)를 구축하였는데, 예를 들어 강규호 외 (2021)에서 사용한 기대 인플레이션, 주택가격지수, 주택담보대출 금리, 미국 소비자 물가지수, 미국 10년 만기 국채 금리, 경기선행지수 등을 추가하였다. 그 결과 본 연구에서는 소비자 물가상승률을 포함하여 총 93개의 국내외 거시경제/금융 변수를 활용하였으며, <Table 1>은 구체적인 변수들을 제시하고 있다.

<Table 1> List of Variables Used for Korean Inflation Forecasting

	Code	Description	Tcode
1	PL_CA	Capital goods Production index (seasonally adjusted)	3
2	SL_CA	Capital goods Shipment index (seasonally adjusted)	3
3	IL_CA	Capital goods Inventory index (seasonally adjusted)	3
4	PL_I	Intermediate goods Production index (seasonally adjusted)	3
5	SL_I	Intermediate goods Shipment index (seasonally adjusted)	3
6	IL_I	Intermediate goods Inventory index (seasonally adjusted)	3
7	PL_CO	Consumers' goods Production index (seasonally adjusted)	3
8	SL_CO	Consumers' goods Shipment index (seasonally adjusted)	3
9	IL_CO	Consumers' goods Inventory index (seasonally adjusted)	3
10	PL_W	Whole country Production index (seasonally adjusted)	3
11	SL_W	Whole country Shipment index (seasonally adjusted)	3
12	IL_W	Whole country Inventory index (seasonally adjusted)	3
13	PL_MM	Monthly survey of Mining and Manufacturing Production index (seasonally adjusted)	3
14	SL_MM	Monthly survey of Mining and Manufacturing Shipment index (seasonally adjusted)	3
15	IL_MM	Monthly survey of Mining and Manufacturing Inventory index (seasonally adjusted)	3
16	PL_MQ	Monthly survey of Mining and Quarrying Production index (seasonally adjusted)	3
17	SL_MQ	Monthly survey of Mining and Quarrying Shipment index (seasonally adjusted)	3

	Code	Description	Tcode
18	IL_MQ	Monthly survey of Manufacturing Production index (seasonally adjusted)	3
19	PI_MF	Monthly survey of Manufacturing Shipment index (seasonally adjusted)	3
20	SI_MF	Monthly survey of Manufacturing Inventory index (seasonally adjusted)	3
21	IL_MF	Monthly survey of Electricity, gas and steam supply Production index (seasonally adjusted)	3
22	SI_E	Monthly survey of Electricity, gas and steam supply Shipment index (seasonally adjusted)	3
23	Unemployment	Unemployment rate	2
24	Act_pop_total	Total Economically active population (Thousand Person)	3
25	Employed_total	Total Employed persons (Thousand Person)	3
26	Unemployed_3mts	Unemployed persons by duration of seeking for work Less than 3 months	3
27	Unemployed_3_6 mts	Unemployed persons by duration of seeking for work 3 Months and over & less than 6 months	3
28	Unemployed_6_12 mts	Unemployed persons by duration of seeking for work 6 Months and over & less than 12 months	3
29	Unemployed_ov6	Unemployed persons by duration of seeking for work 6 months and over	3
30	Unemployed_ov12	Unemployed persons by duration of seeking for work 12 Months and over	2
31	HousePrice	House price index national	3
32	HousePrice_S	House price index - Seoul capital area	3
33	HousePrice_P	House price index - provincial area	3
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	3
40	PI_C	Index of all industry production Construction	3
41	PI_SI	Index of all industry production Service Industry	3
42	PI_PA	Index of all industry production Public administration	3
43	MB	Monetary Base (Average, SA)	3
44	Currency_cir	Currency in circulation	3
45	CBLIAB_DC	Central bank liabilities to depository corporations	3
46	M1	M1	3

	Code	Description	Tcode
47	M2	M2	3
48	CD	CD rate (91 days)	2
49	CP	CP rate (91 days)	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	IR_House	CBs and SBs Weighted Averages of Interest Rates Loans to households (Houses)	2
58	TB_spread	Treasury Bonds spread (10year-1year)	2
59	FX_US	Won per United States Dollar (Basic Exchange Rate)	3
60	FX_CH	Won per Yuan (Basic Exchange Rate)	3
61	FX_JP	Won per Japanese Yen (100Yen)	3
62	FX_EU	Won per Euro	3
63	FX_BR	Won per United Kingdom Pound	3
64	FX_CA	Won per Canadian Dollar	3
65	FX_SW	Won per Swiss Franc	3
66	CPI	CPI: Finished Consumer Goods	3
67	CPI_F	CPI: Food and non-alcoholic beverages	3
68	CPI_C	Consumer Price indices Clothing and footwear	3
69	CPI_H	CPI: Furnishings, household equipment and routine household maintenance	3
70	CPI_HE	CPI: Health	3
71	CPI_T	CPI: Transport	3
72	CPI_E	CPI: Education	3
73	CPI_M	CPI: Miscellaneous goods and services	3
74	SPI_R	Domestic Supply Price index Raw materials	3
75	SPI_I	Domestic Supply Price index Intermediate goods and services	3
76	SPI_F	Domestic Supply Price index Final goods and services	3
77	SPI_RI	Domestic Supply Price index Raw & intermediate materials	3
78	IPI_A	Import Price Index (Basic Groups) Agricultural, forestry & marine products	3
79	IPI_B	Import Price Index (Basic Groups) Bituminous coal	3
80	IPI_C	Import Price Index (Basic Groups) Crude oil	3

	Code	Description	Tcode
81	IPI_L	Import Price Index (Basic Groups) Liquefied natural gas	3
82	IPI_M	Import Price Index (Basic Groups) Manufacturing products	3
83	EPI_A	Export Price Index (Basic Groups) Agricultural, forestry & marine products	3
84	EPI_M	Export Price Index (Basic Groups) Manufacturing products	3
85	KOSPI	KOSPI Index	3
86	KOSDAQ	KOSDAQ Index	3
87	VKOSPI	VKOSPI Index	1
88	Expected_inf	Expected Inflation	2
89	LeadingCI	Composite Index of Business Indicators Leading Composite Index	3
90	LaggingCI	Composite Index of Business Indicators Lagging Composite Index	3
91	Oilprice	Oil Price	3
92	USCPI	US CPI: CPIAUCSL in FRED MD	3
93	USB_10	US 10-year bond yield	2

Notes: The column Tcode denotes the following data transformation for a series x : (1) no transformation; (2) Δx_t ; (3) $\Delta \log(x_t)$.

설명변수로 사용된 시계열 데이터가 단위근을 가지고 있으면 이를 정상시계열로 변환하여 머신러닝 기법들을 적용하였는데, 단위근 검정 결과는 지면 관계상 생략하였다. 단위근 검정 결과 VKOSPI 지수 외 다른 변수들이 모두 단위근을 가지는 것으로 확인되었다. FRED-MD의 데이터 변환 방법을 참고하여 정상시계열로 만들기 위해 지수형 변수들은 로그차분, 이자율 변수들은 차분으로 변환하여 단위근을 제거하였다. 다만 FRED-MD의 데이터 변환 방법에 따르면 가격변수들은 로그이차차분을 실시하는데, 본 연구에서 고려하는 표본 기간의 가격변수들(우리나라 및 미국)은 로그이차차분을 실시할 필요는 없는 것으로 나타나 로그차분만을 취하였다. <Table 1>의 마지막 행에 각 변수를 정상시계열로 만들기 위한 데이터 변환 방법을 써넣었다.

다양한 머신러닝 모형들을 적용할 때 모든 설명변수의 1차에서 4차까지의 시차변수(from 1 lag up to 4 lags of all variables)들을 개별 설명변수로 취급하고 종속변수의 1차에서 4차까지의 시차변수도 설명변수로 추가하였다. 그리고 총 93개 거시경제/금융 변수와 함께 이 93개 거시경제/금융 변수에서 추출한 4개 주성분요소

(Static Factor)의 1차에서 4차까지의 시차변수도 설명변수로 포함하였다.⁵⁾ 이는 Medeiros et al. (2021)이 채택한 방식을 그대로 따른 것이다. 따라서 각 머신러닝 모형에 사용되는 설명변수의 수는 총 388개(97×4)가 되었다.

Ⅲ. 랜덤 포레스트 적용시 부트스트랩 방법에 대한 검토

부트스트랩 방법 중 시계열의 의존성을 고려하는 Block Bootstrap 방법으로 널리 사용되는 방법들은 Künsch (1989)의 Moving Block Bootstrap(MBB) 또는 Politis and Romano (1994)의 Stationary Bootstrap(SB)이다. MBB는 Block의 크기가 고정되지만 SB는 Block의 크기가 고정되지 않고 임의로 정해진다는 차이가 있으며, 이로 인해 원래의 표본이 정상시계열일 때 MBB에 의한 부트스트랩 표본은 안정적(Stationary)이지 않으나 SB에 의한 부트스트랩 표본은 안정적이라는 것을 Politis and Romano (1994)가 보였다. 그러나, Lahiri (1999)는 모든 Block 부트스트랩 방법이 같은 크기의 1차 편의(First-Order Bias)가 있으나 Block 크기가 고정된 부트스트랩 방법이 더 작은 1차 분산(First-Order Variance)를 가지고 있음을 보였다. 이는 MBB가 SB보다 주어진 표본에서 나올 수 있음(Better Finite Sample Performance)을 의미한다.

이 장에서는 이론적 검토⁶⁾ 대신 랜덤 포레스트에 적용하는 부트스트랩 방법을 달리할 때 실제 시계열 데이터에 대한 랜덤 포레스트의 예측력이 어떠한 영향을 받는지를 살펴본다. 독립 부트스트랩을 이용하는 랜덤 포레스트, SB를 이용하는 랜덤 포레스트, MBB를 이용하는 랜덤 포레스트 등 세 모형의 예측력을 비교한다. 앞 장에서 설명한 우리나라 물가상승률 예측 관련 데이터를 이용한 결과와 Medeiros et al. (2021)이 이용한 미국 물가상승률 예측 관련 데이터를 이용한 결과를 모두 살펴보았다. 우선 우리나라 데이터의 경우 전체 표본 크기는 총 211개월(2004. 9월-2022. 3월)이며, 구간이동 방법을 적용할 때 추정 기간의 크기(Window Size)는 120개월(10년)로 설정하고 예측 기간은 총 91개월로 정하였다. 예측 대상 기간은 1개월 단위로 1개월부터 12개월로 설정하였다. 예측 대상기간이 2개월에서

5) 주성분요소들은 구간이동 방법을 적용할 때 매 구간마다 새로 추정하였다.

6) Goehry (2020)는 시계열 자료를 적용하는 랜덤 포레스트에 관한 이론적 연구를 제공하고 있다.

12개월까지인 다단계 예측(Multi-Step Ahead Forecasting)에서는 Medeiros et al. (2021)와 Kim and Han (2022)과 같이 직접 예측(Direct Forecasting) 방식을 이용하였다. 이는 설명변수들의 예측치를 별도로 구할 필요가 없다는 면에서 편리할 뿐만 아니라 예측 관련 기존 연구에서 모형오지정(Model Misspecification)에 강건(Robust)하다는 장점이 있는 것으로 알려져 있기 때문이다. 미국 데이터의 경우 Medeiros et al. (2021)이 이용한 데이터(1960년 2월부터 2015년 12월)와 실행한 예측 절차를 그대로 따랐으며 예측 기간이 1990-2000년(Subsample 1, 1-step ahead 구간이동 예측시 추정 기간의 크기는 359)과 2001~ 2015년(Subsample 2, 1-step ahead 구간이동 예측시 추정 기간의 크기는 491)인 두 경우를 모두 고려하였다.

〈Table 2〉는 Gioacomini-White 검정의 통계치를 제시하는데, 이는 MAE 기준이며 무조건부(Unconditional) 검정이다.⁷⁾ 독립 부트스트랩을 사용하는 랜덤 포레스트 모형과 SB 또는 MBB⁸⁾를 사용하는 랜덤 포레스트 모형이 동일한 예측력을 가졌는지 아닌지를 테스트하는데, 통계치가 음(-)의 값을 가지면 독립 부트스트랩을 사용하는 모형의 예측오차가 낮다는 것을 의미한다. 우선 우리나라 데이터의 SB에 비교한 결과를 보면, 예측 대상기간이 3~4개월과 11개월인 경우를 제외하고는 통계치가 음(-)의 값을 가지고 특히 5개 예측 대상기간의 경우 두 모형의 예측력이 같다는 귀무가설을 기각하는 것으로 나타났다. 이는 SB보다 독립 부트스트랩을 사용하는 모형이 전반적으로 나은 예측력을 가진다는 것을 의미한다. 우리나라 데이터의 MBB에 비교한 결과를 보면, 통계치가 음(-)과 양(+)의 값을 골고루 가지는 것으로 나타났고 모든 경우에 귀무가설을 기각하지 못하는 것으로 나타났다. 이는 MBB 또는 독립 부트스트랩을 사용하는 것이 랜덤 포레스트의 예측력에 유의한 차이를 초래하지 않는다는 것을 의미한다.

미국 데이터를 살펴보면, 랜덤 포레스트에 독립 부트스트랩을 이용하는 것이 전반적으로 우월한 예측력을 가진다고 나타났다. 1990-2000년을 예측 기간으로 삼을 때(Subsample 1), 대부분의 예측 대상기간에 통계치가 음(-)의 값을 가지며, 특히

7) 제V장에 Gioacomini-White 검정의 자세한 설명이 제시되어 있다.

8) Block의 크기는 Politis and White (2004)가 제안하고 이후 Patton et al. (2009)에 의해 수정된 방법을 적용하였는데, 각 변수의 최적 Block 크기들의 최댓값을 이용하였다. 각 변수의 최적 Block 크기들의 평균을 Block의 크기로 정한 경우에도 질적으로 비슷한 결과가 도출되었다. SB의 튜닝 파라미터(Tuning Parameter)는 정해진 최적 Block 크기의 역수를 취해 정하였다.

〈Table 2〉 Giacomini-White Test Results for Random Forests Using Different Bootstrapping

	1Month	2Month	3Month	4Month	5Month	6Month	7Month	8Month	9Month	10Month	11Month	12Month
Korean Inflation Forecasting												
IB vs SB	-1.42	-2.20**	0.60	0.14	-1.20	-2.51**	-2.45**	-2.25**	-2.37**	-1.56	1.54	-0.50
IB vs MBB	-1.42	0.16	0.49	0.88	0.25	0.65	-0.34	-1.33	-1.68	-1.33	0.29	-0.30
US Inflation Forecasting Sample 1												
IB vs SB	-0.13	-0.46	-1.62	0.55	-1.90*	-2.18**	-2.30**	-0.56	-0.36	0.96	-1.26	-0.68
IB vs MBB	-3.08***	-5.18***	-5.77***	-5.60***	-4.20***	-3.74***	-3.46***	-1.97**	-2.03**	-2.08**	-2.95***	-2.73***
US Inflation Forecasting Sample 2												
IB vs SB	0.23	-1.04	-0.51	-1.31	-2.71***	-2.81***	-1.15	-0.70	-0.09	-1.52	-1.64	-1.23
IB vs MBB	0.74	-0.60	-0.85	-2.26**	-3.41***	-1.80*	-0.95	-0.87	-0.08	-0.29	-0.40	-0.54

Notes: IB, SB, and MBB indicate random forest methods using independent bootstrap, stationary bootstrap, and moving block bootstrap, respectively. The table provides unconditional Giacomini-White test statistics between two models for absolute errors. One is the random forest model using independent bootstrapping and the other is the random forest model using either stationary bootstrapping or moving block bootstrapping. A negative statistic implies that the model using independent bootstrapping has a lower forecast loss than the other. *, **, *** indicate rejection of equal predictive ability at the 10%, 5%, 1% significance level, respectively.

MBB에 대해서는 귀무가설을 기각하고 있다. 이는 독립 부트스트랩을 사용하는 모형이 MBB를 사용하는 것보다 통계적으로 유의하게 우월한 예측력을 가진다는 것을 의미한다. 2001~2015년을 예측 기간으로 삼을 때 (Subsample 2)에도 대부분의 예측 대상기간에 통계치가 음(-)의 값을 가지는 것으로 나타났다.

이상 우리나라와 미국의 물가상승률 데이터로 살펴본 바에 따르면, 시계열 예측에 랜덤 포레스트를 적용할 때 부트스트랩 방법을 굳이 Block 부트스트랩으로 변경할 필요 없이 통상적인 독립 부트스트랩을 사용하는 것이 문제가 없으며 오히려 Block 부트스트랩을 사용할 때 예측력이 다소 나빠질 수도 있음을 알 수 있다. 따라서 제 V장의 예측 및 분석에서의 랜덤 포레스트 모형들에는 모두 독립 부트스트랩을 적용하였다. 물론 이러한 결과를 시계열 전체의 경우로 일반화할 수는 없음에 주의해야 한다. 표본의 수 및 예측 절차가 크게 다를 경우에도 이러한 결과가 나타날 것이라고 예상할 수는 없다. 랜덤 포레스트의 임의 노드 분할⁹⁾이 본 논문에서 살펴본 결과와 관계가 있을 것으로 추측되는데, 이에 대한 엄밀한 이론적 연구는 추후 과제로 남긴다.

IV. 예측모형

본 연구에서는 우리나라 물가상승률 예측을 위해 다음과 같은 총 13개의 시계열 모형과 머신러닝 방법들을 고려하였다. 임의보행, 자기회귀 (AR model with lag 4), Ridge regression, LASSO (least absolute shrinkage and selection operator), Adaptive LASSO, Elastic Net, Adaptive Elastic Net, CSR (complete subset regression), 랜덤 포레스트, XGBoost, BS_RF (Boruta Selected Random Forest, 보루타 알고리즘에 의해 중요한 변수로 분류된 변수들만을 랜덤 포레스트 적용, 아래 설명 참고), 인공신경망 (Neural Network with three hidden layers and 32, 16, and 8 nodes in each hidden layer), LSTM (Long Short-Term Memory) 등이다.¹⁰⁾ 이 모

9) 임의 노드 분할은 주어진 부트스트랩 표본으로 결정 트리를 추정할 때 각 노드 분할시 임의로 선택된 설명변수를 이용하는 것이다. 전체 설명변수가 m 개일 때, 보통 $m/3$ 개 설명변수를 임의로 선택한다. 분류 (Classification)의 경우는 \sqrt{m} 개를 선택한다.

10) Bai and Ng (2008)의 Target Factor 모형도 고려하였으나 예측오차가 너무 크게 나와 이를 제외하였다.

형들은 BS_RF 모형을 제외하고는 머신러닝 방법을 활용한 경제/금융 시계열 예측에서 널리 사용되는 방법이다. 모형들에 대한 간단한 설명은 〈부록 1〉에 추가하였다.

BS_RF 모형은 Kim and Han(2022) 과 장나원·한희준(2022) 이 사용한 모형과 비슷하지만 한 가지 차이점이 있다. Kim and Han(2022) 은 보루타 알고리즘¹¹⁾ 과 교차검증(Cross-Validation) 그리고 랜덤 포레스트를 통해 예측치를 도출하는 방법을 제안하고,¹²⁾ 이 방법이 VIX 지수를 예측할 때 예측 대상기간이 길수록 우수한 예측력을 가진다는 것을 보인 바 있다. 장나원·한희준(2022) 은 우리나라 수출입 증가율 예측에서 Kim and Han(2022) 의 모형이 우수한 예측력을 가진다는 것을 보였다. BS_RF 모형은 Kim and Han(2022) 의 방법에서 교차검증을 통해 최적 변수의 수를 결정하는 단계를 생략하고 보루타 알고리즘이 중요한 변수로 분류한 변수들을 모두 사용하여 랜덤 포레스트에 적용하는 것이 특징이다. 이는 예측 방법을 단순하게 하고 교차검증으로 인한 계산 시간을 줄이는 장점이 있다. 본고의 데이터에 적용할 때 BS_RF 모형과 Kim and Han(2022) 의 모형이 비슷한 예측치를 제공하는 것을 확인하였다. 보루타 알고리즘에 대한 자세한 설명은 〈부록 1〉에 추가하였다.

다음 장에서 실행한 모든 예측 및 예측 평가 등은 R 프로그램을 이용하였다. 보루타 알고리즘을 포함하여 모두 공개된 패키지(Package) 들을 사용하였고, 특히 Medeiros et al. (2021) 의 코드(<https://github.com/gabrielrvsc/>) 와 Kim and Han(2022) 의 코드(<https://github.com/bryank2242/Forecasting-VIX>) 를 주로 참고하였다. 논문에서 사용한 머신러닝 방법들의 예측에 채택한 하이퍼 파라미터(Hyper

11) 랜덤 포레스트의 모형으로 변수 중요도(Variable Importance)를 측정하는 방법은 다양한데, Speiser et al. (2019) 는 13개의 다양한 방법들을 여러 데이터 세트에 적용하여 비교하였고 그 중 보루타 알고리즘의 성능이 좋다는 것을 밝혔다.

12) Kim and Han(2022) 은 보루타 알고리즘이 중요한 변수로 구분한 변수들의 Z-score 평균을 내림차순으로 정렬하고 이를 변수 중요도 순위로 취급하였다. 우선 변수 중요도가 가장 높은 변수와 두 번째로 높은 변수들만을 이용하여 랜덤 포레스트를 적용하고, In-sample OOB(out-of-bag) MSE(Mean Squared Error)를 계산하였다. 변수 중요도 순위에 따라 그 다음 순위의 변수를 하나씩 추가하면서 각각의 경우 In-sample OOB MSE를 계산하여 이 중 MSE가 가장 낮은 변수의 수를 최적 변수의 수로 정하였다. 최종적으로 변수 중요도가 높은 변수들을 최적 변수의 수까지만 사용하여 랜덤 포레스트에 적용하고 예측치를 도출하는데, 이때 선택되지 않은 나머지 설명변수들은 모두 제거하고 랜덤 포레스트를 실행하였다.

Parameter)들은 대부분 Medeiros et al. (2021)과 같게 설정하였으나, Ridge regression, LASSO, Adaptive LASSO, Elastic Net, Adaptive Elastic Net 등 Shrinkage 관련 모형들의 하이퍼 파라미터 λ 는 교차검증을 통해 결정하였다. Medeiros et al. (2021)은 표본 수가 커서 계산 시간을 줄이기 위해 BIC 기준으로 λ 를 결정하였으나, 우리나라 데이터에 적용한 결과 교차검증을 통한 변수 선택이 BIC보다 적절한 것으로 나타났다.¹³⁾ LASSO 등 Shrinkage 관련 모형의 경우 설명 변수들 사이의 다양한 교차항과 설명변수의 제곱 및 세제곱 등을 설명변수로 포함하여 비선형 관계를 고려할 수 있는데, 설명변수들의 제곱을 추가하여 예측력을 비교한 결과 비슷한 수준의 예측오차를 나타내는 것을 확인하였다. LSTM 모형은 학습 횟수(Epoch)를 100으로 그리고 한 번에 처리할 데이터의 크기(Batch Size)를 25로 설정하였는데, Batch Size를 32로 하거나 은닉층을 두 개로 바꾸더라도 유사한 예측오차가 나타남을 확인하였다.

다음 장에서는 구간이동 방법을 적용하여 총 13개 모형의 예측치를 생성하고, 예측오차로 RMSE(Root Mean Squared Error)와 MAE(Mean Absolute Error)를 비교하였다. III장에서 설명한 예측 절차에 따라 각 모형의 예측치를 생성하고, 예측오차를 비교한 다음 예측오차 간의 차이가 통계적으로 유의한 지를 검정하기 위해 Giacomini and White(2006)와 Hansen et al. (2011)의 검정 방법을 적용하였다.

V. 예측 및 분석 결과

〈Table 3〉과 〈Table 4〉는 각 모형의 예측 대상기간별 물가상승률 예측치에 대한 예측오차인 RMSE와 MAE를 제시한다. 모형별 RMSE를 비교하면, 12개 모든 예측 대상기간에서 BS_RF 모형의 예측오차가 가장 낮은 것으로 나타났다. 모형별 MAE를 비교하면, 총 10개 예측 대상기간에서 BS_RF 모형의 예측오차가 가장 낮은 것으로 나타났다. 나머지 2개 예측 대상기간(9개월과 11개월)에서는 XGBoost 모형의 예측오차가 가장 낮았다. 이를 통해 보루타 알고리즘을 통해 중요한 변수로 선택된 변수들만을 사용하여 랜덤 포레스트에 적용하는 방법의 예측력이 전반적으로 좋은 것을 알 수 있다.

13) BIC 기준으로 하이퍼 파라미터를 결정하더라도 제 V 장에서의 예측력 평가 결과가 질적으로 달라지지 않았다.

〈Table 3〉 RMSEs for Korean Inflation Forecasting

Model	1Month	2Month	3Month	4Month	5Month	6Month	7Month	8Month	9Month	10Month	11Month	12Month
Random Walk	0.0043	0.0054	0.0055	0.0045	0.0046	0.0045	0.0043	0.0043	0.0054	0.0057	0.0046	0.0036
AR	0.0034	0.0036	0.0035	0.0034	0.0034	0.0035	0.0035	0.0034	0.0032	0.0032	0.0032	0.0032
Ridge	0.0033	0.0035	0.0035	0.0035	0.0036	0.0036	0.0035	0.0036	0.0035	0.0035	0.0036	0.0036
Lasso	0.0033	0.0033	0.0034	0.0034	0.0036	0.0035	0.0034	0.0035	0.0034	0.0034	0.0035	0.0035
Ada-Lasso	0.0033	0.0033	0.0034	0.0035	0.0036	0.0036	0.0034	0.0034	0.0034	0.0033	0.0034	0.0035
Elastic Net	0.0032	0.0033	0.0035	0.0035	0.0035	0.0035	0.0035	0.0035	0.0034	0.0034	0.0035	0.0035
Ada-Elastic Net	0.0032	0.0033	0.0035	0.0035	0.0035	0.0035	0.0034	0.0034	0.0034	0.0033	0.0035	0.0035
CSR	0.0033	0.0034	0.0033	0.0033	0.0033	0.0032	0.0032	0.0032	0.0031	0.0030	0.0030	0.0030
RF	0.0031	0.0033	0.0033	0.0034	0.0034	0.0033	0.0032	0.0032	0.0032	0.0031	0.0031	0.0031
XGBoost	0.0032	0.0034	0.0034	0.0036	0.0035	0.0032	0.0033	0.0032	0.0031	0.0031	0.0030	0.0031
BS_RF	0.0028	0.0030	0.0029	0.0032	0.0031	0.0031	0.0030	0.0030	0.0031	0.0029	0.0030	0.0029
Neural Network	0.0036	0.0041	0.0037	0.0039	0.0038	0.0036	0.0038	0.0037	0.0038	0.0037	0.0034	0.0036
LSTM	0.0033	0.0036	0.0037	0.0036	0.0034	0.0037	0.0037	0.0034	0.0035	0.0034	0.0034	0.0036

Note: Numbers in bold represents the minimum of forecast errors for each forecast horizon.

〈Table 4〉 MAEs for Korean Inflation Forecasting

Model	1Month	2Month	3Month	4Month	5Month	6Month	7Month	8Month	9Month	10Month	11Month	12Month
Random Walk	0.0035	0.0042	0.0043	0.0036	0.0036	0.0035	0.0035	0.0036	0.0041	0.0045	0.0036	0.0029
AR	0.0028	0.0029	0.0028	0.0028	0.0027	0.0028	0.0028	0.0027	0.0025	0.0025	0.0026	0.0027
Ridge	0.0027	0.0028	0.0028	0.0028	0.0029	0.0029	0.0028	0.0029	0.0028	0.0028	0.0029	0.0029
Lasso	0.0026	0.0027	0.0028	0.0028	0.0029	0.0029	0.0027	0.0028	0.0027	0.0027	0.0027	0.0028
Ada-Lasso	0.0027	0.0026	0.0028	0.0028	0.0029	0.0029	0.0027	0.0028	0.0027	0.0027	0.0028	0.0028
Elastic Net	0.0026	0.0027	0.0029	0.0028	0.0028	0.0028	0.0028	0.0028	0.0027	0.0027	0.0028	0.0028
Ada-Elastic Net	0.0026	0.0027	0.0028	0.0028	0.0029	0.0029	0.0027	0.0027	0.0028	0.0027	0.0028	0.0028
CSR	0.0026	0.0025	0.0026	0.0026	0.0026	0.0026	0.0026	0.0026	0.0025	0.0024	0.0024	0.0024
RF	0.0025	0.0027	0.0027	0.0027	0.0027	0.0026	0.0025	0.0025	0.0025	0.0025	0.0025	0.0025
XGBoost	0.0025	0.0028	0.0028	0.0028	0.0026	0.0026	0.0025	0.0024	0.0025	0.0025	0.0023	0.0024
BS_RF	0.0023	0.0023	0.0023	0.0025	0.0024	0.0024	0.0023	0.0024	0.0025	0.0023	0.0024	0.0022
Neural Network	0.0028	0.0031	0.0029	0.0031	0.0030	0.0028	0.0030	0.0030	0.0031	0.0029	0.0026	0.0029
LSTM	0.0026	0.0028	0.0030	0.0029	0.0028	0.0032	0.0031	0.0027	0.0029	0.0028	0.0028	0.0028

Note: Numbers in bold represents the minimum of forecast errors for each forecast horizon.

다음 단계로는 BS_RF 모형의 예측력이 다른 모형들에 비해 통계적으로 유의하게 우월한지를 Giacomini-White 검정과 MCS 검정을 이용하여 검정하였다. Giacomini and White (2006)는 조건부 예측력 검정 방법을 제시하였는데, 귀무가설은 아래와 같다.

$$H_0 : E[L(y_{t+h}, f_{1,t+h}(\hat{\theta}_1)) - L(y_{t+h}, f_{2,t+h}(\hat{\theta}_2)) \mid G_t] = 0$$

위 식에서 $L(\cdot)$ 은 RMSE나 MAE와 같은 예측오차함수(Forecast Loss Function)이며, y_{t+h} 는 h 기 앞의 실제값, $f_{1,t+h}(\hat{\theta}_1)$ 는 첫 번째 모형의 h 기 앞 예측치, $f_{2,t+h}(\hat{\theta}_2)$ 는 두 번째 모형의 h 기 앞 예측치를 나타낸다. 위 귀무가설은 두 모형의 예측력이 동일하다는 것이다. 각 예측치가 모형 추정치의 함수라는 것은 이 검정 방법이 예측절차, 추정 기간의 크기 등의 차이를 고려한다는 사실을 의미한다. Giacomini and White (2006)의 검정 방법은 기존의 Diebold-Mariano-West 검정 방법보다 우월한데, 이는 Giacomini and White (2006)의 검정 방법이 미포괄모형들(Non-Nested Models)과 포괄모형들(Nested Models)을 모두 다룰 수 있기 때문이다. 반면 Diebold-Mariano-West 검정 방법은 미포괄모형들을 위한 것으로 포괄 모형들에 적용할 때는 통계치의 접근분포가 표준정규분포가 아니라는 것을 Clark and McCracken (2001)이 보였다.

〈Table 5〉는 BS_RF 모형과 기타 모형 간의 Giacomini-White 검정 결과를 제시한다. 구체적으로 무조건부¹⁴⁾ Giacomini-White 검정을 시행하였으며, 예측오차함수는 MAE를 기준으로 하였고, 첫 번째 모형 f_1 은 상대 모형이고 두 번째 모형 f_2 는 BS_RF 모형으로 정하였다. 따라서 〈Table 5〉에서 통계치가 양(+)의 값을 가지는 것은 BS_RF 모형의 예측오차가 테스트하는 상대 모형보다 낮다는 것을 의미하고, 반대로 통계치가 음(-)의 값을 가지는 것은 상대 모형의 예측오차가 낮다는 것을 의미한다. 따라서 XGBoost 모형에 대해서 예측 대상기간이 9개월과 11개월인

14) 만약 G_t 가 시간 t 에 따라 변하는 정보집합(information set)이라면 조건부 검정이며, $G_t = \{\emptyset, \Omega\}$ 이라면 무조건부 검정이다. 구간이동 예측을 사용하면서 무조건부 Giacomini-White 검정을 실시하는 것은 전통적인 Diebold-Mariano 검정과 동등하다는 것을 Giacomini and White (2006)이 보였다.

경우를 제외한 나머지 모든 모형과 예측 대상기간에 대해서는 Giacomini-White 통계치가 모두 양(+)의 값을 가진다.

Giacomini-White 검정의 귀무가설은 두 모형의 예측력이 같다는 것인데, 귀무가설을 기각한다면 예측오차가 낮은 모형이 통계적으로 유의하게 상대 모형보다 예측력이 우수하다는 것을 의미한다. 예를 들어 예측 대상기간이 1개월인 경우 LSTM 모형을 제외한 모든 상대 모형에 대해 통계치가 양(+)의 값을 가지며 귀무가설을 기각하는 것으로 나타났는데, 이는 BS_RF 모형이 LSTM 모형을 제외한 상대 모형들보다 통계적으로 유의하게 예측력이 우수하다는 것을 나타낸다. 반면 Giacomini-White 통계치가 음(-)인 두 경우 모두 귀무가설을 기각할 수 없는 것으로 나타나는데, 이는 (비록 예측오차는 낮더라도) BS_RF 모형보다 통계적으로 유의하게 예측력이 우수한 모형은 존재하지 않는다는 것을 의미한다.

〈Table 6〉은 Hansen et al. (2011)의 MCS 검정 결과를 제시하며, Medeiros et al. (2021)과 같이 유의수준(Confidence Level)은 50%로 설정하였다. Giacomini-White 검정은 두 모형의 예측력을 테스트하는 데 반해, MCS 검정 방법은 모든 모형을 한꺼번에 테스트하는 것으로 예측력이 가장 우수한 모형들이 Model Confidence Set에 포함된다. MCS 검정 방법은 Hansen(2005)의 Superior Predictive Ability(SPA) 검정 방법을 발전시킨 것으로, SPA 검정에서는 벤치마크 모형을 정해야 하지만 MCS 검정에서는 그럴 필요가 없다. 〈Table 6〉에 따르면 모든 예측 대상기간에서 BS_RF 모형이 Model Confidence Set에 포함되어 있으며, 특히 예측 대상기간이 9개월과 11개월을 제외한 모든 경우에 BS_RF 모형만이 Model Confidence Set에 포함되어 예측력이 우수한 모형임을 보인다. BS_RF 모형을 제외하고 보면, CSR 모형과 XGBoost 모형이 예측 대상기간이 9개월과 11개월인 경우에 Model Confidence Set에 포함되어 있다.

BS_RF 모형의 예측력이 우수한 것은 1) Speiser et al. (2019)에서 밝혔듯이 보루타 알고리즘이 중요 변수 선별에 뛰어나다는 점, 2) 선별된 중요 변수들만을 랜덤 포레스트에 사용하여 불필요한 노이즈(Noise)를 제거한 점, 3) 랜덤 포레스트 모형이 변수 간의 비선형 관계를 모형화하는 데 뛰어난 점 등이 결합하였기 때문으로 유추된다. Medeiros et al. (2021)은 랜덤 포레스트 모형이 변수 간의 비선형 관계를 모형화하는 데 뛰어나기 때문에 미국 물가상승률 예측에 뛰어나다는 점을 이미 보인 바 있다. 아래에서는 보루타 알고리즘의 변수 선택과 오라클 성향(Oracle

(Table 5) Giacomini-White Test Results for Korean Inflation Forecasting

Model	1Month	2Month	3Month	4Month	5Month	6Month	7Month	8Month	9Month	10Month	11Month	12Month
Random Walk	4.71***	5.50***	7.12***	3.96***	4.77***	4.43***	5.16***	4.74***	4.78***	5.57***	3.38***	2.42**
AR	3.90***	3.18**	3.62***	2.46**	3.99***	2.19**	3.27***	2.72***	0.57	2.14**	1.56	3.03***
Ridge	3.11***	3.51***	3.98***	3.13***	5.15***	3.35***	3.39***	3.80***	2.97***	3.49***	3.98***	4.88***
Lasso	3.17***	3.68***	4.11***	3.00***	5.02***	3.63***	3.87***	3.65***	2.29**	3.06	3.36***	4.53***
Ada-Lasso	4.24***	3.13***	4.82***	3.41***	5.18***	3.80***	3.35***	3.29***	1.95*	2.68***	3.22***	4.71***
Elastic Net	3.48***	3.42***	5.38***	3.79***	4.83***	3.62***	4.00***	3.60***	2.35**	3.08	3.65***	4.39***
Ada-Elastic Net	3.61***	3.78***	4.34***	3.54***	4.93***	3.67***	3.01***	3.20***	2.53**	2.91***	3.59***	4.42***
CSR	2.19**	1.48	2.89***	1.39	3.62***	2.52**	1.98**	1.89*	0.09	0.56	0.17	1.87*
RF	3.64***	3.39***	4.80***	2.65***	4.72***	2.39**	2.30**	1.52	0.66	1.75*	1.31	3.63***
XGBoost	2.10**	2.83***	4.19***	2.70***	2.33**	0.87	1.15	0.82	-0.02	1.25	-0.72	1.98**
Neural Network	2.82***	4.28***	3.08***	3.04***	4.99***	2.06**	4.05***	2.96***	3.93***	4.63***	0.95	3.07***
LSTM	1.51	2.46**	2.97***	2.42**	1.92*	4.33***	4.92***	1.95*	2.70***	3.40***	2.41**	2.71***

Notes: The table provides unconditional Giacomini-White test statistics between the BS_RF model and each competing model for absolute errors. A positive statistic implies that the BS_RF model has a lower forecast loss than a competing model. *, **, *** indicate rejection of equal predictive ability at the 10%, 5%, 1% significance level, respectively.

〈Table 6〉 MCS(Model Confidence Set) Test Results for Korean Inflation Forecasting

Model	1Month	2Month	3Month	4Month	5Month	6Month	7Month	8Month	9Month	10Month	11Month	12Month
Random Walk												
AR									O			
Ridge												
Lasso												
Ada-Lasso												
Elastic Net												
Ada-Elastic Net												
CSR									O		O	
RF											O	
XGBoost									O		O	
BS_RF	O	O	O	O	O	O	O	O	O	O	O	O
Neural Network												
LSTM												

Note: The table provides model confidence set test results with significance level 50%. If a model is included in model confidence set, it means that the model exhibits best predictive ability.

property)을 충족하는 Adaptive LASSO의 변수 선택을 비교하였다. Zou (2006)은 LASSO는 오라클 성향을 충족하지 못하나 Adaptive LASSO는 오라클 성향을 충족함을 보였다.

〈Table 7〉은 BS_RF 모형에 사용된 변수들을 제시하는데, 이는 각 예측 대상기간별로 보루타 알고리즘이 중요한 변수들로 분류한 설명변수들을 의미한다. 〈Table 7〉의 변수들이 우리나라 물가상승률 예측에 중요한 요인임을 의미하는데, 예측 대상기간별로 적게는 8개 많게는 21개의 변수가 선택되었다. 예를 들어 예측 대상기간이 1개월인 경우, 경제활동인구(Act_pop_total), 취업자 수(Employed_total), CPI 식료품 및 비주류음료(CPI_F), 수입물가지수 원유(IPI_C), 국제유가(Oilprice), 기업경기실사지수 업황전망(BSI_BC), 미국 10년 만기 국채금리(USB_10), 국내공급물가지수 최종재(SPI_F), 3개월 CD 금리(CD), 소비자물가지수(CPI), 변수들로부터 추출한 주성분요소(factor1, factor 2) 등이 선택되었다. CPI 식료품 및 비주류음료(CPI_F), 수입물가지수 원유(IPI_C), CPI 보전(CPI_HE), CPI 교육(CPI_E), 국내공급물가지수 최종재(SPI_F) 등 물가지수 증가율의 과거값들이 포함된 것은 시계열의 의존성을 고려할 때 놀랍지 않은 것으로 보인다. 다만 소비자물가상승률 예측에 원유 가격뿐만 아니라 식료품 및 비주류음료, 보전, 교육 등의 세부 항목들이 중요하다는 점을 주목할 만하다.

물가지수 관련 변수들을 제외한 나머지 변수들을 살펴볼 때 가장 큰 특징은 노동시장 관련 변수가 모든 예측 대상기간에 포함된 것이다. 경제활동인구와 취업자 수는 〈Table 7〉의 12개 모든 예측 대상기간의 경우에 선택되었는데, 경제활동인구는 예측 대상기간이 1개월, 4~6개월, 10~12개월인 경우 변수 중요도 순위가 1번째였으며 취업자 수는 예측 대상기간이 7~9개월인 경우 변수 중요도 순위가 1번째였다. 그리고 6개월 예측의 경우 취업자 수의 과거값들이 변수 중요도 순위 2번째, 3번째, 4번째, 7번째를 차지하고 있다. 또한 구직기간 3개월 미만 실업자 수(Unemployed_3mts)도 1개월, 5~6개월, 8개월의 경우 포함되었다. 이는 소비자물가상승률 예측에 노동시장 상황이 커다란 영향을 미친다는 것을 의미한다. 고용과 물가상승률의 관계는 필립스 곡선 등을 통해 오래전부터 연구된 주제인데, 고용 측면에서 대안적 지표를 모색하여 필립스 곡선의 물가상승률 예측력을 높이는 사례로 박강우·이상우(2014)를 들 수 있다.

〈Table 7〉 List of Selected Variables for BS_RF Model

	1Month	2Month	3Month	4Month	5Month	6Month
1	Act_pop_total (t-1)	CPI_F (t-2)	CPI_F (t-1)	Act_pop_total (t-4)	Act_pop_total (t-3)	Act_pop_total (t-2)
2	Employed_total (t-1)	CPI_F (t-1)	BSL_BC (t-1)	Employed_total (t-4)	Employed_total (t-3)	Employed_total (t-2)
3	CPI_F (t-3)	BSL_BC (t-3)	BSL_BC (t-2)	BSL_BC (t-1)	Employed_total (t-4)	Employed_total (t-3)
4	IPI_C (t-1)	BSL_BC (t-2)	Employed_total (t-4)	CPI_HE (t-2)	CPI_HE (t-1)	Employed_total (t-4)
5	Oilprice (t-1)	BSL_BC (t-1)	CPI (t-1)	CPI_F (t-2)	CPI_F (t-1)	Unemployed_3 mts (t-4)
6	CPI_F (t-2)	CPI (t-2)	factor2 (t-1)	Employed_total (t-1)	CPI_E (t-4)	CPI_HE (t-3)
7	BSL_BC (t-4)	CPI_HE (t-4)	CPI_HE (t-3)	factor1 (t-1)	Employed_total (t-2)	Employed_total (t-1)
8	Act_pop_total (t-4)	CD (t-1)	factor1 (t-1)	BSL_SG (t-1)	factor1 (t-4)	factor1 (t-4)
9	BSL_BC (t-3)	Employed_total (t-1)	factor1 (t-2)		BSL_SG (t-1)	CPI_E (t-3)
10	Employed_total (t-4)	factor1 (t-2)	M2 (t-1)		PI_PA (t-3)	BSL_SG (t-4)
11	USB_10 (t-1)	factor2 (t-2)	SPI_F (t-1)		HousePrice (t-2)	BSL_BC (t-4)
12	BSL_BC (t-2)	Act_pop_total (t-1)	Employed_total (t-2)		HousePrice_S (t-2)	CPI_HE (t-4)
13	SPI_F (t-1)	factor1 (t-3)	Act_pop_total (t-1)		Unemployed_3 mts (t-1)	Act_pop_total (t-3)
14	factor2 (t-3)	SPI_F (t-2)	Employed_total (t-1)		PI_PA (t-2)	PI_PA (t-2)
15	CPI (t-4)	Act_pop_total (t-3)			SPI_F (t-2)	CPI_E (t-4)
16	factor1 (t-3)	CPI_E (t-1)				
17	Unemployed_3 mts (t-1)	Employed_total (t-3)				
18	SPI_F (t-3)	Act_pop_total (t-2)				
19	factor1 (t-4)	CPI (t-1)				
20	CD (t-2)	CP (t-1)				
21	CPI (t-3)					

	7Month	8Month	9Month	10Month	11Month	12Month
1	Employed_total (t-4)	Employed_total (t-3)	Employed_total (t-2)	Act_pop_total (t-4)	Act_pop_total (t-3)	Act_pop_total (t-2)
2	Act_pop_total (t-4)	Act_pop_total (t-3)	CPI_F (t-4)	Employed_total (t-1)	CPI_F (t-2)	CPI_F (t-1)
3	CPI_HE (t-4)	CPI_HE (t-3)	Act_pop_total (t-2)	Employed_total (t-4)	Employed_total (t-3)	Employed_total (t-2)
4	Employed_total (t-3)	Employed_total (t-1)	CPI (t-4)	CPI_F (t-4)	CPI (t-2)	CPI (t-1)
5	Employed_total (t-2)	Employed_total (t-2)	BSI_FS (t-2)	Act_pop_total (t-1)	HousePrice_S (t-2)	HousePrice_S (t-1)
6	CPI_HE (t-2)	BSI_FS (t-3)	CPI_HE (t-2)	BSI_FS (t-1)	Employed_total (t-2)	Employed_total (t-1)
7	Act_pop_total (t-1)	CPI_HE (t-1)	Employed_total (t-4)	CPI (t-3)	Act_pop_total (t-1)	Act_pop_total (t-1)
8	BSI_FS (t-4)	factor1 (t-2)	HousePrice_S (t-4)	HousePrice_S (t-3)	Act_pop_total (t-2)	SPI_F (t-4)
9		BSI_BC (t-2)	Employed_total (t-1)	CPI_HE (t-1)	Employed_total (t-1)	BSI_FS (t-1)
10		Act_pop_total (t-4)	Act_pop_total (t-4)	Employed_total (t-3)	Unemployed_3 mts (t-2)	
11		Act_pop_total (t-2)	Act_pop_total (t-3)	factor1 (t-1)		
12		CPI_F (t-3)	factor1 (t-1)			
13		CPI_E (t-4)	BSI_BC (t-1)			
14		factor1 (t-3)	BSI_SG (t-1)			
15		Unemployed_3 mts (t-2)	Employed_total (t-3)			

Notes: The table provides explanatory variables used for the BS_RF model. For each forecast horizon, we select variables that are confirmed to be important by using the Boruta algorithm. Next, we use the selected variables in a random forest model for each forecast horizon, which is the procedure for the BS_RF model. See Table 1 for code and description of each variable. Factor1-factor4 are principal component factors extracted from 93 variables.

기업경기실사지수도 많은 예측 대상기간에 걸쳐 선택되었는데, 특히 예측 대상기간이 1~3개월인 경우는 기업경기실사지수 업황전망(BSI_BC) 만이 선택되었고, 기업경기실사지수 매출증가율전망(BSI_SG)이 5~6개월에 그리고 기업경기실사지수

자금사정전망(BSL_FS)이 7~10개월과 12개월에 선택되었다. 또한 예측 대상기간이 5개월인 경우 주택매매가격지수 전국(HousePrice)과 수도권(HousePrice_S)이 포함되었는데, 9~12개월에는 주택매매가격지수 수도권(HousePrice_S)만이 포함된 것이 특징이다.

〈Table 8〉은 각 예측 대상기간별로 Adaptive LASSO 모형에 의해 선택된 설명변수들을 제시하는데, 이를 보루타 알고리즘이 중요한 변수들로 분류한 설명변수들과 비교하기 위해 추가하였다. 〈Table 8〉을 살펴보면 전반적으로 선택된 변수들의 수

〈Table 8〉 List of Selected Variables by Adaptive LASSO Model

	1Month	2Month	3Month	4Month	5Month	6Month
1	CPI (t-1)	CPI_F (t-1)	CPI_F (t-1)	CPI_HE (t-2)	CPI_HE (t-1)	CPI_HE (t-3)
2	CPI_E (t-1)	CPI_F (t-2)	CPI_HE (t-3)	Employed_total (t-4)	CPI_HE (t-4)	CPI_HE (t-4)
3	CPI_F (t-2)	CPI_HE (t-4)	CPI_M (t-2)	BSL_BC (t-1)	Employed_total (t-3)	Employed_total (t-2)
4	CPI_F (t-3)	M2 (t-2)	IPI_A (t-3)	factor2 (t-1)		Unemployed_3 mts (t-4)
5	SPI_F (t-1)	BSL_BC (t-3)	M2 (t-1)			
6	Oilprice (t-1)	factor2 (t-3)	factor2 (t-2)			
7	M2 (t-3)					
8	Act_pop_total (t-1)					
9	Employed_total (t-1)					
10	BSL_SG (t-4)					
11	BSL_BC (t-4)					
12	USB_10 (t-1)					
13	factor1 (t-4)					

	7Month	8Month	9Month	10Month	11Month	12Month
1	CPI_HE (t-2)	CPI_HE (t-1)	CPI (t-4)	CPI (t-3)	CPI (t-2)	CPI (t-1)
2	Employed_total (t-4)	CPI_HE (t-3)	Employed_total (t-2)	Employed_total (t-1)		
3		Employed_total (t-3)				

Notes: The table provides explanatory variables selected by the adaptive LASSO model. See Table 1 for code and description of each variable. Factor1-factor4 are principal component factors extracted from 93 variables.

가 적은 편인데, 특히 예측 대상기간이 길어질수록 선택된 변수들의 수가 줄어드는 것으로 나타났다. 예를 들어, 예측 대상기간이 1개월인 경우는 13개 변수가 선택되었는데, 예측 대상기간이 4개월 이후인 경우는 선택된 변수들의 수가 4개 이하로 줄어들고, 특히 예측 대상기간이 9~10월인 경우는 2개의 변수만이 그리고 11~12개월인 경우는 1개 변수만이 선택되었다. 이는 보루타 알고리즘이 중요한 변수로 분류한 변수들의 수가 더 많은 것과 대조적인데, 이러한 차이는 Adaptive LASSO 모형이 선형 모형의 제약을 지니지만 보루타 알고리즘은 랜덤 포레스트 기반이라 변수들의 비선형 관계를 반영하기 때문으로 보인다. <Table 8>의 선택된 변수들을 살펴보면, 보루타 알고리즘이 중요한 변수로 분류한 변수들과 겹치는 경우도 있으며 물가지수들의 과거값, 고용시장 관련 변수, 기업경기실사지수, M2 등이 선택된 것으로 나타났다. 한 가지 주의할 사항은 <Table 7>의 변수들은 중요도 순위에 따라 정렬되었으나 <Table 8>의 변수들은 중요도 순위와 상관없이 선택된 변수들이라는 점이다.

LASSO 모형에 의해 선택된 설명변수들도 살펴보았는데, 이 경우 Adaptive LASSO 모형보다 선택되는 변수의 수가 다소 많은 것으로 나타났다. 이는 예상과 부합하는 것으로 Adaptive LASSO 모형이 초기 추정치를 가중치로 사용하기 때문이다. 다만 LASSO 모형을 사용하더라도 예측 대상기간이 7개월부터 12개월인 경우 선택되는 변수의 수는 서너 개에 불과한 것으로 나타났다.

IV. 결론 및 시사점

최근 우리나라를 포함하여 전 세계적으로 물가상승이 가파르고 이에 대한 통화정책의 대응이 커다란 관심을 받고 있다. 물가상승률 예측은 가계/기업의 경제활동 및 중앙은행 등 정책 당국의 정책 결정에 커다란 영향을 미치며, 따라서 이와 관련한 연구가 활발히 진행되어 왔다. 본고는 물가상승률 예측과 관련하여 다음과 같은 학술적 기여를 제공한다. 첫째, 우리나라와 미국의 물가상승률 예측 관련 데이터를 이용해서 살펴본 결과, Block 부트스트랩을 이용하는 것이 통상적인 독립 부트스트랩을 이용하는 것에 비해 랜덤 포레스트 모형의 예측력을 통계적으로 유의한 수준으로 개선하지는 않는 것을 보였다. 본고는 실제 정상시계열 예측의 사례에서 Stationary Bootstrap, Moving Block Bootstrap, 독립 부트스트랩 등 부트스트랩

방법에 따른 랜덤 포레스트 모형의 예측력을 비교한 최초의 연구라는 점에서 의의가 크다. 그러나 이 결과를 정상시계열 예측의 일반적 상황에도 적용된다고 주장하는 것은 무리이며, 이에 대한 엄밀한 이론적 연구는 추후 과제로 남긴다.

둘째, 본 연구는 FRED-MD와 기존 연구를 참고하여 총 93개의 방대한 국내외 거시경제/금융 데이터를 수집/활용하였고, XGBoost 모형이나 LSTM 모형 등 최근 실무에서 활발히 이용되는 머신러닝 기법들을 포함하여 예측력을 비교하였다. 셋째, 본 논문은 Kim and Han (2022)의 모형을 간소화하여 보루타 알고리즘과 랜덤 포레스트를 활용하는 모형을 제안하고, 이 모형이 우리나라 물가상승률 예측에 여타 모형들보다 우월한 예측력을 지닌다는 점을 보였다. 추가로 이 모형의 예측력이 대다수 경쟁 모형들에 비해 통계적으로 유의한 수준으로 우월하다는 것을 Giacomini and White 검정과 MCS 검정을 통해 보였다. 이러한 결과는 보루타 알고리즘이 변수 중요도의 순위를 도출하는 데 우수하고 랜덤 포레스트 모형이 변수 간의 비선형 관계를 모형화하는 데 뛰어난 점 등이 결합하여 있기 때문으로 유추할 수 있다. 본 논문의 결과는 방대한 데이터를 머신러닝 기법들에 적용할 때 불필요한 노이즈를 제거하고 예측하고자 하는 종속변수와 관련된 설명변수만을 선택/활용할 필요가 있다는 의견을 뒷받침하는 사례가 될 수 있다.

마지막으로 본 논문은 우리나라 물가상승률 예측을 위해 중요한 변수들의 순위를 보루타 알고리즘을 통해 도출하였고, 이를 변수 선택에서 널리 사용되는 Adaptive LASSO 모형을 통해 선택되는 변수들과 비교하였다. Adaptive LASSO 모형은 예측 대상기간이 길어질수록 선택된 변수의 수가 크게 줄어드는 한계가 있는 것으로 나타났다. 또한 LASSO 타입의 모형을 통한 변수 선택 방법은 변수 중요도의 순위까지는 보일 수 없는 반면 보루타 알고리즘은 변수 중요도 순위를 제공하므로 활용도가 더 높다고 볼 수 있다. 보루타 알고리즘 결과 중 가장 커다란 특징은 경제활동인구 및 취업자 수의 증가율이 모든 예측 대상기간에 포함되고 변수 중요도의 높은 순위를 차지한 점이다. 본 논문의 예측 및 분석 결과는 우리나라 물가상승률 예측을 위해 경제활동인구, 취업자 수, (구직기간 3개월 미만) 실업자 수 등을 중점으로 한 고용시장의 상황, 기업경기실사지수, 주택가격 등을 면밀히 살펴볼 필요가 있다는 정책적 시사점도 제공한다.

다만 본 논문은 주택매매가격지수를 2004년 8월부터 이용할 수 있으므로 그 이전 데이터를 활용하지 못하였는데, 더 확장된 시계열이 이용 가능한 주택가격 관련 변

수로 주택매매가격지수를 대체할 수 있다면 구간이동 예측시 추정 기간의 크기를 상당폭 늘릴 수 있을 것이다. 또한 본 논문은 Medeiros et al. (2021)과 같이 전월 대비 물가상승률의 예측을 고려하고 있는데, 전년동월대비 물가상승률의 예측에서도 비슷한 결과가 도출되는지를 살펴보는 것도 흥미로울 것이다. 이러한 사항들에 관한 연구는 추후 연구과제로 남긴다.

■ 참 고 문 헌

1. 강규호 · 김정성 · 신세림, “우리나라 소비자물가상승률 예측,” 『경제분석』, 제27권 제4호, 한국은행 경제연구원, 2021, pp. 1-42.
(Translated in English) Kang, Kyu Ho, Jungsung Kim, and Serim Shin, “Forecasting Korean CPI Inflation,” *Economic Analysis*, Vol. 27, No. 4, 2021, pp. 1-42.
2. 김종화 · 이중식, “인플레이션 전망에 있어서 정보변수의 역할,” 『경제분석』, 제11권 제3호, 한국은행 경제연구원, 2005, pp. 31-58.
(Translated in English) Kim, Jong Hwa, and Joong Shik Lee, “Use of Information Variables in Inflation Forecasting,” *Economic Analysis*, Vol. 11, No. 3, 2005, pp. 31-58.
3. 김현학, “예측조합 및 밀도함수에 의한 소비자물가 상승률 전망,” 『BOK 경제연구』, 제 2015-11, 한국은행 경제연구원, 2015, pp. 1-33.
(Translated in English) Kim, Hyun Hak, “Forecasting CPI Inflation using Combination of Point Forecast and Density Forecast,” *Economic Analysis*, Vol. 2015-11, 2015, pp. 1-33.
4. 박강우 · 이상우, “고용지표의 유용성: 인플레이션 예측력을 중심으로,” 『경제분석』, 제20권 제 1호, 한국은행 경제연구원, 2014, pp. 1-32.
(Translated in English) Park, Kangwoo, and Sangwoo Lee, “In Search of Better Labor-market Indicators: Assessment of Inflation Predictability,” *Economic Analysis*, Vol. 20, No. 1, 2014, pp. 1-32.
5. 이진희 · 김덕파, “한국 경제변수에 대한 자기회귀 및 벡터자기회귀 모형의 예측성과 비교,” 『경제분석』, 제20권 제4호, 한국은행 경제연구원, 2014, pp. 114-150.
(Translated in English) Lee, Jinhee, and Dukpa Kim, “Forecasting Korean Macroeconomic Variables with Autoregressions and Vector Autoregressions,” *Economic Analysis*, Vol. 20, No. 4, 2014, pp. 114-150.
6. 장나원 · 한희준, “머신러닝 방법을 활용한 한국의 수출입 증가율 예측 및 분석,” 『국제경제연구』, 제28권 제4호, 국제경제학회, 2022, pp. 59-89.

- (Translated in English) Jang, Nawon, and Heejoon Han, "Forecasting Exports and Imports of South Korea using Machine Learning Methods," *Kukje Kyungje Yongu*, Vol. 28, No. 4, 2022, pp.59-89.
7. 최 인 · 황승진, "한국의 인플레이션 예측," 『응용경제』, 제14권 제3호, 2012, pp.133-168.
(Translated in English) Choi, In, and Seung Jin Hwang, "Forecasting Korean Inflation," *Applied Economics*, Vol. 14, No. 3, 2012, pp.133-168.
 8. Bai, J., and S. Ng, "Forecasting Economic Time Series using Targeted Predictors," *Journal of Econometrics*, Vol. 146, 2009, pp.304-317.
 9. Breiman, L., "Random Forests," *Machine Learning*, Vol. 45, 2001, pp.5-32.
 10. Clark, T., and M. McCracken, "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics*, Vol. 105, 2001, pp.85-110.
 11. Giacomini, R., and H. White, "Tests of Conditional Predictive Ability," *Econometrica*, Vol. 74, No. 6, 2006, pp.1545-1578.
 12. Faust, J., and J. Wright, "Forecasting Inflation," in *Handbook of Economic Forecasting* (Vol. 2A), eds. G. Elliott and A. Timmermann, Amsterdam: Elsevier, 2013.
 13. Goehry, B., "Random Forests for Time-dependent Processes," *ESAIM: Probability and Statistics*, Vol. 24, 2020, pp.801-826
 14. Hansen, P. R., "A Test of Superior Predictive Aability," *Journal of Business & Economic Statistics*, Vol. 23, No. 4, 2005, pp.365-380.
 15. Hansen, P. R., A. Lunde, and J. M. Nason, "The Model Confidence Set," *Econometrica*, Vol. 79, No. 2, 2011, pp.453-497.
 16. Kim, B. Y., and H. Han, "Multi-step-ahead Forecasting of the CBOE Volatility Index in a Data-rich Environment: Application of Random Forest with Boruta Algorithm," *The Korean Economic Review*, Vol. 38, No. 3, 2022, pp.541-569.
 17. Kursa, M. B., and W. R. Rudnicki, "Feature Selection with the Boruta Package," *Journal of Statistical Software*, Vol. 36, 2010, pp.1-13.
 18. Künsch, H. R., "The Jackknife and the Bootstrap for General Stationary Observations," *Annals of Statistics*, Vol. 17, 1989, pp.1217-1241.
 19. Lahiri, S. N., "Theoretical Comparisons of Block Bootstrap Methods," *Annals of Statistics*, Vol. 27, 1999, pp.386-404.
 20. Masini, R. P., M. C. Medeiros, and E. F. Mendes, "Machine Learning Advances for Time Series Forecasting," *Journal of Economic Surveys*, Vol. 37, No. 1, 2023, pp.76-111.
 21. Medeiros, M. C., G. F. Vasconcelos, A. Veiga, and E. Zilberman, "Forecasting Inflation in a Data-rich Environment: The Benefits of Machine Learning Methods," *Journal of Business & Economic Statistics*, Vol. 39, No. 1, 2021, pp.98-119.
 22. Kursa, M. B., "Boruta - A System for Feature Selection," *Fundamenta Informaticae*, Vol. 101, 2010, pp.271-285.
 23. Patton, A., D. N. Politis, and H. White, Correction to "Automatic Block-length Selection for Dependent Bootstrap," by D. Politis and H. White. *Econometric Rev.* 28, 2009, pp.372-375.

24. Politis, D. N., and J. P. Romano, "The Stationary Bootstrap," *Journal of the American Statistical Association*, Vol. 89, 1994, pp.1303-1313.
25. Politis, D. N., and H. White, "Automatic Block-length Selection for Dependent Bootstrap," *Econometric Rev.* 23, 2004, pp.53-70.
26. Speiser, J. L., M. E. Miller, J. Tooze, and E. Ip, "A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling," *Expert Systems with Applications*, Vol. 134, 2019, pp.93-101.
27. Stock, J., and M. Watson, "Modeling Inflation After the Crisis," Technical Report, National Bureau of Economic Research, 2010.
28. Zou, H. White, "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, Vol. 101, No. 476, 2006, pp.1418-1429.

〈부록 1〉

1. 보루타 알고리즘

Kursa and Rudnicki (2010)가 소개한 보루타 알고리즘은 랜덤 포레스트를 기반으로 중요 변수들을 안정적으로 선택하는 기법으로, 각 변수에 대해 순열 중요도 (Permutation Importance Measure)를 그 표준편차로 나누어 계산하는 Z-score¹⁵⁾를 변수 선택의 척도로 사용한다. 보루타 알고리즘은 다음과 같은 단계로 실행된다.

- 1) 모든 변수의 복사본을 추가하여 데이터 세트(Dataset)를 확장한다.
- 2) 추가된 변수들을 섞어서(Shuffle) 종속변수와의 상관관계를 제거한다(이를 Shadow Attributes라 칭함).
- 3) 확장된 데이터 세트를 이용하여 랜덤 포레스트를 실행하고 각 변수들의 Z-score를 수집한다.
- 4) Shadow Attributes의 Z-score 중 최댓값(Maximum Z-score among Shadow Attributes, 이하 ‘MZSA’)을 구한 후, MZSA보다 Z-score가 큰 모든 변수에 히트를 할당(assign a hit)한다.
- 5) 4에 해당하지 않는 변수의 경우, MSZA에 대한 양측 동등성 검정(Two-Sided Equality Test)을 수행한다.
- 6) MZSA보다 Z-score가 유의하게 작은 변수를 중요하지 않은 변수로 간주하고 데이터 세트에서 모두 제거한다.
- 7) MZSA보다 Z-score가 유의하게 큰 변수를 중요한 변수로 간주한다.
- 8) 모든 Shadow Attributes를 제거한다.
- 9) 모든 변수에 중요도가 할당되거나 또는 랜덤 포레스트의 최대 실행 횟수에 도달할 때까지 위 절차를 반복한다.

15) Out-of-bag 샘플은 랜덤 포레스트 실행시 부트스트랩 샘플링에서 제외된 샘플을 의미한다. 랜덤 포레스트에서 각 변수의 중요도는 Out-of-bag 샘플에서 계산한 OOB MSE (Out-of-Bag Mean Squared Error)와 해당 변수를 무작위 치환(Random Permutation)하고 계산한 OOB MSE의 차이를 이용하여 측정하고, 각 변수의 Z-score는 이 차이의 평균(Kim and Han (2022, equation (5) 참조)을 표준편차로 나누어 계산한다.

데이터 세트에 포함된 변수의 수가 많으면 보루타 알고리즘을 실행할 때 모든 변수가 중요하거나 중요하지 않은 두 종류로 분명히 분류되기보다는 중요한 변수(Confirmed Important), 중요하지 않은 변수(Confirmed Unimportant), 결정하기 힘든 변수(Tentative) 등 세 가지로 분류된다. 보루타 알고리즘의 R 패키지는 랜덤 포레스트 최대 실행 횟수의 기본값(Default)을 100으로 정하고 있는데, 이 경우 랜덤 포레스트를 100번 실행하여 각 변수의 Z-score의 평균, 중위값 등을 구할 수 있다.

2. Shrinkage Methods

Ridge regression, LASSO 등의 Shrinkage Methods는 OLS 추정과 같이 잔차의 제곱의 합을 최소화하는 것이지만, 다음과 같은 제약이 추가되어 있다.

$$\hat{\beta}_{shrinkage} = \operatorname{argmin}_{\beta} \left[\sum_{i=1}^n (y_i - X_i \beta) + p(\beta, \lambda) \right]$$

이 식에서 X 는 m 개의 설명변수를 포함하는 벡터이다. 설명변수들은 모두 표준화(Standardize)하여 적용하였다. $p(\beta, \lambda)$ 는 제약 함수(penalty function)인데, 각 Shrinkage method마다 다르게 정의된다.

Ridge regression은 Hoerl and Kennard(1970)에 의해 소개되었으며 다음과 같은 L_2 제약을 둔다.

$$p(\beta, \lambda) = \lambda \sum_{j=1}^m \beta_j^2$$

λ 는 Tuning-parameter 또는 Hyper-parameter이며 BIC 등의 Information Criteria나 교차검증을 통해 결정한다. 본고는 모든 Shrinkage Methods의 λ 를 교차검증을 통해 정하였으며, BIC를 통해 결정할 때에도 질적으로 유사한 예측치가 생성됨을 확인하였다.

LASSO는 Tibshirani(1996)에 의해 소개되었으며, 다음과 같은 L_1 제약을 둔다.

$$p(\beta, \lambda) = \lambda \sum_{j=1}^m |\beta_j|$$

Adaptive LASSO는 Zou (2006)에 의해 소개되었으며, 제약 함수에 가중치 w_j 가 추가된 다음과 같은 제약을 둔다.

$$p(\beta, \lambda) = \lambda \sum_{j=1}^m w_j |\beta_j|$$

w_j 는 OLS, Ridge regression, LASSO 등을 통해 전 단계에서 구한 초기 추정치를 이용하여 $w_j = |\tilde{\beta}_{initial,j}|^{-1}$ 로 설정하는데, 본고에서는 Medeiros et al. (2021)과 같이 LASSO 추정치를 이용한 $w_j = (|\tilde{\beta}_{LASSO,j}| + 1/\sqrt{n})^{-1}$ 로 정하였다.

Elastic Net은 Zou and Hastie (2005)에 의해 소개되었으며, 다음과 같은 L_1 과 L_2 의 가중평균 제약을 둔다.

$$p(\beta, \lambda) = \lambda \sum_{j=1}^m \{\alpha \beta_j^2 + (1 - \alpha) |\beta_j|\}$$

본고에서는 Medeiros et al. (2021)과 같이 $\alpha = 0.5$ 로 설정하였다. 또한 Adaptive Elastic Net도 고려했는데, Adaptive LASSO에서 사용한 $w_j = (|\tilde{\beta}_{LASSO,j}| + 1/\sqrt{n})^{-1}$ 를 사용하였다.

3. Complete Subset Regression(CSR)

CSR은 Elliott et al. (2013)에 의해 소개된 앙상블 방법으로 총 m 개의 설명변수에서 k 개의 설명변수로 이루어진 모든 가능한 조합을 이용하여 선형 추정 및 예측을 하고 이 예측치들의 평균을 최종 예측치로 삼는 것이다. 다만 설명변수의 수가 너무 많은 경우 추정해야 할 수가 급격히 증가하기 때문에 ($m = 25$ 그리고 $k = 4$ 인 경우 12,650번의 추정이 필요함), Medeiros et al. (2021)은 사전에 각 설명변수를 추정하여 t-통계값을 구하고 이의 절대값 크기가 큰 25개의 변수만을 선별하고

$k = 4$ 를 설정하여 CSR을 실시하였다. 본고에서도 이와 같은 방법을 따랐다.

4. Random Forest 모형

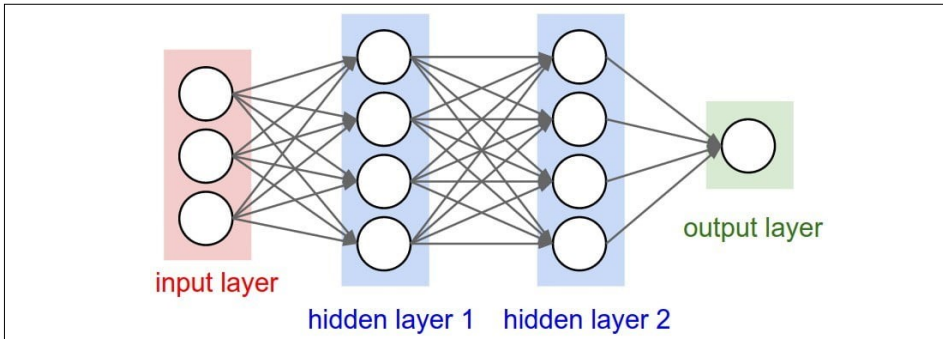
랜덤 포레스트는 Breiman (2001)에 의해 소개되었으며, 결정 트리(Decision Tree)를 기반으로 하는 Bootstrap aggregation (Bagging)을 분산이 작아지도록 개선한 방법이다. 부트스트랩 샘플로 결정 트리를 추정할 때 결정 트리의 모든 분리 (Split)마다 임의로 선택된 설명변수를 고려하도록 하는데, 이는 부트스트랩 샘플에서 만들어지는 결정 트리들을 서로 상관없게 만들고 (Decorrelate) 결과적으로 분산을 감소시키게 된다.

5. XGBoost 모형

XGBoost는 Extreme Gradient Boosting의 약자이다. XGBoost 모형은 결정 트리를 기반으로 하는 Boosting 알고리즘으로 기존 Gradient Tree Boosting 방법에 과적합 방지를 위한 기법이 추가된 방법이다. XGBoost는 병렬 처리로 기존 Gradient Boosting 방법보다 학습 속도가 빠르며, 과적합에 강한 내구성을 지니며, 분류(Classification)와 회귀(Regression) 모두에서 뛰어난 예측 성능을 발휘하는 것으로 알려져 널리 사용되고 있다.

6. 인공신경망(Neural Network) 모형

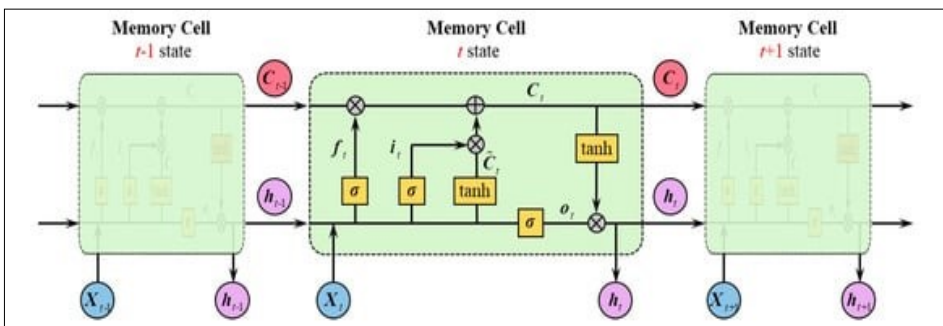
아래 그림은 은닉층(Hidden layer)이 두 개이고 각각이 네 개의 노드(Node)를 가진 인공신경망 모형을 나타낸다.



본고에서 이용한 인공신경망 모형은 세 개의 은닉층과 각 은닉층이 32, 16, 8개의 노드를 지닌 것으로 설정하였으며, 활성화 함수(Activation function)로는 일반적인 Rectifier 함수를 사용하였고 학습 횟수(Epoch)는 100으로 정하였다. 두 개의 은닉층과 각 은닉층이 32개와 16개의 노드를 지닌 경우도 살펴보았으나, 예측오차가 비슷한 것으로 나타났다.

7. LSTM 모형

기존 RNN(Recurrent Neural Network) 모형은 BPTT(BackPropagation Through Time) 방식을 사용하는데, 역전파(Back Propagation)시 Vanishing/Exploding Gradient 문제 때문에 시계열의 장기 의존성(Long-Term Dependence)을 제대로 반영하지 못하는 한계가 있다. LSTM 모형은 이 한계를 극복하기 위해 아래 그림과 같이 장기 기억을 담당하는 Cell State를 도입하여 정보를 장기간에 걸쳐 저장 및 접근할 수 있게 한다.



LSTM의 알고리즘 등 자세한 사항은 Masini et al. (2023)을 참고할 수 있다. LSTM 모형은 분석 시 학습 횟수(Epoch), 한 번에 처리할 데이터의 크기(Batch Size) 등의 하이퍼 파라미터를 조정할 수 있는데, 본고에서는 학습 횟수를 100으로 그리고 한 번에 처리할 데이터의 크기를 25로 설정하였다. 그리고 Batch Size를 32로 한 경우에도 유사한 결과임을 확인하였다.

Random Forest for Stationary Time Series: The Case of Forecasting Inflation in Korea

Heejoon Han*

Abstract

This paper first investigates whether adopting the stationary bootstrap or the moving block bootstrap, instead of the usual independent bootstrap, in the random forest method improves forecasting of stationary time series. It is shown that the block bootstrap procedures adopted in the random forest method do not make any statistically significant improvement in Korean or US inflation forecasting. Secondly, we consider inflation forecasting in Korea using 93 macroeconomic/financial variables and various machine learning methods. The samples are from September 2004 to March 2022. Comparing total 13 models, one model outperforms the rest models for most forecast horizons, which is a simplified method of the model proposed by Kim and Han (2022). The method consists of the following two steps: 1) Select important variables based on the Boruta algorithm, 2) Using only those selected variables, implement the random forest and produce a forecast. The tests by Giacomini and White (2006) and Hansen et al. (2009) show that the model provides significantly better forecasts for most forecast horizons. In particular, the Boruta algorithm selected total economically active population, total employed persons, BSI, house price as important variables for Korean inflation forecasting.

Key Words: random forest, bootstrap, inflation, forecasting, Machine learning

JEL Classification: C32, C50, E31

Received: Feb. 9, 2023. Revised: April 26, 2023. Accepted: August 25, 2023.

* Professor, Department of Economics/Quantitative Applied Economics, Sungkyunkwan University, 25-2, Seonggyungwan-ro, Jongno-gu, Seoul 03063, Korea, Phone: +82-02-760-0428, e-mail: heejoonhan@skku.edu