석사학위 논문 초고

조정효

KOSPI 방향 예측에 대한 랜덤포레스트의 군집화 변수중요도 분석

최근 기계학습을 이용한 주가지수 방향의 예측에 대한 관심이 증가하면서 모형의 성능을 향상시키는 연구가 많이 이루어졌다. 하지만 기계학습 모형을 통해 예측변수를 분석하여 주식시장과 변수 간의 관계를 분석한 연구는 드물다. 본 연구에서는 랜덤포레스트 분류기 모형을 이용해 기술적분석 지표, 투자주체별 수급, 거시경제 변수를 예측변수로 하여 KOSPI 지수 방향을 예측한 후, 학습된 랜덤포레스트 모형의 순열중요도(permutation importance)를 측정해 변수에 대해 분석했다. 이 때 변수 간 상관관계가 존재할 경우 중요도 측정에 편향이 생기므로 변수를 미리 군집화하여 변수군집에 대한 중요도를 측정하는 방법을 사용했다. KOSPI 지수의 1 일, 5 일, 20 일 후의 등락 여부를 각각 목표변수로 하여 예측한 결과 20 일, 5 일, 1 일 순으로 예측성능이 높았으며, 변수중요도를 비교한 결과, 1 일 방향에 대해서는 환율로 대표되는 변수군집이 큰 중요도를 보였으며 5 일 및 20 일 방향에 대해서는 장기적 시장 추세와 시장 변동성으로 대표되는 변수군집이 큰 중요도를 보였다.

목차

I. 서론

주가지수의 방향을 예측하는 것은 시장 매매 전략을 발전시키는데 중요한 역할을 하며(Leung et al., 2000) 주가지수 매매에서 이익을 얻을 수 있게 해준다. 따라서 주가지수의 움직임을 예측하는 것은 금융시장 연구의 중요한 문제로 여겨진다. 또한 주가지수를 예측하는 변수에 대한 분석을 통해 어떤 변수가 주식시장 예측에 많은 영향을 끼치는지를 알 수 있다.

최근 사용 가능한 금융시장의 데이터 양이 급격하게 늘어남에 따라 기계학습을 이용한 주가지수 예측에 대한 관심이 높아졌으며, 기존의 ARIMA(Autoregressive Integrated Moving Average), GARCH(Generalised Autoregressive Conditional Heteroscedasticity)등의 시계열 통계기법과 비교하여 기계학습의 예측력이 대체로 뛰어남을 여러 연구에서 밝혔다(Nti et al., 2019). 따라서 최근 주가지수를 예측하는 기계학습 모형에 대한 연구가 많이 진행되어 왔지만 주로 예측 정확도를 높이는 것에 목적을 두어 모형을 개발하고 예측 변수를 설정하는 연구가 대부분이었다. 반면 예측변수를 해석하는 방법에 대한 연구는 상대적으로 적게 이루어졌다. 기계학습 모형은 기존의 전통적 통계 모형과 달리 결과의 도출 과정을 알 수 없다는 'black box' 문제가 제기되어 왔으나, 기계학습 모형에 대한 발전이 이루어지면서 설명가능한 인공지능에 대한 관심이 증가하고 기계학습 모형의 'black box' 문제를 해결하려는 연구들이 다수 진행되었다. 그 중에서 예측 변수를 해석하는 방법으로 기계학습 모형의 훈련과 예측 과정에 예측변수들이 얼마나 기여했는지를 측정하는 변수 중요도가 있으며, 이는 기계학습 모형을 통한 변수에 대한 해석을 가능하게 한다.

본 연구는 KOSPI 의 증감을 예측 목표변수로 하여 변수 중요도를 측정하고 비교 및 분석한다. 분석하는 예측변수는 총 28 개로 과거 가격의 추세와 변동성, 거래량을 포함하는 기술적 분석 지표와 투자주체별 수급, 그리고 환율, 상품가격을 사용한다. 이를 통해 과거 시장 정보를 이용하여 주가지수를 예측할 수 있는지 살펴보는 것에 더해 시장의 추세와 변동성, 환율 및 상품가격 등의 변수가 예측에 얼만큼의 기여를 하는지를 분석한다. 또한 목표변수인 KOSPI 증감에 대해 각각 1 일, 5 일, 20 일 후로 하여 예측 기간에 따라 비교한다. 분석 자료로는 2012 년 1 월 3 일부터 2021 년 12 월 30 일까지의 일별 자료를 사용하며, 2012 년-2018 년을 훈련 기간으로 2019 년-2021 년을 테스트 기간으로 나누어 훈련 기간에 대해 모형을 학습시키고 테스트 기간에 대해 예측 성능 및 변수 중요도를 구하여 분석한다.

변수 중요도를 도출하는 과정으로 학습된 기계학습 모형을 이용하여 해당 모형의 예측에 각 변수가 얼마나 중요한 영향을 미치는지를 측정한다. 기계학습 모형으로는 주가 및 주가지수 방향 예측에 뛰어난 성능을 보이는 랜덤 포레스트(Random Forest, RF)를 사용하며, 중요도를 측정하는 방법으로는 순열중요도(permutation importance)를 사용한다. RF 의 순열중요도를 이용할 때 예측 변수 간의 강한 상관관계가 존재할 경우 중요도가 하향 편향되는 경향이 있으므로(Breiman, 2001), 변수 간의 상관관계가 변수중요도에 영향을 미치지 않도록 해야 한다. 따라서 변수 간 상관관계를 사전에 차단하고자 변수를 미리 군집화(cluster)한 후 군집에 대해 중요도를 구하는 "군집화 순열 중요도"(De Prado, 2020)를 측정 방법으로 한다.

군집화 과정으로는 선형적 상관관계를 고려한 계층적 군집화를 진행하며 군집화 결과에 따른 각 변수 군집을 특징에 따라 해석하고 중요도를 구한다.

본 논문은 서로 다른 기간의 주가지수 방향에 대한 변수 중요도를 비교하고, 이를 통해 과거 시장 정보 및 거시경제 변수가 향후 주가 지수의 움직임을 예측할 수 있는지, 그리고 어떤 변수가 예측에 기여를 많이 하며, 예측 기간에 따라 중요도가 어떻게 달라지는지를 알아보는데 목적을 둔다. 또한 과거 시장 추세와 변동성 등 공공 시장 자료로 구성된 변수의 중요도가 높은 것을 확인하여 공공 시장 정보로는 주식시장을 예측할 수 없다는 약형 효율적 시장가설을 반박하는 근거를 추가한다. 본 연구는 아직까지 많이 이루어지지 않은 기계학습의 변수 중요도를 이용한 주식시장 분석의 한가지 방법을 제시한다는 점에서도 의미를 갖는다.

II. 선행 연구

효율적 시장 가설(Efficient Market Hypothesis, 이하 EMH) (Malkiel and Fama, 1970)에 따르면 주식 시장은 현재 가능한 모든 정보를 포함하고 있으며 가격이 무작위로 움직이기 때문에 과거 가격, 거래량 등의 시장 정보로 향후의 주가를 예측하는 것은 불가능하다. 하지만 이후 EMH 를 반박하는 연구들이 다수 등장하였는데 그 중 Jegadeesh and Titman (1993)은 과거 승자였던 주식이 패자의 주식보다 앞으로 6 개월-12 개월 동안 더 좋은 성과를 보이는 가격의 모멘텀(momentum) 현상을 언급했으며, Moskowitz et al. (2012)은 주가지수의 과거 가격이 이러한 모멘텀을 보이는 것을 이용해 과거 시장 정보를 통해 금융시장을 어느정도 예측할 수 있음을 보였다. 따라서 과거 시장 정보를 이용하여 금융시장을 예측하고자 하는 연구들이 계속 진행되어 왔으며 사용가능한 데이터의 양이 방대해지고 인공지능에 대한 관심이 고조됨에 따라 기계학습을 이용한 금융시장 예측에 대한 관심 또한 커졌다.

과거 시장 정보를 담은 기술적분석 지표(technical analysis indicator)를 예측변수로 하여 기계학습을 통해 금융 시장을 예측하려는 연구들이 최근 증가하고 있다. 국내 주가지수를 예측하는 연구를 살펴보면 Kim (2003)은 서포트 벡터 머신(Support vector machine, 이하 SVM)을 이용하여 KOSPI 의 증감을 예측했으며, 박재연 외 (2016)는 SVM, ANN(Artificial Neural Networks), 라쏘(Lasso) 회귀모형을 사용하여 KOSPI 를 예측하고 비교했다. 이우식 (2017)은 딥러닝(deep learning) 구조의 모형으로 KOSPI 방향을 예측하였다. 세 연구 모두 기술적지표를 예측변수로 사용했다는 공통점이 있다. 하대우 외(2019)는 XGBoost 모형을 이용하여 KOSPI200 등락 예측의 유용성을 보였다. 또한 본 연구에서 사용하는 기계학습 모형인 랜덤포레스트(Random Forest, 이하 RF) 이용하여 국내 주가지수를 예측한 연구로 박석진, 정재식(2019)은 미시구조론 변수를 이용하여 KOSPI200 선물 등락을 로지스틱 회귀모형(Logistic Regression), SVM, RF 로 예측해 비교했으며 RF 의 성능이 뛰어남을 보였다. 또한 RF 를 포함한 여러 기계학습 모형을 사용하여 성능을 비교한 해외 시장에 대한 연구들이 다수 있는데 그 중 Ballings et al. (2015), Patel et al. (2015)은 RF 의 성능이 주가 혹은 주가지수를 예측하는데 있어서 SVM, ANN 등의 다른 기법보다 뛰어나거나 적어도 떨어지지 않음을 보인 바 있다. 따라서 본 논문은 주가지수 등락의 예측력이 어느정도 보장된 RF 를 예측모형으로 사용하기로 했으며, RF 를 이용한 변수중요도 측정 방법 중 순열중요도(permutation importance)를 이용한다.

기계학습의 변수중요도를 이용하여 금융 시장 변수를 분석한 연구를 살펴보면 우선 Nti et al (2019)은 RF 기반의 변수 선정을 통해 주가예측에 대한 거시경제 변수의 중요도를 측정했다. Haq et al. (2021)은 44 개의 기술적 지표를 변수로 하여 NASDAQ 의 주식들의 추세 예측에 대한 변수 선정을 했다. 이때 LR, SVM, RF 을 각각 이용하여 변수 중요도를 구했다. 본 연구는 Haq et al. (2021)에서 사용한 RF 의 순열 중요도를 응용하여 국내 주식 시장을 예측하는 변수를 분석했다. 국내 연구 중에는 이재응, 한지형 (2021)이 Layer-wise Relevance Propagation (LRP)를 이용하여 KOSPI 증감에 대한 기술적 지표 및 거시경제 지표 영향을 분석한 바

있으나 RF 의 순열 중요도를 사용한 국내 주식 시장에 대한 연구는 아직까지 없다. 또한 모형의 성능을 최대화하는 과정으로 변수 선정(selection)을 위한 목적의 중요도 연구가 주를 이루었기 때문에 변수의 중요도를 통해 경제학적 해석을 한 연구는 거의 없다.

RF 의 순열 중요도는 변수들 간에 선형 또는 비선형의 상관관계가 있을 때 해당 변수의 순열 중요도가 낮게 편향되어 있음을 Strobl et al. (2008), Nicodemus et al. (2010), 신승범, 조형준(2021)에서 밝혔다. 이러한 문제점을 보완하기 위해 변수 간의 상관관계가 변수 중요도에 영향을 미치지 않도록 하는 방법에 대한 연구들 또한 진행되었다. Strobl et al. (2008), Debeer and Strobl (2020)은 RF 를 이용한 조건부 순열 중요도(conditional permutation importance)를, 신승범, 조형준 (2021)은 RF 를 이용한 Max MDI(mean decrease impurity), Max MDA(mean decrease accuracy)[1]를 De Prado (2020)은 예측변수들을 미리 군집화(clustering)하여 변수 군집에 대한 중요도를 계산하는 군집화 피쳐 중요도(Clustered-feature importance) 방법을 제안했다.

본 논문은 여러 연구에서 사용한 기술적분석 지표와 거시경제 변수를 예측변수로 하고, RF 를 변수중요도를 구하는 예측모형으로 정한다. 또한 De Prado (2020)가 제시한 군집화 순열 중요도(Clustered-permutation importance)를 이용하여 국내 주식 시장 예측에 대한 기여도를 알아보고자 KOSPI 증감에 대한 예측 중요도를 측정한다.

[1] MDA(mean decrease accuracy)는 순열중요도와 같은 개념으로 기준이 되는 점수가 정확도(accuracy)인 순열중요도이다.

III. 분석 자료 및 분석 모형

본 장에서는 모형이 훈련하고 예측하고자 하는 목표변수를 설정하는 방법, 변수 중요도를 도출하게 될 예측변수를 추출하는 방법을 나누어 소개한다.

1. 목표변수

예측모형이 예측하고자 목표변수는 KOSPI 일별 수정 종가 기준으로 h거래일 후 대비 증감 여부로 하며, 증가하였으면 1, 감소하였으면 0 으로 하는 이중-클래스 라벨(binary-class label)로 설정한다. 이 때 h에 대해 각각 h가 1, 5, 20 일 때의 결과를 도출하여 서로 다른 기간에 대한 가격 방향에 대한 예측력과 변수중요도를 비교한다. 샘플 거래일수는 2386 이며, 2012 년 1 월 3 일부터 2021 년 12 월 30 일까지의 KOSPI 자료를 사용한다. 훈련 자료로는 그 중 2012 년 1 월부터 2018 년 12 월까지로 총 1665 일, 테스트 데이터로 2019 년 1 월부터 2021 년 12 월까지로 총 721 일로 한다.

$$y_t = \begin{cases} 1, & if\ X_{t+h} - X_t > 0 \\ 0, & if\ X_{t+h} - X_t \leq 0 \end{cases} \quad h = 1, 5, 20 \quad \text{---------- <식 1>}$$

2. 예측변수

사용하는 예측변수 데이터로는 기술적 분석 지표 18 개, 투자주체별 수급 변수 3 개, 거시경제 변수 7 개가 있다.

기술적분석 지표는 과거 일정 기간 동안의 KOSPI 일별 가격(시가, 고가, 저가, 종가)과 거래량[2]을 이용하여 계산되는 지표로 주로 금융자산의 차트 분석에 사용된다. 기술적분석 지표는 시장 참여자의 행동이나 시장의 흐름의 패턴 등을 반영하기 때문에 향후 주식시장의 추세에 대한 신호로 여겨지며 예측변수로 많이 사용되고 있다. 또한 기술적 분석 지표는 과거의 시장 정보를 담고 있으며 각각 다른 기간에 대한 추세, 추세의 강도, 변동성, 거래량의 추세 등을 수치화기 때문에 기술적 분석 지표를 통해 과거의 공공 시장 정보가 주식시장 예측에 어느정도 도움이 되는지를 변수 중요도로 분석할 수 있다. 본 연구에서는 시계열 정상성(stationary)을 가지는 변수만을 사용하고자 훈련 표본 기간에 대해 Augmented Dickey-Fuller 검정을 하여 p 값이 0.05 보다 작은 지표 중 총 9 개의 변수를 선정했다. 또한 각 기술적 지표의 계산 과정의 과거 기간(look-back window)은 이전 연구에서 주로 쓰이는 것을 사용했으며, 그것의 두배 기간으로 계산한 지표를 추가하여 각 기술적 지표를

[2] KOSPI 일별 가격과 거래량은 한국증권거래소(KRX)를 이용해 수집했다.

두 개씩으로 생성하여 기술적 분석 지표 총 18 개를 생성했다. 이는 단기적, 장기적 과거의 시장정보를 나누어 보기 위함이다.

기술적 분석 지표는 크게 추세, 거래량, 변동성을 나타내는 지표로 나눌 수 있다. 우선 추세 지표로는 RSI(Relative Strength Index), WR(William's R) [3], DPO(Detrended Price Oscillator), MACD(Moving Average Convergence Divergence), MACD Difference 가 있으며 이 지표들은 계산해내는 과정은 모두 다르지만 일정 기간 동안의 추세를 나타내는 지표로 구분할 수 있다. 거래량 지표로는 FI (Force Index)와 MFI(Money Flow Index) 가 있다. 단, 거래량 지표는 거래량 자체를 나타내기 보다 과거 가격 추세에 거래량을 곱하여 거래량의 추세를 나타내는 지표로 볼 수 있다. 다음으로 변동성 지표로는 ATR(Average True Range)과 STD(Standard Deviation)가 있다. [표 1]을 통해 모든 기술적 분석 지표의 계산과정을 살펴볼 수 있다.

기술적분석 지표에 더해 투자 주체별 순매수량을 변수로 추가했다. 기관 및 외국인 투자자의 거래와 국내 주식시장은 밀접한 관계가 있으며(정재위,2002; 정현철, 정영우, 2011), 투자주체별 순매수강도는 주가수익률에 영향을 미칠 수 있다(김수경, 변영태; 2011). 본 연구에서는 투자 주체별 수급 변수로 개인, 기관계, 외국인의 순매수량을 사용했으며 모든 KOSPI 종목의 순매수량의 합을 사용했다. 또한 자료의 잡음(noise)을 감소시키고자 각각에 대해 5 일 이동평균화하였다.

다음으로 거시경제 변수로 일별 자료로 구할 수 있는 환율과 상품가격을 사용했다. 환율로는 국내 경제와 큰 관련이 있다고 판단되는 원 기준 달러, 유로, 엔, 위안화를 사용했다. 상품 가격으로는 국제 금, 원유, 천연 가스의 선물 가격을 사용하였다. 금 가격은 COMEX 기준 국제 금 2022 년 4 월 선물 연결가격, 원유는 NYMEX 기준 서부텍사스유(WTI) 2022 년 4 월 선물 연결가격, 천연 가스는 NYMEX 기준 천연가스 2022 년 4 월 선물 연결가격을 사용했다. 또한 각 거시경제 변수에 대한 추세성을 제거하기 위해 일일 변화율을 계산해 변화율을 최종적인 변수로 사용했다. [표 2]에 사용한 28 개의 변수가 나와있다.

마지막으로 위의 자료를 이용하여 모형을 학습시키기 전에, 모든 예측변수에 대해 표준화 스케일러(Standardize scaler)를 사용하여 훈련기간의 데이터를 학습한 후 훈련기간 및 테스트기간의 데이터에 대해 표준화 변형(transform)한 최종 예측변수를 사용한다. 표준화를 하는 이유는 기계학습 모형이 여러 변수가 있을 때 변수 값의 범위가 모두 다르면 효과적으로 분류할 수 없어서 일정한 수준으로 맞추어야 하기 때문이다.

---

[3] RSI, WR 은 추세의 강도를 나타내는 모멘텀(momentum) 지표로 추세 지표와 따로 분류하는 경우도 있지만 본 논문에서는 다른 추세 지표와 높은 상관관계를 가지기에 이를 묶어 추세 지표로 분류한다.

[표 1] 기술적 분석 지표

	지표 (기간)	식
추세 지표	RSI (n =14) (n =28)	$RS = \frac{EMA(n)\ of\ Up}{EMA(n)\ of\ Down}$ $RSI = 100 - \frac{100}{1+RS}$
	WR (n =14) (n =28)	$WR = \frac{Highest(n) - Close}{Highest(n) - Lowest(n)}$
	DPO (n =20) (n =40)	$Close\ \left(\frac{n}{2} + 1\ days\ ago\right) -\ SMA(n)\ of\ Close$
	MACD (n_1, n_2=26,12) (n_1, n_2= 52,24)	$EMA(n_2)\ of\ Close - EMA(n_1)\ of\ Close$
	MACD Difference (n_1, n_2, n_3= 26,12,9) (n_1, n_2, n_3= 52,24,18)	$MACD(n_1, n_2) - EMA(n_3)\ of\ MACD(n_1, n_2)$
거래량 지표	FI (n =13) (n =26)	$FI(1) = (Close - Prior\ Close) * Volume$ $FI(n) = EMA(n)\ of\ FI(1)$
	MFI (n =14) (n =28)	$Money\ Flow = \frac{High + Low + Close}{3} * Volume$ $MFR = \frac{Sum\ of\ Positive\ Money\ Flow}{Sum\ of\ Negative\ Money\ Flow}$ $MFI(n) = 100 - \frac{100}{1+MFR}$
변동성 지표	ATR (n = 14) (n = 28)	$TR = Max(High, PriorClose) - Min(Low, PriorClose)$ $First\ ATR = \frac{1}{n}\sum_{i=1}^{n} TR$ $ATR = \frac{(Prior\ ATR\ *\ (n\ -\ 1) + TR)}{n}$
	STD (n = 20) (n = 40)	$Moving\ Standard\ Deviation\ (n)\ of\ Close$

주: $SMA(n)$, $EMA(n)$은 각각 n 일 이동평균, 지수화 이동평균, Up 는 상승하는 종가 변화, $Down$ 는 하락하는 종가 변화, $Highest(n)$는 n 일 간 가장 높은 가격, $Lowest(n)$는 n 일 간 가장 낮은 가격, $Close$는 종가, $High$는 고가, Low는 저가, $Volume$은 거래량을 나타낸다. 각 지표의 계산과정에 있는 n 을 각각 두 개씩 설정하여 지표를 도출했다.

[표 2] 예측 모형에 사용되는 예측 변수

구분	기술적지표−추세	기술적지표−거래량	기술적지표−변동성	주체별 순매수량	환율	상품가격
변수	RSI (14), RSI (28), WR (14), WR (28), DPO (20), DPO(40), $MACD$ (26,12), $MACD$ (52,24), $MACD\ Diff$ (26,12,9), $MACD\ Diff$ (52,24,18)	FI(13), FI(26), MFI(14), MFI(28)	ATR(14), ATR(28), STD (20), STD(40)	$individuals\ sma5$ $foreigners\ sma5$ $institutions\ sma5$	USD/KRW, EUR/KRW , JPY/KRW, CNY/KRW	$gold$, $crude\ oil$, $natural\ gas$

주: 주체별 순매수량은 5 일 이동평균을, 환율과 상품가격은 일일변화율을 이용했다.

IV. 분석 모형

본 장에서는 랜덤 포레스트 모형에 대해 간략히 설명하고 최적의 모형을 구축하기 위한 하이퍼파라미터 튜닝(hyperparameter tuning) 결과를 살펴본다. 그리고 예측 변수 중요도를 측정하기 위한 방법인 순열 특징 중요도를 설명하고, 변수 간 상관관계를 제거하기 위해 변수들을 계층적 군집화(hierarchical cluster)하는 과정을 살펴본다. 마지막으로 변수 군집을 생성하여 군집 결과를 변수의 특징과 연결 지어서 확인한다.

1. 랜덤 포레스트

랜덤포레스트(Random Forest, 이하 RF)는 다수의 훈련된 의사결정나무(decision tree)를 사용하는 앙상블(ensemble) 모형으로 부트스트랩(bootstrap)을 통해 무작위로 샘플을 여러 번 추출해 결과를 집계하고, 다수결로 예측치를 도출하는 모형이다.

의사결정나무는 각 노드(node)에서 지니(Gini) 지수 등의 불순도를 최소화하는 과정으로 학습하는 분류기 모형이다. 값을 분할하는 단순한 규칙 구조에 의해 분류하기 때문에 직관적이며 설명력을 유지한다는 점이 강점이다. 하지만 의사결정나무는 자료의 소음(noise)이나 조금의 변형에 따라 규칙 구조가 완전히 달라질 수 있으며 학습 과정의 성능과 다르게 표본 외 예측에 대해서는 매우 낮은 성능을 보이는 과적합(overfitting) 문제가 발생한다는 점이 큰 단점이다.

RF 는 이러한 수많은 의사결정나무 모형을 이용해 다수결로 결과를 도출하므로 개별 의사결정나무 모형이 가지는 불안정성 및 과적합 문제를 보완하며, 표본 외 예측 성능을 높이는 강점을 가지고 있다.

효과적인 변수 중요도를 도출하기 위해서는 예측모형의 예측력이 어느정도 높아야 한다. 만약 모형의 예측력이 낮다면 해당 모형을 이용하여 중요도를 구하는 것은 무의미 해진다. 따라서 모형의 성능을 높이기 위해 최대화하는 모형 최적화 방법인 하이퍼파라미터 튜닝(hyperparameter tuning)을 진행한다. 하이퍼파라미터 튜닝이란 기계학습 모형의 여러 파라미터를 시도해본 후 최적의 결과를 보이는 파라미터를 찾아내는 과정을 뜻한다. RF 의 하이퍼파라미터로는 의사결정 나무의 개수(number of trees), 최대 깊이(maximum depth), minimum sample split, minimum sample leaf 등이 있다. 본 연구에서는 RF 의 성능에 큰 영향을 미치는 의사결정 나무의 개수와 최대 깊이 두 가지의 하이퍼파라미터에 대해 후보군을 두고 훈련시켜 튜닝을 진행하며, 각각 1 일, 5 일, 20 일 후 주가지수 방향의 목표변수마다 진행하여 세 개의 최적 모형을 결정한다. 하이퍼파라미터의 후보군으로는 의사결정 나무의 개수는 20, 50, 100, 최대 깊이는 3, 9, 15 로 설정하고 성능 중 정확도(accuracy)를 가장 높이는 조합을 찾는 GridSearch 방법을 사용한다. GridSearch 는 파이썬(python) 프로그램의 라이브러리 *sklearn* 에 내장되어 있는 하이퍼파라미터 튜닝 방법으로 후보군의 모든 조합을 통해 결과를 도출해본 후 최적의 조합을 선정하는 알고리즘이다.

하이퍼파라미터 튜닝의 결과로 최적의 나무의 개수와 최대 깊이 조합은 1 일 후, 5 일 후, 20 일 후의 KOSPI 등락 각각에 대해 (100, 3), (100, 3), (50, 15)이다. 따라서 실증분석에서 각 목표변수를 예측할 때 해당 하이퍼파라미터를 RF 분류기에 적용한다.

[표 3] 하이퍼파라미터 튜닝 결과

	Hyperparameters	
	Number of trees [20, 50, 100]	Maximum depth [3, 9, 15]
y_1	100	3
y_5	100	3
y_{20}	50	15

2. 변수 중요도

기계학습을 이용한 금융 시장의 예측에 대한 연구로는 정확도를 높이는 연구가 주를 이루었다. 본 장에서는 정확도를 높이는 목적에서 더 나아가서 예측 결과에 영향을 끼친 예측변수에 대한 분석을 통해 기계학습 모형이 가지는 'black box' 문제를 완화하고자 한다. 본 장에서는 주가지수의 향후 방향을 예측하는데 어떠한 변수가 많은 기여를 했는지 알 수 있는 예측변수 중요도를 도출하는 방법으로 순열 특징 중요도에 대해 알아보고, 변수 간 상관관계가 있을 때 중요도는 하향 편향을 가지므로 변수들을 군집화하여 변수 군집에 대한 중요도를 도출하는 군집화 순열 중요도 방법을 소개한다.

2.1. 순열 중요도

중요도 측정을 위해 순열 중요도(Permutation importance)를 사용한다. 순열 특징 중요도란 학습된 기계학습 모형을 통해 중요도를 구하는 방법으로 특정한 변수 값(j)을 무작위로 재배열하여 정보를 제거한 후 테스트 데이터에 대한 예측성능(s_j)이 재배열 전(s)에 비해 얼만큼 감소하는지를 측정하는 방법이다. 이 때 기준의 되는 성능은 정확도, F1 점수, ROC-AUC 점수 등 분류기 모형의 성능을 나타내는 어느 지표라도 사용 가능하며, 순열(permutation)을 여러 번(K) 반복해 평균을 구하여 해당 변수의 중요도를 측정한다. 따라서 측정된 감소치는 모형이 해당 예측변수에 얼만큼 의존하는지를 나타낸다.

$$PI_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j} \qquad \text{---------- <식 2>}$$

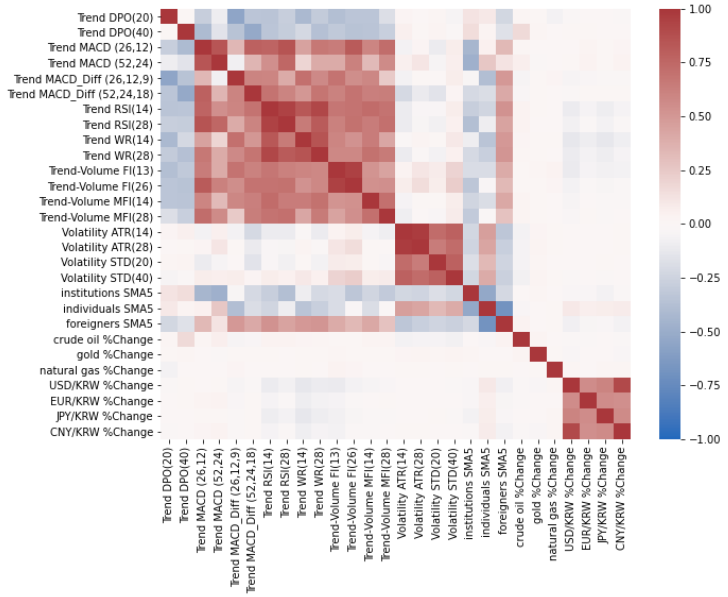
순열 중요도 방법은 사용할 모형과 측정하고자 하는 점수를 자유롭게 설정할 수 있으며, 예측변수 값의 순열을 원하는 만큼 반복하여 평균을 구할 수 있다는 장점이 있다. 또한 순열 중요도는 RF 관련 연구에서 주로 쓰이는 중요도 측정 방식인

MDI(mean decrease impurity)와 비교하여 많은 강점이 있다. MDI는 테스트 자료에 대한 예측이 아닌 모형의 학습과정에서 중요도를 계산하기 때문에 인샘플 편향(in-sample bias)이 존재하며, 불순도(impurity)를 기준으로 중요도를 계산하기 때문에 변수의 카디널리티(cardinality)가 중요도에 영향을 미치는 한계점이 존재한다. 반면 순열중요도는 테스트 자료에 대한 예측과정에서 중요도를 도출하기 때문에 인샘플 편향이 없으며 성능의 변화가 중요도 측정 방법이므로 카디널리티의 문제도 없다. 또한 MDI는 모든 변수의 중요도가 양수이며 합이 1이기 때문에 변수 간의 상대적인 중요도만을 고려하게 되는 문제점이 있는데 반해 순열 중요도는 중요도가 음수일 수도 있으며 변수가 예측에 도움이 되는 정도를 절대적으로 평가할 수 있다.

2.2. 변수 군집화

변수 간에 상관성이 존재하면 중요도를 구하는데 문제가 생긴다. 상관관계가 강한 변수들을 함께 사용하게 되면 해당 변수는 비슷한 빈도로 의사결정 나무에 의해 선택되기 때문에 변수 서로의 변수중요도를 낮추게 된다. 이를 해결하기 위한 방안으로 예측변수들을 미리 군집화한 후 해당 군집을 이루는 변수 하위집합(subset)을 무작위로 섞은 후 모형의 정확도가 얼만큼 감소하는지를 측정하는 순열 중요도를 구한다. 따라서 하위집합인 변수 군집의 중요도를 구하게 되며 변수 군집 간에는 상관관계가 없으므로 변수중요도가 효과적으로 측정될 수 있다. [그림 1]을 보면 본 연구에서 사용되는 변수 간에 상관관계가 강한 것들이 있는 것을 볼 수 있다. 따라서 변수에 대한 군집화가 필요하다.

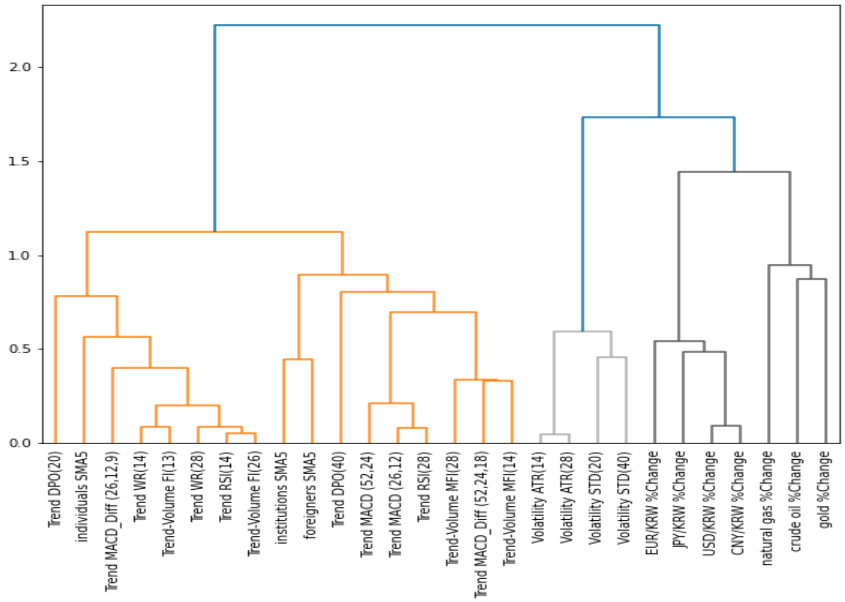
[그림 1] 예측변수 간의 상관관계를 나타낸 그래프



본 연구에서는 변수의 군집화 방법으로 계층적 군집화(hierarchical clustering)를 사용하고자 한다. 계층적 군집화는 변수 간의 거리가 가장 가까운 두 개를 선택한 후 하나로 묶어가며 군집화하고, 군집끼리 거리가 가까운 두 개를 하나로 합치며 군집 개수를 줄여 가는 방법이다. 군집 간의 거리를 계산하는 연결 기준(linkage criterion)으로는 최소 분산을 계산하는 왈드 기준(Wald' s criterion)을 사용하며, Spearman 상관계수를 거리 측도로 하여 선형적 관계를 고려한다. 이 때 거리 유사도의 임계값을 임의로 설정하여 군집 개수를 조정할 수 있는데 본 연구에서는 임계값을 1.0 로 설정한다. 그렇게 하여 30 개의 예측변수를 계층적 군집화를 통해 군집화한 결과 총 5 개의 군집을 생성하였다. [그림 2]를 보면 계층적 군집화에 따라 변수 간 거리가 가까운 것끼리 묶여가는 형태를 볼 수 있으며 변수 간 유사도(y 축)의 임계값이 1.0 일 때 묶여 있는 변수들끼리 그룹으로 만들면 군집화가 완성된다.

[표 4]를 보면 5 개의 군집에 속한 각각의 변수가 나타나 있다. 군집 1 의 경우 과거 시장 추세 지표와 개인 순매수량이 있으며, 군집 2 는 과거 시장 추세 지표와 기관계, 외국인의 순매수량이 포함되어 있다. 군집 1 과 군집 2 의 과거 시장 추세 지표의 차이는 지표를 계산하는 과거 기간에 있으며 군집 1 은 단기적 군집 2 는 장기적 과거의 추세라고 볼 수 있다. DPO, RSI, MACD Difference 는 계산 기간을 두 배로 한 지표가 군집 2 에 속한다. 군집 3 은 주가지수의 변동성 지표로 구성되어 있고, 군집 4 와 군집 5 은 각각 환율 변화율과 상품가격 변화율로 구성된다.

[그림 2] 예측변수 간 유사도를 나타낸 계층적 군집 덴도그램(Dendogram)



주: y 축은 유사도를 나타내며 본 연구에서는 유사도가 1.0 이하에서 생성된 군집을 최종 변수군집으로 설정했다.

[표 4] 예측변수에 대한 계층적 군집화 결과

	구성 변수	특징
군집 1	$DPO(20), MACD\ Diff\ (26,12,9), RSI(14), WR(14), WR(28), FI(13), FI(26), individuals$	시장 추세(단기)
군집 2	$DPO(40), MACD\ (26,12), MACD\ (52,24), MACD\ Diff\ (52,24,18), RSI(28), MFI(14), MFI(28) institutions, foreigners$	시장 추세(장기)
군집 3	$ATR(14), ATR(28), STD(20), STD(40)$	변동성
군집 4	$USD/KRW, EUR/KRW, JPY/KRW, CNY/KRW$	환율
군집 5	$crude\ oil, gold, natural\ gas$	상품가격

V. 실증 분석 결과

본 장에서는 RF 분류기의 예측 성능을 로지스틱 회귀모형(Logistic Regression, 이하 LR)과 비교하여 예측성능을 확인하고, 변수중요도를 측정하는 모형으로 적합한지를 판단한다. 그리고 학습된 RF 모형을 이용해 KOSPI 증감을 예측하는 과정으로부터 예측변수의 중요도를 구하며, 변수중요도 분석은 4 장에서 도출한 변수군집에 대해 진행하며 순열중요도 방법을 사용한다. 마지막으로 과거 시장 추세, 변동성, 거시경제 등이 국내시장 예측에 기여하는 정도를 살펴본다.

1. 분류기 예측 성능

예측 성능에 대한 실증 분석으로 목표변수와 예측변수를 포함한 시계열 자료를 훈련 기간과 테스트 기간으로 나누어 훈련 기간의 데이터를 사용하여 모형을 학습시키고 테스트 데이터에 대해 예측한 결과를 확인했다. 분석 기간으로는 2012 년 1 월 1 일부터 2021 년 12 월 31 일 까지의 표본 기간에 대해 2012 년 1 월 1 일부터 2018 년 12 월 31 일까지를 훈련 데이터로 하고 2019 년 1 월부터 2021 년 12 월까지를 테스트 데이터로 했다. 예측 목표변수는 KOSPI 지수의 1 일, 5 일, 20 일 후 방향으로 총 3 개로 설정하여 결과를 비교했으며, 예측변수로는 3 장에서 설명한 30 개의 생성된 변수를 사용했다. 예측 모형으로는 RF 분류기 모형을 사용하여, RF 와 함께 LR 의 성능을 구한다. LR 은 일반적인 회귀모형과 마찬가지로 종속변수와 독립변수 간의 관계를 구체적인 함수로 나타내어 예측에 사용하며, 이진(binary) 종속변수에 대해 독립변수의 선형 결합을 이용하는 확률적 모형이다. 예측 값을 [0, 1]로 하는 분류기 모형으로 고려하여 그 결과를 RF 와 비교한다. 결과를 살펴보기 전에 분류기 모형의 예측 성능의 척도로 사용되는 점수인 정확도(accuracy), F1 점수, ROC-AUC 점수에 대해 설명한다.

[표 5] 혼돈 행렬

		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

[표 5]는 이진분류기의 혼돈 행렬(confusion matrix)을 나타낸 것으로 주가지수의 등락을 예측한다고 했을 때 TP(true positive)는 실제 주가지수가 상승하며 모형이 주가지수가 상승하는 것으로 맞게 예측하는 것이고, FN(False negative)는 실제 주가지수가 상승하는데 모형은 주가지수가 하락하는 것으로 틀리게 예측하는 것이다. FP(False Positive)는 실제 주가지수가 하락하는데 모형은 주가지수가 상승하는 것으로 틀리게 예측하는 것이고 TN(true negative)은 실제 주가지수가 하락하며 모형 또한 주가지수가 하락하는 것으로 맞게 예측하는 것이다. 본 연구에서는 예측 모형의 성과를 나타내기 위해 사용되는 지표로 정확도, F1 점수, ROC-AUC 점수를

사용한다. 세 점수 모두 0 과 1 사이 값으로 1 에 가까울수록 모형이 성능이 뛰어나다고 할 수 있다. 이 중 정확도와 F1 점수는 혼돈 행렬의 TP, FP, FN, TN 으로 계산되는 점수이며 식(3.1), 식(3.4)와 같이 계산된다.

정확도 $(accuracy) = \frac{TP+TN}{TP+TN+FP+FN}$ ---------- 식(3.1)

정밀도 $(precision) = \frac{TP}{TP+FP}$ ---------- 식(3.2)

재현율 $(recall) = \frac{TP}{TP+FN}$ ---------- 식(3.3)

F1 점수 $(F1-score) = 2 \cdot \frac{precision \cdot recall}{precision+recall}$ ---------- 식(3.4)

정확도는 모든 예측치 중 맞게 예측한 경우를 나타낸 수치이고, F1 점수는 정밀도와 재현율의 조화평균으로 계산한 수치이다. 정확도는 목표변수인 라벨(label)이 불균형일 때 편중(bias)를 가지는 데, F1 점수는 이러한 문제를 보완한 점수이다. 다음으로 ROC-AUC 점수는 ROC 곡선의 아래 면적을 0 과 1 사이로 수치화한 점수이다. ROC 곡선은 x 축인 TPR 와 y 축인 FPR[4]의 관계를 그린 곡선으로, ROC-AUC 점수가 높으면 효과적으로 모형이 학습되었다고 할 수 있으며, 0.5 에 가까울수록 분류 결과가 운에 의해 분류한 것과 마찬가지인 것으로 해석한다(Bradley and Andrew, 1997).

테스트 기간에 대한 RF 와 LR 분류기의 예측 성능을 보면 [표 6], [그림 3]와 같다. LR 의 경우, 예측하는 KOSPI 방향을 정하는 향후 거래일 수가 커질수록 정확도, F1 점수, ROC-AUC 점수가 낮아진다. 또한 ROC AUC 점수가 0.5 에 가깝기 때문에 LR 은 분류 능력이 없다고 할 수 있다. 이처럼 낮은 성능의 분류기 모형을 이용하여 변수 중요도를 구하게 되면 변수중요도가 높더라도 예측이 적절하게 이루어지지 않은 모형이기 때문에 중요도가 가지는 의미가 없다. 반면 RF 의 경우, ROC-AUC 점수가 y_1, y_5, y_{20}(각각 1 일, 5 일, 20 일 KOSPI 방향) 예측에 대해 약 0.61, 0.61, 0.79 으로 분류를 보다 효과적으로 한 것임을 알 수 있다. 또한 정확도와 F1 점수 모두 LR 보다 현저히 높으며 각각 y_1, y_5, y_{20} 의 예측에 대해 정확도는 약 0.60, 0.63, 0.70, F1 점수는 0.69, 0.75, 0.77 이다. 이는 RF 의 변수중요도가 무의미한 의미를 가지지 않는다는 사실을 뒷받침한다. 또한 예측하는 KOSPI 방향 거래일 수에 따라 결과를 살펴보면 예측하는 향후 거래일이 길수록 성능이 높아진다. 즉 1 일 후 주가지수의 방향보다 5 일 후 주가지수 방향이 더 예측 가능하며, 20 일 후 주가지수 방향에 대한 예측력이 가장 높다.

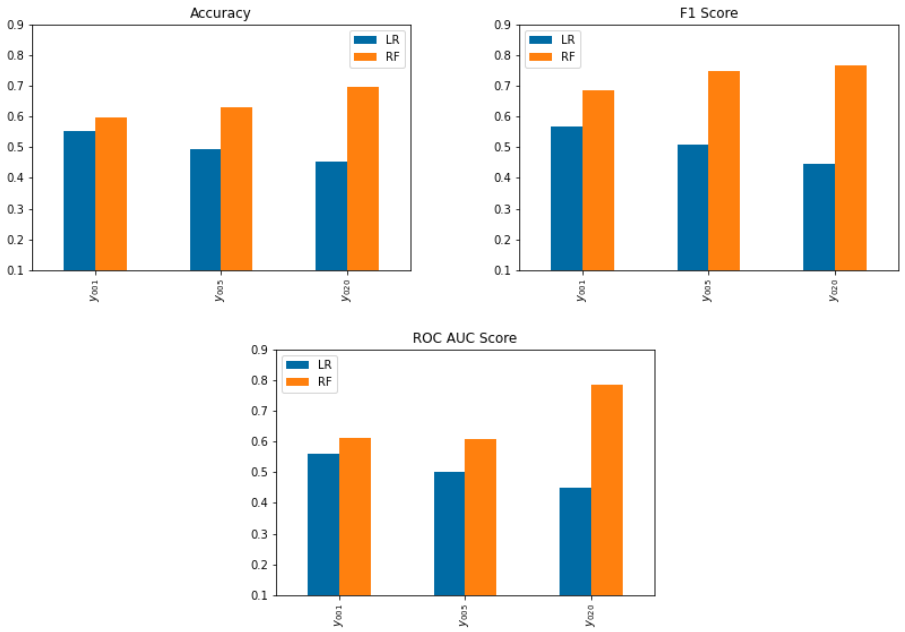
[4] TPR(True positive rate) $= \frac{TP}{TP+FN}$, FPR(False positive rate) $= \frac{FP}{FP+TN}$

[표 6] 예측 성능 결과 비교

Label	Model	Accuracy	F1 score	ROC-AUC score
y_1	LR	0.5520	0.5664	0.5588
	RF	0.5964	0.6861	0.6107
y_5	LR	0.4938	0.5074	0.4998
	RF	0.6297	0.7474	0.6096
y_{20}	LR	0.4521	0.4476	0.4481
	RF	0.6976	0.7666	0.7850

주: Label(목표변수)의 y_1, y_5, y_{20} 은 각각 1 일, 5 일, 20 일 후의 KOSPI 증감을 의미한다.

[그림 3] 예측 성능 결과 비교



2. 변수 중요도 분석

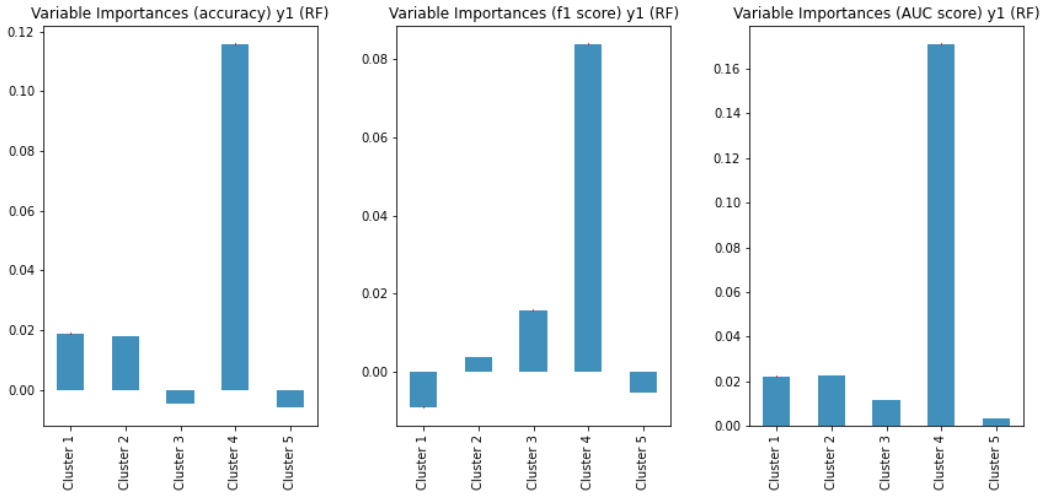
학습된 RF 분류기 모형을 사용하여 테스트 기간(2019 년-2021 년; 3 년)에 대한 KOSPI 증감을 예측하여 변수군집의 중요도를 구할 수 있다. 3 장에서 설정한 5 개의 변수군집에 대해 각각 순열 중요도를 계산했으며, 해당 변수군집에 속한 모든 변수를 무작위로 재배열한 후, 여러 번 반복하여 구한 예측성능의 평균을 재배열하기 전의 성능으로부터의 감소치를 구하여 도출했다. 본 연구에서는 각각 1 일, 5 일, 20 일 후

KOSPI 방향에 대한 예측 시 정확도, F1 점수, ROC-AUC 점수를 기준으로 세 개의 목표값과 세 개의 점수 기준에 따라 순열중요도를 도출했다. 또한 순열 과정은 10 번 반복했다. [표 7]와 [그림 4,5,6]으로 각 변수군집의 중요도를 확인할 수 있다.

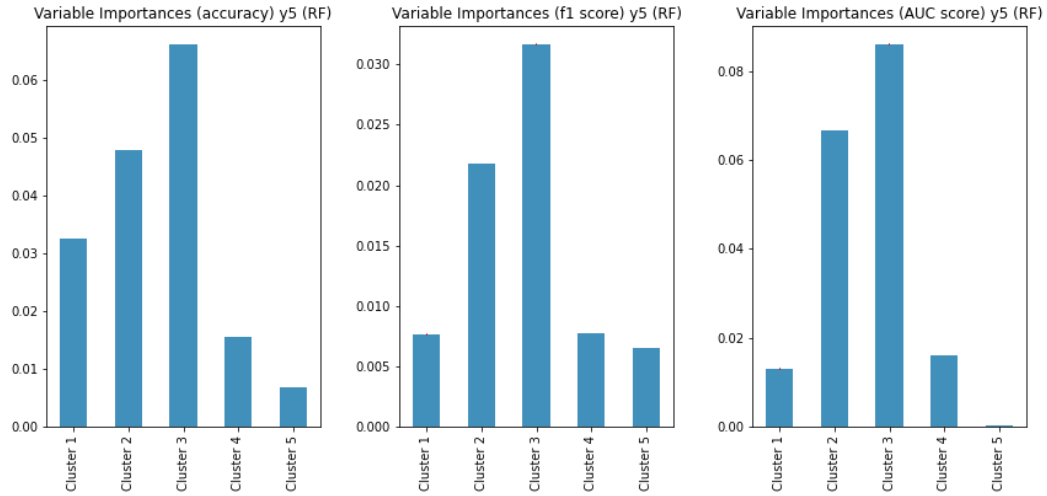
[표 7] 변수 군집의 순열중요도

Label	기준 성능	군집 1	군집 2	군집 3	군집 4	군집 5
y_1	Accuracy	0.0189	0.0178	−0.0046	0.1159	−0.0059
	F1 score	−0.0091	0.0037	0.0158	0.0840	−0.0053
	ROC-AUC score	0.0218	0.0225	0.0116	0.1708	0.0035
y_5	Accuracy	0.0326	0.0480	0.0662	0.0156	0.0069
	F1 score	0.0076	0.0217	0.0316	0.0077	0.0066
	ROC-AUC score	0.0129	0.0667	0.0859	0.0161	0.0002
y_{20}	Accuracy	−0.0834	0.0785	0.1293	−0.0107	−0.0155
	F1 score	−0.1497	0.0709	0.0558	−0.0336	−0.0219
	ROC-AUC score	0.0604	0.2755	0.3256	0.0305	0.0196

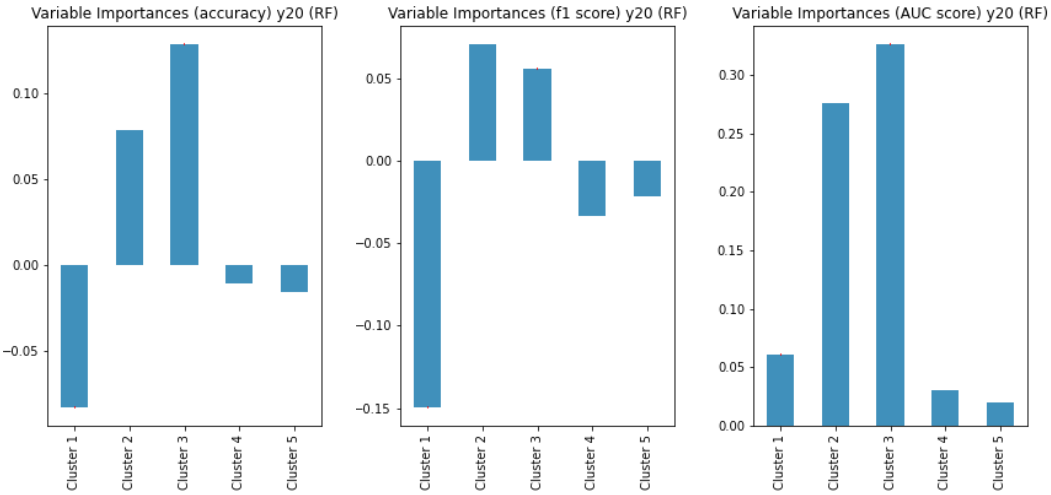
[그림 4] 1 일 KOSPI 방향에 대한 변수 군집 중요도



[그림 5] 5 일 KOSPI 방향에 대한 변수 군집 중요도



[그림 6] 20 일 KOSPI 방향에 대한 변수 군집 중요도



변수중요도 분석 결과를 살펴보면 우선 1 일 KOSPI 방향 예측에 대해서는 환율 변수로 대표되는 군집 4 가 다른 변수군집과 비교해 세 점수 기준 중요도에서 모두 매우 높은 중요도 수치를 보인다. 다른 변수군집의 중요도는 모두 0.02 이하로 매우 낮으며 군집 3 와 군집 5 는 정확도에서, 군집 1 와 군집 5 는 F1 점수에서 음수의 중요도를 보인다. 즉 시장 추세 및 변동성 그리고 상품가격 변수는 1 일 주가지수 방향 예측에 도움이 되지 않거나 오히려 악영향을 미친다.

5 일 KOSPI 방향 예측에 대한 변수중요도를 보면 시장 변동성으로 대표되는 군집 3 의 중요도가 가장 높으며 다음으로 시장 장기 추세로 대표되는 군집 2 의 중요도가 높다. 또한 환율 변수의 중요도가 1 일 KOSPI 방향 예측 중요도와 대조적으로 비교적 낮은 편이다.

20 일 방향에 대해서도 마찬가지로 변동성 변수와 시장 장기 추세와 높은 중요도를 보인다. 특히 두 변수 군집(군집 2, 군집 3)의 중요도가 세 점수 기준에서 모두 5 일 방향 예측보다 높은 것을 확인할 수 있다. 또한 시장 단기 추세로 대표되는 군집 1 의 중요도가 전체적으로 낮은 편인데 특히 20 일 KOSPI 방향에 대해서는 음수의 중요도를 보여 예측에 오히려 악영향을 미치는 것으로 분석할 수 있다. 이를 통해 과거 시장의 단기적 추세보다 장기적 추세가 향후 주가지수를 예측하는데 도움이 된다는 것을 알 수 있다.

각 변수군집의 중요도를 통해 주가지수를 예측하는데 어떠한 변수가 영향을 미치는지 확인할 수 있었다. 주식 시장의 변동성과 추세는 5 일, 20 일의 KOSPI 방향 예측 도움이 된다. 다만 과거의 추세를 단기와 장기로 나눈다면 단기 추세는 예측에 도움이 되지 않는다. 거시경제 변수를 보면 환율 변수는 1 일 KOSPI 방향 예측에 많은 기여를 하는 반면 5 일, 20 일 방향에서는 그렇지 않다. 상품가격은 모든 기간의 예측에 대해 매우 낮은 중요도를 보여 주가지수 예측에 도움이 되지 않음을 확인할 수 있다. 이러한 결과를 통해 과거 시장 추세와 변동성이 5 일 이상의 미래 주가지수의 움직임을 예측하는 정보를 담고 있다고 말할 수 있으며, 이는 공공 시장 정보를 이용하여 금융시장을 예측할 수 있다는 이론을 뒷받침한다. 다만 1 일 주가지수 방향에 대해서는 이러한 시장 정보가 예측에 도움이 되지 않음을 알 수 있다.

VI. 결론

본 논문은 주식시장 예측을 위한 기계학습 모형의 사용이라는 연구 주제에서 더 나아가 예측에 사용되는 변수들의 중요도를 측정하는 방법을 연구하여 설명가능한 기계학습 모형을 구축했다. 변수중요도를 측정하는 방법으로 학습된 랜덤포레스트 분류기를 이용한 순열중요도를 선택했으며 변수 간의 상관관계가 있을 경우 중요도가 하향 편향되는 경향이 있기 때문에 변수들을 군집화하여 변수군집에 대한 순열중요도를 구하는 군집화 순열 중요도를 사용했다. 또한 KOSPI 등락을 예측하는데 어떤 변수가 많은 기여를 하는지를 분석하여 과거 시장의 추세와 변동성, 환율 및 상품가격이 국내 주식시장을 예측하기 위한 정보를 담고 있는지를 살펴보았으며, KOSPI 방향에 대해 1 일, 5 일, 20 일로 나누어 이를 비교했다. 기존의 국내 주식시장에 대한 기계학습 모형 예측에 대한 연구는 주로 예측력에 초점을 맞추어 이루어졌으며 변수에 대한 분석은 활발히 진행되지 않았다. 특히 변수간 상관관계가 있음에도 이를 고려하지 않은 채 중요도를 도출하였기 때문에 본 연구는 군집화 순열 중요도를 이용하여 변수에 대해 분석한다는 점에서 의의가 있다.

본 연구는 2012 년 1 월 3 일부터 2021 년 12 월 30 일까지의 일별 자료를 이용한 실증 분석을 통해 랜덤포레스트의 성능이 로지스틱 회귀모형과 비교하여 뛰어남을 보였으며 KOSPI 방향 기간이 길수록 예측 성능이 높아지는 것을 확인했다. 다음으로 중요도를 구하기 위한 학습된 분류기 모형으로 최적화된 랜덤포레스트를 선택하여 순열중요도를 구했다. 이 때 변수 간의 상관관계가 존재하기 때문에 계층적 군집화를 통해 5 개의 변수군집을 생성했으며 각 군집은 단기적 시장 추세, 장기적 시장 추세, 시장 변동성, 환율, 상품가격으로 대표된다. 변수군집에 대한 중요도 분석 결과, 1 일 KOSPI 방향에 대해서는 환율 변수가 큰 중요도를 보였으며, 5 일, 20 일 KOSPI 방향에 대해서는 장기적 시장 추세와 시장 변동성 변수가 큰 중요도를 보였다. 따라서 5 일 이상의 KOSPI 방향을 예측하는데 있어서 과거의 시장 정보가 도움이 됨을 확인했다.

본 연구에 더 나아가서 조건부 순열중요도와 같은 다른 중요도 기법을 사용하거나 예측성능이 높은 여러가지 기계학습 모형을 추가적으로 사용해 결과를 비교하고, 본 연구에서 사용한 예측변수에 더해 시장 참여자의 거래 행태 등의 정보를 포함하는 시장 미시구조(microstructure) 변수 등을 추가적으로 사용하는 등 본 연구의 중요도 방법을 응용한 다양한 주식시장 분석이 가능할 것으로 기대된다.

참고 문헌

김수경, 변영태. (2011). 외국인 및 기관투자자의 순매수강도와 주식수익률 간의 관계. 경영과 정보연구, 30(4), 23-44.

박석진, 정재식. (2019). 고빈도 자료를 이용한 머신러닝 모형의 예측력 비교 · 분석: KOSPI200 선물시장을 중심으로. 금융연구, 33(4), 31-60.

박재연, 유재필, 신현준 (2016) 기술적 지표와 기계학습을 이용한 KOSPI 주가지수 예측, 정보화연구, 13:2, 331-340

신승범, & 조형준. (2021). 랜덤포레스트를 위한 상관예측변수 중요도. 응용통계연구, 34(2), 177-190.

이우식. (2017). 딥러닝분석과 기술적 분석 지표를 이용한 한국 코스피주가지수 방향성 예측. 한국데이터정보과학회지, 28(2), 287-295.

이재응, 한지형. (2021). 설명 가능한 KOSPI 증감 예측 딥러닝 모델을 위한 Layer-wise Relevance Propagation (LRP) 기반 기술적 지표 및 거시경제 지표 영향 분석. 정보과학회논문지, 48(12), 1289-1297.

정재위. (2002). 기관투자가의 거래가 증권시장에 미치는 경향에 관한 연구. 세무회계연구, 11(0), 237-249.

정현철, 정영우. (2011). 외국인 순투자가 주가에 미치는 영향. 국제경영연구, 22(1), 1-28.

하대우, 김영민, 안재준.(2019).XGBoost 모형을 활용한 코스피 200 주가지수 등락 예측에 관한 연구. 한국데이터정보과학회지,30(3),655-669.

Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. Expert systems with Applications, 42(20), 7046-7056.

de Prado, M. M. L. (2020). Machine learning for asset managers. Cambridge University Press.

Debeer, D., & Strobl, C. (2020). Conditional permutation importance revisited. BMC bioinformatics, 21(1), 1-30.

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. Statistics and Computing, 27(3), 659-678.

Haq, A. U., Zeb, A., Lei, Z., & Zhang, D. (2021). Forecasting daily stock trend using multi-filter feature selection and deep learning. Expert Systems with Applications, 168, 114444.

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. The Journal of finance, 48(1), 65-91.

Kim, K. J. (2003). Financial time series forecasting using support vector machines. Neurocomputing, 55(1-2), 307-319.

Malkiel, B.G. and Fama, E.F. (1970) Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, 25, 383-417.

Malkiel, B. G. (2003). The efficient market hypothesis and its critics. Journal of economic perspectives, 17(1), 59-82.

Moskowitz, T. J., Ooi, Y. H., & Pedersen, L. H. (2012). Time series momentum. Journal of financial economics, 104(2), 228-250.

Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC bioinformatics, 11(1), 1-13.

Nti, K. O., Adekoya, A., & Weyori, B. (2019). Random forest-based feature selection of macroeconomic variables for stock market prediction. American Journal of Applied Sciences, 16(7), 200-212.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert systems with applications, 42(1), 259-268.

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics, 8(1), 1-21.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. BMC bioinformatics, 9(1), 1-11.

부록

[표 A1] 예측변수의 기초통계량

		평균	표준편차	중간값	최소값	최대값	관측치
기술적 분석 지표	DPO(20)	−0.61	28.77	0.19	−254.5	188.42	2386
	DPO(40)	−0.75	39.31	1.37	−372.81	169.7	2386
	MACD (26,12)	2.93	23.32	3.2	−153.59	120.35	2386
	MACD (52,24)	6.04	32.67	5.49	−136.56	159.09	2386
	MACD_Diff (26,12,9)	−0.02	6.96	0.04	−47.26	34.68	2386
	MACD_Diff (52,24,18)	−0.08	10.57	0.28	−69.87	36.02	2386
	RSI(14)	52.16	12.16	52.39	12.15	86.25	2386
	RSI(28)	51.74	8.45	51.91	20.31	79.57	2386
	WR(14)	−42.9	31.24	−38.78	−100	0	2386
	WR(28)	−41.84	30.47	−38.24	−100	0	2386
	FI(13)	3.05E+08	4.79E+09	1.82E+08	−4.7E+10	4.93E+10	2386
	FI(26)	3.11E+08	3.53E+09	1.27E+08	−3.1E+10	3.29E+10	2386
	MFI(14)	53.29	15.76	53.67	6.53	100	2386
	MFI(28)	53.25	11.26	53.83	19.19	84.58	2386
	ATR(14)	25.53	10.83	22.05	12.19	83.02	2386
	ATR(28)	25.53	9.85	22.18	14.41	64.75	2386
	STD(20)	33.25	22.7	26.84	5.7	214.74	2386
	STD(40)	46.83	32.04	37.06	11.8	245.5	2386
주체별 순매수량	individuals SMA5	−2.4E+10	1.74E+11	−1.6E+10	−1.7E+12	7.02E+11	2386
	institutions SMA5	2.8E+10	2.39E+11	−1.3E+09	−1E+12	1.96E+12	2386
	foreigners SMA5	−8.1E+09	2.1E+11	4.78E+09	−1.4E+12	8.16E+11	2386

상품가격	crude oil %Change	0	0.07	0	−3.06	0.38	2386
	gold %Change	0	0.01	0	−0.09	0.06	2386
	natural gas %Change	0	0.03	0	−0.17	0.18	2386
환율	USD/KRW %Change	0	0	0	−0.03	0.03	2386
	EUR/KRW %Change	0	0.01	0	−0.03	0.02	2386
	JPY/KRW %Change	0	0.01	0	−0.04	0.07	2386
	CNY/KRW %Change	0	0	0	−0.03	0.03	2386