랜덤 포레스트(Random Forest)를 이용한 KOSPI 방향 예측과 예측변수의 군집화 순열 중요도(Clustered Permutation Importance) 연구

서강대학교 조정효

연구동기

- 최근 금융 분야에서 인공지능 활용에 대한 관심도가 높아지며, 주가지수를 예측하는 기계학습 모형 개발 연구가 많이 진행되고 있음. 특히 예측 정확도를 높이는 모형 개발 및 변수 설정이 연구의 주를 이룸
- 기존의 전통적 통계 모형과 비교해 기계학습 모형에 대해 'black box' 문제가 제기되어 왔으나, 변수 중요도에 대한 연구 등 설명가능한 AI에 대한 관심 증가. 변수 중요도는 기계학습 모형의 훈련과 예측 과정에 예측변수들이 얼마나 기여했는지를 측정하는 방법으로 모형 개발의 변수 선정 과정에 주로 쓰임
- 본 연구는 변수 중요도 측정 방법 중 랜덤 포레스트(Breiman, 2001)를 이용한 "군집화 순열 중요도"(De Prado, 2020)를 통해 KOSPI의 움직임에 대한 예측변수 중요도를 도출하고자 함.
- 예측변수로는 과거 가격 추세, 변동성, 거래량, 환율, 상품가격 등을 사용하여 서로 다른 기간의 주가지수 방향에 대한 변수 중요도를 비교하고, 이를 통해 과거 시장 정보를 이용하여 향후 주가 지수의 움직임을 예측할 수 있는지, 그리고 어떤 변수가 예측에 기여를 많이 하며, 또한 예측 기간에 따라 중요도가 어떻게 다를지를 알아보고자 함.

선행 연구

- 과거 가격 및 시장정보를 통해 금융시장 예측
  - 효율적 시장 가설(Efficient Market Hypothesis, EMH) (Malkiel and Fama, 1970)에 따르면 과거 가격 및 시장 정보로 주가를 예측하는 것이 불가능하지만, 21세기 이후 EMH를 반박하는 연구들이 다수 등장. Jegadeesh and Titman (1993)은 과거 승자였던 주식이 패자의 주식보다 앞으로 6개월-12개월 동안 더 좋은 성과를 보이는 가격의 "momentum" 현상을 언급, Moskowitz et al. (2012)은 주가지수의 과거 가격이 지속성을 보이는 것을 이용해 **과거 시장 정보를 통해 금융시장을 예측할 수 있음**을 보임.
  - 과거 시장 정보를 담은 **기술적 분석 지표**(technical analysis indicator)를 예측변수로 하여 **기계학습**(machine learning)을 통해 금융 시장을 예측하려는 연구들이 최근 증가하고 있으며, 국내 주가지수를 예측하는 연구들 또한 진행 되어 옴(Kim, 2003; Chong et al., 2017; 이우식, 2017).
  - 특히 Ballings *et al.* (2015), Patel et al. (2015)은 앙상블(ensemble) 방법인 **랜덤포레스트** (Random Forest, 이하 RF)의 성능이 SVM, ANN 등의 다른 기법보다 뛰어남을 보임.

선행 연구

- 기계학습 중요도를 이용한 금융 예측 변수 분석
  - Nti et al (2019): RF 기반 feature selection을 통해 주가예측에 대한 거시경제 변수의 중요도 측정
  - Haq et al. (2021): RF의 순열 중요도(permutation importance)를 이용한 Feature-Ranking 방법으로 주가 추세 예측 (LR, SVM의 feature-selection과 비교)
  - 이재응, 한지형 (2021). Layer-wise Relevance Propagation (LRP)를 이용하여 KOSPI 증감에 대한 기술적 지표 및 거시경제 지표 영향 분석
- 상관관계가 있는 변수에 대한 순열중요도 연구
  - 선형 또는 비선형의 상관관계가 있는 변수의 경우 해당 변수의 순열 중요도가 낮게 편향되어 있음을 밝힘(Strobl et al., 2007; Nicodemus et al., 2010; Gregorutti et al., 2017)
  - 이를 해결하기 위해 조건부 순열 중요도 (Strobl et al., 2008; Debeer and Strobl, 2020), Max MDA (신승범 조형준, 2021), 군집화 피쳐 중요도(Clustered-feature importance) (De Prado, 2020) 등이 제안됨.
  - 금융시장 예측에 대해 "상관관계를 고려한 기계학습 기반 변수 중요도"를 측정한 연구는 거의 없었으며, 본 연구는 **RF**와 **군집화 순열 중요도(Clustered-permutation importance)**를 이용하여 기간에 따른 과거 시장 정보와 거시경제 변수의 군집 중요도를 측정하고자 함.

분석 자료

- 샘플기간
  - 훈련기간(2012년-2018년) 1904개, 테스트기간(2019년-2021년) 721개, 일별 데이터
- 목표변수(y)
  - KOSPI 일별 수정종가 기준으로 h 거래일 후 대비 증감 여부로 하며, 증가하였으면 1, 같거나 감소하였으면 0으로 하는 이중-클래스 라벨(binary-class label)로 설정.
  - 이 때 h에 대해 각각 $h = 1, 5, 20$일 때의 결과를 도출하여 서로 다른 기간의 가격 방향에 대한 예측력과 변수중요도를 비교

  - $$y_t = \begin{cases} 1, & if\ X_{t+h} - X_t > 0 \\ 0, & if\ X_{t+h} - X_t \leq 0 \end{cases} \quad h = 1, 5, 20$$ <식1>

분석 자료

- 예측변수(X) (총 30개)
  - 기술적 분석 지표 (20개): 과거 가격(시가, 저가, 고가, 종가), 거래량을 이용
    - 기술적 지표는 차트 분석가들이 주로 이용하는 지표로 과거 모멘텀 및 추세, 거래량, 변동성 등의 정보를 담고 있음 (단, 거래량지표의 경우 거래량 자체보다는 추세를 나타냄)
    - 각 기술적 지표의 계산 과정의 과거 기간(look-back window)은 이전 연구에서 주로 쓰이는 것을 사용하며, 그것의 두배 기간으로 계산한 지표를 추가하여 각 기술적 지표를 두 개씩으로 함 (10개x2)
  - 투자주체별 수급 (3개)
    - KOSPI 종목의 개인, 기관, 외국인의 순매수량
    - 이상 값을 filtering하기 위해 5일 이동 평균값 사용
  - 환율 및 상품가격 (7개)
    - 환율(원 대비 달러, 유로, 엔, 위안), 상품가격(원유(WTI), 금, 천연가스 선물)
    - 추세를 제거하기 위해 변화율 사용
  - 모든 변수에 대해 ADF 검정을 통해 시계열의 안정성(stationarity)이 확보된 변수 사용 (모든 변수 사용)
  - 훈련데이터의 분포를 이용해 변수의 크기를 표준화 스케일링(“standard scaling”) 함

분석 자료

- 예측변수(X)

구분	TA-Trend(12 개)	TA-Volume(4 개)	TA-Volatility(4 개)	주체별 순매수량(3 개)	환율(4 개)	상품가격(4 개)
변수	RSI (15), RSI (30) WR (15), WR (30) ADX (15), ADX (30) DPO (20), DPO(40) MACD (26,12), MACD (52,24) MACD Difference (26,12,9), MACD Difference (52,24,18)	FI(15) FI(30) MFI(15) MFI(30)	ATR(14) ATR(28) STD (20, STD(40)	개인(5 일 이동평균) 외국인(5 일 이동평균) 기관(5 일 이동평균)	USD/KRW 변화율 EUR/KRW 변화율 JPY/KRW 변화율 CNY/KRW 변화율	금 가격 변화율 원유 가격 변화율 천연가스 가격 변화율

표1. 예측 모형에 사용되는 예측 변수

분석 모형 – 랜덤 포레스트

- Random Forest Classifier (Breiman, 2001)
  - RF는 다수의 훈련된 의사결정나무(decision tree)를 사용하는 앙상블(ensemble) 모형으로 부트스트랩(bootstrap)을 통해 무작위로 샘플을 여러 번 추출해 결과를 집계하고, 다수결로 예측치를 도출하는 모형. 개별 의사결정나무 모형의 불안정성 및 과적합(overfitting) 문제를 보완하며, OOS의 예측 성능을 높임.
- 모형 최적화 진행
  - 훈련 기간에 대해 GridSearch(hyperparameter 후보군을 설정 한 뒤 정확도를 가장 높이는 parameter 조합을 찾는 알고리즘)를 통한 hyperparmeter tuning 진행
  - 각 목표값(1일, 5일 20일 KOSPI방향)에 따라 최적의 hyperparameter 도출

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

표2. 하이퍼파라미터 튜닝 결과

분석 모형 – 군집화 순열 중요도

- 순열 중요도(Permutation importance)
  - 학습된 기계학습 모형을 통해 중요도를 구하는 방법으로 특정한 변수 값(j)을 무작위로 재배열하여 정보를 제거 한 후 테스트 데이터에 대한 예측성능(s_j)이 재배열 전(s)에 비해 얼만큼 감소하는지를 측정. 이 때 기준의 되는 성능은 정확도, F1, AUC 등 분류기 모형의 성능을 나타내는 어느 지표라도 사용 가능하며, 순열(permutation)을 여러 번(K) 반복해 평균을 구하여 해당 변수의 중요도를 측정.

  - $PI_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j}$ <식2>

  - vs. MDI(Mean decrease impurity)
    - MDI는 RF 관련 논문에서 주로 쓰이는 중요도 측정 방식으로 학습과정에서 중요도를 계산하기 때문에 인-샘플 편향(in-sample bias)이 존재하며, 변수의 cardinality가 중요도에 영향을 미치는 한계점이 존재.

분석 모형 – 군집화 순열 중요도

- 군집화 순열 중요도 (Clustered permutation importance)
  - 변수 간의 선형 및 비선형 상관관계가 존재할 경우 해당 변수의 중요도가 낮게 나오는 문제점 (Strobl, 2007) 존재. 이를 해결하기 위한 방안으로 Clustered-feature importance(De Prado, 2020)를 이용.
  - Clustered-feature importance란 예측변수를 미리 군집화(cluster)하여 중요도를 계산하는 방법으로 학습된 모형이 해당 변수 군집(feature cluster)에 대해 무작위 재배열하여 예측한 결과를 바탕으로 변수군집에 대한 중요도를 계산하여 예측변수 간의 상관관계를 사전에 차단할 수 있음.
- 계층적 군집화(Hierarchical Clustering)
  - 예측변수의 군집화 방법으로 계층적 군집화 사용. 계층적 군집화는 변수 간의 거리가 가장 가까운 두 변수를 선택한 후 하나의 군집으로 묶고, 또 거리가 가까운 두 군집을 하나로 합치며 군집 개수를 줄여 가는 방법.
  - 본 연구에서는 변수 혹은 군집 간 거리를 Spearman 상관계수로 계산하며, 거리 행렬(distance matrix)을 이용해 군집하는 연결 기준(linkage criterion)으로 분산을 최소화하는 "Wald's criterion"을 사용함.
  - 최종 군집을 결정하는 임계점으로는 1.0을 선택 (임계점을 달리하여 원하는 군집 개수를 조정할 수 있음.)

분석 결과 – 예측성능 비교

- 이진 분류 성능 측도
  - 정확도(accuracy), F1-점수, ROC-AUC 점수 세 가지를 이용하여 성능을 측정. 모두 0과 1사이의 값으로 1에 가까울수록 예측력이 높음을 의미.

		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

  - $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
  - $F1-score = 2 \cdot \frac{precision \cdot recall}{precision+recall}$ $(precision = \frac{TP}{TP+FP}, recall = \frac{TP}{TP+FN})$
  - $ROC-AUC\ score$ = ROC(Receiver operating characteristic) 곡선의 아래 면적
    - ROC 곡선은 TPR($\frac{TP}{TP+FN}$) 와 FPR($\frac{FP}{FP+TN}$)의 관계를 그린 곡선으로, $AUC\ score$ 가 높으면 효과적으로 모형이 학습되었다고 할 수 있으며, 0.5에 가까울 수록 분류 결과가 운에 의한 것으로 해석.

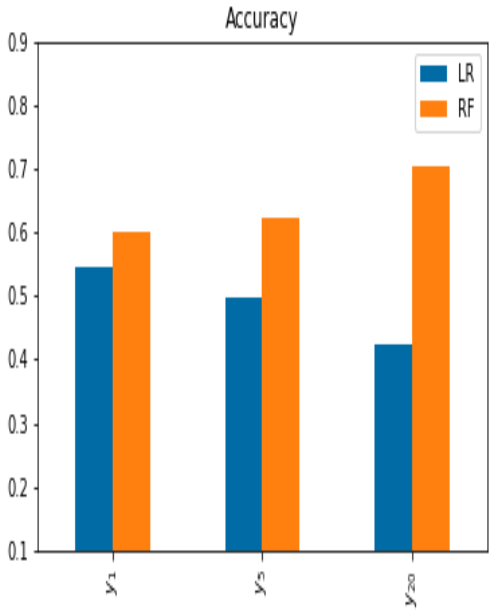
분석 결과 – 예측성능 비교

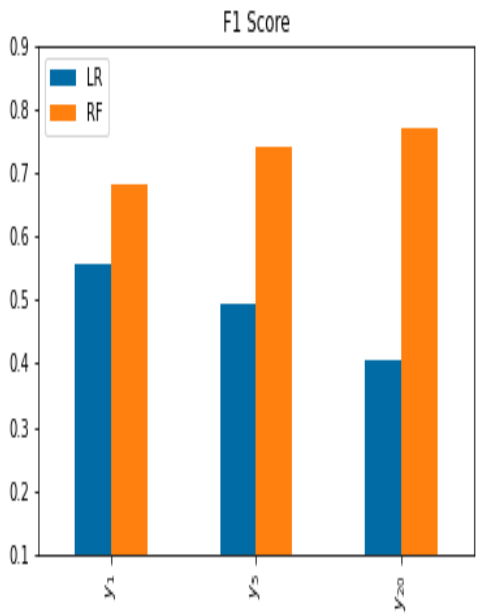
- RF의 성능을 다른 분류기 방법인 로지스틱 회귀모형(Logistic Regression)과 비교
  - LR(logistic regression)
    - LR은 일반적인 회귀모형과 마찬가지로 종속변수와 독립변수 간의 관계를 구체적인 함수로 나타내어 예측에 사용하며, 이진(binary) 종속변수에 대해 독립변수의 선형 결합을 이용하는 확률적 모형. 예측값을 [0, 1]로 하는 분류기 모형으로 고려하여 그 결과를 RF와 비교함.

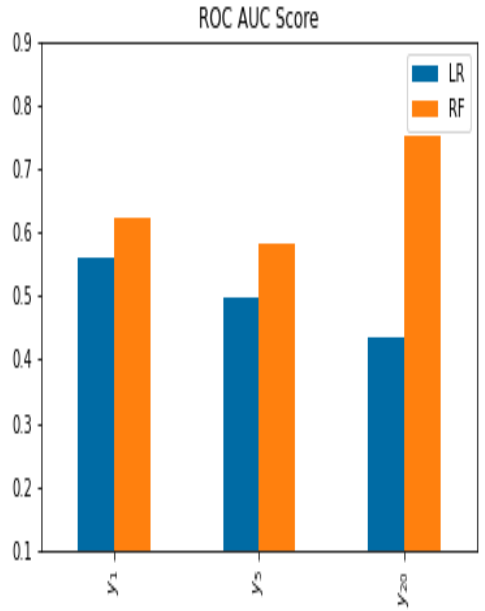
Label	Model	Accuracy	F1 score	ROC−AUC score
y_1	LR	0.5465	0.5551	0.5588
	RF	0.5936	0.6826	0.6138
y_5	LR	0.4979	0.4930	0.4978
	RF	0.6255	0.7404	0.5887
y_{20}	LR	0.4230	0.4057	0.4334
	RF	0.6963	0.7627	0.7583

표3. 각 목표변수 별 예측성능 결과 (RF, LR비교)

분석 결과 – 예측성능 비교



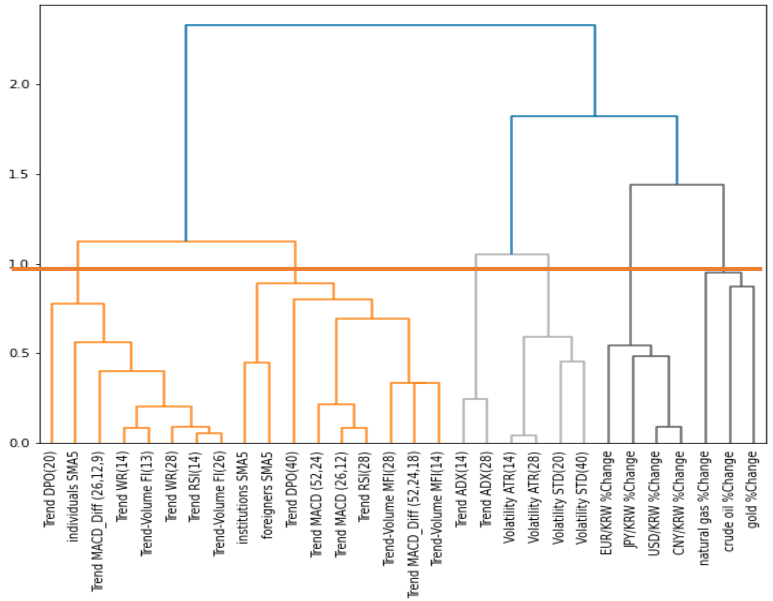




- 모든 목표변수에 대해 세 가지 성능이 모두 **RF가 LR보다 뛰어남**
- 20일, 5일, 1일 순으로 정확도와 F1 score가 높음 (즉 예측하는 **KOSPI 방향의 기간이 길수록 예측력이 높음**)
- 20일 방향에 대해서는 정확도와 F1점수가 0.7을 넘고 AUC 점수가 0.75을 넘어 강한 예측 가능성을 보임.

분석 결과 – 변수 군집화

- 예측변수의 계층적 군집화 (훈련 데이터 이용)



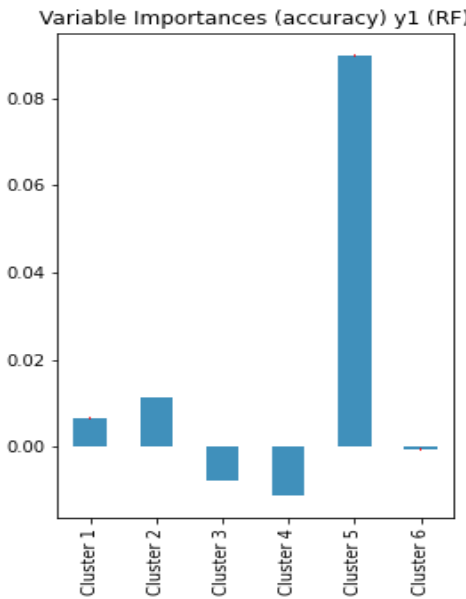
	구성 변수
군집 1	DPO(20), MACD_Diff (26,12,9), RSI(14), WR(14), WR(28), FI(13), FI(26), individuals
군집 2	DPO(40), MACD (26,12), MACD (52,24), MACD_Diff (52,24,18), RSI(28), MFI(14), MFI(28), institutions, foreigners
군집 3	ADX(14), ADX(28)
군집 4	ATR(14), ATR(28), STD(20), STD(40)
군집 5	USD/KRW %Change, EUR/KRW %Change, JPY/KRW %Change, CNY/KRW %Change
군집 6	Crude oil %Change, Gold %Change, Natural gas %Change

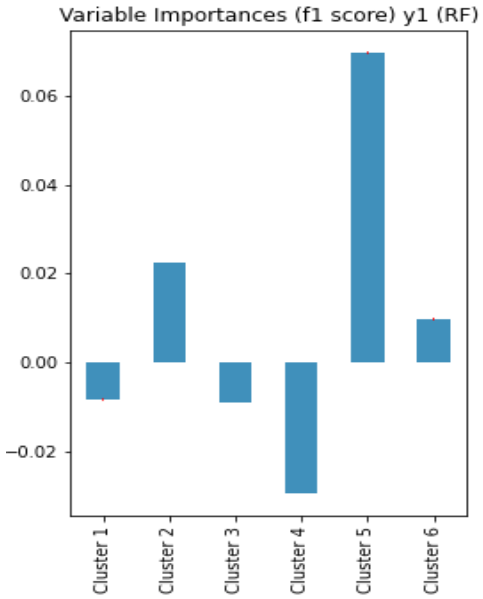
표4. 예측변수의 계층적 군집화 결과

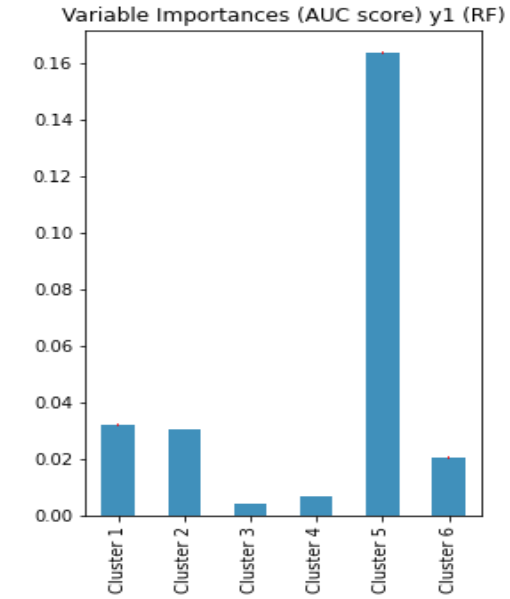
- 높은 상관관계를 보이는 변수끼리 군집화된 결과, 미리 변수의 특징에 따라 구분해 놓은 표1과 비슷하게 구성됨. 다만 투자주체별 수급은 추세 지표와 묶이고, 단기적, 장기적 과거 추세가 나뉘며, ADX지표가 독립적인 군집으로 존재함.
- 군집별 특징 정리: **군집1 (단기 가격 거래량 추세) 군집2(장기 가격 거래량 추세), 군집3(ADX), 군집4 (변동성 지표), 군집5(환율 변화율), 군집6(상품 가격 변화율)**

분석 결과 – 변수 군집 중요도

- 1일 KOSPI 방향에 대한 변수 군집(clustered-variable) 중요도



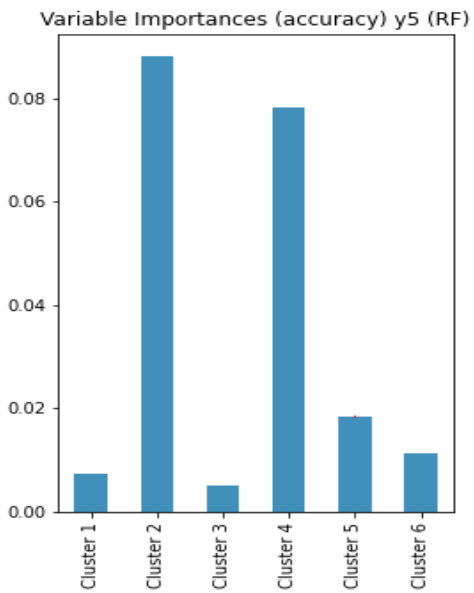


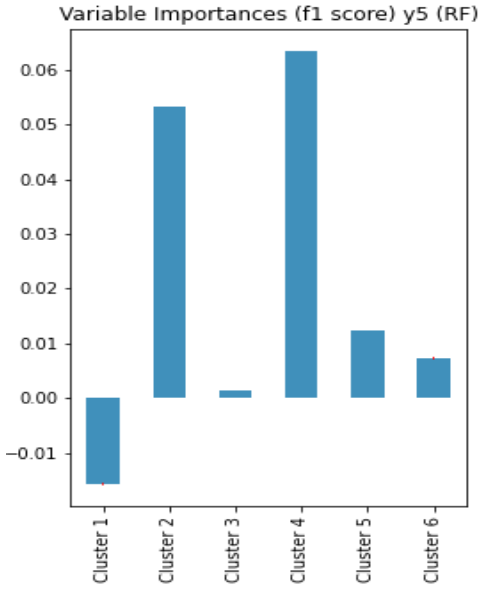


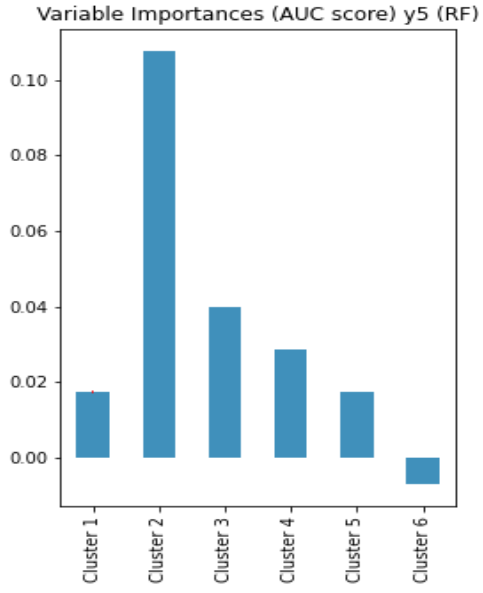
- 군집5(**환율 변화율**)의 중요도가 눈에 띄게 **높음**
- 군집4(**변동성 지표)**의 중요도는 **음수**/ 매우 **낮음**

분석 결과 – 변수 군집 중요도

- 5일 KOSPI 방향에 대한 변수 군집(clustered-variable) 중요도



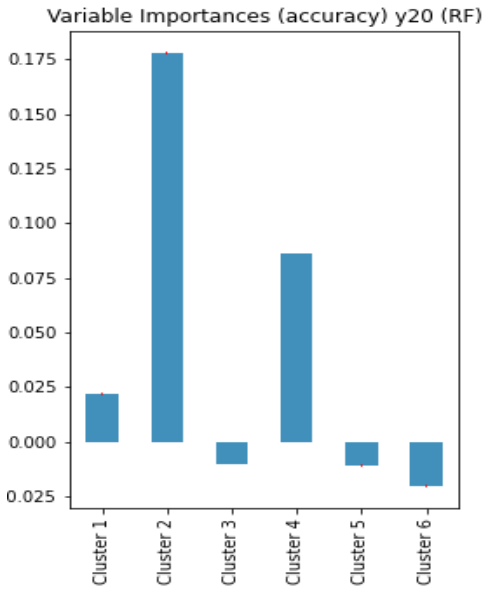


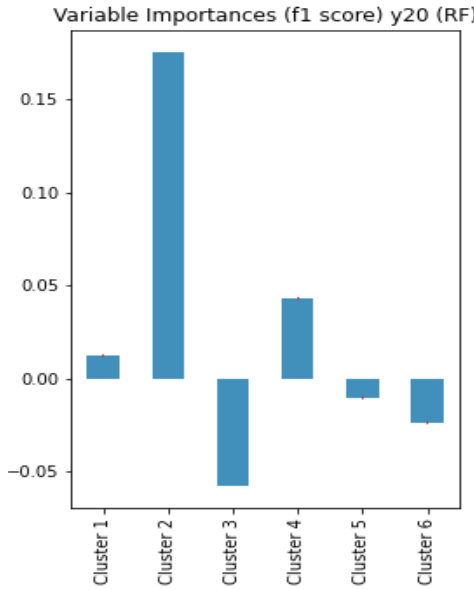


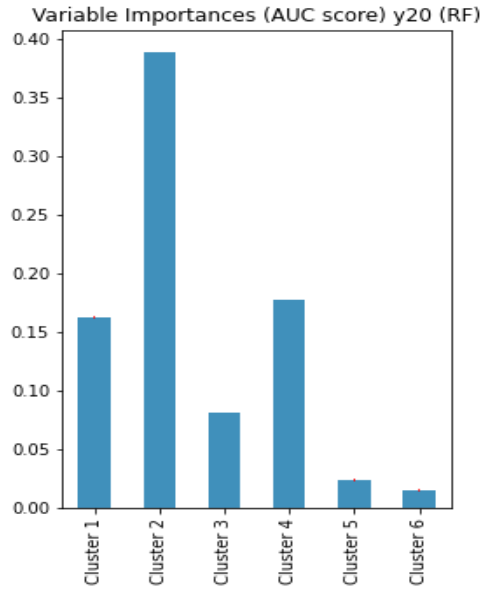
- 군집2(**장기 가격 거래량 추세**)와 군집4(**변동성 지표**)의 중요도가 **높음** (acc, f1)

분석 결과 – 변수 군집 중요도

- 20일 KOSPI 방향에 대한 변수 군집 (clustered-variable) 중요도







- 5일 방향과 마찬가지로 군집2(**장기 가격 거래량 추세**)와 군집4(**변동성 지표**)의 중요도가 **높음**
- 특히 군집2(장기 과거 추세)의 중요도(0.175)가 5일 방향 보다 **2배 이상 높음**

분석 결과 – 변수 군집 중요도

- 각 변수 군집 별 결과
  - 가격 및 거래량 추세 변수 (군집1, 2)
    - 장기적 추세가 단기적 추세보다 더 높은 중요도를 보이며, 1일 방향에 대해서는 낮은 중요도를 보이지만 5일, 20일에 대해서는 가장 높은 중요도를 보임. 특히 20일 방향에서는 정확도 기준으로 중요도가 0.175으로 5일 방향보다 2배 이상 높음
  - ADX (군집3)
    - 모든 결과에서 전체적으로 낮은 중요도를 보임
  - 변동성 변수 (군집4)
    - 1일 방향에 대해서는 음수의 중요도를 보이지만, 5일, 20일 방향에서는 과거추세 변수와 함께 중요도가 높음
  - 환율 (군집5)
    - 1일 방향에 대해 높은 중요도를 보이나, 5일 방향에 대해 비교적 낮게, 20일 방향에서는 음수의 중요도를 보임.
  - 상품가격(군집6)
    - 모든 결과에서 전체적으로 낮은 중요도를 보임

분석 결과 – 변수 군집 중요도

- 측정한 중요도에 따라 변수 선정(feature-selection)을 하고, 다시 학습시키고 예측하였을 때의 결과를 도출하여, 중요도를 검증
  - 각 목표변수에 대해 중요도(acc, f1, auc의 평균)가 높은 세 개 변수 군집, 낮은 세 개 변수 군집, 전체 변수를 예측변수로 사용한 모형의 예측 성능을 비교

Label	Variables	Accuracy	F1 score	ROC-AUC score
y_1	All	0.5936	0.6826	0.6138
	Top 3 Clusters	0.5964	0.6659	0.6117
	Bottom 3 Clusters	0.5576	0.6708	0.5403
y_5	All	0.6255	0.7404	0.5887
	Top 3 Clusters	0.6186	0.7363	0.5754
	Bottom 3 Clusters	0.5603	0.6755	0.5553
y_{20}	All	0.6963	0.7627	0.7583
	Top 3 Clusters	0.7060	0.7720	0.7838
	Bottom 3 Clusters	0.6755	0.7310	0.7283

표5. 중요도에 따라 변수 선정 후 예측 성능 비교

  - 상위 중요도 변수(top 3 clusters variables)를 사용하였을 때 성능은 전체 변수를 사용할 때와 비교해 유사.
  - 하위 중요도 변수(bottom 3 clusters variables)를 사용하였을 때 성능은 전체 변수를 사용할 때와 비교해 많이 낮음.
  - → 중요하지 않은 변수가 포함된다고 성능이 떨어지지 않을 수 있지만, 중요한 변수가 포함되지 않으면 성능은 떨어짐. 이는 다수의 변수를 사용해도 되는 앙상블의 특징과 부합함.

결론

- KOSPI 방향 예측에 RF가 LR보다 눈에 띄게 높은 성능을 보이며 20일, 5일, 1일 순으로 (기간이 길수록) 정확도가 높다.
- 1일 방향에 대해서는 환율 변수가 높은 중요도를 보인다. 하지만 1일 방향 예측모형은 성능이 낮기(50%대) 때문에 중요도가 무의미할 수 있다.
- 5일, 20일 방향에 대해서는 가격 및 거래량 추세와 변동성 변수가 높은 중요도를 보이며, 특히 20일 방향에서 가격 및 거래량 추세의 중요도가 매우 높다. 이러한 결과를 통해 과거 가격 및 거래량의 추세가 한달 이상의 미래 주가지수의 움직임을 예측하는 정보를 담고 있다고 볼 수 있으며, 이는 공공 시장 정보를 이용하여 금융시장을 예측할 수 있다는 이론을 뒷받침한다.

References

- 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.
- 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.
- Chong, E., Han, C., & Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. Expert Systems with Applications, 83, 187-205.
- 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.

References

- 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.