KOSPI 방향 예측에 대한 랜덤포레스트의 군집화 변수중요도 분석

서강대학교 조정효

연구동기 및 연구목표

- 최근 금융 분야에서 인공지능 활용에 대한 관심도가 높아지며, 주가지수를 예측하는 기계학습 모형 개발 연구가 많이 진행되고 있음. 예측 정확도를 높이는 모형 개발 및 변수 설정이 연구의 주를 이룬 반면 **예측변수를 해석하는 방법에 대한 연구**는 비교적 적게 이루어짐.
- 기존의 전통적 통계 모형과 비교해 기계학습 모형에 대해 'black box' 문제가 제기되어 왔으나, 변수중요도에 대한 연구 등 설명가능한 AI에 대한 관심이 증가함. **변수중요도**는 기계학습 모형의 훈련과 예측 과정에 예측변수들이 얼마나 기여했는지를 측정하는 방법.
- 금융시장 예측의 변수중요도 연구는 많이 이루어지지 않았으며, 특히 변수 간 상관관계가 있을 때 중요도가 하향 편향되는 등 기존의 변수중요도 기법의 한계점을 고려하지 않은 방법론을 사용한 것이 대부분임.
- 본 연구는 변수중요도 측정 방법 중 랜덤 포레스트(Breiman, 2001)를 이용한 순열중요도(permutation importance) 기법을 기반으로 한 "군집화 순열 중요도"(De Prado, 2020)를 통해 KOSPI 방향 예측에 대한 변수중요도를 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 변수의 중요도가 어떻게 달라지는지를 알아보고자 함.

선행 연구

- 과거 가격 및 시장정보를 통해 금융시장 예측
  - 효율적 시장 가설(Efficient Market Hypothesis, EMH) (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.

분석 자료

- 샘플기간
  - 훈련기간(2012년-2018년) 1665개, 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

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

분석 자료

- 예측변수(X) (총 28개)
  - 기술적 분석 지표 (18개): 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 변수는 ADF 검정을 통해 시계열의 안정성(stationarity)이 확보된 변수
  - 훈련데이터의 분포를 이용해 변수의 크기를 표준화 스케일링(“standard scaling”) 하여 모형이 효과적으로 학습하도록 함

분석 자료

- 예측변수(X)

[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* *foreigners* *institutions*	*USD/KRW*, *EUR/KRW* , *JPY/KRW*, *CNY/KRW*	*gold*, *crude oil*, *natural gas*

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

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

[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

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.

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

  - 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.

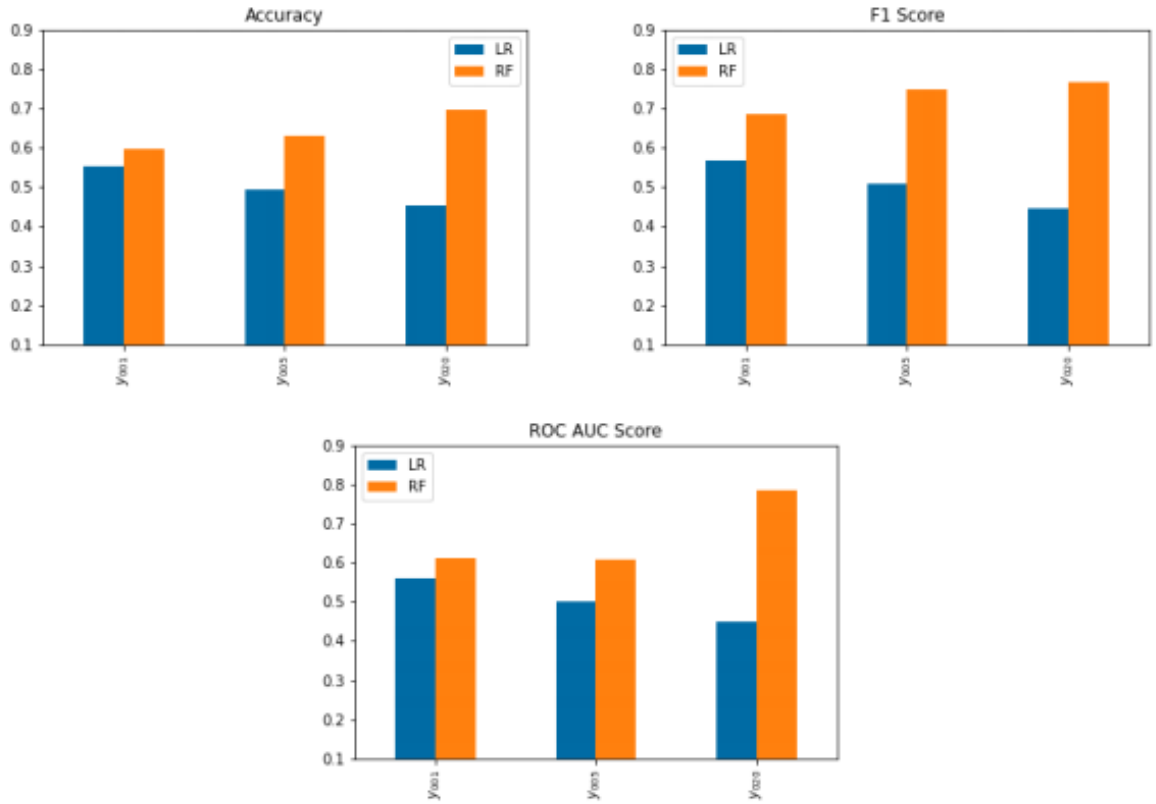
[표 6] 예측 성능 결과 비교

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 등락을 의미한다.

분석 결과 – 예측성능 비교

[그림 3] 예측 성능 결과 비교

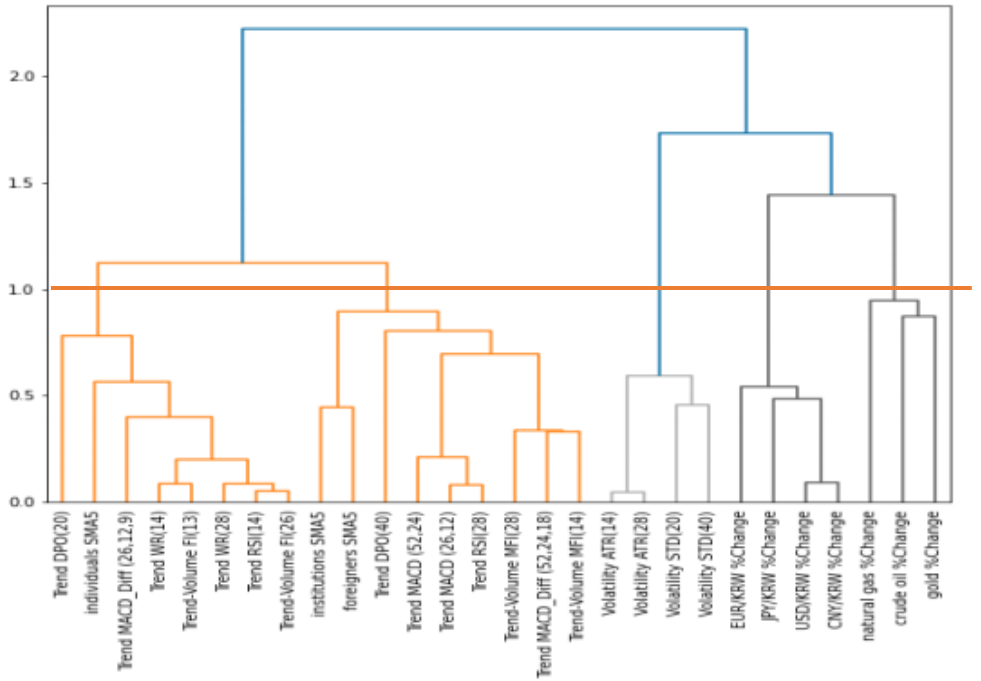


- 1. 모든 목표변수에 대해 세 가지 성능이 모두 **RF가 LR보다 뛰어남 → 예측력이 높은 RF 모형을 기반으로 중요도를 구하는 것은 의미 있음**
- 20일, 5일, 1일 순으로 점수가 높음 (즉 예측하는 **KOSPI 방향의 기간이 길수록 예측력이 높음**)

분석 결과 – 변수 군집화

- 예측변수의 계층적 군집화 (훈련 데이터 이용)

[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.

분석 결과 – 변수 군집화

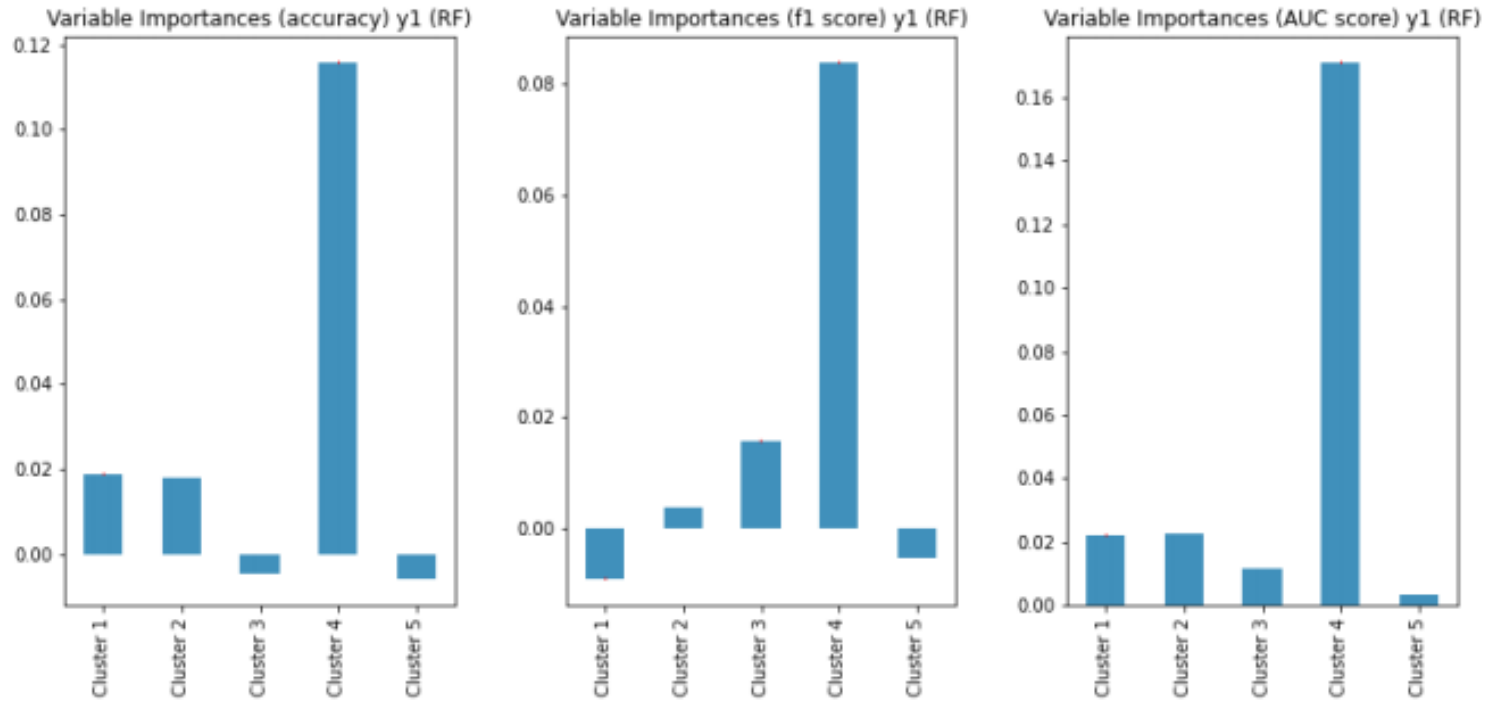
- 예측변수의 계층적 군집화 (훈련 데이터 이용)

[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

분석 결과 – 변수 군집 중요도

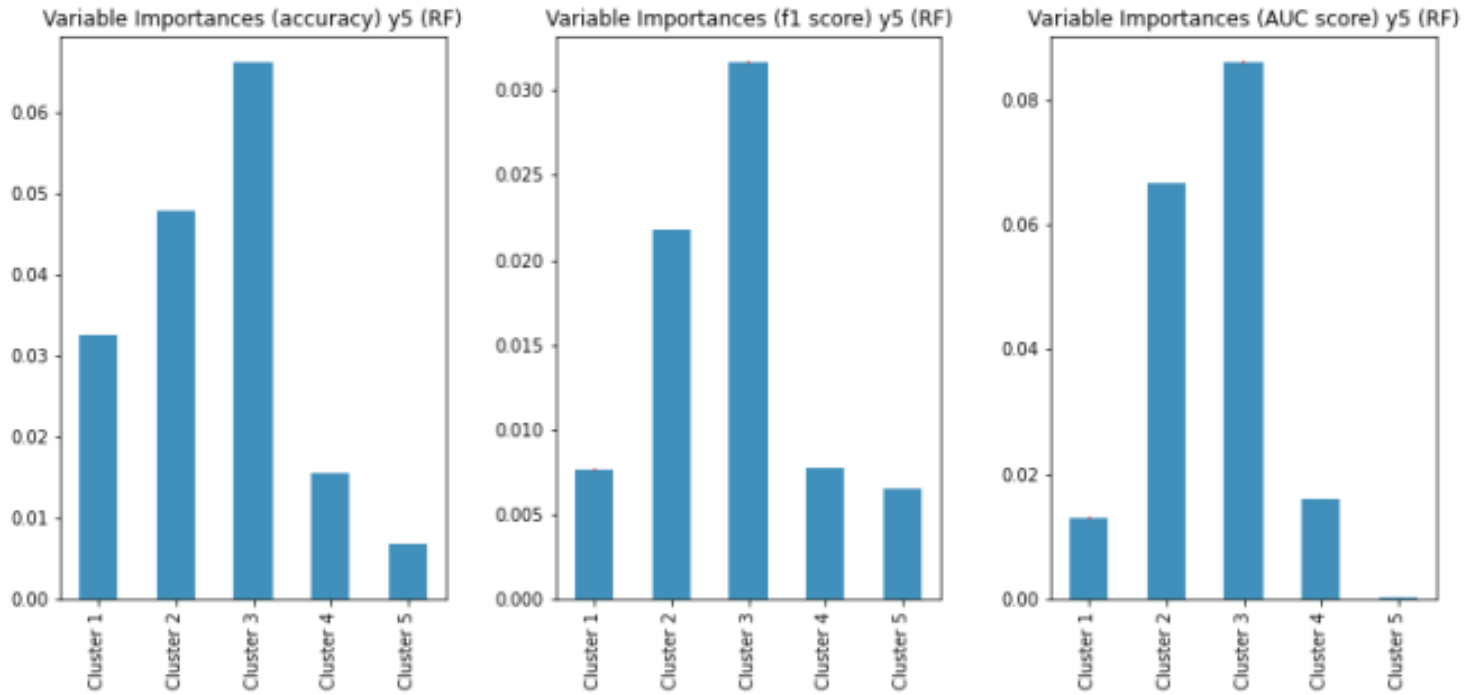
[그림 4] 1 일 KOSPI 방향에 대한 변수 군집 중요도



- 군집4(**환율**)의 중요도가 **높음**

분석 결과 – 변수 군집 중요도

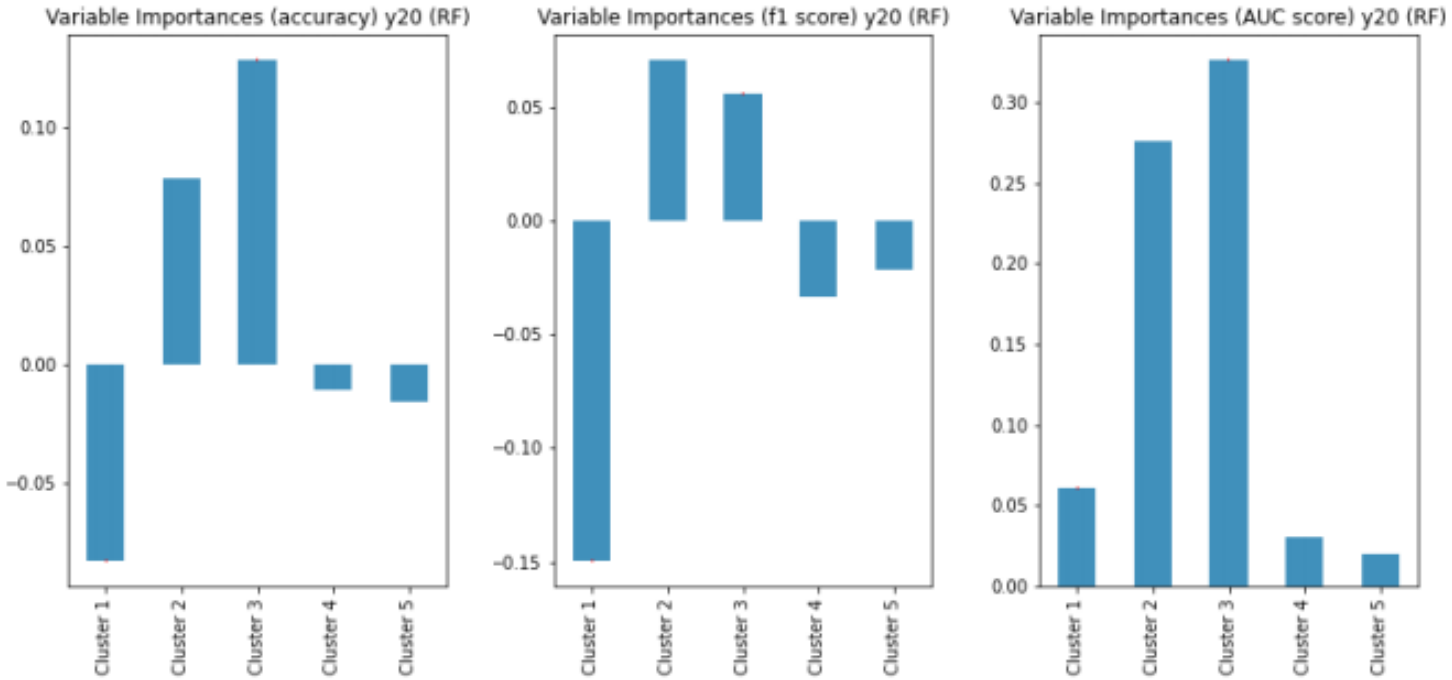
[그림 5] 5 일 KOSPI 방향에 대한 변수 군집 중요도



- 군집2(**장기 시장 추세**)와 군집3(**변동성 지표**)의 중요도가 **높음**

분석 결과 – 변수 군집 중요도

[그림 6] 20 일 KOSPI 방향에 대한 변수 군집 중요도



- 5일 방향과 마찬가지로 군집2(**장기 시장 추세**)와 군집3(**변동성 지표**)의 중요도가 **높음**
- 반면 군집1**(단기 시장 추세)**의 중요도는 **낮음**

결론

- 주식시장 예측을 위한 기계학습 모형의 활용을 넘어 예측에 사용되는 변수들의 중요도를 측정하는 방법을 연구하여 설명가능한 기계학습 모형을 구축함.
- 기존의 국내 주식시장에 대한 기계학습 모형 예측에 대한 연구는 주로 예측력에 초점을 맞추어 이루어져, 변수에 대한 분석은 활발히 진행되지 않음. 특히 변수간 상관관계가 있음에도 이를 고려하지 않은 채 중요도를 도출. 본 연구는 이러한 점을 보완한 방법인 군집화 변수중요도를 이용하여 변수에 대해 분석한다는 점에서 의의가 있음.
- 실증분석을 통해 랜덤포레스트를 사용한 KOSPI 방향 예측이 가능함을 보임. 특히 기간이 길수록 예측 성능이 높아지는 것을 확인함. 해당 예측모형을 사용해 변수중요도를 구하는 것은 의미 있음을 보임.
- 변수군집에 대한 중요도 분석 결과, 1일 KOSPI 방향에 대해서는 환율 변수가 큰 중요도를 보였으며, 5일, 20일 KOSPI 방향에 대해서는 장기적 시장 추세와 시장 변동성 변수가 큰 중요도를 보임. 특히 예측모형의 성능이 좋았던 5일, 20일 KOSPI 방향 예측에 하는 데 있어서 거시 변수보다 과거 시장 정보가 더 많은 기여를 함을 확인함.

참고문헌

- 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.

참고문헌

- 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.
- Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 25(2), 383-417.
- 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.
- 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.

참고문헌

- 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.