석사 과정 논문 Proposal

조정효

연구 제목(가제): = 기계학습 분류기의 KOSPI 방향 예측과 예측변수의 순열 특징 중요도 연구

0. 요약

본 연구는 SVM, Random Forest 과 같은 기계학습 분류기를 이용하여 코스피 지수의 방향을 예측하고, 예측에 사용되는 변수들의 중요도를 측정하고자 한다. 예측을 위한 정답값으로 각각 1 일, 5 일, 20 일 후의 KOSPI 방향에 대해서 상승(1) 또는 하락(0)으로 설정한다. 정답값을 예측하는데 필요한 예측변수로는 과거 가격 및 거래량 자료를 바탕으로 추출한 기술적 분석 지표, 투자주체별 순매수량 그리고 거시경제 지표로서 환율과 상품가격을 사용한다. 표본 기간은 2012 년 1 월부터 2021 년 12 월까지의 10 년간 일별 데이터에 대해 훈련(2012-2018; 7 년), 테스트(2019-2021; 3 년)로 나누었으며, 서로 다른 분류기와 서로 다른 기간의 KOSPI 방향에 대한 결과를 비교한다. 다음으로 예측 변수의 중요도를 측정하기 위해 순열 특징 중요도(Permutation feature importance)를 사용하여 예측변수의 예측중요도를 분석한다. 이 때, 예측변수 간의 상관성이 유의미한 중요도를 측정하는데 해를 끼칠 수 있으므로, 예측변수 간 상관성을 미리 제거하는 방법으로 계층적 군집화(Hierarchical clustering)를 사용하여 군집별 대표 변수에 대한 중요도 분석을 진행한다.

1. 선행 연구

1.1. 기계학습을 사용한 금융시장 예측 연구

- 기술적 지표, ANN, SVM, RF 등 이용 (Kim, 2003; Kara et al.,2011; Patel et al.,2015)
- 향후 추가

1.2. 예측변수 중요도 및 피쳐 선정(feature selection) 연구

- RF 기반 Permutation Importance 를 이용해 단기 전기량 예측에 대한 Feature selection 연구 (Huang et al., 2016)
- RF 를 이용해 거시경제 변수의 중요도를 측정 (Nti et al., 2019)
- LRP(Layer-wise Relevance Propagation)를 이용한 KOPSI 증감에 대한 기술적 지표 및 거시경제 지표 등 설명변수 분석 (이재응, 한지현, 2021)
- LR, RF, SVM 를 이용한 Feature-Ranking 방법으로 피쳐 선정 후 일별주가 트렌드 예측 (Haq et al, 2021)
- 다중공선성을 가지는 변수의 Permutation Importance 분석의 한계 지적 (Gregorutti et al., 2017; Drobnič et al, 2020)
- 향후 추가

2. 자료

2.1. 정답값

예측모형이 예측하고자 목표변수에 대해 KOSPI 일별 수정종가 기준으로 h 일 후 대비 증감 여부를 정답값으로 한다. 증가하였으면 1, 감소하였으면 0 으로 하는 이중-클래스(binary-class) 라벨로 설정한다. 이 때 h=1, 5, 20 으로 하여 서로 다른 기간에 대한 가격 방향에 대한 예측력과 변수중요도를 비교한다. 샘플 거래일수는 2386 이며, 2012 년 1 월부터 2021 년 12 월까지의 KOSPI 자료를 사용한다. 2012 년 1 월부터 2018 년 12 월까지를 훈련 데이터로 하고 2019 년 1 월부터 2021 년 12 월까지를 테스트 데이터로 한다.

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

2.2. 예측변수

사용하는 예측변수 데이터로는 기술적 분석 지표, 투자주체별 수급, 거시경제 변수가 있다. 기술적 분석 지표는 과거 가격과 거래량을 통하여 계산되는 지표로 금융자산의 차트 분석에 사용된다. 본 연구에서는 Kara et al.(2011)에서 쓰인 지표를 참고하였으며, 시계열 정상성(stationary)을 가지는 변수만을 사용하고자 Augmented Dickey-Fuller Test 를 통해 p 값이 0.05 보다 작은 지표 중 총 10 개의 변수를 선정했으며 각 지표를 구하기 위해 사용되는 window size 를 각각 1 배, 2 배 크게 해 계산하여 총 20 개의 기술적 지표 변수를 생성했다. 기술적 분석 지표는 크게 추세, 거래량, 변동성을 나타내는 지표로 나눌 수 있다. 단, 거래량 지표는 거래량만을 나타내기 보다 과거 거래량을 이용하여 추세를 나타내는 지표로 구성되어 있다.

- 추세 지표:
  - RSI(Relative Strength Index) 14(28), 과거 14(28)일 동안의 종가의 상승 변화율과 하락 변화율의 상대적인 강도를 나타내는 트렌드 지표
  - WR(William's R) 14(28): 과거 14(28)일 동안 최고가-현재 종가 폭과 최고가-최저가 폭을 비교하는 트렌드 지표
  - ADX(Average Directional Movement Index) 14(28): 과거 14(28)일 간 가격 방향 지표(DI)를 활용하여 14일 이동 평균화한 트렌드 지표
  - DPO(Detrended Price Oscillator) 20(40): 고점에서 고점 또는 저점에서 저점으로 이어지는 가격 주기 기간을 추정하여 20(40)일 동안 고점과 저점 사이 거리를 측정하는 지표
  - MACD(Moving Average Convergence Divergence) 26(52), 12(24): 종가의 26(52)일 EMA(exponential moving average)와 12(24)일 EMA의 차이로 트렌드를 나타내는 지표
  - MACD Difference 26(52), 12(24), 9(18): MACD 26(52),12(24)와 이를 9(18)일 이동평균한 값의 차이로 트렌드를 나타내는 지표
- 거래량 지표:
  - FI (Force Index) 15(30): 종가 변화율과 거래량의 곱을 15(30)일 이동평균한 지표로 거래량 기반 트렌드 지표
  - MFI(Money Flow Index) 15(30): 가격과 거래량을 이용한 15(30)일 간 상승 압력과 하락 압력의 상대적 강도를 나타내는 거래량 가중의 트렌드 지표
- 변동성 지표:

- ATR(Average True Range) 10(20): 각각 현고가와 현저가, 현고가와 전종가, 현저가와 전종가의 차이 중 최댓값을 True Range로 하여 이를 10(20)일 기준으로 평활화한 지표로 변동성을 나타내는 지표
- STD(Rolling Standard Deviation) 20(40): 과거 20(40)일 간 종가를 이용하여 이동 표준편차를 구하여 변동성을 측정한 지표

주체별 수급으로는 개인, 기관계, 외국인의 순매수량을 사용했으며, 데이터의 노이즈를 감소시키고자 각각에 대해 5 일 이동평균화하여 총 3 개의 변수를 생성했다.

- Individuals (5): 개인 순매수량의 5 일 이동평균
- Institutions (5): 기관 순매수량의 5 일 이동평균
- Foreigners (5): 외국인 순매수량의 5 일 이동평균

다음으로 거시경제 변수로는 환율과 상품가격을 사용했다. 환율로는 원 기준 달러, 유로, 엔, 위안을, 상품 가격으로는 국제 금, 원유, 천연 가스의 선물 가격을 사용하였다. 또한 각 거시경제 변수에 대한 추세성을 없애고자 변화율을 계산해 총 7 개의 변수를 생성했다.

- 환율:
  - USD/KRW : 일별 원-달러
  - EUR/KRW : 일별 원-유
  - JPY/KRW : 일별 원-엔
  - CNY/KRW : 일별 원-위안
- 상품가격:
  - gold: COMEX 기준 국제 금 2022 년 4 월 선물 연결가격
  - crude oil: NYMEX 기준 서부텍사스유(WTI) 2022 년 4 월 선물 연결가격
  - natural gas: NYMEX 기준 천연가스 2022 년 4 월 선물 연결가격

구분	TA-Trend(12 개)	TA-Volume(4 개)	TA-Volatility(4 개)	주체별 순매수량(3 개)	환율(4 개)	상품가격(4 개)
Features	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

표 1. 예측 모형에 사용되는 예측 변수

모든 예측변수에 대해 표준화(Standard Scaler)를 사용하여 훈련기간의 데이터를 fit 한 후 훈련기간 및 테스트기간의 데이터에 대해 transform 한 최종 예측변수를 사용한다.

자료 출처

- KOSPI 일별 시가, 고가, 저가, 종가, 거래량 (기술적 지표 계산에 사용됨): 한국거래소(KRX) 스크래핑 – 파이썬 API *pykrx* (*https://github.com/sharebook-kr/pykrx*)
- KOSPI 개인, 기관계, 외국인 순매수량: 한국거래소(KRX) 스크래핑 – 파이썬 API *pykrx* (*https://github.com/sharebook-kr/pykrx*)
- 환율 (원 기준 달러, 유로, 엔, 위안): 네이버 금융 스크래핑 – 파이썬 API *FinanceDataReader* (*https://github.com/FinanceData/FinanceDataReader*)
- 상품 가격: 국제 금(COMEX 연결선물 22-04), WTI(NYMEX 연결선물 22-04), 천연가스(NYMEX 연결선물 22-04) – 파이썬 API *yfinance* (*https://github.com/ranaroussi/yfinance*)

3. 예측 모형

예측을 위해 기계학습 분류기 모형을 사용하여 비교, 분석한다. 분류기 기법으로는 SVM 과 Random forest 를 사용하며, 로지스틱 회귀모형과 성능을 비교하여 머신러닝 모형의 예측력이 기존의 전통적인 계량 모형보다 높은 가를 알아본다.

각 분류기 모형의 주요 하이퍼파라미터(hyperparameter)는 다음의 표와 같다.

모형	SVM	Random Forest
하이퍼파라미터	$C = [1]$ $Kernel = [RBF]$	$n\_estimators = 1000$ $max\_depth = 7$

표 2. 예측 모형 분류기의 주요 하이퍼파라미터

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 한다. 예측 정답값은 코스피 지수의 1 일, 5 일, 20 일(거래일) 후 방향으로 총 3 개 설정하여 결과를 비교한다. 예측변수로는 앞서 설명한 40 개의 생성된 변수 혹은 피쳐(feature)를 사용한다.

4. 예측변수 중요도 분석

4.1. 순열 특징 중요도(Permutation feature importance)

기계학습을 이용한 금융 시장의 예측에 대한 연구로 정확도를 높이는 연구가 주를 이루었다. 본 연구에서는 정확도를 높이는 목적에서 더 나아가서 예측 결과에 영향을 끼친 예측변수에 대한 분석을 통해 기계학습 모형이 가지는 'black box' 문제를 완화하고자 한다. 예측변수의 분석을 위해 각 변수의 예측변수 중요도를 구하여 순위를 매겨 각 정답값(1 일, 5 일, 20 일 방향)을 예측하는데 어떠한 변수가 많은 기여를 했는지 알아본다.

중요도 측정을 위해 순열 특징 중요도(Permutation feature importance)를 사용한다. 순열 특징 중요도란 학습된 모형을 통해 중요도를 구하는 방법으로 하나의 예측변수 값이 무작위로 섞였을 때 점수(score)가 얼만큼 감소하는지를 측정하는 방법이다. 측정된 감소치는 모형이 해당 예측변수에 얼만큼 의존하는지를 나타낸다. 이 중요도 방법은 사용할 모형과 측정하고자 하는 점수를 자유롭게 설정할 수 있으며, 예측변수 값의 순열(permutation) 여러 번 반복하여 평균을

구할 수 있다. 본 연구에서는 SVM, Random Forest 를 이용하여 정확도의 감소치를 기준으로 중요도를 측정한다.

4.2. 변수 간 상관성 문제와 변수 군집화

변수의 상관성 문제: 변수 간의 다중공선성 및 상관성을 가지면 중요도를 구하는데 문제가 생긴다. 이를 해결하기 위한 방안 중 변수를 미리 군집화하여 무작위로 군집내의 대표 변수를 선정하여 상호 상관성이 없는 변수 하위집합(subset)만을 중요도 측정에 사용한다. (de Prado, 2020)

본 연구에서는 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. 본 연구에서는 거리를 계산하는 연결 기준(linkage criterion)으로 최소 분산을 계산하는 Wald' s criterion 을 사용하며, Spearman 상관계수를 거리 측도로 하여 선형적 관계를 고려한다.

5. 실증 분석 결과

5.1. 예측 성능

예측 모형의 성과를 나타내기 위해 사용되는 지표로 정확도(Accuracy), F1 점수, ROC-AUC 점수를 사용한다. 세 점수 모두 0과 1사이 값으로 1에 가까울수록 모형이 성능이 뛰어나다고 할 수 있다.

- 정확도 (Accuracy): 전체 예측 결과 중 실제와 예측 결과가 일치하는 경우의 비율
- F1 점수: 정밀도(precision)와 재현율(recall)의 조화평균
- ROC-AUC 점수: ROC 곡선의 아래 영역

Label	Methods	Accuracy	F1	ROC AUC
y1	Logistic Regression	0.55	0.59	0.56
	SVM	0.63	0.73	0.65
	RF	0.64	0.71	0.71
y5	Logistic Regression	0.49	0.49	0.49
	SVM	0.65	0.74	0.64
	RF	0.65	0.74	0.67
y20	Logistic Regression	0.42	0.40	0.43
	SVM	0.65	0.74	0.65
	RF	0.64	0.72	0.64

표 3. 예측 성능 결과 비교

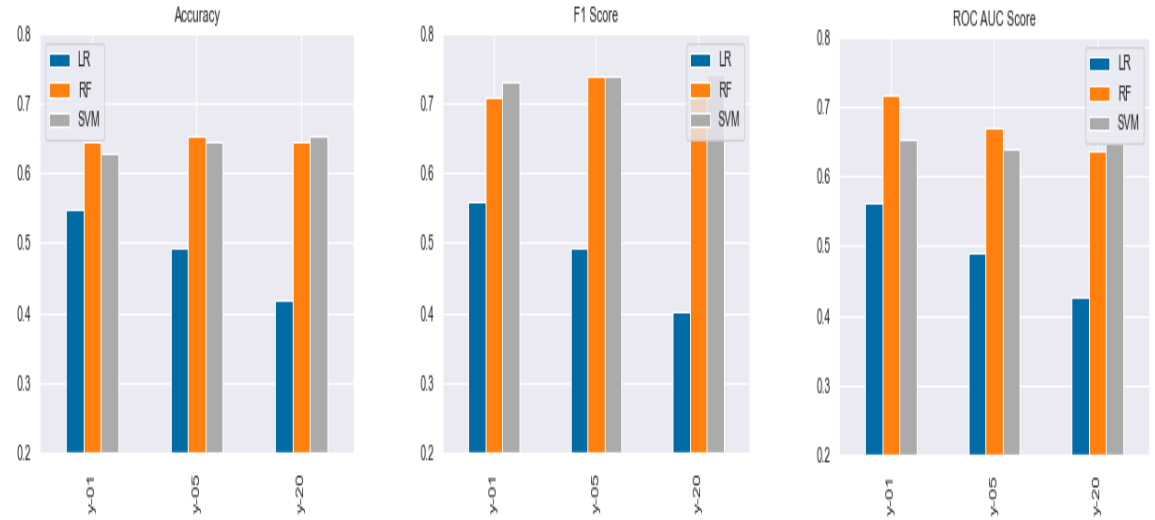

그림 1. 예측 성능 결과 비교

5.2. 변수 중요도 분석

다음으로 40 개의 예측변수를 계층적 군집화를 통해 군집화하며, 이 때 거리 유사도의 임계값을 0.5로 설정한다. 결과적으로 14개의 군집이 생성되었다. 생성된 군집 내에 대표 변수를 파이썬 라이브러리 *numpy* 를 이용하여 무작위로 선정한다. 단, 군집 내에 변수가 하나이면 그 변수를 대표 변수로 한다.

	대표 변수	나머지 변수
군집 1	individuals SMA5	foreigners SMA5
군집 2	institutions SMA5	
군집 3	MACD (52,24)	MACD (26,12), RSI(28)
군집 4	MFI(28)	MFI(14), MACD_Diff (52,24,18)
군집 5	RSI(14)	MACD_Diff (26,12,9), WR(14), WR(28), FI(13), FI(26)
군집 6	DPO(20)	
군집 7	DPO(40)	
군집 8	ADX(28)	ADX(14)
군집 9	ATR(28)	ATR(14), STD(20), STD(40)
군집 10	CNY/KRW %Change	USD/KRW %Change
군집 11	EUR/KRW %Change	JPY/KRW %Change
군집 12	crude oil %Change	
군집 13	natural gas %Change	
군집 14	gold %Change	

표 4. 예측변수에 대한 계층적 군집화 결과

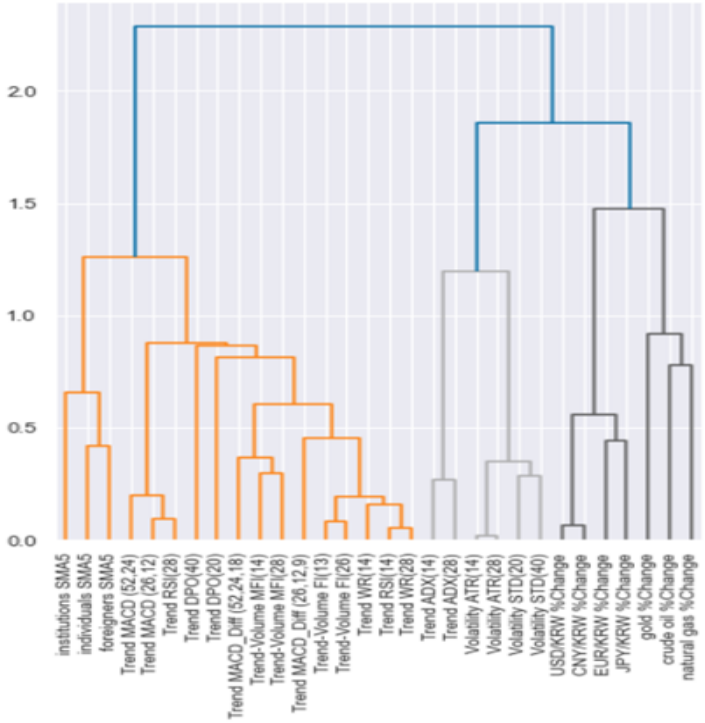

그림 2. 예측변수 간 유사도를 나타낸 덴도그램(Dendogram). 본 연구에서는 0.5 거리 기준으로 군집화했다.

각 군집의 특징:

- 군집 1: 개인, 외국인 수급
- 군집 2: 기관계 수급
- 군집 3: 장기 모멘텀 지표
- 군집 4: 중기 모멘텀 지표
- 군집 5: 단기 모멘텀 지표
- 군집 6, 7: DPO 추세지표 각각 20 일, 40 일
- 군집 8: ADX 추세지표
- 군집 9: 변동성 지표
- 군집 10: 달러, 위안화 환율 변화율
- 군집 11: 유로, 엔화 환율 변화율
- 군집 12,13,14: 각각 원유, 천연가스, 금 가격 변화율

다음으로 14 개의 대표 변수에 대해 테스트 구간에 대한 순열 중요도를 구한 결과를 살펴본다. RF, SVM 을 이용하여 각각 1 일, 5 일, 20 일 후 KOSPI 방향에 대한 예측 시 정확도 기준으로 중요도를 도출한다.

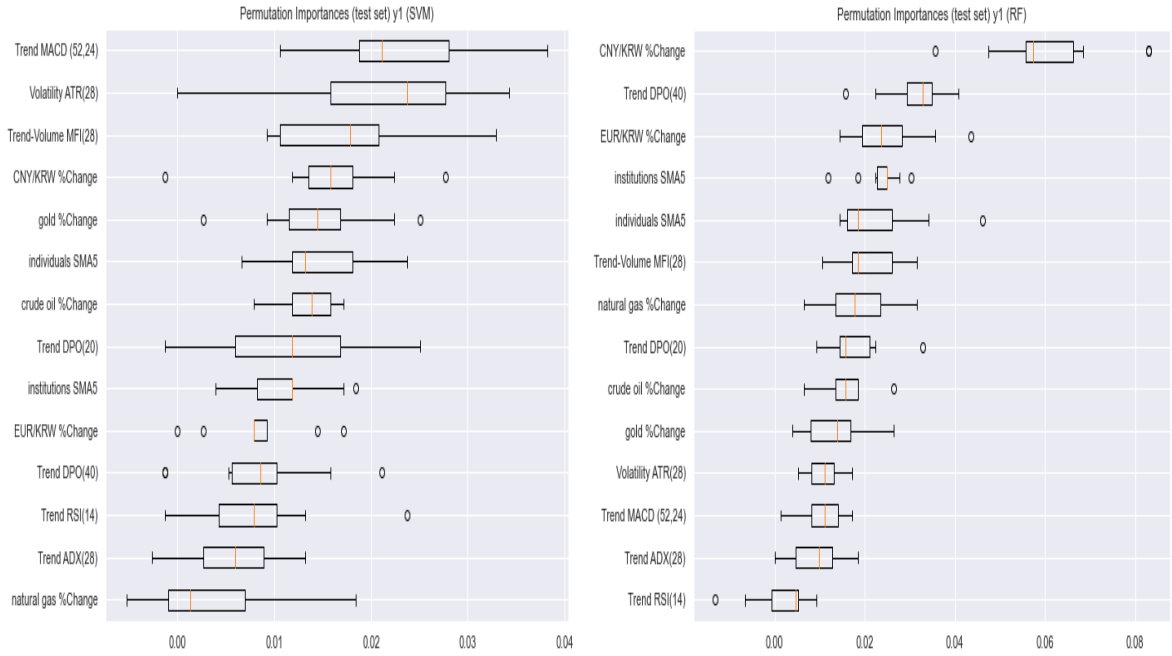

그림 3.1. 1 일 KOSPI 방향에 대한 변수 중요도

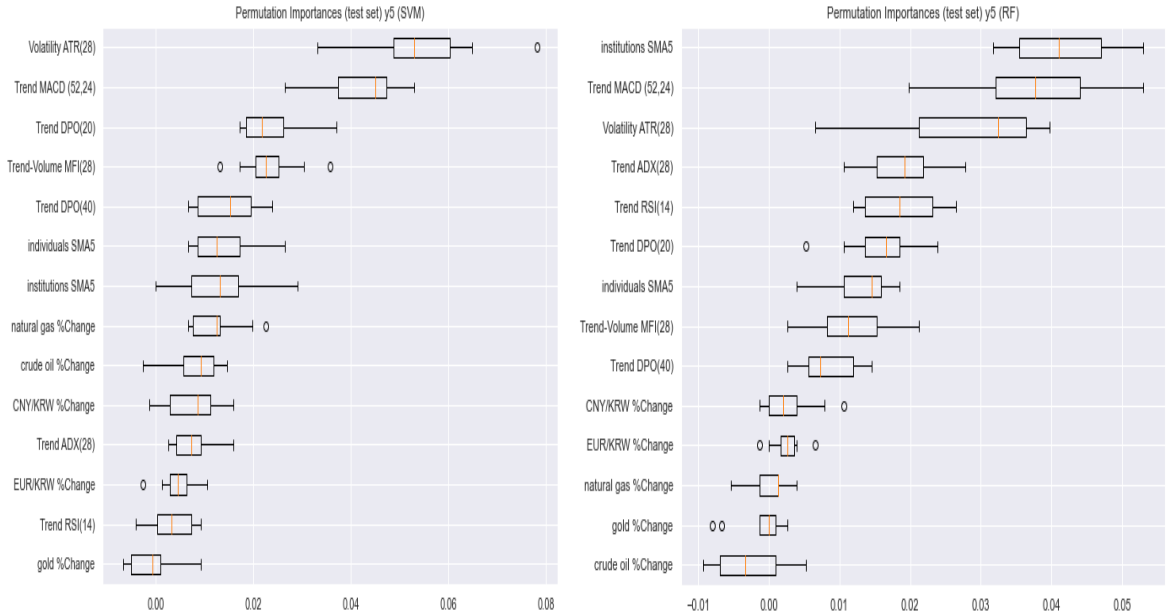

그림 3.2. 5 일 KOSPI 방향에 대한 변수 중요도

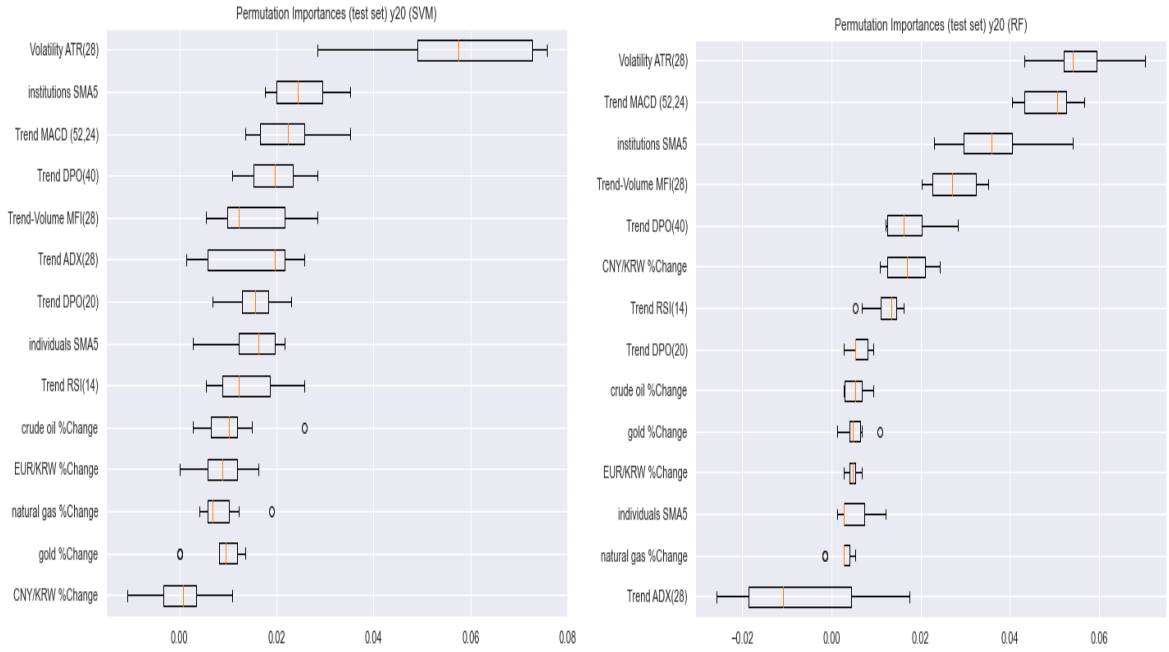

그림 3.3. 20 일 KOSPI 방향에 대한 변수 중요도

5 일 방향, 20 일 방향에 대해서는 두 모형 모두 변동성 지표(ATR(28))와 중장기 모멘텀 지표들(MACD(52,24), DPO, MFI(28)) 그리고 기관계 수급이 높은 중요도를 보이며, 거시경제 변수와 단기 모멘텀 지표(RSI(14))는 상대적 낮은 중요도를 보인다. 특히 20 일 방향에서는 거시경제 변수는 0 에 가까운 중요도를 가진다. 단 거시경제 변수는 환율 및 상품가격을 차분한 것으로 데이터가 가지는 정보가 제거될 수 있었음 유의할 필요가 있다.

반면 1 일 방향에 대해서는 다른 목표값과 비교하여 환율 및 상품가격 변수가 높은 중요도를 보인다. 특히 RF 분류기로 측정하였을 때 위안과 유로는 각각 1 위, 3 위이며, SVM 분류기로 측정하였을 때 위안과 금 가격은 4 위, 5 위이다.

SVM 과 RF 를 비교했을 때 SVM 은 모든 정답값에 대해 변동성 지표(ATR(28))와 장기 모멘텀 지표(MACD(52,24))가 높은 중요도를 가지며, 단기 모멘텀 지표(RSI(14)는 낮은 중요도를 가진다. RF 는 각 정답값에 대해 다른 중요도 순위를 가지며, 긴 기간의 가격 방향일수록 변동성, 장기 모멘텀 지표의 순위가 높고, 짧은 기간의 가격 방향일수록 거시경제 변수와 단기 모멘텀 지표의 순위가 높다.

6. 참고 문헌

박석진, & 정재식. (2019). 고빈도 자료를 이용한 머신러닝모형의 예측력 비교 분석: KOSPI200 선물시장을 중심으로. *금융연구*, *33*(4), 31-60.

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.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

De Prado, M. L. (2018). *Advances in financial machine learning*. John Wiley & Sons.

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

Drobnič, F., Kos, A., & Pustišek, M. (2020). On the interpretability of machine learning models and experimental feature selection in case of multicollinear data. *Electronics*, *9*(5), 761.

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.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, *124*, 226-251.

Huang, N., Lu, G., & Xu, D. (2016). A permutation importance-based feature selection method for short-term electricity load forecasting using random forest. *Energies*, *9*(10), 767.

Kara, Y., Boyacioglu, M. A., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert systems with Applications*, *38*(5), 5311-5319.

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

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.

Permutation Importance with Multicollinear or Correlated Features[Website], (2021, March 06) https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance_multicollinear

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236-244.