

기계학습과 기술적 분석 지표를 이용한 KOSPI 지수 변동성 예측

이승헌¹ 이승수¹ 문혜민¹ 김순태¹

전북대학교 소프트웨어공학과

201614008@jbnu.ac.kr; 201546151@jbnu.ac.kr; 201710566@jbnu.ac.kr;
stkim@jbnu.ac.kr;

Predicting KOSPI volatility using Machine learning and Technical analysis indicators

Seunghoon Lee¹ Seungsu Lee¹, Hyemin Moon¹, Suntae Kim¹
Jeonbuk National University

요 약

본 연구에서는 기술적 분석 지표 및 머신러닝 모델인 SVM 과 Decision Tree, Random Forest 를 활용하여 KOSPI 주가지수의 변동성을 예측하고 검증을 수행한다. 실험은 학습 데이터와 실험 데이터로 나눠서 진행하였으며 소프트웨어 R 을 활용하여 36 개의 종속변수를 추출하고 기계학습 기법을 이용하여 KOSPI 지수를 예측하는 모델을 개발한다. 제시한 기계학습 기법들의 예측력을 비교, 분석한 결과 SVM 이 Decision Tree 와 Random Forest 보다 더 높은 예측력을 보였다.

키워드: SVM, Decision Tree, 주가지수 예측, KOSPI, 기계학습

1. 서 론

최근 한국의 금융시장은 예금과 적금 주식시장은 이자율이 매우 저조한 상태이며, 계속되는 금리 인하로 현재 기준금리 0.50%로 역대 최저 수준으로 낮아졌다. 저금리 시대에 목돈마련과 노후 대비를 위해 금융투자가 필수가 된 상황이다. 주가 예측은 투자자들에게 많은 관심을 받으며 다른 금융 분야와 함께 주가 예측에 대한 다양한 방법으로 주가 예측에 대한 연구를 수행하고 있으며 핀테크의 발달 등 금융산업의 새로운 흐름으로 인공지능과 투자자문 전문가의 합성어인 로보어드바이저(Robo-Advisor)에 대한 관심이 높아지고 있다.

주식에 대한 주가 예측 방법으로 기본적 분석(fundamental analysis)과 기술적 분석(technical analysis)이 존재한다. 기본적 분석은 기업의 내재가치에 영향을 미치는 모든 요소들을 분석하는 방법이지만 개별 기업의 내재가치 정의 및 측정방식 등을 기반으로 미래의 주가를 예측하는 것은 한계가 있다. 기술적 분석은 ‘주가는 수요와 공급의 원리에 따라 결정되며 수급의 결과는 그래프로 나타난다.’는 이론을 배경으로, 그래프 분석을 통해 미래 주가를

예측하는 방법이다. 컴퓨터 알고리즘의 발달과 더불어 기본적 분석과 기술적 분석을 주가 및 지수를 예측하는 방법은 위에서 기술했듯이 기본적 분석과 기술적 분석 등을 거쳐서 기술적 분석(technical analysis) 지표와 인공지능(artificial intelligence)을 접목한 다양한 연구들이 진행되고 있다. 또한, 시계열 데이터에 기계학습(machine Learning)을 접목한 다양한 연구들이 진행되며 높은 예측 성능들을 도출함으로써 주식시장에서 기계학습의 유용성에 대한 가능성을 제시하고 있다[2].

[2]은 SVM 과 라쏘 회귀분석 그리고 인공신경망 학습 모델을 갖고 KOSPI 지수의 예측력을 연구함으로써 현재 트렌드(trend)가 실효성이 있는지 연구를 수행하였다. 이 연구에서는 학습 데이터(training data)와 실험 데이터(test data)로 나눠 실험을 진행하는데 전자는 2000 년 1 월 1 일부터 2009 년 12 월 31 일까지이며 후자는 2010 년 1 월 1 일부터 2015 년 9 월 15 일로 지정하였다. Training 을 위해서 input data 와 output data 를 수집하였으며 전자는 독립변수(independent variable)그리고 후자는 종속변수(dependent variable)가 된다. 과거 시계열의 독립 및 종속변수를 바탕으로 기계학습을 하고 학습된

모형에 새로운 input data 를 적용하였으며 예측된 output data 를 추출하여 학습 방법론 간의 예측성에 관하여 분석하였다. 그 결과 라쏘 회귀분석의 경우 예측력이 저조하지만 SVM 과 인공신경망 학습 모델의 경우에는 예측력이 학습 데이터와 실험 데이터 모두 0.5 보다 높은 결과를 입증하였다.

이에 본 연구에서는 기존 연구에서 사용된 6 개의 기술적 분석 지표 외에 트렌드와 모멘텀을 기반으로 기술적 분석 지표 36 개를 추출하여 머신러닝 모델인 SVM(Support Vector Machines), Decision Tree, Random forest 를 이용해 KOSPI 지수의 예측력을 분석을 수행하였다. 실험기간은 2000 년 05 월 02 일부터 2020 년 04 월 30 일까지 일별(daily) 지수의 증가 값을 기초로 한다. Training 을 위해서 input data 와 output data 를 수집하였는데 전자는 독립변수(independent variable) 그리고 후자는 종속변수(dependent variable)가 된다. 독립 및 종속변수를 바탕으로 기계학습을 하고 학습된 모형에 새로운 input data 를 적용하면 예측된 output 을 추출하여 다양한 모델에 관하여 실험한 결과 SVM 의 성능이 0.57 로 가장 우수하였으며 Random Forest 의 경우 0.56, Decision Tree 는 0.54 의 정확도를 확인할 수 있었다.

본 논문의 2 장에서는 기계학습 기법과 관련 연구에 관해서 기술하며, 3 장에서는 실험을 위한 데이터 설명과 함께 2 가지 학습 방법론 간의 예측성에 관해서 설명한다. 4 장에서는 본 연구의 실험 결과에 관하여 기술하고 5 장에서는 결론으로 마무리한다.

2. 배경지식 및 관련 연구

이번 장에서는 본 논문을 이해하기 위한 핵심적인 기계학습 알고리즘과 주가 예측에 대한 관련 연구에 관하여 소개한다.

2.1. 배경지식

2.1.1 Support Vector Machines

서포트 벡터 머신(support vector machines, SVM)은 기계 학습의 알고리즘 중 하나로 패턴 인식, 자료 분석을 위한 지도학습(supervised training)모델이며, 주로 분류와 회귀 분석을 위해 사용한다(Vapnik, 1995). 러시아의 통계수리학자인 Vapnik 에 의해 개발된 분류알고리즘으로, 입력공간과 관련된 비선형문제를 고차원의 특징 공간에서 선형문제로 대응시켜 적은 수의 샘플만으로도 상대적으로 우수한 분류성능을 도출할 수 있는 가장 효율적인 분류알고리즘 중의 하나이다(Ahn et al. 2006; Vapnik, 1995). 두 개의 카테고리 중 어느 한 카테고리에 속한 데이터 집합이 주어졌을 때, SVM 은 주어진 데이터 집합을 기반으로 하여 새로운 데이터가 어느 카테고리에 속하는지를

판단하는 비확률적 이진선형분류 모델을 만들어 준다. 이렇게 만들어진 모델은 데이터가 사상된 공간에서 하나의 경계로 표현되는데 SVM 알고리즘은 그 경계 중 가장 큰 폭(margin)을 갖는 경계를 찾는 알고리즘이다. 이러한 SVM 은 구조적 위험을 최소화하는 원리를 가지고 있기 때문에, 과적합(overfitting)의 위험성이 높은 인공신경망과 비교해 이론적으로는 더 우월한 기계학습 알고리즘으로 인식되고 있다[3].

2.1.2 Decision Tree

Decision Tree (이하 결정트리)는 자료 내에 존재하는 관계 및 규칙 등을 찾아내어 그 의사결정규칙을 나무 구조로 도표화하여 탐색적 절차에 유용할 뿐 아니라 예측 및 분류를 수행할 수 있다는 장점을 가지고 있다(Kwak 과 Rhee, 2016). 의사결정나무모형은 분석의 목적과 자료구조에 따라 적절한 분리기준과 정지규칙을 지정하고 의사결정가지치기와 같은 특징 선택을 통해 분류에 가장 필요한 특징들만 추출함으로써 원자료에 비해 줄어든 자료를 얻을 수 있으며, 분류의 기준이 되기에 기여도가 떨어지는 잡음, 중복자료 그리고 규칙을 제거할 수 있다. 의사결정나무모형을 만드는 데는 카이제곱통계량을 분리 기준으로 하는 카이 제곱 자동 상호 작용 검출모형, 지니계수를 분리 기준으로 하는 분류회귀나무모형, 그리고 엔트로피 지수를 분리 기준으로 하는 C4.5 등 다양한 모형이 있다[2].

2.1.3 Random Forest

랜덤 포레스트(Random Forest)는 무작위 복원추출과 무작위 변수선택으로 구축되는 기계학습 알고리즘이다(Breiman, 2001). 무작위 복원추출로 학습 데이터를 n개의 샘플로 나누어, 각 샘플에서 의사결정나무를 구축한다. 또한 결정나무의 노드는 무작위로 선택된 m 개의 변수 중에서 최적의 분할기준을 찾음으로써 이진(binary) 분할된다. 결과적으로 각기 다른 구조를 가지지만 예측 성능이 뛰어난 의사결정나무 집합이 구축되며, 출력변수에 대한 입력변수의 상관성이 변수중요도로 출력된다[4].

2.2. 관련 연구

주가와 관련하여 다양한 모델을 기반으로 주가 예측 연구가 진행되었다.

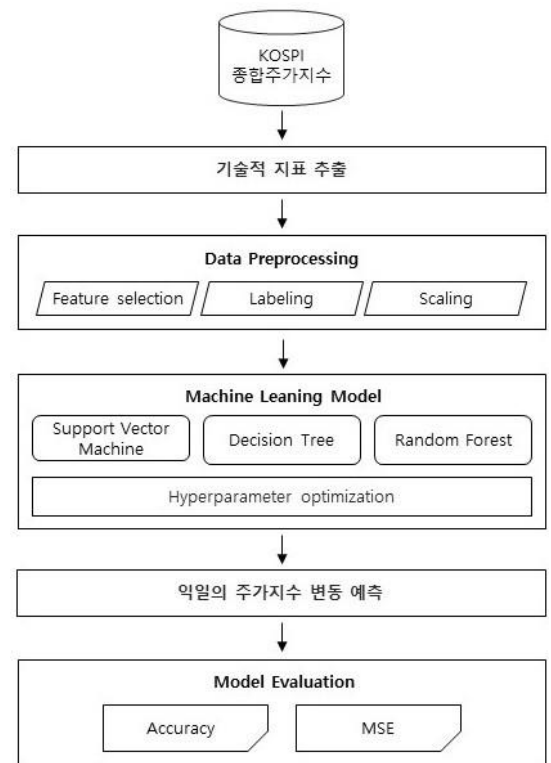
첫 번째로 박재연(2016)은 KOSPI 지수를 예측하는 기계학습 모델인 SVM, 회귀분석 그리고 인공신경망 학습 모델을 갖고 KOSPI 지수의 예측력을 연구함으로써 현재 트렌드(trend)가 실효성이 있는지 분석하였다. 약 15 년의 일별 지수 데이터를 학습기간과 실험 기간으로 나눠서 실험해본 결과 라쏘 회귀분석의 경우에는 예측력이 매우 저조하지만

SVM 과 인공지능망 학습 모델의 경우에는 예측력이 학습 데이터와 실험데이터 모두 0.5 보다 높았다. 이와 같이 아무리 우수한 예측 변수(predictor)를 가지고 있다고 하더라도 미래의 주가의 움직임을 예측하기는 쉽지 않다는 결과를 보였다[3].

두번째로 이우식(2017)은 기술적 주가분석기법을 의사결정나무모형, 서포트벡터머신모형 그리고 딥러닝모형에 결합한 새로운 기법을 제안하였고, 한국 코스피 주가지수의 방향성 예측에 대해 비교 분석하였다. 분석 결과, 첫째, 전체 변수를 사용하지 않고 선정된 변수만으로도 한국 코스피 주가지수 방향을 예측할 수 있다는 결과를 얻었다. 둘째, 제시한 모형들의 비교실험을 수행한 결과, 세 모형의 방향성에 대한 예측력이 비슷하다는 결과를 얻었다. 이는 수많은 금융시장 변수들이 직간접적으로 서로 복합하게 얽혀 불규칙적으로 변화하기 때문에 딥러닝 모형을 이용하더라도 기술적 분석 지표만으로 미래 주가지수의 움직임을 예측하는 데에는 한계가 있음을 확인할 수 있었다[1].

3. 접근방법

본 연구에서는 KOSPI 지수를 예측하는 기계학습 모형을 개발하고 성능을 시험하기 위해서 크게 학습데이터(Train Data)와 실험데이터(Test Data)로 나누어 성능 Test 를 수행하였다. 본 장에서는 앞서 기술한 학습 및 실험 데이터의 수집과 종속변수에 관하여 설명하며 기계학습 모형으로 사용하는 SVM(Support Vector Machine), Decision Tree, Random Forest 모델에 관해서 설명하고자 한다.



(그림 1) 본 연구의 전체적인 프로세스

3.1. KOSPI 종합주가지수

본 연구에서는 KOSPI 종합주가지수 예측을 위하여 실험 데이터와 학습 데이터로 구분하였다. 학습 데이터는 2000년 05월 02일부터 2020년 04월 30일까지이며 실험 데이터는 2015년 05월 04일부터 2020년 04월 30일까지로 전자인 학습 데이터는 기계학습 모형을 구성하기 위한 기간이고 후자인 실험 데이터 기간은 학습된 모형에 입력변수를 투입해서 추출되는 값과 실제 값과의 비교를 통해서 학습 모형을 검증하기 위한 기간이다. 데이터는 한국거래소에서 일별(daily) 지수의 종가, 시가, 고가, 저가, 거래량, 변동률에 대한 데이터를 수집하였다. 수집한 데이터를 기반으로 통계 소프트웨어인 R의 quantmod 패키지를 통해 36개의 기술적 분석 지표를 변수로 추출하였으며 종속변수는 익일 KOSPI 지수의 변동률로 정의한다. 따라서 오늘 입력변수를 통해 상승 신호(signal)가 나왔다면 익일 KOSPI 지수의 시가는 상승 신호가 나온 날의 종가 대비 상승하고, 반대로 하락 신호가 나오면 익일 KOSPI 지수의 시가는 하락 신호가 나온 날의 종가 대비 하락하는 것을 의미한다. KOSPI 지수로 산출된 변수는 추세(Trend)와 모멘텀(momentum) 등으로 변환한 기술적 분석 지표로 아래 표 1과 같다.

기술적 지표	설명	
Volume	거래량	
TR	이전 시점의 가격과 현재 가격의 실제 변동 폭을 나타낸 것으로, 당일 고가에서 저가를 뺀 값, 당일 고가에서 전일 종가를 뺀 값, 당일 저가에서 전일 종가를 뺀 값 중 절대값이 가장 큰 값을 True Range 라 한다.	
ATR	TR 값의 평균을 낸 것으로 현재 추세의 강도를 측정하는 데 사용된다.	
True High	금일 고가와 전일 종가 중 더 높은 값	
True Low	금일 저가와 전일 종가 중 더 낮은 값	
ADX(Welles Wilder's Directional Movement Index)	오늘의 주가 움직임의 범위가 전일의 범위를 위로 벗어났는지, 아니면 아래로 벗어났는지를 관찰함으로써 매수매도의 추세를 보여준다.	
	Dip	Positive Directional Indicator 는 ADX 지표의 구성요소이며 상승추세를 판단하는 데에 사용된다.
	Din	Negative Directional Indicator 는 ADX 지표의 구성요소이며 하락추세를 판단하는 데에 사용된다.
	DX	directional movement Index, 또는 DMI 는 가격의 방향을 판단하는 데에 사용된다.
MACD(moving average convergence divergence)	이동평균수렴·확산지수. 기간이 다른 이동평균선 사이의 관계에서 추세변화의 신호를 찾으려는 진동자 지표. 주가의 움직임과 방향을 판단하는 데에 사용된다.	
	MACD signal	MACD 의 9 일 이동평균선으로 매수매도 시점을 판단하는 데에 사용된다. 보편적으로 MACD 가 signal 선을 상향돌파 시 매수, 하향돌파 시 매도 신호로 인식된다.
SAR(Parabolic Stop-and-Reverse)	SAR 은 지수가 움직이는 방향을 판단하는 데에 사용된다. SAR 은 stop and reverse 의 약자로, 가격이 움직이는 방향의 전환을 나타낸다.	
60 volatility	60 일 평균 주가 변동을	
Bollinger Band	추세중심선인 이동평균선을 중심으로 하여 이동평균선 위쪽에 상한선, 아래쪽에 하한선을 설정함으로써 형성되는 가격변동폭의 띠	
	BB up	볼린저밴드의 상단밴드
	BB dn	볼린저밴드의 하단밴드
	BB pctB(%B)	%B 지표는 상단밴드와 하단밴드에 대한 가격의 위치를 이용해 추세 및 트레이딩 시그널을 찾아내는 데 유용하게 쓰일 수 있다.
RSI(Relative Strength Index)	상대강도지수. RSI 는 시장가격 변동폭 중에서 시장가격의 상승폭이 어느 정도인지를 분석하는 것으로 주가가 상승 추세일 경우 얼마나 강한 상승세인지, 하락추세라면 얼마나 강한 하락세인지를 퍼센트로 나타낸다.	
Stochastic	주가의 마감 가격이 일정기간 어느 곳에 있었는지를 관찰하기 위해 백분율로 나타낸 단기 기술적 지표를 말한다. 스토캐스틱은 두 개의 선으로 표시되는데, %K 와 %K 의 이동평균을 낸 %D 로 이루어진다.	
Williams AD	Williams % R (또는 % R)은 지난 n 일의 높고 낮은 것과 관련하여 현재 종가를 보여준다.	
Aroon	Aroon oscillator 지표는 추세의 변화와 강도를 판단하는 데에 사용된다. 본 지표는 특정 기간 고가가 발생한 후 다음 고가까지의 기간과 저가 발생 후 다음 저가까지의 기간을 측정한다.	
	Arron Up	상승 추세의 강도를 나타내는 지표
	Arron Dn	하락 추세의 강도를 나타내는 지표
	Oscillator	Aroon Up 에서 Aroon Dn 을 뺀 값
CCI(Commodity Channel Index)	최근 가격과 이동평균선의 차이를 나타내는 지표. 이를 토대로 가격의 방향과 탄력성을 판단할 수 있는 근거를 제시한다.	
Chaikin AD	차이킨오실레이터는 단기 ADI 이동평균에 장기 ADI 이동평균을 빼서 오실레이터화 한 지표이다. 이 지표는 가격이 상승하기 전에 거래량이 먼저 상승한다는 것을 전제로 만들어졌으며 오실레이터화해서 보다 빠르게 신호가 나타난다.	

CMF(Chaikin Money Flow)	당일 가격 변동폭에 대한 당일 종가의 비율에 거래량을 가중한 것으로 시장 강도를 측정하기 위한 지표이다.	
EMV(Arms' Ease of Movement Value)	거래량 대비 가격이 움직이는 용이함의 정도를 나타내는 지표이다. 거래량 대비 가격 움직임이란 거래량 대비 가격의 움직임이 큰 것을 말한다.	
	maEMV	EMV 지표의 이동평균선
TDI (Trend Detection Index)	추세의 시작과 끝을 나타내는 지표	
DI(Direction Indicator)	추세의 방향을 나타내는 지표	
TRIX(Triple Smoothed Exponential Oscillator)	주가의 증가에 대해 3 번(triple) 이동평균한 값을 구해 단기적인 가격의 급등락을 완화해주는 보조지표이다. 0 을 기준으로 0 을 상향 돌파하면 매수로, 하향 돌파하면 매도로 해석한다.	
VHF (Vertical Horizontal Filter)	추세의 유무를 구분하는데 유용한 지표로서 해당 시장에 사용할 지표를 결정하는데 도움을 준다. VHF 값이 상승한다면 시장이 추세적 방향을 땀을 의미하며 VHF 값이 하락하는 것은 시장이 추세가 없는 수평 방향을 띄는 것을 의미한다.	
ROC (Price Rate of Change)	일정 시점의 가격 변화율을 백분율로 나타낸 것으로 0 선을 기준으로 하여 추세의 반전을 알려주는 지표이다.	

(표 1) 본 연구에서 활용하는 36 개 기술적 지표

3.2. Data Preprocessing

본 절에서는 데이터 전처리를 위하여 Feature selection, Labeling, Scaling 을 수행하였다. 첫 번째로 feature selection 은 모든 특징의 부분 집합을 선택하거나, 불필요한 특징을 제거하여 간결한 특징 집합을 만드는 차원 축소 방법으로 10 개의 종속변수를 도출하였다. 도출된 10 개의 종속변수는 SVM 모델에 활용되며 도출된 변수는 아래 표 2 와 같다.

Priority	feature	priority	Feature
1	BB_up	6	TRIX
2	chaikinAD	7	MACD_signal
3	BB_dn	8	TDI
4	ATR	9	MACD
5	DI	10	trueHigh

(표 2) Feature Selection 으로 도출된 10 개의 feature

두 번째로 예측력을 분석하기 위해 CP, RP 두가지로 나눠서 실험을 진행하는데 Labeling 은 변동률에 관하여 다음과 같이 음수일 경우 -1, 양수일 경우 +1 로 클래스(class)를 나눠서 분류(classification)를 하였다. 이는 본 연구에서 CP(classification problem)라고 정의한다. 다음은 회귀(regression) 분석을 통해서 익일의 지수 변동률을 예측하는 방법으로 이는 본 연구에서 RP(regression problem)라고 정의한다. 세 번째로 Scaling 은 자료의 오버플로우나 언더플로우를 방지하고 독립 변수의 공분산 행렬의 조건 수를 감소시켜 최적화 과정에서의 안전성 및 수렴 속도를 향상시키는 데이터 전처리 과정이다. Scaling 방법으로는 Scikit-learn 패키지의 StandardScaler 를 이용하여 모든 자료에 선형 변환을 적용한 후 전체 자료의 평균은 0으로, 표준편차는 1 이 되도록 변환하였다.

3.3. Machine Learning Model

머신러닝 기계학습을 통한 KOSPI 지수 예측 능력을 시험하기 위해 SVM 과 Decision Tree, Random Forest 모델로 학습한다. 이 때 최적의 Hyperparameter 를 찾기 위해 Hyperparameter 에 대한 테스트를 수행했다. 테스트 수행 결과, 모델별로 다음과 같이 Hyperparameter 를 적용하였다.

Model	Type	Hyperparameter
SVM	C	0.001
	Kernel	rbf
	Gamma	0.001
Decision tree	Max_depth	2
	Max_Features	auto
	Random_state	0

Random Forest	N_estimators	100
	Max feature	Log2
	Randomstate	0

(표 3) 예측 모델의 Hyperparameter

4. 실험결과

본 연구를 통해 앞서 설명한 3 가지 기계학습 방법을 통해서 개발한 학습 모델을 통해 익일 KOSPI 주가지수 변동을 예측을 위하여 실험을 진행한 결과는 아래 표 4 와 같다.

		SVM	Decision Tree	Random Forest
In sample	RP	0.57	0.58	1.00
	CP	1.07	0.96	0.14
Out of sample	RP	0.60	0.59	0.55
	CP	0.43	0.96	0.99

(표 4) 예측 모델별 정확도와 MSE 결과

회귀 모델의 경우, MSE(Mean Squared Error)를 통하여 측정하였으며 값이 작을수록 오류가 작은 것을 나타낸다. 분류 모델의 경우, 실제 값과 예측 값이 일치한 경우를 판단하여 정확도를 측정하였으며 1 에 가까울수록 높은 정확도를 나타내고 0 에 가까울수록 낮은 정확도를 나타낸다.

실험 결과 In sample 의 경우에 RP 기반은 Random Forest 가 1.00 으로 가장 우수하며, CP 기반은 SVM 이 1.07 로 가장 높았다. Out of sample 의 경우 RP 기반은 SVM 이 0.60으로 가장 우수했으며, CP 기반은 Random Forest 가 0.99 로 가장 우수했다.

5. 결론

본 연구는 기술적 분석 지표와 기계학습을 활용하여 익일 KOSPI 주가지수 변동을 예측하고 그 성능을 분석하였다. 기계학습 방법으로는 SVM 과 Decision Tree 그리고 Random Forest 학습 모델을 통해 KOSPI 지수의 예측력을 연구함으로써 현재 트렌드(trend)가 실효성이 있는지 분석하였다. 약 20 년간의 데이터를 기반으로 실험데이터와 학습데이터를 나눠서 실험해본 결과 SVM 의 성능이 Out of sample 일 때 0.60 으로 가장 높았다. 이처럼 기계 학습을 이용하여 지수의 방향성을 예측하기는 쉽지 않으나 적절한 기술적 분석 지표와 기계학습 모델을 이용한다면 높은 정확도를 기대할 수 있을 것으로 예상된다. 더불어, 본 연구에서 제시한 3 가지 예측 모델 외 다양한 머신러닝(machine Learning) 모델과 비교해보고 더 우월한 모델을 제시, 연구해 볼 수 있다.

6. REFERENCES

- [1] 이우식, “딥러닝분석과 기술적 분석 지표를 이용한 한국 코스피주가지수 방향성 예측”, 한국데이터정보과학회지, 제 28 권, 제 2 호, pp.287-295, 2017.
- [2] 하대우, 김영민, 안재준 “XGBoost 모형을 활용한 코스피 200 주가지수 등락 예측에 관한 연구”, 한국데이터정보과학회지, 제 30 권, 제 3 호, pp. 655-669, 2019.
- [3] 이종식, 안현철, “입력변수 및 학습사례 선정을 동시에 최적화하는 GA-MSVM 기반 주가지수 추세 예측 모형에 관한 연구”, 지능정보연구, 제 23 권, 제 4 호, pp. 147-168, 2017.
- [4] 신한술, 박철수, “랜덤 포레스트 모델을 위한 데이터 전처리 기법의 적용”, 대한건축학회 학술발표대회 논문집, pp. 633-634, 2016.