

설명 가능 인공지능 기법을 활용한 주가 전망 예측

김지현*, 이연수*, 정수민**, 조설아***, 안정은*, 김현희*

*동덕여자대학교 정보통계학과

**동덕여자대학교 컴퓨터학과

***동덕여자대학교 문예창작과

wilkil526@gmail.com, ahh0123@naver.com, sumin00716@naver.com, joisnow96@gmail.com,
jn6430@naver.com, heekim@dongduk.ac.kr

A Stock trend Prediction based on Explainable Artificial Intelligence

Ji Hyun Kim*, Yeon Su Lee*, Su Min Jung**, Seol A Jo***, Jeong Eun Ahn*, Hyun Hee Kim*

*Dept. of Statistics and Information Science, Dongduk Women's University

**Dept. of Computer Science, Dongduk Women's University

***Dept. of Creative Writing, Dongduk Women's University

요 약

인공지능을 활용한 주가 예측 모형을 실제 금융 서비스에 도입한 사례가 많아지고 있다. 주식 데이터는 일반적인 시계열 데이터와 다르게 예측을 어렵게 하는 복합적인 요소가 존재하며 주식은 리스크가 큰 자산 상품 중 하나이다. 주가 예측 모형의 활용 가능성을 높이기 위해선 성능을 향상시키는 것과 함께 모형을 해석 가능한 형태로 제시해 신뢰성을 향상시킬 필요성이 있다. 본 논문은 주가 전망 결정 방법에 따른 예측 결과를 비교하고, 설명 가능성을 부여해 모형 개선했다는 것에 의의가 있다. 연구 결과, 주가 전망을 장기적으로 결정할수록 정확도가 증가하고, XAI 기법을 통해 모형의 개선 근거를 제시할 수 있음을 알 수 있었다. 본 연구를 통해 인공지능 모형의 신뢰성을 확보하고, 합리적인 투자 결정에 도움을 줄 수 있을 것으로 기대한다.

1. 서론

근래에 투자의 위험을 줄이고, 합리적인 투자 결정을 돕기 위해 주가 예측 모형을 실제 서비스에 적용한 사례가 많아지고 있다. 주가 데이터는 시계열 데이터에 속한다. 그러나 일반적인 시계열 데이터와 달리 비정상성(non-stationarity), 장기 의존성(long-term dependency), 비선형성(non-linearity) 등의 특성을 가지고 있어 예측하는데 어려움이 있다[1].

최근에 시계열 예측이 아닌 기계 학습 기법을 사용한 주가 등락 예측 모형이 제안되고 있다[2]. 하지만 정확도, 정밀도 같은 모델 성능 평가 지표 외에 인공지능 모형을 신뢰하고 투자 결정을 내리기 위한 근거는 부족한 상황이다. 주식은 변동성이 큰 자산으로 리스크가 큰 금융 상품에 속한다. 투자 위험을 완화하며 주가 예측 모형을 활용하기 위해서 인공지능을 신뢰할 수 있는 근거를 제시하고, 모형에 설명 가능성을 부여할 필요성이 있다.

본 논문은 설명 가능 인공지능(XAI, Explainable Artificial Intelligence) 기법 중 SHAP(SHapley Additive

exPlanations) 기법[3]을 활용해 인공지능이 학습한 내용을 해석 가능한 형태로 제시하고, 모델 개선 근거¹를 마련하였다. 이때, 분류 예측을 위해서 XGBoost[4]를 사용하였다. 또한, 전망 결정에 참고하는 데이터 범위를 중·장기적으로 확장하여 주가 전망 결정 방법에 따른 예측 정확도 차이를 폭넓게 분석하였다.

첫번째 실험으로 주가 전망 결정 방법에 따른 성능을 비교하고, 각 변수가 미치는 영향을 확인했다. 이 실험의 목적은 모형의 개선 근거를 마련하기 위함이다. 앙상블 알고리즘[5]을 사용하여 윈도우 크기를 1, 7, 30, 90 으로 설정하여 학습한 후 예측 정확도를 비교했다. 윈도우 크기는 주가 전망을 결정하기 위해 참고한 일수를 말한다. 가격 데이터, 기술 지표 데이터, 시장 데이터로 특징 벡터를 구성하여 SHAP 기법으로 변수의 영향력을 측정하였다.

실험 결과 윈도우 크기가 커질수록 예측 정확도가 향상되고, 가격보다 기술 지표와 시장 지수가 전망 예측에 중요한 요소로 해석됨을 알 수 있었다. 이를 근거로 하여 가격 데이터와 시장데이터로부터 추출된

* 본 연구는 교육부의 대학혁신지원사업 사업비를 받아 진행하는 것입니다.

기술 지표 변수 조합을 최종 변수로 선정하였다. 최종 모형 적용 결과 주가 전망을 30 일을 기준으로 결정했을 때 93.64%의 성능을 보였다. 이것은 설명 가능 인공지능 기법을 통해 변수를 선별하여 모델을 개선하고, 예측 결과의 신뢰성을 판단하여 투자 결정을 내리는데 도움을 주는 모델을 제안하였다는 의의가 있다.

본 논문은 다음과 같이 구성된다. 먼저 제 2 장에서 분석 방법을 기술하고 제 3 장에서 실험 결과를 자세히 설명한다. 마지막으로 제 4 장에서 결론 및 향후 연구를 제시한다.

2. 분석방법

2.1 실험설계

본 논문은 2011 년부터 2020 년까지 10 년간의 네이버 주식 데이터를 사용했다. 사용한 변수는 가격, 기술 지표, 시장 지수로 나눌 수 있다. 첫번째 실험으로 주가 전망 결정 방법에 따른 예측 정확도 차이를 비교하고, 주가 전망 예측에 변수가 미치는 영향을 확인했다. 이 실험의 특징 벡터는 가격 데이터, 기술 지표 데이터, 시장 지수 데이터를 모두 포함하여 구성하였다. 기술 지표로 MA5, MA10, MA20, MA60 of close, MA20, MA60 of volume, RASD5, RASD10, MACD, CCI14, ATR14, Bollinger band ub, middle, lb, MTM1, MTM3, ROC60, WPR14, ADX14, ADXR14 를 사용하였다. 기술 지표 변수 이름 뒤에 추출 시 설정한 기간을 표기하였다. 예를 들어, MA5 는 5 일동안의 이동 평균을 뜻한다. 최종 모형엔 첫번째 실험 결과에 따라 영향력 있다고 판별된 변수 조합을 사용하였다. 모델 개선과 해석은 윈도우 사이즈가 30 일 때를 기준으로 하였다. 최종 모형에 사용된 특징 벡터는 Table 4 에서 확인할 수 있다. Table 1 은 실험에 사용한 변수들을 나타낸다.

price	Open, High, Low, Close, Volume
Technical Indicator	CCI, MAs of {Close, volume}, RASD, MACD, ATR, Bollinger band {ub, middle, lb}, MTM, ROC, WPR, ADX, ADXR
Market	S&P500, SOX, VIX, KOSPI

Table 1. 실험에 사용한 변수

데이터에 대한 전처리 과정은 다음과 같다. 먼저 네이버 주식 데이터에 대해 결측치를 제거한 후 2360 개의 영업일 데이터를 Min-max scaler 를 사용해 0 과 1 사이로 정규화 하였다. 첫번째 실험엔 Xgboost, RandomForest, Adaboost 와 같은 앙상블 분류기를 사용했으며 이 중 Xgboost 분류기에 설명 가능성을 부여해 모델을 개선하였다. 성능 평가 지표로는 accuracy 를 사용하였다. 네이버 주식 외에 삼성전자, 현대차,

아모레퍼시픽, LG 화학 주식 데이터를 사용해 네이버 주식과 유사한 실험 결과가 도출되는지 확인했다. Table 2 에서 실험에 사용된 데이터, 데이터 수집 기간, 전처리, 분류 알고리즘, 실험 환경을 요약하였다.

Stock	네이버, 삼성전자, 현대차, 아모레퍼시픽, LG 화학
Period	2011.1.03~2020.12.30 (영업일 2360 일)
Classifier	XGBoost, RandomForest, Adaboost
CPU	apple silicon m1
RAM	8GB
OS	mac

Table 2. 실험 환경

본 연구는 다음 수식으로 주가 전망을 결정했다. 일반적으로 주가 전망은 기준 값과 대상 값을 비교해 ‘상승’ 과 ‘하락’으로 결정한다. 본 연구에서도 기준 값과 대상 값을 비교해 전망이 ‘하락’으로 결정될 경우를 0, ‘상승’으로 결정될 경우를 1 로 설정했다. 윈도우의 사이즈를 1, 7, 30, 90 으로 설정해 폭넓게 예측 실험을 했다. 윈도우 내의 추세로 전망을 결정하기 위해 다음날부터 n 일 동안의 평균 증가를 대상 값으로 사용하였다. 해당 일의 증가를 기준 값으로 다음 날부터 n 일까지의 평균 증가를 대상 값으로 설정하면 수식 (1)과 같다.

$$target = \begin{cases} 0 & \text{if } close_t \geq \frac{\sum_{i=1}^n close_{t+i}}{n} \\ 1 & \text{if } close_t < \frac{\sum_{i=1}^n close_{t+i}}{n} \end{cases} \quad (1)$$

3. 실험 결과

3.1 주가 전망 결정 방법에 따른 예측 정확도 분석

Table 3 은 첫번째 실험 결과로 네이버 주식에 대해 앙상블 분류기로 학습한 모델의 테스트 데이터 셋 분류 정확도를 보여준다.

n	Xgboost	RandomForest	Adaboost
1	55.08%	52.75%	53.18%
7	80.93%	82.20%	66.31%
30	91.95%	93.43%	81.14%
90	97.03%	97.03%	93.22%

Table 3. 네이버 주가 예측 정확도

랜덤포레스트 모형의 성능이 가장 우수했으며, 주가 전망을 장기적으로 결정해 예측할 수록 성능이 향상되었다. 실험 기간의 양성 레이블은 n이 1일에서 90으로 증가할 때 47.97%에서 60.3%로 증가하였고, 랜덤포레스트 정확도는 52.75%에서 97.03%로 증가하였다. 다른 분류기도 양성 레이블 비율이 증가한 것에 비해 예측 정확도는 더 큰 폭으로 상승한 것을 확인할 수 있다. 이를 통해 주가 전망 결정 방법에 따라 정확도가 달라지며, 전망 결정 방법이 성능 향상에 영향을 미칠 수 있다는 것을 알 수 있다.

3.2 모델 해석

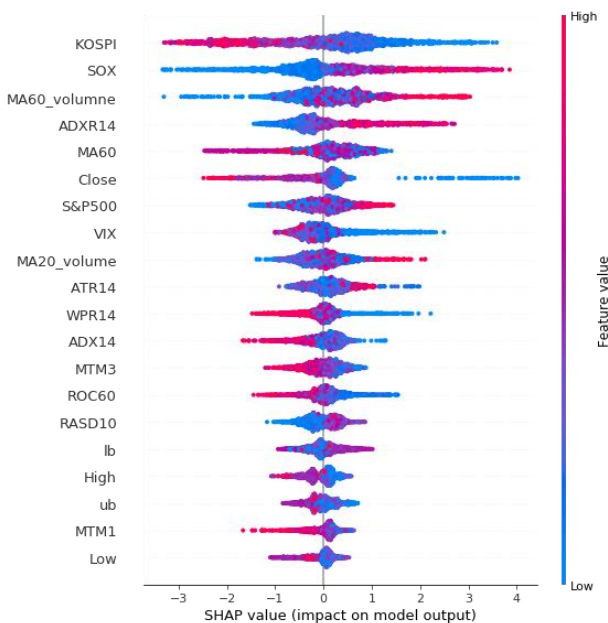
윈도우 사이즈가 30인 경우의 XGBoost 모형에 설명가능성을 부여해 주가 예측에 영향력 있는 변수를 확인하였다. <Figure 1>는 전체 입력 피쳐들이 주가 전망 예측에 어떤 영향을 미치는지 새플리 값의 절댓값순으로 시각화한 결과이다. X 축은 주가 예측에 부정적인 요소부터 긍정적인 요소까지 각 요소의 크기를 표시한다. Y 축엔 변수가 나열 되어있고, 색깔은 해당 변수의 값이 작아지거나 커질 때의 상황을 표시한 것이다. 값이 커질수록 붉은 색을 띠며, 작을수록 푸른 색을 띤다[6]. 코스피 지수는 작을수록 주가 상승 예측에 긍정적인 영향을, 커질수록 부정적인 영향을 준 것을 확인할 수 있다. 반면에, 반도체 지수와 ADXR14는 커질수록 긍정적인 영향을, 작을수록 부정적인 영향을 준 것을 확인할 수 있다. 코스피 지수, 반도체 지수, ADX, ADXR, MA60 등의 시장 지수 데이터와 기술 지표 데이터가 전망을 예측하는데 영향력 있는 변수로 계산되었음을 확인할 수 있다.

3.2 모델 개선 실험

첫번째 실험으로 생성된 모델은 가격 데이터와 시장 지수 장 마감 가격 데이터가 입력 피쳐로 사용된다. 가격 데이터는 주가 예측에 어떤 영향을 주었는지 해석하기 어렵다. 따라서 <Figure 1> 결과를 바탕으로 시장 데이터와 가격 데이터에서 기술 지표를 추출하였고, 기술적 분석을 통해 해석이 가능하도록 모델을 개선하는 실험을 했다. 특히, 가격으로 표현되지 않는 기술 지표를 변수로 사용했다. 주식 증가와 시장 지수 증가로부터 기술 지표를 추출하여 특징 벡터를 구성하였다. 시장 지수는 모델 해석에서 가장 중요한 변수로 측정되었던 KOSPI 지수를 사용하였다. Table 4는 새로운 특징 벡터를 사용하고, 하이퍼 파라미터 튜닝을 진행하여 개선된 모델을 요약한 것이다.

Feature vector	kosADX14, kosADX30, kosADX60, kosADX90, CCI30, CCI60, MTM1, MTM3, ROC60, ROC90, WPR60, WPR90, ADX14, ADX30, ADX60, ADX90, ADXR14, ADXR30, ADXR60, ADXR90	
Classifier	Xgboost	
Parameter	learning_rate	0.3
	max_depth	6
	n_estimators	110
	booster	gbtree
	objective	binary:logistic

Table 4. 개선된 모델



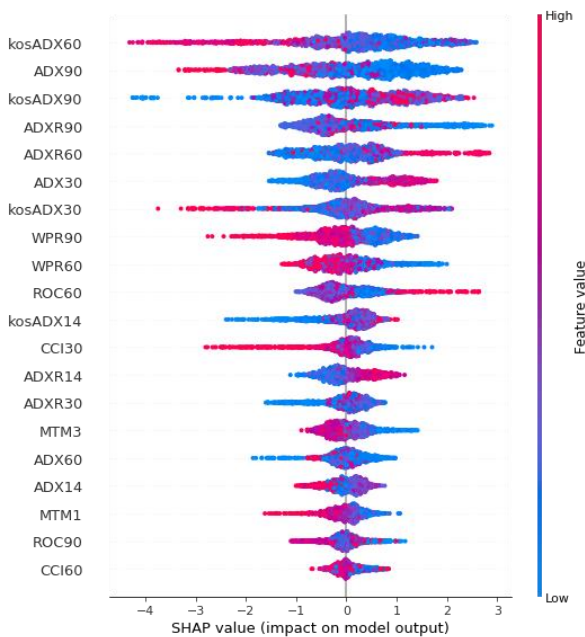
<Figure 1>

Table 5는 5개의 종목에 대해 학습한 최종 모형의 테스트 데이터 셋 분류 정확도를 보여준다. 네이버 주식과 비슷한 코스피 시가총액 상위 종목에서 유사한 결과를 얻었음을 확인할 수 있다. 네이버 주가 예측 최종 모형으로 30일 동안의 전망 예측을 실험한 결과 검증 데이터 셋에 대해 93.6%의 정확도를 얻을 수 있었다. 이는 가격 데이터를 입력 피쳐로 사용한 모델에 비해 개선된 수치이다. 주가 예측에 비효율적인 변수를 제거하고 모델이 학습한 과정에 해석이 용이한 변수를 사용하여 모델이 개선된 것이다. 최종 모형은 윈도우 사이즈가 7, 30인 경우에 첫번째 실험의 XGBoost 모델보다 향상된 정확도를 보였다. 이는 윈도우 사이즈가 30인 경우를 기준으로 모형을 개선할 방향을 정했기 때문인 것으로 보인다.

n	네이버	삼성 전자	현대차	아모레 퍼시픽	LG 화학
1	54.24%	52.52%	54.41%	53.47%	51.68%
7	81.57%	80.88%	83.40%	82.53%	82.35%
30	93.64%	91.60%	94.12%	92.21%	93.70%
90	96.19%	96.85%	95.17%	97.26%	97.48%

Table 5. 개선된 모델 예측 실험 결과

<Figure 2>은 최종 모형에 사용된 피처에 대해 주가 전망 예측에 주는 영향력을 SHAP value 순으로 시각화한 것이다. 코스피의 추세 강도를 나타내는 변수가 가장 영향력이 큰 변수로 계산되었다. KosADX60 은 커질수록 부정적인 영향을, 작을수록 긍정적인 영향을 준 것으로 보인다. 반면, KosADX90은 커질수록 주가 상승 예측에 긍정적인 영향을, 작을수록 부정적인 영향을 준 것으로 보인다. 코스피와 주가 추세의 강도를 나타내는 ADX, ADXR 지표도 영향력 있는 변수로 계산되었다.



<Figure 2>

<Figure3>은 특정 일의 입력 변수가 예측에 어떤 영향을 주었는지 시각화한 그림이다. 주가 상승 예측에 붉은색은 긍정적인 영향을 준 변수이며, 푸른색은 부정적인 영향을 준 변수이다. 이 그림을 해석하면 주가 상승에 긍정적인 영향으로 작용한 변수는 Williams percent range, ROC, ADXR, CCI 와 같다. 이 결과는 기술

적 분석을 통해 투자 전략을 세울 때 중요한 기술 지표를 선택하고, 모형의 예측 결과의 신뢰성을 확보하는데 도움을 줄 수 있다.



<Figure 3>

4. 결론 및 향후 연구

본 논문에서는 주가 전망 결정 방법에 따른 예측 정확도의 차이를 분석하고, 설명 가능 인공지능 기법을 활용하여 개선된 모델을 제시했다. 연구 결과, 주가 전망 결정 방법에 따라 예측 정확도의 차이가 있다는 것을 확인했다. 또한, 설명 가능 인공지능 기법 중 하나인 SHAP 기법을 활용하면 예측에 변수가 어떤 영향을 주는지 확인할 수 있어 변수를 선별하고 모델을 개선하는데 용이하다는 것을 알 수 있었다. 이것은 모형의 예측 결과의 신뢰성을 확보하고, 투자 전략의 방향을 수립하는데 도움을 줄 수 있다. 향후 연구에는 기술 지표 변수를 확장하여 다양한 기술적 분석을 가능하게 하고, 전망 결정 방법에 따른 해석 차이를 비교할 예정이다.

참고문헌

- [1] 천성길, 이주홍, 최범기, 송재원, “대규모 외생 변수 및 Deep Neural Network 기반 금융 시장 예측 및 성능 향상”, 스마트미디어저널, Vol.9, No.4, pp. 26-35, Dec.2020.
- [2] 임청수, “A study on Stock Trend Determination in Stock Trend Prediction”, 한국컴퓨터정보학회논문지, Vol.25, No. 12, pp.35-44, Dec.2020.
- [3] S. M. Lundbergand, S. I. Lee, “A Unified Approach to Interpreting Model Predictions”, Advancesin Neural Information Processing Systems, Vol. 2017-December, No. Section 2, pp. 4766-4775, Dec. 2017.
- [4] Tianqi Chen, Carlos Guestrin, “XGBoost: A Scalable Tree Boosting System”, KDD’16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp785-794, Aug.2016
- [5] G. Sarah, A. C. Mueller, “Introduction to Machine Learning with Python”, O’Reilly Media, 2016.
- [6] 안재현, “XAI 설명 가능한 인공지능, 인공지능을 해부하다”, 위키북스, 2020.