미래사회
Journal of Future Society
2024. Vol. 15, No. 1, 21~33.

가상자산 시계열 데이터의 분석과 예측
- 시계열 모델 및 인공지능 모델의 비교 및 개선점 제시

안 상 선*

국민대학교

비트코인, 이더리움 등의 가상자산은 주식, 채권, 옵션 등 기존의 투자 자산과 달리 가격 산정의 근거가 되는 기초자산이 존재하지 않는다. 따라서 자본자산가격 결정 모델(CAPM)이나 배당할인(DCF) 모형 등의 가격 결정 매커니즘 분석을 사용하기 어렵다. 따라서 본 연구에서는 펜데믹 기간을 포함해서 2019년부터 현재까지 가상자산 시계열 데이터를 이용해 과거 정보의 의존경향을 반영하는 LSTM, 페이스북(메타)에서 개발한 구조적 상태 변화를 포착하는데 사용되는 Prophet, 시계열 데이터의 구조적인 변화를 모델링하는데 사용되는 HMM(Hidden Markov Model)을 선정했다. 또한 각 분석 기간을 구분하기 위해서 팬데믹 선언 시점, 미국 연준(FRB)의 기준금리 결정 등의 요인을 고려했다. 이를 통해 각 모델의 성능을 평가하고, 기존 시계열 모델인 ARIMA와 비교를 통해 개선점과 한계점을 도출했다. 분석 결과, 신경망 모형을 사용한 인공지능 모델인 RNN, LSTM이 ARIMA 모형에 비해 높은 정확도를 보였다. 이러한 결과는 시계열 데이터의 시간적 의존성, 장기 의존성, 비선형 등의 특징을 잘 반영하기 때문에 예측 정확도가 높은 것으로 보인다. 다만. 분석 기간 동안 코로나19로 인한 팬데믹 선언, 미국의 급격한 기준금리 인상 및 비트코인을 비롯한 가상화폐의 저변 확대 등 환경적 요인이 존재한다. 따라서 보다 정확한 모델의 성과 평가를 위하서는 이러한 환경적인 요인이 시계열 및 인공지능 모델에 어떻게 작용하지 분석이 필요하다.

주제어: 가상자산, 비트코인, 인공지능, 시계열, 암호화폐

* 주저자: 안상선/국민대학교 겸임교수, 매일경제 사외벤처 (주)M-Robo 대표
/서울시 강서구 마곡중앙로 161-17/Tel: 0507-1360-6139/E-mail: sangsun.ahn@m-robo.com

Ⅰ. 서론

가상화폐(Virtual Currency)는 물리적 형태가 없는 디지털 형태의 화폐이다. 경제적 가치를 지닐 뿐만 아니라 결제 수단으로 사용돼, 투자, 거래, 송금 등 다양한 용도로 활용되기 때문에 자산(Asset)이 아니라 화폐(Currency)라 할 수 있다.

가상화폐는 중앙 기관이나 정부에 의해 발행되거나 관리되는 것이 아니며, 대부분 분산 원장 기술인 블록체인을 기반으로 한다. 기존의 전자 화폐, 전자 지불 시스템에서는 높은 중개 수수료 등 거래 비용이 발생했지만, 비트코인, 이더리움 등의 가상화폐는 디지털 서명, 복제와 이중 지출 등의 문제를 블록체인 기술을 이용해서 해결하고 있다(Polasik et, al. 2015). 이 과정에서 거래 내역 등이 암호화 알고리즘을 통해서 관리되기 때문에 암호화폐라고도 불린다.

비트코인, 이더리움 등의 가상자산은 주식, 채권, 옵션 등 기존의 투자 자산과 달리 가격 산정의 근거가 되는 기초자산이 존재하지 않는다. 따라서 자본자산가격 결정 모델(CAPM)이나 배당할인(DCF) 모형 등의 가격 결정 메커니즘 분석을 사용하기 어렵다. 또한, 가상자산은 중앙화된 거래소 없이 세계 여러 곳에 분산된 거래소에서 24시간 거래가 된다. 특히, 우리나라의 거래소에서 거래되는 가상화폐의 가격이 국제적으로 인정받는 다른 거래소의 가격보다 높게 형성되는 '김치 프리미엄'과 같은 현상이 종종 나타난다.

이러한 특징으로 인해서 가상화폐 시장의 효율성이나 균형가격 등을 다룬 연구에서는 경제학의 분석 모형 외에 시장 참여자의 심리적 특성을 다룬 연구(박민정・채상미, 2019), 빅데이터 분석 기법을 활용한 연구, 신경망 모형을 활용한 연구 등 다양한 방법을 적용한 사례가 많다.

기존 연구는 가상화폐의 성격을 기존의 화폐 및 파생상품 자산과 비교해서 그 특징을 도출했다. 또한 이러한 특징을 근거로 기존의 금융 시계열 모형을 적용해보거나, 인공지능 모델을 도입해서 가격 움직임을 분석했다.

본 연구에서는 기존의 연구를 확대해서 기존 금융 시계열 모형과 여러 인공지능 모델을 활용해서 가상화폐 가격의 움직임을 예측하고자 한다. 이를 위해 가장자산 데이터의 시계열(Time Series) 특성에 주목해서 다음과 같은 모형을 활용했다. 과거 정보의 의존경향을 반영하는 LSTM, 페이스북(메타)에서 개발한 구조적 상태 변화를 포착하는데 사용되는 Prophet, 시계열 데이터의 구조적인 변화를 모델링하는데 사용되는 HMM(Hidden Markov Model)을 선정해 각 모델의 미래 가격 예측 성능을 비교했다. 또한 경제학에서 사용하는 시계열 모델인 ARIMA와 비교를 통해 각 모델의 가격 예측 과정에서의 특징을 살펴보았다.

Ⅱ. 가상자산의 특성과 주요 연구

가상화폐는 디지털 정보로 표현되는 형태의 화폐로, 물리적인 형태가 없는 상태로 온라인 및 오프라인 거래에서 사용된다. 이러한 가상화폐는 전자적 매체에 화폐가치가 저장되어 있다는 점에서 이미 1990년대 중반에 출현하였던 전자화폐의 일종이라 할 수 있다. 그러나 발행자가 없고, 발행자가 아닌 네트워크 참여자에 의해 공동으로 관리 및 운영된다는 점에서 전자화폐와 구별된다.

가상화폐는 유럽중앙은행(ECB)에서 정의한 대로 "개발자에 의해 발행되고 통상 관리되며, 특정한 가상 커뮤니티의 회원들 간에 사용되고 수령되는 규제되지 않은 디지털화된 화폐의 한 유형"으로 분류된다.

한편, 가상화폐를 암호화폐라고 부르기도 한다. 암호화폐(Cryptocurrency)는 '암호화'라는 뜻을 가진 'crypto-'와 통화란 뜻으로 화폐란 뜻을 가진 'currency'의 합성어이다, 블록체인(blockchain) 기

술로 암호화되어 분산 발행되고 일정한 네트워크에서 화폐로 사용할 수 있는 전자정보라 할 수 있다. 따라서 발행 국가라는 개념이 존재하지 않아 특정 통화로 환전할 필요가 없다. 이러한 특성으로 가상화폐는 금융 시장에서 투자 상품으로 여겨진다.

비트코인, 이더리움 등의 가상화폐는 처음에는 지급 수단으로 고안되었지만, 투자 목적을 가진 거래가 증가함에 따라 가격이 시장에서 형성되고, 투자자들 사이에서 거래되면서 투자 이익 또는 손실이 발생했다. 이 때문에 가상 화폐는 재화 교환의 매체로 사용되는 동시에 투자 상품 역할을 동시에 하는 특수한 지급 수단이라 할 수 있다.

이 때문에 가상화폐를 가상자산, 또는 디지털 자산(Digital Asset)으로 부리는 경우가 많다. 하지만 이 견해를 따를 경우에는 항공마일리지, 금융권 포인트 등 디지털로 적립 및 사용될 수 있는 기존의 전자화폐와 구별하기 어렵다. 따라서 최근에는 디지털 화폐(CBDC)라는 개념을 사용해서 이를 구분한다.

2015년 이후, 비트코인 가격이 1억 원에 근접하는 등 상승세를 보임에 따라, 가상화폐는 연구자 뿐만 아니라 일반 투자자에게도 큰 관심을 끌었다. 이후부터는 국제적으로 암호화폐의 실제 가치에 대한 연구가 활발하게 진행되고 있다.

가상자산에 대한 연구는 크게 두 가지 범주로 나눌 수 있다. 첫 번째 범주에서는 비트코인을 비롯한 가상화폐를 “화폐”로 취급하고, 이러한 인식이 가격에 미치는 영향을 다룬 것이다. 초창기 비트코인이 등장했을 때, 이를 화폐로 볼 것인지에 대한 논의와 함께, 기존 제도권 금융 시스템 내에서 통용될 수 있는지를 주로 다뤘다.

두 번째 범주는 가상자산의 가격 결정 메커니즘과 이를 활용하는 방안에 대한 연구이다. 여기서는 채굴 동기, 수요-공급 이론, 시장 크기, 수익 변동성, 채굴 비용 등을 다루었다. 그리고 최근에는 비트코인 가격과 경제/금융 지표와의 상관성을 조사하며, 다우존스지수, 미국 달러 선물 지수, 원유 가격, 금 가격과의 관계를 분석하는 연구로 확대되는 추세이다.

<표 2-1> 주요 선행 연구

연구자	주요내용
Kristoufek (2015)	암호화폐에 대한 화폐의 특성 분석 및 가격의 결정 요인에 대한 연구
Ciaian et al. (2016)	암호화폐의 수요와 공급 요인 분석 및 가격 형성 과정에 대한 연구
Wang et al. (2016)	암호화폐를 비롯한 가상자산의 금융시장에서의 리스크 헤징 가능성에 대한 연구
Hayes (2018)	가상자산의 채굴 비용 등 공급에 미치는 요인에 대한 연구
장성일 · 김정연 (2017)	가상자산의 가격에 영향을 미치는 요인 및 금 등 선물 시장과의 관계 대한 연구
이준식 외 2인 (2018)	가상자산의 가격과 국내외 주식 시장과의 상관관계에 대한 연구
Thies and Molnár (2018)	가상자산의 수익 변동성과 리스크 측정 및 변동성에 대한 연구

2023년 12월 현재 미국 증시에서는 가상자산의 채굴 및 블록체인을 보유하고 있는 기업에 투자한 ETF가 상장돼 있으며, 2024년 1월에는 비트코인에 직접 투자하는 현물 ETF가 본격적으로 운영될 예정이다.

이에 따라 증권시장의 애널리스트처럼 투자 대상인 가상 자산의 가격 움직임을 분석하고 이에 대한 예측하는 시도가 늘어날 것으로 예상된다.

기존 연구는 ARIMA(Auto-Regressive Integrated Moving Average)나 벡터자기회귀 모형(Vector AutoRegressive Model)과 같이 전통적인 시계열 모형을 사용한 것이 대부분이었다. 본 연구에서는 시계열 데이터 적합한 여러 인공지능 모델을 활용하여 가상자산의 가격을 예측하고자 한다. 본 논문의 연구기간은 2019년부터 2023년으로 했는

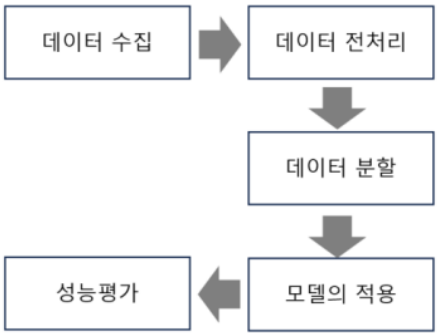
데 이는 가상자산이 어느 정도 정착되고, 이를 활용한 ETF 등이 상장된 시기이다. 따라서 기존 연구의 분석 기간에 비해 성숙된 시장을 다뤘다는 점에서 차별점이 있다.

아울러 시계열 데이터 특성을 반영한 정확도 지표를 산정했으며, 이는 최근 상장된 비트코인 현물 ETF 투자 전략 수립 등에 기여할 것으로 예상된다.

Ⅲ. 연구 방법 및 모형

1. 연구 방법

본 연구에서는 야후 파이낸스의 금융 데이터를 사용했다. 모형에 사용한 데이터는 모두 일단위 데이터로, 종가(Close Price)를 사용하였다. 이후의 연구 순서는 아래 [그림 3-1]과 같다. 모형에 사용한 데이터는 전일 대비 가격 변화율(%)로 변환해 사용했다. 가장자산 중에 가장 거래규모가 큰 비트코인을 산정했다.



[그림 3-1] 연구 순서

시계열 모형을 비롯한 각 인공지능 모형의 성능검증을 위해서 전체 데이터의 95%를 훈련데이터로 사용해 모형을 학습했다. 나머지 5% 데이터를 검증데이터로 사용해 모형의 성능을 평가했다. 모델의 성능평가를 위해서 정확도, 손실 등의 지표를 계산했다.

2. 변수 선정 및 분석 기간

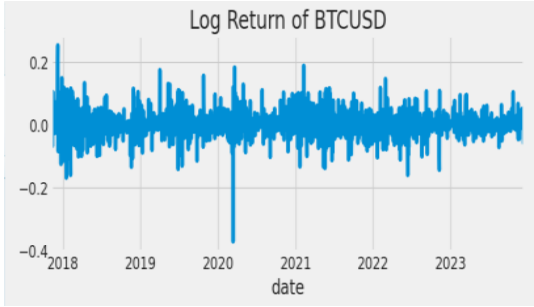
본 연구에서 사용한 변수는 시계열 변수로서 그 값을 그대로 사용할 경우 가성회귀 및 자기상관성 등의 통계적인 문제가 발생한다. 이 때문에 아래 <식 4.1>과 같이 전기 대비 증감액인 차분 변수를 사용했다.

$$R_{i,t} = r_{i,t} - r_{i.\,t-1} \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad <식\ 3.1>$$

(여기서 $R_{i,t}$는 t시점의 값에서 t-1시점의 값을 차감한 차분변수로 순 증감을 나타냄)

따라서 본 연구에서 사용된 모든 변수는 가격 지수의 경우에는 전일 대비 가격 변화율(%)이 되며, 거래량의 경우는 전일 대비 거래량 변화율이다. 또한 시작일자는 2019년 11월 10일이며, 종료 일자는 2023년 12월 13일이다.

[그림 3-2] 비트코인의 기간별 가격 추이



[그림 3-3] 비트코인의 기간별 가격 변화율 추이

<표 3-1>은 연구 모형의 구성 및 평가를 위해 사용한 데이터 세트로 Training Set은 2017년 11월 10일부터 2023년 8월 24일까지의 비트코인 가격으로 총 개수는 2003개이다. Test Set은 2023년 8월 25일부터 2023년 12월 13일까지로 개수는 112개이다.

<표 3-1> 데이터 셋 및 기간 구분

데이터 세트	기 간	표본일수
Training Set	2017.11.10. ~ 2023.08.24	2002일 (95%)
Test Set	2023.08.25. ~ 2023.12.13	112일 (5%)
전체기간	2017.11.10.~ 2023.12.13	2114일 (100%)

3. 연구 모형

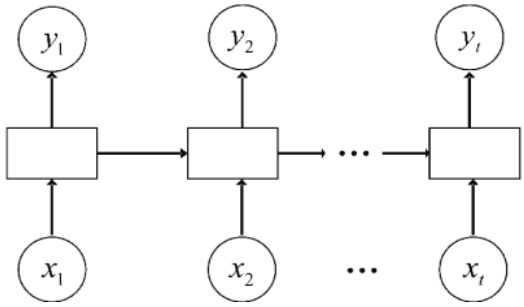
본 연구에서 사용하는 인공지능 모형은 LSTM (Long Short-Term Memory), RNN (Recurrent Neural Network), HMM (Hidden Markov Model), Prophet 모형이다. 이들 모형은 다양한 방식으로 시퀀스 데이터를 다루고 시간적 또는 순차적인 정보를 처리한다.

본 연구에서 사용한 시계열 모형은 ARIMA (AutoRegressive Integrated Moving Average) 모형이다. ARIMA는 AR(Autoregressive), I(Integrated), MA(Moving Average)의 세 가지 구성요소로 이루어져 있다. AR 요소는 자기 자신의 과거 값을 참조하여 예측하는 것이며, MA 요소는 과거의 오차항을 참조하여 예측하는 것이다. I 요소는 시계열 데이터가 비정상(non-stationary)일 때, 이를 정상성을 가지도록 차분하는 단계를 의미한다.

ARIMA 모형은 데이터 탐색 → 차분 → ACF 및 PACF 분석의 순으로 모형을 도출한다.

첫 번째 단계인 시계열 데이터 탐색에서는 주어진 시계열 데이터를 시각화하고 기본 통계량을 살펴봅니다. 데이터의 경향성, 계절성, 이상치 여부 등을 파악한다. 이후 차분(Differencing) 과정을 정상성을 확보한다.

이후에는 ACF(Autocorrelation Function)와 PACF (Partial Autocorrelation Function) 그래프를 그려서 자기상관 및 부분 자기상관의 시차(lag) 구조를 확인한다. 이후 도출된 각 파라미터 값들을 사용해서 모델을 훈련한다.



[그림 3-4] RNN 모형

RNN(Recurrent Neural Network)모형은 시계열 데이터, 자연어 처리, 음성 인식 등 순차적 데이터를 처리하는 데 사용되는 인공 신경망의 한 유형이다. RNN은 순환 구조를 가지며, 이전 단계의 출력을 현재 단계의 입력으로 사용하여 순차적 정보를 처리하고 기억할 수 있다.

RNN 모델은 위 그림과 같이 첫번째 셀(cell)의 output이 두 번째 셀의 input으로 사용된다. 이 과정에서 두 번째 셀은 이전 셀의 상태(status)와 x^2를 입력받아서 y^2를 반환한다. 즉 이전의 상태가 다음의 결과값을 반환하는 데 사용된다.

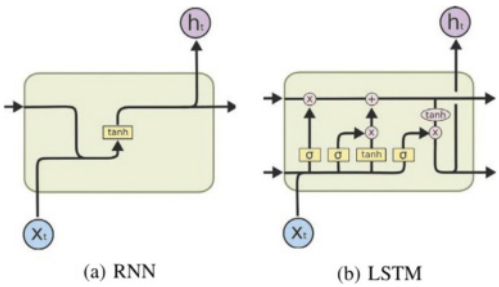
따라서 이러한 모델 특성 때문에 시계열데이터, 자연어를 다루는 모형에서 많이 사용된다. 다만 자연어의 경우 문장 길이가 매우 길어지거나 시계열 데이터의 경우 데이터셋이 커질 때, 즉 데이터가 너무 길어 이전 입력값을 잊어 성능이 저하되는 문제가 발생한다. 이를 장기의존성 문제(Long Term

Dependency)라고 하는데 이 문제를 해결하기 위해 도입된 모형이 LSTM이다.

LSTM(Long Short-Term Memory)모형 은RNN 모델과 마찬가지로 시계열 데이터, 자연어 처리, 음성 인식 등 순차적 데이터를 처리하는 데 사용되는 인공 신경망 모델이다.

RNN 모델은 시간에 따라 의존성이 길어질수록 그래디언트 소실((Gradient Vanishing Problem) 또는 폭발(Gradient Explosion Problem) 문제가 발생하여 제대로 학습하지 못할 수 있다. RNN 모형처럼 하나의 network를 계속 복사해서 순서대로 정보를 전달하는 하는 방식은 과거의 정보를 기반으로 해서 계속해서 축적하는 방식이라 할 수 있다.

하지만 우리가 생각하는 과정은 단순히 과거의 정보를 계속적으로 반복하는 것에 머무르지 않는다. 즉 특정 정보가 계속해서 영향을 미치게 되는 경우나, 아니면 반대로 특정 정보가 일정 기간을 넘어서면 더 이상 영향을 미치지 못하는 경우가 발생한다.

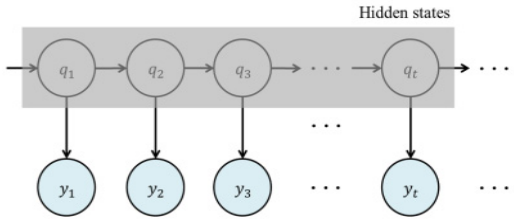


[그림 3-5] RNN과 LSTM 모형 비교

LSTM (Long Short-Term Memory)은 위 그림과 같이 과거의 정보가 향후 프로세스에 영향을 미치지만, 이의 지속성 여부를 “판단”하는 게이트 메커니즘이 추가돼 있다.

따라서 정보의 영향력을 일정 기간 이상 유지하거나 제어할 수 있도록 하기 때문에 RNN의 한계점을 극복한 것이라 평가받는다.

은닉 마르코프 모델(Hidden Markov Model, HMM)은 시계열 데이터나 시퀀스 데이터를 다루는 확률적 모델로, 주로 자연어 처리, 음성 인식, 음악 분석, 생물 정보학 등 다양한 분야에서 활용된다.



[그림 3-6] HMM모델

HMM은 학습 디코딩, 예측의 과정의 프로세스로 구성돼 있다. 우선 학습(Learning) 단계는 관측 데이터를 통해 HMM의 모델 파라미터(상태 전이 확률 및 관측 확률)를 추정하는 과정이다.

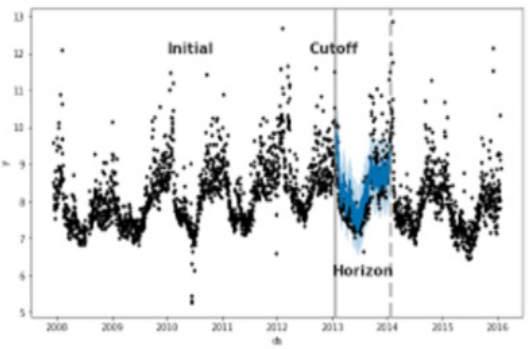
디코딩(Decoding) 단계에서는 주어진 모델과 관측 데이터로부터 가장 확률 높은 순서의 숨겨진 상태 시퀀스를 찾는다. 마지막으로 예측(Prediction) 단계에서는 도출된 모델을 사용하여 미래의 관측 데이터나 상태 시퀀스를 예측한다.

HMM은 이전 상태의 정보만 고려하여 미래 상태나 관측 데이터를 예측하므로, 일부 시나리오에서는 제약이 있을 수 있다. 하지만 많은 응용 분야에서 유용하게 활용되며, 특히 시계열 데이터나 시퀀스 데이터에 대한 모델링에 적합한 것으로 평가받고 있다.

마지막으로 Prophet 모델은 페이스북에서 개발한 모형으로 시계열 데이터에 특화됐다.

$$y(t) = g(t) + h(t) + s(t) + e_t$$

$y(t) : Addictive\ Regressive\ Model$
$g(t) : Trend\ Factor$
$h(t) : Holtday\ component$
$s(t) : Seasionality\ component$
$e(t) : Error\ Term$



[그림 3-7] Prophet 모형

Prophet은 [그림 3-7]와 같이 주기성 패턴, 연간, 월간, 주간 및 일간의 계절성을 고려한다. 자동으로 시계열 데이터를 분석해서 감소 추세, 주기성, 휴일 영향과 같은 시계열 데이터의 구성 요소를 모델링 하기 때문에 사용하기 쉽다는 장점이 있다.

이 때문에 prophet 모델은 주식 가격 예측, 판매 예측, 수요 예측, 마케팅 데이터 분석, 트래픽 예측 등 다양한 분야에서 활용되고 있다.

이상 각 모형의 주요 내용을 정리하면 아래 <표 3-2>와 같다.

<표 3-2> 연구모형 및 주요 특징

연구모형	주요특징
ARIMA 모형	시계열 모형으로 트렌드, 계절성 요인을 잘 반영하는 것으로 평가 받고 있음
RNN	시계열 데이터에 적합한 신경망으로, 과거 정보를 기억하면서 새로운 입력에 반응하는 메커니즘을 사용
LSTM	RNN의 모델의 장기 의존성 문제를 해결한 인공지능 모형
HMM	시계열 데이터에서 시간에 따라 변화하는 시스템의 상태를 모델링하기 위해 사용
Prophet 모형	페이스북에서 개발한 시계열 예측 모델로, 계절성 요인과 휴일 효과를 잘 반영하는 것으로 평가 받고 있음

3. 모형의 성능 평가 기준

일반적으로 어떤 값(value)의 정확도를 예측하는 회귀모형에서는 아래와 같이 RMSE나 SMAPE로 성능을 평가한다.

[식 3-2] 모형의 성능 평가기준 지표 사례

$$\mathrm{RMSE} = \sqrt{\frac{1}{nT}\sum_{i,t}(\hat{y}_{i,t} - y_{i,t})^2}$$

$$SMAPE = \frac{100}{n} \times \sum_{i=1}^{n} \frac{\left|Y_i - \widehat{Y_i}\right|}{\left(|Y_i| + |\widehat{Y_i}|\right)/2}$$

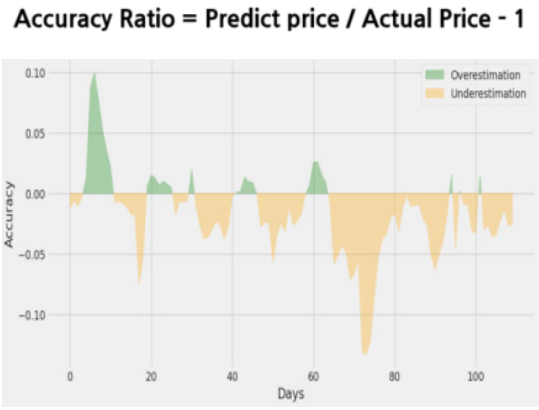
RMSE(Root Mean Squared Error)[1]는 시계열 모델의 성능을 평가하는 데 사용되는 일반적인 평가 지표 중 하나로, 예측한 값과 실제 관측값 사이의 차이를 측정하여 모델의 예측 정확도를 뜻한다.

이외에도 SMAPE를 사용하는데, SMAPE는 우리말로 대칭 평균 절대 백분율 오차(Symmetric Mean Absolute Percentage Error)로, 시계열 예측의 정확도를 측정하는 데 사용되는 지표 중 하나이다.

본 연구에서는 위의 두 지표 대신에 실제값과 정확한 값의 괴리율로 지표를 구성하여 비트코인의 가격을 예측했다.

예측한 값이 실제 값과 일치할수록 Accuracy Ratio는 0에 가까워진다. Test기간에 대해서 Accuracy Ratio를 계산했으며, 이 때 Accuracy Ratio의 평균, 표준편차 등을 계산하면 최종적으로 아래 [그림 3-8]와 같이 결과를 얻을 수 있다.

1) Hands-On Machine Learning with Scikit-Learn and TensorFlow(2017), 42p



[그림 3-8] 모형의 정확도 지표

이와 같은 지표를 사용하면 다음과 같은 기대효과가 있다. 첫 번째로 예측 오차를 해석하기 쉽다. 특히 오차가 얼마나 큰지 또는 어떤 방향으로 편향되어 있는지 쉽게 이해할 수 있어 각 모델의 비교에 용이하다.

또한 이 같은 괴리율 지표를 사용하면 모델의 편향을 탐지하기 용이하다. 어떤 예측이 실제 값과 얼마나 멀리 떨어져 있는지를 나타내므로, 모델이 특정 방향으로 편향되어 예측하는지 여부를 파악할 수 있다.

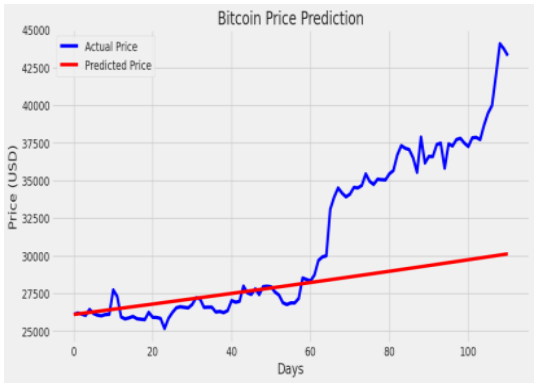
Ⅳ. 실증 분석

1. 각 모형 적용 분석 결과

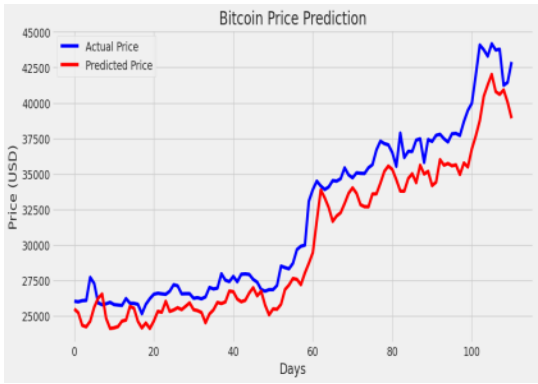
아래 [그림 4-1]부터 [그림 4-5]까지는 Test Set에 대한 비트코인 가격의 실제치(Actual Price)와 예측치(Prediction)를 나타낸 것이다.

비트코인 가격 데이터는 시계열 데이터로, 이전 시점의 데이터가 현재 가격에 영향을 미칠 수 있다. 따라서, LSTM 및 RNN과 같은 순차 모델은 이전 시점의 정보를 고려하여 현재 시점의 예측을 수행할 수 있다. 특히 이 두 모델은 데이터의 시간적 의존성을 잘 찾아낼 수 있다.

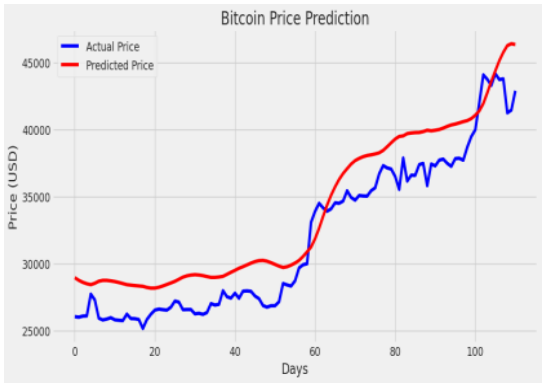
즉, 신경망 모형은 다른 모형에 비해서 시계열 데이터의 시간적 의존성, 장기 의존성, 비선형 등의 특징을 잘 반영하기 때문에 예측 정확도가 높은 것으로 보인다.



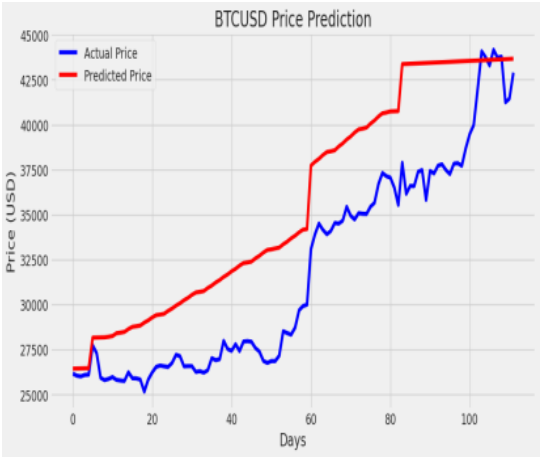
[그림 4-1] ARIMA모형



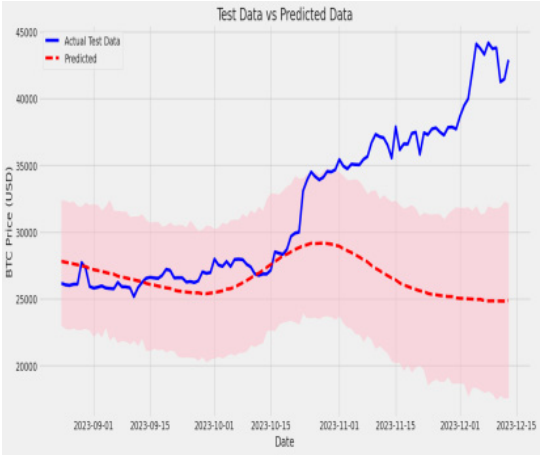
[그림 4-2] RNN모형



[그림 4-3] LSTM모형



[그림 4-4] HMM모형



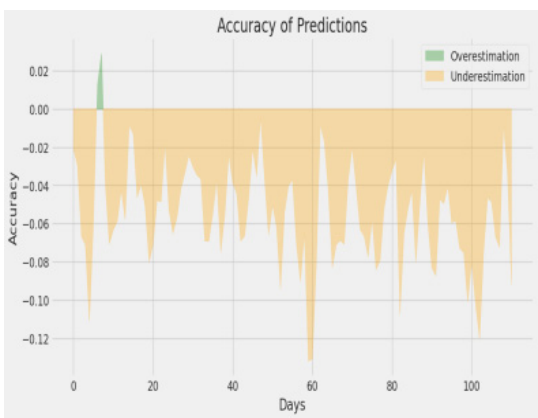
[그림 4-5] Prophet모형

이번에는 모형의 성능 평가 지표인 Accuracy Ratio를 이용해서 LSTM, RNN 모델 중 어느 모형이 더 정확하게 예측하는지 살펴봤다.

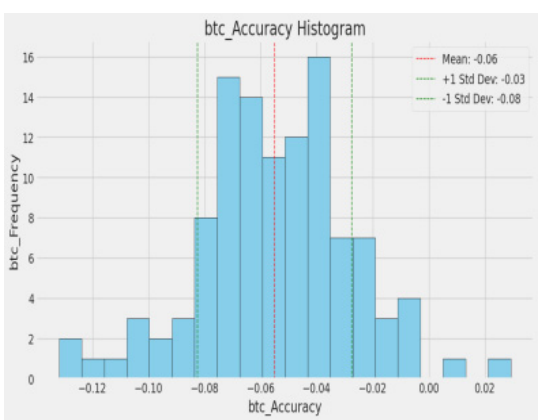
<표 4-1>에서와 같이 RNN 모델은 실제 비트코인 가격에 비해서 -5.5% 과소 예측한 것으로 나타났는데, 이는 모형 중에서 가장 높은 값이다. LSTM 모형은 RNN 모델과 반대로 실제 비트코인 가격에 비해서 6.7%로 과대 예측한 것으로 나타났다.

<표 4-1> 모형 별 Accuracy-Ratio 결과

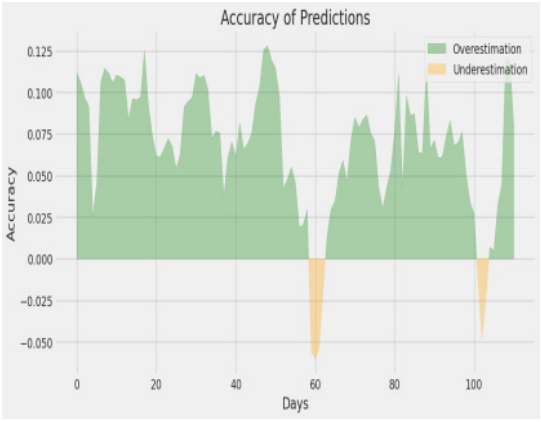
모형	평균	분산	표준편차
RNN 모델	-0.055 (-5.5%)	0.001	0.02 (2%p)
LSTM 모델	0.0670 (6.7%)	0.002	0.04 (4%p)
ARIMA모델	-0.076 (-7.6%)	0.0013	0.11 (11%p)
HMM	0.123 (12.3%)	0.003	0.056 (5.6%p)
Prophet 모델	-0.137 (-13.7%)	0.024	0.16 (16%p)



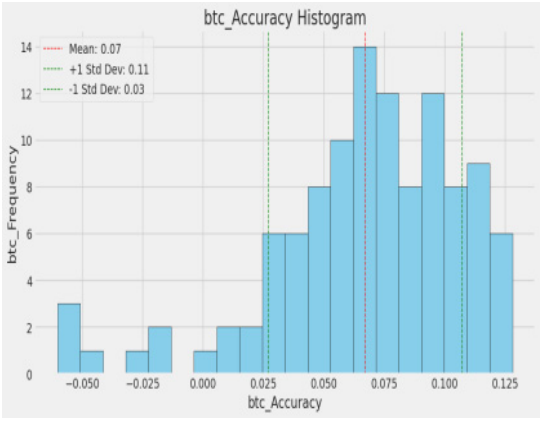
[그림 4-6] RNN모형의 정확도



[그림 4-7] RNN모형의 정확도 분포



[그림 4-8] LSTM모형의 정확도



[그림 4-9] LSTM 모형의 정확도 분포

이번에는 과소 및 과대 예측의 경향 정도를 파악하기 위해서 RNN모델과 LSTM모델의 Accuracy -Ratio의 분포도를 살펴보았다.

[그림 4-6]에서와 같이 RNN은 전체적으로 과소 예측하는 경향이 있는 것으로 나타났다. 반면 LSTM 모형은 [그림 4-8]에서와 같이 과대 예측하는 경향을 보였다.

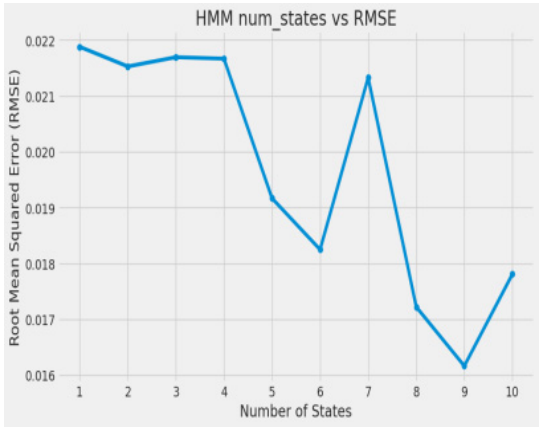
신경망 모형을 제외한 ARIMA모형, HMM. Prophet 모형을 평가할 경우에는 ARIMA모델이 HMM, Prophet모형에 비해 정확도가 높았으며, HMM, Prophet모형은 대체로 낮은 정확도를 보였다.

2. 각 모형의 비교 결과

각 모형에 대한 평가 결과 LSTM 및 RNN 모델이 다른 모형에 비해 가격을 정확히 예측하는 것으로 나타났다. 이는 이들 모형의 시계열 데이터의 특성, 즉 이전 시점의 정보를 다른 모델에 비해 잘 고려했기 때문으로 보인다.

예를 들어 RNN, LSTM 모형에서는 현재 t시점을 기간으로 일정 기간 이전의 정보(여기서는 60일 전)를 반영한다.

즉, 오늘이 12월 1일이라면 60일전의 가격정보까지가 오늘에 미치고, 내일은 어제 보다 하루 뒤의 정보와 바로 어제 정보가 영향을 순차적으로 미치는 방식이다.



[그림 4-10] HMM의 최적의 상태 수 도출

본 논문에서는 모형의 예측 정확도를 기준으로 각 모형의 성과를 평가했지만, 모형의 예측 정확도만으로 모형을 평가하는 것은 한계가 있다.

예를 들어 HMM 모형의 경우에는 상태의 개수(Number of states) 값에 따라서 정확도가 달라질 수 있다. 여기서는 위 [그림 4-10]과 같이 최적상태의 수를 9로 산정해 모형을 도출했다.

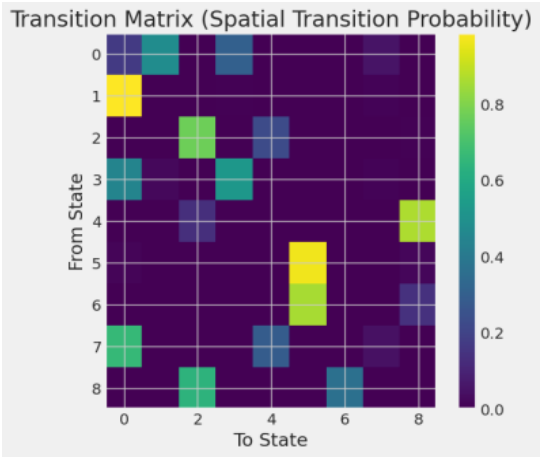
HMM은 어떤 가격의 정확한 값을 예측하기 보다는 어떤 상태로 이동하는지를 예상하는 문제에

더 적합하다.

위 [그림 4-11]는 “Transition Matrix (Spatial Transition Probability)”를 시각화 한 것이다. 이는 상태 간의 전이 확률을 나타내는 것으로 ‘From State’와 ‘To State’는 각각 특정 상태에서 출발하여 다른 상태로 이동할 확률을 나타내는 축이다.

여기서 색상 막대는 확률 값을 나타낸다. 연한 색 (노랑색 계열)은 높은 전이 확률을, 어두운 색 (보라색 계열)은 낮은 전이 확률을 나타낸다.

색상 막대의 오른쪽에 있는 범례는 색상과 확률 값 사이의 대응 관계를 보여 준다. 0(0%)에서 1.0(100%)까지의 확률 값이 있으며, 이 범위 내에서 색상의 변화가 확률의 변화를 나타낸다.



[그림 4-11] HMM의 상태 전이 확률

대각선에 해당하는 셀들이 밝은색 보이는데, 이는 어떤 상태가 같은 상태로 유지될 확률이 높음을 나타낸다. 다른 셀들은 상태가 다른 상태로 변할 확률을 보여주고, 이 셀들 중 일부는 높은 확률을, 다른 일부는 낮은 확률을 나타낸다.

이 행렬은 마르코프 체인, 게임 이론, 경제 모델링, 생물학적 과정 모델링, 웹 페이지 랭킹 시스템 (예: 구글의 PageRank 알고리즘) 등 다양한 분야에서 상태 간의 전환을 모델링할 때 사용될 수 있다.

HMM은 어떤 가격의 정확한 값을 예측하기 보다는 어떤 상태로 이동하는지를 예상하는 문제에 더 적합하다.

한편 Prophet 모형은 아래 [그림 4-7]와 같이 어떤 정확한 값을 예측하기 보다는 상한/하한 값의 범위를 산정한다.

따라서 Prophet 모형은 어떤 가격의 정확성 보다는 가격의 상한, 하한의 확률 즉, 범위를 예측하는데 적합한 모형이라 할 수 있다.

본 연구에서 사용한 ARIMA 모형은 비트코인의 가격을 예측하기 위해 과거의 비트코인 가격 외에 다른 정보를 반영하기가 어렵다. 실제로 ARIMA 모형으로는 분석 기간 동안의 기준금리, VIX지수 (공포지수), S&P500, 나스닥종합 지수 등의 시장 지수 등 다양한 정보를 반영하기가 어렵다.

Ⅴ. 결론

본 논문에서는 가상자산인 비트코인의 시계열 데이터를 이용해 시계열 모형을 비롯한, 여러 인공지능 모형을 활용해 미래 가격을 예측했다. 이를 위해 시계열 데이터의 특성을 이용해, 과거 정보의 의존경향을 반영하는 LSTM, 페이스북(메타)에서 개발한 구조적 상태 변화를 포착하는데 사용되는 Prophet, 시계열 데이터의 구조적인 변화를 모델링하는데 사용되는 HMM(Hidden Markov Model)을 선정했다. 이후 각 모델의 성능을 평가하고, 기존 시계열 모델인 ARIMA와 성과를 비교했다.

전체 분석 기간을 2개로 구분해서 Training 기간과 Test 기간 2개로 구분했다. 각각의 비중은 Training 기간이 95%이고 Test 기간이 5%이다. 각 모형의 성과 평가는 Test 기간에 대해서 각 모형으로 산출한 Predict price와 실제 가격인 Actual price를 차이를 나타내는 정확도 지표를 사용했다. 분석 결과 신경망 모형인 RNN, LSTM이 높은 정확도를 보였다.

이는 시계열 데이터의 시간적 의존성, 장기 의존성, 비선형 등의 특징을 잘 반영하기 때문에 예측 정확도가 높은 것으로 보인다.

RNN 모델은 실제 비트코인 가격에 비해서 -5.5% 과소 예측한 것으로 나타났는데, 이는 모형 중에서 가장 높은 값이다. LSTM 모형은 RNN 모델과 반대로 실제 비트코인 가격에 비해서 6.7%로 과대 예측한 것으로 나타났다.

RNN, LSTM 모형에서는 현재 시점을 기간으로 일전 기간 이전의 정보를 반영하는 과정이 포함돼 있다.

예를 들어, 오늘이 12월 1일이라면 T기간 전의 가격정보까지가 오늘에 미치고, 내일은 어제 보다 하루 뒤의 정보와 바로 어제 정보가 영향을 순차적으로 미치는 방식이다.

하지만 모형의 예측 정확도만으로 모형을 평가하는 것은 한계가 있다.

HMM은 어떤 가격의 정확한 값을 예측하기 보다는 어떤 상태로 이동하는지를 예상하는 문제에 더 적합하다. 가격의 하락, 상승 가능성과 같이 움직임의 방향을 판단하는 경우에는 HMM모델이 보다 적합하나도 할 수 있다.

또한 Prophet 모형은 어떤 가격의 정확성 보다는 가격의 상한, 하한의 확률 즉, 범위를 예측하는데 적합한 모형이라 할 수 있다. 따라서 가격의 변동으 범위를 예측하는데 보다 적합한 모형이다.

본 논문에서 고려한 시계열 모형은 ARIMA모형으로 비트코인 가격 외에 다른 변수를 고려하기 어렵다. 따라서 VAR 모형처럼 기준금리, VIX지수(공포지수), S&P500, 나스닥종합 지수 등의 시장 지수 등 다양한 정보를 반영한다면 더 높은 성과를 보일 것으로 기대한다.

참고문헌

이수용, 이경중 (2011). 시계열 자료의 데이터마이닝을 위한 패턴분류 모델설계 및 성능비교, 한국지능시스템학회 21(6), 730 - 736.

이준식, 김건우, 박도형 (2018). “비트코인 가격변화에 관한 실증분석”, 지능정보연구, 24(2), 195-220.

임병권, 윤평식, 박순홍(2016). “애널리스트의 정보력과 투자자별 거래행태: IPO 기업을 대상으로”, 한국증권학회지, 45(5)(2016), 971-999.

장성일, 김정연 (2017). 비트코인의 자산성격에 관한 연구, 한국전자거래학회지, 22*(4),* 117-128.

Hands-On Machine Learning with Scikit-Learn and TensorFlow (2017).

Hayes, A. S(2018), Bitcoin price and its marginal cost of production: support for a fundamental value, *Applied Economics Letters 26*(7), 554-560.

Kristoufek, L.(2015), “What are the main drivers of the Bitcoin price? Evidence from wavelet coherence analysi”s, *PloS one 10* (4), p. e0123923.

Polasik, M., A. I. Piotrowska, T. P. Wisniewski, R. Kotkowski, and G. Lightfoot(2015), “Price fluctuations and the use of Bitcoin: An empirical inquiry”, *International Journal of Electronic Commerce 20*(1), pp. 9-49.

투고일자: 2024. 1. 2.

심사일자: 2024. 1. 23.

게재확정일자: 2024. 2. 13.

미래사회
Journal of Future Society
2024. Vol. 15, No. 1, 21~33.

Analysis and Prediction of Crypto Currency Time Series Data

Comparison of time series models and artificial intelligence models and suggestions for improvement

Ahn, Sang-Sun

Kookmin University

This study examines various artificial intelligence models' effectiveness in predicting virtual asset prices such as Bitcoin and Ethereum, which, unlike traditional assets, lack a tangible basis for valuation. Given the absence of a reliable pricing mechanism like CAPM or DCF models for virtual assets, the research leverages time series data since 2019, incorporating the pandemic period. It employs models such as LSTM, which accounts for past information dependencies; Prophet, developed by Facebook (Meta) for detecting structural changes; and HMM for modeling data structural shifts. These models were tested against ARIMA, considering factors including pandemic timelines and U.S. Federal Reserve rate decisions. Results showed that the neural network-based models, RNN and LSTM, outperformed ARIMA, attributed to their superior handling of time series data characteristics. However, further investigations into the impact of external factors like pandemic declarations and interest rate changes on model accuracy and virtual asset price prediction are needed.

Keywords: Virtual assets, Bitcoin, artificial intelligence, time series, cryptocurrency