JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 25, No. 10: 1287~1295, Oct. 2021

On-Chain Data를 활용한 LSTM 기반 비트코인 가격 예측

안유진[1] · 오하영[2*]

Utilizing On-Chain Data to Predict Bitcoin Prices based on LSTM

Yu-Jin An[1] · Ha-Young Oh[2*]

[1]Graduate Student, Department of Fintech, Sungkyunkwan University, Seoul, 03063 Korea
[2*]Associate professor Professor, College of Computing & Informatics, Sungkyunkwan University, Seoul, 03063 Korea

요 약

최근 10여 년 동안 가장 가파르게 가치가 상승한 자산군을 꼽자면 단연 비트코인이라고 할 수 있을 것이다. 특히 비트코인은 중앙통제 기관이 없음에도 불구하고 첫 등장을 한 2009년의 사실상 0달러에서 2021년 최고점인 65,000 달러 수준까지 치솟아 역사에 길이 남을 가치 상승을 보여주었다. 이에 따라 비트코인의 가능성에 대해서 반신반의 했던 상당수 투자자들의 포트폴리오에도 비트코인이 상당한 비중을 차지하는 경우가 많아졌으며, 제도권 내의 금융권에서도 이런 비트코인의 움직임에 주목하고 있다. 비트코인에 대한 관심과 더불어 비트코인의 가격에 거시경제 변수나 센티멘트가 비트코인의 가격이 어떻게 움직이는가에 대한 연구 또한 상당히 진전되었다. 하지만, 이들 연구에서 활용한 변수들은 비트코인만의 특징적인 데이터라고 할 수 있는 블록체인 내의 데이터를 취합하여 가공한 온체인 데이터를 적극적으로 활용하지는 않았다. 따라서, 본 논문에서는 시계열 데이터 예측에 적극적으로 활용되고 있는 LSTM을 기반으로 온체인 데이터를 활용하여 비트코인의 가격을 예측해보고자 한다.

ABSTRACT

During the past decade, it seems apparent that Bitcoin has been the best performing asset class. Even without a centralized authority that takes control over, Bitcoin, which started off with basically no value at all, reached around 65000 dollars in 2021, showing a movement that will definitely go down in history. Thus, even those who were skeptical of Bitcoin's intangible nature are stacking bitcoin as a huge part of their portfolios. Bitcoin's exponential growth in value also caught the attention of traditional banking and investment firms. Along with the spotlight Bitcoin is getting from the investment world, research using macro-economic variables and investor sentiment to explain Bitcoin's price movement has shown progress. However, previous studies do not make use of On-Chain Data, which are data processed using transaction data in Bitcoin's blockchain network. Therefore, in this paper, we will be utilizing LSTM, a method widely used for time-series data prediction, with On-Chain Data to predict the price of Bitcoin.

키워드 : 비트코인, 디지털 금, 온체인 데이터, LSTM, 가격

Keywords : Bitcoin, Digitial-gold, On-Chain Data, LSTM, Price

Received 26 May 2021, **Revised** 22 June 2021, **Accepted** 9 July 2021

*** Corresponding Author** Hayoung Oh (E-mail: hyoh79@gmail.com Tel:+82-2-583-8585)
Associate Professor, College of Computing & Informatics, Sungkyunkwan University, Seoul, 03063 Korea

Open Access http://doi.org/10.6109/jkiice.2021.25.10.1287 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

최근 몇 년간 비트코인은 투자자뿐만 아니라 정치계 그리고 미디어로부터 수없이 조명받아 왔다. 2009년 비트코인이 보급된 이래로 1달러에서 2017년 1월, 20,000달러 선까지 넘보면서 비트코인을 가치저장 수단으로 인정하는 분위기가 형성되기 시작했다. 정치계에서는 비트코인의 탈중앙성과 익명성 그리고 전통적인 자산과 비교해 심한 변동성을 두고 우려 섞인 시선을 보내기도 하였다. 이런 시선에도 불구하고 비트코인은 높은 이윤을 추구하고자 하고자 하는 투자자들로부터 꾸준한 관심을 받아왔으며 2021년 4월 22일 기준 2017년 전고점인 20,000달러를 훨씬 상회하는 54,000달러 수준을 유지하고 있다.

이런 랠리와 함께 비트코인이나 알트코인(altcoin)에 대한 관심이 나날이 치솟고 있다. 대규모 금융 투자 기관들뿐만이 아니라 일반 기업체들도 암호화폐를 결제 수단으로 인정하고자 하고 있으며 이에는 2021년 2월의 랠리를 이끈 주역인 테슬라도 포함된다. 또한, 현재 세계 시총 1위를 다투고 있는 애플도 애플 페이(Apple Pay)를 기반으로 암호화폐 결제를 받아들일 수 있다는 얘기[1]가 나오고 있다. 2018년 초 급락을 목격하여 불신에 사로잡혀 있던 개인들조차 다시금 비트코인에 대해 눈독을 들이고 투자를 시작하고 있다.

이처럼 호재가 넘치는 상황이고 암호화폐의 기반이 전과 비교해 탄탄해지고 있으나 여전히 암호화폐 시장이 전통적 자산과 비교해 변동성이 크고 그에 따라 위험이 크다는 것은 의심할 여지 없는 사실이다. 앞서 언급한 대형 금융 회사들이나 애플, 테슬라 같은 자금 여력이 있는 회사들은 큰 하방 변동성에 상대적으로 여유를 갖고 대처할 수 있겠지만 대체로 자금의 시간적 여유가 없는 개인은 암호화폐의 작은 하락에도 훨씬 크게 반응한다. 또한, 반응하지 않았을 때 폭락장이 찾아오면 이는 일반적인 투자자들에게 금전적으로 치명적인 손실을 입힐 수도 있다. 특히 중앙통제 기관이 없는 암호화폐 시장의 특성상 가격 하락에 대한 압박이나 공포가 여타 자산보다 훨씬 크게 다가온다.

그동안 많은 연구가 암호화폐를 기존의 자산시장에 대입하여 연구를 지속해왔다. 특히 여러 자산 중 비트코인 등의 암호화폐 시장과 가장 비슷해 보이는 주식 시장과 비슷하게 거시 경제 지표를 변수로 두고 연구를 진행해왔다[2-5]. 하지만 해당 연구는 비트코인이나 암호화폐도 암호화폐만의 특징을 잘 드러내는 변수들이 있다고 보고 그들 중 블록체인 데이터를 직접적으로 분석하여 생성되는 온체인 데이터(On-Chain)를 통해 비트코인 가격 예측을 시도해보고자 한다. 이 시도는 개인 투자자들에게 암호화폐에 투자할 시에 어떤 부분에 집중해야 하는지에 대한 새로운 관점을 선사해 변동성이 큰 시장에서 투자의 위험을 줄이는 데 도움이 되리라 기대한다.

Ⅱ. 선행 연구 및 연구 동향

비트코인이 주류 자산으로 차츰 자리매김함에 따라 이에 대한 다양한 연구가 이루어졌다. 비트코인 연구 초기에는 가격 예측 연구보다 무엇이 비트코인에 가치를 부여하고 어떤 변수들이 가격을 결정하는 데 영향을 주는지에 대한 연구가 선행되었다. 이러한 변수로는 금 가격[2], 인플레이션이나 고용률 지표[3-4], 주가지수 같은 거시 경제 지표[5] 등과 트위터나 뉴스 센티멘트[2, 5, 6]가 있다.

최근에는 거대 기업, 투자 은행, 헤지 펀드 등이 비트코인을 인플레이션에 대한 헤지(Hedge) 수단으로 비트코인 잔고를 쌓기 시작하면서 '디지털 금(Digital Gold)'이라는 속칭까지 생겨나기 시작하였다[7]. 이러한 대형 기관들의 개입은 필연적으로 현재의 주식이나 파생상품 시장처럼 트레이딩 알고리즘에 대한 관심을 불러일으켰다. 그리고 이러한 트레이딩 알고리즘이 만들어지기 전에 필수적으로 거쳐야 하는 단계인 가격 예측에 대한 관심 또한 상당해졌다. 학자들과 퀀트 트레이더들은 머신러닝, 딥러닝 같은 문제해결 알고리즘을 활용하여 정확한 가격 예측을 목표로 지속해서 연구해왔다.

앞서 언급한 대로 상당수의 연구는 비트코인의 가격에 영향을 줄 만한 트레이딩 볼륨과 같은 거래소에서 보이는 변수들이나 거시경제적인 변수들, 혹은 센티멘트를 위주로 연구가 진행됐다. 그러나 가격[8]만을 활용하거나 Hash Rate나 Mining Difficulty[9] 등의 온체인 데이터를 활용한 연구도 있다. 하지만 논문[9]은 비트코인 블록에 명시적으로 기록된 데이터만 활용했을 뿐, 거래소 내 비트코인 잔고처럼 비트코인 체인에 기록된 거래를 '분석'하여 얻어낸 데이터는 아니다. 그렇기에 해당 연구는 블록체인 내에 거래 데이터를 분석한 CryptoQuant

(www.cryptoquant.com)의 온체인 데이터를 활용하여 LSTM(Long-Short Term Memory)을 통해 비트코인 가격 예측을 진행하고 그 성능을 평가해보고자 한다. 이 연구를 기반으로 변동성이 큰 시장에서 온체인 데이터를 통한 가격 예측이 실효성이 있는지 검증하고자 한다.

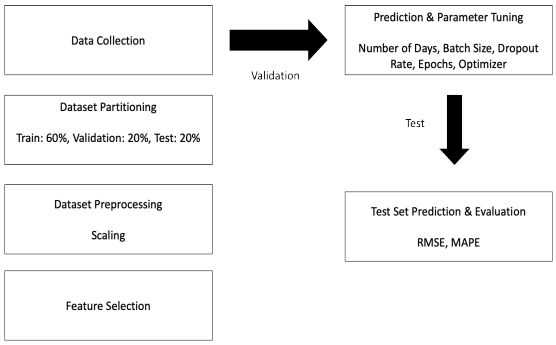

Fig. 1 Experiment Process

해당 연구는 우선 CryptoQuant API Docs[10]를 바탕으로 온체인 데이터와 데이터에 관해 설명한다. 그리고 전체적인 실험 과정을 보여주는 그림. 1에서처럼 데이터를 모으는 과정과 피처(Feature)들을 어떻게 스케일링(Scaling)하였는지 그리고 피처 선택 과정에 관해 설명한다. 후에 검증 세트를 바탕으로 하이퍼파라미터 튜닝(hyper-parameter tuning)을 거치며, 마지막으로 모델에 데이터를 적용한 후 테스트 세트 예측에 대한 성능 평가를 통해 해당 모델이 얼마나 비트코인 가격을 정확하게 예측했는지 평가하고자 한다.

Ⅲ. 연구 데이터 및 이론적 배경

3.1. 데이터 설명

1) Exchange Flows

거래소의 움직임은 비트코인 가격 방향 예측에 있어 큰 역할을 한다. 기본적으로 비트코인이 OTC(Over the Counter) 시장에서 거래가 되지 않는 이상 비트코인 거래는 거래소를 통하여 거래가 이루어진다. 그중에서 주목해볼 데이터는 Total Reserve, Reserve Netflow, Address Count, Transactions Count 데이터이다. Total Reserve와 Reserve Netflow가 높을 수록 거래소에서 매도를 기다리는 비트코인이 많음을 뜻하며 반면에 Total Reserve가 하락하거나 Netflow가 음의 값이라면 비트코인을 팔려는 경우보다 지갑으로 인출되어 매도를 위해 대기하는 비트코인의 개수가 적어졌음을 의미한다. Address나 Transactions 데이터는 각각 비트코인 거래소로부터 비트코인이 유입/유출되는 거래를 한 지갑의 수와 거래량이다.

2) Flow Indicator

Flow Indicator는 비트코인 소유자들의 움직임에 대해 나타내주는 지표이다. 이 안에서 첫 번째로 주목해볼 데이터는 Exchange Whale Ratio이다. Whale은 주로 1,000개 이상의 비트코인을 소유하고 있는 자를 일컫는다. 온체인 데이터를 참고하는 많은 트레이더가 whale들의 움직임에 많은 관심을 기울인다. 이는 whale들이 비트코인 가격을 주도하고 있다고 여기기 때문이다. 해당 연구에서 사용하게 될 데이터인 Exchange Whale Ratio는 거래소 내의 총 비트코인 입금량 중 상위 10개 입금량의 비중으로 한다. 또 하나의 지표는 Stablecoins Ratio 이다. 이는 거래소 내의 비트코인 잔고를 스테이블 코인(Stable Coin) 잔고로 나눈 값이다. 많은 투자자가 스테이블 코인을 매개로 비트코인 거래의 매개로 사용하기에 이 값이 높을 수록 매도 압박이 더 크다고 볼 수 있다. 마지막으로 살펴볼 데이터는 Fund Flow Ratio 이다. 이는 비트코인 네트워크 중 거래소의 거래량 비중을 나타낸다.

3) Market Indicator

Market Indicator는 거래 시장과 온체인 데이터를 활용해 산출된 지표이다. 우선 Estimated Leverage Ratio는 선물 시장 미결제 거래 잔고(open-interest)를 거래소의 잔고로 나눈 값이다. 특히 Leverage를 이용하는 선물 시장은 비트코인을 향한 센티멘트를 측정하기 위한 좋은 시장이다. 사람들이 비트코인 가격의 미래에 대해 긍정적으로 평가한다면 공격적으로 레버리지를 더 많이 두고 선물 계약을 할 것이기 때문이다. 두 번째로 살펴볼 데이터는 Stablecoin Supply Ratio이다. 이는 전체 암호화폐 공급량 중 스테이블 코인 공급량의 비중을 나타낸다.

4) Miner Flows

Miner Flow는 주요 채굴풀의 비트코인이 어떤 흐름을 보이는지 나타내는 지표이다. 이 중 Reserve 데이터와 Netflow 데이터를 이용할 것이다. 이를 통해 얼마나

많은 채굴자가 비트코인을 쌓아놓고 있는지 그리고 얼마나 수익 실현을 위해 출금되고 있는지 알 수 있다.

5) Market Data

Market Data는 시장 내의 특정 기간 동안 비트코인의 고점, 저점, 시작가, 마감가 등을 포함하는 데이터이다. 이 연구에서 타깃(target)으로 사용하는 Price USD 데이터는 하루 단위 마감가 미국 달러 기준으로 나타낸 것이다. 여기에서 원화를 사용하지 않고 달러를 기준으로 활용하는 이유는 비트코인의 경우 아비트라지(arbitrage) 기회가 각종 규제로 인해 상대적으로 적기 때문에 나라마다 비트코인 가격의 차이가 심하기 때문이다. 특히 한국 시장의 경우 시장이 과열될 때마다 '김치 프리미엄(Kimchi Premium)'이 20~30%까지 치솟을 때가 있다.

6) Network Data

네트워크 데이터 안에는 transaction fee, mining difficulty, hashrate 등 [9]에서 이미 활용되었던 온체인 데이터가 포함된다. Transaction fee는 비트코인 거래를 통해 발생하는 거래 수수료 개념이며, mining difficulty는 채굴 난이도로, 얼마나 채굴이 어려운지를 나타내는 지표이다. 채굴 난이도는 채굴자들이 많아질수록 올라가는 특성이 있다. Hashrate는 채굴자들이 얼마나 빨리 비트코인 채굴을 위한 hash problem을 풀어내는지를 보여주는 지표이다.

3.2.1. RNN(Recurrent Neural Network)

그림. 2에서 볼 수 있듯이 RNN은 기존의 뉴럴 네트워크와 달리 네트워크 내에 루프를 형성하여 입력된 정보가 지속될 수 있도록 한다. 예를 들어, A라는 뉴럴 네트워크가 있고 X_t라는 인풋, 그리고 H_t라는 아웃풋이 존재한다면 RNN은 지속해서 X_t라는 인풋을 넣은 A를 후계자(successor)에 다시금 전달한다. 따라서, 체인(Chain)이나 리스트(List) 형태로 된 데이터처리 적합한 특성을 보인다. 특히, 음성인식이나 언어 모델링 등에 많이 사용된다.

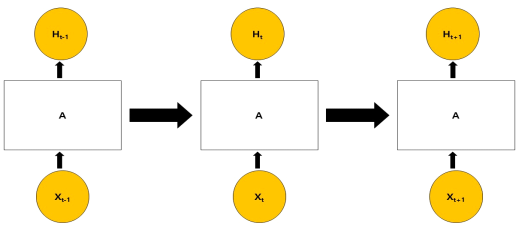

Fig. 2 RNN Structure

3.2.2 LSTM(Long-Short Term Memory)

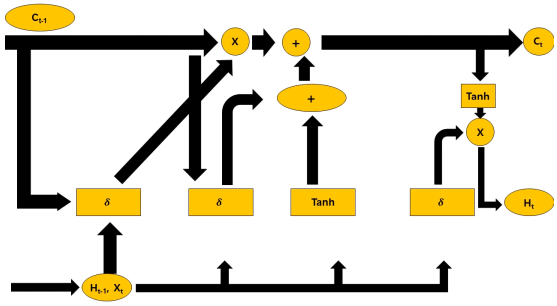

Fig. 3 LSTM Structure

LSTM은 RNN의 하위분류로 보면 된다. 기존의 RNN은 직전의 정보를 다음의 처리를 위해 사용한다는 장점이 있지만, 동시에 거리가 멀어질수록 이에 대한 고려가 사라진다. 이러한 문제를 해결하기 위해 등장한 것이 LSTM이다.

RNN에는 A의 내부가 Tanh만으로 이루어지는 매우 단순한 구조로 이루어져 있었다면 LSTM 안에는 다양한 선형 상호작용이 발생한다. **그림. 3**은 LSTM의 A를 확대해서 보여준다. 맨 위의 연속적으로 이루어지는 화살표에서는 이전의 Cell State인 C_{t-1} 를 입력받고 크게 변하지 않는다. 하지만, **그림. 3** 밑 부분의 σ(sigmoid) 층을 입력값이 통과하면 0부터 1 사이의 결과가 나오며 이는 얼마나 많은 양의 정보가 해당 층을 통과하게 될지 결정한다. 이를 망각 게이트라고 한다. 첫 번째 σ층에서는 이전 출력값 H_{t-1}과 현재의 입력값 X_t를 입력받아 f_t를 출력하고 C_{t-1}와 곱셈 연산을 진행한다. 두 번째 σ층에서도 같은 입력값을 받으며, 어떤 값을 업데이트할지 다시 한번 결정하여 i_t를 출력한다. Tanh 층 또한 같은 인풋값을 받아 Tanh 연산 후 입력 후보를 $\widetilde{C}_t$를 저장한다. 그 후 i_t와 $\widetilde{C}_t$값을 곱셈 연산 후 앞서서 곱셈 연산을 하여 나온 값과 f_t를 덧셈하면 C_t가생성된다. 마지막으로 X_t와 H_{t-1}을 마지막 σ층에 입력하여 그 값과 Tanh 층을 거친 Ct 값을 곱셈 연산하여 H_t를 출력한다.

LSTM의 필요한 데이터는 지속해서 사용한다는 특성은 긴 기간의 데이터를 바탕으로 예측에 활용하는 데 적합하다. 특히 해당 연구에서 사용하는 모든 데이터는

시계열 데이터로 어느 정도 시간적 간격이 있더라도 의존성이 두드러질 수밖에 없다. 그렇기에 여타의 알고리즘보다 시계열의 데이터라는 특성을 살려 예측을 진행할 수 있는 LSTM이 다른 딥러닝 알고리즘보다 해당 주제에 더욱 적합하다고 볼 수 있다. 또한, 모든 데이터를 지속해서 쓰는 것이 아니라 앞서 언급하였던 망각을 통해 중요하지 않다고 보는 부분은 자체적으로 사용하지 않기 때문에 시시각각 변화하는 비트코인 시장에 적합하다. 특히, 전체적인 가격 흐름의 추세가 전환되었을 때는 이전의 모든 기록을 기억하기보다 최근의 흐름에 더욱 큰 가중치를 부여하는 것이 중요하기에 LSTM은 해당 연구에 더없이 적합하다고 볼 수 있다.

Ⅳ. 실험 과정 및 분석

4.1. 데이터 전처리

필요한 온체인 데이터와 가격 데이터는 Crypto Quant API access를 통해 추출하였다. CryptoQuant는 거래소 비트코인 잔고 데이터뿐만 아니라 비트코인 네트워크, 거래소, 채굴자 등 다양한 데이터를 제공한다. 하지만 유의해야 할 점은 피처별로 제공되는 시간 단위가 다르다는 것이다. 그렇기에 해당 연구는 대부분의 데이터에서 제공하는 하루 단위 데이터를 기반으로 진행하였다. 취합한 데이터는 우선 훈련, 검증, 테스트를 위해 훈련 세트 (2019년 4월 19일-2020년 7월 5일; 총 444일), 검증 세트(2020년 7월 6일 – 2020년 11월 30일; 총 144일), 훈련-검증 세트(2019년 4월 19-2020년 11월 30일; 총 588일), 테스트 세트(2020년12월 1일-2021년 4월 27일; 총 144일)로 나눈다.

훈련 세트, 검증 세트, 테스트 세트를 나눈 후에 훈련 세트를 바탕으로 피처 선택(Feature Selection)을 진행한다. **표. 1**에 피처 선택 과정의 결과가 나와 있는데, 우선 트레이닝 데이터를 기준으로 파이썬의 StandardScaler를 이용하여 스케일링을 진행한 후, sklearn.model_selection의 SelectKBest의 'score_func'를 mutual information regression으로 설정한 후, 점수가 가장 높은 10개의 피처를 선정한 것이다.

연구에서 가장 중요한 부분은 모델에 적용할 데이터의 스케일링 방법이었다. 만약 전체 기간의 데이터를 단번에 스케일링을 진행한다면 약 2년이라는 긴 기간에 해당하는 데이터에 대해 한꺼번에 스케일링하므로 단기적인 트렌드를 모델이 제대로 잡아낼 수 없게 된다. 그렇기에 해당 연구는 [11]을 참고하여 우선, 모델을 훈련하기 위한 데이터에 대해 스케일링을 진행한 후, 예측에 활용할 피처들에 대해서는 다른 방법으로 스케일링을 진행하였다. 훈련 세트는 한 번에 스케일링하였지만, 테스트 세트는 LSTM 모델이 t 시점에 대한 예측을 하기 위해 t-9~t-1일까지의 데이터를 활용한다면 피처별로 t-9~t-1에 해당하는 데이터에 대해서 스케일링을 진행하며 이 과정을 반복하는 형식의 스케일링을 진행한다. 이로써 모델이 트레이닝 후 예측을 진행할 때 단기적인 트렌드를 더 정확하게 예측하게 한다. 구체적인 방법은 위의 **그림. 4**에 나타나 있다.

Table. 1 Selected Features

Feature	Definition	Feature Type
Exchange Reserve	Total number of Bitcoin in exchanges	Exchange Flows
Exchange Transactions Count Outflow	Total number of transactions flowing out of Bitcoin exchanges	Exchange Flows
Addresses Count Inflow	Total number of addresses involved in Inflow Transactions	Exchange Flows
Fund Flow Ratio	The total BTC amount flowing into or out of exchanges divided by the total BTC transferred on the whole Bitcoin network	Flow Indicator
Estimated Leverage Ratio	Open interest of exchange divided by exchange's Bitcoin reserve	Market Indicator
Stablecoin Supply Ratio	Ratio of stablecoin supply in the whole cryptocurrency market	Market Indicator
Miner's Reserve	Total number of Bitcoin miners hold	Miner Flows
Miner's Reserve in USD	USD total of Bitcoin miners hold	Miner Flows
Open Interest	Number of open positions currently on derivative exchanges	Market Data
Hashrate	The mean speed at which miners in the network solve hash problems	Network Data

Algorithm for Scaling Training and Test Data

Requirements: Training Dataset, Test Dataset, Look Back Days

```
split features(X) and target(y) for both training and test datasets
make two new lists for features and target

# Making Training Dataset Suitable for LSTM
for i in range(from=lookback days, to=dataset length):
  append features[i-lookback:i] to features list
  append target[i] to target dataset

# test set scaling
make new empty list for features and two new lists for target

for  i in range(from=len(training set), to=len(whole dataset)):
  standard scaling for features[i-lookback:i]
  append to new features list
  standard scaling for target[i-lookback:i]
  append to new target list(for inverse-scaling)
  append target[i] to new target list2
```

Fig. 4 Scaling Methodology

4.2. 검증 및 하이퍼파라미터 튜닝

스케일링을 완료한 후, 훈련 세트를 이용하여 모델을 훈련한 후에 검증 세트에 대한 예측을 진행한다. 예측은 파이썬 딥러닝 라이브러리 중 하나인 케라스(Keras)를 이용하여 진행하였다. 또한 실험 과정에서 하이퍼파라미터 튜닝을 진행하여 테스트 세트에 대한 예측을 진행하기 위한 최적의 하이퍼파라미터를 설정하였다.

모델의 성능은 회귀 모델 평가에서 주로 사용하는 RMSE, MAPE, R-squared를 사용하기로 한다. RMSE 오차의 제곱 평균의 제곱이고, MAPE는 실제값과 오차 비율의 평균을 퍼센트로 나타낸 것이며, R^2는 독립변수가 종속변수에 대해 얼마만큼의 설명력을 가지는지 나타내는 지표이다.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_{true} - y_{pred})^2}{n}} \quad (1)$$

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \| \frac{y_{true} - y_{pred}}{y_{true}} \| \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_{true} - y_{pred})^2}{\sum_{i=1}^{n} (y_{true} - y_{mean})^2} \quad (3)$$

이들 지표를 기반으로 하이퍼파라미터 튜닝을 진행하였으며, 튜닝 결과 Lookback Day는 3, Epoch은 1000, Batch Size는 64, unit의 개수는 7, dropout rate는 0.5, optimizer는 adam이 최적의 결과를 나타낸다는 것을 확인할 수 있었다.

4.3. 테스트 결과 및 평가

Table. 2 Hyper-parameter Tuning Result

N	Epochs	Batch Size	Units	Dropout	Optimizer
3	1000	64	7	0.5	adam

표. 2의 하이퍼파라미터 튜닝 결과를 바탕으로 테스트 셋에 대한 예측을 진행한다. **그림. 5**는 모델을 훈련하고 예측을 진행하는 일련의 과정을 보여준다. 훈련-검증 세트를 기반으로 LSTM 모델을 훈련한 후에 나오는 대로 테스트 세트를 스케일링한다. 스케일링을 완료한 테스트 세트를 기반으로 한 예측값을 역스케일링한 후 통해 실제값과 비교 및 평가를 진행한다.

Algorithm for Training Model and Prediction

Requirements: Scaled Training Dataset, Test Dataset

```
make model
add model(LSTM(units, input shape))
add model(dropout)
add model(Dense(1))

compile model(loss function, optimizer)
fit model(scaled training data, epochs, batch size)

prediction = predict(scaled test data)
```

Fig. 5 Model Training and Prediction Process

표. 3를 검증과 테스트 시의 성능을 더욱 직관적으로 확인할 수 있다. 검증 단계에서는 RMSE가 436.894, MAPE가 2.074%, R^2가 0.971로 나왔으며 테스트 단계에서는 RMSE가 2162.380, MAPE가 3.993%, R^2가 0.976으로 나타났다. 검증 세트의 예측에 비해 테스트

Table. 3 Validation/Test Evaluation

Evaluation Metrics	Validation	Test
RMSE	436.894	2162.380
MAPE	2.074%	3.993%
R^2	0.971	0.976

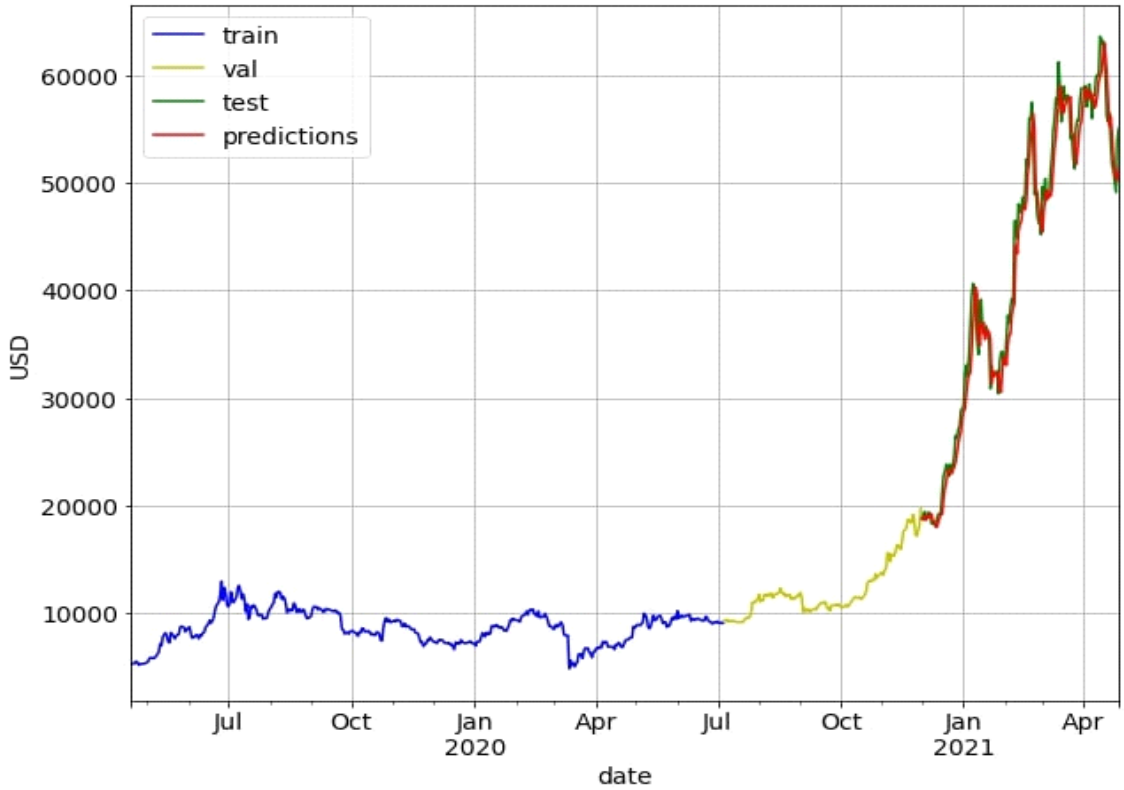

Fig. 6 Prediction Result(Full Data Length)

값의 예측 성능과 차이가 날 수 있다. 이에 대해서는 과 검증 데이터에 해당하는 기간의 가격 변동성과 테스트 데이터에 해당하는 기간의 가격 변동성이 크게 차이가 나기 때문으로 보인다. **표. 3**와 **그림. 6**를 통해 이를 명확

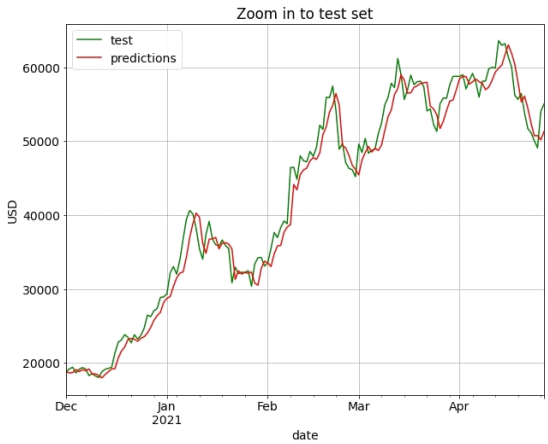

Fig. 7 Comparison Between Prediction Results and Actual Data

하게 확인할 수 있는데, 검증 세트의 기간에는 상대적으로 비트코인이 평탄하게 상승하고 있으나, 테스트 세트에 해당하는 구간에는 비트코인이 가파르게 상승하고 있음을 확인할 수 있다. 이는 테스트 구간과 훈련 기간에 따라서 예측 성능에 차이가 존재할 수 있음을 시사한다고 볼 수 있지만 그런데도 전체적으로 비트코인의 예측 가격과 실제 가격의 움직임 자체는 큰 차이가 없어 보인다는 것 또한 확인할 수 있다. **그림. 7**을 통하여 예측값과 실제값의 흐름을 더욱 구체적으로 확인해 볼 수 있다.

Ⅴ. 결 론

본 연구는 CryptoQuant를 통해 확보한 블록체인 온체인 데이터를 활용하여 비트코인 가격을 예측하였다. 예측 결과와 평가 지표를 보았을 때 기존 연구들에서 활용되었던 가격이나 각종 경제 지표뿐만 아니라 온체인 데이터 또한 비트코인 가격의 향방을 예측하는 데 있어

충분히 활용할 수 있음을 시사한다. 특히 hashrate 같은 네트워크 데이터 외에도 거래소 내의 비트코인의 이동이나 비트코인 채굴자의 비트코인 보유량 등 이전에 잘 활용되지 않았던 데이터들도 비트코인 가격 예측에 유용하다는 것을 보여준다. 아직은 비트코인 가격의 예측을 하는 데 있어서 다른 지표들만큼 대중화되지 않은 상황이지만 온체인 데이터를 LSTM과 같은 알고리즘을 활용했을 때 변동성이 큰 암호화폐 시장에서 좋은 가이드라인을 제공할 수 있을 것이라는 가능성을 확인할 수 있었다.

더 나아가, CryptoQuant를 통해 확보를 할 수 있는 지표 외에도 예측에 활용될 만한 온체인 데이터들이 존재한다. 특히 가격 측면의 진입장벽으로 인해 확보할 수 없었던 여러 가지 데이터가 존재한다. 특히 비트코인 다량 보유자들의 움직임에 대한 데이터는 이들 비트코인 가격에 큰 영향을 끼친다는 인식이 있다는 것을 고려했을 때 향후 연구가 진행될 필요성이 존재한다. 더 나아가, 전통적으로 많은 연구에서 활용되었던 주식 시장 지수, 인플레이션, 구글 트렌드 등의 변수와 함께 가격 예측을 했을 때의 예측 성능에 대한 연구가 진행된다면 향후 비트코인의 움직임을 파악하는 데 있어서 적극적으로 활용될 수 있을 것이다.

또한, 고빈도 온체인 데이터를 활용한 예측에 대한 연구도 진행될 필요가 있다. 특히 데이 트레이딩(day-trading)이 비트코인 시장에서 중요한 자리를 차지함과 동시에 짧은 시간 내에 가격이 급격히 변동할 수 있는 비트코인의 특성상 짧은 기간의 온체인 데이터를 활용한 예측의 가치가 매우 크리라 생각할 수 있다.

마지막으로, 타 암호화폐의 온체인 데이터를 이용한 예측 연구이다. 타 암호화폐는 비트코인과 다른 데이터가 존재한다. 특히 hashrate를 찾는 것이 아닌 지분율에 따라 의사결정권이 주어지는 지분 증명 방식(Proof of Stake)을 채택한 차세대 암호화폐들의 경우 hashrate나 difficulty와 같은 변수들은 존재하지 않는다. 이들의 경우 비트코인 가격 예측에 활용되었던 변수들을 대체할 만한 변수들을 찾아야 할 것이다. 대부분의 비트코인보다 변동성이 매우 크고 이에 따라 위험성도 더 크기에 연구의 중요성이 매우 크다고 본다.

ACKNOWLEDGEMENT

I would like to thank the Department of Fintech, Sungkyunkwwan University (SKKU) for providing me with the environment necessary to accomplish this research.

Special thanks should be given to CryptoQuant for giving me the opportunity to use its state-of-the-art data.

This work was supported by the BK21 FOUR Project.

Following are results of a study on the "Convergence and Open Sharing System" Project, supported by the Ministry of Education and National Research Foundation of Korea

REFERENCES

[1] T. Bradshaw. (2021, May). Apple ad for alternative pay ments job signals cryptocurrency interest. *Financial Times*. [Internet]. Available: https://www.ft.com/content/2b30e69d-6662-40e8-9735-eaea8662906e.

[2] A. Aggarwal, I. Gupta, N. Garg, and A. Goel, “Deep learning approach to determine the impact of socio economic factors on bitcoin price prediction,” 2019 *Twelfth International Conference on Contemporary Computing (IC3)*, pp. 1-5, 2019. doi: 10.1109/IC3.2019.8844928.

[3] S. Pyo and J. Lee, “Do FOMC and macroeconomic announcements affect bitcoin prices?,” *Finance Research Letters*, vol. 37, 2020.

[4] J. Lee, K. Kim, and D. Park, “Empirical analysis on bitcoin price change by consumer, industry and macro-economy variables,” *Journal of Intelligence and Information Systems*, vol. 24, no. 2, pp. 195-220, 2018.

[5] K. Lee, S. Cho, G. Min, and C. Yang, “The determinant of bitcoin prices in Korea,” *Korean Journal of Financial Studies*, vol. 47, no. 4, pp. 393-415, 2019.

[6] E. Kim and J. Hong, “The prediction model of cryptocurrency price using news sentiment analysis and deep learning,” *Korea Society of IT Services, 2020 Fall Conference*, pp. 122-126, 2020.

[7] B. Jafar. (2021, April). Bitcoin is not a bubble, BTC is like digital gold, says Bill Miller. *Finance Magnates | Financial*

and business news [Internet]. Available: https://www.financemagnates.com/cryptocurrency/news/bitcoin-is-not-a-bubble-btc-is-like-digital-gold-says-bill-miller/.

[8] S. McNally, J. Roche, and S. Caton, "Predicting the price of bitcoin using machine learning," *2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*, pp. 339-343, 2018.

[9] Z. Chen, C. Li, and W. Sun, "Bitcoin price prediction using machine learning: an approach to sample dimension engineering," *Journal of Computational and Applied Mathematics*, vol. 365, 2020.

[10] CryptoQuant. CryptoQuant Data API (1.3.0) [Internet]. Available: https://cryptoquant.com/docs.

[11] Y. Ng. (2019, January). Machine learning techniques applied to stock price prediction [Internet]. Available: https://towardsdatascience.com/machine-learning-techniques-applied-to-stock-price-prediction-6c1994da8001.

안유진(Yu-Jin An)

성균관대학교 대학원 핀테크융합전공 석사 재학 (2021~)
성균관대학교 글로벌경제학과, 데이터사이언스융합학과 학사 졸업 (2021)
※관심분야 : 금융 데이터, 자산가격결정, 알고리즘 트레이딩, 자연어 처리

오하영(Ha-Young Oh)

Sungkyunkwan University Professor (2020~)
Ajou University Professor (2016~2020)
Soongsil University Professor (2013~2016)
Ph.D. in computer engineering at Seoul National University (2013)