

대표적인 데이터 예측 기법

01. 전통적인 통계 모델

ARIMA (AutoRegressive Integrated Moving Average) : 시계열 데이터에 기반하여 과거 데이터를 이용해 미래를 예측하는 통계 모델

SARIMA (Seasonal ARIMA) : 계절성이 포함된 시계열 데이터를 다룰 때 유용한 통계 모델

VAR (Vector AutoRegression) : 여러 시계열 변수가 상호 영향을 미칠 경우 사용

GARCH (Generalized Autogressive Conditional Heteroskedasticity) : 변동성이 중요한 데이터에서 사용되는 기법

02. 머신러닝 기반 기법

선형 회귀 (Linear Regression) : 독립 변수와 종속 변수 간의 관계를 모델링하는 기법

랜덤 포레스트 (Random Forest) : 다수의 의사결정 트리를 결합하여 예측하는 기법

그라디언트 부스팅 (Gradient Boosting) : 예측 정확도를 높이기 위해 연속적으로 약한 모델을 결합하는, (예를 들면 XGBoost, LightGBM, CatBoos 등) 기법

SVN (Support Vector Maching) : 데이터를 분류하거나 회귀를 통해 예측하는 기법

03. 딥러닝 기반 기법

LSTM (Long Short-Term Memory) : 순환 신경망 (RNN)의 변형으로, 시계열 데이터에서 장기 의존성을 잘 학습하는 알고리즘

GRU (Gated Recurrent Unit) : LSTM의 단순화된 버전으로 계산 효율이 높은 알고리즘

Transformers 기반 모델 : 시계열 데이터의 패턴을 모델링 하는 용도

변수의 조작적 정의

High :고가, 해당 날짜 동안 기록된 가장 높은 가격, 투자자들의 심리와 특정 시점의 수요를 반영
Low :저가, 해당 날짜 동안 기록된 가장 낮은 가격, 매도 압력과 하락 추세를 보여주는 지표로 사용.
Open :개장가, 거래가 시작될 때의 가격, 투자자들의 기대치를 반영.

고가, 저가, 개장가는 종가에 대해 완벽한 상관관계를 가지나, 다중 공선성이 존재해 독립변수가 아니며, 종가 예측에 과적합을 발생시킬 가능성이 있다.
고가, 저가, 개장가는 선택 시 성능향상에 도움이 되나 항상 폭이 높지 않으며 고가, 저가 중 하나만 선택되면 다른 변수는 필요하지 않다.

* 암호화폐 종가 예측 성능과 입력 변수 간의 연관성 분석, 박재현, 2022, 8 page

Price :종가, 거래가 종료될 때의 가격, 해당 일의 최종적인 평가 가치를 나타냄.
Ma_n :n 일간의 평균 가격을 나타내는 이동 평균 (Moving Average), 단일 시점의 가격 데이터보다 전반적인 추세를 알 수 있기 때문에 사용한다.
Volatility_nd :n일간의 가격 변화 정도를 나타내는 변동성 (Volatility), n 일간 자산 가격이 얼마나 흔들렸는지 (분산이 큰지) 를 나타낸다.
Volume :거래량. 해당 날짜동안 거래된 총량, 시장의 유동성과 투자자들의 관심을 나타냄.

거래량은 다중공선성이 존재하지 않아 독립변수이며, 예측 성능이 긍정적인 영향을 끼칠 수 있다. * 암호화폐 종가 예측 성능과 입력 변수 간의 연관성 분석, 박재현, 2022, 5page

Active_address :활성화된 지갑 주소의 수, 네트워크 활동 정도를 파악할 수 있음
Avg_txn_size :평균 거래 크기, 한 거래당 평균적으로 얼마나 많은 코인이 교환되었는지를 나타내는 값
Avg_txn_fee :평균 거래 수수료, 거래 비용의 경향을 보여 줌.

Dollar_Index :미국 달러지수, 미국 달러 USD의 강세 또는 약세를 나타내는 지표.
달러지수 변화량인 Dollar Index Futures 는 비트코인 가격 변화와 음의 상관관계를 가진다. 즉 달러지수가 상승하면 비트코인 가격은 하락하는 패턴을 보인다.

* 비트코인 가격 변화에 관한 실증분석: 소비자, 산업, 그리고 거시변수를 중심으로, 2018, 이준식, 김건우, 박도형, 17 page

S&P500 : Standard and Poor’s 500, 미국 주식 시장에서 가장 영향력 있는, 주식시장 시가총액 상위 500개 기업 주가를 기반으로 산출된 주식 지수
S&P500 은 미국 경제 상태와 투자 성과를 평가하는 지표로 사용된다.
암호화폐와 S&P 는 서로 다른 시장 특성을 가지나, 글로벌 경제 이벤트(금융 위기, 금리 변화)에는 공통적인 영향을 받는다.

(하락장 이전, 매수 빈도가 약세, An Information Theory Approach to the Stock and Cryptocurrency Market: A Statistical Equilibrium Perspective, 2023, 22page)

두 시장간의 상관관계는 점차 높아지고 있다. 즉, 암호화폐가 점점 전통자산과 같은 자리를 잡고 있다.

NASDAQ :대표적인 미국 주식 시장 중 하나로, 기술주 중심의 주가로 알려져있음.
Gold :금, 비트 코인은 달러지수와 반대 방향으로 움직이는 경향, 즉 금과 유사한 안전 자산의 특성을 가진다.

* 비트코인 가격 변화에 관한 실증분석: 소비자, 산업, 그리고 거시변수를 중심으로, 2018, 이준식, 김건우, 박도형, 1 page

금의 가격은 비트코인과 부(-)의 관계를 가진다. * 비트코인과 주요 투자 자산 간의 장기 균형관계 추정, 박준기, 2020, 1 page
* 위 다른 두 논문을 종합해 보았을 때, 수집된 데이터 차이, 시간이 지날 수록 코인이 더 자산으로서 자리를 잡은 점, 달러와 비교했을 때는 금-코인 모두 대체 자산이 되는 공통 점이 있지만, 금과 비트코인은 다시 서로 대체재가 된다는 특징이 보인다.

코인 가격 예측에 관한 선행 연구 논문 – 전처리 및 데이터 분석

01. 암호화폐 증가 예측 성능과 **입력 변수 간의 연관성 분석** (2022, 정보처리학회논문지, 박재현)

데이터 수집 방법

데이터 수집 도구: .

암호화폐 선택: Bitcoin, Ethereum, Binance, Cardano, Chainlink, Litecoin

변수: 시가, 증가, 고가, 저가, 볼륨(거래량), change(증가변동률)

시계열 기준 단위: 논문에 표기되어있지 않음

시계열 데이터 범위: 논문에 표기되어있지 않음

분석 방법

통계기반 분석 – 상관분석: 피어슨 상관계수를 사용

통계기반 분석 – 다중 공선성 분석: VIF 를 사용, 100이상일 경우 다중공선성이 있다고 판단

분석 결과

고가, 저가, 개장가는 완벽한 상관관계, 독립 변수가 아님.

고가, 저가 중 하나만 선택되는 경우가 많고, 선택한다 해도 성능 향상 폭이 높지 않으며, 한쪽을 선택하면 다른 쪽은 불필요하다.

거래량은 다중공선성이 존재하지 않음, **독립변수**.

증가 변동률은 다중공선성이 존재하지 않으나, 상관관계도 존재하지 않음.

02. 암호화폐 가격 정보 데이터에 대한 상관관계분석 및 회귀테스트 (2018, 정보처리학회논문지, 권도형, 허주성, 김주봉, 임현교, 한연희)

데이터 수집 방법

데이터 수집 도구: 빗썸 API

암호화폐 선택: 12개의 암호화폐 (BTC, ETH, XRP, BCH, LTC, EOS, DASH, XMR, ETC, QTUM, BTG, ZEC)

변수: 시가, 종가, 고가, 저가, 볼륨(거래량)

시계열 기준 단위: 10분 간격

시계열 데이터 범위: 2017년 6월 – 2018년 2월 약 8개월, 하지만 코인마다 다름

데이터 전처리 방법

이상치: 직전 행의 데이터를 복사하고, 새로 만든 label 컬럼에 이를 명시, 이 외에 대한 언급은 없음

정규화: (0,1)

결론

(1) 암호화폐들은 서로 다른 가치관을 가지나, 가격은 독립적으로 형성되지 않는다.

연구 당시, 비트코인이 지배적인 암호화폐로서, 다른 암호화폐의 거래에도 쓰여 가격에 영향을 미친다.

(2) 암호화폐 가격들은 데이터가 무작위성을 띤다. 즉 평균으로 회귀한다는 보장이 없다.

때문에 회귀 모델을 적용하기 적합한지의 여부 결정을 위해

Augmented Dickey-Fuller 단위근 검정, Hurst Coefficient 분석, Regression half life 를 계산해보았고

세 분석 결과 모두 회귀 모델 사용이 비적합함을 보였다.

코인 가격 예측에 관한 선행 연구 논문 – 전처리 및 데이터 분석

03. On-Chain Data를 활용한 LSTM 기반 비트코인 가격 예측(2021, 한국정보통신학회논문지, 안유진, 오하영)

데이터 수집 방법

데이터 수집 도구: CryptoQuant API

암호화폐 선택: -

변수: * 해당 논문 5page

Exchange Flows – Exchange Reserve (거래소 비트코인 총량),

Exchange Flows – Exchange Transactions Count Outflow (거래소에서 유출된 거래 횟수),

Exchange Flows – Addresses Count Inflow (거래소로 유입된 거래에 관련한 주소 수),

Flow Indicator – Fund Flow Ratio (비트코인 네트워크의 총 거래량 대비 거래소 거래량 비중)

Market Indicator – Estimated Leverage Ratio (선물 시장의 미결제 거래 잔고 대비 거래소 비트코인 잔고 비율)

Market Indicator – Stablecoin Supply Ratio (암호화폐 시장에서 스테이블 코인 공급량 비율)

Miner Flows – Miner's Reserve (채굴자가 보유한 비트코인 총량)

Miner Flows – Miner's Reserve in USD (채굴자가 보유한 비트코인의 USD 가치)

Market Data – Open Interest (파생상품 거래소에서의 미결제 포지션 수)

Network Data – Hashrate (채굴 네트워크에서의 해시 문제를 푸는 평균 속도)

* 변수 선택의 기준은 정규화 후, SelectKBest의 score_func 를 mutuak information regression 로 설정 후 점수가 높은 변수들을 선택.

시계열 기준 단위 : 1일 간격

시계열 데이터 범위: 2019년 4월 – 2020년 7월 (총 444일 훈련 셋), 2020년 7월 – 2020년 11월 (총 114일, 검증 셋)

데이터 전처리 방법

정규화: Python StandardScaler 로 훈련 데이터 기준 스케일링, 해당 스케일링 기준을 테스트 데이터에 적용.

(훈련 데이터를 통해 구한 mean 과 std를 사용하여 테스트 데이터를 스케일링 했다는 뜻)

실험 방법과 그 결과

Lookback Days = 3일 로 설정 (3일간의 데이터를 기반으로 미래의 1일치를 예측한다.)

평균 오차 3.99, 설명력 R^2 0.976으로,

온체인 데이터를 기반으로 LSTM 모델을 사용하여 비트코인 가격 예측 가능성이 있음 (즉 모델 자체의 유용성을 보여주는 목적의 논문)

코인 가격 예측에 관한 선행 연구 논문 – 머신러닝 기반

04. **그래디언트 부스팅**을 활용한 암호화폐 가격 동향 예측 (2018, 정보처리학회논문지, 허주성, 권도형, 김주봉, 한연희, 안채현)

데이터 수집 방법

데이터 수집 도구: 빗썸 API

암호화폐 선택: 데이터가 많은 상위 7개의 암호화폐 (BTC, BCH, ETC, DASH, ETH, XRP, LTC)
개별 코인 각각에 대해 학습한 것이 아닌, 7종 데이터를 통합해 사용 한 것.
7종 코인 데이터로 전체 패턴을 학습 후, 특정 코인의 타겟을 예측하는 모델을 설계.

변수: 시가, 종가, 고가, 저가, 볼륨(거래량)

시계열 기준 단위: 10분 간격

시계열 데이터 범위: 2017년 8월 4일 – 2018년 5월 18일 약 10개월

데이터 전처리 방법

손실 데이터: (10분 간격으로 수집되지 않은 데이터를 뜻함) 이전 데이터를 복사

결측 데이터: (10분 간격으로 수집되었으나, 값이 들어오지 않은 Null 을 뜻함) 이전 데이터를 복사

부적합 데이터: (10분 간격으로 수집되지 않은, 9분 간격 등으로 수집된 데이터를 뜻함)

해당 시간에 10분가격으로 수집된 데이터가 있을 경우 부적합 데이터를 삭제,
없을 경우 시간을 조정해 (19분->20분) 사용

정규화: **암호화폐마다 크기가 다르기 때문에**, 정규화를 수행.

가격을 비율로: 각 코인마다 종가를 기준값으로, 시가, 종가, 고가, 저가, 볼륨을 종가로 나누어서 가격->비율로 변환

Mix-max: 상대적 변화를 **(0,1) 범위로** 스케일링

코인 가격 예측에 관한 선행 연구 논문 – 머신러닝 기반

04. **그래디언트 부스팅**을 활용한 암호화폐 가격 동향 예측 (2018, 정보처리학회논문지, 허주성, 권도형, 김주봉, 한연희, 안채현)

학습 데이터 구성

전체 데이터 개수: 41098 개 (시계열 row 수를 의미함)

학습 데이터 개수: 36988 개 (약 90%)

테스트 데이터 개수: 4110 개 (약 10%)

하이퍼 파라미터:

Θ : 예측 대상의 시간 간격을 의미, $\theta=10$ 이면, 예측 데이터도 10분 간격으로 예측하겠다는 의미.

n : 윈도우 크기를 의미, 모델 학습 시 참조하는 과거 시점 데이터의 개수.* Gradient boosting은 입력 데이터로 행렬 형태를 받음

τ : 가중치 조정 계수, $\tau=1$ 이면 모든 데이터에 균등한 가중치를 주었다는 의미.

ε : 가격 변화의 임계값, 즉 $\varepsilon=0.1$ 이면 현재 시점 대비 10% 상승/하락을 의미, 상승했을 경우를 1로 분류.

실험 방법과 그 결과

시간에 따른 비교:

$n=25, \tau=1, \varepsilon=0.1$ 로 고정 후, 시간 단위를 10분, 30분 60분으로 변경해가며 실험.

결과: 60분 기준으로 모델이 좋은 성능을 보였으나 정확도 차이가 크지는 않았다. (약 0.01~02)

윈도우 사이즈에 따른 비교: $\theta=10, \tau=1, \varepsilon=0.1$ 로 고정 후, (논문에 시간 단위, 윈도우 사이즈, 상승률을 고정한 후 n 을 변경했다고 적혀있으나, 문맥 상 오타로 보임. 논문 내 5.4 절 3번째 줄 참고)

윈도우 사이즈를 10, 25, 50, 75 로 변경하며 실험.

결과: 25 기준으로 모델이 좋은 성능을 보이며 값이 클 수록 성능이 떨어짐.

암호화폐의 경우 한번에 너무 많은 오랜 기간 데이터를 반영 할 수록 좋지 않은 것으로 보임.

상승률에 따른 비교:

ε 를 0.1, 0.25, 0.5 로 변경하며 실험 진행

결과: 0.5 일 때 가장 높은 성능, 0.1 일 때 가장 낮은 성능을 보임

최종 정확도 0.60

* 논문 내 내용이 아닌, ChatGPT 에 의한 지식은 자료 신뢰문제로 따로 첨부합니다. 참고만 부탁드립니다.

이 연구 논문과 관련해서 ϵ 에 대해 좀 더 자세히 설명해주시겠어요?



논문에서 사용된 ϵ 은 **가격 변화율의 임계값(Threshold)**을 나타냅니다. 이 값은 예측하려는 대상 (목표)을 설정하는 데 중요한 역할을 합니다. 구체적으로, $\epsilon = 0.1$ 로 설정되었으므로, **현재 가격 대비 10% 이상의 상승을 "상승(1)"으로 분류한 것을 의미합니다.**

1. ϵ 의 역할

1. 라벨링 기준 설정:

- ϵ 은 미래 시점의 가격 변화가 예측 목표를 충족했는지를 판단하는 기준입니다.
- 현재 시점의 종가(P_t)와 미래 시점의 종가($P_{t+\theta}$)를 비교하여 **변화율**을 계산한 후, 이 값이 ϵ 이상이면 상승으로 간주합니다.

$$\text{변화율 (Rate of Change)} = \frac{P_{t+\theta} - P_t}{P_t}$$

- 변화율 $\geq \epsilon \rightarrow 1$ (상승)
변화율 $< \epsilon \rightarrow 0$ (하락)

2. 예측 문제 단순화:

- 암호화폐는 가격 변동이 매우 크기 때문에, 변화율이 **10% 이상**일 때만 관심 대상으로 간주합니다.
- 이는 단순히 상승/하락 여부를 판단하는 이진 분류 문제로 전환하기 위함입니다.

4. ϵ 과 모델 성능

1. ϵ 이 성능에 미치는 영향:

- ϵ 값이 너무 작거나 크면, 모델의 **Precision(정밀도)**와 **Recall(재현율)** 간의 균형이 무너질 수 있습니다.
- 연구자는 실험적으로 $\epsilon = 0.1$ 이 최적임을 확인한 것으로 보입니다.

2. 모델 학습의 의도:

- 모델은 특정 수준 이상의 상승(변화율 $\geq \epsilon$)을 예측하도록 학습하며, 이는 투자 의사결정 (예: 매수/매도)을 지원하는 데 유용합니다.

5. ϵ 의 실제 사용 이유

• 시장 특성을 반영:

- 암호화폐 시장에서는 10% 이상의 변동이 중요한 이벤트로 간주되며, 이를 기준으로 전략적 결정을 내릴 가능성이 높습니다.

• 실제 투자 활용:

- $\epsilon = 0.1$ 설정은 가격 변동의 실질적 의미를 반영하며, 변동성이 큰 시장에서 노이즈를 제거하고 주요 패턴에 집중하도록 돕습니다.

코인 가격 예측에 관한 선행 연구 논문 – 딥러닝 기반

05. 암호화폐 가격 예측을 위한 딥러닝 앙상블 모델 개발 연구 (2020, 최수빈)

해당 논문이 선행연구 조사를 한 2018년 기준, 암호화폐 가격 예측은 수치 데이터만을 사용한 경우 * 해당 논문 30page

Linear Regression + Logistic Regression + SVM + NN 연구 -> 정확도 0.55

RF + SVM + Logistic Regression -> 정확도 0.50 ~ 0.55 / Gradient boosting -> 정확도 0.60 등 대부분 낮은 정확도를 보였다.

데이터 수집 방법

데이터 수집 도구: coinmarketcap API

암호화폐 선택: 당시 시가총액 2위 이더리움(ETH)

변수: 시가, 종가, 고가, 저가, 볼륨(거래량)

시계열 기준 단위: 24시간 간격 * 연구하고자 하는 목표가 다음날 가격 예측이기 때문, 해당 논문 43page

시계열 데이터 범위: 2015년 8월 – 2020년 1월 약 5년

데이터 전처리 방법

결측 데이터: (Null 을 뜻함) 평균 값으로 대체

부적합 데이터: (2015년도의 경우 이더리움이 출시된 지 얼마 되지 않아 데이터가 0인 것이 많았다) 부적합 데이터를 삭제

정규화: Mix-max: 상대적 변화를 (0,1) 범위로 스케일링

코인 가격 예측에 관한 선행 연구 논문 – 딥러닝 기반

05. 암호화폐 가격 예측을 위한 딥러닝 앙상블 모델 개발 연구 (2020, 최수빈)

학습 데이터 구성

전체 데이터 개수: 약 1600개 정도로 추측 * 수집된 데이터가 1624 row, 해당 논문 42page

학습 데이터 개수: (70%, train0.8 vali0.2)

테스트 데이터 개수: (30%)

하이퍼 파라미터:

Sequence Length: 30 (30일 만큼의 데이터)

모델

앙상블 모델, Deep 4-LSTM Ensemble model : D4LE * 해당 논문 54page

앙상블 학습이란 여러 개의 개별적인 모델을 생성하고, 예측 결과들을 결합하여 더 정확한 예측을 도출하는 방법을 말한다.

실험 방법과 그 결과

실험적으로 하이퍼파라미터를 채택 *해당 논문 57 – 59 page

Epoch(10, 30, 50, 100, 150) + 뉴런(16, 32, 64, 128, 200) 조합으로 정확도 0.7960, Epoch와 뉴런 수 고정 후, window size(10, 7, 5, 3) + gate activation(tanh, relu, linear, softmax) 조합으로 정확도 0.8736

Batch size 2, dropout ratio 0.25, optimizer adam, loss fuction mse 등 * 해당 논문 60page 표 12 참고

네, **시퀀스 길이(Sequence Length)**와 **윈도우 크기(Window Size)**는 거의 같은 의미로 사용됩니다. 둘 다 **시계열 데이터를 모델에 입력하기 위해 사용하는 과거 데이터의 개수**를 정의합니다. 다만, 사용되는 맥락에 따라 용어가 조금 다르게 들릴 수 있습니다:

1. 시퀀스 길이(Sequence Length)와 윈도우 크기(Window Size)의 공통점

- 공통된 역할:
 - 과거 데이터를 일정 개수만큼 묶어서 하나의 입력 데이터로 만드는 방식.
 - 시계열 데이터에서 시간 의존성을 반영하기 위해 사용.
- 결과 데이터의 형식:
 - 윈도우 크기나 시퀀스 길이를 사용하면 입력 데이터는 (배치크기, 시퀀스길이, 특성개수) 형태의 3차원 배열로 변환됩니다.

예시:

원본 데이터:

[100, 105, 110, 115, 120, 125, 130]

- 윈도우 크기 또는 시퀀스 길이 = 3으로 설정하면:
 - 입력 (X):

[100, 105, 110]

[105, 110, 115]

[110, 115, 120]
 - 출력 (y):

↓

[5, 120, 125]

2. 두 용어의 차이점

(1) 윈도우 크기(Window Size):

- 일반적인 데이터 처리 관점에서 사용:
 - 윈도우 크기는 시계열 데이터 전처리 과정에서, 과거 데이터를 묶는 크기를 나타냅니다.
 - 머신러닝(예: Gradient Boosting) 또는 딥러닝 모두에서 사용할 수 있는 개념입니다.
 - 딥러닝 모델을 사용할 때는 **시퀀스 길이**로 이어집니다.

(2) 시퀀스 길이(Sequence Length):

- 딥러닝 모델에서 사용되는 용어:
 - 시퀀스 길이는 주로 LSTM, GRU, Transformer와 같은 시계열 딥러닝 모델의 입력 길이를 정의합니다.
 - 시퀀스 길이는 모델의 입력 구조(*input_shape*)와 밀접하게 연결됩니다.
 - 딥러닝에서는 시퀀스 길이와 특성 개수(*features*)를 반드시 명시적으로 정의해야 합니다.

• 뉴런 수(Neurons):

- 일반적으로 신경망의 은닉층(hidden layer)에 있는 노드(node)의 개수를 의미합니다.
- 예를 들어, LSTM 레이어에서 뉴런 수는 해당 레이어의 출력 차원(output dimension)을 정의합니다.
- 뉴런 수가 많아지면 모델의 복잡도가 증가하며, 더 많은 패턴을 학습할 수 있지만 과적합(overfitting)의 위험도 높아집니다.

5조 자료조사 목록

1주차

- (1) Tracing Cryptocurrency Scams: Clustering Replicated Advance-Fee and Phishing Websites
- (2) The 2024 Crypto Crime Report and 2024 가상자산 연계 투자사기 사례집
- (3) Enhancing Security in Blockchain Networks: Anomalies, Frauds, and Advanced Detection Techniques
- (4) 암호화폐 펌프와 덤프 탐지 모델 연구 : SMOTE 오버샘플링을 적용한 트리 앙상블 기법을 중심으로
- (5) Significant Digits: Using Large-Scale Blockchain Data to Predict Fraudulent Addresses
- (6) LSTM 학습모델의 성능 향상을 위한 디지털 신호 필터의 비교연구
- (7) 블록체인 기반 가상화폐 거래의 보안 위험 및 대응 방안
- (8) AI를 활용한 FDS(사기거래탐지시스템)의 연구
- (9) 전자금융 불법이체사고 방지를 위한 실시간 이상거래탐지 및 분석 대응 모델 연구
- (10) 의사 결정 나무를 이용한 이상 금융 거래 탐지 정규화 방법에 관한 연구
- (11) AI-powered Automation of Fraud Detection in Financial Services
- (12) Enhancing Fraud Detection in Cryptocurrencies with Deep Learning (2024)
- (13) Cryptocurrencies and Future Financial Crime (2022)

2주차

- (14) 암호화폐 종가 예측 성능과 입력 변수 간의 연관성 분석, 박재현, 2022
- (15) 암호화폐 가격 정보 데이터에 대한 상관관계분석 및 회귀테스트 회귀테스트 (2018, 정보처리학회논문지, 권도형, 허주성, 김주봉, 임현교, 한연희)
- (16) On-Chain Data를 활용한 LSTM 기반 비트코인 가격 예측, 2021, 안유진, 오하영
- (17) 그래디언트 부스팅을 활용한 암호화폐 가격 동향 예측 (2018, 정보처리학회논문지, 허주성, 권도형, 김주봉, 한연희, 안채현)
- (18) 암호화폐 가격 예측을 위한 딥러닝 앙상블 모델 개발 연구 (2020, 최수빈)
- (19) An Information Theory Approach to the Stock and Cryptocurrency Market: A Statistical Equilibrium Perspective, Emanuele Citera, Francesco De Pretis, 2023)