감성분석 및 다양한 모델들을 활용한 가상화폐 가격예측:

여러가지 머신러닝, 딥러닝 모델들의 성능 비교를 기준으로

2023020369 장희중

I. 초록

본 연구는 감성분석 및 시계열 모델들을 이용하여 가상화폐 가격을 예측하는 데 초점을 맞추었다. 최근 가상화폐 시장은 급격한 성장을 경험하면서 주요 금융 자산으로 부상하였으며, 이에 따라 정확한 가격 예측의 필요성이 더욱 커지고 있다. 본 연구에서는 비트코인을 대상으로 15분 간격의 시간별 데이터를 분석하였고, 트위터에서 크롤링한 데이터를 활용하여 시장 참여자들의 감성을 분석하였다. VADER 감성 분석 도구를 이용해 각 트윗의 감성 점수를 산출하고, 이를 가상화폐 가격 예측 모델의 입력 변수로 활용하였다.

분석을 위해 전통적 시계열 모델인 ARIMA부터 머신러닝, 딥러닝 모델까지 다양한 모델들을 적합하고 성능을 비교하였다. 연구 결과, GRU 모델이 가장 낮은 RMSE를 기록하며 가장 우수한 예측 성능을 보였으며, 이는 GRU의 시계열 데이터 처리 능력이 가상화폐 시장의 변동성을 잘 포착함을 시사한다. RandomForest Regressor 또한 딥러닝 모델들과 나쁘지 않은 성능을 내었다. 반면, 전통적인 ARIMA 모델은 높은 변동성을 가진 가상화폐 시장의 특성을 충분히 반영하지 못했다.

본 연구는 가상화폐 시장의 동향을 이해하고, 관련된 투자자 및 기관에게 유용한 정보를 제공함으로써, 해당 시장의 건전한 발전에 기여할 수 있을 것으로 기대된다. 또한, 연구의 한계점을 인식하고, 이를 극복하기 위한 향후 연구 방향을 제시하였다.

II. 연구배경 및 목적

최근 몇 년간, 가상화폐는 급속한 발전과 함께 실물화폐의 대안으로서, 그리고 가치 저장 수단으로서의 가능성을 점차 확장하고 있다. 특히 비트코인과 같은 주요 가상화폐들은 전 세계적인 관심을 받으며 금융 시장에서 중요한 역할을 담당하게 되었다. 이러한 추세는 코빗리서치센터의 '2024년 가상자산 시장 전망' 리포트에서도 확인할 수 있다. 리포트에 따르면, 비트코인을 포함한 가상화폐는 내년에 글로벌 자산으로서의 입지를 더욱 강화하고, 전통 금융권에까지 그 영향력을 확대할 것으로 전망된다.

이러한 배경 하에, 본 연구의 목적은 감성분석 및 머신러닝, 딥러닝 기술을 이용하여

가상화폐의 가격 변동을 예측하는 것이다. 가상화폐 시장은 전통적인 금융 시장과는 다른 독특한 동향과 변동성을 보이며, 이는 기존의 예측 모델로는 충분히 설명되지 않는 경우가 많다. 따라서, 이 연구는 가상화폐 시장의 데이터를 분석하고, 감성 분석을 통해 시장 참여자들의 정서적 경향성을 파악함으로써, 가상화폐 가격의 변동성을 보다 정확히 예측하고자 한다. 또한, 다양한 시계열 모델들을 활용하여 이러한 예측의 정확도를 높이고, 실제 시장에 적용 가능한 모델을 개발하는 것을 목표로 한다.

본 연구는 가상화폐 시장의 이해를 깊게 하고, 투자자 및 관련 기관에게 유용한 정보를 제공함으로써, 가상화폐 시장의 건전한 발전에 기여할 수 있을 것으로 기대된다.

III. 데이터 수집

1. 가상화폐 가격 데이터

Binance 사이트에서 API를 통해 15분 단위의 비트코인 데이터를 받아왔다.

2. 트위터 데이터 크롤링

본 연구의 데이터 수집 과정은 트위터의 공개 데이터를 활용하여 진행되었다. 트위터는 실시간으로 사용자의 의견과 반응이 반영되는 소셜 미디어 플랫폼으로, 가상화폐 시장의 동향과 투자자들의 감성을 파악하는 데 매우 유용하다.

데이터 수집을 위해 트위터 API를 통한 크롤링을 구현하였다. snscrape 모듈을 활용하여 사용자가 지정한 검색어, 날짜 범위, 사용자 이름 등에 따라 데이터를 수집할 수 있도록 하였다. 수집된 데이터는 트윗의 날짜, ID, 본문, 사용자 이름, 언어, 해시태그, 답글 수, 리트윗 수, 좋아요 수 등 다양한 정보를 포함한다.

IV. VADER를 통한 감성분석

VADER (Valence Aware Dictionary and Sentiment Reasoner)는 특히 소셜 미디어 텍스트에 최적화된 감성 분석 도구이다. 이 도구는 긍정적, 부정적, 중립적 감성을 포함하는 텍스트를 분석하여, 각각의 감성 점수와 종합적인 감성 지수(compound score)를 산출한다. VADER는 감성 사전에 기반하여 작동하며, 소셜 미디어에서 자주 사용되는 이모티콘, 축약어, 비속어 등을 포함한 다양한 언어 사용 패턴을 고려한다.

- 분석과정

1. 데이터 전처리: 트위터에서 수집한 텍스트 데이터는 먼저 전처리 과정을 거친다. 이 과정에서 URL, 멘션, 해시태그 등이 제거되고, 텍스트가 정제된다.

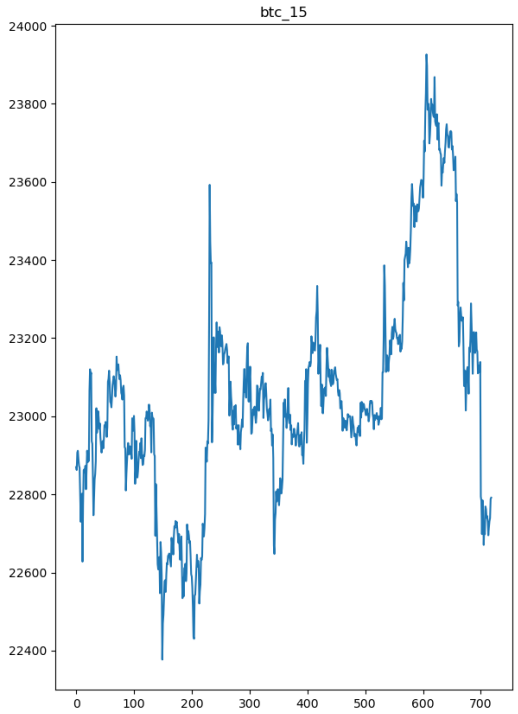
2. VADER 적용: 전처리된 텍스트에 대해 VADER의 'SentimentIntensityAnalyzer' 클래스를 사용한다. 이 클래스는 각 트윗의 텍스트를 입력받아, 감성점수를 분석한다.

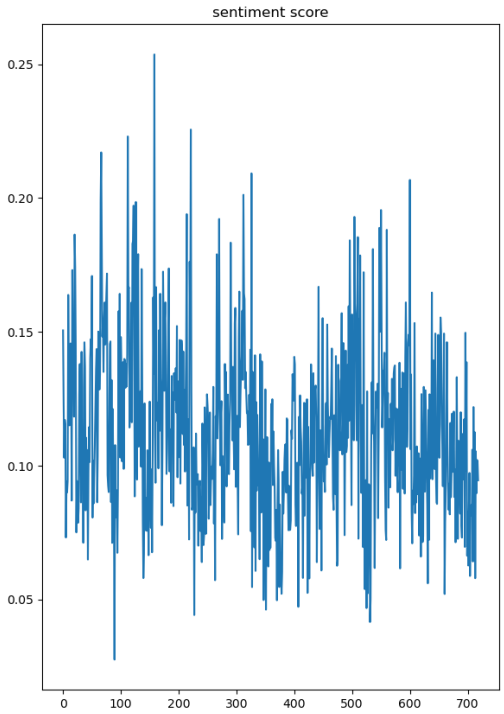
3. 감성점수 산출: VADER는 각 트윗에 대해 'Positive', 'Negative', 'Neutral', 'Compound' 총 네 가지 점수를 산출한다. 'Compound'는 종합적인 감성지수로, -1(매우 부정적)에서 +1(매우 긍정적) 사이의 값을 가진다.

이후 가격예측 모델링에는 Compound 감성지수를 이용한다.

V. 가격예측 모델링

데이터는 비트코인 Open가, Compound 감성점수 두 개의 열로 이루어져 있으며 2023-01-23 12:15:00 ~ 2023-01-31 23:45:00 까지 15분 간격의 일주일치 데이터이다. 과거 168개의 데이터를 가지고 15분 후의 비트코인 open가격을 예측하는 것을 목표로 한다. 비트코인 Open가와 감성점수 각각의 plot은 다음과 같다.





- SVR, RandomForest를 위한 전처리

x열에 시차특성을 추가하여 과거 값을 예측에 활용할 수 있도록 하였다.

- 시계열 데이터 전처리 ; ARIMA, LSTM, GRU, AdaBoost-LSTM, AdaBoost-GRU

1. 이동평균 적용

   데이터의 단기적 변동성을 줄이고 장기적인 추세를 강조하기 위해 이동평균을 적용하였다. 이동 평균의 윈도우 크기는 10으로 설정하였다. 이는 각 시점에 대해 그 이전 10개의 데이터 포인트의 평균을 계산함을 의미한다. 이를 통해 데이터에서의 노이즈 및 임시적 변동을 줄이고, 보다 안정적인 시계열 데이터를 생성한다. 또한 과거 168개의 데이터를 가지고 다음날의 가격을 예측하는 구조로 data를 구성하였다.

2. 데이터 스케일링

   비트코인의 Open가와 감성점수는 서로 다른 범위를 가질 수 있으므로, 데이터를 0과 1 사이의 값으로 조정하는 Min-Max Scaling을 적용하였다. 이는 모델의 학습 효율성을 높이고, 예측 성능을 개선하는 데 기여한다.

- 모델 적합

1. ARIMA

   ARIMA는 자기회귀(AR), 차분(I), 이동평균(MA)를 결합한 시계열 예측 모델이다. AR은 과거 시점의 자기 자신의 데이터가 현 시점의 자기 자신에게 영향을 미치는 모델이며 MA모델은 이동평균 모델로 추세가 변하는 상황에서 적합한 모델이다. I는 차분을 뜻하며 차분을 통해 데이터의 정상성을 만족시킨다.

   ARIMA는 다변량 변수를 넣는 것이 불가능하여 비트코인의 Open가만 이용하여 분석하였다.

   ■ 예측 결과

   Test RMSE: 567

보다시피 예측 성능은 매우 좋지 않다. 변동성이 매우 큰 가상화폐에 대한 가격예측에 ARIMA 모델은 적합하지 않음을 알 수 있다.

2. Support Vector Regressor

Support Vector Regressor(SVR)은 Support Vector Machine(SVM)의 원리를 회귀 분석에 적용한 모델이다. SVM이 주로 분류 문제에 사용되는 것과 달리, SVR은 연속적인 값을 예측하는 데 사용된다. SVR의 기본 아이디어는 데이터 포인트를 고차원 특징 공간으로 매핑하고, 이 공간에서 최적의 회귀 선(또는 초평면)을 찾는 것이다.

SVR에서는 먼저 데이터를 고차원 공간으로 변환하는 커널 함수를 정의한다. 가장 일반적인 커널 함수는 선형, 다항식, 방사형 기저 함수(RBF), 시그모이드 등이 있다. 이 커널은 입력 데이터를 더 높은 차원의 공간으로 변환하여, 선형으로 분리 가능한 형태로 만든다. 본 모델에서 rbf커널을 사용했다.

모델은 두 가지 주요 요소를 사용하여 학습한다:

1) 손실 함수: SVR에서는 ε-감도 손실 함수(ε-insensitive loss function)를 사용한다. 이 함수는 정해진 ε 범위 내에서의 오차는 무시하며, 범위를 벗어난 오차에 대해서만 패널티를 부여한다. 이는 모델이 데이터의 작은 변동성에 과도하게 반응하지 않도록 하는 동시에, 큰 오차에 대해서는 민감하게 반응하도록 한다.

2) 정규화: SVR은 정규화 파라미터 C를 사용하여 모델의 복잡도를 조절한다. C 값이 크면 모델은 훈련 데이터에 더 잘 맞춰지려 하고, C 값이 작으면 모델은 더

많은 오류를 허용한다. 이는 모델의 일반화 능력과 과적합 경향 사이의 균형을 찾는 데 중요하다.

본 모델은 다음과 같은 파라미터를 사용하였다. 이 파라미터는 gridsearchCV 를 통해 결정되었다.
C=20, gamma=0.001, epsilon=0.2, kernel='rbf'

■ 예측 결과

Test RMSE: 262.51

3. RandomForest Regressor

Random Forest Regressor는 앙상블 학습 방법의 일종으로, 여러 결정 트리(Decision Trees)를 결합하여 작동한다. 이 방법은 개별 결정 트리의 예측을 평균내어 단일 예측을 생성한다. Random Forest는 각각의 결정 트리가 데이터의 다른 부분에 대해 학습하도록 함으로써 과적합을 줄이고 일반화 성능을 향상시킨다. Random Forest 모델은 다음과 같은 주요 특징을 가진다.

1) 부트스트랩 샘플링

Random Forest는 훈련 데이터에서 여러 부트스트랩 샘플(즉, 중복을 허용한 무작위 샘플)을 생성한다. 각각의 결정 트리는 이러한 부트스트랩 샘플 중 하나를 사용하여 훈련된다.

2) 특성의 무작위 선택

모델은 각 분할에서 후보 특성의 무작위 서브셋을 사용한다. 이는 트리들이 서로 다른 특성의 조합을 고려하도록 하여, 트리들 간의 상관관계를 줄이고 모델의 다양성을 증가시킨다.

3) 결정 트리의 앙상블

모든 트리의 예측은 평균 되어 최종 예측을 형성한다. 이는 개별 트리의 과적합을 상쇄시키고 전체 모델의 안정성과 정확성을 향상시킨다.

4) 파라미터 조정

Random Forest의 성능은 트리의 개수, 트리의 최대 깊이, 분할에 사용되는 최소 샘플 수 등과 같은 하이퍼파라미터에 의존한다. 이러한 파라미터는 주어진 문제에 따라 조정되어야 한다.

본 모델은 다음과 같은 파라미터를 사용하였다. 이 파라미터는 gridsearchCV 를 통해 결정되었다.
n_estimators=100, max_depth=6

■ 예측 결과

Test RMSE: 67.3

4. LSTM

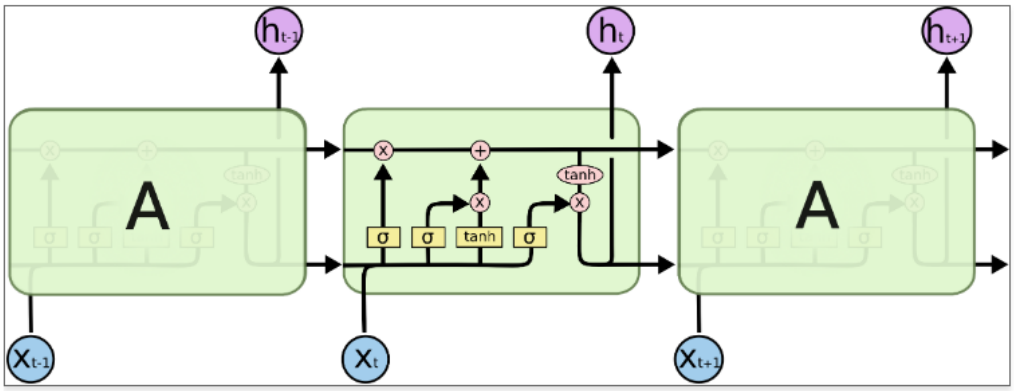
LSTM은 순환 신경망(Recurrent Neural Network, RNN)의 한 종류로, 특히 장기 의존성 문제를 해결하기 위해 설계된 모델이다. 이 모델은 시계열 데이터에서 장기간에 걸친 패턴을 학습하고 기억하는 능력이 뛰어나다.

LSTM은 셀 상태와 게이트라는 두 주요 구성 요소를 가지고 있다. 셀 상태는 정보를 장기간 동안 유지하는 데 도움을 주며, 게이트는 셀 상태에 어떤 정보를 추가하거나 제거할지 결정한다. 게이트는 아래와 같이 세 종류로 구성된다:

1) 망각 게이트(Forger Gate): 셀 상태에서 불필요한 정보를 제거한다.

2) 입력 게이트(Input Gate): 새로운 정보를 셀 상태에 추가하는 역할을 한다.

3) 출력 게이트(Output Gate): 셀 상태를 바탕으로 다음 hidden state를 생성한다.



■ 예측 결과

Test RMSE: 38

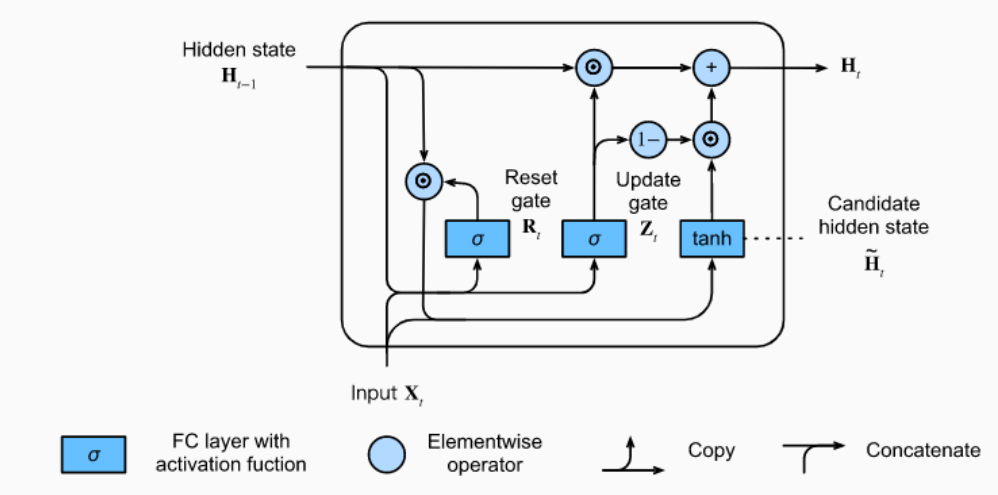
5. GRU

GRU는 순환 신경망(Recurrent Neural Network, RNN)의 일종으로, LSTM과 유사한 구조를 가지지만 더 간소화된 형태이다. 특히 시계열 데이터의 장기 의존성을 처리하고, 중요한 정보를 효과적으로 학습하는 데 사용된다.

GRU의 주요 특징은 '업데이트 게이트'와 '리셋 게이트' 두 가지 유형의 게이트를 포함한다는 것이다. 이 게이트들은 시계열 데이터에서 중요한 정보를 어떻게 유지하고, 불필요한 정보를 어떻게 제거할지 결정한다.

1) 업데이트 게이트(Update Gate): 이 게이트는 셀의 정보를 얼마나 유지할지 결정하며 LSTM의 망각 게이트와 입력 게이트의 기능을 함께 수행한다.

2) 리셋 게이트(Reset Gate): 이 게이트는 과거의 정보를 얼마나 무시할지 결정하며 이를 통해 모델이 과거 데이터의 중요도를 조절할 수 있다.



가상화폐 시장은 고도로 변동적이며, 시간에 따라 정보의 중요성이 달라질 수 있다. GRU는 이러한 변동적인 데이터에서 중요한 정보를 식별하고, 불필요한 정보를 제거함으로써 효율적으로 시계열 데이터를 분석한다.

■ 예측 결과

Test RMSE: 29.9

6. AdaBoost

AdaBoost는 앙상블 학습 방법 중 하나로, 여러 '약한 학습기'(weak learners)를 결합하여 강력한 예측 모델을 구성하는 기법이다. 이 방법은 각각의 약한 학습기가 데이터의 다른 측면을 학습하여, 결합될 때 전체적인 예측 성능을 향상시킨다.

AdaBoost는 여러 라운드에 걸쳐 학습기를 순차적으로 추가한다. 각 라운드에서는 이전 라운드의 오류를 분석하고, 잘못 분류된 데이터에 더 많은 가중치를 부여한다. 이를 통해 알고리즘은 오류를 줄이는 방향으로 학습을 계속 진행한다. 학습이 진행될수록, 각 학습기는 특정 부분의 데이터에 더 집중하게 되며, 이는 모델의 강점과 약점을 보완하는 데 도움이 된다.

AdaBoost를 사용하면 여러 학습기의 결합으로 오류가 줄어들고, 전체 모델의 예측 정확도가 향상된다. 또한 잘못 예측된 데이터에 대해 학습기가 더 집중하게 되어 모델의 오류를 줄이는 데 효과적이다.

A. AdaBoost-LSTM

AdaBoost 알고리즘을 LSTM 모델에 적용한 앙상블 학습 방법이다. 이 접근법은 LSTM의 장기의존성 학습 능력과 AdaBoost의 오류보정능력을 결합하여, 시계열 데이터의 예측정확도를 높인다.

■ 예측 결과

Test RMSE: 49.3

B. AdaBoost-GRU

AdaBoost 알고리즘을 GRU모델에 적용한 앙상블 학습 방법이다. 이 방법은 GRU의 시계열 데이터 처리 능력과 AdaBoost의 오류보정기능을 결합하여, 시계열 데이터의 예측 정확도를 향상시킨다.

■ 예측 결과

Test RMSE: 109

- 모델 비교

Model	SVR	RF	ARIMA	LSTM	GRU	AdaBoost-LSTM	AdaBoost-GRU
Test RMSE	262.5	67.3	567	38	29.9	49.3	109

SVR, RF, ARIMA, LSTM, GRU, AdaBoost-LSTM, AdaBoost-GRU 총 7가지의 모델들을 비교해 본 결과, 가장 낮은 RMSE를 갖는 GRU가 제일 좋은 예측 성능을 가진다.

VI. 결과 및 한계점

- 결과

  본 연구에서는 감성분석 및 다양한 머신러닝 모델, 시계열 딥러닝 모델들을 활용하여 가상화폐 가격을 예측하였다. 연구 결과, GRU 모델이 가장 낮은 RMSE 값을 보여 가장 우수한 예측 성능을 나타냈다. 이는 GRU가 가상화폐 시장의 복잡한 시계열 데이터에서 중요한 정보를 효과적으로 포착하고, 장기적인 데이터 패턴을 학습하는 데 뛰어난 능력을 가지고 있음을 시사한다.

  RandomForest도 나쁘지 않은 성능을 보였으나 피처 추가 등 모델의 개선이 필요해 보인다. LSTM과 AdaBoost를 결합한 AdaBoost-LSTM모델도 상대적으로 낮은 RMSE 값을 보였으나, 순수 LSTM이나 GRU 모델에 비해 성능이 떨어지는 것으로 나타났다. 이는 AdaBoost 알고리즘의 가중치 조정 방식이 복잡한 시계열 데이터에서는 예상보다 효과적이지 않았을 수 있음을 의미한다.

  전통적인 시계열 분석 모델인 ARIMA는 예측 성능이 가장 낮았다. 이는 ARIMA 모델이 가상화폐 시장의 높은 변동성과 비정상성을 효과적으로 처리하지 못함을 보여준다.

- 한계점

  1. 데이터 범위

     본 연구는 제한된 시간 범위의 데이터에 기반하여 진행되었다. 이는 모델의 일반화 능력에 대한 평가를 제한할 수 있다. 더 넓은 범위의 데이터를 사용하여 추가적인 연구가 필요하다.

  2. 단일 가상화폐 분석

     연구는 비트코인 가격 데이터에 초점을 맞추었다. 다른 가상화폐의 가격 데이터를 분석하여 결과의 일반성을 검증할 필요가 있다.

  3. 감성 분석의 한계

     감성 분석은 주관적이며, 다양한 사회적, 정치적 요인에 의해 영향을 받을 수 있다. 감성 데이터의 해석과 이의 모델에 대한 영향력은 보다 심도 있는 분석이 필요하다.

VII. 향후 연구방향

더 광범위한 데이터와 다양한 가상화폐에 대한 연구를 통해 모델의 강점과 약점을 더욱 명확히 파악할 수 있을 것이다.

추가적인 앙상블 기법과 더 발전된 머신러닝, 딥러닝 기법을 탐구하여 가상화폐 가격 예측의 정확도를 높일 수 있는 가능성을 모색한다.

VIII. 참고자료

- 커지는 비트코인 낙관론…"내년 금융권으로 저변 확대"

  https://kr.investing.com/news/cryptocurrency-news/article-978923