학 사 학 위 논 문

비트코인과 투자자산과의 상관성 분석 및 이를 활용한 비트코인 시세 예측

지도교수 주 용 성

동국대학교 통계학과

김정의

2021

목차

표 및 그림 목차

Ⅰ.서론

비트코인은 2008년 10월 사토시 나카모토라는 익명의 프로그래머가 만든 가상화폐로써 블록체인기술을 바탕으로 해킹이 불가능한 특성인 “안전성” 거래내역이 모두 공개되는 “투명성”이 큰 장점이다. 비트코인은 가격변동의 제한이 있는 주식과 달리 하루 동안에 가격이 큰 폭으로 증가하고 감소하는 등 예측 불가능성으로 유명하였다. 하지만 최근에 여러 보고서에서 비트코인이 주식, 금 등의 투자자산과 상관성이 보인다는 여러 보고서가 나타나고 있다. 최근 비트코인 분석업체인 코인메트릭스의 보고서에 따르면 금과 비트코인의 60일 상관관계는 0.5를 상회하는 수준으로 높아졌으며 이는 역대 최고치라는 보고가 있었다(Coin metrics, 2020). 하지만 두 시계열 데이터의 추세를 비교함에 있어 상관계수를 이용한 방법은 선형성만을 나타내는 척도로써 시간에 따른 두 시계열 데이터의 추세를 비교하기 위한 가장 좋은 척도라고 말할 수 없다. 따라서 시계열 데이터에서 두 추세를 비교하기 위해 다양한 방법으로 시계열데이터의 유사성을 파악하고자 한다. 또한, 이를 바탕으로 비트코인의 시세를 예측함에 있어 비트코인의 시세만을 활용하여 예측하는 경우와 투자자산의 시세를 추가적으로 활용하여 시세를 예측하는 경우들 중 어느 모델이 가장 성능이 좋은지를 테스트해보고자 한다. 예측모델로써는 최근 가장 큰 주목을 받고 있고 성능이 뛰어난 딥러닝 모델 LSTM을 사용할 것이다. ARIMA와 같은 통계적모형 대신 LSTM을 활용하는 이유는 비트코인가격의 예측에 있어 LSTM이 ARIMA등의 통계적 모형보다 성능이 더 뛰어나다는 여러 연구결과가 있기 때문이다(지세현, 2020). 이러한 모델링을 통하여 투자자산과 비트코인과의 상관성이 높다면 이러한 정보를 LSTM모델이 학습할 때 반영하여 예측성능이 더 높아질 것이라고 생각하였다.

본 논문의 1장에서는 사용데이터의 시기 및 분석에 사용된 투자자산의 종류 등 데이터들에 대해 간략히 설명할 것이다. 2장에서는 비트코인과 주식, 금, 달러 등 투자자산과의 상관성을 다양한 방법을 통하여 파악하고자 할 것이다. 이를 바탕으로 각 방법에 따라 가장 상관성이 높은 자산은 무엇인지를 파악해볼 것이다. 마지막으로 3장에서는 비트코인의 시세를 예측하기 위해 LSTM을 사용하여 모델을 만들고자 한다. 이때 비트코인의 과거 시세만을 사용한 모델과 다른 투자자산을 사용한

모델을 비교하여 가장 좋은 성능을 보여주는 모델이 어떤 것인지 파악하고자 한다.

Ⅱ.본론

1장 사용데이터 설명

1.1 데이터의 시기 및 단위

본 연구에서는 2020년 1월1일부터 논문을 쓰고 있는 현재 날짜인 9월 30일까지의 비트코인 및 투자 자산들의 데이터를 사용하였다. 그동안 비트코인의 가격은 지나치게 변동이 심하였고 여러 재화와의 연관성이 있다는 보고서들 대부분이 2020년의 일로써 그전의 데이터를 사용하는 것은 최근에 높아진 비트코인과 투자자산의 상관성을 파악하기에 너무 과거의 데이터라 판단하였다. 또한, 비트코인은 시간, 분별로 가격이 짧은 시간에도 매우 크게 변하는 특성을 가졌지만 단위를 시간으로 잡을 경우 지나치게 불균형성이 커질 것을 우려하여 비트코인의 가격 추세를 파악하는 데 좋지 않을 것으로 생각하였다. 따라서 일별 시세를 기준으로 하였으며 이에 따라 비트코인 외에 투자자산도 하루를 기준으로 하기로 하였다.

마지막으로 365일 거래할 수 있는 비트코인과 달리 휴일에는 거래할 수 없는 주식, 금, 달러 와 같은 투자자산과의 시점을 맞춰 주기 위해 이들 자산의 거래가 없는 날짜들은 제외하였다.

1.2 비트코인과의 상관성 파악을 위한 주요 투자자산

비트코인과의 투자자산 사이의 상관성 파악을 위하여 다우존스 산업평균지수, S&P 500 , 나스닥 , Dollar Index , 금을 주요 투자자산으로 사용하였다.

다우존스 산업평균지수, S&P 500 , 나스닥은 세계 경제를 대표하는 미국의 3대 주가지수라고 불릴 만큼 대표성을 가지고 있다고 생각하였다. 또한, Dollar Index의 경우 유로, 엔, 파운드, 캐나다(달러), 스웨덴(크로나), 스위스(프랑) 에 대한 달

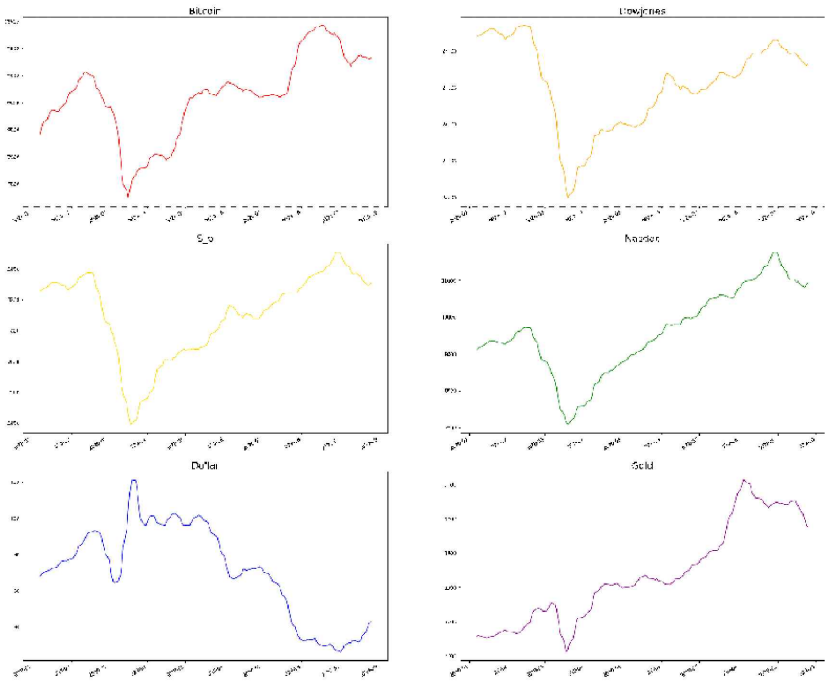
러의 가치를 바탕으로 구해진 값으로써 비트코인과 달러와의 상관성을 파악하기 위하여 주요 투자 자산으로 포함했다.

마지막으로 안전자산으로써 대표성을 가지는 금을 포함했다. 금은 오래전부터 안전자산의 대표 격이며 또한 최근 다양한 보고서에서 금과 비트코인의 상관성이 높아지고 있다는 보고가 많은 만큼 주요 투자자산으로써 포함했다.

2장 비트코인과 투자자산과의 상관분석

2.1 시각화를 통한 분석

우선 각 시계열 데이터에서 추세성분과 오차 등을 분리하여 추세성분을 분리하였다. 이를 통하여 각 자산의 추세를 시각화를 통하여 비교하기 쉽게 파악할 수 있도록 하였다. 분해 방법으로는 다음과 같은 Multiplicative Model을 사용하였는데 이는 비트코인과 투자자산들의 주기와 변화폭이 일정하지 않아 Addittive 모델보다 Multiplicative 모델이 더 적합하다고 생각하였기 때문이다.



[그림 1] 비트코인과 투자자산들의 시계열 그래프 왼쪽 위부터 시계방향으로 Bitcoin , Dowjonse, Nasdac , Gold , Dollar index , S&P 500.

그 결과 Dollar index를 제외하고 2020년 4월 초에 큰 낙폭이 있는 것이 유사하였으며 낙폭이 있고 난 뒤 4개의 그래프 모두 유사한 패턴을 보여주었다.

2.2 수치적 비교

서론에서 금과 비트코인의 상관관계가 0.5에 도달했다는 보고서에 대해 언급하였다. 하지만 앞서 말했듯이 시계열 데이터에서 상관성 파악을 위해 일반적인 상관분석에서 사용하는 상관계수를 사용하는 것은 선형성만을 나타내므로 시계열데이터의 시간에 따른 추세를 비교하기에는 좋은 방법이 아니다.

이를 해결하기 위해 시계열 데이터에서 두 데이터를 비교하는 다양한 방법들을 사용해보고자 하였다. 유클리드 거리, 코사인유사도, DTW를 이용하여 두 시계열 데이터의 유사성을 비교하고자 한다. 유사성을 비교하기에 앞서 투자자산들마다 단위가 다르기 때문에 최소 최대 정규화를 통하여 값을 0부터 1의 값으로 표현하였다.

1. 유클리드 거리

유클리드거리 =

$$\sum_{t=1}^{n} d_t = \sum_{t=1}^{n} |y_t - x_t|$$

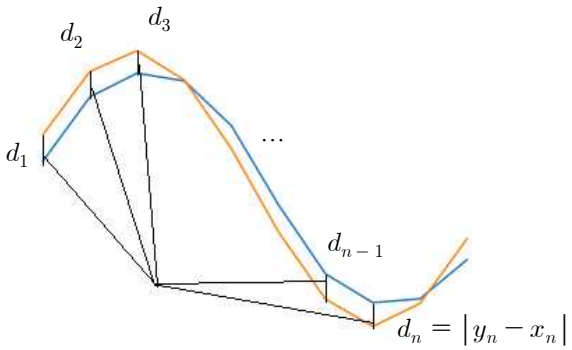


그림 <2> 유클리드 거리를 통한 시계열데이터의 비교

유클리드 거리를 통한 방법은 각 t시점에서 두 시계열데이터의 값의 차에 절댓값을 씌운 후 이를 모두 합하여 구한다.

2. 코사인 유사도

$$\text{코사인 유사도} = \frac{\sum_{t=1}^{n} x_t \times y_t}{\sqrt{\sum_{t=1}^{n} (x_t)^2 \times \sum_{t=1}^{n} (y_t)^2}}$$

코사인 유사도는 두 시계열 데이터를 각각 벡터로 표현한 후 두 벡터를 내적을 통해 구한 값을 각각의 벡터의 크기의 곱으로 나눠서 구한 것과 같다. 두 데이터가 완전히 같다면 값이 1이 나오게 되고 유사하지 않다면 0에 가까운 값이 나오게 된다.

3.DTW (Dynamic Time Warping)

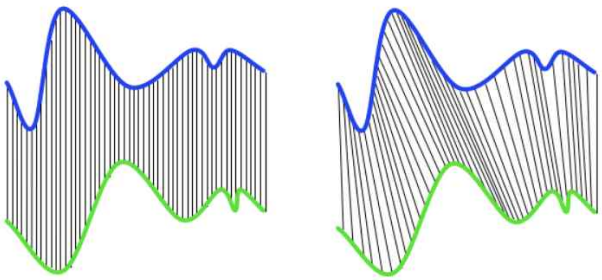


그림 <3> 유클리드거리 방법과 DTW 방법

DTW는 두 시계열의 시점이 완전히 일치하지 않더라도 두 시계열 데이터의 추세를 잘 비교할 수 있게 해준다. 위의 그림을 보면 두 그래프의 추세가 매우 유사하지만 각 시점의 차이에 의하여 유클리드 거리를 기준으로 판단하게 된다면 그 유사성이 떨어진다고 나올 수 있다. 하지만 오른쪽 그림과 같이 DTW는 두 시계열데이터의 추세가 유사하지만 시간의 차이에 의하여 비교가 어려울 때 매우 유용한 지표이다. 두 자산 간의 가격변동이 상관성이 있다고 할 때 그날에 바로 영향을 주는 경우가 아니라 하루 뒤 혹은 이틀 뒤에 가격이 반응한다고 하였을 때 앞의 두 알고리즘보다 더 효율적인 비교 방법이라 할 수 있다. DTW를 구하는 공식은 다음과 같다.

$$d(i,j) = |x_i - y_j|$$

$$D(i,j) = d(i,j) + \min(D(i-1,j-1), D(i-1,j), D(i,j-1))$$

$$W_n = D(n,m)$$

$$W_{n-1} = \min(D(n-1,m), D(n-1,m-1), D(n,m-1)$$

.....

$$W_1 = \min(D(1,2), D(1,1), D(2,1))$$

$$DTW = \sum_{i=1}^{n} W_i$$

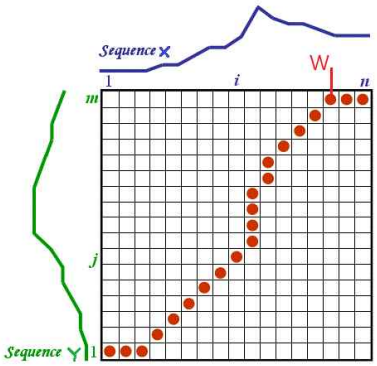


그림 <4> DTW를 통한 두 시계열 데이터비교

	Euc	Euc30	Euc60	Euc90	Dtw	Dtw30	Dtw60	Dtw90	Cos	Cos30	Cos60	Cos90
Dowjones	27.890309	5.952045	6.787677	4.703425	2.430368	1.071634	0.736669	0.692282	0.966525	0.884458	0.895463	0.889042
S_p	19.902158	5.658347	4.596066	5.645359	1.452813	1.057285	0.696374	0.919933	0.981226	0.889596	0.944956	0.882970
Nasdac	18.406314	5.222986	5.118835	8.137938	0.787705	1.025692	0.685040	1.299223	0.984932	0.898798	0.921623	0.780034
Dollar	64.673494	12.449251	18.671855	8.247649	4.899482	1.938920	3.435511	1.421713	0.707561	0.590249	0.332833	0.728055
Gold	31.991348	8.519405	3.111821	10.372271	1.708955	0.924857	0.602764	1.775484	0.945552	0.834563	0.970891	0.713320

<표 1> 비트코인과 투자자산과의 유사성 파악.

	Euc	Euc30	Euc60	Euc90	Dtw	Dtw30	Dtw60	Dtw90	Cos	Cos30	Cos60	Cos90	Total
Dowjones	3.0	3.0	4.0	1.0	4.0	4.0	4.0	1.0	3.0	3.0	4.0	1.0	35.0
S_p	2.0	2.0	2.0	2.0	2.0	3.0	3.0	2.0	2.0	2.0	2.0	2.0	26.0
Nasdac	1.0	1.0	3.0	3.0	1.0	2.0	2.0	3.0	1.0	1.0	3.0	3.0	24.0
Dollar	5.0	5.0	5.0	4.0	5.0	5.0	5.0	4.0	5.0	5.0	5.0	4.0	57.0
Gold	4.0	4.0	1.0	5.0	3.0	1.0	1.0	5.0	4.0	4.0	1.0	5.0	38.0

<표 2> 순위로 표현한 비트코인과 투자자산과의 유사성.

3가지의 방법대로 비트코인과의 유사성을 비교한 결과 다음의 표와 같았다. 첫 번째 표는 자산별 비트코인과의 유사도를 기간에 따라 구별하여 구한 것이다. 예를 들어, EUC는 2020년 1월1일부터 9월 30일까지 기간 전체에 대하여 유사도를 구한 것이며 EUC30, EUC60, EUC 90은 가장 최근인 9월 30일을 t-1 시점이라 하면 t-1 ~ t-30 , t-31 ~ t-60 , t-61 ~ t-90 의 크기로 나눈 것이다. 두 번째 표에선 이를 자산별로 등수를 매겨 표시하고 순위의 합을 표시하였다. 전체적인 순위를 봤을 때 비트코인과 가장 유사한 자산은 Nasdaq으로 나타났으며 S&P 또한 비슷한 결과를 보여주었다. 나머지 3변수는 비트코인과의 유사성이 S&P 와 Nasdaq에 비해 크다고 말할 수 없었다. 특히 Dollar index의 경우 모든 방법, 기간에서 좋지 않은 결과를 보여주었다. 또한, 눈여겨볼 점으로 금의 경우 30일 구간, 60일 구간에서 DTW 값이 가장 높은 경향을 보여줬으며 이는 비트코인과 금이 비교적 최근에 유사성이 높아졌으며 약간의 시차를 두고 경향을 따라가는 경향을 가질 수 있다고 생각하였다.

3장

3.1 모델 생성

1. LSTM

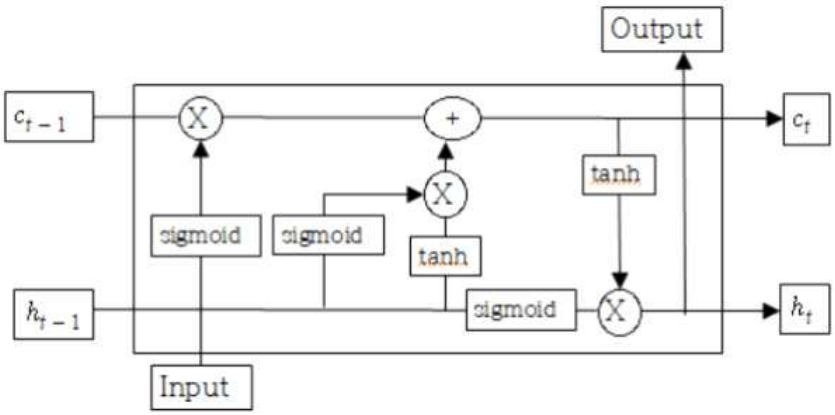


그림 < 5 > LSTM 모델 구조

주가 예측 모델로써 최근 가장 주목을 받는 모델인 LSTM을 사용하였다.
LSTM은 딥러닝의 한 종류로써 RNN(순환신경망)의 특수한 버전이다. 정보가 들어오게 되면 단기정보를 가지고 있는 h_{t-1} 히든층 과 함께 현재의 정보를 장기 기억층 c_t에 반영할지를 확률을 통하여 표현한다. 예컨대, 반영할 확률이 0.1이라면 잊히는 정보가 0.9라면 계속 기억할 정보로 표시하는 방식이다. 또한, c_{t-1}는 h_{t-1}, Input값과 함께 다음 층에 반영할 h_t 값을 결정하고 Output값을 생성하게 된다.
LSTM은 이렇게 장단기 정보를 반영할 수 있다는 장점 때문에 주가예측과 같은 시계열 데이터 예측에서 자주 사용되는 모델이다.

2. 모델 구조 설명

```
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 10, 30)            3960
_________________________________________________________________
lstm (LSTM)                  (None, 40)                11360
_________________________________________________________________
dense (Dense)                (None, 1)                 41
=================================================================
```

다음과 같이 LSTM층을 2개로 쌓은 모델을 만들었다.

예측하고자 하는 날짜의 하루 전부터 10일 전까지의 비트코인의 종가와 재화의 종가가 입력층에 들어가게 된다. 10일보다 많은 20일, 30일 전의 데이터까지 반영하게 되면 지나치게 변수가 많아지게 되어 그림 <6> 과 같이 예측이 잘 되지 않았다.

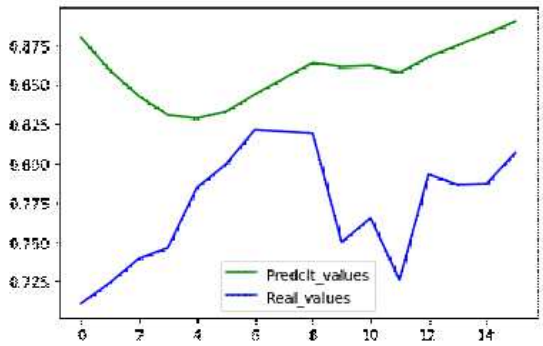


그림 < 6 > 30일 데이터를 바탕으로 예측한 모델

첫 번째 LSTM 층을 통해 30개의 Output이 나오게 되며 나온 아웃풋들은 다시 LSTM 층에 들어가 40개의 아웃풋을 가지게 된다. 최종적으로 마지막 층에서 linear 한 활성화 함수를 거침으로써 최종적으로 하나의 값이 나오게 되며 이 값이 10일간의 데이터를 바탕으로 하루 뒤의 시세를 예측한 값이 되게 된다.

모델의 손실함수로는 Mse (Mean Squared Error)를 사용하였으며 Optimizer로는 Adam을 사용하였다. Adam은 RMSprop과 Momentum 방식의 장점을 합쳐 만들어진 알고리즘으로써 가장 널리 사용되고 있는 Optimizer이다.

마지막으로 모델을 학습시키는 수인 Epochs의 값은 기본적으로 100으로 하되 모델별로 Loss_function의 값이 더 감소하지 않고 충분히 수렴했다면 학습을 멈추어 과적합이 되는 것을 막았다.

3. 모델 평가

1월 1일부터 9월 30일까지 날짜 중 주식시장이 개장하지 않은 날짜들을 제외하였으며 또한 학습 과정에서 10일 전의 데이터들이 필요하였기 때문에 1월 16일부터 9월 30일까지 총 179개의 데이터 중에서 163개를 Train data로 약 10% 정도인 16개를 테스트 데이터로 사용하였다. 시계열 데이터로써 순차적으로 배열되어야 했기 때문에 Cross Validation을 사용할 수 없었고 테스트 데이터를 많이 잡을 수 없

었지만 16개의 데이터면 약 3주간을 예측하는 것으로서 모델의 성능을 검증하기에 충분하다고 생각하였다. 모델의 성능을 판단하는 기준으로는 MSE와 MAE를 사용하였다. MSE 외에 MAE를 판단기준에 넣은 것은 MSE의 경우 모델 예측값에 극단치가 있을 경우 지나치게 영향을 받을 것으로 생각하여 MAE를 같이 고려하기로 하였다.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \hat{Z}_i)^2 \qquad \text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Z_i - \hat{Z}_i\right|$$

마지막으로 딥러닝은 블랙박스모델이라는 수식어가 붙을 만큼 모델내부의 동작과정을 알기가 어려우며 역전파를 통하여 수정되는 가중치들이 같은 모델이라 할지라도 매 시행마다 예측값이 약간씩 차이가 나게 된다. 따라서 5번의 시행을 통하여 평가하고 최종적으로 이를 평균을 내어 결과를 표시하기로 하였다. 가장 먼저 비트코인의 과거시세만을 사용하여 현재의 비트코인 시세를 예측하는 단일모델과 [비트코인의 과거시세 , S&P 지수], [비트코인의 과거시세 , Dowjones] 와 같이 투자자산을 같이 학습한 모델을 만들어 총 6개의 모델을 만들고 평가하였다.

시행 횟수 = n		n=1	n=2	n=3	n=4	n=5	평균
단일모델	MSE	0.002634	0.002180	0.007620	0.002785	0.003079	0.003660
	MAE	0.046323	0.042808	0.063790	0.045085	0.050452	0.049692
S&P	MSE	0.002308	0.000729	0.006116	0.003997	0.000809	0.002792
	MAE	0.041127	0.019729	0.062772	0.056426	0.021550	0.040321
Dow jones	MSE	0.002306	0.001874	0.002167	0.002297	0.004217	0.002573
	MAE	0.038496	0.038761	0.036387	0.043783	0.051136	0.041713
Nasdac	MSE	0.002890	0.001007	0.001098	0.002203	0.000754	0.001591
	MAE	0.047858	0.026701	0.028924	0.041994	0.024225	0.033941
Gold	MSE	0.000965	0.001839	0.000747	0.000923	0.000795	**0.001055**
	MAE	0.024980	0.038240	0.020928	0.023898	0.021931	**0.025996**
Dollar	MSE	0.002934	0.002827	0.005472	0.005100	0.004592	0.004185
	MAE	0.042913	0.048401	0.067360	0.0646801	0.061516	0.056974

표 < 3 > 모델들의 성능 비교 표

5번의 시행 후 MSE와 MAE의 평균을 구한 결과 금을 사용하였을 때 MSE의 평균은 0.001055, MAE의 평균은 0.025996으로 가장 낮았으며 그 다음으로 나스닥을 사용하였을 경우 MSE의 평균은 0.001591, MAE의 평균은 0.033941으로 금과 유사하지만 약간 낮은 결과를 보였다.

가장 좋은 성능을 보여주었던 금의 경우 비트코인만을 사용한 모델보다 MSE 값

의 평균을 0.002605만큼 낮추었으며 MAE의 값은 0.023696 만큼 감소시켰다. 또한 상관성이 가장 낮게 나온 투자자산이었던 Dollar의 경우 오히려 모델의 정확도가 떨어지는 결과가 나오게 되었다. 가장 좋은 예측력을 보여주었던 모델인 금은 다음과 같은 결과를 보여주었으며 사용한 세 번째 시행 모델로 다음과 같은 결과를 보여 주었다. (그림 <7> 의 위 부터 오른쪽방향 3번째 그림 초록선은 예측결과, 파란선은 실제값이다.)

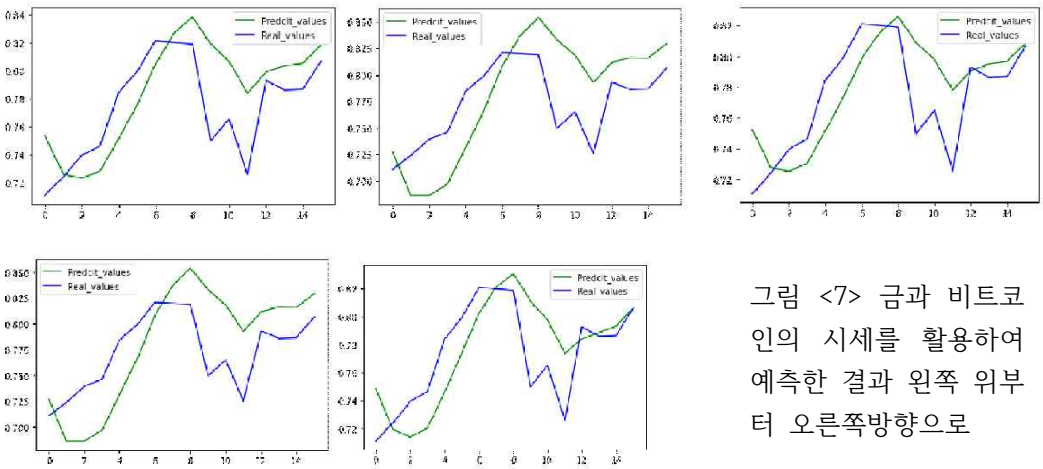


그림 <7> 금과 비트코인의 시세를 활용하여 예측한 결과 왼쪽 위부터 오른쪽방향으로

시각화를 통하여 모델의 성능을 확인해보았을 때도 모델이 실제 가격의 추세를 어느 정도 따라가는 것으로 보였다. 오차가 가장 큰 시점은 9시점과 11시점으로 약 0.06의 차이를 보였다. 하지만 Bitcoin의 가격이 Min_max 정규화를 통하여 표현한 값이기 때문에 이를 원래의 단위인 $로 다시 바꾸어 값을 계산할 경우 오차가 어느 정도 존재하였다. 예컨대, 오차가 가장 큰 t=11 의 시점의 경우 0.78과 0.72를 원래 값으로 복원하여 차이를 구하면 실제 비트코인의 가격인 10739.4$인 11시점에서 약 447달러의 오차가 발생하였다.

Ⅲ.결론

비트코인과 여러 투자자산의 상관성을 파악해보고 투자자산을 사용한 이변량 LSTM모델의 성능을 테스트해보았다. 유사성이 가장 낮게 나왔던 달러 index를 제외하고 모든 투자자산의 경우에 비트코인의 시세만을 사용한 단일 모델보다 성능이 올라간 것을 확인할 수 있었다. 유사성을 파악할 때 S&P와 나스닥은 종합적인 지표에서는 높은 성적을 보여주었지만 LSTM을 통해 모델을 만들었을 때 가장 좋은

성적을 보여준 것은 금을 활용한 모델이었다. 9월 9일부터 9월 30일까지의 테스트 데이터를 통하여 시세를 예측함에 있어 금은 테스트 데이터의 시기와 가까운 시기인 30일과 60일 기간에서 DTW를 통해 유사성을 파악할 때 가장 좋은 성능을 보여주었다. 이러한 결과를 바탕으로 했을 때 LSTM 모델을 사용하여 비트코인의 가격을 예측할 때 DTW값이 높은 변수를 사용함으로써 정확도를 높일 수 있으며 추가적으로 시계열데이터의 유사성을 파악할 때 DTW를 통한 방법의 효과성과 LSTM과 DTW의 상관성에 대하여 추가로 연구해볼 필요성을 제시한다.

참 고 문 헌

[1] Coin Metrics (2020, 06) , CMBI Single Asset Indexes add Binance. US and remove bitFlyer as Constituent Exchanges.

[2] 지세현, 백의준, 신무곤, 구영훈, 윤성호 (2019), 비트코인 트랜젝션 수 예측을 위한 LSTM 모델 설계.