Encoder Decoder LSTM 구현 및 실험 결과 보고 (Time series data split / some issues)

Tae-Won Lee,

leetee95@gmail.com

2020. 03. 23.

Introduction

- 기존의 LSTM등 RNNs의 모형에서 input length와 output length가 다른 데이터를 다룰 수 없는 문제점이 있음.

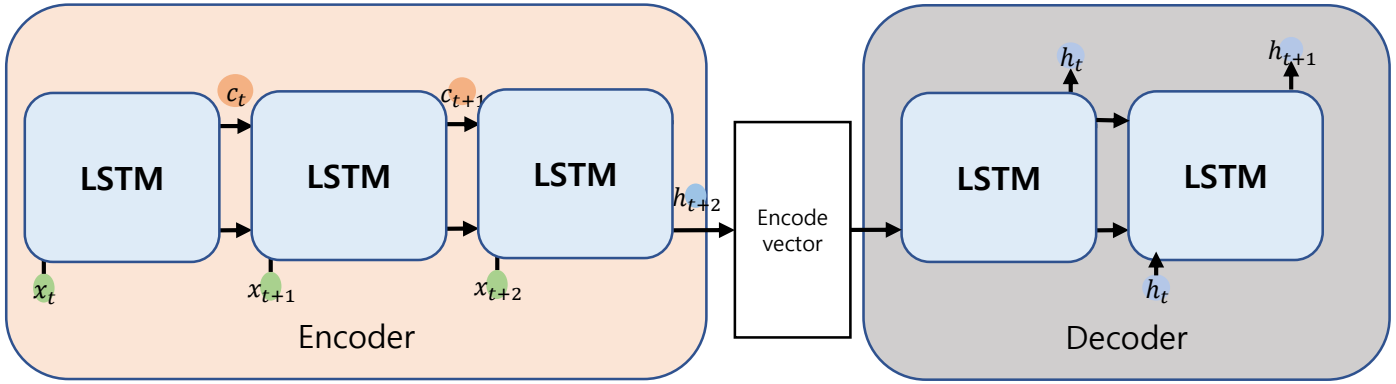
- 이를 해결하기 위해 Encoder-Decoder 구조를 LSTM에 적용함.

- input lengt와 output length가 다른 길이가 되어도 출력이 가능하게 됨.

Related works

- Encoder-Decoder LSTM

  - 두개의 LSTM 아키텍쳐가 결합된 구조.



  - Encoder아키텍쳐에서 셀의 마지막 시점의 hidden state를 Decoder로 넘겨줌

  - 받은 hidden state를 Decoder에서 최초의 hidden state로 사용함.

Proposed method

Experimental setting

- Dataset

stocks	N	Train	val	test
KOSPI	5156	3210	979	967
KOSDAQ(2013-03-03 ~)	1898	938	474	486
NASDAQ	5265	3263	1002	1000
S&P 500	5265	3263	1002	1000
다우존스	5265	3263	1002	1000
홍콩 항생	5151	3196	976	979
일본 니케이	5124	3182	971	971
독일 DAX	5306	3300	1005	1001
프랑스 CAC	5343	3313	1015	1015
스페인 IBEX	5314	3283	1017	1014
대만 기권	5144	3206	971	967
네덜란드 AEX	5347	3316	1016	1015
인도 선섹스	5151	3204	970	977
브라질 보베스파	5265	3263	1002	1000
Gold price	5050	3067	994	989
Oil price(2000-08-23 ~)	5060	3076	994	990
10년 만기 미국 국채	3264	3264	1027	688
10년 만기 한국 국채				
비트코인 암호화폐(2014-09-17 ~)	2275	1560	359	356
이더리움 암호화폐(2015-08-07 ~)	1951	1236	359	356

KOSPI INDEX log difference

2000.01.01 – 2020.12.31

Experimental setting

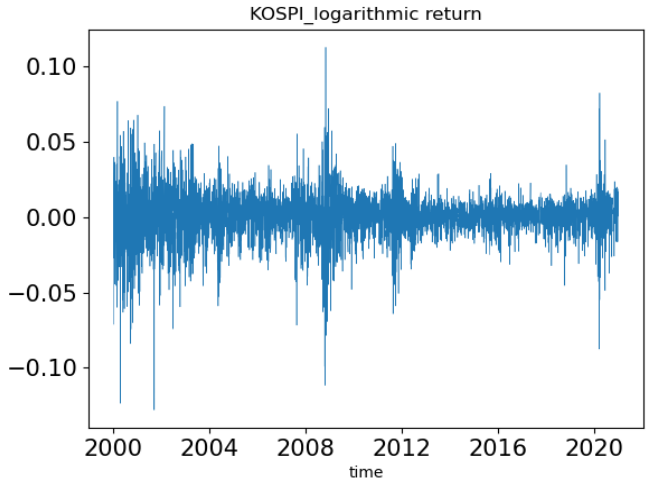
- Statistics

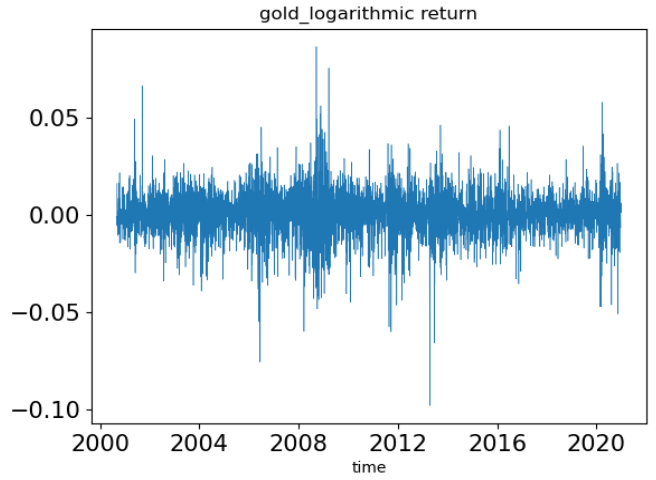
stocks	Mean	Standard deviation	Skewness	Kurtosis(3)	Jerque Bera	ADF
KOSPI	1582.518	587.343	-0.421	1.861	432.61***	-0.438834
KOSDAQ	668.623	103.98	0.46	2.632	78.4663***	-1.173884
NASDAQ	3791.769	2367.556	1.388	4.395	2125.9183***	3.346301
S&P 500	1653.635	674.270	1.033	3.067	941.5057***	1.678640
다우존스	14655.366	5792.134	1.043	2.906	959.6158***	0.870942
홍콩 항생	19919.806	5726.487	-0.18	2.076	211.8634***	-1.579436
일본 니케이	14640.566	4694.21	0.393	2.009	342.692***	-0.491717
독일 DAX	7661.465	3006.491	0.385	1.978	363.2483***	-0.506127
프랑스 CAC	4469.135	918.824	0.153	2.207	161.6942***	-2.233112
스페인 IBEX	9680.838	2017.084	0.667	3.565	466.6305***	-2.262163
대만 기권	7903.82	2055.782	0.243	2.707	69.2831***	-0.264858
네덜란드 AEX	435.655	108.924	0.265	2.222	197.9852***	-1.899507
인도 선섹스	17921.698	11385.273	0.401	2.120	305.4783***	0.679944
브라질 보베스파	3791.769	2367.556	1.388	4.395	2125.9183***	3.346301
Gold price	997.637	479.258	-0.105	1.796	315.6375***	-0.459757
Oil price	61.785	25.912	0.400	2.381	216.5233***	-1.997886
10년 만기 미국 국채	118.257	9.575	-0.285	2.145	219.8195***	-2.216594
10년 만기 한국 국채						
비트코인 암호화폐	4925.441	4830.747	1.056	4.332	595.9625***	1.440318
이더리움 암호화폐	223.111	231.702	1.601	6.123	1642.5896***	-2.082944
KOSPI INDEX log difference	0.000193	0.0149	-0.572	10.03	10943.0983***	-33.339841***

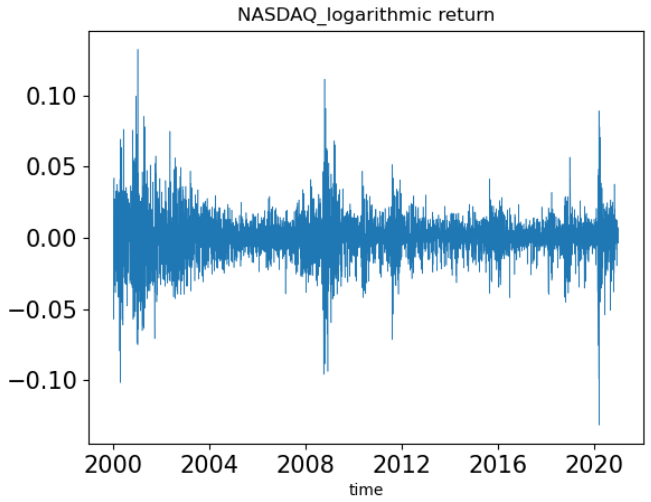
Used Close(종가) variable.

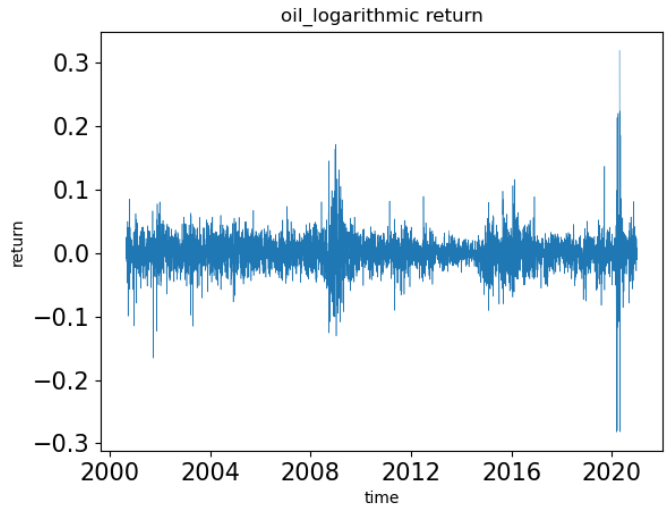
Experimental setting

- logarithmic return : ln((t+1)/t) t : 시점이 t일때 종가(close)









- logarithmic return

  - 수익률을 나타내는 지표 중의 하나.

  - 일반 수익률보다 실제 수익률에 더 가까운 지표.

Experimental setting

- Stock Dataset
  - data normalization
    - $\log(x_i + 1) - \log(x_7 + 1), i = 1, .., 8$
    - $+1$ prevents the denominator to be zero
    - Value added by 1 approximates to value without 1 added

- Metric
  - MSE($e^{predicted\ value}, e^{true\ value}$) or MAE($e^{predicted\ value}, e^{true\ value}$)

- Train, validataion
  - Used MSE Loss as loss function

- Test.
  - Ran model final evaluation by metric

Hyperparameters	Number
hidden state, cell state dimension	10
input sequence length	7
output sequence length	1
epoch	20
Batch size	128

Experimental setting

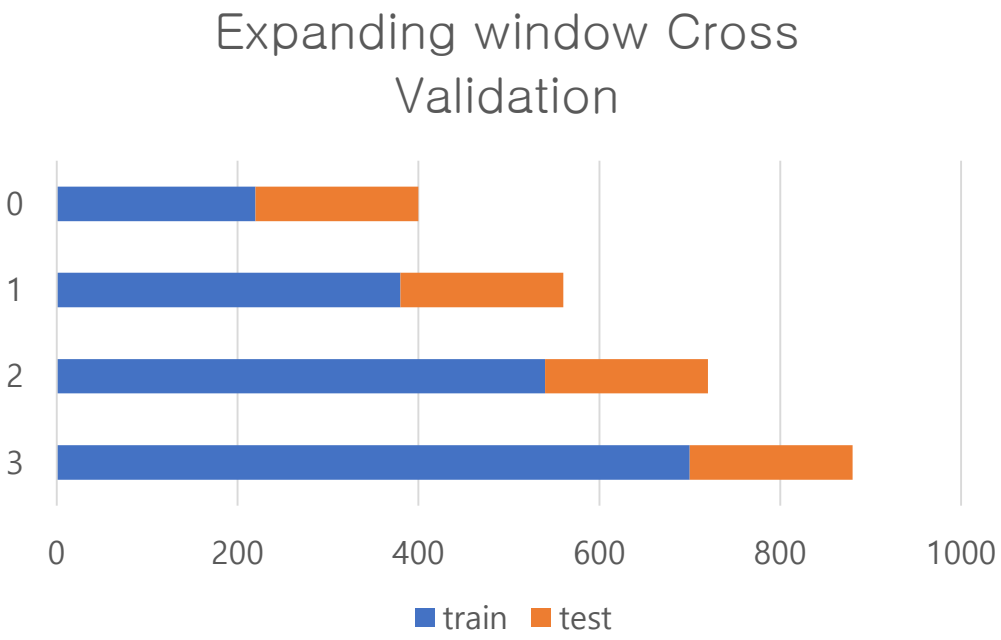
- To determine the detailed performance of a model
  - Add another model evaluation metric.
    - Mean Absolute Percentage Error (MAPE) = $\frac{\sum_{t=1}^{n}\left|\frac{y_t-\hat{y}_t}{y_t}\right|}{n} \cdot 100(\%)$
      - MAPE is undefined for data points where the value is zero.
      - MAPE is biased towards prediction that are systemically less than the actual values themselves

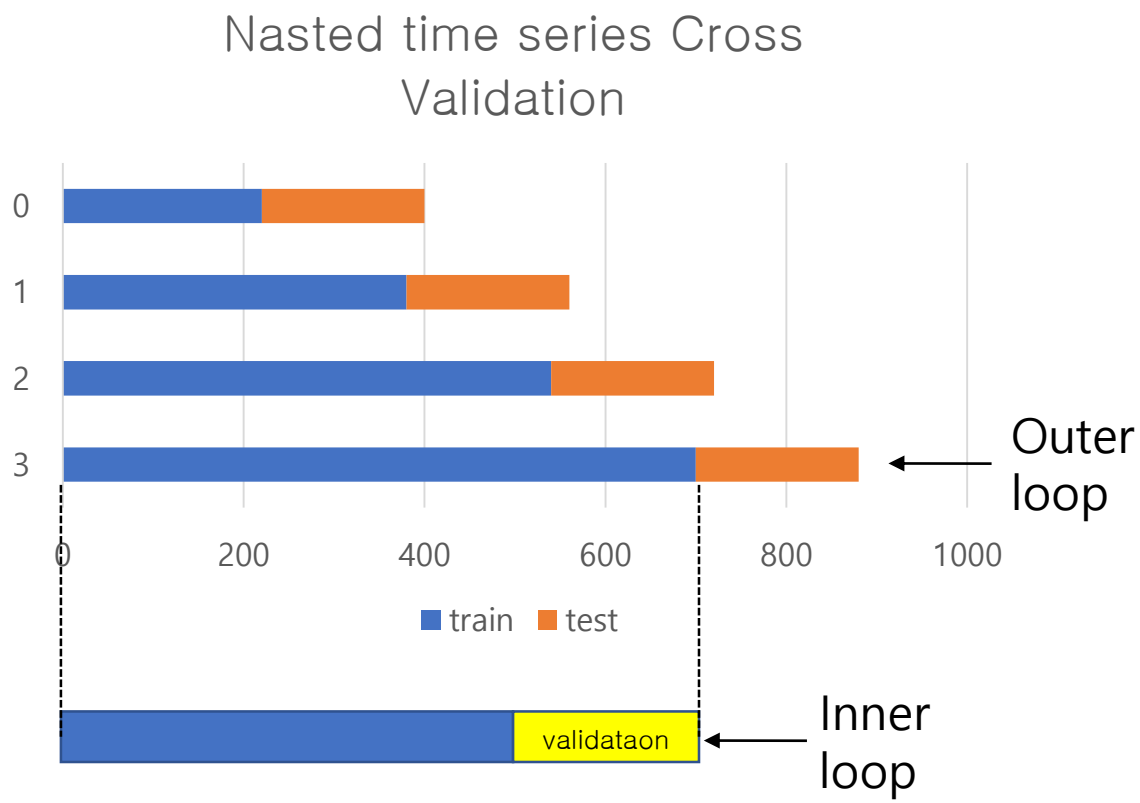
$$y = 20, \hat{y} = 10, n = 1 \rightarrow MAPE = 50\%$$
$$y = 10, \hat{y} = 20, n = 1 \rightarrow MAPE = 100\%$$

  - Compare three metrics.
    - MAE
    - MSE
    - MAPE

Experimental setting

- Stock dataset split in timeseries data





Experimental setting

- Stock dataset split in timeseries data
  - In many papers

"The first eight years of the data as a training set, the next two years of the data as a test set."
- Qiang Jiang, Chenglin Tang, Chen Chen, Xin Wang and Qing Huang(2019), Stock Price Forecast Based on LSTM Neural Network. Proceedings of the Twelfth International Conference on Management Science and Engineering Management

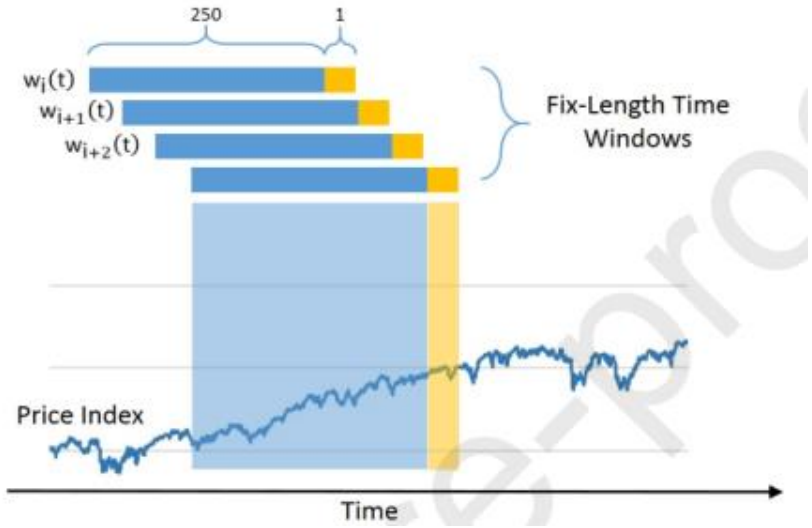
"In both neural network models, 20% of the training set was used for the validation set, and the realized volatility was set as the target value."
- Ha Young Kim, Chang Hyun Won(2018), Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models, Expert Systems With Applications

"training and test datasets where 70% of each dataset was used for training and the remaining 30% of each dataset was used for testing the accuracy of models." - Sima Siami-Namini, Neda Tavakoli, Akbar Siami Namin(2018), A Comparison of ARIMA and LSTM in Forecasting Time Series. IEEE International Conference on Machine Learning and Applications

"Each sample data of agricultural commodity futures prices is split into a training set (60%), validation set (20%), test set (20%) in chronological order." - HUI LI, YUNPENG CUI, SHUO WANG, (Member, IEEE), JUAN LIU, JINYUAN QIN, AND YILIN YANG (2020), Multivariate Financial Time-Series Prediction With Certified Robustness. IEEE Access

Experimental setting

- Stock dataset split in timeseries data



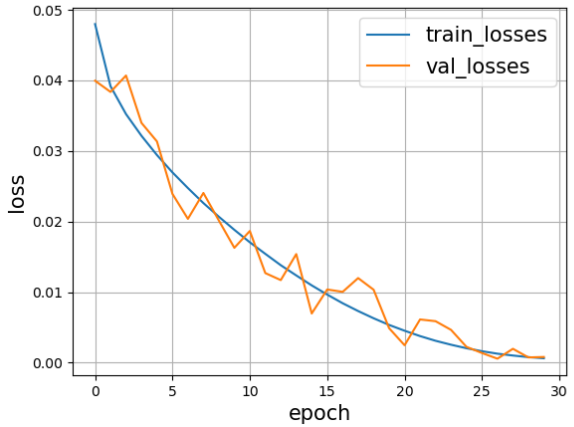
"we select 150 data for testing (out-of-sample data), and split the remaining data into training set (80%) and validation set (20%)." - Hadi Rezaei, Hamidreza Faaljou, Gholamreza Mansourfar (2020), Stock Price Prediction using Deep Learning and Frequency Decomposition. Expert Systems with Applications

- Conclusion

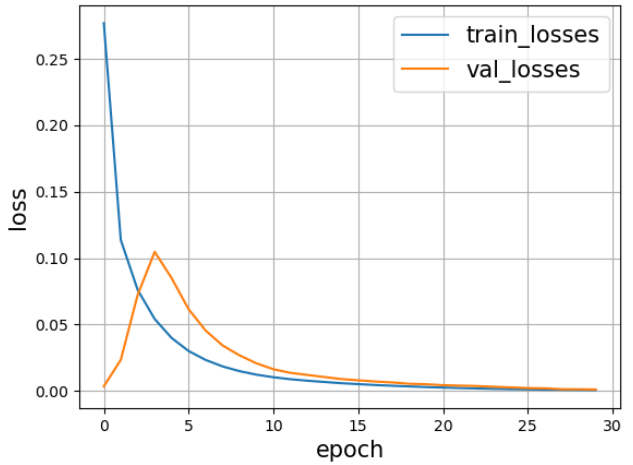
We investigated 14 time series prediction papers between 2019 and 2021, and they adopted chronological data split method.

Experimental setting

- Some issue
  - learning curve
    - Each is a learning curve of the same data, different iterations of the same model.
    - In most of learning curve, the scale that unexpected shape of the learning curve was large.



Expected shape

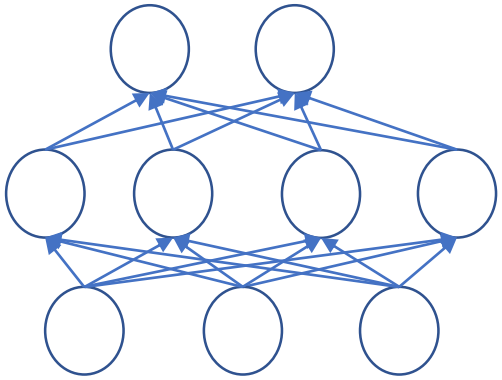


Unexpected shape

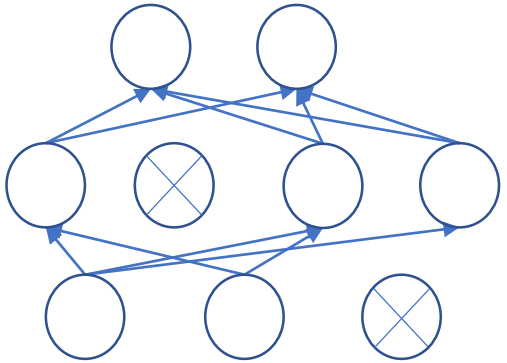
- Expected cause
  - input sequence length is short

Experimental setting

- Model training speed
  - dropout
    - dropout can reduce model accuracy



MLP with one hidden layer



Hidden layer after dropout

- A Model for Applying the Stock Market
  - Momentum effect in stock market.
    - Asset prices follow a trend for a rather long time.
  - Attention LSTM may reflect a tendency
    - Attention mechanism may serve as a weight to reflect the trend.

Comparison results

- Mean Absolute Error (mean of 30 iteration)

Data	Proposed	Standard Deviation	RNN	Standard Deviation	LSTM	Standard Deviation	GRU	Standard Deviation	Enc-Dec LSTM	Standard Deviation
KOSPI(^KS11)			0.021195	0.019049	0.020116	0.025005	0.018809	0.020418		
KOSDAQ(^KQ11)			0.102302	0.071357	0.078273	0.078779	0.108261	0.080757		
NASDAQ(^IXIC)			0.016808	0.012544	0.025125	0.027928	0.020856	0.019298		
S&P 500(^GSPC)			0.020992	0.025467	0.016364	0.016857	0.021337	0.020431		
다우존스(^DJI)			0.022184	0.022629	0.013185	0.013078	0.016047	0.015028		
홍콩 항생(^HSI)			0.022545	0.020101	0.013334	0.009264	0.026282	0.031992		
일본 니케이(^N225)			0.016286	0.015128	0.018038	0.015499	0.021768	0.023830		
독일 DAX(^GDAXI)			0.017664	0.017576	0.013985	0.011829	0.021262	0.019056		
프랑스 CAC(^FCHI)			0.022548	0.021522	0.016373	0.015951	0.031835	0.035117		
스페인 IBEX(^IBEX)			0.016077	0.023092	0.019598	0.020462	0.018918	0.019273		
대만 기권(^TWII)			0.016150	0.018184	0.014470	0.017367	0.019759	0.023773		
네덜란드 AEX(^AEX)			0.023351	0.023820	0.019398	0.021147	0.023697	0.028435		
인도 선섹스(^BSESN)			0.017187	0.021773	0.011863	0.011989	0.018064	0.022021		
브라질 보베스파(^BVSP)			0.034658	0.029081	0.023900	0.028509	0.022915	0.013754		
Gold price(GC=F)			0.025382	0.029737	0.020862	0.024010	0.026176	0.028744		
Oil price(음수 존재)										
10년 만기 미국 국채										
10년 만기 한국 국채										
비트코인 암호화폐(BTC-USD)			0.077903	0.052344	0.090269	0.068051	0.053026	0.043577		
이더리움 암호화폐(ETH-USD)			0.092046	0.049431	0.079820	0.044870	0.091226	0.059508		

Comparison results

- Mean Squared Error (mean of 30 iteration)

Data	Proposed	Standard Deviation	RNN	Standard Deviation	LSTM	Standard Deviation	GRU	Standard Deviation	Enc-Dec LSTM	Standard Deviation
KOSPI(^KS11)			0.001275	0.002006						
KOSDAQ(^KQ11)			0.010130	0.015369						
NASDAQ(^IXIC)			0.000987	0.001767						
S&P 500(^GSPC)			0.001236	0.001681						
다우존스(^DJI)			0.001389	0.002728						
홍콩 항생(^HSI)			0.001570	0.003137						
일본 니케이(^N225)			0.001250	0.002907						
독일 DAX(^GDAXI)			0.000778	0.001146						
프랑스 CAC(^FCHI)			0.000654	0.000965						
스페인 IBEX(^IBEX)			0.001022	0.001471						
대만 기권(^TWII)			0.000705	0.001687						
네덜란드 AEX(^AEX)			0.000935	0.001855						
인도 선섹스(^BSESN)										
브라질 보베스파(^BVSP)										
Gold price(GC=F)										
Oil price(음수 존재)										
10년 만기 미국 국채										
10년 만기 한국 국채										
비트코인 암호화폐(BTC-USD)										
이더리움 암호화폐(ETH-USD)										

Comparison results

- Mean Absolute Percentage Error (mean of 30 iteration)

Data	Proposed	Standard Deviation	RNN	Standard Deviation	LSTM	Standard Deviation	GRU	Standard Deviation	Enc-Dec LSTM	Standard Deviation
KOSPI(^KS11)										
KOSDAQ(^KQ11)										
NASDAQ(^IXIC)										
S&P 500(^GSPC)										
다우존스(^DJI)										
홍콩 항생(^HSI)										
일본 니케이(^N225)										
독일 DAX(^GDAXI)										
프랑스 CAC(^FCHI)										
스페인 IBEX(^IBEX)										
대만 기권(^TWII)										
네덜란드 AEX(^AEX)										
인도 선섹스(^BSESN)										
브라질 보베스파(^BVSP)										
Gold price(GC=F)										
Oil price(음수 존재)										
10년 만기 미국 국채										
10년 만기 한국 국채										
비트코인 암호화폐(BTC-USD)										
이더리움 암호화폐(ETH-USD)										

Reference

1 Ha Young Kim, & Chang Hyun Won (2018). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. Expert Systems With Applications, Vol.103. 25-37. https://doi.org/10.1016/j.eswa.2018.03.002

2 Qiang Jiang, Chenglin Tang, Chen Chen, Xin Wang and Qing Huang(2019), Stock Price Forecast Based on LSTM Neural Network. Proceedings of the Twelfth International Conference on Management Science and Engineering Management. Vol. 32. 393-408. https://doi.org/10.1007/978-3-319-93351-1_32

3 Sima Siami-Namini, Neda Tavakoli, Akbar Siami Namin(2018), A Comparison of ARIMA and LSTM in Forecasting Time Series. IEEE International Conference on Machine Learning and Applications, Vol.17. 10.1109/ICMLA.2018.00227

4 Hui li, Yunpeng cui, Shuo wang, (Member, IEEE), Juan liu, Jinyuan qin, And Yilin Yang (2020), Multivariate Financial Time-Series Prediction With Certified Robustness. IEEE Access, Vol. 8. 109133 - 109143. 10.1109/ACCESS.2020.3001287

5 Hadi Rezaei, Hamidreza Faaljou, Gholamreza Mansourfar (2020), Stock Price Prediction using Deep Learning and Frequency Decomposition. Expert Systems with Applications, Vol. 169. 114332. https://doi.org/10.1016/j.eswa.2020.114332

Appendix

Appendix

- Jarque-bera test
  - 왜도와 첨도가 정규분포와 일치하는지 판단하는 적합도 검정
- ADF test
  - 시계열 데이터가 정상성(stationary)을 만족하는지에 대한 검정
- LM-ARCH 검정
  - ARCH, GARCH와 같은 모형을 적용할 수 있는지에 대한 검정.