세미나 발표

목차

1. LSTM 학습의 문제점

2. 해당 문제점 관련한 저널 2가지

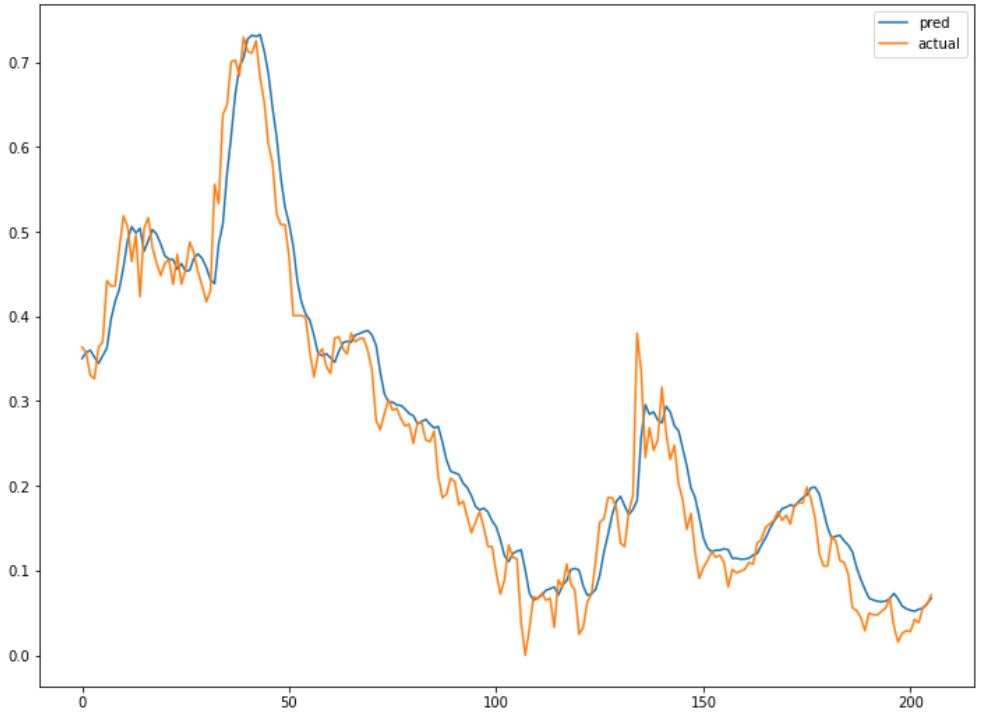
3. LSTM 시계열 예측모델의 논문 소개

4. 논문 방향성 결정

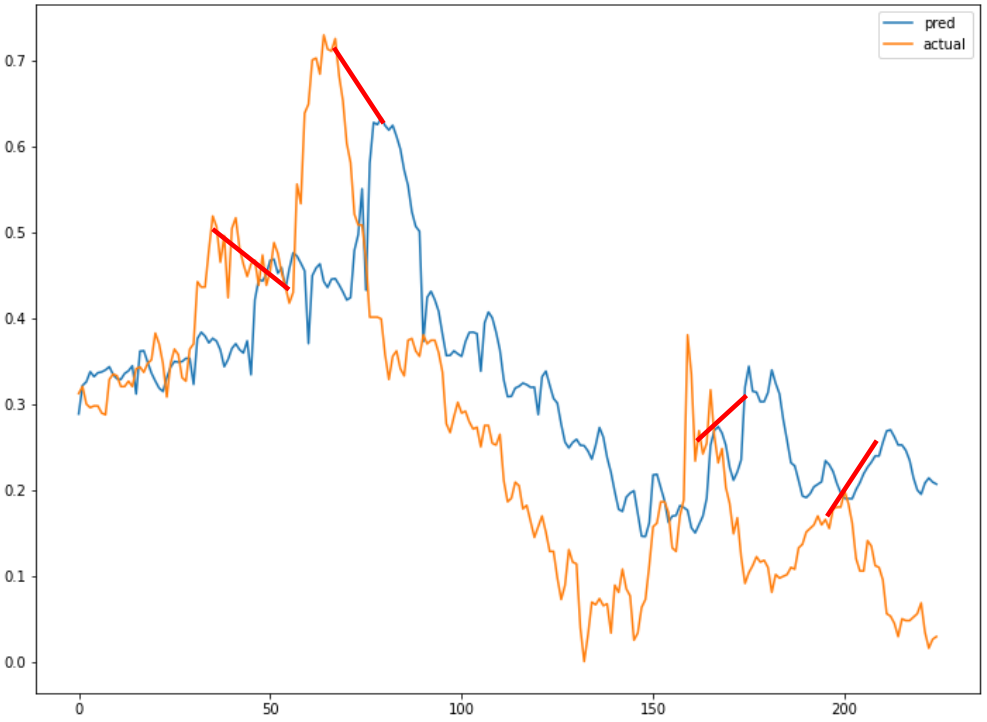
LSTM 학습의 문제점

시계열 데이터 실습

학습의 문제점



many to one



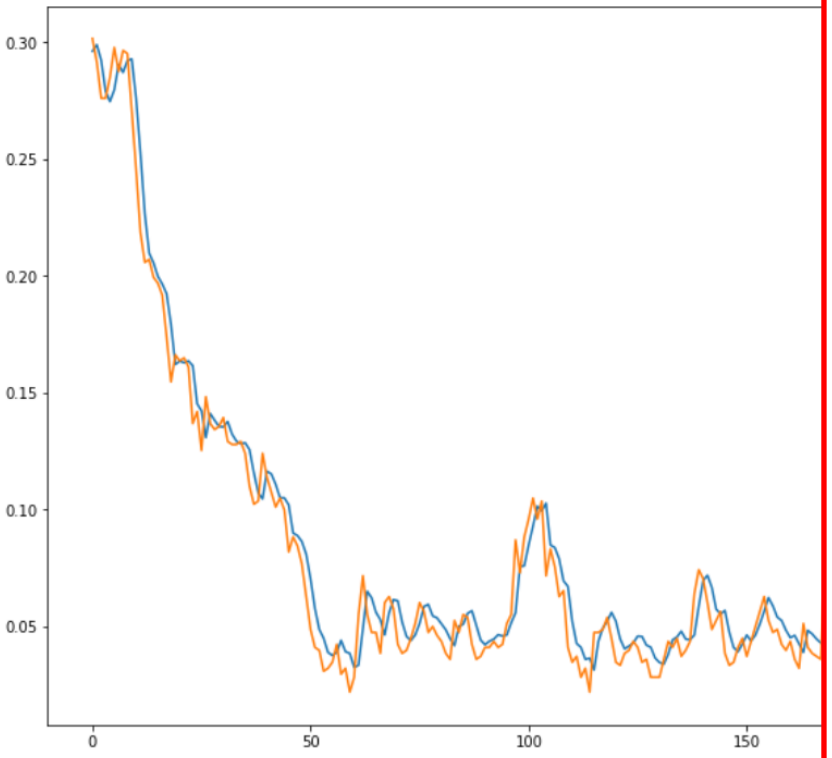
many to many

Many to one??

Many to (될까?

Many to one??

Many to one 에서는 학습이 잘 될까?



Pred를 1-time lagging 시킨 후 결과 값은 일치하는 것을 볼 수 있습니다.

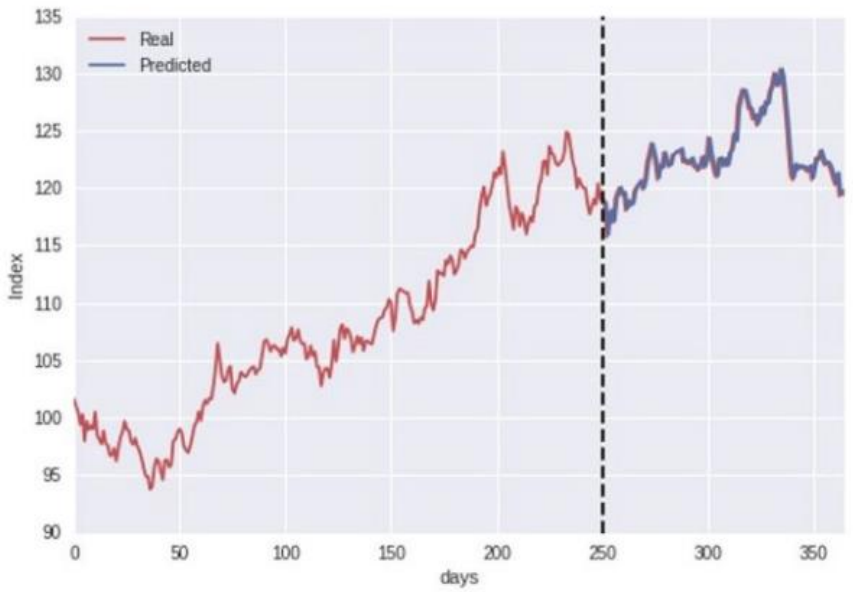
해당 문제점 관련한 저널 2가지

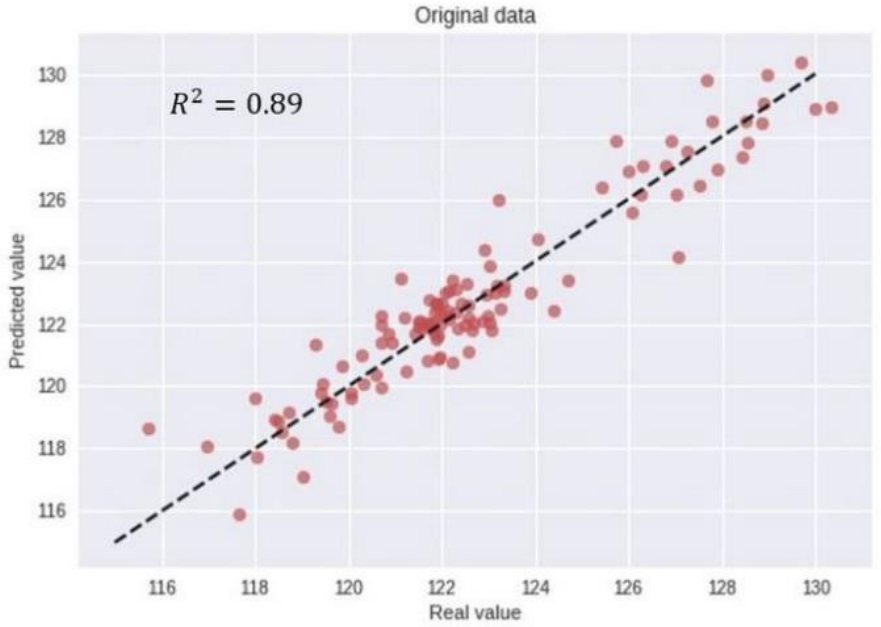
비슷한 경향 조사

How (not) to use Machine Learning for time series forecasting: Avoiding the pitfalls

towards data science

비슷한 경향 조사





주가 데이터의 시계열 예측을 잘하는 것처럼 보이며,
R^2 값을 통계적 수치로 확인할 경우 0.89로 신뢰적인 수치이고,
평균 오류율 및 예측 정확도는 높은 것을 확인 가능합니다.

비슷한 경향 조사

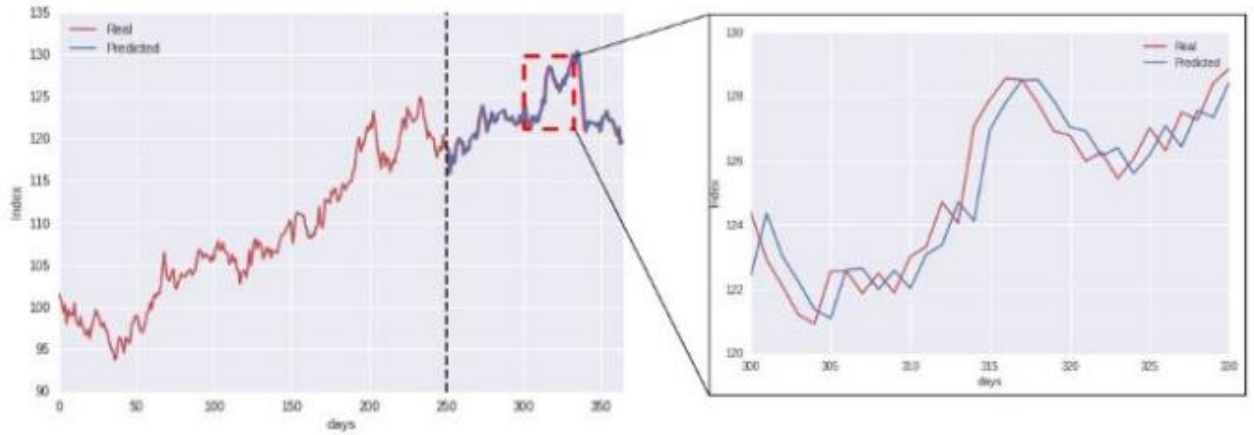
방금 본 주가 데이터 모델은 모델 성능을 평가할 때
정확도 평가지표를 선택하는 것이 얼마나 오해의 소지가 있는지를 보여주는 예

저자의 주장 : 주가 데이터(랜덤 프로세스)는 완전히 확률적인 과정이다.
이러한 이유로, 행동을 배우고 미래의 결과를 예측하기 위해
과거 데이터를 훈련 세트로 사용하는 아이디어는 불가능합니다.

이런 상황에서 모델이 그렇게 정확한 예측을 제공하는 것처럼 보이는 이유는 무엇일까요?

비슷한 경향 조사

이런 상황에서 모델이 그렇게 정확한 예측을 제공하는 것처럼 보이는 이유는 무엇일까요?

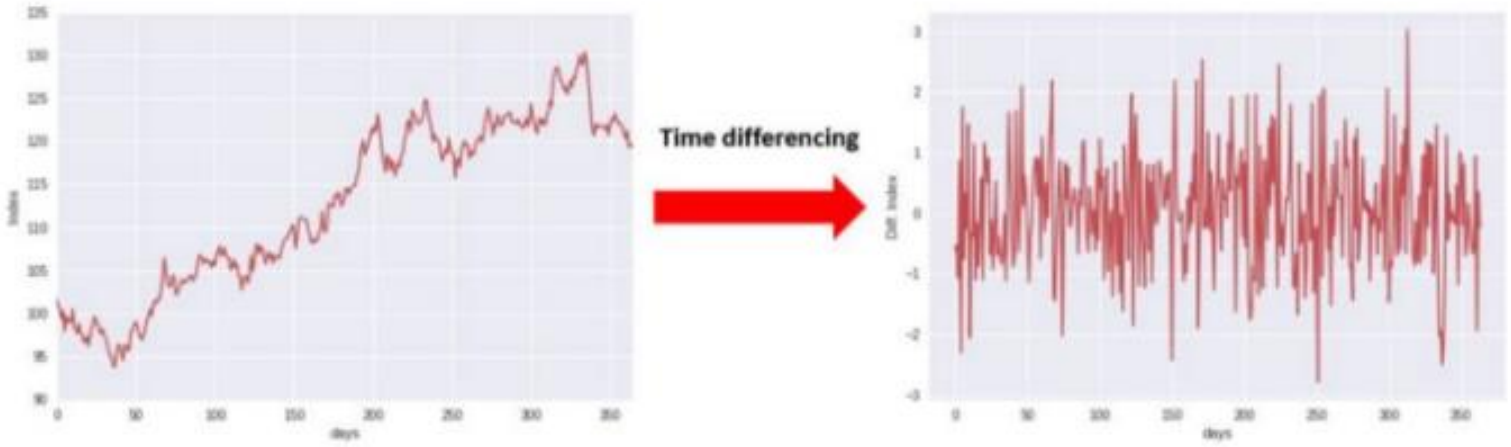


시계열 데이터는 시간 상관 관계가 있는 경향이 있으며 유의한 자기 상관 관계를 나타냅니다.
"t+1"의 값이 시간 "t"의 지수와 굉장히 비슷하다.

즉 오른쪽 위의 그림에 표시된 것처럼, 모델은 시간 " t+1 " 의 값을 예측할 때 단순히
시간 " t " 의 값을 예측(종종 지속성 모델이라고도 함)으로 사용합니다.

이는 따라서 모델이 단순히 미래에 대한 예측으로 이전 값을 사용한다는 것을 나타냅니다.

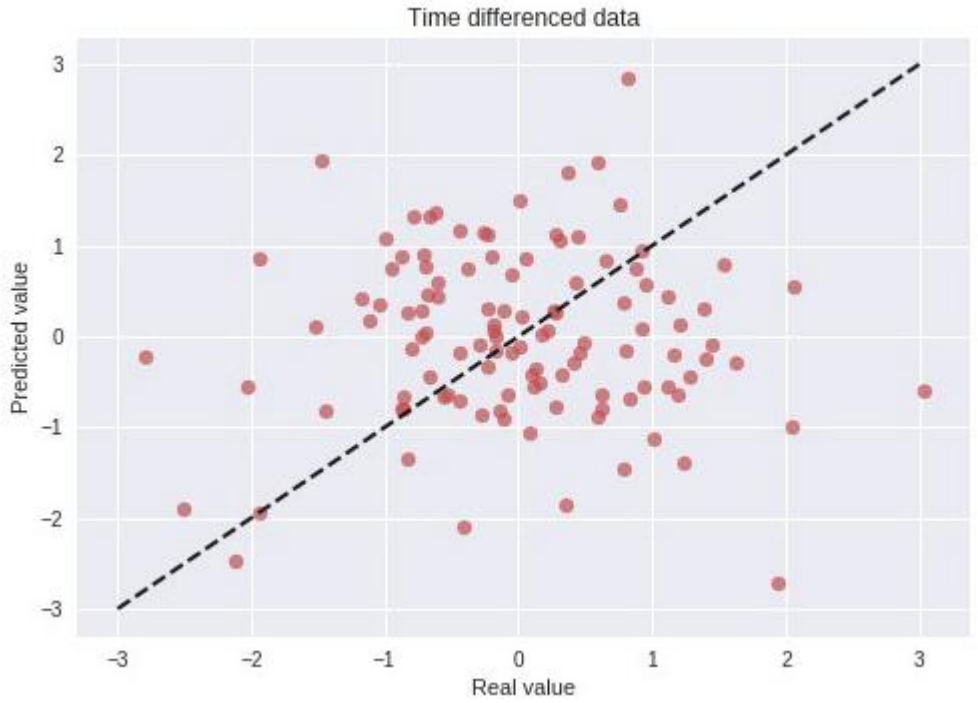
비슷한 경향 조사



저널 추천 : 값 자체보다는 시간 단계 간 값의 차이를 예측하도록 모형을 정의하는 것이 모형의 예측 검정력에 대한 훨씬 강력한 검정일 것입니다.

시간단계 값의 차이를 사용할 경우 :
데이터의 자기 상관 관계가 강하다고 단순하게 사용할 수 없으며
결국 시간 "t"의 값을 "t+1"에 대한 예측으로 사용할 수 있습니다.

비슷한 경향 조사



이 테스트의 결과는 시간 단계별 차이를 고려하여
실제 값과 예측 값의 산점도를 보여 주는 그림입니다.
-> 완벽하게 예측하지 못함

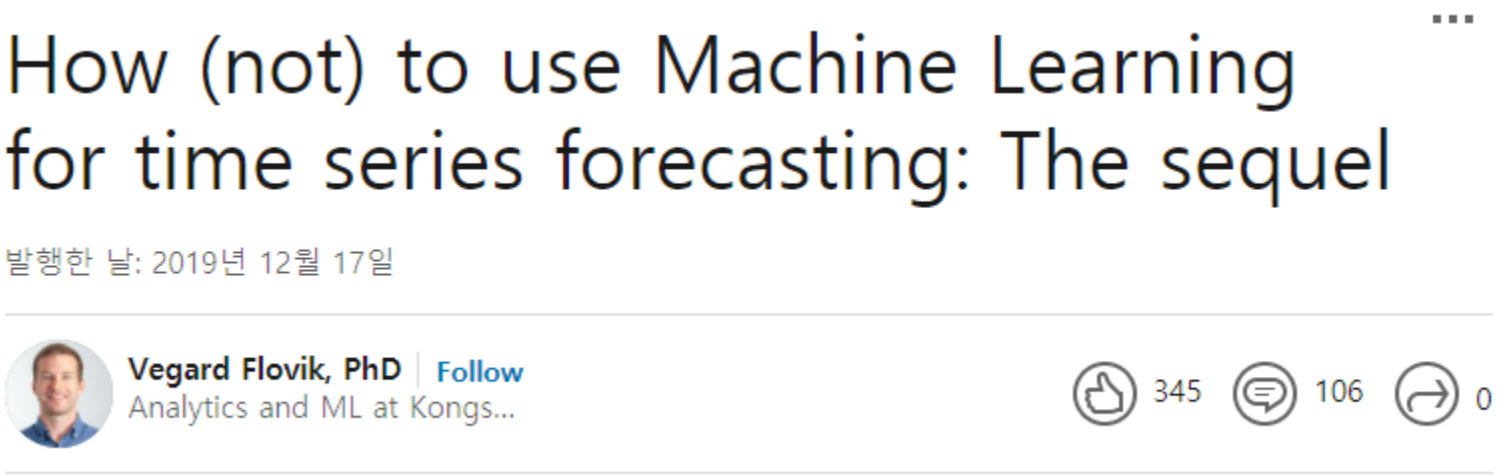
비슷한 경향 조사

결론 : 예측 정확도 측면에서 모델 성능을 평가할 때는 회의적이어야 합니다. 주가 데이터에서 보이듯,
미래의 결과를 예측하는 것이 정의상 불가능한 과정의 경우에도, 결과에 속을 수 있습니다.
실제로는 모델에 예측력이 전혀 없을 수 있습니다.

베스트 코멘트 :
LSTM 모델이 주가 값을 예측하기 위한 행동은 완벽하게 수행 중입니다.

즉 LSTM 모델이 t시점의 주가를 t+1시점이 따라간다는 것은 학습이 잘 되었다.
문제는 암묵적인 목표가 가격을 예측하는 것이 아니라 수익을 예측하는 것이라는 사실에서 비롯되는 것입니다.

비슷한 경향 조사



How (not) to use Machine Learning for time series forecasting: The sequel

발행한 날: 2019년 12월 17일

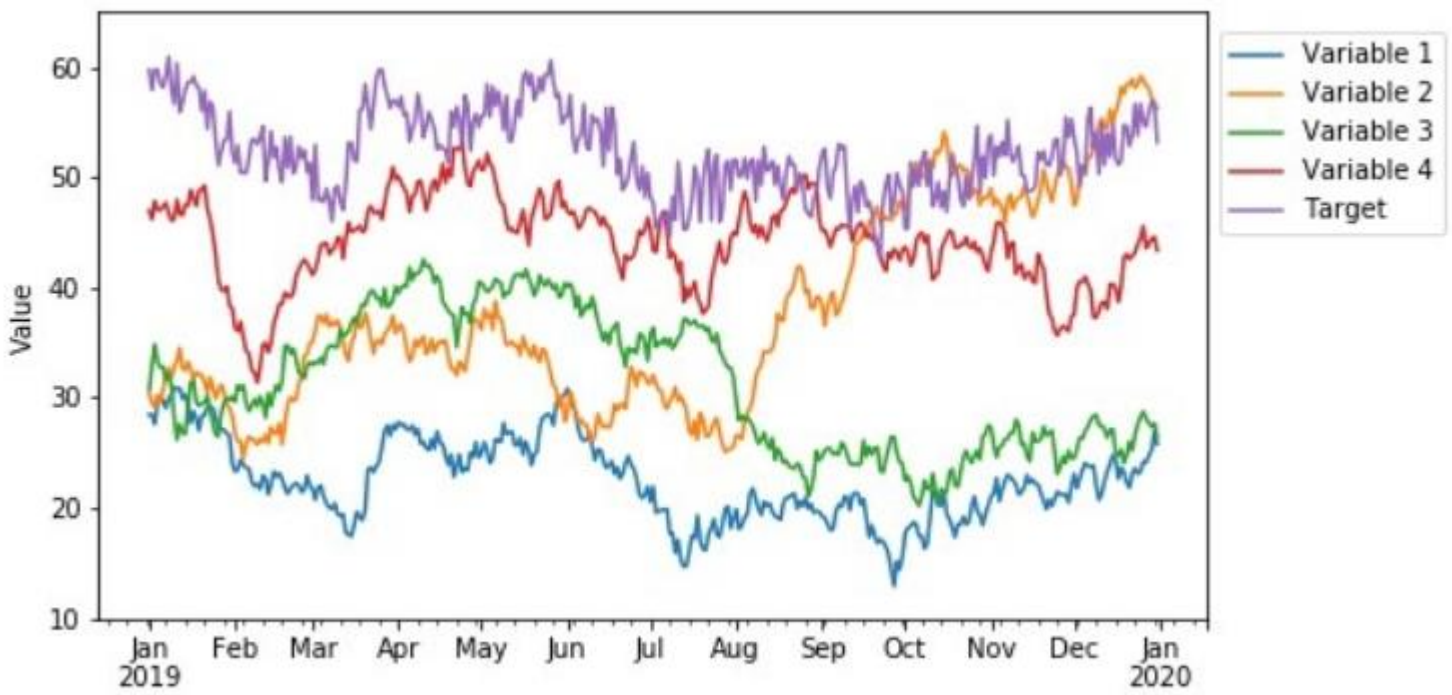
Vegard Flovik, PhD | Follow
Analytics and ML at Kongs...

345 106 0

This post will also go through the task of time series forecasting using machine learning, and how to avoid some of the common pitfalls.

비슷한 경향 조사

주장 : 이전 주가 데이터가 다음 주가를 예측하는 데에 실제로 영향을 미치지 않는다는 것입니다.

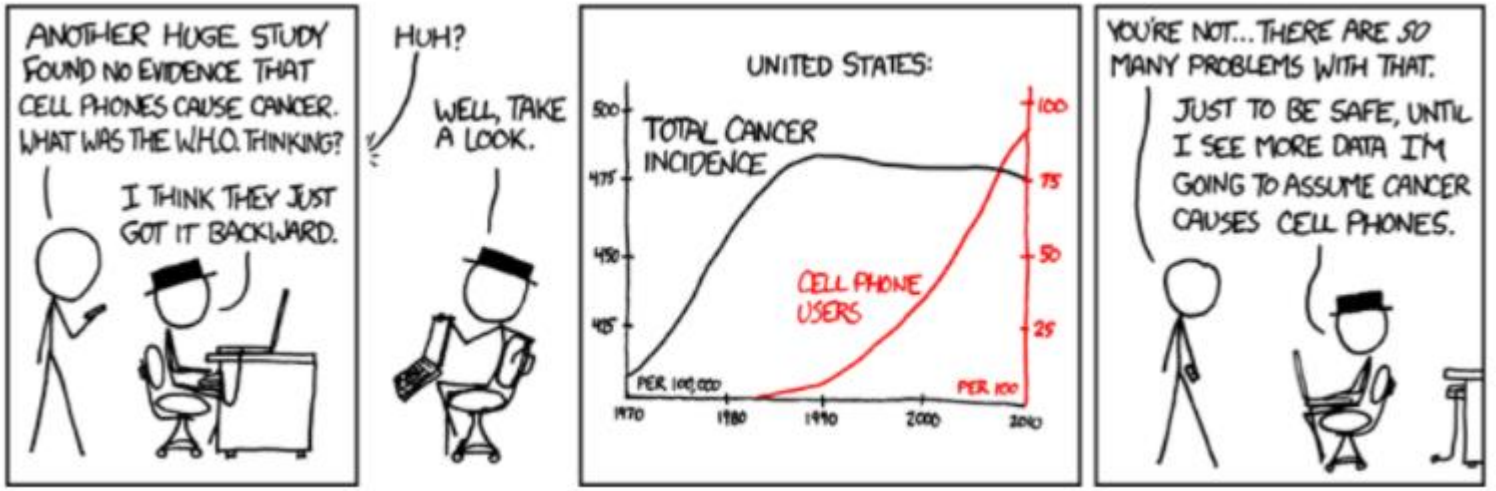


1. 연간 주가지수의 변화
2. 제품의 판매
3. 수요, 일부 센서 데이터
4. 장비 상태

가정 : 모델에 대한 입력 변수가 표적 변수를 예측할 수 있는 유용한 정보를 포함하고 있을 수 있다.
(정보가 포함될 수도 있고, 아닐 수도 있다.)

비슷한 경향 조사

주요한 쟁점 : 데이터 간의 상관관계와 인과관계를 이해해야 한다.



상관관계 및 의존성은 인과관계를 불문하고 모두 통계적 관계입니다.
상관 관계는 실제로 활용할 수 있는 예측 관계를 나타낼 수 있습니다.
ex) 날씨가 매우 더워지거나 추워질 때, 전력의 관계
하지만 일반적으로 상관관계가 존재한다고 해서 인과관계의 존재를 추론하기에 충분하지 않습니다.
(즉, 상관관계가 인과관계를 의미하지는 않는 것을 알아야 합니다.)

비슷한 경향 조사

변수 간의 상관 계수를 나타내는 상관 행렬을 구한다면 **시각화**가 가능합니다.

where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

참고 : 코시-슈바르츠 부등식

비슷한 경향 조사

Train 과정

모델은 예측하려는 프로세스의 **Target** 데이터가 무엇이어야 하는지 (희망적) 교육받습니다

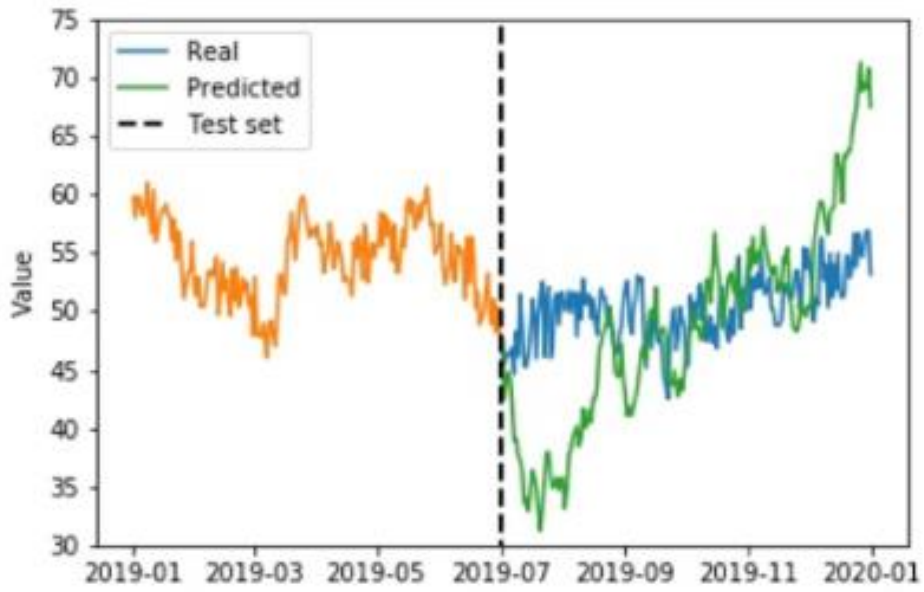
모델은 입력 변수와 대상 간의 특성 패턴/상관을 활용하여 새로운 예측을 제공할 수 있는 관계를 설정합니다.

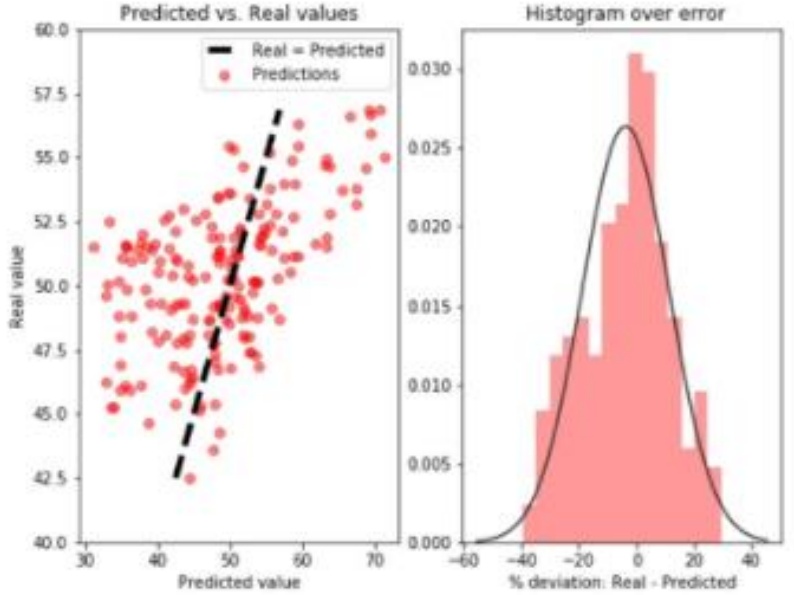
따라서 상관관계 자체가 통계상의 허점일 뿐이며, 상관관계 사이에 인과관계가 없다는 것일 수 있다.

비슷한 경향 조사

6개월의 데이터를 Train

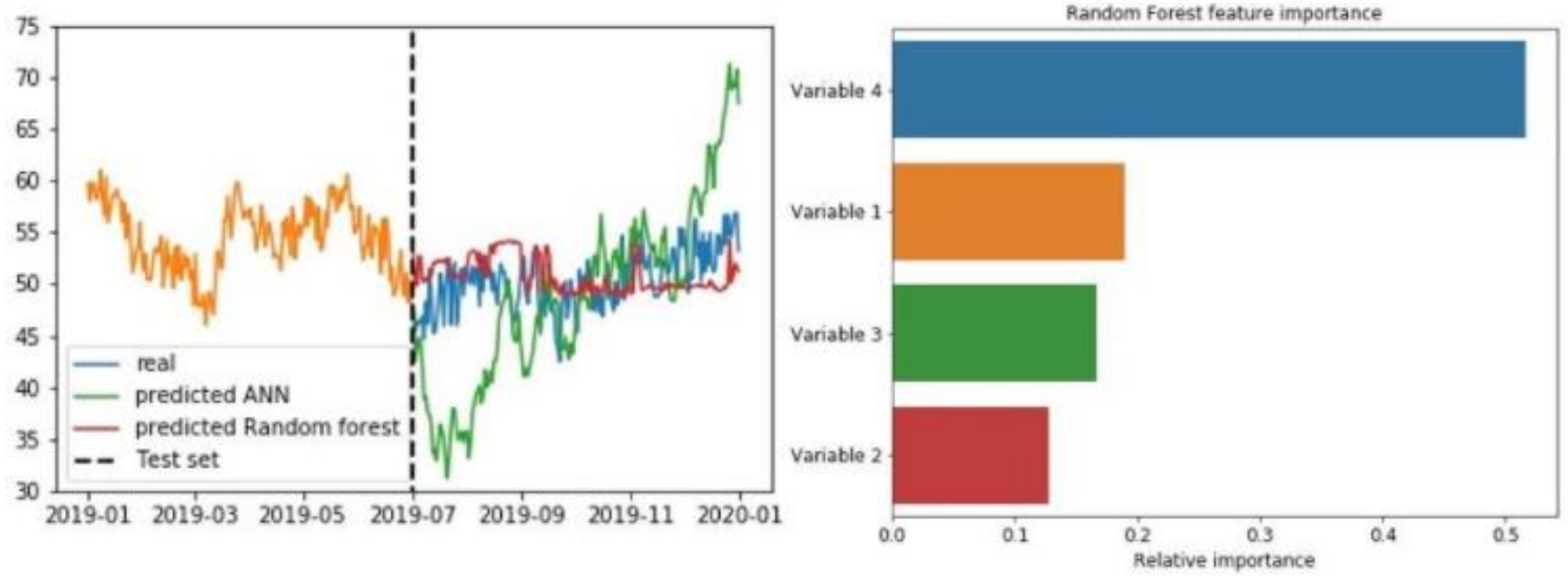
10%는 Validation 데이터





실제 Test data로 예측시 잘못된 예측이 다수 발생한다.

비슷한 경향 조사

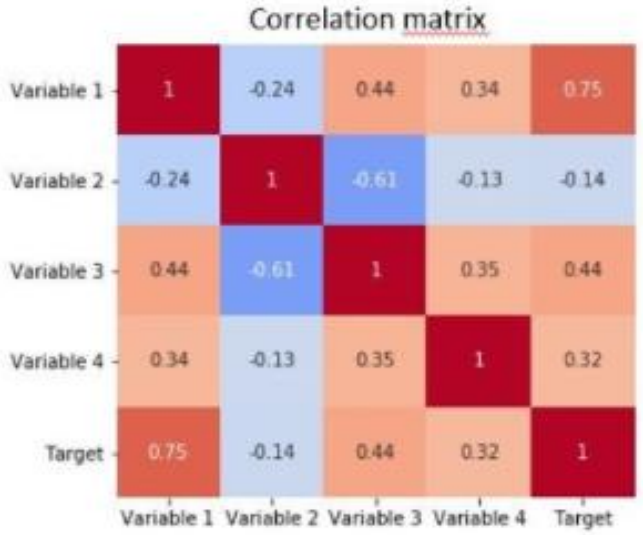


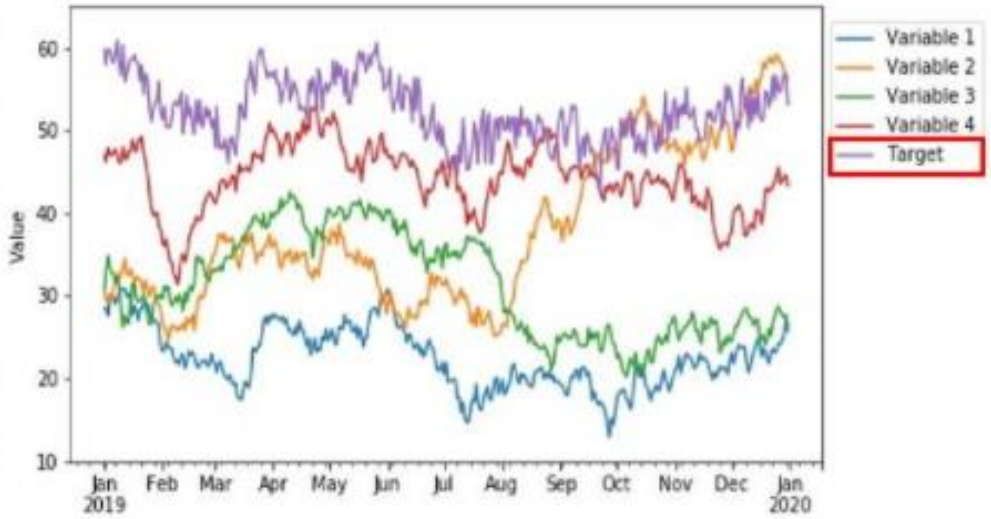
Random forest model과 LSTM model을 비교하여 Test

저자가 Random forest를 사용한 이유 :

1. 간단한 모델이 더 좋은 성능을 내는 경우들이 다수 존재하기 때문
2. 입력데이터와 타겟데이터의 상관성을 나타낼 수 있기 때문

비슷한 경향 조사





상관계수로는 Target는 Variable1과 가장 연관되어 있고,
Random forest에서는 Variable4가 가장 연관되어 있다.
(Variable 1는 2번째로 연관되어 있음)

따라서 실제 Target data는 Variable 1의 추세를 따른다.

비슷한 경향 조사

입력 변수가 타겟 변수와 상관관계가 있다고 해서 인과관계가 있다는 것이 아님.

⇩

상관 관계만으로 딥러닝 모델을 구축할 시 실제로 예측력이 전혀 없을 수 있음.

따라서 시계열 데이터가 인과관계가 존재하는지 검증하는 과정이 필요합니다.

⇩

그레인저 인과관계 검정

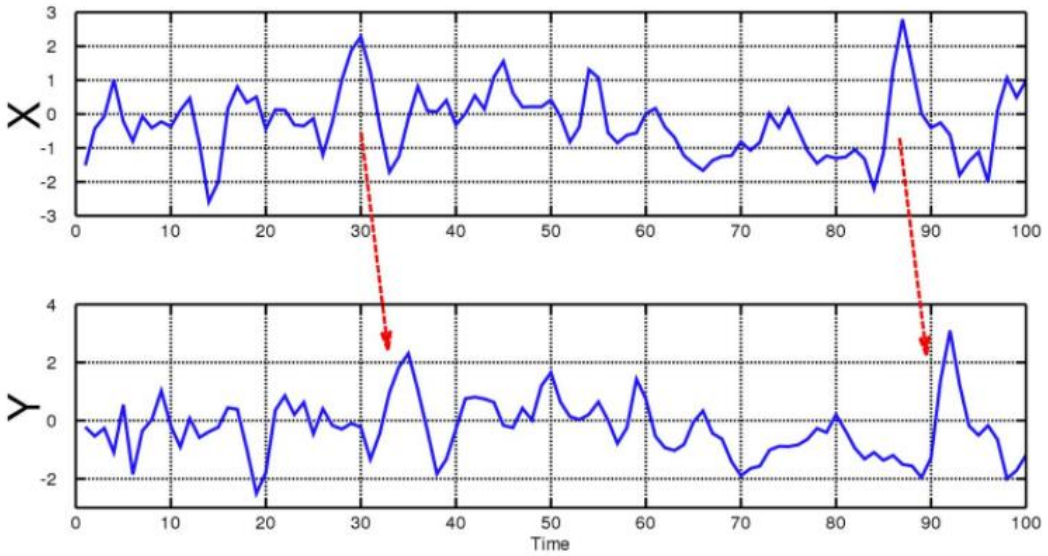
비슷한 경향 조사

그레인저 인과관계 검정이란 ?

입력 시계열 데이터가 타겟 시계열 데이터을 예측하는 데 유용한지 여부를 결정하기 위한 통계적 가설 검정

조건

1. 원인(입력)은 결과(타겟)보다 먼저 발생합니다
2. 원인(입력)에는 미래 가치에 대한 고유한 정보가 있어야 합니다.



비슷한 경향 조사

두가지 저널에 대한 저자의 결론 :

1. 주가 값 자체가 아닌 시간 단계 값의 차이를 사용하여 예측해야 함

2. 데이터가 제공하는 것에 대해 항상 회의적인 시선이 중요함

3. 비판적인 질문들을 하고 어떠한 성급한 결론도 내리지 말아야 함.

논문 소개

논문 소개

Prediction of Stock Price Based on LSTM Neural Network

Dou Wei
College of Management and Ecomomics
Tianjin University
douweidw1113@163.com

2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)

I. INTRODUCTION

개인 투자자 및 분석가는 개인적인 경험과 직관에 의존하여 투자를 결정하는 " 비합리적 " 경향이 있습니다.
비합리적인 방법으로 투자하는 행동은 더 큰 위험을 초래하고 투자자에게 경제적 손실이 발생할 수 있습니다.
따라서 주가를 정확하게 분석, 판단 및 예측하는 방법은 매우 중요합니다.
이를 딥러닝 LSTM으로 모델을 구현하여 시계열 예측을 실시합니다.

논문 소개

입력데이터 : 거래 일자, 시가, 종가, 최저가, 최고가, 주식의 일일 거래량 (6가지)

데이터 전처리 - 정규화

```
import tensorflow as tf
import numpy as np

x = np.array([[1,2,3], [2,3,4], [3,4,5], [4,5,6], [5,6,7],
              [6,7,8], [7,8,9], [8,9,10], [9,10,11], [10,11,12],
              [20,30,40],[21,22,23], [22,23,24], [25,26,27], [30,40,50],
              [31,32,33],[32,33,34],[35,36,37], [40,50,60]])
y = np.array([[4,5,6],
              [5,6,7],
              [6,7,8],
              [7,8,9],
              [8,9,10],
              [9,10,11],
              [10,11,12],
              [11,12,13],
              [12,13,14],
              [13,14,15],
              [50,60,70],
              [24,25,26],
              [25,26,27],
              [28,29,30],
              [60,70,80],
              [34,35,36],
              [35,36,37],
              [38,39,40],
              [70,80,90]])

print(x.shape) # (20,3)
print(y.shape) # (20,3)
x = x.reshape((x.shape[0], x.shape[1], 1))
print(x.shape) # (20,3,1)

model = tf.keras.Sequential([
      tf.keras.layers.LSTM(15, activation = 'relu', return_sequences = False, input_shape=(3,1)),
      tf.keras.layers.Dense(100),
      tf.keras.layers.Dense(3)
])

model.compile(optimizer='adam', loss='mse')
model.fit(x, y, epochs=110, batch_size=3)

x_input = np.array([25,35,45]) # predict용
x_input = x_input.reshape((1,3,1))

yhat = model.predict(x_input)
print(yhat)
```

문제점

학습이 된 범위 -> 출력 잘됨

학습이 안된 범위
-> 출력이 이상함

정규화를 해주어서 학습을 시켜야함

-> **MinMax Scaler (0부터 1사이 정규화)**

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

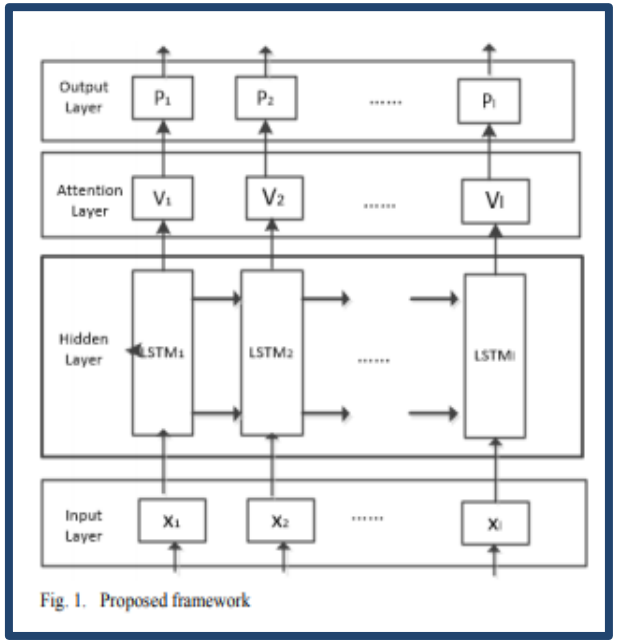


Fig. 1. Proposed framework

Optimizer :
mini-batch gradient descent
(batch size = 64)

Learning Rate : 0.001

논문 소개

RESULTS : Train의 크기(기간)를 다르게 해서, TEST실시

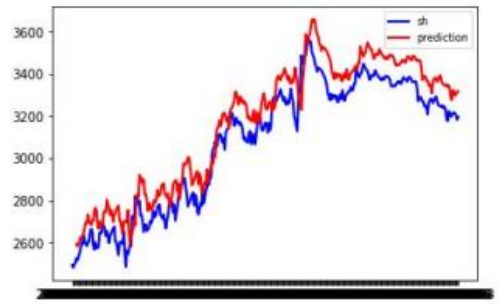


Fig. 3. Shanghai stock index over 18 months

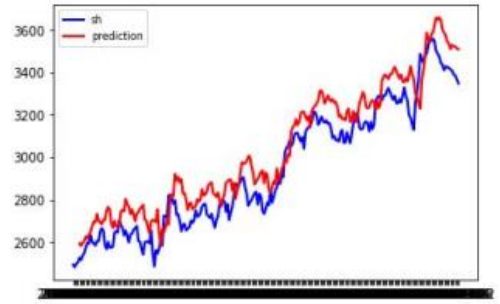


Fig. 2. Shanghai stock index over 1 year

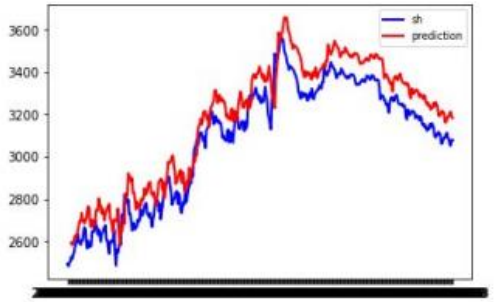


Fig. 4. Shanghai stock index over 2 years

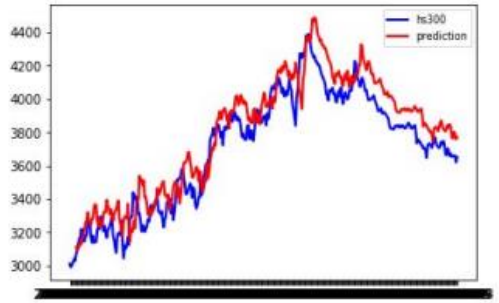


Fig. 9. HS300 index over 18 months

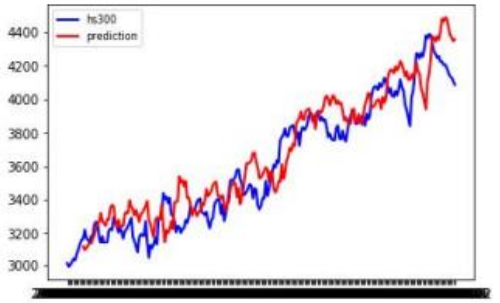


Fig. 8. HS300 index over 1 year

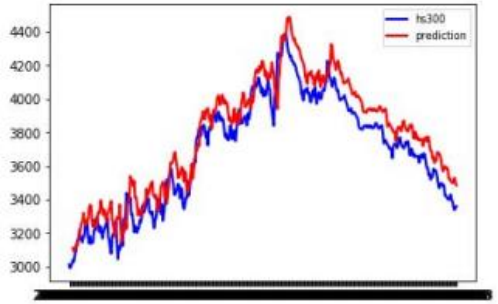


Fig. 10. HS300 index over 2 years

예측 데이터가 실제 데이터의 시간 지연이 보이며
입력 데이터의 범위(시간의 범위)가 증가하면 시간 지연이 점차 감소합니다.

논문 소개

Stock Market Prediction based on Deep Long Short Term Memory Neural Network

Xiongwen Pang[1], Yanqiang Zhou[1], Pan Wang[2], Weiwei Lin[3] and Victor Chang[4]
[1]*School of Computer, South China Normal University, Guangzhou, China*
[2]*China Merchants Bank Branch in Wuhan, Wuhan, China*
[3]*School of Computer Science and Engineering, South China University of Technology, Guangzhou, China*
[4]*International Business School Suzhou, Xi'an Jiaotong-Liverpool University, Suzhou, China*

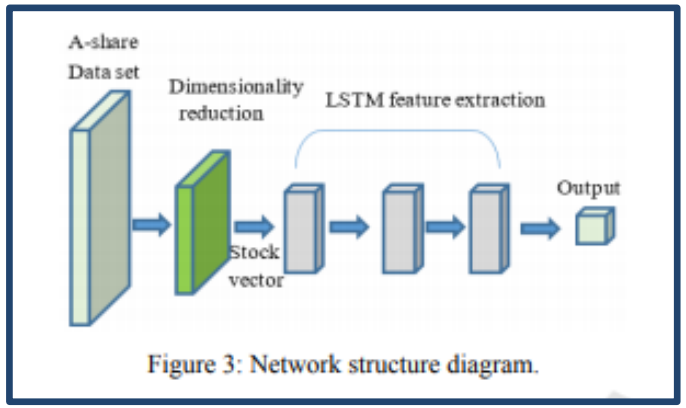
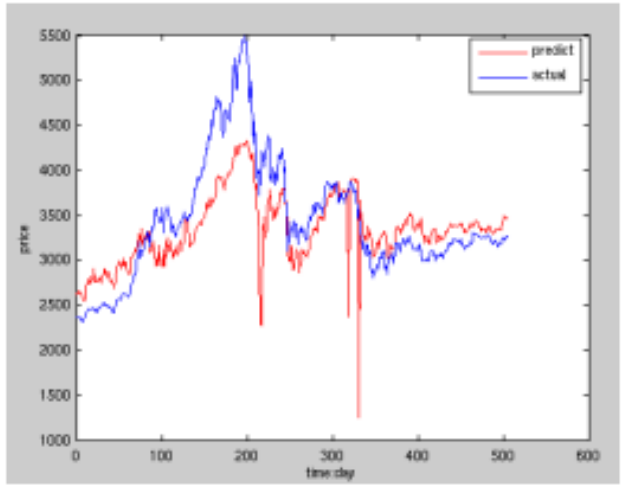


Figure 3: Network structure diagram.

Embedding layer + LSTM



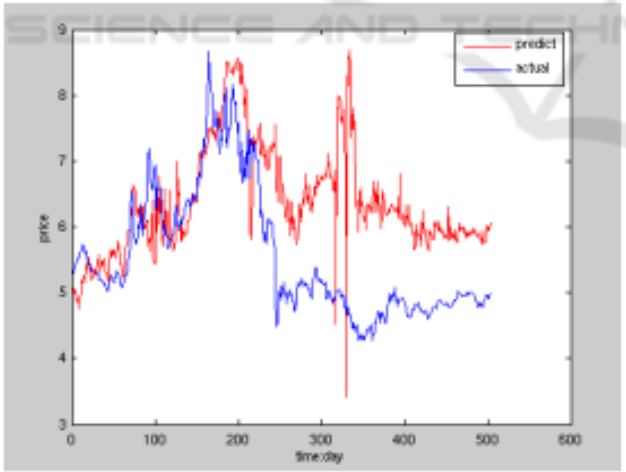


Figure 10: Comparison of predictive value and true value of the Sinopec.

논문 소개

applied sciences

MDPI

Article

Predicting the Trend of Stock Market Index Using the Hybrid Neural Network Based on Multiple Time Scale Feature Learning

Yaping Hao * and Qiang Gao

School of Electronic and Information Engineering, Beihang University, Beijing 100191, China; gaoqiang@buaa.edu.cn

* Correspondence: haoyaping@buaa.edu.cn; Tel.: +86-138-1096-8583

Received: 25 April 2020; Accepted: 4 June 2020; Published: 7 June 2020

check for updates

스위스 온라인 학술지 출판연구소

회사

MDPI

IMPACT FACTOR 2.679

CITESCORE 3.0 SCOPUS

논문 소개

I. INTRODUCTION

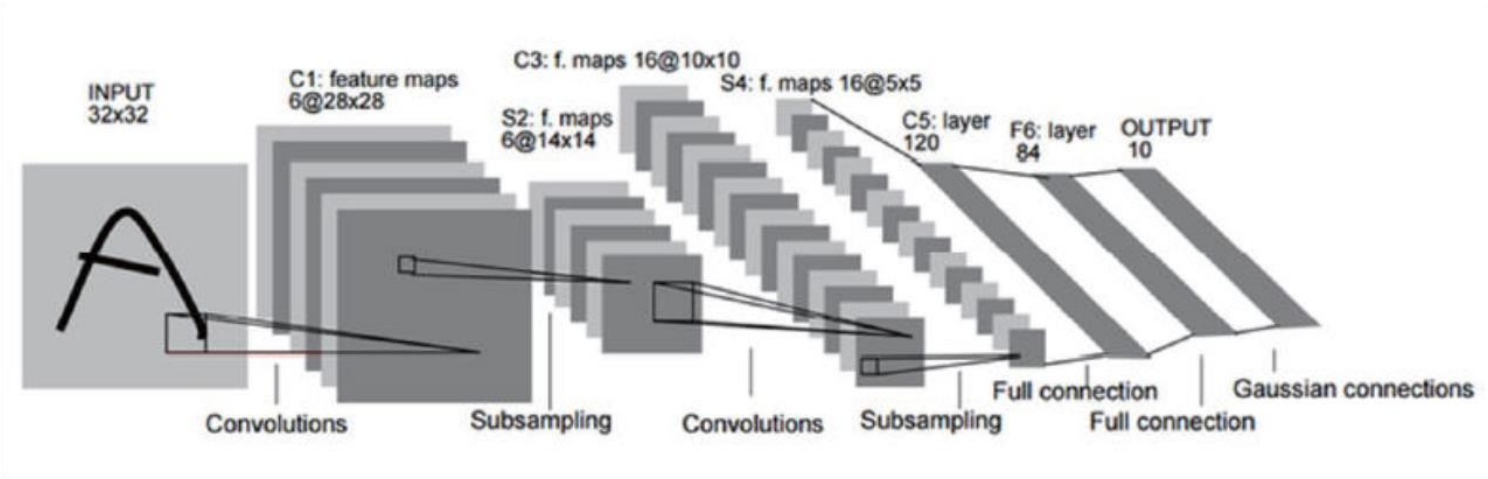
현재까지의 대부분의 연구는 주식 시장 지수의 단일 시간 척도 측면에서 수행되었습니다.
주식 시장은 경제 환경, 정치 정책, 산업 발전, 뉴스, 자연 요인 등과 같은 많은 요인의 영향을 받기 때문에
요인이 주가에 영향을 끼치는 기간은 서로 다릅니다.

따라서 우리는 주식 시장 지수에서 여러 시간 척도의 특징을 관찰 할 수 있습니다.
그 중 한달 뒤 주가를 학습한다면 모델은 주가의 장기적인 추세를 반영 할 수 있는 반면,
하루 뒤의 주가를 학습한다면 주가의 단기적인 변동을 반영 할 수 있습니다.

따라서 다중 스케일을 조합하여 모델을 만든다면 더욱 정확한 예측을 할 것이라고 판단,
주가 데이터의 시간 별 특징을 추출하여, 이를 학습시킬 예정입니다.

논문 소개

데이터에서 짧은(일주일) 주기 - 긴 주기(한 달) 추출



‘Gradient-based learning applied to document recognition’에 실린 블럭도

컨볼 루션 신경망 (CNN)은 특징 추출에서 높은 성능을 보여 왔습니다.
기존 연구에서 영감을 받아 CNN을 사용하여 여러 시간 척도 기능을 추출합니다.

논문 소개

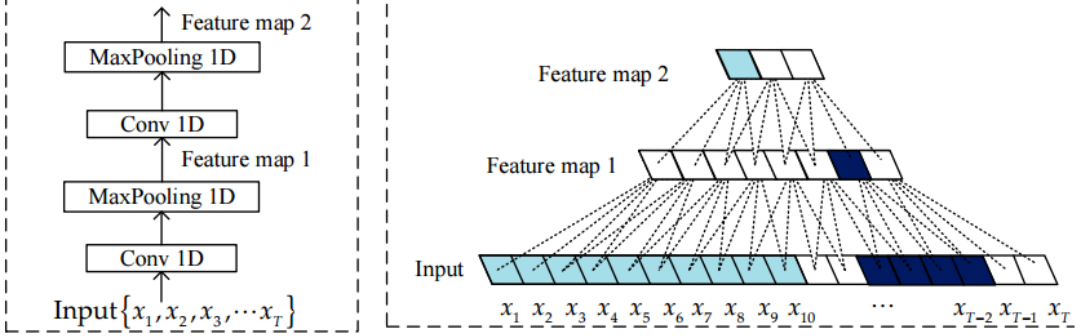


Figure 1. An example showing that CNN can extract features of different time scales. (**a**) The daily closing price series learned by a convolutional neural network; (**b**) Description of the time scale of the features.

여기서는 CNN을 통한 특징 추출로 총 두개의 Feature map을 만들었습니다.
Feature map1은 단기적인 추세를 나타내며, Feature map2는 장기적인 추세를 나타냅니다.

1. Basic Input data - 40개 간의 데이터 (unit = day)
2. Feature map1 data - 8개의 데이터 (unit = week)
3. Feature map2 data - 2개의 데이터 (unit = month)

논문 소개

End-To-End Hybrid Neural Network 제안

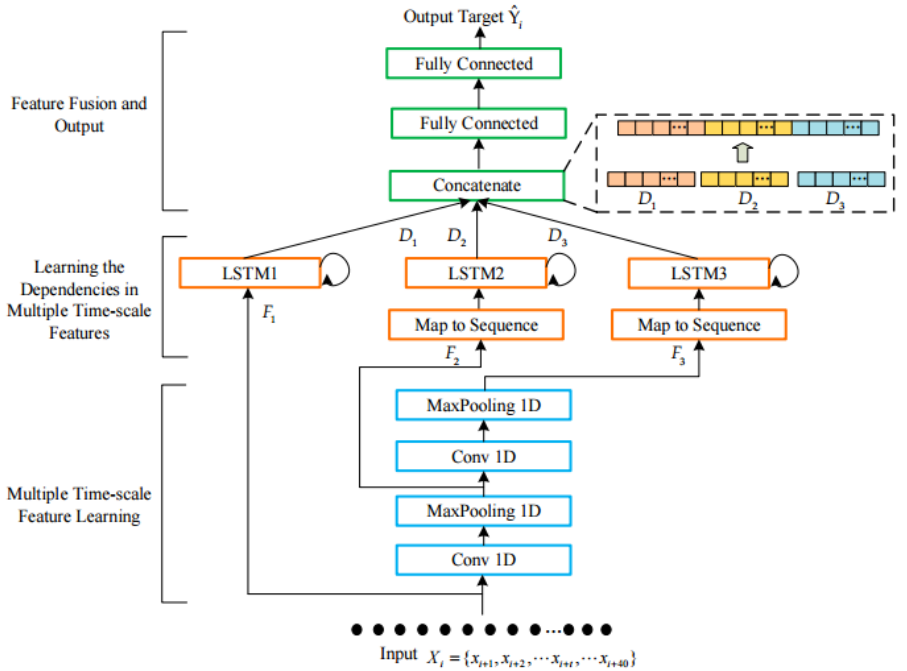


Figure 2. The proposed hybrid neural network based on multiple time scale feature learning.

1. Basic Input data - 40개 간의 데이터 (unit = day)
2. Feature map1 data - 8개의 데이터 (unit = week)
3. Feature map2 data - 2개의 데이터 (unit = month)

⇨ 3가지의 벡터를 Concatenate해서 Model 구축

논문 소개

$$Y_i = \begin{cases} 1 & x_{i+40} \le x_{i+40+n} \\ 0 & x_{i+40} > x_{i+40+n} \end{cases}$$

해당 논문에서는 직접적인 종가 예측이 아닌, 추세를 예측하는 모델을 목표로 함
n = 1(a day), 5(a week), 20(a month) 이며
종가를 한달 or 일주일 or 하루 뒤 비교했을 때 상승 -> 1 ,하락 -> 0 으로 학습

Standard & Poor's 500 Index

1999년 1월 30일 ~ 2015년 1월 30일 train data
2015년 1월 30일 ~ 2017년 1월 30일 validation data
2017년 1월 30일 ~ 2019년 1월30일 test data

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Model Train
optimizer : Adam
loss function : binary cross-entropy

Result

논문 소개

Table 1. Comparisons with models based on single time scale features in accuracy.

Forecast Horizon	Model	Accuracy (%)
One week	Model based on F_1	64.40
	Model based on F_2	63.30
	Model based on F_3	61.76
	The proposed model	**66.59**
One month	Model based on F_1	71.14
	Model based on F_2	70.23
	Model based on F_3	73.64
	The proposed model	**74.55**

F1 : a day model, F2 : a week model , F3 : a month model
The proposed model = F1 + F2 + F3

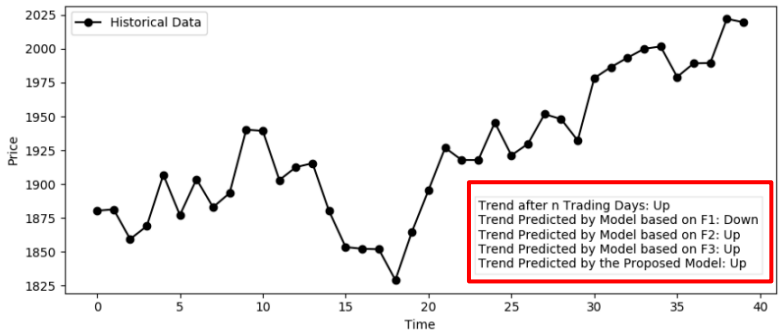


Figure 6. Visualization of the trend prediction by different models on test example 1.

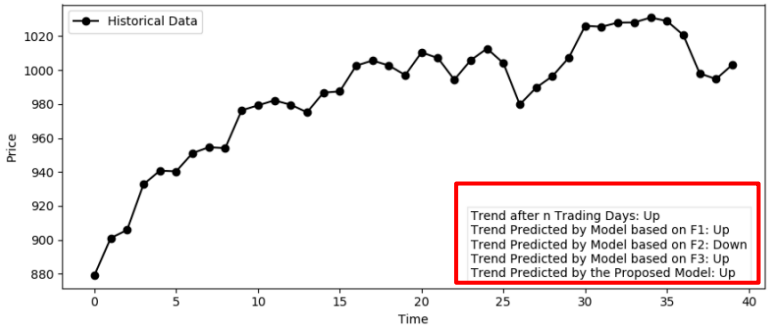


Figure 7. Visualization of the trend prediction by different models on test example 2.

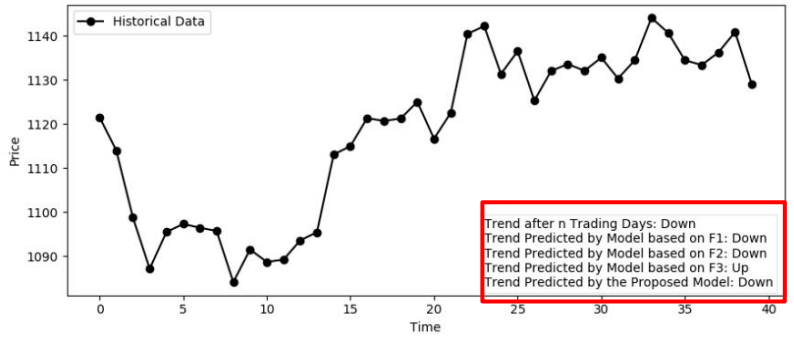


Figure 8. Visualization of the trend prediction by different models on test example 3.

Day(Figure 6), week(Figure 7), month(Figure 8) 일 때
각기 잘못 예측하는 경우가 발생할 수 있는 것을 보여줌

논문 소개

Table 2. Comparisons with existing models in accuracy.

Forecast Horizon	Model	Accuracy (%)
One week	Simplistic Model	54.82
	SVM	61.98
	LSTM	65.05
	CNN	59.34
	Multiple Pipeline Model	63.30
	NFNN	65.93
	The proposed model	**66.59**
One month	Simplistic Model	56.92
	SVM	70.91
	LSTM	71.59
	CNN	67.95
	Multiple Pipeline Model	72.05
	MFNN	72.27
	The proposed model	**74.55**

기존의 모델들의 예측 정확도를 수치화 시킴

제시한 모델이 기존 모델보다 좋은 정확도를 나타낸다. 라는 것을 확인시켜줍니다.

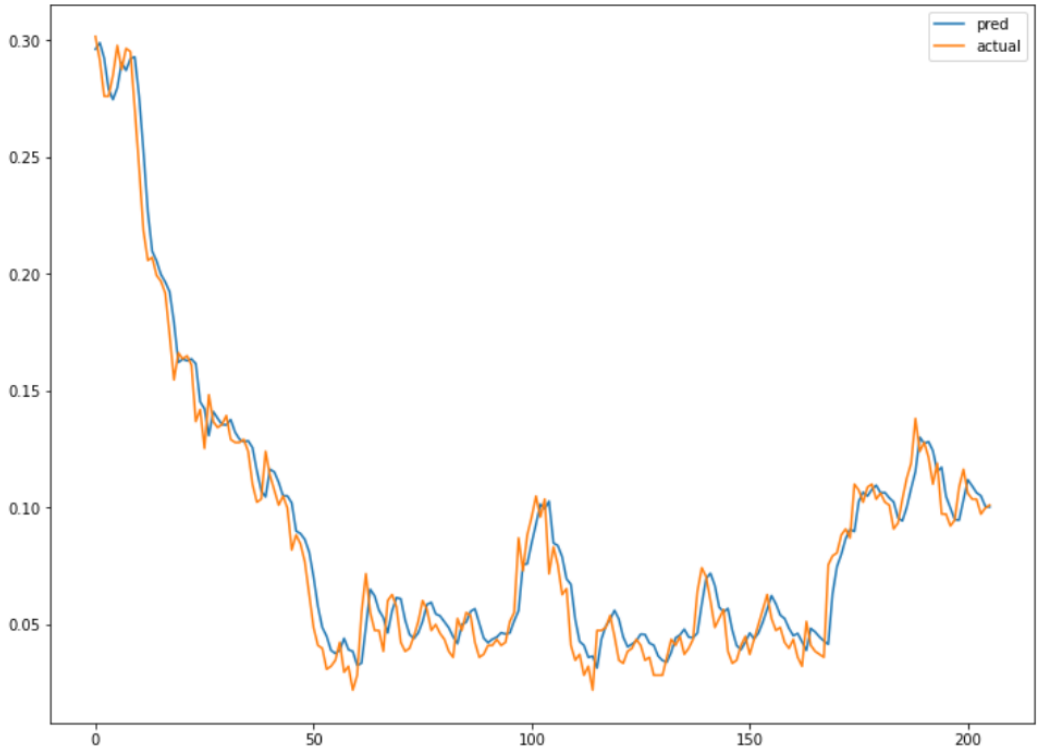
해당 논문에서 고려해야하는 부분

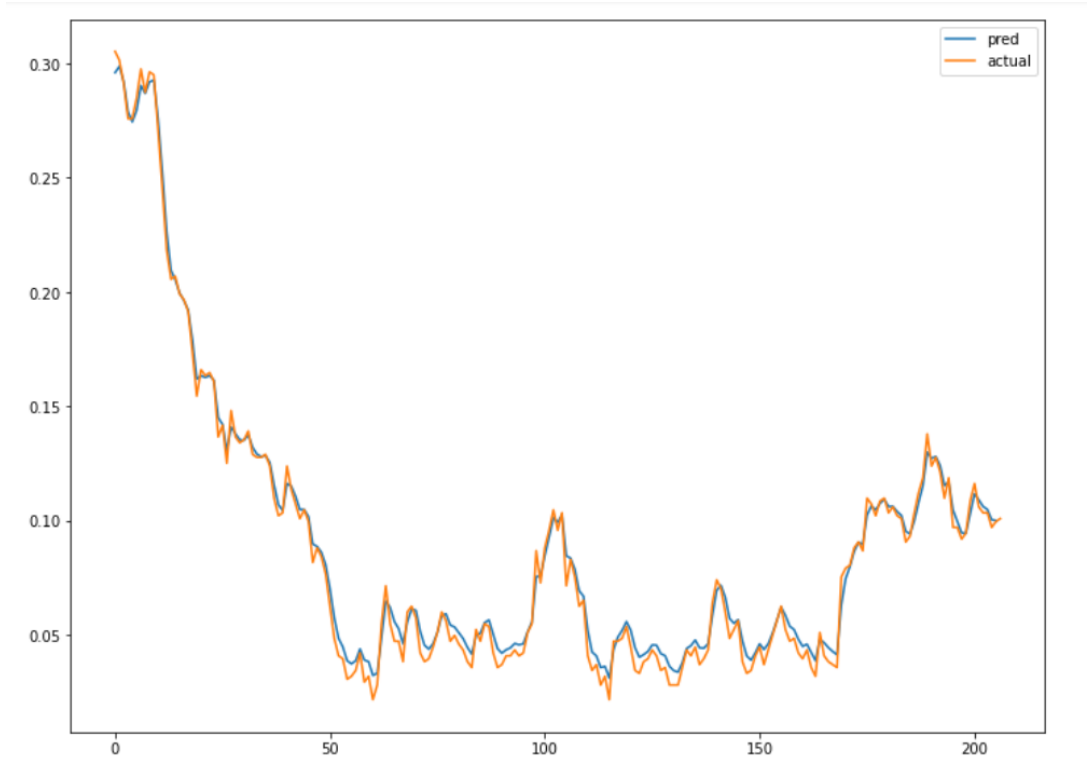
1. concatenate 부분을 보팅방식으로 바꿔도 될 것 같다.
2. 다양한 모델에 대한 평가가 없었다.

논문 결정

논문 결정

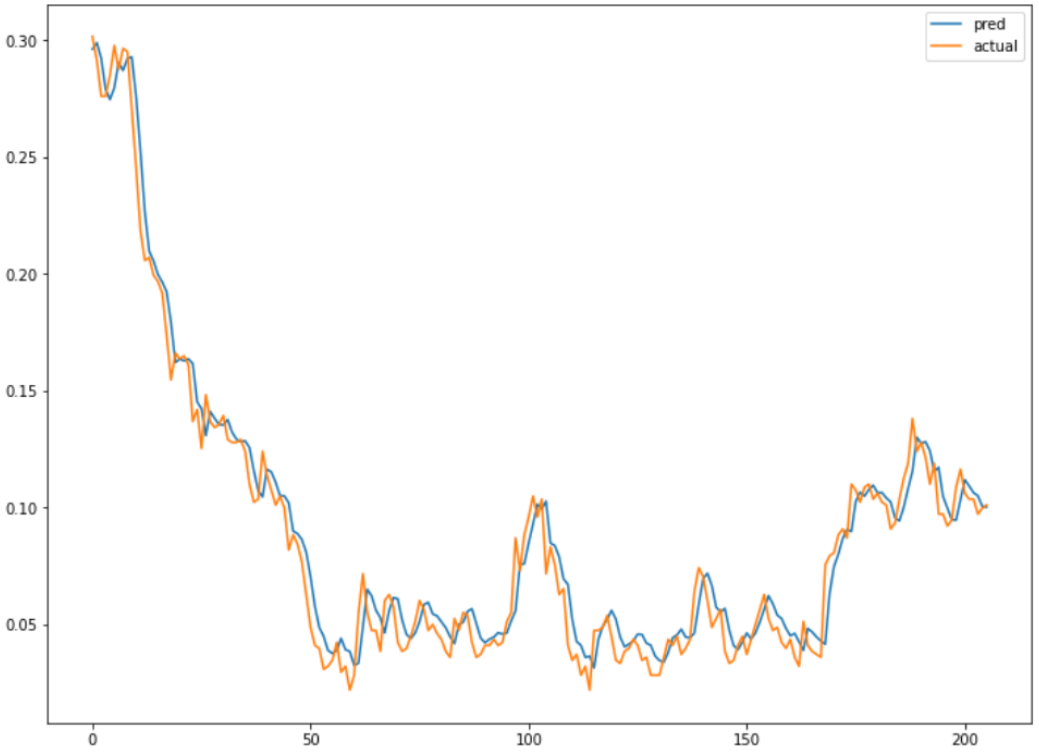
추세를 따라간다 -> LSTM의 효용성 체크가 필요함





논문 결정

추세를 따라간다 -> LSTM의 효용성 체크가 필요함



a week, a month : 확인해서,
결과를 내고 LSTM과 별 차이가 없다는
것을 테스트 하고 결론 내고 싶습니다.

Table 1. Comparisons with models based on single time scale features in accuracy.

Forecast Horizon	Model	Accuracy (%)
One week	Model based on F_1	64.40
	Model based on F_2	63.30
	Model based on F_3	61.76
	The proposed model	**66.59**
One month	Model based on F_1	71.14
	Model based on F_2	70.23
	Model based on F_3	73.64
	The proposed model	**74.55**

Q & A