Stock Analysis and Prediction Based On LSTM Deep Learning

Yujing Song
Dec. 2020

Outline

01 Backgroud
Motivation and objectives

02 Model Establish
Apple Inc. and four sectors in S&P 500

03 Special Time Period
Discoveries in the covid-19 period

04 Conclusion
The summary and future work direction

PART 01

Background

- Problem Statement

Problem Statement

How to predict the stock price?

How to apply deep learning to stock research to get more accurate prediction results?

Can stock prices be predicted?

Stock prices change according to time series. Does this mean that stock prices change regularly over time?

What can we learn from this project?

From the predicted results and the exploration of the market, can we get some guiding opinions on investment?

PART 02

Model Establish

- Start with Apple Inc.
- 4 Major Sectors
    - Technology
    - Finicial
    - Health Care
    - Transportation

Start with Apple Inc.

1 Loading the data

2 Cutting time series into sequences

3 Spliting training and testing sets

4 Building RNN regression model

5 Checking model performance

Pre-processing

Pre-processing

Loading the data → **Cutting time series into sequences** → **Splitting training and testing sets**

Loading the data

- Stock of Apple Inc. from Feb. 2013 to Feb. 2018
- Source: Super DataScience
- Daily close price
- Predicted Apple's stock price 7 days in advance

Cutting time series into sequences

Splitting training and testing sets

- 80% training data
  - 1000 records
  - Feb. 2013 - Feb. 2017
- 20% testing data
  - 251 records
  - Feb. 2017 - Feb. 2018

Pre-processing —— Cutting time series in sequences

The time series is a sequence of numbers that we can represent in general mathematically as:

$$s_0, s_1, s_2, \dots, s_P$$

where s_p is the numerical value of the time series at time period p and where P is the total length of the series.

Window of size

T = 5

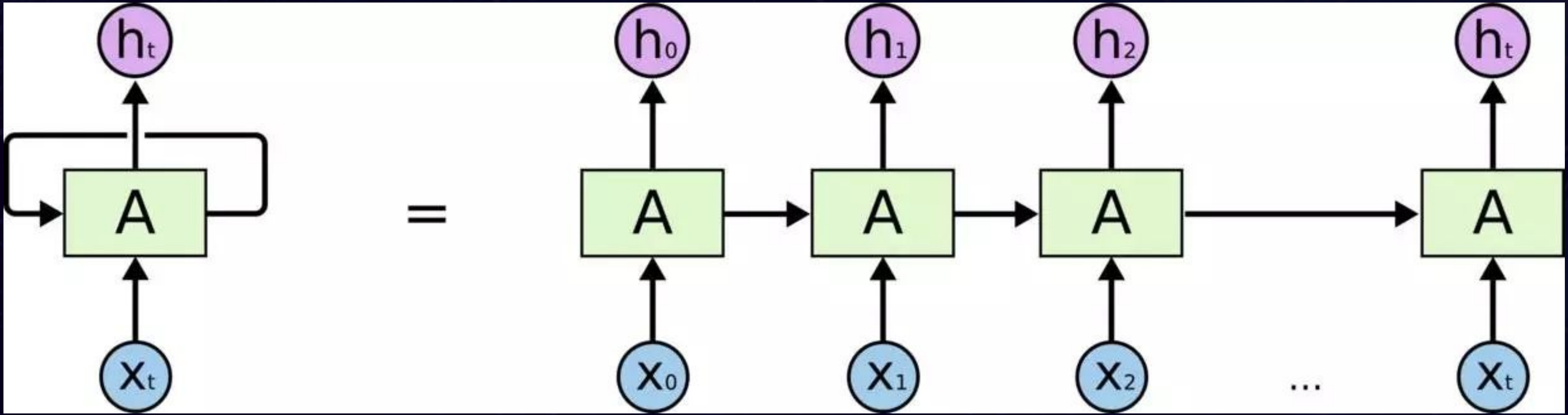
Input	Output
$\langle s_1, s_2, s_3, s_4, s_5 \rangle$	s_6
$\langle s_2, s_3, s_4, s_5, s_6 \rangle$	s_7
$\vdots$	$\vdots$
$\langle s_{P-5}, s_{P-4}, s_{P-3}, s_{P-2}, s_{P-1} \rangle$	s_P

Start with Apple Inc.

1. Loading the data
   Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor. Donec facilisis lacus eget mauris.
2. Cutting time series into sequences
   Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor. Donec facilisis lacus eget mauris.
3. Spliting training and testing sets
   Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor. Donec facilisis lacus eget mauris.
4. Building RNN regression model
   Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor. Donec facilisis lacus eget mauris.
5. Checking model performance
   Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor. Donec facilisis lacus eget mauris.

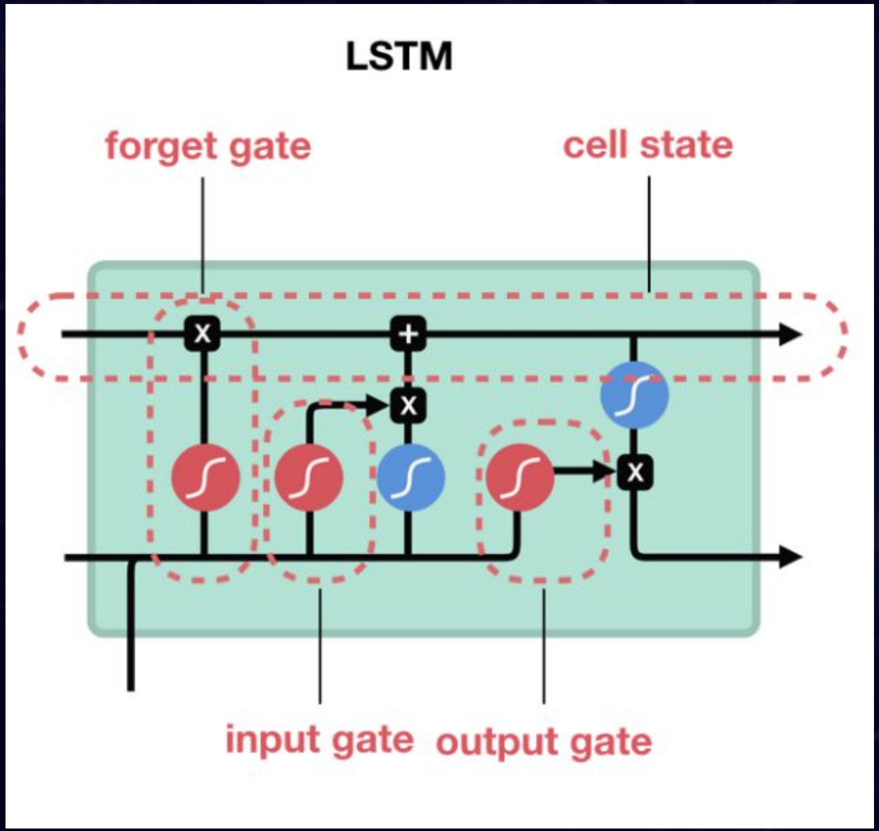
Pre-processing

Basic Structure of RNN

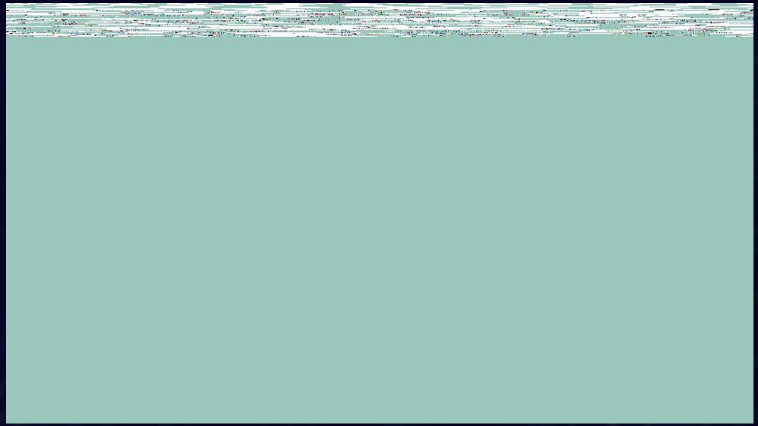


- Problem that RNN solve: sequence problem
- Elements are not independent of each other. They have dependencies.

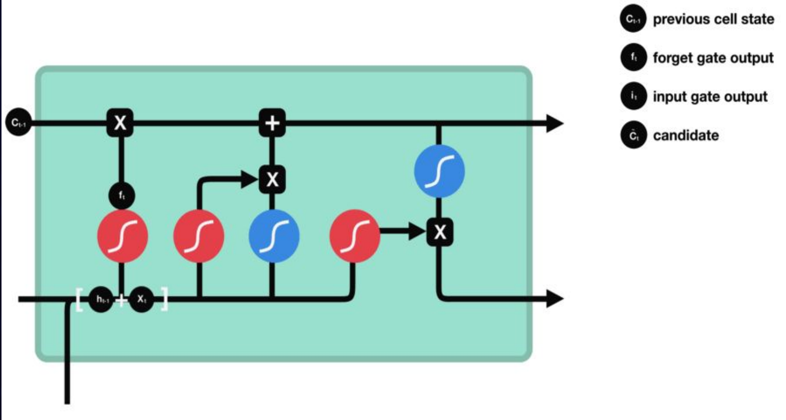
Basic Structure of LSTM



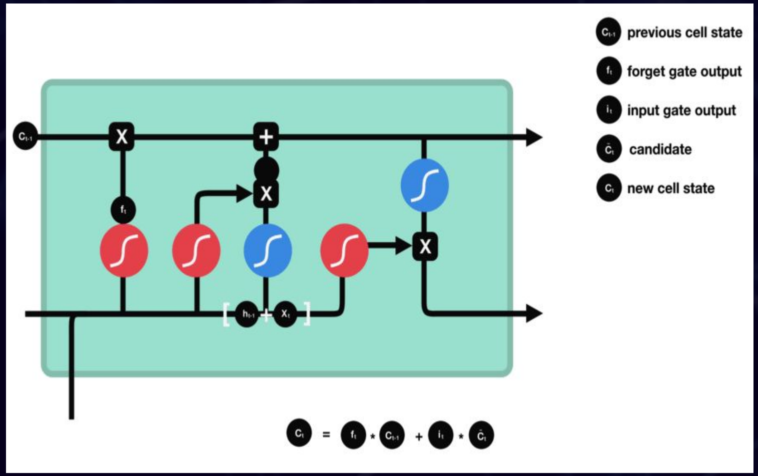
Forget Gate:



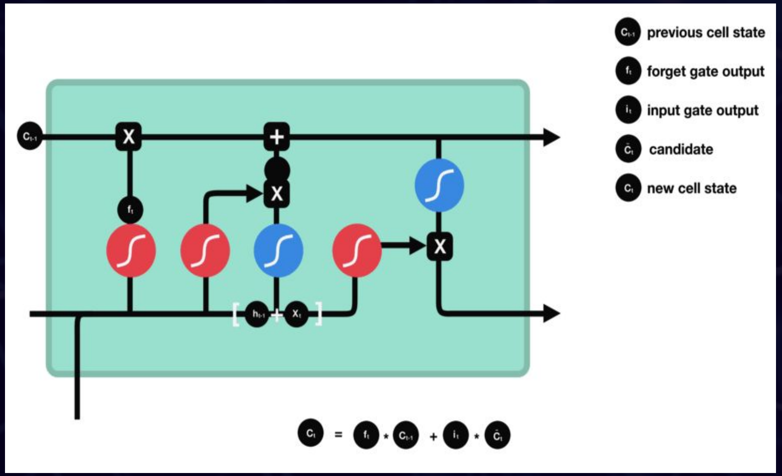
Input Gate:



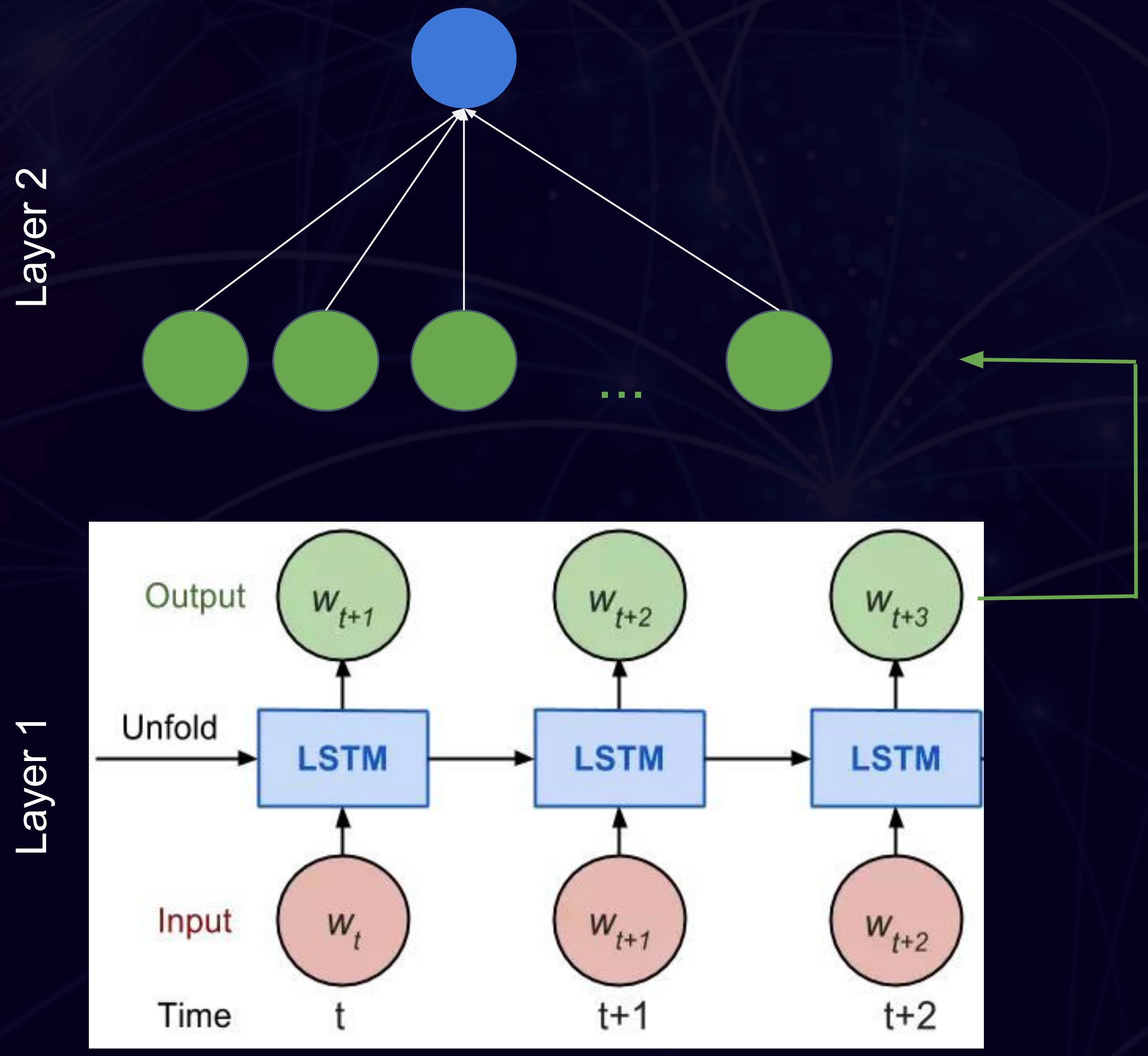
Cell State:



Output Gate:



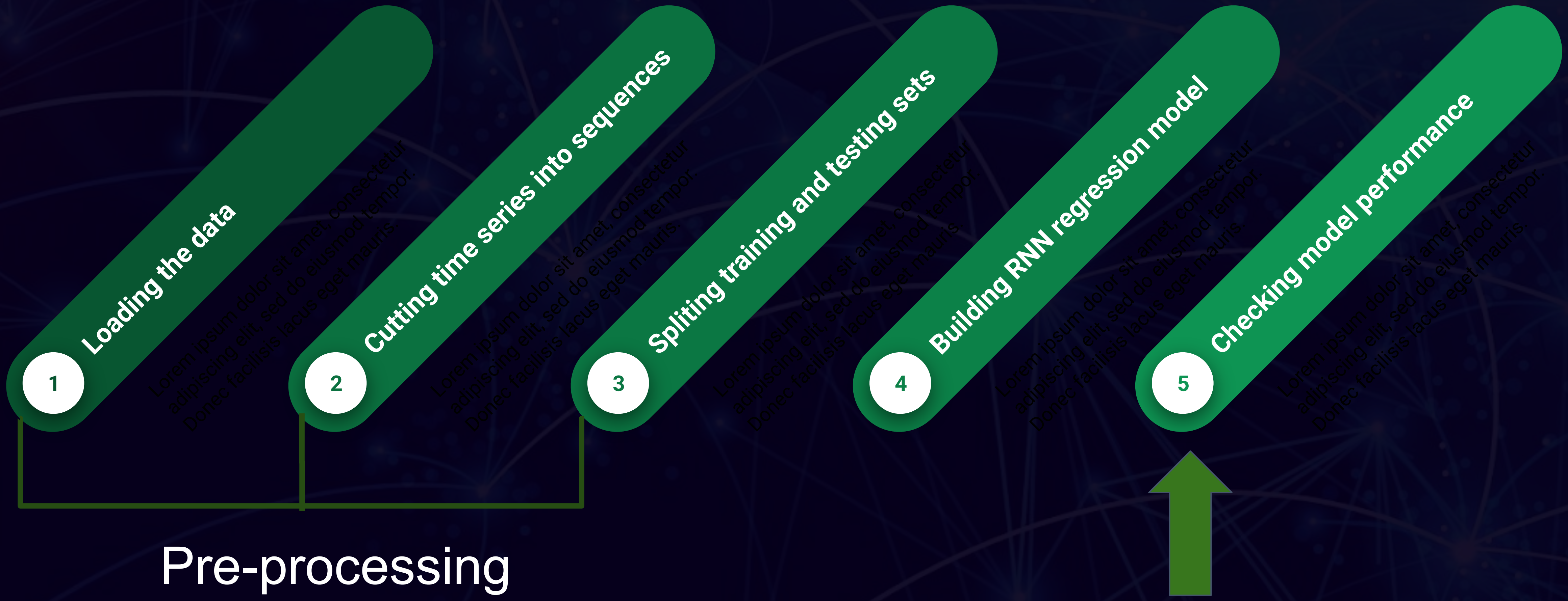
Basic Structure of LSTM



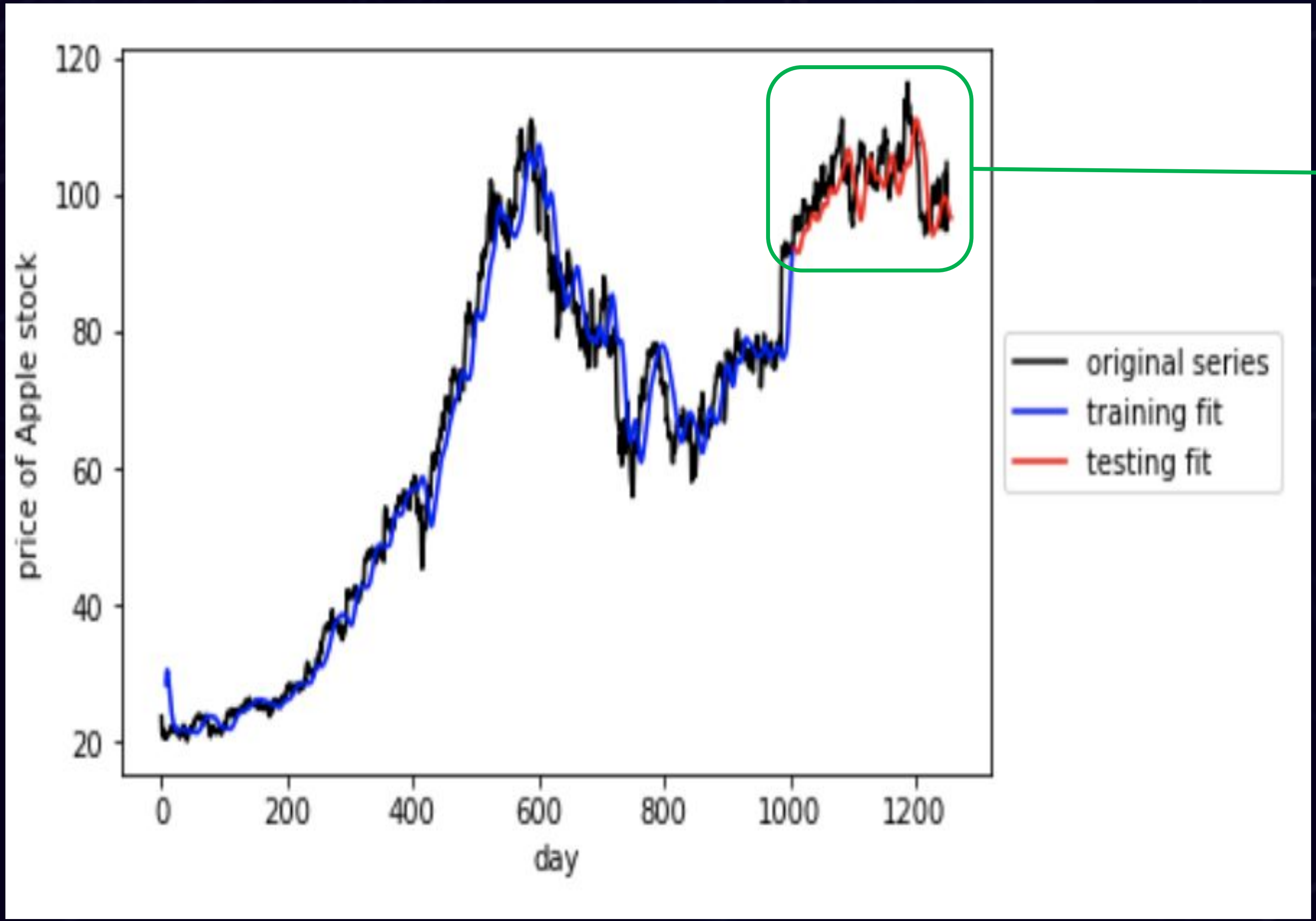
Two hidden layer RNN of the following specifications:

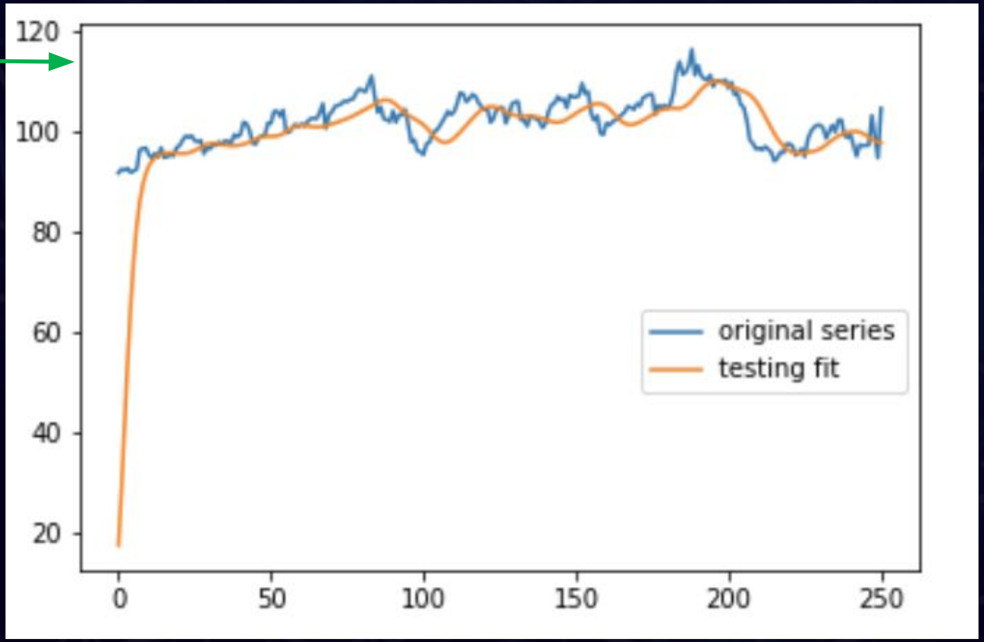
- layer 1 uses 3 LSTM module with 64 hidden units, input size is 7.
- layer 2 uses a fully connected module with one unit
- Loss function: MSE

Start with Apple Inc.



Prediction Result





Apply Same Modle to Four Main Sectors

Loading the data	Cutting time series into sequences	Spliting training and testing sets	Building RNN model	Check model performance

Loading the data

- Source: Yahoo Finance
- 4 Sectors:
  - XLK: Technology
  - XLV: Hearlth Care
  - XLF: Finicial
  - XTN: Transportation
- Dec. 2010 - Dec. 2020
- Daily close price

Cutting time series into sequences

- Window size = 7
- Predicted 7 days in advance

Spliting training and testing sets

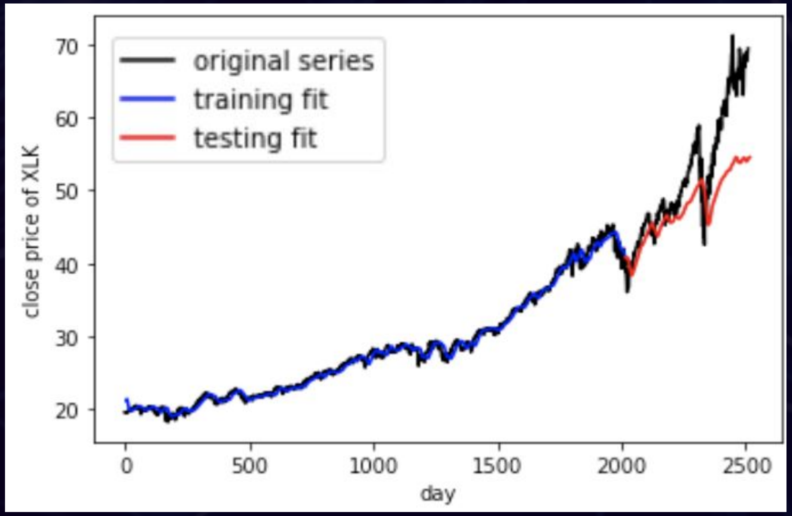
- 80% training data
  - 2008 records
  - Dec. 2010 - Dec. 2018
- 20% testing data
  - 502 records
  - Dec. 2018 - Dec. 2020

Building RNN model

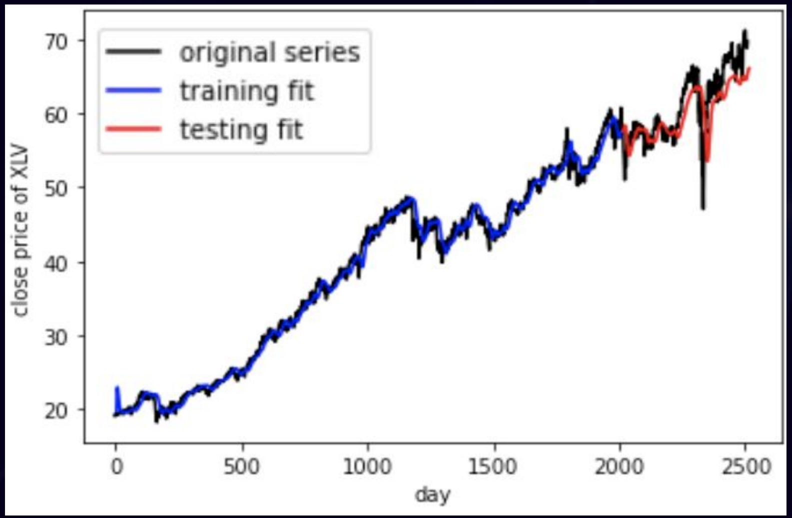
- RNN + LSTM
- 2 layers:
  - LSTM
  - Fully Connection

Model Performance for 4 Sectors

XLK



XLV

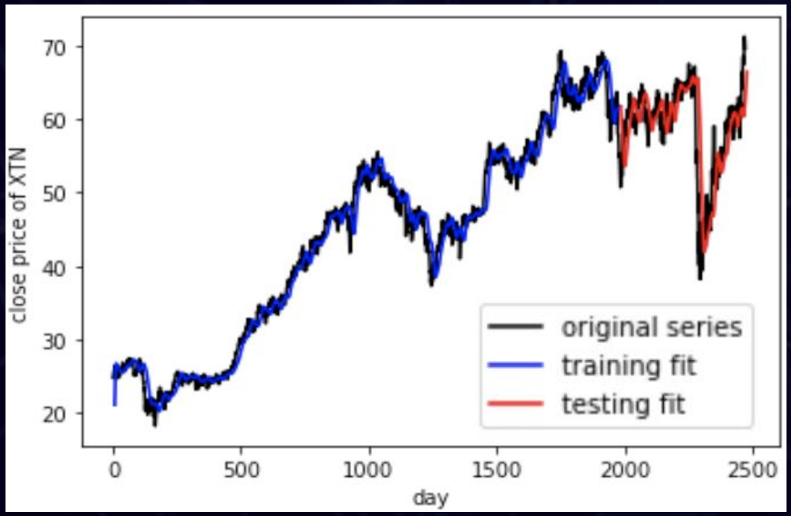


RMSE for test set:

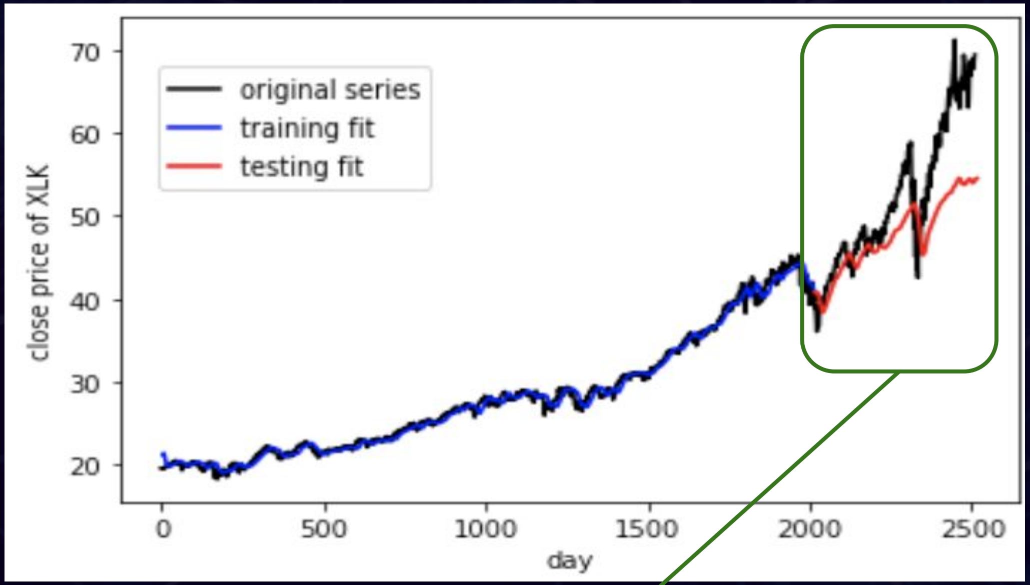
- XLK: 6.46
- XLV: 3.95
- XLF: 4.46
- XTN: 3.99

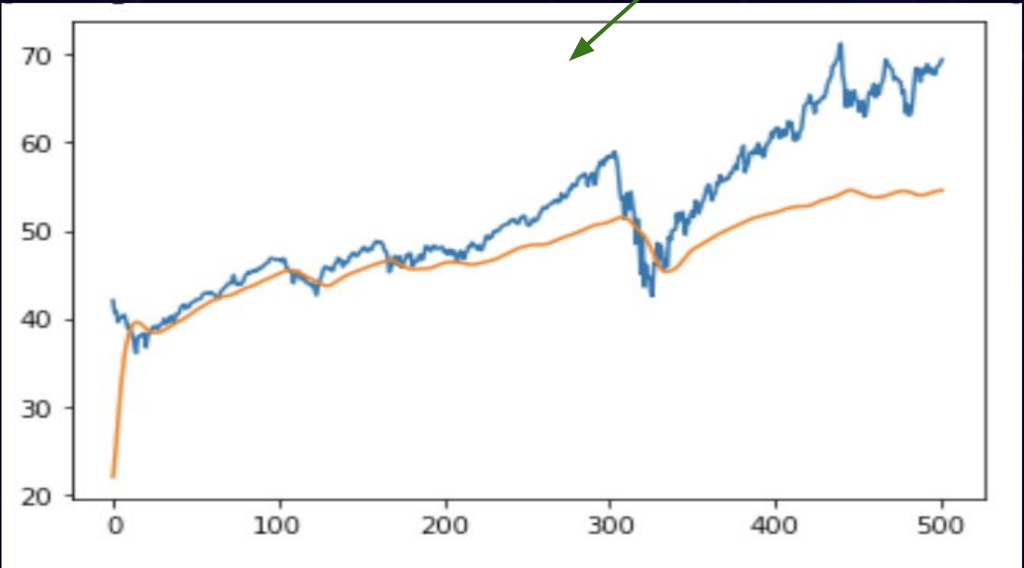
XLF

XTN



Refine the LSTM Model for XLK



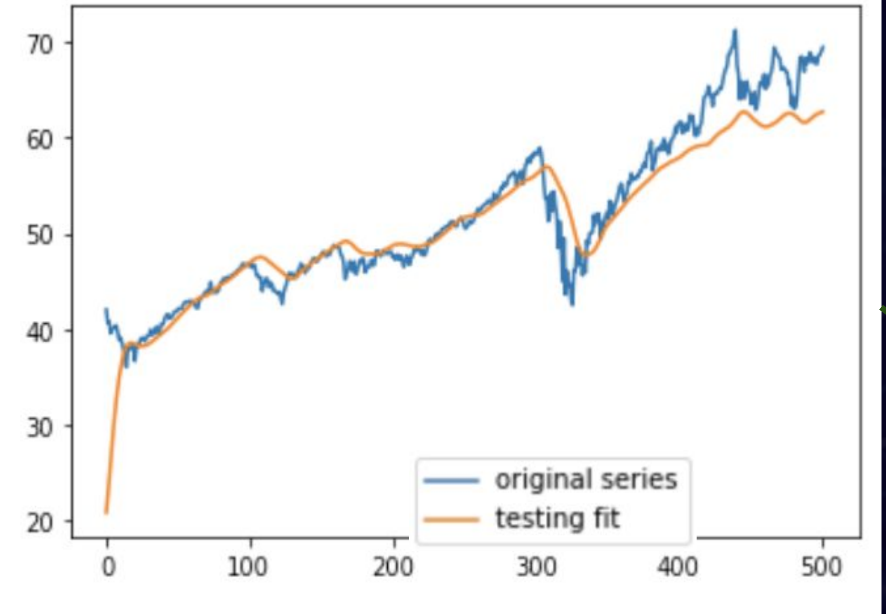


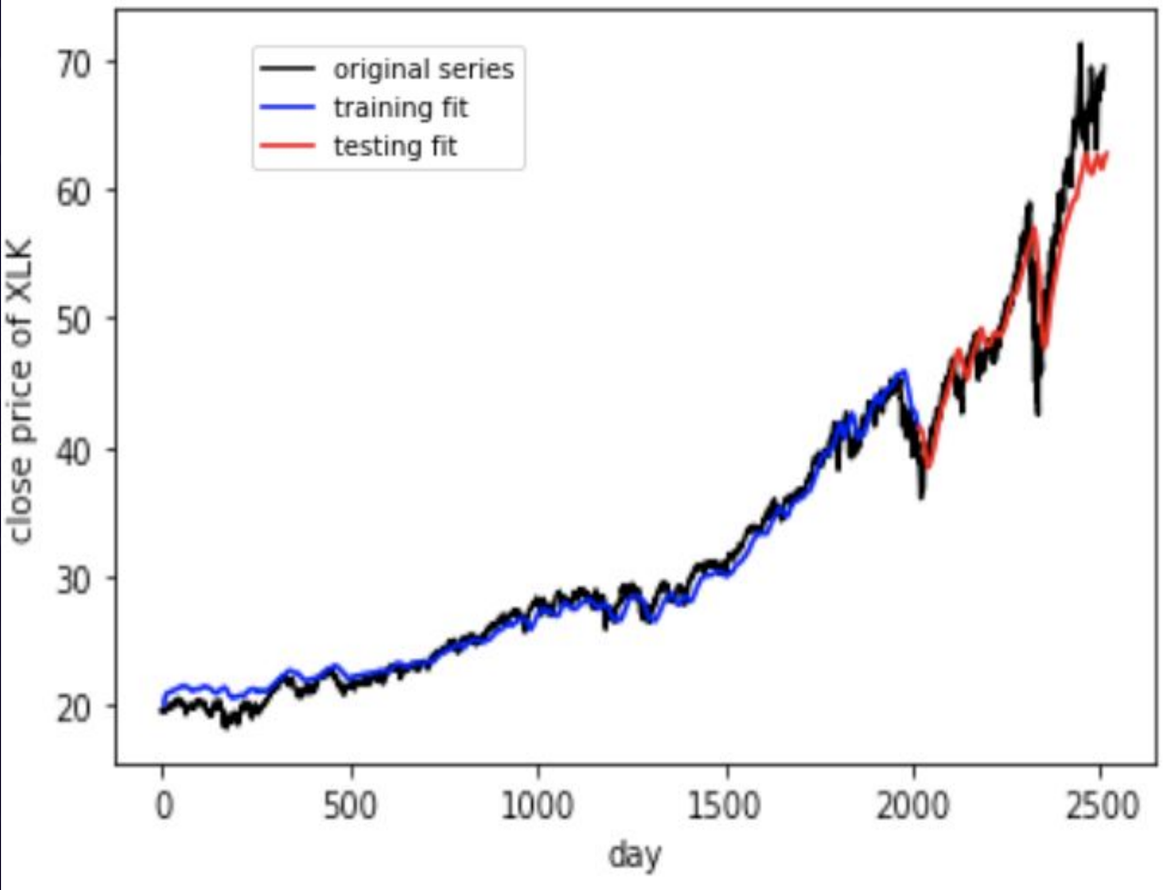
Performance well in training set but not well in testing set

Overfitting

- L1 and L2 Regression
- Dropout ✦
- Early Stopping
- Simplier model structure ✦
- Increase data
- ...

Refine the LSTM Model for XLK





Method:

- Decrease the layers number (layer_number = 2)
- Increase the dropout rate (dropout = 0.2)
- Adjust the learning rate (learning_rate = 0.001)

Result:

RMSE for testing set = 3.33

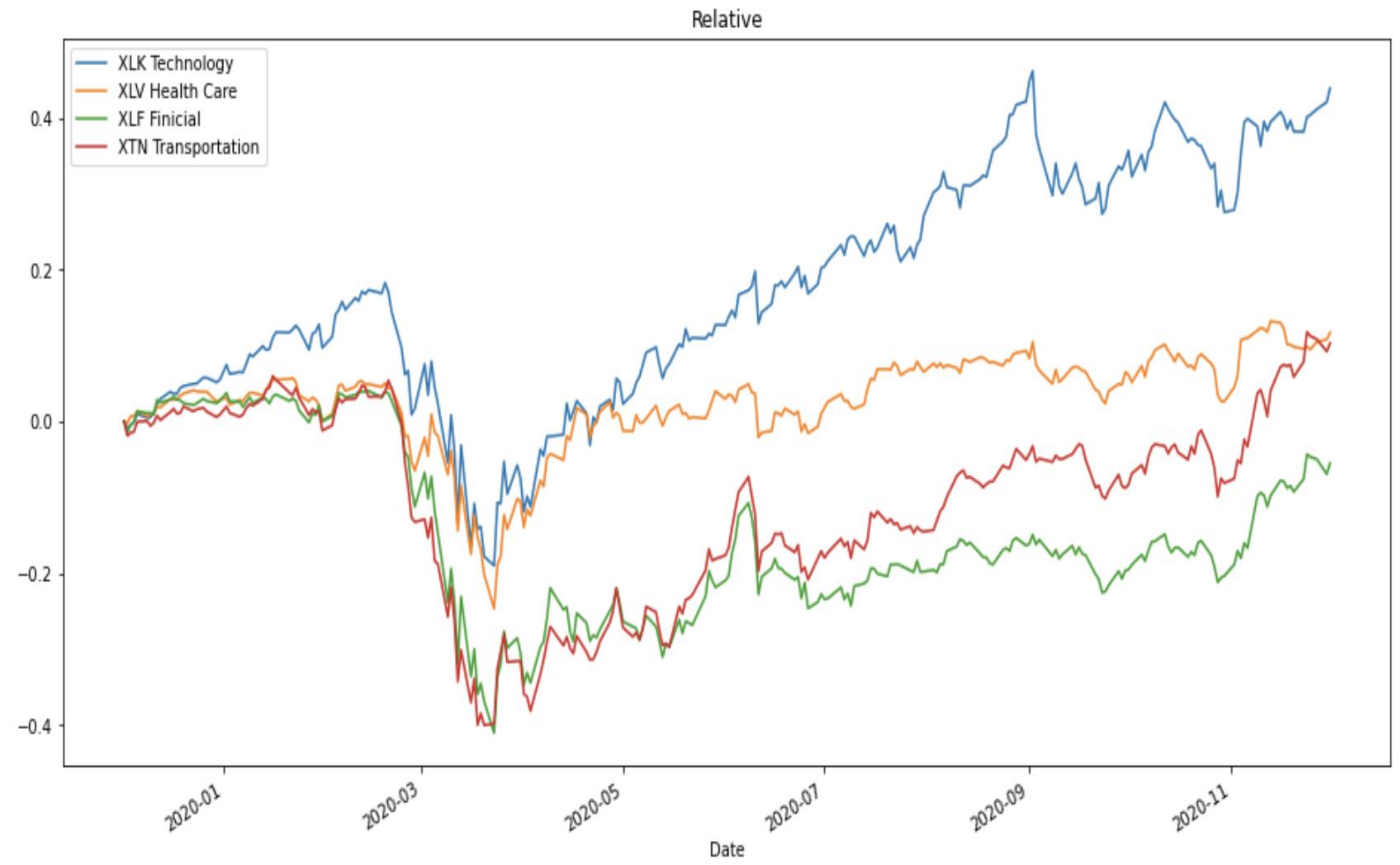
PART 03

Covid-19 Period

- The Whole Market
- Relative Change
- Daily Percentage Change
- Sharpe Ratio

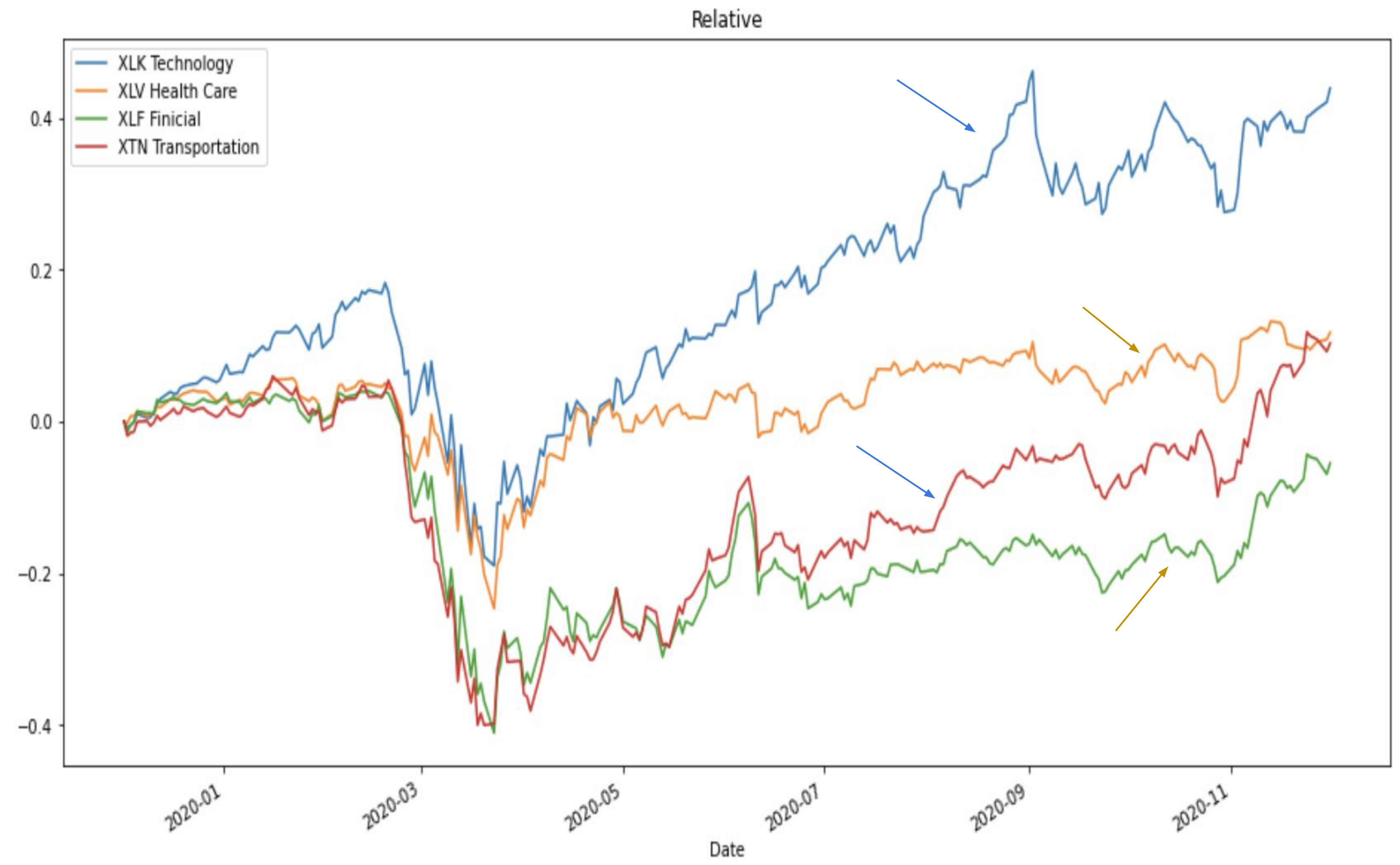
Sepecial Time Period: Covid-19 Period

Relative Change in Covid-19 Period



- All sectors experienced precipitous declines in April, then gradually recovering.
- XLK:
  - One of the fastest to recover.
  - Returned to normal levels in July
  - Continuing to rise rapidly.
- XLV:
  - Fastest to return to the level before covid-19
  - Stay horizontal
- XLF:
  - Recovery, but not much
- XTN:
  - Increase rapidly after July
  - Another quickly growth after November

Relationship Among Four Sectors



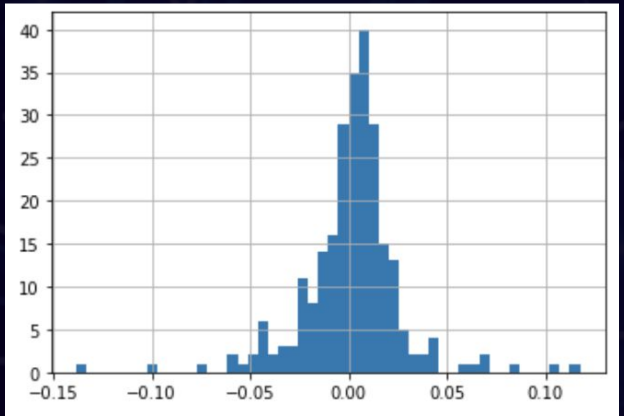
Pearson Correlation:

- XLK and XTN: 0.897
- XLV and XLF: 0.798

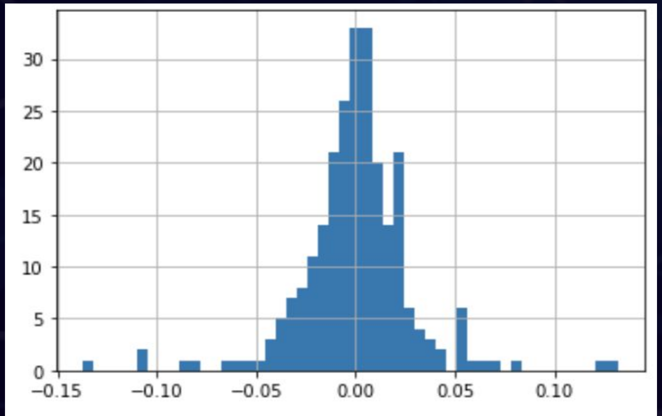
The relative change of XLK and XTN have strong correlation, so as XLV and XLF.

Daily Percentage Change in Covid-19 Period

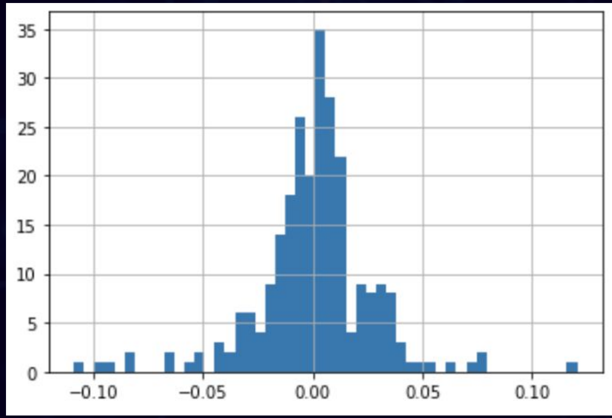
XLK



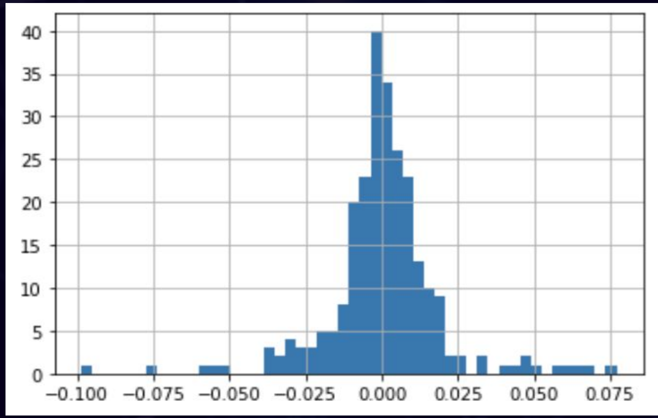
XLF



XTN

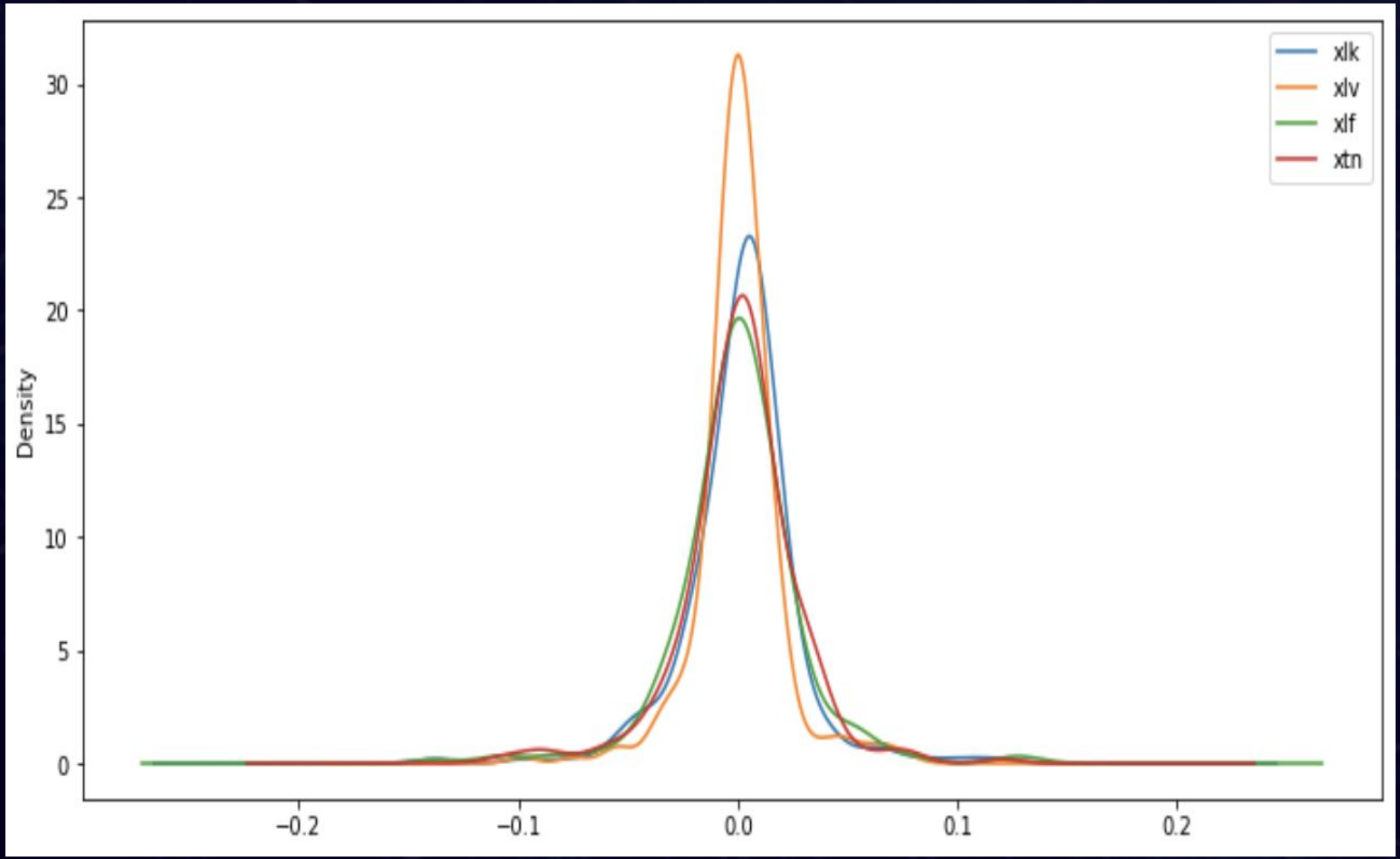


XLV



Daily Percentage Change in Covid-19 Period

- Most close price are unchanged
- The change percentage of XLV is the most concentrated
- XLF and XLK are really close to normal distribution
- XTN has more days with price decreasing
- XLV more days with price increasing.



Sharpe Ratio in Covid-19 Period

$$Sharpe\ Ratio = \frac{R_p - R_f}{\sigma_p}$$

R_p = Return of portfolio

R_f = Risk-Free rate

σ_p = Standard deviation of portfolio's excess return

	Sharpe Ratio before Covid-19	Sharpe Ratio in Covid-19 Period
XLK	2.685	1.108
XLV	2.001	0.515
XLF	1.388	0.103
XTN	1.296	0.444

For every extra unit of risk, the premium that investor can get in covid-19 period is lower than previous years. However, the technology sector still managed to make a decent return this year.

PART 04

Conclusion

- Conclusion for This Project
- Suggestions for Investment
- Future Work

Conclusion

- The mothod to predict stocks: LSTM
  - Basic structure of RNN
  - Basic structure of LSTM
  - Combine together and apply in data
    - Apple Inc
    - 4 sectors in S&P500: XLK, XLV, XLF, XTN
- Special time period: Covid-19
  - Relative change:
    - XLK increased fast after April. The stock price even higher than before.
    - XTN had an extremly rapid increase
    - The relative change of XLK and XTN have strong correlation, so as XLV and XLF
  - Correlation:
    - The relative change of XLK and XTN have strong correlation
    - The relative change of XLV and XLF have strong correlation
  - Daily percentage change:
    - Most investment are unchanged
    - XTN has more days with price decreasing
    - XLV more days with price increasing
  - Sharpe ratio:
    - The sharpe ratio in covid-19 is lower than normal
    - XLK is still the best choice

Conclusion

Suggestions for investors based on the project

- Investoment should be diversified.
- Technology sector is always the best choice.
- Health Care is the best choice for investors who want to invest in the lowest risk.
- Transcription has a rapid increase recently.
- Financial is slowly recovering.
- Stock prices can be predicted by LSTM algorithm.

Directions for the future work

- Try multiple models other than LSTM.
- For each sector, detect which company is the best choice to invest.
- Studying stock market changes in recent years as all pandemics occur. It may contains some regularities.

Thank You!

Any questions or comments?

You can find me by email: ys3251@columbia.edu

Code for this project: click to get the code
Previous projects:
https://github.com/YujingSong/Personal-Portfolios