Stock Price Prediction Using Deep Learning Model

Lu, Tao
Department of Information Systems,
Business Statistics and Operations Management
Hong Kong University of Science and Technology
tluae@connect.ust.hk
20297398

Abstract

In this project, a deep learning model is used in stock price predicting, the advatages and shortcomings of the deep learning model are discussed. Data specification and model specification are introduced and backtest of the porfolio constructed based on deep learning model are listed.

1 Introduction

Stock price prediction is a high profile problem which is intrigue and fascinating, since the task is related to real money directly, and the return can be magnified by leverage to a great extent. On the other hand, the machine learning is a hot topic recently, especially deep learning. In this project, a stock prediction model trained by deep learning method will be constructed and discussed.

According to the hypothesis of efficient market, proposed and summarized by (Fama, 1970), the stock market will reflect all the information which will drive the stock price immediately. If the hypothesis is true, the stock prices can not be predicted by historical trading data or announced firm information. Empirically, the efficient market hypothesis describes the stock market approximately, and scholars have found some anomalies such as momentum factor, accounting factors, and some behavior factors. So, we can have a basic expectation about the stock prediction project, there exists some signals can be mined from historical and announced data, but the prediction power will be weak and can not be very robust.

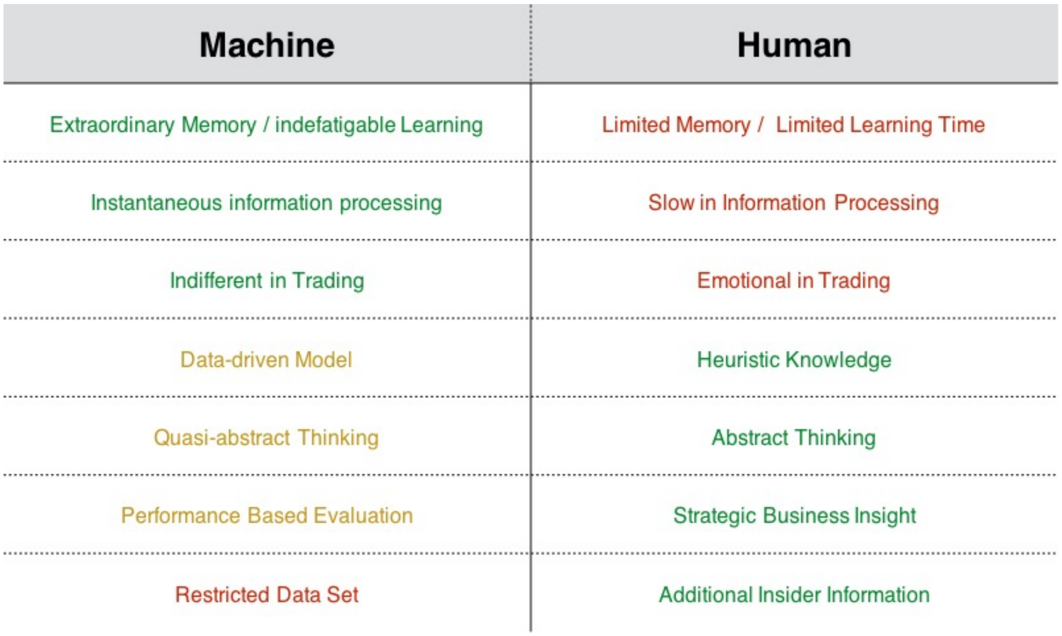
2 Analysis

Whether the deep learning model is suitable for stock preediction, this is the first question we have to ask. So, I first compare the deep learning with human decision makers, asset pricing model, and traditional machine learning methods. From Figure 1 to Figure 3, the tables are the results I get. All the subjects or methods have their advantages which is marked as green and shortcomings which is marked as red, and neutral points which is marked as yellow. Given these discussions, I think the deep learning model is deserved to test whether it is useful in predicting the stock price.

3 Data

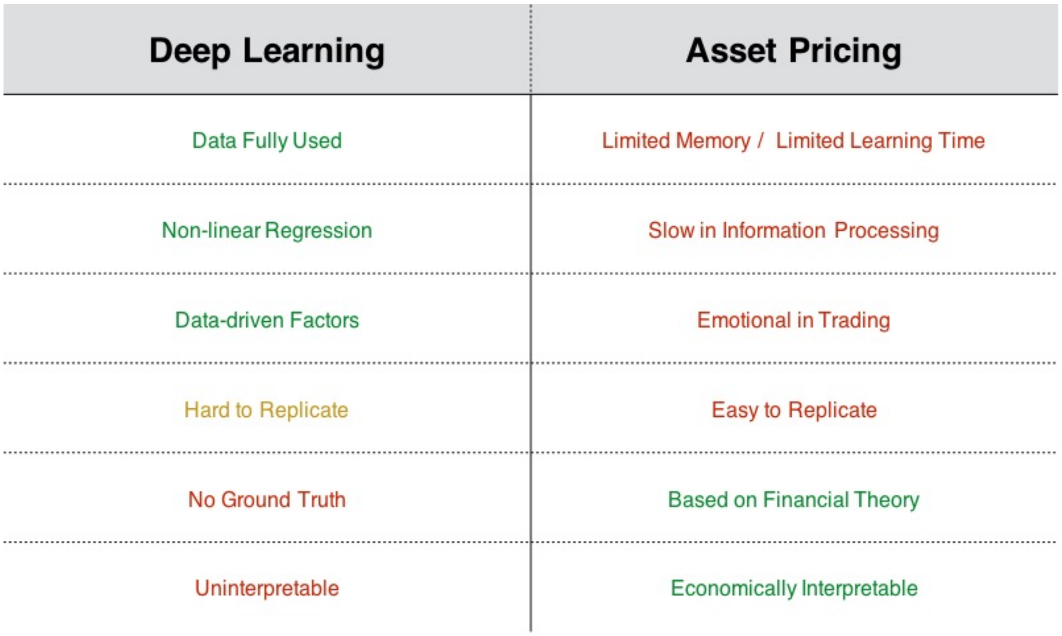
In this project, data are from chinese stock market from 2000 to 2016, within more than 2900 stocks. There are more than 7,000,000 observations in the data set. The data is basically composed by two parts, the time-series part and the dynanmic part. For the time-series data, 200 days trading history are used. Totally, there are more than 3,000 dimensions in the input data set.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.



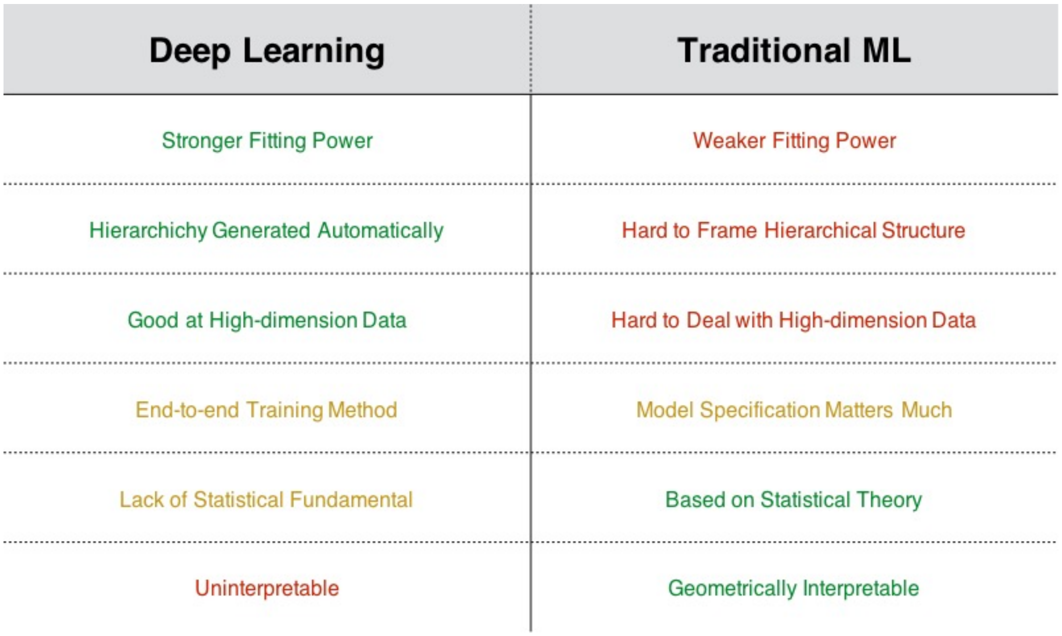
Machine	Human
Extraordinary Memory / indefatigable Learning	Limited Memory / Limited Learning Time
Instantaneous information processing	Slow in Information Processing
Indifferent in Trading	Emotional in Trading
Data-driven Model	Heuristic Knowledge
Quasi-abstract Thinking	Abstract Thinking
Performance Based Evaluation	Strategic Business Insight
Restricted Data Set	Additional Insider Information

Figure 1: Comparison between Machine and Human



Deep Learning	Asset Pricing
Data Fully Used	Limited Memory / Limited Learning Time
Non-linear Regression	Slow in Information Processing
Data-driven Factors	Emotional in Trading
Hard to Replicate	Easy to Replicate
No Ground Truth	Based on Financial Theory
Uninterpretable	Economically Interpretable

Figure 2: Comparison between Deep Learning and Asset Pricing Models



Deep Learning	Traditional ML
Stronger Fitting Power	Weaker Fitting Power
Hierarchichy Generated Automatically	Hard to Frame Hierarchical Structure
Good at High-dimension Data	Hard to Deal with High-dimension Data
End-to-end Training Method	Model Specification Matters Much
Lack of Statistical Fundamental	Based on Statistical Theory
Uninterpretable	Geometrically Interpretable

Figure 3: Comparison between Deep Learning and Traditional Machine Learning Methods

3.1 Time-series Data

3.1.1 Stock Level

1. Percentage Change (Open-price Benchmark), the difference between close price and open price divided by open price
2. Percentage Change (Pre-close Benchmark), the difference between close price and close price of last trading day divided by close price of last trading day
3. Trend Deducted Percentage Change, percentage change (Pre-close Benchmark) subtracted by the market percentage change times beta coefficient
4. Free Turnover, stock turnover proportion of tradable stock
5. High Price Percentage Change, the difference between high price of the trading day and open price divided by open price
6. Low Price Percentage Change, the difference between low price of the trading day and open price divided by open price
7. Suspend Indicator, whether the stock is suspended or not of the trading day
8. Percentage Order Position, the rank of the percentage change of the stock among all the stocks in the market

3.2 Market Level

1. Percentage Change, shanghai composite index percentage return of the trading day
2. Turnover, shanghai composite index turnover
3. Equity Value Big-Small Factor, daily SMB factor acording to Fama French 3-factor model
4. PE High-Low Factor, daily PE HML factor
5. PB High-Low Factor, daily PB HML factor
6. Free Turnover High-Low Factor, daily turnover HML factor
7. Market-wide Variance, daily variance of percentage changes of stocks among the market

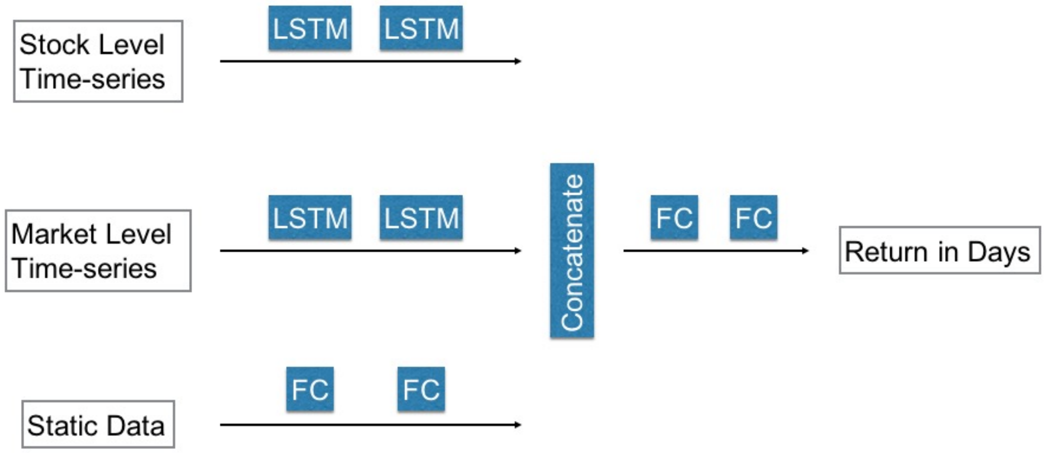


Figure 4: Network Structure

3.3 Static Data

3.3.1 Accounting Data

1. Equity Value, the tradable stock value of the stock
2. Price to Book-value, the price divided by book value per share
3. Price to Earning, the price divided by earning per share

3.3.2 Timing Data

1. Weekdays, Monday, Tuesday, ...
2. Weather, weather of the day in Beijing, Shanghai, Shenzhen
3. Temperature, temperature of the day in Beijing, Shanghai, Shenzhen

3.3.3 Segment Data

1. Industry, the industry segment of the stock
2. Geography, the province where the basis of the firm form
3. State-owned, whether the firm is owned by state or not

4 Model Structure

The deep learning model is constructed as Figure 4. Three sorts of data: 1) the stock level time-series data, 2) the market level time-series data, and 3) the static data are fed into model respectively. The first two sorts are connected with two layers of LSTM layer, which is one kind of recurrent neural network, usually used for time-series data. And the third sort, the static data is connected with two layers of fullly connected network. Then, the three streams of data are concatenated into one, and finally go through two fully connected layers with dropout layers to overcome the overfitting problem to the output layer.

5 Results and Discussion

5.1 The Backtest Performance

There are a lot of model are trained, I pick one representative model's backtest performance to give an illustration about the performance. This model is on the average level among all the models. Figure 5 is the comparison between the porfolio return in red and the SSEC Index return in green. The sharpe ratio of the portfolio is bigger than 2. Figure 6 is another illustration of the model, we can see that

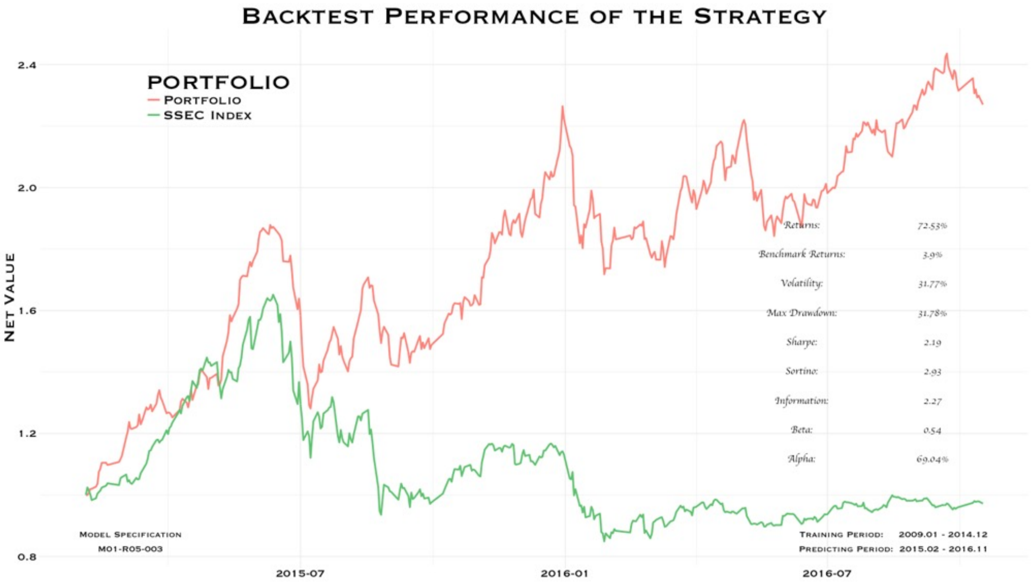


Figure 5: Backtest Performance of the Strategy

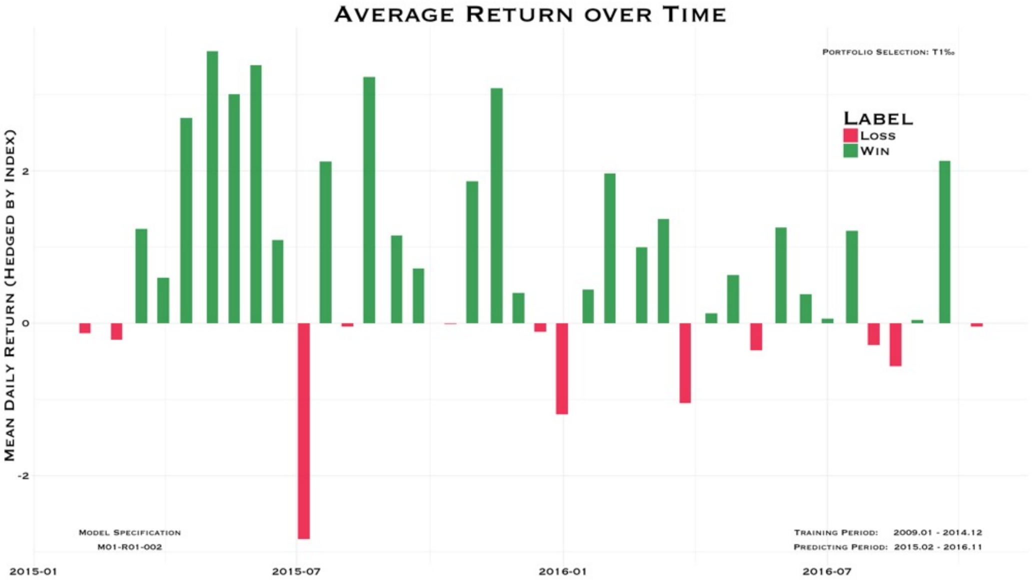


Figure 6: Average Return Over Time

in most periods of time, the return of the portfolio is larger than the index, and also there are some periods that the portfolio didn't perform well.

5.2 The Prediction Power

Two dimensions of the model are tested on their prediction power: 1) the percentile of the stocks to be choosed, and the 2) prediction period of time. In figure 7, there are several curves, from top to buttom, the porfolio are composed by the top 1 ‰, 2 ‰, 5 ‰, 1 %, 2%, 5 %, 10 % of the stocks which have the higher win rate. Value on x-axis is the amount of days after the prediction is given and the value of y-axis is the mean return of the porfolio hedged by the return of index. As we can see, the higher percentile we choose to compose the portfolio, the higher the return, and the closer the date the prediction given, the higher the return. So, we can have some conlusions which are not

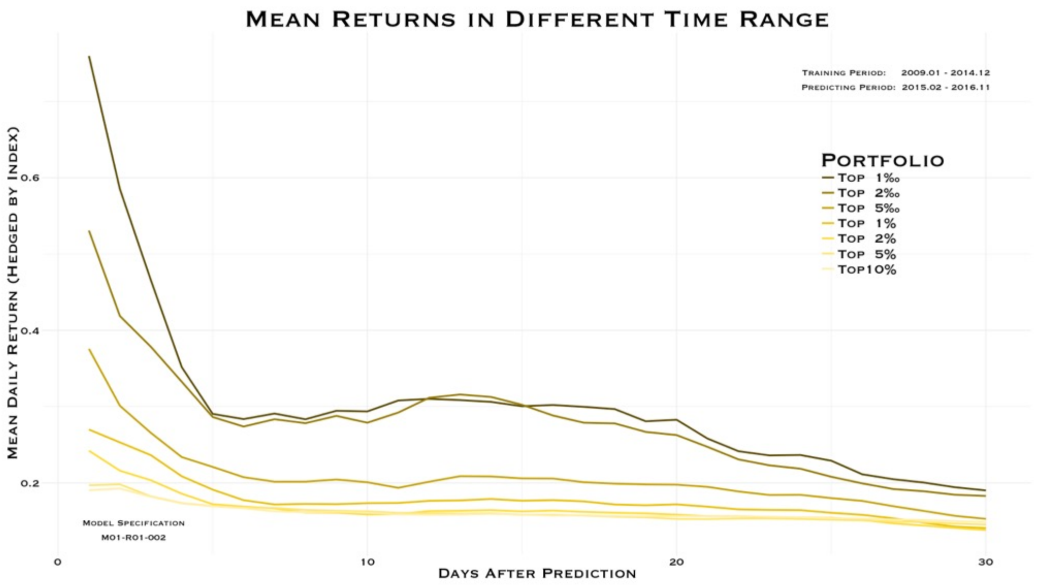


Figure 7: Mean Returns in Different Time Range

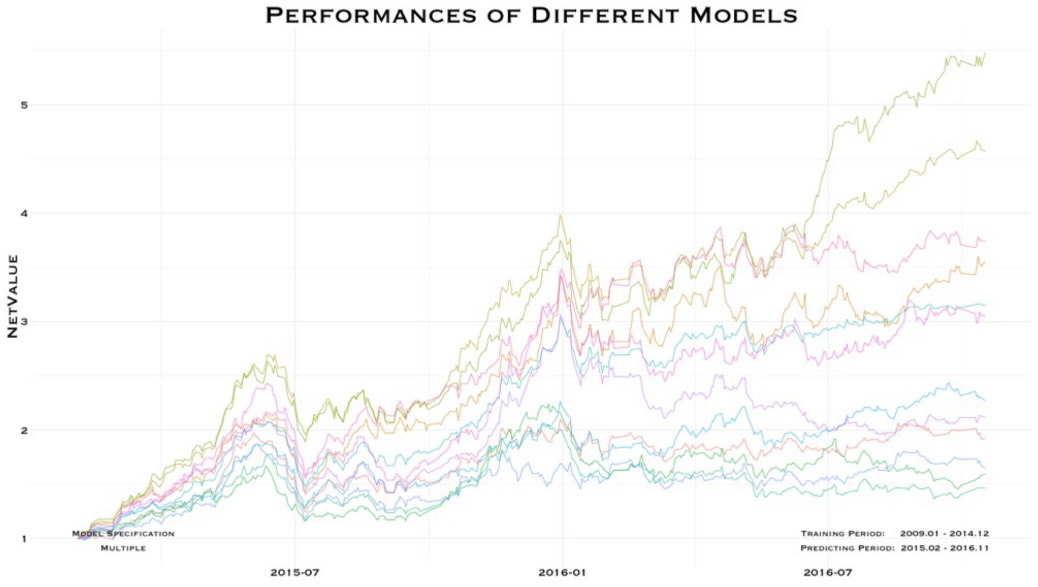


Figure 8: Perfomance of Different Models

very rigourous, 1) the deep learning model do have some prediction power, 2) the prediction power is restricted with a short period of time, in another word, the short-term return is easier to predict.

5.3 Hard to Control

As we all know that the deep learning model lack of explanation, which is just a black box and we do not know which factor affect the performance of the model. In Figure 5, there are some different portfolios constructed by different deep learning models. These models have same data and similar model structures but have very different performances in the backtest, or in statistical words, out sample test. So, we should still be cautious when using the stock prediction model using deep learning techneques.

6 Conlusions and Remarks

From the trails I have done in the project, I can get these rudimentary results:

- The deep learning model do have prediction power on chinese stock market
- The deep learning model will have less prediction power when the prediction period increase
- The deep learning model should be cautious used, because we can hardly tell when it will work, and when it will lose the prediction power
- The deep learning model can fit the data very well in in-sample test

And also, there are something we can further investigate:

- We can clean the data in deepth, including eliminating outlier samples, rectified the stock return according to the annual reports, or some other methods
- We can increase the data source, for example, we can introduce some sentimental data extrated from social network.
- We should futher invested the model and try to get control and try to understand the model.

References

[1] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance,* **25 (2), 383-417.**