Tokyo Stock Exchange Prediction with a Hybrid Model of Lightgbm and DNN

Yishuai Yang[1,a], Xuan Zhang[2,b], Shuyi Liu [3,c], Wenke Du[*4,d]

[a]yangyishuai0801@163.com, [b]xzhang9979@monroecollege.edu, [c]shuyi.liu@grenoble-em.com, [d]dwksdmndb@163.com

[1]Chongqing University of Posts and Telecommunications, Chongqing, China
[2]Monroe college, New Rochelle, United States
[3]Tongji University,Shanghai,China
[4]Renmin University of China, Beijing, China

Abstract—As stock investment has become an increasingly mainstream way of wealth management, researchers have increasingly attached importance to the study of stock price prediction, and constantly used a variety of methods to predict its price trend. In this paper, we pay attention to the JPX Tokyo Stock Exchange Prediction. The dataset is provided by Kaggle platform. We hybrid LightGBM and DNN to predict the stock price. Sharpe Ratio is our evaluation metrics. The results show that our hybrid model owns the best performance with the highest Sharpe Ratio score 0.152, which is 0.041, 0.032, 0.004 higher than Xgboost, Lightgbm and DNN respectively.

Keywords: JPX Tokyo Stock Exchange, investment, LightGBM, DNN, Sharpe Ratio

1. INTRODUCTION

As stock investment has become an increasingly mainstream way of wealth management, researchers have increasingly attached importance to the study of stock price prediction, and constantly used a variety of methods to predict its price trend. In the traditional methods to solve asset pricing problems, most of the asset pricing models used are linear models, with strict assumptions, weak explanatory power and insufficient effectiveness. In the era of artificial intelligence, cutting-edge algorithms such as machine learning and deep learning can be used to process and analyze financial data, excavate potential nonlinear relationships, and continue to explore and study problems that cannot be solved by traditional financial models.

At present, among the mainstream research methods for stock price forecasting, machine learning methods have improved the accuracy better than traditional methods, but they are still insufficient. The deep learning model is good at capturing the nonlinear relationship between factors and stock price trends, and has better adaptability to complex and high noise data. It shows high accuracy and superiority in stock price forecasting, and has certain research value.

In this paper, we pay attention to the JPX Tokyo Stock Exchange Prediction. In the following part, we firstly introduce related work on the Quantitative investment. In the

MSEA 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2334222

section III and section IV, we describe our methods and experiments. In the final part, we conclude our work and put forward our improvement expected.

2. RELATED WORK

There are some limitations in the use of linear models in traditional multi factor stock selection. When the number of selected factors increases, there may be some factors that have nonlinear correlations with stock returns. Linear models cannot accurately describe this nonlinear relationship. Machine learning algorithm can effectively deal with a large number of data, and is good at mining the nonlinear relationship between data, so it was introduced into the field of quantitative investment.

Up to now, some classical algorithms have been applied to quantitative stock selection by many scholars. The more common algorithms include random forests, support vector machines, logical regression, naive Bayes and deep learning algorithms represented by neural networks have made great progress in the field of quantification.

[1] used the support vector machine algorithm to classify stocks and forecast the stock price trend, which proved that the support vector machine algorithm is effective in forecasting the stock price trend and has high accuracy.

[2] used the prediction model constructed by the regression algorithm to predict the short-term change trend and the medium and long-term change trend of the stock index, and the empirical results show that the model constructed by the regression algorithm can also achieve certain effects in the trend prediction of stock indexes. [3] combines the SVM algorithm in the financial time series prediction model, and the results show that the stability of the model is effectively improved after adding the SVM algorithm. [4] introduced genetic algorithm into the investment model based on neural network, which greatly improved the accuracy of the model. [5] proposed a DeepLOB network for high-frequency order book data in the day, which extracts interactive information through convolution layers. [6] combines image analysis technology, and uses stock trend images and time series data to effectively predict.

3. METHODS

- LightGBM

LightGBM [7] model is a powerful development in boosting set model. Its basic principles are the same as those of GBDT, XGBoost and other models. They first derive and calculate the negative gradient of the objective function, replace the residual of the current tree model with the calculated value, and then generate new tree model fitting data again to predict the residual. In this way, they keep approaching the true value, and finally achieve the goal of reducing the model error. However, because the model is innovative in many aspects, it will perform better than XGBoost [8] model. The specific innovations and characteristics of the model are as follows:

(1) The decision tree algorithm based on histogram no longer uses the pre-sorting method in XGBoost model. After Histogram calculation, it can improve the stability of the model, speed up the running speed and reduce the memory space occupation.

(2) Sampling data through GOSS algorithm. In the face of large data sets, in order to improve the training speed and efficiency, we can use the method of weighting and sampling. LightGBM uses the gradient based unilateral sampling method (GOSS) to sample, sort the samples according to the gradient size, select the values with larger gradient of the first a%, then randomly select b% of the training samples from the samples with smaller gradient of (1-a%), and finally select a% * # samples + b% * (1-a%) * # samples as the training samples. This data collection method can keep the data sampling approximately consistent with the original distribution, and can ensure that samples with smaller gradient can be trained.

(3) The independent feature merging (EFB) method is used to reduce the dimension of features and improve the calculation efficiency of the model. Due to the unique hot coding of features, it is easy to lead to problems such as high dimension of features and sparse features. In general, features that can be bundled and merged must meet the mutually exclusive conditions, so that the two features will retain their original information after merging. At the same time, considering that some features are not completely mutually exclusive, the model can use conflict ratio to calculate the correlation degree of features. When the conflict ratio is small, features can be bundled.

(4) Before that, most decision tree algorithms used the 'Level wise' idea to grow the decision tree. That is, each time when splitting, all leaf nodes in the same layer were split, but not all leaf nodes needed to be split. The information gain of some leaf nodes was very small, and splitting would only increase the waste of computing resources.

- Deep neural network (DNN)

Deep neural network (DNN) [9] is a framework of deep learning, which is a neural network with at least one hidden layer. Similar to the shallow neural network, the deep neural network can also provide modeling for complex nonlinear systems, but the extra levels provide a higher level of abstraction for the model, thus improving the ability of the model. In our paper, we utilize DNN for quantitative investment.

- Hybrid model

The structure of whole model is shown in figure 1.

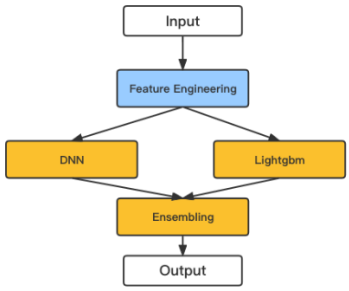

Figure 1: hybrid model structure

4. EXPERIMENTS

- Experimental Data

Our dataset is provided by Japan Exchange Group, Inc. (JPX), which is a holding company operating one of the largest stock exchanges in the world, Tokyo Stock Exchange (TSE), and derivatives exchanges Osaka Exchange (OSE) and Tokyo Commodity Exchange (TOCOM). This dataset contains historic data for a variety of Japanese stocks and options.

The files we use and their descriptions are show in table 1.

Table1. Dataset description

stock_prices.csv	The core file of interest. Includes the daily closing price for each stock and the target column.
options.csv	Data on the status of a variety of options based on the broader market. Many options include implicit predictions of the future price of the stock market and so may be of interest even though the options are not scored directly.
secondary_stock_prices.csv	The core dataset contains on the 2,000 most commonly traded equities but many less liquid securities are also traded on the Tokyo market. This file contains data for those securities, which aren't scored but may be of interest for assessing the market as a whole.
trades.csv	Aggregated summary of trading volumes from the previous business week.
financials.csv	Results from quarterly earnings reports.
stock_list.csv	The number of shares on the most/second most competitive buy level.

The stock price contains 12 columns, which is the core input of our model. Figure 2 shows the change in stock returns over time with different stock codes. Table 2 shows the main features.

Table2. Main features

RowId	Date
SecuritiesCode	Open
High	Low
Close	Volume
AdjustmentFactor	ExpectedDividend
SupervisionFlag	Target

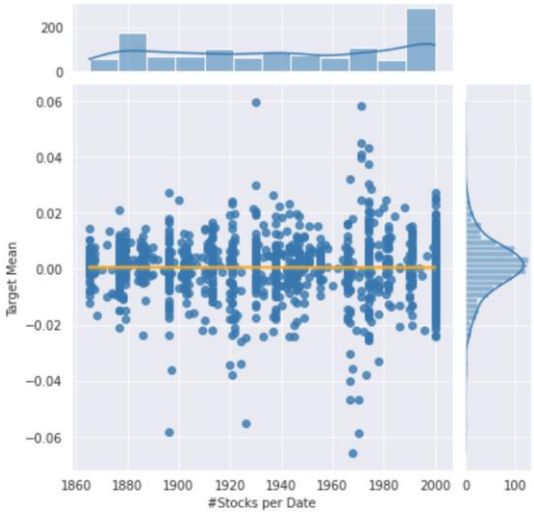

Figure 2. Target Distribution versus per Date

- Feature engineering

Use DNN to perform feature engineering for two feature types. One is normal feature, and the other is ID feature. Normal features include some Open, High, Low, Close, Volume AdjustmentFactors and their corresponding derived features. For ID characteristics, compress the securityCode into 4 dimensions. The process is shown in figure 3.

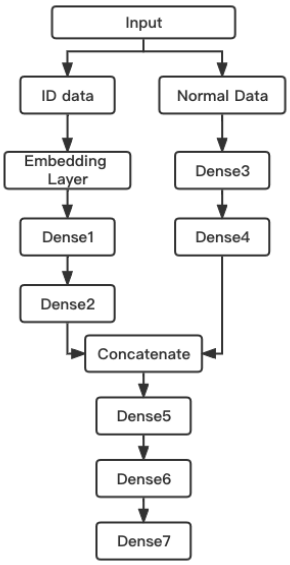

Figure 3: Feature Engineering

- Experimental settings and Experimental results

Our experimental settings for DNN are shown in the following table 3, we train our model using Pytorch.

Table 3: experimental settings

Epoch	5
batch size	2048
Adam	0.05

To compare our model with other models, we evaluate our result using the Sharpe ratio [10]. Sharp ratio in the financial field measures the performance of an investment (such as securities or portfolios) relative to risk-free assets after adjusting its risk. It is defined as the expected value of the difference between investment return and risk-free return, divided by the investment standard deviation (i.e. its volatility). It represents the additional return of each unit of risk that the investor additionally bears.

$$S_a = \frac{\mathrm{E}[R_a - R_b]}{\sigma_a} = \frac{\mathrm{E}[R_a - R_b]}{\sqrt{var[R_a - R_b]}}$$

The experimental results of competing models and our model are shown in table 4.

Table 4: performance of different models

Models	Sharpe Ratio
Xgboost	0.111
Lightgbm	0.120
DNN	0.148
Hybrid model	0.152

From Table 4, we can see that our hybrid model owns the best performance with the highest Sharpe Ratio score 0.152, which is 0.041, 0.032, 0.004 higher than Xgboost, Lightgbm and DNN respectively.

5. CONCLUSION

As a classical and challenging problem, the prediction of stock market has attracted the attention of economists and computer scientists. In order to establish an effective prediction model, linear and machine learning tools have been explored in the past decades. Recently, deep learning has been introduced as a cutting-edge technology in this subject and has developed rapidly. In this paper, we focus the JPX Tokyo Stock Exchange prediction in this paper. In the following part, we firstly introduce related work on the Quantitative investment. In the section III and section IV, we describe our methods and experiments. In the final part, we conclude our work and put forward our improvement expected. Our hybrid

model owns the best performance with the highest Sharpe Ratio score 0.152, which is 0.041, 0.032, 0.004 higher than Xgboost, Lightgbm and DNN respectively.

ACKNOWLEDGEMENT

Thanks to the Kaggle platform, we could advance our research in the long period and produce this paper documenting the work. And due to Yishuai Yang and Xuan Zhang's encouragement, the work could finish efficiently. Liu Shuyi and Wenke Du do some experiments and calculate sharpe ratio of each models.

REFERENCES

[1] Hu Z, Liu W , Jiang B , et al. Listening to Chaotic Whispers: A Deep Learning Framework oriented Stock Trend Prediction[J]. Papers, 2019.

[2] Kim T, Kim H Y, Hernandez Montoya A R. Forecasting stock prices with a feature fusion L model using different representations of the same data[J]. PLoS ONE. 2019, 14(2):255-267.

[3] Long W, Lu z, Cui L. Deep learning -based feature engineering for stock price movement Knowledge-Based Systems. 2019, 164(JAN.15): 163-173.

[4] Eapen J, Bein D, Verma A. Novel Deep Learning Model with CNN and Bi-Directional Improved Stock Market Index Prediction[C] 11 2019 IEEE 9th Annual Computing and Com Workshop and Conference (CCWC). IEEE, 2019

[5] K. c, Peng, And, et al. Multiple-time scales analysis of physiological time series under neural Physical A: Statistical Mechanics and its Applications. 1998, 249(1-4): 491-500.

[6] Papadimitriou S , Yu P S . Optimal multi-scale patterns in time series streams[C]/ Proceed ACM SIGMOD International Conference on Management of Data, Chicago, Illinois, USA, : 2006. ACM, 2006.

[7] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[J]. Advances in neural information processing systems, 2017, 30.

[8] Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting[J]. R package version 0.4-2, 2015, 1(4): 1-4.

[9] Shah D, Campbell W, Zulkernine F H. A comparative study of LSTM and DNN for stock market forecasting[C]//2018 IEEE international conference on big data (big data). IEEE, 2018: 4148-4155.

[10] Sharpe W F. The sharpe ratio[J]. Streetwise–the Best of the Journal of Portfolio Management, 1998: 169-185.