JPX Tokyo Stock Exchange Prediction with LightGBM

Mingda Huo[1,a], Sen Wang[2,b], Tianxiao Xu[3,c], Daniel Boxiao Huang[4,d], Tong Zhou[5,e*],

sydney678@stu2019.jnu.edu.cn[a], wangsen01@zybank.com.cn[b], tx19718n@pace.edu[c], HumbleBeyondX@gmail.com[d], tongzhoufuture@gmail.com[e*]

Jinan University, Guangzhou, China[1]
ZhongYuan Bank Co,LTD, Zhengzhou, China[2]
Pace university, New York, United States[3]
Beijing Normal University-Hong Kong Baptist University United International College (UIC), Zhuhai, China [4]
Johns Hopkins University, Baltimore, United States [5]

Abstract— The stock market is in an environment where risks and rewards coexist. A well-designed stock portfolio can make profits or even windfall profits through trading. Stock prediction refers to the use of scientific methods, using mathematical statistical methods as the means. In this paper, we pay attention to the JPX Tokyo Stock Exchange Prediction. The dataset is provided by Kaggle platform. We first do feature engineering and extract import features for model input. We use the LightGBM 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 0.288, which is 0.023, 0.098, 0.048 higher than Xgboost, SVM and DNN respectively.

Index Terms— JPX Tokyo Stock Exchange, investment, LightGBM, Sharpe Ratio

1. INTRODUCTION

The stock market is in an environment where risks and rewards coexist. A well-designed stock portfolio can make profits or even windfall profits through trading. The market is also vulnerable to various factors. In addition to the company's internal operating conditions, the overall development status and prospects of the industry and the level of regional economic development, it will also be affected by external interest rate levels, balance of payments, policy situation and even political and regional conflicts. Complex information is mixed in various stock data.

Stock prediction refers to the use of scientific methods, using mathematical statistical methods as the means, based on accurate data and data information, fully understanding the history of the stock market, starting from the law of development, mining useful information from the existing large amount of data to make an analysis and prediction of the future development trend of the stock.

ICBBEM 2023, May 19-21, Hangzhou, People's Republic of China

DOI 10.4108/eai.19-5-2023.2334371

LightGBM algorithm looks at the relationship between factors and stock returns from the perspective of classification and establish a learning model to study the effectiveness and robustness of factors. By constructing the stock selection model, the performance of the stock selection model in terms of income, risk and investment cost performance is quantified.

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 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

Ensemble learning is a model framework that combines multiple weakly supervised models according to a certain strategy to complete a task together and reduce prediction bias and variance. It is not only favored by many scholars, but also widely used in various fields. Chen T and Guestrin C [1] proposed a new integrated learning algorithm Xgboost. It improves efficiency based on gradient lifting algorithm.

Zhang [2] uses Adaboost algorithm to predict the future trend of stocks, which proves that integrated learning has a prominent performance in practice. Li [3] applies dynamic Xgboost algorithm in the field of stock selection to periodically divide factors and screen and evaluate stocks based on it, which solves the problems of poor timeliness of traditional strategies and difficult to determine the proportion of characteristics. Lohrmann [4] selected the random forest algorithm to divide stocks from the perspective of classification, optimized the trading strategy of traditional experiments, and formulated a unique position building and position adjustment scheme. Zhang [5] built a combination strategy integrating Adaboost algorithm, probabilistic support vector machine (Ps V M) and genetic algorithm (GA) and other algorithms to classify the stock inflection point trend and predict its future changes, and effectively solve the imbalance problem of stock inflection point classification. Lightgbm [6] is a systematic implementation of the Gradient Upgrading Decision Tree (GBDT) proposed by Microsoft in 2017. Ke g, Meng Q [7] and others used common data sets for experiments. The results show that Lightgbm algorithm can improve the training speed by more than 20 times on the premise of ensuring the same accuracy as traditional GBDT algorithm. Zhou [8] first applied the cascade integrated learning architecture in the field of index tracking and simulation. The architecture cascaded the logical regression model LR to the gradient enhancement decision tree (GBDT) model, and simulated the position building and position changing, which proved the effectiveness of the algorithm in real market transactions.

3. METHODS

We use LightGBM for stock prediction. LightGBM is a new boosting framework model proposed by Microsoft, is an improved lightweight Gradient Boosting algorithm. The core idea of the algorithm is to reduce the number of split points, samples and features and control the complexity of the model, so as to give consideration to efficiency and accuracy.

Compared with the traditional GBDT algorithm, LightGBM algorithm has two relatively core lifting points, namely, gradient based unilateral sampling technology and independent feature merging technology.

Gradient-based One - Side Sampling is a new sampling method based on GBDT algorithm, which can achieve a good balance between reducing the number of data instances and maintaining the accuracy of the decision tree, thus effectively improving the generalization ability of the model. Generally speaking, the value of data sampling information will decrease with the decrease of gradient. Therefore, the core of GOSS technology is to sample data by reserving a few large gradient data. Through the above strategy, the algorithm model will pay more attention to the insufficiently trained instances during training, and will not change the distribution of the original data too much.

The core of the independent feature binding (EFB) technology is to reduce the dimensions of the original data. High-dimensional data is usually very sparse. The sparsity of feature space provides us with the possibility of designing an almost lossless method to reduce the number of features. Specifically, in the sparse feature space, many features are mutually exclusive. In other words, their values will not be taken as non-zero values. Therefore, we can reasonably bind exclusive features into a single feature, that is, exclusive feature binding. By sampling the carefully designed feature scanning algorithm, the model can build the same feature histogram from the feature bundle as the single feature, which can greatly speed up the training speed of the GBDT algorithm model without affecting its accuracy.

In terms of improving the generalization ability of the algorithm model, LightGBM algorithm uses the histogram algorithm, uses a table for each feature to record non-zero data, and builds a histogram based on it. The algorithm can obtain the best splitting point for a series of basic functions. Secondly, LightGBM algorithm uses a special strategy for leaf growth, which is different from other strategies in that it limits the depth and avoids the additional resource cost caused by the low splitting gain of the traditional strategy.

The whole structure of our method is shown in figure 1. First, we do feature engineering and extract the volatility feature, the moving average feature and growth feature. Then we feed them into the LightGBM.

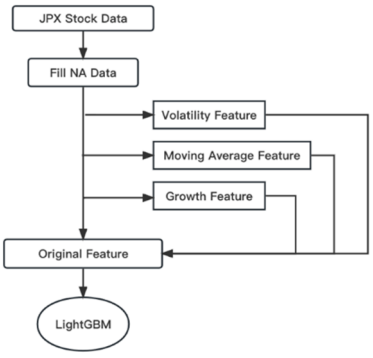

Figure 1. Lightgbm structure

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.

Table 1: Dataset Files

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: Important Input

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

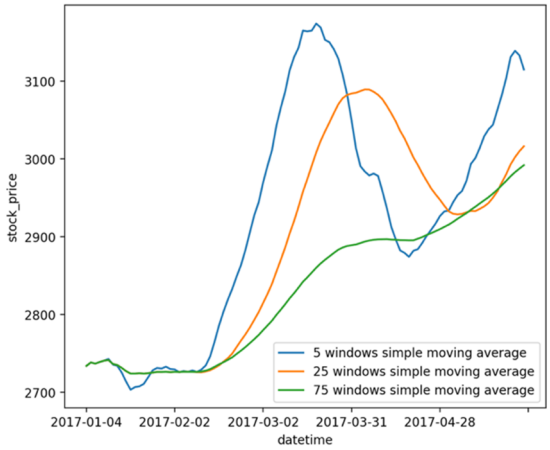

Figure 2: Input

- **Feature engineering**

We process the input into important features and feed it into model. The important features we extracted and their descriptions are as follows.

Table 3: Important Features

return_1month	The month-on-month growth rate of the closing price in the past month
return_2month	The month-on-month growth rate of the closing price in the past two months
return_3month	The month-on-month growth rate of the closing price in the past three months
volatility_1month	Volatility in the past month
volatility_2month	Volatility in the past two months
volatility_3month	Volatility in the past three months
MA_gap_1month	Moving average in the past month
MA_gap_2month	Moving average in the past two months
MA_gap_3month	Moving average in the past three months

- **Experimental settings**

Our experimental settings for LightGBM are shown in the following table 4, we train our model using Pytorch.

Table 4: Experimental Settings

learning_rate	0.05
metric	None
'objective'	'regression'
'boosting'	'gbdt'
'verbosity'	0
'n_jobs'	-1
force_col_wise	True

- **Experimental results**

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.

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

The experimental results of competing models and our model are shown in table 5.

Table 5: Performance of Different Models

Models	Sharpe Ratio
Xgboost	0.265
SVM	0.190
DNN	0.240
LightGBM	0.288

From Table 5, we can see that our LightGBM model owns the best performance with the highest Sharpe Ratio score 0.288, which is 0.023, 0.098, 0.048 higher than Xgboost, SVM and DNN respectively.

5. CONCLUSION

Stock returns are closely related to many aspects, such as macroeconomic policy, economic development index, industry development prospects, corporate financial data and investor sentiment. These indicators that affect stock factors are also known as stock factors. How to make effective use of these factors and then provide reference for the stock selection

model with high efficiency and high yield has become the focus of research. In recent years, factor-based stock trend prediction algorithms have been proposed continuously.

In this paper, we focus the JPX Tokyo Stock Exchange prediction in this paper. In the section I, 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 0.288, which is 0.023, 0.098, 0.048 higher than Xgboost, SVM 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.

REFERENCES

[1] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]/Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794

[2] Guoying Z, Ping C. Forecast of yearly stock returns based on Adaboost integrational gorithm[C] /2017 IEEE International Conference on Smart Cloud(Smart Cloud). IEEE, 2017263-267

[3] Jidong L, Ran Z Dynamic weighting multi-factor stock selection strategy based on Xgboostmachine learning algorithm[C]/2018 IEEE International Conference of Safety ProduceInformatization(IICSPI), IEEE, 2018: 868-872

[4] Lohrmann C, Luukka P Classification of intraday S&P500 returns with a Random Forest[J]International Journal of Forecasting, 2019, 35(1): 390-407

[5] Zhang X, Li A, Pan R Stock trend prediction based on a new status box method and Ada Boostprobabilistic support vector machine[J]. Applied Soft Computing, 2016, 49: 385-398

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

[7] Ke G, Xu Z, Zhang J, et al. Deep GBM: A deep learning framework distilled by GBDT for online prediction tasks[C]/Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining 2019: 384-3941

[8] Zhou F, Zhang Q, Sornette D, et al. Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices[J]. Applied Soft Computing, 2019, 84: 105747