Stock Price Prediction Based on XGBoost and LightGBM

Yue Yang[(1st)1,a], YangWu[2,b], Peikun Wang[(1st)3,c], Xujiali[4,d]

[1]Southwest Minzu University, Chengdu, China
[2]JinagXi University Of Fianance and Economics Jiangxi, China
[3]South China University of Technology GuangZhou, China
[4]Shanghai University of Finance and Economics Shanghai, China
Yue Yang and Peikun Wang are co-first author.

Abstract. Stock trading, as a kind of high frequency trading, generally seeks profits in extremely short market changes. And effective stock price forecasting can help investors obtain higher returns.Based on the data set provided by Jane Street, this paper makes use of XGBoost model and LightGBM model to realize the prediction of stock price.Since the given training set has a large amount of data and includes abnormal data such as missing value, we first carry out feature engineering processing on the original data and take the mean value of the missing value, so as to obtain the preprocessed data that can be used in modeling.
The experimental results show that the combined model of XGBoost and LightGBM has better prediction performance than the single model and neural network.

1 Introduction

Stock price prediction refers to the prediction of the trading operations at a certain time in the future.It is based on the historical and real data of the stock market according to a certain forecasting model. This prediction plays an important and positive role in improving the efficiency of the trading market and giving play to market signals. Accurate stock price forecast can help investors adjust their trading strategies in time, and effectively avoid investment risks, so as to obtain higher returns.

Price prediction has long appeared in all kinds of trading markets. However, due to the influence of many factors, including not only the internal change rules of the stock market, but also the sudden impact of the external market, the prediction results of some existing stock price prediction models are not perfect. Using the existing technology and the improvement of the existing algorithm, the prediction result can be closer to the actual situation.

Therefore, we need to further improve the algorithm and model, make use of the historical data given, and extract valuable data information to achieve more accurate stock price prediction.

1.1 Related Work

For this kind of price prediction problem, regression model can be used for modeling and solving. Firstly, the model is used to fit the given historical data, and a model with a high degree of fitting is obtained.Then use the last part of the data to detect and verify the model, adjust and optimize the model.Finally, according to the given data, the model is used to predict the results.

In [1], the machine learning algorithm of gradient-enhanced decision tree is mainly introduced. It estimates all possible division points of information gain by scanning all data instances, but this method has a high time complexity. This article introduces the LightGBM model that combines GOSS and EFB technologies. It eliminates data instances with small gradients and bundles mutually exclusive features to improve the accuracy and efficiency of the model.

[2] and [3] mainly used XGboost and LightGBM algorithms to predict and analyze the sales volume of product sales data sets, study the principles of XGboost and LightGBM algorithms, fully analyze prediction objects and conditions, and compare algorithm parameters and data set features.

[3] is on the basis of XGBoost, a new sparse data perception algorithm and a weighted quantile approximate tree learning algorithm are proposed, thereby constructing a scalable number enhancement system, which can be expanded from fewer resources Lots of examples.

[4] and [6] proposed a prediction model based on convolutional neural network and LightGBM. First, a new feature set was constructed by analyzing the characteristics of the original data in the time series, and then CNN was used to obtain information from the input data. , Adjust the network parameters by comparing with the actual results; then integrate the LightGBM model into the above model to further improve the accuracy and robustness of the prediction.

[7] and [8] mainly introduced the use of neural networks and SVM to predict stock prices. As a data mining technology widely used in commercial areas, artificial neural networks have the ability to learn and

[a]tracyyyang@gmail.com, [b]wuyangwebdeveloper@163.com
[c]peikunwang45@gmail.com, [d]zhenlanda@126.com

detect the relationship between nonlinear variables. The SSA method decomposes stock prices into characteristic data with different time ranges and different economic characteristics, and then introduces these characteristics into SVM for price forecasting. The results show that the above methods have good forecasting capabilities.

1.2 Our Contribution

The data provided in this paper mainly contains the real stock market data from Jane Street composed of several anonymous characteristics, including date, operation value and other specific information. Each line of data represents a trading process.We need to use this historical data to predict what trades will take place at a tradable point in time.Usually, the data given by the material can not be used directly as the original data, and the data need to be preprocessed before modeling. This paper mainly carries out feature engineering processing on the data, and converts the data into data that can be used directly in modeling.

The rest of this paper is organized as follows: The second part introduces the feature engineering and stock price data, and mainly preprocesses the original data.The third part introduces the concrete implementation of the XGBoost model.

The fourth part gives the experimental results and makes a concrete analysis and comparison to them.

Finally, the fifth part is the summary of the overall results and the future work.

2 FEATURE ENGINEERING

In this section, we will focus on data sets and feature engineering.

The data set in this paper is mainly from the real data of the stock market provided by Jane Street, and there are three CSV files related to the data.

Where, the train. CSV file is the training data set, which contains the historical data of the trading market and the returned results.The features.csv file is some metadata related to anonymous features.And the example_test.csv file is a set of mock tests.

Since the data set provided by the material has a large amount of data and many attributes, we need to select appropriate features to preprocess the data, so as to reduce the calculation cost of the model, reduce the noise interference and improve the training efficiency.

Using feature engineering can produce data from raw data that can be used for modeling while reducing data redundancy and dimensionality.

Commonly used feature engineering methods include timestamp processing, discrete variable processing, partitioning processing, cross feature, feature selection, feature scaling and feature extraction, etc., which can be used for feature preprocessing of time data, discrete data and continuous data respectively.

Among them, the timestamp processing is mainly to separate the data into multiple dimensions, such as seconds, minutes, hours, days, months, years, etc., and at the same time to standardize the data processing.

Discrete variable processing is generally the unique heat encoding of data, such as transforming character data into floating point data, so as to meet the input requirements of random forest model.

Partitioning operations are usually used to partition contiguous data, dividing the data into certain blocks according to a certain range, and then proceeding to the next operation.

Cross feature refers to the combination of two or more category attributes into one attribute to obtain a better combination of features than a single feature.

Feature selection refers to the use of a certain algorithm to sort the features, select the features of higher importance, so as to achieve the purpose of reducing noise and redundancy.

Feature scaling is applicable to the situation that one feature has a much higher span value than other features, which can effectively avoid the weight value of some features with a large numerical difference.

Feature extraction refers to a series of algorithms that automatically obtain some new feature sets from original attributes, commonly used including principal component analysis, clustering algorithm, etc.

In addition, for the processing of data with missing values, we choose to use the mean value to fill in the missing values to prevent anomalies in the subsequent data processing.

After processing, the feature distribution of feature_1 is shown in Figure 1.

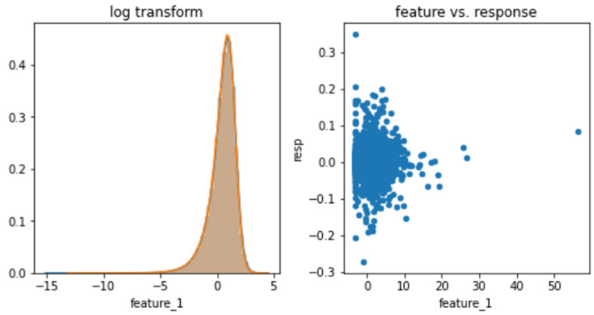

Figure 1. The Feature Distribution of feature_1.

3 XGBoost model

XGBoost makes some improvements to Boosting algorithm on the basis of GBDT model, and its internal decision tree uses regression tree.Boosting algorithm's inherent idea is to group several weak classifiers together to form a strong classifier.

XGBoost, as a decision tree promotion model, integrates several tree models together to form a strong classifier.A tree is grown by constantly adding trees, constantly splitting features, and each time a new tree is added, the residuals obtained from the previous training are fitted again.Finally, the scores corresponding to several trees obtained by training are added up to be the final predicted value of the sample.

The main advantages of the XGBoost model include a regular term in the target function to prevent overfitting.Its parallel optimization is on the feature granularity, which is more efficient.Supporting column

sampling can not only reduce the over-fitting, but also reduce the amount of calculation.Considering that the training data is sparse, the default direction of the branch can be specified for some values.

4 LIGHTGBM model

LightGBM, as a framework to implement GBDT algorithm, combines GOSS algorithm and EFB algorithm and is used in data sets with large sample data and high dimensions.

Its optimization features include Leaf-Wise based decision tree growth strategy, optimal segmentation of category eigenvalues, feature parallelism and data parallelism.It not only reduces the communication overhead and time complexity between data, but also ensures the rationality of data.

The main advantages of LightGBM model include its relatively fast training speed, low memory consumption, high accuracy, distributed support, and the ability to process large amounts of data quickly.

We also use HyperOpt tool to debug the parameters of the above machine learning model.Typically, a machine learning engineer or data scientist will perform some form of manual tuning (grid search or random search) for a small number of models (such as decision trees, support vector machines, and k-nearest neighbors), then compare the accuracy rates and choose the best one to use.However, this approach usually selects the optimal model, because when we find the optimal parameters of one model, we may miss the optimal parameters of another model. HyperOpt can perform serial and parallel synchronization optimization in the search space and adjust the parameters through Bayesian optimization, so the parameter tuning speed is faster and the effect is better.

5 Experiments

In order to verify the effectiveness of the model, we chose XGBoost single model, LightTGBM single model, their fusion model and neural network for comparison under the same data set and evaluation index.

In this article, we evaluate in terms of a utility score, each row in the test set represents a trade opportunity, and we need to predict the value of the operation, where 1 represents trade and 0 represents pass.

The evaluation index calculation formula we choose is as follows:

$$p_i = \sum_j (weight_{ij} * resp_{ij} * action_{ij})$$

$$t = \frac{\sum p_i}{\sqrt{\sum p_i^2}} * \sqrt{\frac{250}{|i|}}$$

Where for each date i, each trade j will correspond to the associated weight and resp, and $|i|$ is the number of unique dates in the test set.The utility is then defined as:

$$u = \min(\max(t,0),6)\sum p_i$$

Table 1 shows the prediction results of the four models. According to the results, the utility value of the Proposed Model was 7852. Compared with the other three models, the Model obtained by 1:1 fusion of XGBoost and LightTGBM had more obvious advantages in stock price prediction.

Table1. The Prediction Results of the Four Models.

Models	**Unity**
Proposed Model	**7852**
Xgboost	6008
Lightgbm	7487
Nerual Network	5019

In addition, we also obtained the average rate of return for resp_2 by date, as shown in Figure 2.

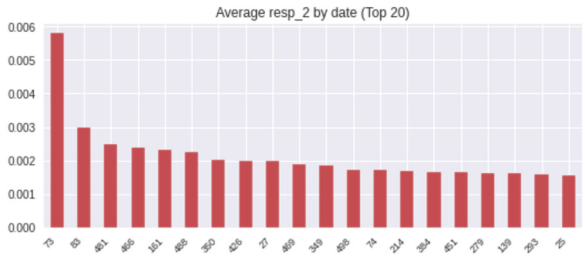

Figure 2. The Average Rate of Return for resp_2 by Date.

6 conclusions

In this paper, we propose a stock price prediction method based on XGBoost and LightTGBM models. Firstly, we preprocess the original data with feature engineering, and conduct average processing for the missing values.Then the predicted results of three single models and a combined model were compared. Among them, the utility value of the Proposed Model was the highest (7852), the utility value of XGBoost was 6008, the utility value of LightGBM was 7487, and the utility value of Nerual Network was 5019. It is found that the XGBoost and LightGBM model with 1:1 fusion has better prediction effect than the single model and the neural network model.

References

1. Thomas Finley.LightGBM: A Highly Efficient Gradient Boosting Decision Tree.2011.
2. Xingyuan Tang.A Blending Model Combined DNN and LightGBM for Forecasting the Sales of Airline Tickets.2018.
3. Khaula Qadeer, Moongu Jeon.Prediction of PM10 Concentration in South Korea Using Gradient Tree Boosting Models.2017.
4. Chen Tianqi,Guestrin,Carlos. XGBoost:A Scalable Tree Boosting System.2016.

5. Ju Yun,Sun Guangyu.A Model Combining Convolutional Neural Network and LightGBM Algorithm for Ultra-Short-Term Wind Power Forecasting.2019.
6. Chen Cheng,Zhang Qingmei.LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion.2019.
7. Thanh Tung Khuat.An Application of Artificial Neural Networks and Fuzzy Logic on the Stock Price Prediction Problem.2017.
8. Wen Fenghua,Xiao Jihong.Stock Price Prediction based on SSA and SVM.2014.