Stock Volatility Prediction Based on 1D-CNN and LightGBM

Ningyun Dan[1a], Yuxin Li[1b], Zimo Nie [3c], Yuan Li[d*]

[a]3368709754@qq.com

[b]yl3924@columbia.edu

[c]zimo_nie@126.com

[d*]li13305272907@163.com

[1a]Zhaotong University, Ningyun Dan, China

[1b]Columbia University New York, United States

[3c]Beijing Language and Culture University Beijing, China

[d*]Nanjing University of Posts and Telecommunications, Nanjing, China

Abstract—Stock price fluctuations often bring opportunities to investors. Predicting the trend of stock price fluctuation effectively can bring effective and feasible suggestions to investors. This paper uses the real data of the stock market as the data set, and adopts the fusion model based on 1D-CNN model and LightGBM model to predict the stock fluctuations. We first preprocess the original data and extract the important information in the data. Then we train the model and the prediction results are obtained. Experimental results show that the prediction performance of 1D-CNN and LightGBM fusion model is better than that of naive Bayes model and single XGBoost and LightGBM model.

Keywords—Stock Volatility, CNN, LightGBM

1 INTRODUCTION

Predicting the price fluctuation of the stock market can help investors better understand and grasp the operation law of the stock market, stock price fluctuation and its influence mechanism and degree on the real economy. This kind of forecast is generally based on some historical real data, according to the forecast model to predict the future fluctuation. Predicting the trend of stock price fluctuation effectively can bring effective and feasible suggestions to investors.

Because the price fluctuation of stock market is affected by many factors, it contains very complex information. Deep learning provides powerful tools for capturing these information. Therefore, we use deep learning modeling to obtain models with stronger predictive power.

ICIDC 2022, June 17-19, Qingdao, People's Republic of China

DOI 10.4108/eai.17-6-2022.2322733

1.1 Related Work

Predicting the price fluctuation of stock market is a very important and valuable problem, so many works and models have been carried out to study it. The 1D-CNN model and LightGBM model have also been improved for many times and applied to a variety of scenarios.

Both [3] and [4] present a prediction model based on the improved LightGBM. [3] mainly proposed the protein prediction model related to fertility. Firstly, the initial feature space is constructed by combining various data information, and then redundant features are removed by LASSO algorithm. Finally, the improved LightGBM is used as a classifier to predict the results. This paper [4] proposes a bayesian optimal light gradient enhancement prediction model in time series mode. In this paper, a decision tree model with input variable gradient propulsion is proposed to predict the foundation load capacity of buildings. The experimental results show that the MAE of GBDT model is 24% lower than that of SVM model under high load mode. It is about 37% lower than the DNN model. In paper [7], it determines all possible information gain segmentation points by scanning real data given by materials. But at the same time, this method also leads to a lot of time cost of the experiment. This paper mainly introduces the LightGBM model using GOSS and EFB technology. It first determines the gradient size of a data instance and then excludes data with a smaller gradient. The mutual exclusion feature is used to improve the efficiency and precision of the model. In papers [8] and [9], all proposed models are based on CNN and LightGBM. Firstly, a new feature set is constructed by using the data features in the form of time series. Then use the characteristics of CNN to obtain useful information from the input data set. Adjust network parameters based on the comparison between the predicted information and actual information. Finally, the above LightGBM model is combined with the existing model.

1.2 Our Contribution

We first preprocessed the features of the original data set, and then used 1DCNN and LightGBM to train the fusion model. Our stock prediction model has the correlation of time dimension, but some segments are highly independent, so 1D-CNN conforms to such usage scenarios. By optimizing and adjusting the parameters, a model with high prediction accuracy is obtained.

We used model fusion to improve the performance of the final model. Different models have their own advantages and differences, while model integration can give full play to the advantages of each model, so that these relatively weak learners can become stronger learners by combining some strategies.

2 FEATURE ENGINEERING

In this section, we focus on data sets and feature engineering. The data set of this experiment comes from stock market data related to the actual execution of financial market transactions. The dataset contains characteristics from multiple dimensions, such as ‘time_id’ for the point in time of transaction and ‘bid_price1’ for the highest normalized

bid price. ‘Ask_price1’ represents the normalized minimum asking price. Where, the number at the end of the attribute represents the N-th highest or lowest price.

We also propose the WAP feature, which is computed as follows:

$$WAP = \frac{BidPrice_1 * AskSize_1 + AskPrice_1 * BidSize_1}{BidSize_1 + AskSize_1}$$

After feature engineering, the resulting feature importance is shown in Figure 1 and Figure 2 below.

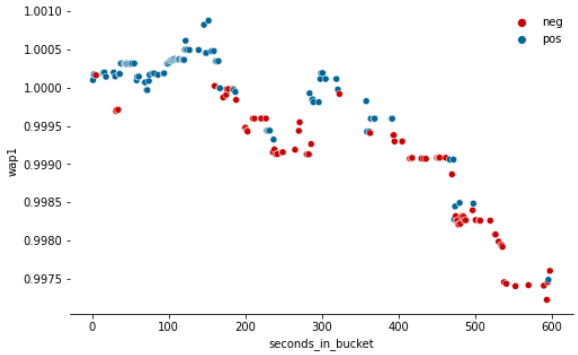

Figure 1. Wap fluctuation Versus Time.

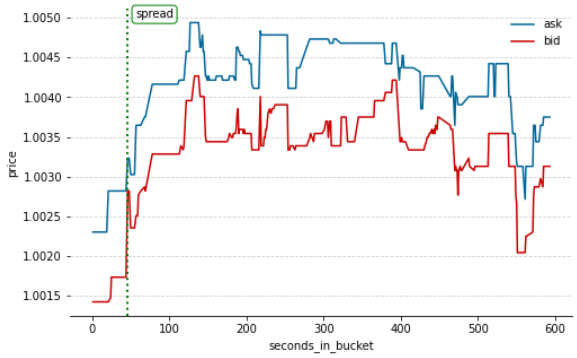

Figure 2. Bid And Ask Price versus Time.

3 XGBOOST AND LIGHTGBM MODEL

3.1 1D-CNN Model

CNN can do a good job of identifying simple patterns in data set. 1D-CNN is very effective when the task needs to obtain interesting features from shorter fragments of the overall data

set, and the position of the features in the fragments is not highly correlated. Our stock prediction model has the correlation of time dimension, but some segments are highly independent, so 1D-CNN conforms to such usage scenarios. Meanwhile, 1D-CNN is also commonly used in the field of sequence model and natural language processing. Some important parameters are involved in the use of 1D-CNN. Filters represent the number of convolution, kernel_size indicates the airspace or time window length of the convolution kernel_size, strides indicates the strides for the convolution step.

3.2 LightGBM

LightGBM is also a framework for implementing models such as GBDT. It can make up for the deficiency of XGBoost [1,2] to some extent and accelerate the training speed of the model without affecting the accuracy. LightGBM [5,6] uses histogram algorithm to solve the problem of excessive number of split points, and unilateral gradient sampling algorithm to solve the problem of excessive number of samples. Meanwhile, LightGBM uses mutually exclusive feature binding algorithm to solve the problem of excessive number of features, as follows.

(1) Leaf-decision tree generate strategy. At present, most of the existing decision tree algorithms are based on the level-wise strategy to generate trees. In this way, some leaves with low fission gain were also divided. This approach incurs some unnecessary overhead. LightGBM, on the other hand, splits the leaves with the maximum splitting gain each time. Thus, the cost is reduced, the error is reduced and the accuracy is improved.

(2) GOSS algorithm. Using the GOSS algorithm can reduce the number of instances to a certain extent, and the gradient is small. When calculating the information gain, only data with a high gradient is used. This saves a lot of space and time .

(3) EFB algorithm. Using EFB algorithm, dimensionality reduction can be achieved by binding multiple mutually exclusive features into the same feature.

(4) Parallelism of feature and data. When processing large data sets, the parallel results of some traditional feature parallel algorithms cannot meet the needs of subsequent operations. Improvements are therefore required. The LightGBM is to find the optimal segmentation, then use the information between different machines to obtain the optimal partition. In this way, it can effectively reduce the cost of information exchange, and at the same time, it also ensures the rationality and authenticity of the data.

3.3 Model Fusion

Model fusion is to build and combine multiple learners to complete the task, we call it model fusion or integrated learning. Different models have their own strengths and differences, while model fusion can give full play to the advantages of each model, so that these relatively weak learners can be combined through some strategy to achieve a stronger learner. There are two requirements for selecting a fusion model: 1) The accuracy of the model used is good. 2) Generally, there are multiple models to be selected. Only when there are differences between these models, can they give full play to their advantages through the fusion model.

4 EXPERIMENTS

We finally chose the 1D-CNN and LightGBM fusion model. The input data were first treated with data augmentation, then trained with 1D-CNN and LightGBM models, which were then fused to produce the final output. The model fusion diagram is shown in the following Figure 3.

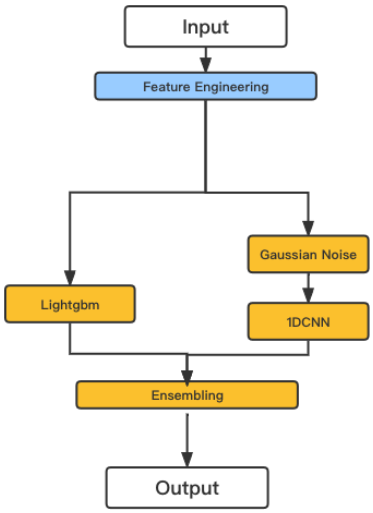

Figure 3. Model fusion.

Through learning and training, we finally adjusted the determined important parameters of 1D-CNN as follows. Where the batch_size is about 2048, the epochs is 1000, and the learning_rate is 0.0001. The important parameters of LightGBM determined by final adjustment are as follows. Where learning_rate is 0.05, min_DATA_IN_leaf is 500, max_depth is -1, and Feature_fraction is 0.5, subsample is 0.72. We also compare the results of model fusion with naive Bayes and the single 1D-CNN and LightGBM models. The metric chosen is RMSPE, which is used to measure the difference between the target value and actual value. Table 1 below is the RMSPE results for the four models. The Naive Bayes model has a RMSPE value of 0.362, the XGBoost model has a RMSPE value of 0.312, the LightGBM model has a RMSPE value of 0.301, and the XGBoost LightGBM fusion model has a RMSPE value of 0.297. The results show that compared with the other three models, 1D-CNN and LightGBM fusion model has a better effect on stock volatility prediction.

Table 1. Results of different models for volatility forecast task.

Models	**RMSPE**
Naïve Bayes	0.351
Xgboost	0.318
lightgbm	0.268
Xgboost+Lightgbm	**0.210**

5 CONCLUSIONS

This paper proposes a stock volatility prediction model based on 1D-CNN and LightGBM. We firstly perform the feature engineering operation on the given dataset. Then, Naive Bayes model, the XGBoost model and LightGBM mode alone, the LightGBM model and the fusion model were compared. We find that the RMSPE value of 1D-CNN and LightGBM fusion model is the smallest, which is 0.210. This shows that the fusion model has a better prediction of volatility than the other three models.

REFERENCES

[1] Xing Liu, TianQiao Liu, Peng Feng.Long-term performance prediction framework based on XGBoost decision tree for pultruded FRP composites exposed to water, humidity and alkaline solution.Composite Structures.2022.

[2] Kyung Keun Yun, Sang Won Yoon, Daehan Won.Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process.Expert Systems with Applications.2021.

[3] Minghui Wang, Lingling Yue, Xinhua Yang, Xiaolin Wang, Yu Han, Bin Yu.Fertility-LightGBM: A fertility-related protein prediction model by multi-information fusion and light gradient boosting machine.Biomedical Signal Processing and Control.2021.

[4] Xiaochen Hao, Zhipeng Zhang, Qingquan Xu, Gaolu Huang, Kun Wang.Prediction of f-CaO content in cement clinker: A novel prediction method based on LightGBM and Bayesian optimization.Chemometrics and Intelligent Laboratory Systems.2022.

[5] Ali Shehadeh, Odey Alshboul, Rabia Emhamed Al Mamlook, Ola Hamedat.Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression.Automation in Construction.2021.

[6] Wanhu Zhang, Junqi Yu, Anjun Zhao, Xinwei Zhou.Predictive model of cooling load for ice storage air-conditioning system by using GBDT.Energy Reports.2021.

[7] Thomas Finley.LightGBM: A Highly Efficient Gradient Boosting Decision Tree.2011.

[8] Ju Yun,Sun Guangyu.A Model Combining Convolutional Neural Network and LightGBM Algorithm for Ultra-Short-Term Wind Power Forecasting.2019.

[9] Chen Cheng,Zhang Qingmei.LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion.2019.