Stock Price Anomalies Analysis and Detection Based on Machine Learning

Tiancheng Xia[1, †, *], Zhidao Chen[2, †]

[1]Shanghai World Foreign Language Academy, Shanghai, China

[2]Beijing Haidian International School, Beijing, China

*Corresponding author: 20204258@stu.hebmu.edu.cn

†These authors contributed equally to this work and should be considered as co-first authors

Abstract. Through long years of development in the financial industry, more people and investments have been immersed here. However, the highly volatile stock market often presents investors with significant surprises. Therefore, identifying anomalies within stock prices in a timely manner is a widely-discussed problem, successfully detecting anomalies can help investors avoid loss, even gaining profit in some cases. In this essay, the performances of the Autoencoder and Light Gradient Boosting Machine are compared in predicting anomalies, employing supervised learning and unsupervised learning respectively. For Autoencoder, we construct layers and use a target loss to make decisions. As for the Light Gradient Boosting Machine, we split the original data into a test set and training set and trained the model. To assign labels, a statistical approach is employed to pre-test the data for normality. From an overall picture, a threshold of 4 achieves the best accuracy performance, proving that the number around four would be a good choice while detecting the strange turning trend in stock price.

Keywords: Anomaly detection, machine learning, autoencoder, LightGBM.

1. Introduction

Global stock markets have become an indispensable and important part of the global economic system. According to statistics, the total trading amount of the Shanghai Stock Exchange in 2022 has reached 9,6255,627.47 million yuan in 242 trading days, which is such a unprecedented figure. This means that the amount traded on the Shanghai Stock Exchange alone already accounts for more than half of China's GDP [1]. This makes both global and Chinese stock markets a large, complex, opportunity and challenging venue. In addition, there is a growing number of participants in global stock markets. From large institutional investors to retail investors, from multinational corporations to startups, everyone is looking for investment opportunities and value in this market [2]. The stock markets are constantly influenced by a multitude of factors, with microeconomic and macroeconomic factors being the most prominent [3]. Key indicators such as the debt-to-equity ratio, earnings per share, return on assets, and total assets turnover reflect a company's internal situation. By analyzing these data points, we can determine whether a particular company is operating favorably. Furthermore, an examination of the broader national and global economic environment aids in assessing the company's potential to generate profits within a given period. However, anomalies that arise during emergencies should also be taken into consideration as significant factors; they may occur due to changes in market conditions [4], while investor sentiments can also play a role [5]. It is in such a stock market environment that timely identification of anomalies in stock prices has become a crucial issue, as it will directly affect the value of stocks in people's hands.

There are two approaches to combating emergencies, which correspond to two distinct relationships between breaking news and stock prices. The first approach entails the occurrence of a sudden event followed by public panic, resulting in significant fluctuations in stock prices. To address this issue, we employ predictive analysis to determine the speed at which specific breaking news spreads after the event occurs and subsequently analyze public sentiment towards it, enabling us to

forecast price trends. Here's an article which talks about combining financial data with news events and sentiment trends to build strategies [6]. The second approach involves abnormal fluctuations in stock prices caused by market sentiment, investor buying and selling behavior, or technical factors unrelated to publicly disclosed breaking news. These situations may arise due to emotional fluctuations among investors or technical trading strategies that do not rely on company fundamentals. Breaking news may be made public after a period, which will further affecting the stock price. In such scenarios, astute investors should possess the ability to detect anomalies before any breaking news emerges as a means of preventing substantial losses. The strategies in < A survey of anomaly detection techniques in financial domain> is trying to find patterns in available data to detect anomalies by multiple models [7]. After conducting extensive research, the majority of academic papers have focused on elaborating upon the first situation. Therefore, the objective of this paper is to delve deeper into the second scenario.

In case of approaches for finding anomalies, statistical methods, deep learning-based detection, and machine learning-based detection… are all feasible [8]. Now, analysis of each type of methods are efficient, especially in discovering their drawbacks and advantages.

In the article of trying to recognize risk statues, author mainly compare the usage of SVM and XGBoost (one type of Gradient lifting algorithm, ensemble learning), in fact, the later one shows better performance, due to the reason of being able to find non-linear relationship, not to be affected by imbalanced data. Another way of detecting anomalies is by using Autoencoder (AE), which shows well performance while detecting fraud and potential financial misstatement [9]. Therefore, this paper compares the performance of Autoencoder and Light Gradient Boosting machine through supervised learning and unsupervised learning, and gives data on the performance and accuracy of each method at different thresholds.

2. Methodology

2.1. Mathematical principle of statistical method

In the field of stock analysis, the individual price of a stock on a certain day may not convey much information, but its change rate is of great significance. Therefore, the calculation of the rate of change, usually expressed as percentage change (PCT), becomes very important in distinguishing the rate of increase and decrease between prices for several consecutive days. When the PCT ratio in a specific period reaches a high level that is beyond the acceptable range, it will issue a red alarm of potential abnormality.

In order to assign accurate labels to each data point, statistical methods are needed. According to the principle of Gaussian distribution (also known as normal distribution), with the expansion of data set, the distribution of data tends to adopt characteristic bell curve. The peak of the curve represents the average, and the standard deviation provides a rough estimate of the proportion of data points located in a certain range around the average. Generally, data falling within one standard deviation contains about 68.27% of the whole data set, while two standard deviations contain about 95.45% and three standard deviations contain about 99.73%. However, although the area beyond these ranges is relatively small, it needs careful consideration to determine the optimal range of p value.

It is very important to balance the p value. If the value of p is set too high, there is a risk of classifying normal data as abnormal data, resulting in false positives. On the contrary, if the range of p value is too narrow, it may lead to ignoring some anomalies that should be identified as abnormal data. In this case, determining the most suitable P value becomes a key task, because it will affect the accuracy of anomaly detection and the reliability of the whole analysis process.

2.2. Auto Encoder (AE)

An autoencoder is model which composed of great number of neurons, an unsupervised machine learning model used for data classification based on inherent data characteristics it learns [10]. The encoder component is responsible for transforming input feature data into a higher or lower

dimensional feature space. Typically, during dimensionality reduction, the encoder compresses the input data, leading to more efficient processing. In the context of anomaly detection, the encoder is trained to learn the characteristics of normal data and subsequently reconstruct them by expanding the data's dimensions and then compressing them back.

The autoencoder comprises three essential components: the encoder, the bottleneck, and the decoder. The encoder focuses on capturing underlying patterns within input normal data and reconstructing them with minimal error. The bottleneck arises as the encoder generates a new-dimensional representation of the input data. Subsequently, the decoder takes this resultant representation and endeavors to predict the original data as closely as possible. A final reconstruction error is computed to establish an acceptable range of deviations.

Finally, both normal data and anomalies are input into the encoder. If the loss exceeds the acceptable threshold, the instance is classified as an anomaly; otherwise, it is classified as normal. This mechanism forms the basis for effective anomaly detection using autoencoders. By harnessing the encoder's capacity to capture subtle patterns and deviations, this approach aids in distinguishing anomalies from normal data instances.

2.3. Light Gradient Boosting Machine (LightGBM)

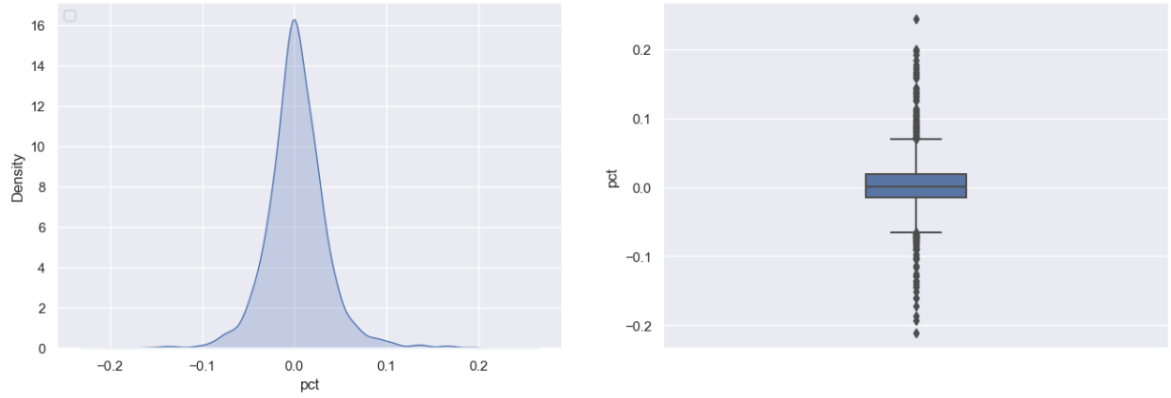
LightGBM is a form of supervised ensemble learning, categorized in the boosting method, which aims to effectively classify data with complex characteristics of different importance levels and optimize the algorithm to produce accurate output [11]. The workflow of LightGBM includes several basic stages: first, the data is divided into different training sets and test sets, and then the parameters are carefully adjusted. This adjustment includes a series of parameters, including determining the number of leaves and the intensity of regularization. In this process, the algorithm systematically explores different parameter combinations and finally infers the optimal configuration. This will eventually generate an informed decision tree.

The challenge of unbalanced data is always a problem in forecasting analysis. However, the inherent ability of integrated learning, as utilized by LightGBM, provides a strategic solution. This is achieved by fine-tuning the weights of each category, thus reducing any deviation of model propagation. In addition, the implementation of cross-validation further enhances the generalization ability of the model and improves its effective execution ability on unknown data. The interaction of these complex technologies makes LightGBM a powerful tool to solve the inherent complexity in modern data analysis and enhances the accuracy and robustness in the presence of complex feature landscapes and unbalanced data sets.

3. Experiment

3.1. Data Process and Analysis

The data came from Kaggle.com, it consists 2843 rows of data, with each row containing the opening prices, the highest prices along with the lowest prices, and the closing prices and volume of the Tesla stock [12]. The time period for the data is between 2010 and 2020. Since no strict formula or rule can used to determine whether a certain value in stock price is an anomaly or not [13], this report first uses a statistical approach firstly to help recognize and labeling a particular data is a normal data or an anomaly. Therefore, the percentage change(PCT) is first calculated to determine the change rate of each day, which equals the difference between today’s price and yesterday’s price, and then over yesterday’s price. After calculation, putting these PCT value onto the same graph, to show the density.



(a) (b)

Figure.1 PCT distribution graph: (a) Normal distribution graph (b) box plot

Figure 1 shows two types of graph to show the distribution of pct, indicating that most of the data lies within -0.1~0.1 percentage change. In order to investigate at which threshold do the two algorithms have the best performances, we set dynamic standard deviations on the graph. For each iteration of a standard deviation, the data which stand away from a certain level of standard deviations are seen as anomalies, the data which stand within this level of standard deviations are considered as normal. Here, normal data are labeled as 0, while anomalies are labeled as 1.

3.2. Determination of Model Threshold

The next step is to set the dynamic thresholds in order to compare the accuracy of these different algorithms. The goal is to search for a reasonable range in which the ratio between the anomalies and the normal data are suitable, neither too high nor too low. The density graph is the perfect tool to observe the data within the range.

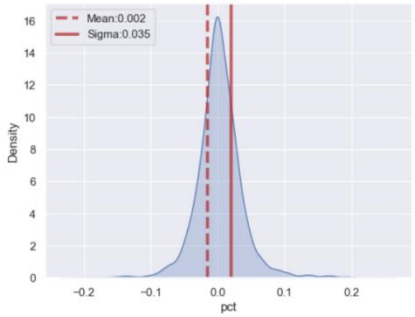


Figure.2 Density distribution when boundary is set on 1 SD

Figure 2 is the graph when threshold is set at 1. It has been labelled with 2 red lines, one stands for the mean and the other stands for the sigma line. The region between these 2 lines are seen as normal. After much verification, 1 standard deviation(around 0.03 pct change) is the lowest boundary available. Any boundary lower than 1 deviation will cause great influence on the accuracy of the algorithms, and there are more anomalies than normal ones.

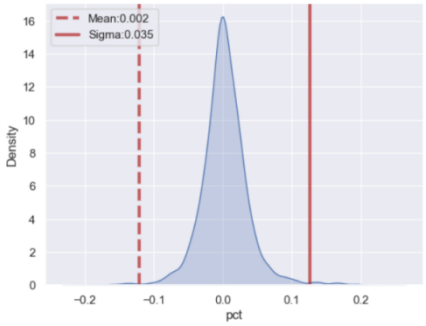


Figure.3 Density distribution when boundary is set on 3.5 SDs

Figure 3 shows the boundary when threshold is set at 3.5, which is the highest deviation that can ensure enough accuracy (about 0.13 percentage change). When the deviation slowly reaches 3.5, the ratio of normal data against anomalies increases greatly, as the area between the 2 red lines increases and the area which is left out decreases.

After determining the lowest boundary and the highest boundary, we select the values from 1 to 3.5 and 0.5 as a single separation.(0.5, 1, 1.5, 2…) Table 1 shows the exact number of normal data and anomalies that are calculated.

Table 1. Normal data over anomalies ratio

Deviation	Number of Normal Data	Number of Anomalies	Ratio
1	2251	591	3.81
1.5	2576	266	9.68
2	2700	142	19.01
2.5	2760	82	33.66
3	2791	51	54.73
3.5	2802	40	70.05

Table 1 shows the number of different situations and the ratio between normal data and anomalies. The ratio increases from around 4 times to 70 times. Although the difference is huge, but it is acceptable because both provide a reasonable amount of data.

3.3. Model Training

The first model here used is Autoencoder. Autoencoder is a model which tries to reconstruct the input original data by using encoder, which transform the data to a lower-dimensional space, and then using a decoder, which attempts to predict the original input.

In this case of anomaly detection, there are 4 layers, when all normal data are input, Autoencoder will capture the internal patterns and features existing in them, expand the dimension to 32 and 64, and then shrink back to 32 and 1. This process makes the model learn a compact representation of the normal data. Meanwhile, the automatic encoder learns to minimize the reconstruction loss, which measures the acceptable interval between the output and input. This loss value is not only used to evaluate the quality of the model, but also used as a threshold to determine whether a certain data is normal during testing. There is a significant difference between Autoencoder and other common predicting algorithms which split the original data into the training set and the test set, Autoencoder uses the original data for training, and gives a target loss as an output. Then the algorithm re-tests the original dataset and any data that has a greater loss than the target loss is considered as an anomaly.

The second model used here is the Light Gradient Boosting Machine (LightGBM). LightGBM is a gradient boosting method based on decision tree. It is necessary to use gradient advancing algorithm to build decision tree integration, which involves iteratively improving the model by adding trees to the integration, predicting the target value with initial parameters, and then calculating the loss of a pair of parameters. In order to make the loss most accurate, all possible parameter pairs have to go through the cross-validation process. During the whole process, the machine itself will learn to determine the weights of each feature and how to get the best classification. After that, the model can help make predictions with test data. In this certain case, we set the test size to 0.2, then we use the training set to train the model and use it to predict the test set.

The final step of the investigation phase is to compare the results with the results using the traditional statistical method, then calculate an accuracy score for each algorithm.

4. Results

In the final result, we will further visualize the previously calculated data, compare the accuracy changes of the two methods and select the optimal threshold. In the first part, we set the threshold to

1 and calculated the accuracy of the two methods in each case by training the two models and plotted them. Figure 4 shows the accuracy of each method.

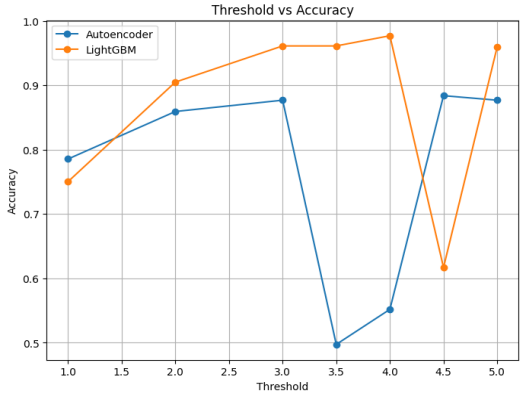


Figure.4 Line chart of accuracy at different threshold

From this line chart (Figure.4), we can figure that the overall accuracy of the LightGBM method is higher than that of the Autoencoder method. After that, we evaluate the performance of the two methods by calculating the value of AUC and also show it in the form of a statistical graph. Figure 5 shows this graph.

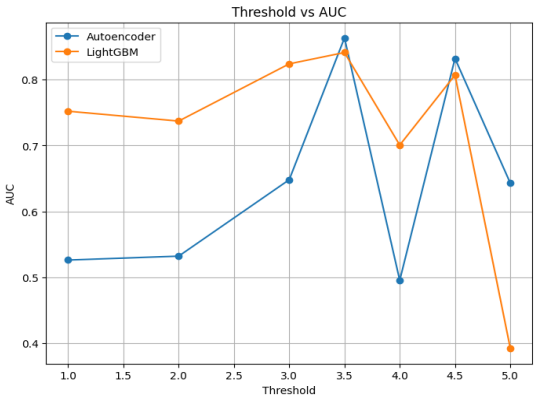


Figure.5 AUC chart of each method at different threshold

From Figure.5, we can notice that when the threshold is less than 3.5, the performance of the LightGBM method is significantly higher than that of Autoencoder. At thresholds of 3.5 and 4.5, the peak performance of the two methods is almost equal. However, at a threshold of 4.0, there is a significant decrease in AUC feedback for both. Finally, as can be seen from Figure 4, the accuracy is highest at a threshold of 4.0.

5. Conclusion

The aim of this review is to determine the optimal threshold for statistical labeling methods aimed at detecting anomalies, using the capabilities of two powerful machine learning methods, the Autoencoder and the light gradient boosting machine (LightGBM). This study includes six different datasets to allow for comprehensive experiments. Our results show that setting the threshold at 4 standard deviations is the most effective way to identify abnormal stock price movements. This choice strikes a balance between minimizing the inclusion of normal data in the anomalous data and ensuring that most instances of anomalous data are accurately detected. In addition, this investigation opens the way for further exploration. The overall process can be divided into two distinct phases. First, the previous tagging process can be optimized by fine-tuning the thresholds. In addition, other statistical and labeling methods are worth exploring. Second, the subsequent prediction phase can be enhanced. This requires exploring optimal machine learning parameters and finding models with higher prediction accuracy. The realization of these two goals is expected to synergistically improve the overall accuracy of anomaly detection.

References

[1] Summary of stock transactions: Shanghai Stock Exchange. Trading overview | Shanghai Stock Exchange. (n.d.). http://www.sse.com.cn/market/stockdata/overview/yearly/

[2] The volume of A-shares is approaching that of China's total GDP capital market, which highlights the steady growth of services. A-share volume is approaching China's total GDP capital market outstanding service stable growth - Xinhua Net. (n.d.). http://www.news.cn/fortune/2022-01/19/c_1128276299.htm

[3] Gursida, Hari. "The influence of fundamental and macroeconomic analysis on stock price." Jurnal Terapan Manajemen dan Bisnis 3.2 (2017): 222-234.

[4] Safeer, Mohammed, and S. Kevin. "A study on market anomalies in Indian stock market." Int. J. Bus. Admin. Res. Rev 1 (2014): 128-137.

[5] Summers, Barbara, and Darren Duxbury. "Decision-dependent emotions and behavioral anomalies." Organizational Behavior and Human Decision Processes 118.2 (2012): 226-238.

[6] Daradkeh, Mohammad Kamel. "A hybrid data analytics framework with sentiment convergence and multi-feature fusion for stock trend prediction." Electronics 11.2 (2022): 250.

[7] Ahmed, Mohiuddin, Abdun Naser Mahmood, and Md Rafiqul Islam. "A survey of anomaly detection techniques in financial domain." Future Generation Computer Systems 55 (2016): 278-288.

[8] Algorithmic Approaches to Anomaly Detection in Financial Markets - Linkedin." Https://Www.Linkedin.Com/Pulse/Algorithmic-Approaches-Anomaly-Detection-Financial-Markets-#:~:text=In%20the%20financial%20markets%2C%20anomalies,Events%20or%20patterns%20in%20data., www.linkedin.com/pulse/algorithmic-approaches-anomaly-detection-financial-markets-. Accessed 5 Aug. 2023.

[9] Bakumenko, Alexander, and Ahmed Elragal. "Detecting anomalies in financial data using machine learning algorithms." Systems 10.5 (2022): 130.

[10] Wang, Yasi, Hongxun Yao, and Sicheng Zhao. "Auto-encoder based dimensionality reduction." Neurocomputing 184 (2016): 232-242

[11] Sewell, Martin. "Ensemble learning." RN 11.02 (2008): 1-34.

[12] Bozsolik, Timo. "Tesla Stock Data from 2010 to 2020." Kaggle, 4 Feb. 2020, www.kaggle.com/datasets/timoboz/tesla-stock-data-from-2010-to-2020.

[13] Anandakrishnan, Archana, et al. "Anomaly detection in finance: editors' introduction." KDD 2017 Workshop on Anomaly Detection in Finance. PMLR, 2018.