Exploring the Effectiveness of Dimensionality Reduction Techniques for Stock Price Prediction

Dongyu Zhuo

zhuody@shanghaitech.edu.cn

School of Information Science and Technology, ShanghaiTech University, Shanghai, 201210, China

Abstract. The research's objective is to anticipate fluctuations in average stock prices for short-term forecasts. The investigation provides valuable insights for investment decision-making, risk management, and evaluating industry and company performance. The study employs a range of techniques, such as data visualization, data cleaning, and dimensional reduction, to accomplish these goals. The models are trained using six machine-learning approaches and evaluated using six metrics. The primary focus of this research is to identify the crucial factors in predicting stock prices and to find the most effective combination of dimensional reduction techniques and machine learning methods.

Keywords: Stock price prediction, Dimension reduction, Machine learning

1 Introduction

Predicting stock prices is critical for making informed investment decisions, managing risks, and assessing industry and company performance. Accurate price predictions assist investors in determining the optimal time to buy, sell, or hold stocks, reducing risks, and maximizing returns. This data also helps companies and financial institutions manage risks by identifying potential market conditions that could affect their operations. By analyzing stock price movements, investors and analysts can evaluate company performance, compare it with competitors and industry trends, and make informed investment decisions. The paper focuses on predicting short-term changes in the mean stock price, and it uses uniform models and parameters to forecast all stocks equally since the aim is to identify similar patterns among stocks.

The data was obtained, cleaned, and visualized before using dimensionality reduction techniques to tackle multicollinearity. Six different machine learning algorithms were then implemented to evaluate how dimensionality reduction affected the model's performance. The models were assessed using six different metrics, and the significance of features was computed for the best model after optimizing hyperparameters.

2 Literature Review

The literature review provides an overview of the methods and indicators used by previous researchers to predict stock prices.

ICPDI 2023, September 01-03, Chongqing, People's Republic of China

DOI 10.4108/eai.1-9-2023.2338744

Studies have suggested that sentiment indicators are useful for forecasting stock prices as they reflect investor sentiment and market trends. Li et al. found that sentiment analysis of financial news articles can improve the accuracy of predicting stock price movements [1]. Ranco et al. observed a significant correlation between Twitter sentiment and abnormal returns during the peaks of Twitter volume [2]. Financial indicators are a crucial area for predicting stock prices as they provide valuable information about a company's financial performance. Tseng discovered that portfolios with low price, low P/E ratios, and small market prices resulted in greater excess returns [3]. Chiu et al. developed an algorithm that enhances traditional PB investment strategies and helps identify good stock portfolios [4].

Machine learning and dimensionality reduction techniques are commonly employed in predicting stock prices. For instance, Vijh et al. utilized Artificial Neural Network and Random Forest techniques to forecast the next day's closing price for five companies [5]. Hegazy et al. proposed a machine-learning model for stock market prediction that utilized Particle Swarm Optimization and Least Square Support Vector Machine algorithms [6]. Ghorbani et al. presented a method for predicting stock prices that employed Principal Component Analysis and time-varying covariance information [7].

3 Data

The dataset has been directly obtained from the GitHub repository hosted at [https://github.com/zhiaozhou/Chinese-Stock-Prediction-Using-Weibo-Baidu-News-Sentiment.] The data is from January 1st, 2018 through April 8th, 2018, and consists of 4564 samples from 200 stocks. The data is stored in the form of a table and is saved as a .csv file.

Data visualization is then achieved through the use of a heatmap.

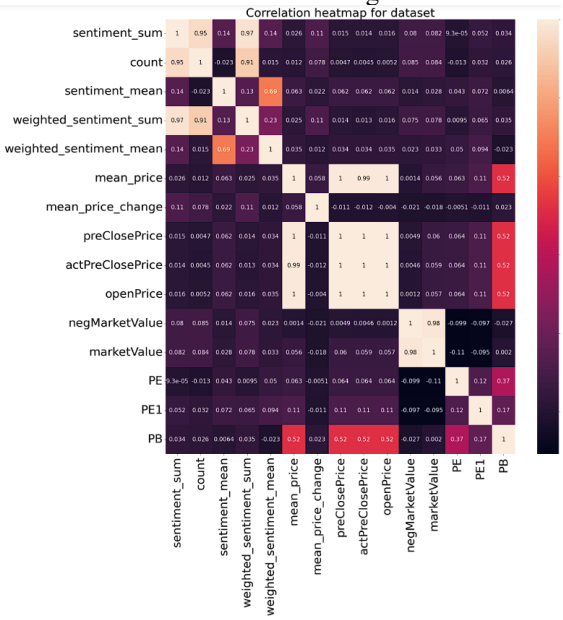

Fig. 1. Heatmap for features

As shown in Figure 1, the heatmap displays the correlation between all the features, highlighting those with high correlation through a light color. The numbers in the graph represent the correlation between a column and a row. For example, the value 0.95 located at the intersection of the second row and first column represents the correlation between the features 'sentiment_sum' and 'count'.

4 Methodology

The methodology section of the essay describes the steps involved in data cleaning, dimensional reduction, and training algorithms.

4.1Data cleaning

Data cleaning involves addressing missing or duplicate data and normalizing the data using z-score normalization.

4.2Dimensional reduction

Five redundant columns, including 'Unnamed: 0', 'ticket', 'date', 'tradeDate_x', and 'tradeDate_y', that are unrelated to stock price prediction in the context of this essay are eliminated. Three groups of features that show extremely high correlation within each group have been identified. Group 1 includes 'sentiment_sum', 'count', and 'weighted_sentiment_sum'. Group 2 includes 'mean_price', 'preClosePrice', 'actPreClosePrice', and 'openPrice'. Group 3 includes 'negMarketValue' and 'marketValue'. To provide a basis for comparison, a control group is created without the implementation of any dimensional reduction techniques. The dataset is labeled as "Dataset 1". In "Dataset 2", certain features from each group have been dropped while retaining others. In Group 1, 'weighted_sentiment_sum' is considered financially more significant as it reflects sentiment and confidence level. In Group 2, 'mean_price' is considered more meaningful as it reflects the average stock price throughout the day, while other features only reflect the price at a specific time. In Group 3, 'negMarketValue' is considered more significant as it considers both assets and liabilities. All other features within each group have been discarded.

PCA is a data analysis technique that reduces the complexity of high-dimensional datasets by identifying the principal components that account for the most significant variation in the data and projecting the data onto a lower-dimensional space. The author applied PCA in two ways: first, by setting a threshold of 0.95 to determine the minimum number of features that account for at least 95% of the variance in the original data, and second, by selecting dimensionality values of 2, 3, and 1, respectively, to minimize the removal of essential features. The resulting datasets were labeled as "Dataset 3.1" and "Dataset 3.2", respectively, and the trade-off in dimensional reduction will be discussed later in the essay.

4.3Training Algorithms

The essay then used six different methods for training in stock price prediction: Linear regression with Ordinary Least Squares (OLS) regression, Linear regression with Weighted Least Squares (WLS), Ridge regression, random forest, support vector machine, and neural network. Linear regression with OLS is used to model the relationship between dependent and

independent variables, while WLS is used when the assumption of constant variance is violated. Ridge regression is used when there is multicollinearity among independent variables. Random forest is an ensemble learning method, SVM finds the hyperplane that separates data points, and neural networks are inspired by the structure and function of the human brain.

5 Results

5.1Evaluation

The essay discusses various metrics used to evaluate the performance of models in predicting stock prices, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R2, mean, and standard deviation. The essay also explains the K-fold cross-validation technique, which is used to evaluate the performance of a machine learning model. This technique provides a reliable estimate of the model's performance, and it is particularly useful for small datasets or models with many parameters.

Furthermore, the essay employs line graphs to evaluate the precision and dependability of the prediction model.

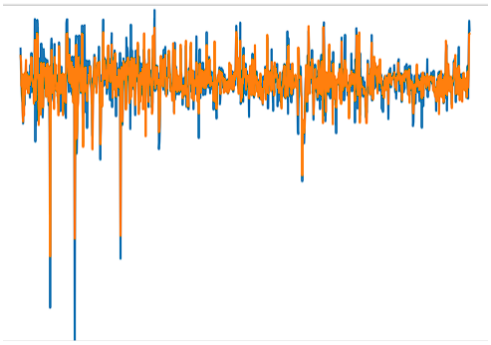

Fig. 2. Line graph for the best model

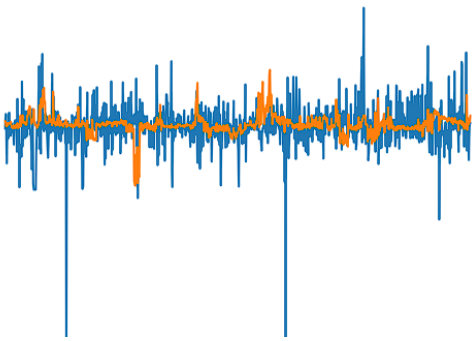

Fig. 3. Line graph for the worst model

Figure 2 and Figure 3 show the actual and predicted prices for the best and worst models, respectively. Both graphs feature an orange line to represent predicted values and a blue line to

represent actual values. The graphs illustrate that the estimates made by the most accurate model are very similar to the actual values, while those made by the least accurate model are quite different.

5.2The performance difference of the models and related explanation

The best model among all models is random forest using Dataset 3.2. It has MAE:0.3460021243440936, MSE: 0.22923942138118888, RMSE: 0.47878953766888943, R2 Score: 0.7474466049143784, Mean: 0.07751180008286569, Standard deviation: 0.9229834804781238. The best hyperparameters for the model are also selected to be: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}.

The worst two models among all models is a neural network using Dataset 2. It has MAE:0.9747263570997134, MSE: 2.0435937158681536, RMSE: 1.4295431843313282, R2 Score: -1.2514300900276791, Mean: -0.15708138703318555, Standard deviation:1.1279007482831156

The essay notes that the models in Dataset 2 and Dataset 3.1 perform poorly as they eliminate many features, resulting in underfitting. Neural networks, which are sensitive to low-dimensional feature spaces, also perform poorly. However, Dataset 1 performs moderately well as it avoids underfitting, although it does not address high correlation issues among features. The majority of models perform well on Dataset 3.2 due to two factors: the utilization of PCA to tackle multicollinearity and the controlled elimination of features to prevent underfitting. These outcomes imply that caution should be taken when selecting dimension-reduction methods for datasets with a restricted number of features.

5.3Feature importance

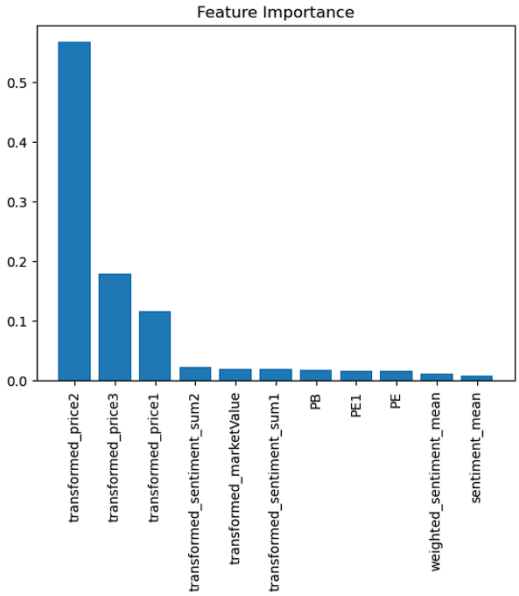

Fig. 4. Feature importance

Figure 4 shows that 'transformed_price2', 'transformed_price3', and 'transformed_price1' are the most significant factors, while the contribution of other factors is relatively low. These three factors are price indicators that have been transformed using PCA, which suggests that price indicators are fundamental for predicting short-term stock prices. However, this does not imply that other indicators are unimportant, as they may have a long-term impact on stock prices.

6 Conclusion

The essay utilizes multiple techniques, such as dimensional reduction methods and six distinct prediction models, to analyze stock prices. It identifies the models with the best and worst performance and highlights the significance of carefully choosing dimension reduction methods and features to attain optimal model performance. The study's importance lies in its potential to enhance various areas, such as investment decision-making, risk management, industry, and company performance evaluation, given the critical role of stock price prediction in these domains. To further improve the essay, it could incorporate recent news about companies using web crawlers to increase the accuracy of stock price prediction.

Reference

[1] Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. Frontiers in Artificial Intelligence and Applications, 263, 373-382.

[2] Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., & Mozetič, I. (2015). The effects of Twitter sentiment on stock price returns. PloS one, 10(9), e0138441.

[3] Tseng, K. (1989). Low price, price-earnings ratio, market value, and abnormal stock returns. Journal of Financial Research, 12(3), 203-214.

[4] Chiu, Y.-J., Chen, K.-C., & Che, H.-C. (2020). Patent predictive price-to-book ratio (PB) on improving investment performance--Evidence in China. PloS one, 15(7), e0236183.

[5] Vijh, M., Chandola, D., Tikkiwal, V., & Kumar, A. (2021). Stock Closing Price Prediction using Machine Learning Techniques. [Unpublished manuscript].

[6] Hegazy, O., Soliman, O., & Salam, M. (2018). Stock market prediction using PSO-optimized LS-SVM. Expert Systems with Applications, 92, 160-172. DOI: 10.1016/j.eswa.2017.09.050

[7] Ghorbani, M., & Chong, E. (2018). Stock price prediction using principal components. arXiv preprint arXiv:1804.04222.