Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2022, 10, 1-21
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

A Machine Learning Approach: Enhancing the Predictive Performance of Pharmaceutical Stock Price Movement during COVID

Beilei He[1], Weiyi Han[1], Suet Ying Isabelle Hon[2]

[1]Jacobs School of Engineering, University of California San Diego, San Diego, USA
[2]Xiaohongshu, Inc., Shanghai, China
[3]Centre of Industrial Relations and Human Resources, University of Toronto (St. George), Toronto, Canada
Email: behe@ucsd.edu, weiyih@alumni.sjtu.edu.cn, suet.hon@mail.utoronto.ca

How to cite this paper: He, B.L., Han, W.Y. and Hon, S.Y.I. (2022) A Machine Learning Approach: Enhancing the Predictive Performance of Pharmaceutical Stock Price Movement during COVID. *Journal of Data Analysis and Information Processing*, **10**, 1-21.
https://doi.org/10.4236/jdaip.2022.101001

Received: October 18, 2021
Accepted: December 26, 2021
Published: December 29, 2021

Abstract

Predicting stock price movement direction is a challenging problem influenced by different factors and capricious events. The conventional stock price prediction machine learning models heavily rely on the internal financial features, especially the stock price history. However, there are many outside-of-company features that deeply interact with the companies' stock price performance, especially during the COVID period. In this study, we selected 9 COVID vaccine companies and collected their relevant features over the past 20 months. We added handcrafted external information, including COVID-related statistics and company-specific vaccine progress information. We implemented, evaluated, and compared several machine learning models, including Multilayer Perceptron Neural Networks with logistic regression and decision trees with boosting and bagging algorithms. The results suggest that the application of feature engineering and data mining techniques can effectively enhance the performance of models predicting stock price movement during the COVID period. The results show that COVID-related handcrafted features help to increase the model prediction accuracy by 7.3% and AUROC by 6.5% on average. Further exploration showed that with data selection the decision tree model with gradient, boosting algorithm achieved 70% in AUROC and 66% in the accuracy.

Keywords

Machine Learning, Stock Price Trend, Prediction, Feature Engineering

1. Introduction

This is a special time because of COVID. As cases and deaths surge across the

globe, people are confronting uncertainty through and beyond the global health crisis. This uncertainty may cause anxiety, fear, and other irrational reactions from the general public. Because of the sensitivity, concern, and other difficulties including more restrictive policies and shutdowns, stock markets are affected in an unrepresented way. This instability caused a great loss for investors, more fragile and volatile price returns and stock performance, and other social problems including wealth inequality [1]. Thus, improving the performance of models aiming to analyze and predict stock prices in the market is meaningful for both private investors and the public interest.

Predicting the stock price change is a challenging task, involving various factors and non-linear relationships among them [2] [3]. Some of these factors are not easily understood and quantifiable, like people's subjective emotions. Things become more challenging and volatile after the pandemic. But even with these difficulties, stock price prediction is a trending topic.

Since the outbreak of the pandemic, there is an increasing number of research analyzing the connection between COVID-19 and the movements in the capital market [4] [5] [6]. More specifically, most of these studies utilize statistical analysis to anticipate COVID-19 data and leading pharmaceutical stock performance [4] [5]. There are some reviews that explore the potential of machine learning models by examining the correlation between COVID-19 statistics and general stock market performance [6] [7]. For instance, Adekoya and his team employed decision tree models to examine the correlation. And Rouf's team focused more on the Indian general stock market.

Nonetheless, these studies and research are different from the metric and goal of our research topic, which emphasizes more about whether COVID-19 statistics can help improve the prediction performance of machine learning models in terms of future pharmaceutical stock price directions. In this study, we mainly focus on two research questions:

1) How to predict the next-day direction of pharmaceutical companies' stock prices movement using a machine learning approach?

2) During the pandemic, to what extent could external information, especially COVID-related features improve the performance of the machine learning models that this study has developed?

Our research intends to fill the research gap by examining whether the inclusion of COVID-related data and some other stock environmental data in the features with prepossessing during pandemics will increase the performance of different machine learning models.

The main contributions of our paper can be summarized as follows:

1) We collected detailed company-specific financial statistics, including current and previous stock prices and key financial measuring metrics, selected global industry indexes, and the COVID-related data. We combined all these data together and transformed them into a single-day oriented dataset that can be used to train various models;

2) We developed a series of handcrafted features for COVID-related data and

selected global industry indexes to aid the prediction of stock price change direction. The results suggest an average of 1.7% increase in AUROC with the inclusion of industry indexes features and an average of 6.5% increase in AUROC with the introduction of COVID-related features;

3) We further explored the potential of our model performance through data selection by eliminating the early COVID period data. After dropping the early COVID period, the results suggest an average of 3.9% increase in AUROC and an average of 1.9% increase in the accuracy. This helps the decision tree model with gradient boosting algorithm achieve 70% in AUROC and 66% in the accuracy.

2. Related Work

Various methods in the stock price prediction field can be classified into two categories. The first category aims to improve the performance by developing new structure or combinations of the models. The second class focuses on finding new informative features related to the stock market to help the stock price movement prediction.

In particular, in the first category that focuses the model construction, many different computational models have been used, including Convolutional Neural Network (CNN), Artificial Neural Network (ANN), Long Short Memory Networks (LSTM), Support Vector Machine (SVM), and Random Forests. In 2011, Kara's team implemented and compared ANN and SVM models performance [8]. They selected ten comprehensive parameters and passed into both models to predict daily stock price movement of Istanbul Stock Exchange National 100 index. The results suggested that ANN models constantly outperform SVM in terms of accuracy in the stock price movement prediction.

Recurrent Neural Network is designed to have internal memories that could help to extract historical characteristics and information. This internal memory could be effective in the stock price prediction domain as stock price database are mostly time oriented and the historical trend information is important. Several research teams have conducted experiments on various Recurrent Neural Networks. In the research from Nelson and his team, the LSTM model was constructed and leveraged with technical indicators as inputs to predict the stock price movement direction of the near future in the Brazilian stock market [9]. The results suggested an average accuracy from LSTM of 55.9%, which outperforms Multilayer Perceptron (MLP) Model. In 2018, Ficher and Krauss applied LSTM with RMSProp as an optimizer to predict stock price movement in S&P 500 from Dec 1989 and Sep 2015. The authors found that the performance from LSTM is better than deep neural network and logistic regression classifier. The extended coverage of the stock market leveraged the function of the long term memory unit in the LSTM. McNally's research team implemented a combination of LSTM and a Bayesian based RNN to forecast the price of Bitcoin [10]. Their dataset ranges from August 2013 to July 2016. Their results showed a similar performance comparing to random forest classifier.

Persio and Honchar implemented and compared different models including MLP, CNN, and RNN to predict the daily based S&P 500 index [11]. The results showed that the CNN model with wavelet transform outperforms other models with accuracy of 53.6%.

Ramos-Pérez and his team developed a combination of different models to predict S&P 500 volatility [12]. In particular, they stacked Gradient Descent Boosting, Random Forest, SVM, and ANN to make the prediction of S&P 500 and evaluated the performance on an annual basis. The result suggested a better performance from stacked networks than other individual models in term of RMSE.

In addition to the improvement on the models, some researches focus more on the feature selection to improve the model performance. Zhang's team established an approach to extract user sentiments and web news from social media to improve their stock price movement prediction performance in China A-share and HK stock data in 2015 [13]. They employed a coupled matrix and tensor factorization framework to investigate the impacts of the events (user sentiments and web news) on the stock price movements. Nam and Seong developed Multiple Kernel Learning (MKL) that depends on the financial newsfeed to predict the stock price movement [14]. The MKL model is used to combine the features of target companies and normal companies. They also employed context-aware text mining based on company-specific financial news to gain a better performance. In 2021, Zhang and his team implemented a Convolutional Neural Network model based on a deep factorization machine and attention mechanism (FA-CNN) with different feature engineering techniques to predict the stock price movement [15]. Their results suggested that the inclusion of the industry index information helps to improve the accuracy by 10.2% on average with FA-CNN. Based on the work from Zhang and his team, we also included the industry index as our independent feature along with other COVID-related features.

There are also some data reprocessing techniques that is relevant to our experiment. In the study from Kotsiantis' team, they explained that the data normalization is important especially when there is a large difference between the minimum and maximum [16]. In particular, they mentioned two normalization techniques: min-max normalization and z-score normalization. In our experiments, we have to collect an extended range of data. Within the same feature, the differences between value two data points could be enormous, such as different stock price value across different companies. In our study, we implemented the min-max normalization to our features.

Based on these previous work, in this study we will compare the performance of three different models: MLP with Logistic regression, Decision Tree with Gradient Boosting, and Decision Tree with Random Forests, with different feature selection combinations, data selection techniques, and online machine learning methods, on 7 selected pharmaceutical stock price over the COVID period.

3. Study Design and Methodologies

In this section, we will explain our research design for the stock pricing direction prediction problem, including our methodologies on dataset prepossessing, main and secondary evaluation metrics, and feature engineering.

3.1. Problem Formalization

For each company k and day i, where $i \in \{1, \cdots, M\}$ and $k \in \{1, \cdots, N\}$, first sorted by the company k and then sorted by the date i, there is a corresponding feature F_i^k, where $F_i^k = A_i^k + C_i^k + I_i$ (+means appending in this case). A_i^k represents the daily financial statistics of the company k at day i, C_i^k represents the daily COVID related data of the company k at day i, and I_i represents the industry index stock key values at day i. These index selections are independent of companies, meaning for all features with same date i, I_i is the same. Stock key values covers conventional metrics, including open, low, high, adjusted close, and volume. Given the feature F_i^k, its corresponding label y_i^k is defined as follows:

$$y_i^k = \begin{cases} 1 & \text{if } CLOSE_i > CLOSE_{i-1} \\ 0 & \text{if } CLOSE_i \le CLOSE_{i-1} \end{cases}$$

3.2. Data Prepossessing

The first step in this study is to construct a comprehensive dataset. The pandemic happened in November 2019, which means we only have historical data of less than 20 months available. In order to expand our dataset and fulfill the generalization purpose of our conclusion, we selected 9 COVID vaccine companies from the market. We then collected daily stock price data and COVID related data from different sources, including yfinance (https://pypi.org/project/yfinance/) and our world in data (https://ourworldindata.org/). Because of the nature of daily stock prediction, the dataset is daily-oriented, meaning that each row is a day of data. The dates ranged from 2019.11.30 to 2021.08.30. The break out of the COVID happened around March in 2021, but there were emerging news on an influx of cases of pneumonia around middle of December. The later analysis on the start of the COVID suggested that the start of the COVID spread may be dated back to November. Therefore, we extended the starting date to end of November to cover all possible range of stock prices that may be affected by COVID.

We also pay closer attention to the later dates of the data (2020.11.1 to 2021.08.30), looking for an increase in performance of our model, as we anticipate that people were impacted by more significant experiences of discomfort, fear and uncertainty at the start of the pandemic. In addition, from 2020.11.01 onward, companies started to reveal vaccine experiment data. Pfizer released phase 3 results of over 90% efficiency on November 9th; Moderna and Sinapharm also followed and revealed their phase-3 results at the end of November and December 2020. This could be a relief to people's concern over pandemics. At the same time, the daily increase in new cases also started to slow down around November 2020. Thus we anticipate that, compared to the onset of the pandemic, the stock

market in the later COVID period should have been more predictable as panic dissolved. In the later experiments, we explored the performances of various machine learning models in different data ranges.

There is a feature associated with each row/day. We organized the features into two categories: inside the company and outside. Inside the company mainly includes the internal financial situation and daily and historical stock key points. Stock key points cover conventional metrics, including open, low, high, adjusted close, volume. Meanwhile, financial statistics include price-to-book ratio and quarterly EBIT and EBITDA values.

$$\text{EBIT} = \text{Net income} + \text{Interest expense} + \text{Tax expense}$$

$$\text{EBITDA} = \text{Net profit} + \text{Interest} + \text{Tax} + \text{Depreciation} + \text{Amortization}$$

To match the daily oriented data to these quarterly sampled data, we implemented quarter-to-day mapping. Specifically, all days belonging to the specific quarter share the value of quarterly sampled data. In the later section, we will discuss our feature engineering in detail.

To accommodate the problem of extended range of some features, especially for the company-specific stock price and COVID-related statistics, we employed the min-max normalization from Kotsiantis' study independently across features (column-wise).

We have one data point partial feature after normalization from our dataset attached in the Appendix **Table A1** as an example. The full dataset is available at https://www.dropbox.com/s/fwe7vw1y66orcvd/data_normalized.csv?dl=0here.

3.3. Evaluation Metrics

To better understand the results and the performance of the selected models, we adopted both the Area Under the Curve for the ROC plot [17] (AUROC) and accuracy as our evaluation metrics. These probability statistics give us a representative performance evaluation of different models in different situations. As we expected the imbalance between the different labels would become larger, AUROC can help to give a similar weight between two classes (0 or 1 in our case). The detailed distribution of our dataset is shown in **Table 1**. Accuracy helps us to evaluate a more balanced evaluation on the precision of both positive and negative labels. For the decision tree model with random forest bagging algorithm, the metrics we use is accuracy, as AUROC does not apply to the categorical output label. Similar implementation has been done in the statistical study of random forests [18].

Table 1. Dataset label distribution.

	Number of samples	Positive labels	Negative labels
Training	3396	1764	1632
Testing	378	194	184
In total	3774	1958	1816

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}}$$

In our study, we will compare AUROC and accuracy in different situations to evaluate whether certain features are helpful for improving the performance, as well as which model has the best performance in this case. We first input only financial features including the stock key values of the current day, 5 previous days, 3 previous weeks, and 2 previous months. Next, we add external handcraft features individually including sub-industry index stock key values and COVID statistics to observe if there is any improvement in prediction performance. We repeat the same process of comparison for different selected machine learning models to evaluate their results.

3.4. Feature Engineering

Using the same notation as before, there are mainly three categories in our feature selection: financial statistics of the company (A_i^k), COVID related statistics (C_i^k), and industry stock key values (I_i).

In total, the feature size is 188, which includes feature size of 58 for financial statistics, 75 for COVID related statistics, and 55 for industry stock key values.

1) Financial Statistics. In this class of features, we mainly incorporated current and historical stock price key values. We also included financial performance statistics in this section. Conventional stock price prediction models usually have this section of features [15]. Stock price key values include open, high, low, adjusted close, and volume.

- Current day stock key values;
- Previous 5 days, 3 weeks, and 2 months stock price key values. Because of the nature of our dataset structure, it is hard for common machine learning algorithms to capture the historical characteristics of the current stock price. We included some historical stock price data to help the model to learn from a wider range of dates. To capture more accurate and detailed historical data, we strictly enforced time intervals, without taking the predetermined weeks and months it belongs to;
- Daily price to book ratio. Price to book value is the ratio of the market value of the company's share price over its book value. This value is used to evaluate whether a stock is properly priced. Higher PB ratio could mean the stock price at this point is overvalued, and vice versa for a lower PB ratio;
- Quarterly EBIT and EBITDA data. EBIT and EBITDA are commonly used to evaluate the probability of a company. EBIT and EBITDA are released every quarter. To match with our daily oriented dataset structure, we mapped this quarterly sampled data to daily rows by repetition. All the days belonging to the specific quarter will share the same quarterly sampled data.

2) Related Index Stock Key Values. In this section of features, we selected several comprehensive industrial indices and some pharmaceutical or biotechnology industry indices. These industry indexes are independent of the company. This

means that each row of company k and day i, they share the same selected industry key values. We expected these industry indices to provide more comprehensive information on the stock market environment. We sampled these industry indices daily to match with our existing dataset structure.

- Current day stock key values;
- Previous 5 days, 3 weeks, and 2 months stock price key values. Because of the nature of our dataset structure, it is hard for common machine learning algorithms to capture the historical characteristics of the current stock price. We included some historical stock price data to help the model to learn from a wider range of dates. To capture more accurate and detailed historical data, we strictly enforced time intervals, without taking the predetermined weeks and months it belongs into account;
- Daily price to book ratio. Price to book value is the ratio of the market value of the company's share price over its book value. This value is used to evaluate whether a stock is properly priced. Higher PB ratio could mean the stock price at this point is overvalued, and vice versa for a lower PB ratio;
- Quarterly EBIT and EBITDA data. EBIT and EBITDA are commonly used to evaluate the probability of a company. EBIT and EBITDA are released every quarter. To match with our daily oriented dataset structure, we mapped this quarterly sampled data to daily rows by reputation. All the days belonging to the specific quarter will share the same quarterly sampled data;
- Quarterly EBIT and EBITDA data: EBIT and EBITDA are commonly used to evaluate the probability of a company;
- Company-specific vaccine data. This includes the vaccine efficiency rate (from phase 3 results), number of countries grant (emergency) approval, phase 3 results release date, onset of emergency use date, full approval date, and number of shots required. For some key date event data, such as the date of phase 3 results being released, we assigned a corresponding indicator variable z_p for each of them.

$$z_p = \begin{cases} 1 & \text{if the event } p \text{ happened in the last 30 days} \\ 0 & \text{otherwise} \end{cases}$$

4. Experiments

In this study, we compared two different types of machine learning models: MLP with logistic regression and decision trees. The structure of the MLP model we used is shown in **Figure 1**. Decision trees are one of the most commonly used models for stock price direction prediction. We also compared boosting and bagging algorithms with random forest: gradient boosting algorithm and random forest bagging algorithm.

To further explore the potential of the dataset, we implemented online machine learning techniques. In our case, we update the model parameters and decision boundaries while evaluating on each test date. In the real world deployment, as the new data points coming, including new daily COVID statistics and stock

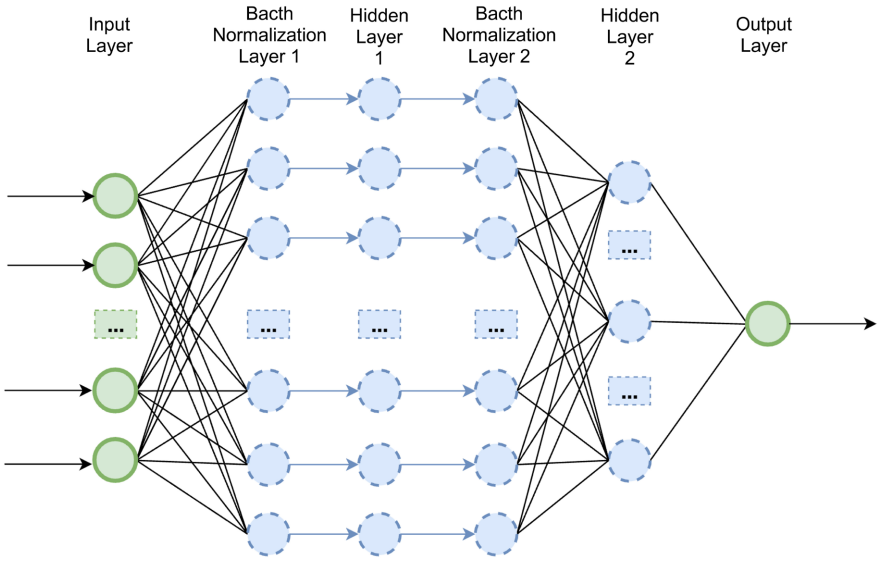

Figure 1. MLP model structure.

price key values, it is important to keep the models being updated to learn the latest connection. The offline learning pattern may hurt the model performance when the new relationships and connections between the features and stock price movements are different from what the models learned using the past dataset. Even though the test dataset is relatively small thus the performance gain from online machine learning techniques may be limited. But we believe this technique will help to keep the model up to date in the real world implementation.

We explored which combination of model and algorithm performs better and results in higher level of AUROC and accuracy. The loss function used by logistic regression and decision tree with gradient descent both are both Binary Cross-Entropy loss. The cross entropy can be use to derive the total entropy between two distributions, which in our case the labels and the predictions. As training process decreases the cross entropy loss, the model learns to make the prediction distribution close to the label distribution.

$$L = \sum_{i=1}^{T} y_i \times \log\left(p\left(y_i\right)\right) + \left(1 - y_i\right) \times \log\left(1 - p\left(y_i\right)\right)$$

After dividing the dataset into training and testing, we inputted those into these three different computational models. For ablation study purpose, we compared the model performance by removing features I_i and C_i^k individually and together with the model performance with full features we collected. We also compared the model prediction performance on full COVID period (2019.11.30-2021.08.30) with later stage of COVID period (2020.10.30-2021.08.30) to examine if dropping the early stage of COVID periods helps the models to perform better. In the model comparison, we also included the baseline of random guessing, which randomly pick 1 or 0 as the output label.

With the setup above, we conducted several empirical tests. The results of decision tree model with gradient boosting (DCGB) are shown below in **Table 2**.

Table 2. Evaluation results of D.T. with G.B.

	AUROC %	ACC %
Random Guess	50	50
Full COVID Prd. W/O COVID & Ind. Feat	61	58
Full COVID Prd. with COVID	64	62
Full COVID Prd. with Ind. Feat	61.4	60
Full COVID Prd. Full Feat	**65.08**	**63.2**
Later COVID Prd. Full Feat	70	66

The baseline DCGB without COVID and Industry index features gives a 58% accuracy, which is relatively lower accuracy comparing to the an average of 59% accuracy of state-of-art FA-CNN model [15]. But their training data include the period far before COVID. In addition to the smaller dataset, the pandemic introduced more unpredictable factors into the stock market, thus a slight decrease of approximately 1% is within our expectation.

The introduction of handcraft features with outside information (COVID and Industry stock price) helps to increase both the metric dimensions: a 4% increase in AUROC and 5.2% increase in accuracy compared to the baseline performance. Comparing to the state-of-art FA-CNN model [15] with the industry stock price, our results with full feature is 3.2% higher than the average accuracy from FA-CNN model with only industry stock price information.

The model performance improvement from the baseline from COVID related features in both the metric dimensions are greater than that from industry stock price feature: 3% increase in AUROC from COVID handcraft features comparing to 0.4% increase in AUROC from industry stock price features; 4% increase in Accuracy from COVID handcraft features comparing to 2% increase in Accuracy from industry stock price features. Figure 2 shows the AUROC improvement with addition of COVID feature, addition of industry stock price feature, and with full features that include both.

The AUROC and Accuracy are further improved after removing the initial period of COVID. With full features including COVID statistics and Industry stock price statistics, AUROC increases by 5% and the accuracy increases by 2.8%.

The results of Decision Tree model with random forests (DCRF) are shown in Table 3. We only include the accuracy in the table as the AUROC does not apply to the outputs from DCRF. The overall performance of DCRF is relatively lower than that of DCGB, but only with a small difference in terms of accuracy. We explored possible reasons behind the performance difference between DCRF and DCGB in the later section.

For the MLP with logistic regression, the results are shown in Table 4. From the results, the introduction of COVID-related statistics increase both measuring metrics: increase AUROC by 10% and increase accuracy by 9%. The industry

stock price statistics improve the metrics performance as well, but with smaller amount. With all of those features, as we expect, the overall performance of the Multi-Layer Perceptron Neural Networks (MLP) obtained a similar improvement. Like the decision tree models, after dropping the early period of COVID, there is a constant enhancement through both measuring metrics. These changes all align with previous models results, even with different absolute values of change.

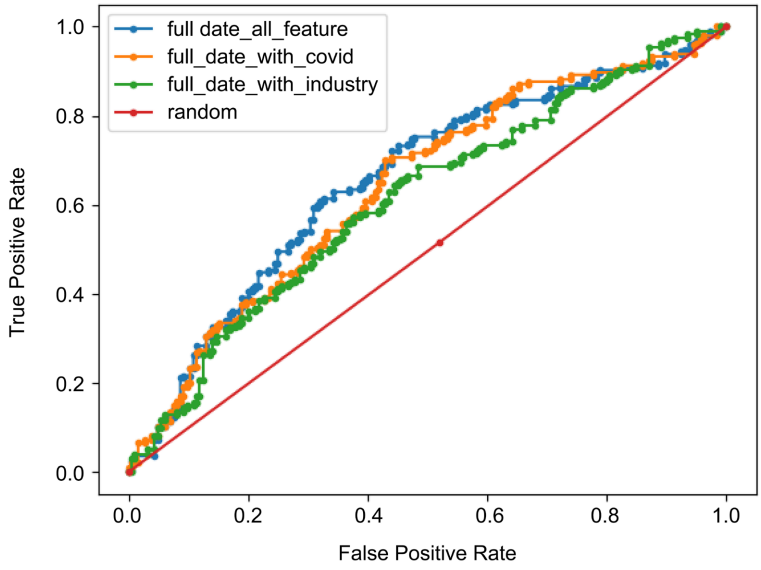

Figure 2. ROC curves of different feature compositions.

Table 3. Evaluation results of D.T. with R.F.

	ACC %
Random Guess	50
Full COVID Prd. W/O COVID & Ind. Feat	52
Full COVID Prd. with COVID	63
Full COVID Prd. with Ind. Feat	56.8
Full COVID Prd. Full Feat	**62.69**
Later COVID Prd. Full Feat	63.35

Table 4. Evaluation results of MLP.

	AUROC %	ACC %
Random Guess	50	50
Full COVID Prd. W/O COVID & Ind. Feat	50	49
Full COVID Prd. with COVID	60	58
Full COVID Prd. with Ind. Feat	53	51
Full COVID Prd. Full Feat	60	56
Later COVID Prd. Full Feat	60	57

We also implemented the online machine learning on full COVID period with Full features across all models. The performance gain by adding online machine learning techniques are shown in Table 5. The result suggests an increase of 1.4% in AUROC for DTGB model. The performance gain for the other two models in AUROC and ACC are relatively smaller. We think it is because of our limited testing data. The online machine learning is most effective to improve the model performance when the new coming data has different connections and relationships between the features and labels comparing to the training data. In our dataset, we think the test data is very similar to the training data because they are close in time. However, in the real world application, the online machine learning techniques is necessary as it helps to avoid keeping outdated relationships and connections learned from the old training set.

The best model in this situation is the decision tree with gradient boosting algorithm (DCGB), as DCGB outperforms other models in both AUROC and accuracy consistently in all feature compositions. In the next section, we will dive into the meaning of performance increment difference between COVID statistics and industry stock price information.

5. Discussion and Analysis

Stock price prediction is a comprehensive task. A huge amount of information can affect the stock market. It became more challenging when the pandemic started. The stock market became more unstable and there were lots of subjective feelings, like anxiety and fears, affecting the stock price. Because of the limited length of observation on COVID effects on the stock market and the limited number of pharmaceutical companies conducting COVID vaccines, the scale of our dataset is not very large. However, with the inclusion of COVID-related handcraft features and other information, we are able to increase our model performance comparable to other state-of-art models. We further take a closer look into the performance gain introduced by those outside-of-company handcraft features.

5.1. Performance Gain

From the results in the previous section, we find out that there are constant increases in both of the measuring metrics with the addition of outside handcraft features, including industry stock price features and COVID statistics features. For the performance gain from industry stock price features, the average gain in AUROC is 1.7%, and the average gain in accuracy is 2%. This is consistent with the accuracy gain introduced by industry index information for FA-CNN model during the non-COVID period [15]. This means that even during the COVID period, the industry index information is still helpful to increase the model performance to predict the stock price change direction.

We also noticed that for all models with different boosting or bagging algorithms, there is a consistent and noticeable increase in performance when we

Table 5. Performance differences with online algorithm.

	Offline	Online
DTGB, AUROC %	65.08	66.48
MLP, AUROC %	60	60.23
DTRF, ACC %	62.69	62.96

only include the COVID statistics handcraft features. The average gain in AUROC is 6.5% and the average gain in accuracy is 7.3%. Comparing to the increase caused by the inclusion of industry features, this gain in performance is significantly greater. This means that during the pandemic, COVID-related status and news does cast an influence on the pharmaceutical stock market. This is perhaps because the COVID-related information is deeply interacted with the reaction and the expectation towards the stock market from the general public, including investors.

5.2. Feature Importance

To examine which features are more important among others, we explored the feature importance after training the DCGB on the full COVID period with full features. The Top 10 important features are shown in Table 6, and the Top 50 feature importance table is available in the Appendix Table A2. The top three most important features are the Africa precious day new cases, current-day stock volume, and Africa precious day new deaths. The current day stock volume's importance is reasonable because it is established factors investors keep tracking and make decisions accordingly. COVID-related data takes 2 of the top three most important features. The most important feature is the precious daily case of Africa. Given the financial condition of most countries in Africa, the testing capability of COVID is relatively low. This low capability may cause to form a high representation of the global COVID situation. More specifically, a small increase in Africa's daily new cases may indicate a greater surge in the global COVID status. The Africa daily new cases and world daily new cases trend is shown in Figure 3. Other continents' COVID statistics, including America and Europe, followed with relatively lower importance. In the top 10 important features (as shown in Table 5), there is only 1 internal financial feature (current day volume) and 9 outside of company features. Within these 9 environmental features, 6 of them are COVID statistics and 2 of them are industry index stock information (NIKKEI 225 INDEX and NASDAQ Biotechnology Index). COVID statistics features are taking more importance weights compared to other features.

5.3. Data Selection by Dropping COVID Early Stage

To further explore the potential of our model performance, we tried to perform data selection by eliminating the early COVID period data. Through all models with different algorithms, there is a constant increase in both measuring metrics

Table 6. TOP 10 feature importance score.

	Feature Importance Score
Africa Previous Day New Cases	24
Current Day Stock Volume	23
Africa Previous Day New Deaths	23
America Previous Day New Cases	19
Daily Price to Book Value	17
Africa Current Day New Cases	17
Europe Current Day New Cases	17
NIKKEI 225 INDEX Current Day Volume	17
NASDAQ Biotechnology Index Current Day Volume	17
Africa Second Previous Day New Cases	15

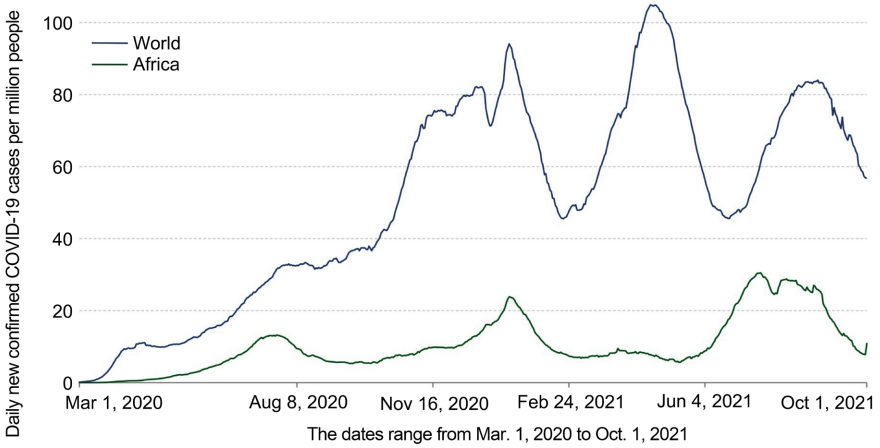

Figure 3. Daily case in africa and world, from our world in data database.

when we removed the initial dates of the COVID period (2019.11.30-2020.10.29). Comparing to the full COVID period with full features, the later COVID period with full features has an average 3.9% increase in AUROC and an average 1.9% increase in accuracy. This is likely because the initial months of the COVID period are more unpredictable as the pandemic just started and the markets are still digesting how the global health crisis may unfold. Because of these anxious and volatile environments, the stock market in the early COVID period was more fragile and more unstable in comparison to the later COVID period. This unsteadiness in the early COVID periods created some noises and augmented correlation between public reaction to COVID and stock market prices. After excluding these relatively low quality data in the early COVID data, we expect the stock price direction is more predictable. The constant increase in all measuring metrics after dropping early COVID periods corroborates with this expectation.

5.4. Performance Difference between DCGB and DCRF

As mentioned before, there is a relatively small but consistent performance dif-

ference between DCGB and DCRF. From Table 2 and Table 3, the accuracy of DCGB is about 1% higher than that of DCRF. This is interesting as the underlying models are both based on decision trees. We believe the reason is related to the nature of the algorithm differences between Gradient Boosting and Random Forest. For the Random Forests, the algorithm trained dozens of fully developed trees to make the prediction. Each fully developed tree could have a high depth, which can help to identify more subtle relationships and connections, which in turn reduces the bias in the prediction. However, the high depths also bring higher variance to the prediction. With a large number of deep and fully developed predictors, Random Forest aims to reduce the variance in the prediction error by aggregating the predictions from all individual predictors, but it could not reduce the bias of the final prediction. The algorithm is very helpful and effective for overfitting problem. On the contrary, the Gradient Boosting algorithm developed many weak learners, which means very shallow decision trees in our case. These weak learners may not identify complicated and subtle connection but it does not deviate from the ground truth relation by its nature. From another perspective, each weak learner brings a relatively higher bias but low variance in the prediction error. The Gradient Boosting algorithm targets to reduce the bias from the prediction error. This means the algorithm is efficient to improve underfitting prediction. In our case, as mentioned before, there are many other factors, including some seemingly unrelated or unquantifiable elements, which are not included in the dataset. Therefore, the relationship prediction between the current features and final prediction may suffer from underfitting rather than overfitting. As a result, Gradient Boosting gives a slightly better performance in terms of accuracy.

6. Conclusion and Future Work

The objective of this paper is to tackle the challenging problem of predicting stock price change direction in the pharmaceutical sub-industry during the COVID period. We explored the performance of different models with different bagging and boosting algorithms. We also examined the performance gain of including a variety of external handcraft features in addition to the conventional internal financial information. These external handcraft features include industry and sub-industry index stock key values and COVID-related statistics.

The result has shown the COVID-related statistics and information, along with related industry index, helps to provide a constant increase in performance across different models in the COVID period. The decision tree with Gradient Boosting algorithm outperforms other models, including MLP with Logistic Regression and decision tree with random forest. The data selection by dropping the first a few months helps the DTGB achieve 70% in AUROC and 66% in the accuracy. The performance gain from online machine learning is less significant. We think it is because of the relatively high similarity between our training and test data sets in the limited time coverage.

In the future, with more data points available, we believe there is a great potential in more complicated machine learning models, specifically LSTM. Given the nature of our dataset (time-series oriented), LSTM could learn from continuous historical trends, instead of dataset rows being learned independently in the current models. As the feature size increases rapidly (189 currently) as we collect more data from different perspectives, feature selection and reduction is another direction that is worth exploring for the next step.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Hong, H., Bian, Z.C. and Lee, C.-C. (2021) COVID-19 and Instability of Stock Market Performance: Evidence from the US. *Financial Innovation*, **7**, 1-18. https://doi.org/10.1186/s40854-021-00229-1

[2] Vijh, M., Chandola, D., Tikkiwal, V.A. and Kumar, A. (2020) Stock Closing Price Prediction Using Machine Learning Techniques. *Procedia Computer Science*, **167**, 599-606. https://doi.org/10.1016/j.procs.2020.03.326

[3] Usmani, S. and Shamsi, J.A. (2021) News Sensitive Stock Market Prediction: Literature Review and Suggestions. *PeerJ Computer Science*, **7**, e490. https://doi.org/10.7717/peerj-cs.490

[4] Aravind, M. and Manojkrishnan, C.G. (2020) COVID 19: Effect on Leading Pharmaceutical Stocks Listed with NSE. *International Journal of Research in Pharmaceutical Sciences*, **1**, 31-36. https://doi.org/10.26452/ijrps.v11iSPL1.2014

[5] Vierlboeck, M. and Nilchiani, R.R. (2021) Effects of COVID-19 Vaccine Developments and Rollout on the Capital Market—A Case Study. https://arxiv.org/abs/2105.12267

[6] Adekoya, A.F. and Nti, I.K. (2020) The COVID-19 Outbreak and Effects on Major Stock Market Indices across the Globe: A Machine Learning Approach. *Indian Journal of Science and Technology*, **13**, 3695-3706. https://doi.org/10.17485/IJST/v13i35.1180

[7] Rouf, N., Malik, M.B. and Arif, T. (2020) A Machine Learning Based Approach to Unleash the Impact of COVID-19 on Indian Stock Market. https://doi.org/10.21203/rs.3.rs-54882/v1

[8] Kara, Y., Boyacioglu, M.A. and Baykan, M.K. (2011) Predicting Direction of Stock Price Index Movement Using Artificial Neural Networks and Support Vector Machines: The Sample of the Istanbul Stock Exchange. *Expert Systems with Applications*, **38**, 5311-5319. https://doi.org/10.1016/j.eswa.2010.10.027

[9] Nelson, D.M.Q., Pereira, A.C.M. and de Oliveira, R.A. (2017) Stock Market's Price Movement Prediction with LSTM Neural Networks. 2017 *International Joint Conference on Neural Networks* (*IJCNN*), Anchorage, 14-19 May 2017, 1419. https://doi.org/10.1109/IJCNN.2017.7966019

[10] McNally, S., Roche, J. and Caton, S. (2018) Predicting the Price of Bitcoin Using Machine Learning. 2018 26*th Euromicro International Conference on Parallel, Distributed and Network-Based Processing* (*PDP*), Cambridge, 21-23 March 2018, 339. https://doi.org/10.1109/PDP2018.2018.00060

[11] Persio, L.D. and Honchar, O. (2016) Artificial Neural Networks Architectures for Stock Price Prediction: Comparisons and Applications. *International Journal of Circuits, Systems and Signal Processing*, **10**, 403-413.

[12] Ramos-Prez, E., Alonso-Gonzlez, P.J. and Nez-Velzquez, J.J. (2019) Forecasting Volatility with a Stacked Model Based on a Hybridized Artificial Neural Network. *Expert Systems with Applications*, **129**, 1-9. https://doi.org/10.1016/j.eswa.2019.03.046

[13] Zhang, X., Zhang, Y., Wang, S., Yao, Y., Fang, B. and Yu, P.S. (2018) Improving Stock Market Prediction via Heterogeneous Information Fusion. *Knowledge-Based Systems*, **143**, 236-247. https://doi.org/10.1016/j.knosys.2017.12.025

[14] Nam, K. and Seong, N. (2019) Financial News-Based Stock Movement Prediction Using Causality Analysis of Influence in the Korean Stock Market. *Decision Support Systems*, **117**, 100-112. https://doi.org/10.1016/j.dss.2018.11.004

[15] Zhang, X., Liu, S. and Zheng, X. (2021) Stock Price Movement Prediction Based on a Deep Factorization Machine and the Attention Mechanism. *Mathematics*, **9**, 800. https://doi.org/10.3390/math9080800

[16] Kotsiantis, S.B., Kanellopoulos, D. and Pintelas, P.E. (2007) Data Preprocessing for Supervised Leaning.

[17] Fawcett, T. (2004) ROC Graphs: Notes and Practical Considerations for Researchers. *Machine Learning*, **31**, 1-38.

[18] Shang, Y. (2016) On the Likelihood of Forests. *Physica A*: *Statistical Mechanics and Its Applications*, **456**, 157-166. https://doi.org/10.1016/j.physa.2016.03.021

Appendix

Table A1. Partial features of one sample in our dataset.

Feature	Value
Label	1.0
Current day open	0.4461791866471421
Current day high	0.4459989331208081
Current day low	0.4574889800622074
Current day adjusted close	0.4479121502946152
Current day volume	0.09400673217699296
First previous day open	0.4461791866471421
First previous day high	0.4459989331208081
First previous day low	0.4574889800622074
First previous day adj close	0.4479121502946152
First previous day volume	0.09400673217699296
Second previous day open	0.4461791866471421
Second previous day high	0.4459989331208081
Second previous day low	0.4574889800622074
Second previous day adj close	0.4479121502946152
Second previous day volume	0.09400673217699296
Third previous day open	0.4372811597358554
Third previous day high	0.4452350888419846
Third previous day low	0.4669823667551453
Third previous day adj close	0.4488203643022887
Third previous day volume	0.06406516721755864
Fourth previous day open	0.4372811597358554
Fourth previous day high	0.4452350888419846
Fourth previous day low	0.4669823667551453
Fourth previous day adj close	0.4488203643022887
Fourth previous day volume	0.06406516721755864
Fifth previous day open	0.4533470774531733
Fifth previous day high	0.4469336179326001
Fifth previous day low	0.4460352422907489
Fifth previous day adj close	0.4396557051278942
Fifth previous day volume	0.09476194433966488
First previous week open	0.7599483884036485
First previous week high	0.1013171403470904
First previous week low	0.7780830538992592
First previous week close	0.6526674346866228
First previous week volume	0.4362714931485329

Continued

Second previous week open	0.7599483884036485
Second previous week high	0.1013171403470904
Second previous week low	0.7780830538992592
Second previous week close	0.6526674346866228
Second previous week volume	0.4362714931485329
Third previous week open	0.7599483884036485
Third previous week high	0.1013171403470904
Third previous week low	0.7780830538992592
Third previous week close	0.6526674346866228
Third previous week volume	0.4362714931485329
First previous month open	0.7599483884036485
First previous month high	0.1013171403470904
First previous month low	0.8166015404072918
First previous month close	0.7548529681037454
First previous month volume	0.5749372182939876
Second previous month open	0.7599483884036485
Second previous month high	0.1013171403470904
Second previous month low	0.8166015404072918
Second previous month close	0.7548529681037454
Second previous month volume	0.5749372182939876
Vaccine whether phase 3 released	0.0
Vaccine whether emergency use	0.0
Vaccine whether approved	0.0
Vaccine num shots	0.017038099322023822

Data sources: yfinance (https://pypi.org/project/yfinance/), ycharts, (https://ycharts.com/), and our world in data (https://ourworldindata.org).

Table A2. Top 50 feature importance score.

	TOP 50 Feature Importance Score
Africa Previous Day New Cases	24
Current Day Stock Volume	23
Africa Previous Day New Deaths	23
America Previous Day New Cases	19
Daily Price to Book Value	17
Africa Current Day New Cases	17
Europe Current Day New Cases	17
NIKKEI 225 INDEX Current Day Volume	17
NASDAQ Biotechnology Index Current Day Volume	17
Africa Second Previous Day New Cases	15
First Previous Week Stock Volume	14
Africa Second Previous Day New Deaths	14
America First Previous Day New Cases	14
Asia Second Previous Day New Deaths	13
Europe Second Previous Day New Cases	13
America Current Day New Cases	11
Europe Current Day New Cases	11
Third Previous Day Stock Adjusted Close Price	10
Fifth Previous Day Stock Volume	10
First Previous Month Stock Volume	10
Africa Previous Week New Deaths	10
Europe Previous Week New Deaths	10
Google Trends Score	10
Current Day Stock Adjusted Close Price	9
America Second Previous Day New Deaths	9
America Current Week New Deaths	9
Europe Second Previous Day New Deaths	9
BSE SENSEX Current Day Stock Volume	9
Current Day Stock Highest Price	8
Previous Week Stock Highest Price	8
Previous Week Stock Lowest Price	8
Second Previous Week Stock Volume	8
Asia Current Day New Cases	8
Africa Current Week New Cases	8
Africa Current Day New Deaths	8
Africa First Previous Month New Deaths	8
America Current Day New Deaths	8

Continued

America Current Day New Deaths	8
Vanguard Health Care Fund Admiral Shares Stock Volume	8
Fifth Previous Day Stock Adjusted Close Price	8
Company-Specific Quarterly EBIT Value	7
Asia Second Previous Day New Cases	7
America First Previous Month New Deaths	7
Europe First Previous Week New Deaths	7
Europe First Previous Week New Deaths	7
EURO STOXX 50 Current Day Stock Volume	7
Third Previous Day Stock Lowest Price	6
Third Previous Day Stock Volume	6
First Previous Week Stock Opening Price	6
Africa Previous Week New Cases	6