Using Machine Learning Techniques to Study Economic Trends in Various U.S. Industries in the Post-Epidemic Era

Hongyi Tang
School of Economics and Management, Beijing University of Aeronautics, 2522977507@qq.com

ABSTRACT

The aim of the project is to predict and analyse broad trends across the US economy using stock data from mainstream companies in six industries on Forbes 2000 and data from COVID-19. A time series analysis approach was used to predict the daily increases in each company's share price. The following five supervised learning techniques (logistic regression, random forest, decision tree, neural network and XGBoost) were used. As the accuracy of the results predicted by the different models for each company varies considerably, only the results predicted by the most accurate model for each company have been selected for analysed. The results show that the Electronic Pleased Technology Industry and the Social Entertainment Internet Industry remain break-even for COVID-19; the E-Commerce Industry shows a significant increase; The Financial Services Industry shows a significant drop in share price, while the Insurance Industry and Pharmaceutical Industry show a small drop in share price.

Keywords: machine learning, COVID-19, stocks, prediction

1. INTRODUCTION

As a barometer of the economy, the stock market is an important part of a country's economy [1], so it can be used as a good indicator reflecting the social and economic conditions. This is one of the biggest investment chances for enterprises and investors, small businesses, brokerage firms, and banking departments all rely on this agency to generate revenue and distribute risk [2], while stock traders need to predict trends in stock market behavior in order to make the right decision to sell or hold what they own or sell shares. If stock traders correctly predict stock price trends, they can achieve considerable profits [1]. So accurate prediction is very important for stock traders. For businessmen mixed in financial markets, the predicted stock price can also reflect the future economic trend of the company and even the industry, and can make corresponding solutions in advance.

Yet stocks market fluctuated volatile [3]. Affected by multiple factors. In particular, in 2021, COVID-19 was arguably one of the largest environmental factors affecting the stock market, and Covid-19 is a sudden and unfamiliar disease that primarily affects the respiratory system [4]. The disease gradually worsened and engraved the world. This poses a huge challenge to the economies of various countries and industries. People from all walks of life have also been affected by economic disasters to varying degrees, thus predicting that subsequent economic and social trends under COVID-19 will be important to guiding their subsequent decisions. In order to accurately predict the stock price, I have to consider this environmental factor.

To make profits, stock traders need to buy stocks that are expected to rise in the near future and sell those stocks that are expected to fall [1]. So I chose the top flow companies in the top six industries in Forbes 2000 to represent the overall economic trends of various industries to reflect the overall economic situation of the United States

Stock factors, which reflect the properties and qualities of a stock, are mostly used to evaluate the performance of a stock. For example, low, open, volume, high, close, adjusted close, which come from the financial statements of the company issuing the stock. These factors are robust and reliable, as they directly reflect the value of the company.

COVID-19 factors, which reflect the attributes and qualities of an outbreak, can be used to assess the severity of an outbreak. Examples include confirmed, deaths, recovered, active, incident_rate, total_test_results, case_fatality_ratio, and testing_rate, which are derived from data from the Johns Hopkins CSSEGIS.

Traditional stock prediction methods are using statistical principles and econometric models to depict financial time series data, such as the autoregressive conditional heterovariance model (ARCH), the autoregressive differential moving average model (ARIMA), etc [5]. Many scholars conduct regression modeling based on the time-series data of the historical stock price trends, and use regression models to predict short-term changes in stock prices. For example, Wu

Yuxia and Wen Xin [6] used ARIMA model to predict the stock price changes of Huatai Securities, which achieved good results in the short term; Yang Qi and Cao Xian [7] used ARMA model and GARCH model to predict the stock price of Volkswagen Public (600635) model, which achieved high prediction accuracy. However, the traditional regression methods have high requirements for the sample size and the distribution degree, and it is difficult to guarantee the accuracy and stability of the prediction, resulting in the lack of universality of the results. Therefore, the applicability is very limited [5].

In the context of big data, machine learning technologies are developing rapidly, and are being widely used in various fields. Machine learning algorithm is the process of conducting multi-dimensional statistical analysis of massive data, and the machine itself removes the interference information and makes correct decisions [5]. With the assistance of machine learning, previous emotional judgments and irrational decisions can be replaced by rigorous predictions by various learning algorithms based on quantitative data. Moreover, hidden features in massive data can be extracted by these algorithms, while individual judgments may overlook some key points. Therefore, this project relies on machine learning algorithms for prediction, which is determined by its objectivity and powerful decision-making ability.

And the development of machine learning has a strong dependence on data. Among many data, time-series data is an important and complex data type, which plays an important role in various industries. Time-series data, i.e., time-series data, is data that is arranged according to the sequence of time occurrence and has a strong temporal correlation. Time-series data forecasting method is to predict the long-term trend of data by capturing the pattern between historical time-series data.

This paper applies machine learning techniques to stock price prediction to overcome these difficulties. Since these data are large and highly complex, more efficient machine learning models are always needed for day-to-day predictions [8]. Therefore, the accuracy of each model is compared, and each company is given the highest accuracy model to predict the results. These models are based on five supervised learning techniques, namely linear regression, random forest, decision tree, neural networks and XGBoost.

Taken together, this project aims to predict and analyze trends across the U. S. economy, using stock data from mainstream companies in multiple industries on Forbes 2000. The motivation is to use time-series analysis to produce a relatively accurate prediction under a relatively objective and robust approach. Specifically, the theoretical basis of the project and the data used are first discussed.Then, the features and modeling procedures used for the analysis are summarized in detail. Finally, this conclusion is presented.

2. DATA

This section evaluates the data used in this project from different perspectives. It will first describe the data sources and reliability. Then, it presents the detailed variable data used for further analysis for further analysis. Next, it shows the empirical distribution of these variables and highlights the importance of data cleaning. Finally, it shows the results after cleaning.

2.1 Data Source

All financial data used in this project were extracted from Kaggle. The data is first extracted into CSV files. Then, these data become local data and can be imported into R and Python indefinitely. And all the data of COVID-19 used in this project are extracted from CSSEGISandData and go through data cleaning, pre-processing, including outlier processing, missing value processing, normalization processing, etc., and then all the data are integrated in chronological order. Finally, it is transferred to CSV text, which becomes local data and can be imported into R and Python indefinitely.

2.2 Data Reliability

The data obtained from Kaggle is considered to be very reliable, and it is a trusted website that provides a platform for developers and data scientists to hold machine learning competitions, host databases, write and share code, and has the attention of 800,000 data scientists, archives and BigQuery, which are basically directly usable datasets.

The data obtained from CSSEGISandData is also considered to be very reliable and comes from the Johns Hopkins CSSE Global Data, which have an impressive dashboard detailing the global COVID-19 case data. These data are updated regularly, thus giving everyone a global view of the spread of the disease and its mortality, and are therefore very reliable and extensive.

2.3 Data Variables Introduction

In summary, the data collected from Kaggle and CSSEGISandData can be used directly in future processes without further validation.

The first category of variables is the basic information about the stocks I am looking at. Instead of using a single stock for forecasting, we selected six stocks from the Forbes 2000 above to represent six different sectors of the industry, AAPL (Apple Computer inc.) Electronic Pleased Technology Industry, AMZN (Amazon) E-Commerce Industry, BRK-A (Berkshire Hathaway) Insurance Industry, PFE (Pfizer, Inc) Pharmaceutical Industry, FB (Facebook) Social Entertainment Internet Industry, and JPM (JPMorgan Chase & Co.) Financial Services Industry. I use the economic trend of these stocks to represent the economic development of each industry. There are two advantages of using this value: 1. These data are recorded in Forbes 2000, the data are real and reliable and complete, and as the leading companies in the industry, they are the most convincing for the economy of the industry. 2. The variables selected are the most basic information of the stocks, such as Low, Open, Volume, High, Adjusted Close, creation of Daily_ Amplitude, Average7_Close, Max7_Close, Min7_Close, Varp7_Close. these values give a good picture of the basic situation of a stock. Max7_Close, Min7_Close.

second category of variables is the basic information about COVID-19 I am interested in. I use the values Confirmed, Deaths, Case_Fatality_Ratio, Incident_Rate, Total_Test_Results for the epidemic from January to August 2021 as impact factors to study the impact of COVID-19 on the stock market economy. There are two benefits to using these values: 1. These data are from the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) data storage project that collects publicly available COVID-19 from national health agencies. The project is open source for educational and academic research purposes, and also provides a visual dashboard web interface that is very reliable. 2. The variables selected are the most basic information in the information of COVID-19, such as Confirmed, Deaths, Case_Fatality_Ratio, Incident_Rate, Testing_Rate. These values are a good reflection of the basic situation of COVID-19 over time. All the variables and their respective definitions are seen in the Table 1.

Table 1. Detailed interpretation of all the variables.

Stock Variable	
Low	The lowest price of the stock for the day
Open	The opening price of the stock on that day
Volume	Stock Daily Volume
High	The highest price of the stock for the day
Adjusted Close	Adjusted closing price of the stock for the day
Average7_Close	Average closing price of a stock in seven days
Max7_Close	Stock maximum in seven days Closing Price
Min7_Close	Minimum closing price of a stock in seven days
Varp7_Close	The variance of the closing price of a stock over a seven-day period
Daily_Amplitude	Stock Daily Amplitude
COVID-19 Variable	
Confirmed	Number of people who confirmed COVID-19
Deaths	Number of people who died from COVID-19
Case_Fatality_Ratio	Proportion of all cases and deaths under COVID-19
Incident_Rate	Obtain the incidence rate of COVID-19

Testing_Rate	Ratio of positive tests to total tests per day during COVID-19
Total_Test_Results	Number of nucleic acid tests per day during COVID-19

2.4 Data Adjustments

Since COVID-19 data were downloaded by individual U.S. states on a single day, they need to be adjusted to sort by order of day for the entire year 2021, and then use the total for the entire U.S. as the specific value to derive the CSV table.

The stock values are only recorded during the opening of the market, so if you need to study the impact of the epidemic on the stock market, you need to correspond the times one by one, and this problem is handled here using the R language merge function.

Then since some of the data were rates and some were the number of people who obtained the epidemic across the United States, the values between the variables were too far apart, so the data were normalized again using R to come up with a usable data. Here is how the normalization was done:

Normalization involves subtracting the series mean and divided by series standard deviation. Then the series becomes standard and can be approximately regarded as a normal distribution, which facilitates most analysis. The detailed formula is listed below.

$$x_{ij,new} = \frac{x_{ij} - \overline{x_i}}{\delta_i} \quad (1)$$

There are no missing values in the dataset, so the missing value adjustment link is omitted.

3. METHODOLOGY

This section describes the details of the basic modeling part. It first introduces the model settings and modeling procedures. Then, it shows the modeling details for the 5 basic models. Finally, it displays the prediction results for those 5 models.

3.1 Goals and Dataset

Using the model to predict the daily increase of the stocks, I used the Daily_Increase of the data as the dependent variable, used 80 percent of the data. As the test set, and the remaining 20 percent of the data. As the validation set to predict the future situation of each stock, then predicted the economic situation of the industry according to the total predicted growth. The detailed settings are shown in the Table 2.

Table 2. Training set and test set setup

Dataset	Duration	No. of Obs	Functions
Training Set	2021.01.05-06.30	123 (random 80%)	Train the models
Test Set	2021.07.01-08.13	31(random 20%)	Test the algorithm performance

3.2 Models

There are 5 basic classification models applied to make predictions including logistic regression, Decision Tree, Random Forest, Neural Network and XGBoost.

LR (logistic regression) is a classical classification algorithm [10], which is mainly used for secondary classification problems. The LR classifier (logistic regression classifier) uses a linear combination of the sample features as the independent variable using the logistic function to (0,1) with the aim to obtain a 0 / 1 classification model by learning from the training data features [9].

Decision Tree is a decision analysis method to evaluate the probability of net present value by forming the decision tree, which is a graphical method to intuitively use probability analysis.Because this decision branch is drawn much like the branches of a tree, it is called a decision tree.

Random Forest (random forest (RF) is an integrated learning model based on Bagging ideas by constructing multiple learning learners to jointly complete the learning task, originally proposed by Ho [10]. The classification results most predicted by all the base learners in the classification task as the final result of the model. To ensure the diversity of decision trees, each decision tree was sampled put back from the entire training sample and randomly selected some features for training, without pruning. Often decision trees are susceptible to extreme values, and RF occurs by combining multiple decisions tree, using random row sampling and column sampling methods reduces the variance of the individual classifier, thus improves generalization ability and is more robust to noise effects [9].

Artificial neural network is a biological neural network in some simplified sense of technical recovery, as a discipline, its main task is according to the principle of biological neural network and the need of practical application to build a practical artificial neural network model, design the corresponding learning algorithm, simulate a certain intelligent activity of the human brain, and then technically implemented to solve practical problems.

XGBoost (XGB) is an efficient implementation of the gradient lifting tree (GBDT) and can be used to solve the classification and regression problems [11].

4. RESULTS

4.1 Predictions of each model for each company

Different models predict the stock changes of each company, and finally derive the mean square error for all models of each company, as shown in Table 3.

Table 3. mean square error for all models of each company.

Modeling	**AAPL**	**AMZN**	**BRK-A**	**FB**	**JPM**	**PFE**
Logistic Regression	0.7254283	1.640219	2.441206	10.79851	4.025731	0.5872132
Decision Tree	4.317612	4.244714	3.975448	4.225854	3.374221	2.761231
Random Forest	3.347021	3.225405	3.619958	4.751116	2.888293	3.517503
Neural Network	1.578247	3.133008	1.432003	6.080501	2.976331	4.895285
Xgboost	1.884418	2.687175	2.973089	5.659643	1.859703	2.620883

4.2 The most accurate prediction results

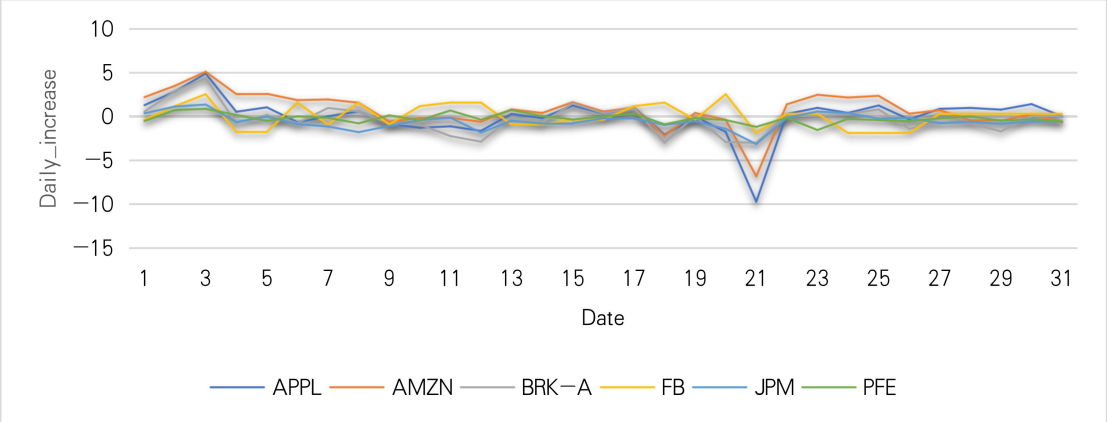

Figure 1. Each company's most accurate prediction results in stock market chart.

The monthly growth (31 days) of APPL was 1.74%, hardly any big change. It can be predicted that the economic operation of AAPL will be stable in the post-epidemic era, which also represents that the electronic glad technology industry is not too much affected by COVID-19.

The monthly growth (31 days) of AMZN was 23.55%, the huge rise, which can be predicted in the post-epidemic era, will be a gradual rise in the economic operation of AMZN, which also represents a more profitable e-commerce industry after being affected by COVID-19

The monthly growth (31 days) of BRK-A was -8.61%, the small decline. It can be predicted that in the post-epidemic era, economy of BRK-A will be operating at a slight loss, which also represents the impact of COVID-19 on the insurance industry, resulting in a small loss.

The monthly growth (31 days) of FB was 2.53%, hardly any big change. It can be predicted that the economic operation of FB will be stable in the post-epidemic era, which also represents that the social entertainment internet industry is not too much affected by COVID-19.

The monthly growth (31 days) of JPM was -16.30%, substantial loss. It can be predicted that in the post-epidemic era, economy of JPM will operate at a significant loss, which also represents a significant economic decline in the financial services industry after the impact of COVID-19.

The monthly growth (31 days) of PFE was -6.24%, the small decline. It can be predicted that in the post-epidemic era, the economy of PFE will be operating at a slight loss, which also represents the impact of the epidemic on the pharmaceutical industry, resulting in a small loss.

5. DISCUSSION AND CONCLUSION

5.1 discussion

AAPL has been able to maintain a steady and small growth despite the market decline, most likely due to the strategy of AAPL that shifting from hardware to software. While COVID-19 set back the hardware, the software business was unaffected, which may be a rainy day for AAPL. I can also see from the most accurate model---linear regression, that the effect of COVID-19 variables on the dependent variable are almost non-existent, which directly indicates that the impact of COVID-19 on AAPL is minimal.

The performance of AMZN can be considered strong during COVID-19, due in large part to COVID-19 consumption dividend, which stimulated the company's sales growth as people were more willing to shop at home, and cloud-based marketplace of AMZN became a major profit maker during COVID-19. Although the variables of COVID-19 do not have much to do with the dependent variable from the most accurate model---linear regression. I think it is caused by insufficient data on the variables.

COVID-19 has led to an economic downturn for most companies, which has seriously affected the economic performance of BRK-A. For example, the company's auto insurance company has proposed a premium credit strategy to face the impact of COVID-19. And I see from the results predicted by the most accurate model---neural network, the predicted situation will also decline by about 8%.

For Facebook, COVID-19 did lead to a shrinkage of the advertising business, which led to a decline in profits, but then the impact of COVID-19 led to a deeper reliance on online connectivity methods, which in turn led to profitability, so on a consolidated basis, economic of FB situation is running smoothly without much fluctuation. And as I can see from the results predicted by the most accurate model---decision tree, the forecast is about a small increase, about 2.5%。

As a result of COVID-19, there are most Americans are losing their jobs, which leads to a significant economic decline in JPM, and the banks have to overdraw their current money to protect their customers, so the economic state of JPM is very bad during COVID-19. And from the results predicted by the most accurate model---XGBoost, JPM has to continue the economic downturn for some time.

It seems a bit odd that economic situation of PFE did not fluctuate much, although during COVID-19, the general public thought that the pharmaceutical industry would certainly gain a lot of profit. I think there are two points to explain this problem, firstly, the storage conditions of the vaccine are so harsh that it causes most of the good to be wiped out, and secondly, the profit of COVID-19 vaccine does not play a big role in the overall profit of a large pharmaceutical company like PFE. The most accurate model---linear regression, shows that only one epidemic variable, the number of

deaths, has a slight effect on the dependent variable, so the economic situation of PFE will not fluctuate much in the subsequent time.

Significance: 1. The study observes the economic situation of various industries in the U.S. in the post-epidemic era, giving some guidance to businessmen from all walks of life. 2. It is for market watchdogs, as during the post-epidemic period, there may be people who are engaging in insider trading being covered up under the guise of COVID-19.

5.2 future research directions

5.2.1 Dataset

The stock market data I have obtained is from Jan to Aug in 2021, but since the stock market has a lot of closed hours and the independent variables selected are the most basic information, I feel that the data is not particularly complete. In the case of stock market variables, I can add the lowest price in seven days, the highest price in seven days, the average opening price in seven days, etc. In the case of COVID-19 variables, I can add the average number of diagnoses per month, the average number of deaths per month, the average death rate per month, etc. The variables of interest can be added to improve the accuracy of the overall model prediction.

5.2.2 Ensemble of models

So far we have only talked about the single ability of different model, but I think it is possible to try to assemble different them to improve the accuracy. I can try the following 4 methods: (a) Combining prediction results of all the 5 models and use arithmetic mean to make final prediction. (b) Combining prediction results of all the 5 models and use the weighted mean according to the reciprocal of the MSE to make final prediction. (c) Combining prediction results of the top 2 models with the highest accuracy to make final prediction.

5.2.3 Additional ideas

Due to the uncertainty and volatility of the stock market, in order to get more accurate forecast results, I need to start not only from the stock itself, adding the general environment data of the epidemic only increases its accuracy a little, I also need to add more influencing factors, such as the political situation in various countries, such as war, economic war, etc.; various celebrities triggered by the effect, such as the Mast currency effect, so-and-so release, etc.; various environmental factors around the world, such as global warming, earthquake and tsunami, etc. The more factors I add, the higher the accuracy I can get, and I think this is a direction that is worth studying in depth.

References

[1] Khan, W., et al. (2020). "Stock market prediction using machine learning classifiers and social media, news." Journal of Ambient Intelligence and Humanized Computing(prepublish).

[2] Pahwa, K. and N. Agarwal (2019). Stock Market Analysis using Supervised Machine Learning. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon).

[3] Bastianin A, Manera M (2018) How does stock market volatility react to oil price shocks? Mach Dyn 22(3):666–682.

[4] Tahir, H., et al. (2021). Forecasting COVID-19 via Registration Slips of Patients using ResNet-101 and Performance Analysis and Comparison of Prediction for COVID-19 using Faster R-CNN, Mask R-CNN, and ResNet-50. 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT).

[5] Zhang Shijie (2021). " The Application of machine learning under big data in stock market forecasting. "Guiyang College Journal (Social Science Edition) 16 (04): 43-48 [in chinese].

[6] Ni Yunwei. Civil Law Analysis, Applications, and Implications of Smart Contracts under Blockchain Technology [J]. Journal of Chongqing University (Social Sciences Edition), 2019 (3): 170-181.

[7] Primavira de Phillippi, Aaron Wright. Regulatory Blockchain: Code Governance [M]. Wei Dongliang, translation. Beijing: Citic Publishing Group, 2019:75.

[8] Vazirani, S., et al. (2020). Analysis of various machine learning algorithm and hybrid model for stock market prediction using python. 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE).

[9] CAI Jingbo and CAI Zhijie (2020). " predicts high transfer in the A-share market based on a machine learning model. "Mathematical modeling and its Application 9 (04): 74-84 [in chinese].
[10] Li Hang. Statistical learning methods [M]. Beijing: Tsinghua University Press, 2012.
[11] Zhu Husi. artificial intelligence [M]. Version 3. Beijing: Tsinghua University Press, 2012.