

Evaluating the social impact of COVID-19 with a big data approach

Jiacheng Chen^{1,*}

¹International school, Beijing University of Posts and Telecommunications, Beijing, 102200, China

Abstract. According to the CNN news, until the first day of year 2021, the total number of COVID-19 infections in the U.S. has exceeded 20 million and resulted in 350,000 deaths. A review of the literature shows that COVID-19 has created a huge crisis in various industries such as offline department stores, tourism, airlines, and restaurants, but also contributes to the online service industry, medical and biopharmaceuticals. The quantitative assessment of the social impact of COVID-19 is based on various types of data. In this paper, stock prices of listed companies are used as indicators to explore the impact of the epidemic on stock prices, which further reflects the impact on different industries. Since the infection information and stock price data of listed companies are easily accessible, this article combines these data and conduct two analyses: correlation analysis and performance analysis, taking 468 listed companies in the U.S. stock market. In the correlation analysis, it is confirmed that the impact of COVID-19 on different industries or companies is different. In the performance analysis, this article predicts the performance of company stock prices before and after the outbreak by using different companies' basic information and find that the XGBoost model works best in the 2-classes case and the random forest model works best in the 5-classes case.

1 Introduction

Since the global outbreak of COVID-19, the United States has been negligent in protecting the citizens, leading to a growing domestic outbreak. As of Jan. 19, 2021, more than 400,000 people have died in the U.S. from the COVID-19 outbreak, with a daily average of upwards of 2,800 deaths over the past two months, the equivalent of one 9/11 terrorist attack per day. The latest statistics released by NBC show that the rate of human death infected with COVID-19 appears to be accelerating. According to the University of Washington's Institute for Health Metrics and Evaluation, more than 115,000 more deaths could occur next month. COVID-19 is a serious problem for the United States and all of humanity.

COVID-19 has also been a devastating economic blow to the whole country. Industries have suffered billions of dollars in losses, make the record of the worst number in a decade. A total of 424 large companies in the U.S. filed bankruptcy petitions in 2020. Yet the impact on large companies is just the tip of the iceberg; there are a large number of self-employed or home-based businesses that are hard to count, and this loss truly hurts families and individuals.

Recreational activities and non-essential consumer goods industries are the worst affected. Since March 15, the global Disneyland suspended operations, the U.S. Disneyland will lay off 28,000 employees, accounting for 25% of the park's total workforce. On Aug. 21, 2020, American Airlines announced plans to temporarily suspend flights to 15 U.S. cities during its "October

schedule" from Oct. 7 to Nov. 3, citing a lack of capacity in those areas due to tourism disruptions caused by the COVID-19 outbreak, but did not indicate when service would resume. Online retail giant Amazon reported its latest quarterly earnings on Oct. 29, 2020, and the company expects its operating income to be between \$1 billion and \$4.5 billion in the fourth quarter of the year. This includes about \$4 billion in costs related to the COVID-19 outbreak for precautionary measures such as testing, cleaning and ensuring social proximity.

Although there have been some corresponding analyses of the impact of COVID-19 on society, they are often based on a particular industry or sector. This paper aims to investigate the impact of the COVID-19 on different industries and analyze the impact of COVID-19 on stock prices based on US outbreak data and stock price data of listed companies in the S&P 500, reflecting the impact on different industries. Daily stock close price data is used for a year to demonstrate the impact of the COVID-19 on the stocks of S&P 500 companies by means of data analysis to better reveal the extent of the impact of the COVID-19 on different industries. Specifically, this article conducted the correlation analysis and performance analysis, and found that: in the correlation analysis, some companies were more affected by the outbreak and therefore had a greater correlation between stock price changes and the number of new diagnoses per day; in the performance analysis, the researcher used basic information about different companies to predict the performance of their stock prices before and after the outbreak, and found that the XGBoost model worked best in the 2-classes case, while the random forest model worked best in the 5-classes case.

*Email: chen_jc0311@bupt.edu.cn

This paper is then organized as follows: Section 2 is an introduction to related work; Section 3 is the correlation analysis; Section 4 is the performance analysis; and Section 5 is the conclusion of the paper.

2 Related Work

In this section, some related work is presented. The impact of COVID-19 on the food industry is analyzed in [1]. After comparing the food retailing and food service sectors in Canada, it is found that roughly 30% of the food dollar that Canadians have been spending on food away from home has shifted to retail [1]. Using the Retail Food Environment and Customer Interaction Model to describe the impact of COVID-19 in state, customer experience etc., the impact of COVID-19 on the healthy food retail is analyzed in [2]. It is found that a more just and equitable RFE is required. The impact of COVID-19 on the U.S. electricity demand and supply is analyzed in [3]. By examining electricity demand in three states, California, Florida and New York, it was found that the impact of the epidemic on electricity demand varied significantly between regions, with an increase of approximately 10% in the cities analyzed [3].

COVID-19 also has a controversial impact on human mobility. A questionnaire survey on the bike sharing usage was carried out in Greece with 223 people involved, showing that COVID-19 will not have a significant impact on the number of people using shared bikes for trips. And for some users, bike sharing is now more attractive [4]. The impact of COVID-19 on the flight networks is analyzed in [5]. They used Opensky's network data to clarify flight patterns and flight densities. In the second half of March 2020, the number of daily flights gradually decreases and abruptly drops by 64%. During this period, the global flight network density drops by 51% [5]. By analyzing the taxi travel data set, the impact of COVID-19 on urban mobility is evaluated from the perspective of taxi travel and social vitality [6]. The taxi travel volume decreases significantly, while the travel speed, travel time and the spatial distribution of taxi trips are significantly affected by the epidemic situation, and social vitality may take 3-6 months as a normal period to fully recover [6]. COVID-19 led to a drop in the demand for transport services, including city public transport. The impact of COVID-19 on the sustainability of transport system of large Russian cities is presented in [7]. Basic methods for assessing the sustainability of transportation services are also presented, with particular reference to urban passenger public transport (CPPT) [7].

In addition, COVID-19 has a great impact on the environment and energy industry. The impact of COVID-

19 on air quality in the Guanzhong Basin is analyzed in [8]. COVID-19 restricted human mobility and non-essential economic activities, which, as a side effect, resulted in the reduction of the emission of pollutants and thus the improvement of the air quality in many cities in China. The PM2.5 concentrations decreased substantially during the lockdown period, with a strong initial decrease and a slower one thereafter [8]. With less travel caused by the lock-down policies internationally and domestically, the motor gasoline demand is highly affected by the COVID-19 pandemic in US. This phenomenon is analyzed in [9]. The gasoline demand went negative in April and grows slowly with a quick rebound in May. However, it is impossible to recover fully before October. Two situations are discussed in [9], namely, the optimistic scenario in which the demand will go back to the non-pandemic level and the pessimistic scenario with the opposite assertion. The impact of COVID-19 on air quality and health in Brazil is analyzed in [10]. After 90 days of isolation, air pollution decreased significantly by 45% for PM10, 46% for PM2.5 and 58% for NO2 [10]. The U.N. reports that tourism losses will range from \$910 billion to \$12 trillion in the wake of the COVID-19 pandemic. In a policy brief released in August 2020, COVID-19 and Transforming Tourism (COVID-19 and Transforming Tourism), the U.N. predicts that international tourist numbers will decline by 58 to 78 percent.

3 Correlation Analysis

3.1. Dataset Description

This study used data from the U.S. epidemic, from January 22, 2020 until December 31, 2020, for the daily increase in the infected number of people in the epidemic. The data source is publicly available: <https://github.com/CSSEGISandData/COVID-19/>

This public dataset contains epidemic data for all countries in the world, but this study used only the U.S. data from it and calculated the number of daily additions during the epidemic. This study also uses the complete historical trading data of 468 companies indexed in S&P 500, which is obtained from Yahoo Finance, the study also obtains the data of open price, high price, low price, close price, and trading volume. The daily change of the close price is mainly used.

Since the epidemic data is updated daily, while the stock data is only recorded on trading days, the dates used in our study are the intersection of the two date sets. A sample format of the stock dataset is shown in Table 1.

Table 1. The stock dataset format.

Date	Open	High	Low	Close	Adj Close	Volume
2020/1/21	79.297501	79.754997	79	79.142502	78.464882	110843200
2020/1/22	79.644997	79.997498	79.327499	79.425003	78.744957	101832400
2020/1/23	79.480003	79.889999	78.912498	79.807503	79.124184	104472000
2020/1/24	80.0625	80.832497	79.379997	79.577499	78.896149	146537600
2020/1/27	77.514999	77.942497	76.220001	77.237503	76.576187	161940000
2020/1/28	78.150002	79.599998	78.047501	79.422501	78.742477	162234000

3.2. Results and Discussions

This study calculate the standard correlation between the epidemic daily increases and daily change of close price.

After this correlation analysis, it is found that the companies with the strongest positive correlation include Regeneron (38.77%), NVidia Corporation (64.25%), PayPal (50.75%), Amazon.com Inc. (45.27%). Their close prices are shown in Figure 1.

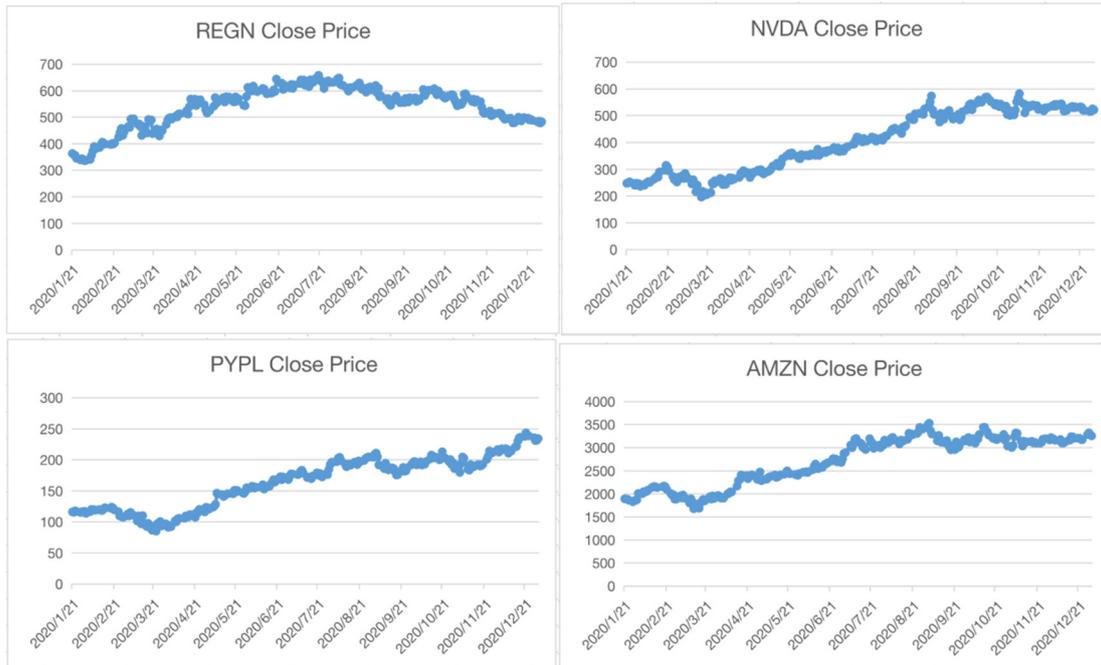


Figure 1. The close prices of REGN, NVDA, PYPL, AMZN.

The companies with the strongest negative correlations include Macerich (-62.94%), Occidental Petroleum (-59.86%), Coty, Inc (-54.73%), United Continental

Holdings (-50.81%). Their close prices are shown in Figure 2.

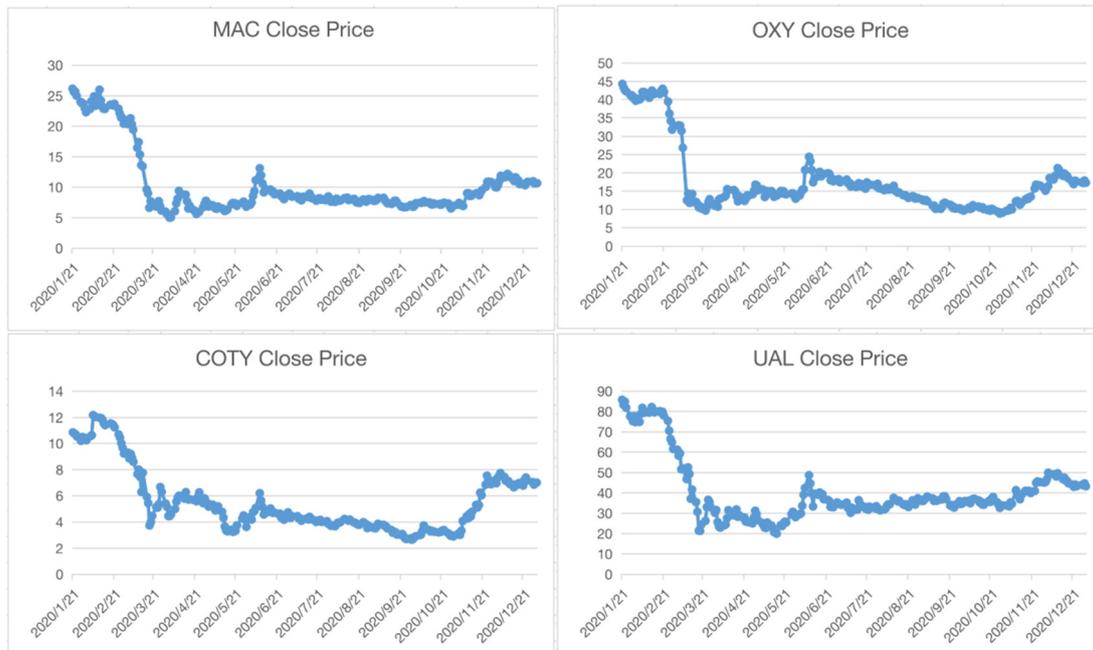


Figure 2. The close prices of REGN, NVDA, PYPL, AMZN.

Furthermore, this study have selected a few of them to analyze the reasons behind them.

- Regeneron, a health care company, saw its share of

stock grow as people were more willing to invest in their own and their families' health management during the epidemic

- Alipay is a financial management company commonly used in China. Since it only needs to be operated online through cell phones, people do not need to go out to banks, so Alipay has become a common choice for people's financial management during the epidemic, showing a positive correlation.

- Macerich, a real estate company, is negatively impacted by the epidemic as people do not focus too much on real estate investments during the epidemic.

- Occidental Petroleum's stock is trending downward as people's travel is restricted during the epidemic and the extraction and use of oil and energy is reduced.

- For United Continental Holdings, the U.S. travel ban on Europe in March and the interruption of people's travel plans during the epidemic led to a significant reduction in air travel and other airline travel, resulting in a sharp decline in the company's revenue, which was negatively correlated with the severity of the epidemic as evidenced by the stock price.

4 Performance Analysis

4.1. Dataset Description

The data obtained from Yahoo finance for the S&P500 is incomplete, and only 468 companies with complete historical trading data were actually used, of which 433 companies had basic information available for classification. Therefore, we obtained the basic information of 433 companies based on Yahoo Finance, including the following fields:

- Symbol: company abbreviation
 - Name: company name
 - Sector: industry in which the company operates
 - Percentage Increase/Decrease: percentage increase/decrease due to the epidemic
 - Full Time Employees: total number of employees
 - Enterprise value: market value of the company
 - State: location in the United States
- A sample format of the dataset is shown in Table 2.

Table 2. The stock basic information format.

Symbol	Name	Sector	Percentage Increase/Decrease	Full Time Employees	Enterprise Value	Profit Margins	Market Cap	Ask Size	State
MMM	3M Company	Industrial	0.561556	96163.0	1.120620e+11	0.15653	9.611008e+10	1000.0	MN
AOS	A.O.S. Smith Corp	Industrial	14.186421	15100.0	8.718930e+09	0.11246	9.068043e+09	800.0	WI
ABT	Abbott Laboratories	Health Care	15.444266	107000.0	2.118540e+11	0.10496	1.978130e+11	800.0	IL
ABBV		Health Care	3.616482	47000.0	2.685360e+11	0.18157	1.893820e+11	2200.0	IL
ACN	Accenture plc	Information Technology	8.465193	514000.0	1.628110e+11	0.11739	1.750220e+11	800.0	NaN

We then analyzed the change in the average stock price of different companies before and after the outbreak (before and after March 2020), and based on this change, the researcher classified the performance of companies into two and five categories, constituting 2- and 5-classes classification problems, respectively.

(1) 2-classes classification problem: Increase and Decrease; for the data in the Percentage Increase/Decrease column, the division interval is [-100,0) for Decrease and (0,100] for Increase.

(2) 5-classes classification problem: Major Decrease, Minor Decrease, The same level, Minor Increase, Major Increase. The corresponding intervals are [-100,-30), [-30,-10), [-10,10), [10,30), [30,100).

Machine learning models are used to solve the classification problems, which have been proven effective in a series of problems, including classification and

prediction [11-14]. The input features include 'Full Time Employees', 'Enterprise Value', 'Profit Margins', 'Market Cap', 'Ask Size', 'Sector', 'State'. Different data preprocessing techniques are used before training the machine learning model. For the categorical features, the one-hot encoding is used. For the numerical features, the standardization normalization is used.

4.2. Results and Discussions

This study tried and compared different machine learning models, including random forest [15], AdaBoost [16], and XGBoost [17], with basic information of the company as input and performance under different definition methods as output, and their 10-fold classification accuracies are shown in Table 3.

Table 3. The accuracy of different machine learning models.

Model	2-classes classification	5-classes classification
Random Forest	0.7205	0.5056
AdaBoost	0.7160	0.4249
XGBoost	0.7342	0.4595

From Table 3, it is found that:

(1) the finer the performance classification, the lower the accuracy.

(2) The XGBoost model works best in the 2-classes case;

(3) The random forest model works best in the case of 5 classes.

5 Conclusion

This study analyzed the impact of the COVID-19 on various industries in the U.S. based on epidemic data and stock price data. It is found that the impact of the COVID-19 on different industries or companies is different. The COVID-19 has had a significant negative impact on industries such as transportation, energy, and travel services, while it has also boosted online sales, biopharmaceuticals, and other companies to a certain extent. Since the impact of the COVID-19 will last for a period of time, based on our analysis, the researcher believes that different remedial measures should be applied to different industries or companies, the government and companies should take the COVID-19 seriously.

References

1. Goddard, Ellen. "The impact of COVID-19 on food retail and food service in Canada: Preliminary assessment." *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie* (2020).
2. Leone, Lucia A., et al. "Healthy food retail during the COVID-19 pandemic: Challenges and future directions." *International journal of environmental research and public health* 17.20 (2020): 7397.
3. Agdas, Duzgun, and Prabir Barooah. "Impact of the COVID-19 pandemic on the US electricity demand and supply: an early view from data." *IEEE Access* 8 (2020): 151523-151534.
4. Nikiforiadis, Andreas, Georgia Ayfantopoulou, and Afroditi Stamelou. "Assessing the Impact of COVID-19 on Bike-Sharing Usage: The Case of Thessaloniki, Greece." *Sustainability* 12.19 (2020): 8215.
5. Suzumura, Toyotaro, et al. "The Impact of COVID-19 on Flight Networks." *arXiv preprint arXiv:2006.02950* (2020).
6. Nian, Guangyue, et al. "Impact of COVID-19 on Urban Mobility during Post-Epidemic Period in Megacities: From the Perspectives of Taxi Travel and Social Vitality." *Sustainability* 12.19 (2020): 7954.
7. Petrov, Artur, and Daria Petrova. "Sustainability of transport system of large Russian city in the period of COVID-19: Methods and results of assessment." *Sustainability* 12.18 (2020): 7644.
8. Zhang, Kainan, et al. "The Impacts of the COVID-19 Lockdown on Air Quality in the Guanzhong Basin, China." *Remote Sensing* 12.18 (2020): 3042.
9. Ou, Shiqi, et al. "Machine learning model to project the impact of COVID-19 on US motor gasoline demand." *Nature Energy* 5.9 (2020): 666-673.
10. Debone, Daniela, Mariana V. da Costa, and Simone GEK Miraglia. "90 days of COVID-19 social distancing and its impacts on air quality and health in Sao Paulo, Brazil." *Sustainability* 12.18 (2020): 7440.
11. Jiang, Weiwei. "Time series classification: nearest neighbor versus deep learning models." *SN Applied Sciences* 2.4 (2020): 1-17.
12. Jiang, Weiwei. "Applications of deep learning in stock market prediction: recent progress." *arXiv preprint arXiv:2003.01859* (2020).
13. Jiang, Weiwei, and Lin Zhang. "Geospatial data to images: A deep-learning framework for traffic forecasting." *Tsinghua Science and Technology* 24.1 (2018): 52-64.
14. Jiang, Weiwei, and Jiayun Luo. "Graph Neural Network for Traffic Forecasting: A Survey." *arXiv preprint arXiv:2101.11174* (2021).
15. Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.
16. Freund, Yoav, and Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting." *Journal of computer and system sciences* 55.1 (1997): 119-139.
17. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.