Data Mining Application: COVID-19's Impact on Stock Prices

Carson Kai-Sang Leung
Department of Computer Science
University of Manitoba
Winnipeg, MB, Canada
kleung@cs.umanitoba.ca

Jiachi Sun
Department of Computer Science
University of Manitoba
Winnipeg, MB, Canada
sunj3459@myumanitoba.ca

Jingyi Huang
Department of Computer Science
University of Manitoba
Winnipeg, MB, Canada
huangj20@myumanitoba.ca

Sammul To
Department of Computer Science
University of Manitoba
Winnipeg, MB, Canada
duy5@myumanitoba.ca

Yi Peng
Department of Computer Science
University of Manitoba
Winnipeg, MB, Canada
pengy346@myumanitoba.ca

Yipeng Liu
Department of Computer Science
University of Manitoba
Winnipeg, MB, Canada
liuy3468@myumanitoba.ca

Abstract— **The demand for data mining has increased in recent years as businesses and large organizations begin to appreciate the importance of data. Data companies have large databases that are updated regularly to keep up with the latest events. However, sorting through large databases can be overwhelming, and this is where data mining is effectively used to make things easier.**

Keywords: data mining; airline stock; prediction; stock price; COVID-19; biomedical data mining; linear regression; SVM

I. Introduction

Data mining can be defined as the process where various software techniques are used to analyze and extract useful information from large volumes of data[1]. During data mining, certain patterns and regularities are targeted by the computer, and any underlying rules are identified. It is also known as knowledge discovery in data (KDD). Data mining has gained popularity in recent years as companies and businesses all over the world are realizing the importance of data. It has been used to provide information and expose trends that can be used to make important business decisions. Artificial intelligence and machine learning make major contributions to data mining since they can be used to analyze big batches of data[2]. Other techniques such as data visualization have been used in conjunction with data mining to present information in ways that are easy to interpret and understand. Data mining uses software and other techniques to conduct an analysis and provide needed information[3]. This research is a discussion of data mining and how it has been used to provide updates on the current COVID-19 pandemic. Data mining tools include Oracle Data mining, Rapid Miner, IBM SPSS Modeler, Orange, and Python[4], among others. In this research paper, Python will be the preferred data mining tool. It will be used in conjunction with other techniques to provide data on other related outbreaks such as SARS.

This research paper is a discussion of the role played by data mining to showcase the stock trends from 3 major industries, such as retail, aviation, and telecommunications, during the pandemic in North America. We used machine learning to create our prediction model. Many businesses were directly affected by the pandemic especially during the lockdown since that meant all businesses had to close down and there was no customer traffic especially for those businesses that fully depended on walk-ins. On the other end, online businesses were thriving, and they became more popular as customers turned to online shopping since it was more convenient in these times. Therefore, data obtained showed an increase in stock value for online companies such as Zoom as people shifted to the online platforms to attend lectures and attend meetings. Data mining can be used to expose trends in the stock market and their relationship with other factors. For example, data mining can be used to show the relationship between the increasing number of COVID-19 infections and the stock value of some companies such as hand sanitizers and companies which make hygiene products. What's more, after the outbreak of the COVID-19, many people who have not returned to their own country are more concerned about the price of air tickets. This report proceeds as follows: The next section provides some background information about biomedical data mining, COVID-19, and some industries' stock prices. Section 3 is referred to as related works. After that, main body is presented in Section 4, data analysis, and conclusion in Section 5 and Section 6. Future works are in the end. Our *key contributions are* we analyzed how the COVID-19 impacts the stock prices by finding the relationship between each day's new COVID positive cases and the next day's stock prices. We selected 16 key stocks that are in the most impacted industries, such as retail industry and aviation industry and created indexes for those industries so we can analysis how the pandemic impact the industries. We also created some models to predict the stock prices based on the number of new daily confirmed cases.

II. Background

Given that the main focus of this paper is on biomedical data mining, the ongoing COVID-19 prioritized with the intention of making predictions of future trends in infections.

[1] 1 Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. Introduction to data mining. Pearson Education India, 2016.

[2] 2 Harrington, Peter. Machine learning in action. Manning Publications Co., 2012.

[3] 3 Mikut, Ralf, and Markus Reischl. "Data mining tools." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1, no. 5 (2011): 431-443.

[4] 4 Demšar, Janez, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina et al. "Orange: data mining toolbox in Python." the Journal of machine Learning research 14, no. 1 (2013): 2349-2353.

Biomedical data mining will be used to retrieve data on other related pandemics in the past such as the Spanish flu, the bubonic plague, and SARS which claimed the lives of millions. The COVID-19 disease is a painful reminder of the level of unpreparedness against such attacks and pandemics and how they could easily threaten the survival of humanity. The importance of data mining, in this case, is that through the analysis of data from similar pandemics in the past, scientists and researchers will be in a better position to know which measures were successfully used to protect humanity in the past, and which measures were not as useful. For example, during the SARS attack that hit Texas in 2003, measures such as quarantine, self-isolation, and social distancing were used in addition to wearing protective wear such as face masks to minimize the rate of transmission[5]. Now, what is COVID-19? it is the disease caused by a new coronavirus[6]. WHO first learned of this new virus on 31 December 2019. It is not only caused human infections and deaths; it is also affected the stock prices badly. The outbreak and spread of novel COVID-19 disease have seriously affected our economies, life, and more than that. Companies need to understand, accept, and deal with the economic impact of COVID-19, although it has emerged as a bane for the financial markets with unexpected levels. The impact of stock prices in North America is turning negative but least volatile[7]. The traditional economic and financial theory holds that stock prices are mainly affected by market characteristic-based factors. At this particular time, according to the theory of behavioral finance, emergencies will have an impact on investors' psychological and behavioral factors, which means it will have an impact on stock price[7]. The airline industry is one of the first industries that was affected by this event. Some similar events could be found in the catastrophic shocks from airline disasters and the impact of SARS on airline stock[14,15]. According to a report that showed the impact on the global airline business, they found airlines stocks in North America, Australia, and the U.K. are the worst[16]. However, there is no accurate relationship between airline stock and new positive cases. More than that, the impact of COVID-19 on retail stock prices and telecommunications stock prices have not yet been investigated. To fill those gaps, we build some models to estimate and predict those stock prices.

Data mining can be used to compare the number of infections before and after the measures was reinforced. Therefore, if a similar attack happens again in the future, the world will be better prepared since they will already know which measures were more effective than the others. Some of the basic patterns to watch out for include the patterns in the number of new COVID-19 infections. A decrease in the number of new infections with the use of PPEs and observation of the set measures will indicate the effectiveness of the measures. Additionally, an increase in the stock value with a decrease in the cases of new infections could translate to more secure and safer stock to invest in. If there is no significant effect on the stock value with changes in the number of cases of new infections, this could indicate a stable company that people should consider investing in. In biomedical or clinical researches, the researcher often tries to understand or relate two or more variables to predict an outcome or other variable[8]. Therefore, we analyzed how the COVID-19 impacts the stock prices currently by using linear regression to determine the relationship between each day's new COVID-19 positive cases and the next day's stock price movement.

III. Related works

We selected 12 different companies and separate them into 2 parts: America and Canada. More specifically, some key stocks that are in the most impacted industries, such as Amazon (online retailer) and Air Canada (Airline). To predict stock prices movement, we use the data based on the number of new COVID-19 positive cases. One of the models we used is linear regression. Linear regression is a linear approach to modeling the relationship between a scalar response and one or more explanatory variables in statistics[9]. At this time, we use the linear regression with one single explanatory variable, which is called *simple regression[10]* and coefficient of determination[11] called R^2 to evaluate the models. It shows how strong the linear relationship is between our two variables and how well the model fits the data. In general, a high value indicates that the model is a good fit for the data, although the interpretation of fit is different based on the context of analysis. For example, an of 0.3 might be a very high portion of variation to predict in a field such as social sciences[12]. For us, if R^2 equals to 0, then there is no relationship between stock prices and new positive cases; The value is at least greater than 0.5 or above to indicate a strong relationship between them. However, when there is more than one variable is available to explain the outcome, often, as variables added, the R^2 will increase

[5] 5 Heymann, David L. "The international response to the outbreak of SARS in 2003." Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 359, no. 1447 (2004): 1127-1129.

[6] *Coronavirus disease (COVID-19)*. (2020). World Health Organization. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19

[7] He, P., Sun, Y., Zhang, Y., & Li, T. (2020). COVID–19's Impact on Stock Prices Across Different Sectors—An Event Study Based on the Chinese Stock Market. *Emerging Markets Finance and Trade*, *56*(10), 2198–2212. https://doi.org/10.1080/1540496x.2020.1785865

[8] Kumari, K., & Yadav, S. (2018). Linear regression analysis study. *Journal of the Practice of Cardiovascular Sciences*, *4*(1), 33. https://doi.org/10.4103/jpcs.jpcs_8_18

[9] Linear regression. (2020, December 17). Retrieved December 18, 2020, from https://en.wikipedia.org/wiki/Linear_regression

[10] Lane, David M. "When there is only one predictor variable, the prediction method is called simple regression" *Introduction to Statistics: An Interactive e-Book.* Chapter 14: Regression. p.462.

[11] Bloomenthal, A. (2020, August 29). How the Coefficient of Determination Works. Retrieved December 16, 2020, from https://www.investopedia.com/terms/c/coefficient-of-determination.asp

[12] Enders, F. B. (2020, May 26). *coefficient of determination | Interpretation & Equation*. Encyclopedia Britannica. https://www.britannica.com/science/coefficient-of-determination

[13] Saunders, L. J., Russell, R. A., & Crabb, D. P. (2012). The Coefficient of Determination: What Determines a UsefulR2Statistic? *Investigative Opthalmology & Visual Science*, *53*(11), 6830. p.6830.

[14] Gillen, D., & Lall, A. (2003). International transmission of shocks in the airline industry. *Journal of Air Transport Management*, *9*(1), 37–49. https://doi.org/10.1016/s0969-6997(02)00068-6

[15] Loh E. The impact of SARS on the performance and risk profile of airline stocks. Int. J. Transp. Econ. 2006;33(3):401- 422. https://www.jstor.org/stable/42747811?seq=1 -metadata_info_tab_contents

[16] Maneenop, S., & Kotcharin, S. (2020). The impacts of COVID-19 on the global airline industry: An event study approach. *Journal of Air Transport Management*, *89*, 101920. https://doi.org/10.1016/j.jairtraman.2020.101920

even if the variable is not important[13]. Specifically, linear regression is not accurate in some cases. The other model we used is the support vector machine(SVM). Support vector machine produces significant accuracy with less computation power[17].

IV. MAIN BODY

Recall from the Introduction, we used some representative stock data of retail, airline and general industries in the US and Canada as experimental data. When using the linear regression algorithm to build the prediction model, the correlation between the stock data and the number of confirmed new coronavirus cases was very low, so we adopted the prediction model built by the support vector machine algorithm. After testing, the prediction model built by the support vector machine algorithm performed better in terms of correlation.

In addition, we also changed the initial prediction value from the percentage increase of the stock price over the previous day to the current day's stock price, which made the prediction results clearer and showed a stronger correlation.

In Section 3.1 and Section 3.2, we will introduce the construction methods of the two forecasting models separately.

3.1 Linear Regression Algorithm

First of all, when building the predictive model, we used the LinearRegression() method in the sklearn library to perform the linear regression model for modeling. First, we sort the collected data according to time and divide them into retail, aviation and general industries in the United States and Canada. Due to the suspension of the stock market, data on different dates will be missing and the data will be incomplete. In this regard, let the y value of all stock indexes of N/A equal to the average of the stock prices of all non-empty dates, deleting the row where the number of confirmed new coronavirus cases is 0.

Finally, we used the first 80% of the data for model training and model building, and the remaining 20% of the data for testing and prediction.

When using 80% of the data for model training and model building, we calculated according to the following formula:

$$\hat{y} = wx + b \ (1)$$

Each term in the formula represents:

- Y is the data to be predicted, and in our project is the stock price
- X is the data provided, that is, the number of confirmed new coronavirus cases per day
- w is the slope, the value that needs to be determined to pass the first 80% of the data when building the model
- b is a constant value that needs to be determined to pass the first 80% of the data when building the model

In order to calculate the optimal solution for w and b obtained by the first 80% of the data, we need to calculate by the following formula:

$$L(w, b) = \frac{1}{n} \sum_{i=1}^{n} (wx_i + b - y_i)^2 \quad （2）$$

Each term in the formula represents:

- $L(w, b)$is the sum of squared errors of the required solution. Our goal is to find the smallest possible value
- $\sum_{i=1}^{n} (wx_i + b - y_i)^2$ is the sum of the squared differences between each predicted value $\hat{y}_i$ and the actual value
- $\frac{1}{n}$is the mean square error obtained by dividing the square difference between the predicted value obtained at the second point and the true value by the total number of terms.

According to the least square method, we can get the formula (3) (4)

$$w = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \quad （3）$$

$$b = \frac{1}{n}\sum_{i=1}^{n} (y_i - wx_i) \quad （4）$$

Through (3) (4) we can get the minimum value of w and b that is the best value of the two constants calculated in our linear regression. After the model is built, we can measure the accuracy of the predictive model through the built model and the remaining 20% of the data.

3.2 Support Vector Regression (SVR) Algorithm

SVR is an application of SVM that can solve non-linear regression problems efficiently. SVR shares the same principle as SVM that it projects the data into higher dimensional feature space by the kernel functions and then perform linear separation.

Cost function:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} (\xi_i + \xi_i^*)$$

Linear SVR:

$$y = \sum_{i=1}^{N} (a_i - a_i^*) \cdot \langle x_i, x \rangle + b$$

Polynomial kernel function:

$$K(X_i, X_j) = (X_i . X_j)^d$$

Mapped Non-linear SVR:

$$y = \sum_{i=1}^{N} (a_i - a_i^*) \cdot K(x_i, x) + b$$

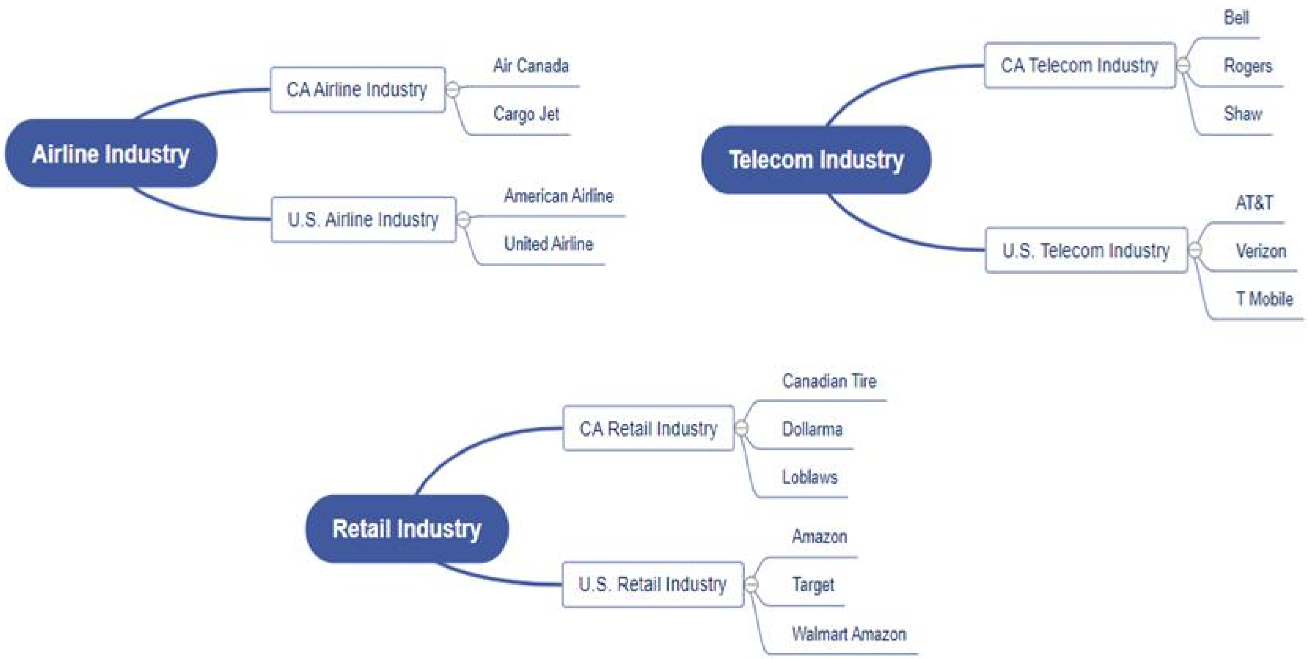

Figure 1

V. DATA ANALYSIS

Recall from Model Creation. We made stock selections in the U.S. (Nasdaq and NYSE) and Canadian (TSX). The specific stock selections are shown in Figure 1.

This section will choose the most representative stocks of each country in each category for a detailed analysis. Other stocks and corresponding data can be found in the output of the code.

When considering the correlation, besides comparing the actual value with the model's actual trend, looking at the Coefficient of determination of the model is another way. Since the Coefficient of determination $\leqslant 1$ (sometimes could be negative). When $0.5 \leqslant$ Coefficient of determination $\leqslant 1$, then a strong relationship is demonstrated.

4.1 Airline Industry

During the COVID pandemic, countries imposed varying degrees of air traffic control, and not only were many domestic flights canceled, but many international flights were affected to varying degrees. The border between the United States and Canada was even closed in March 2020 and will remain closed until at least January 21, 2021 due to the pandemic[18]. Also, required strict questioning before boarding. As a result, many flight companies have been forced to lay off staff. Therefore, we have chosen to study the stock prices of several airlines in the US and Canada.

4.1.1 Air Canada Stock Price Data Analysis

Each graph in Figure 2, and its corresponding coefficient of determination, in turn, is

- Figure 2.1: Relationship between new COVID-19 cases and Air Canada stock price. coefficient of determination: -0.002
- Figure 2.2: Relationship between new COVID-19 cases and Air Canada stock price changes. coefficient of determination: 0.022
- Figure 2.3: Relationship between new COVID-19 cases and Air Canada stock price by SVM. coefficient of determination: 0.130
- Figure 2.4: Relationship between new COVID-19 cases and Canadian Airline industry Index stock price by SVM. Coefficient of determination: 0.703

First, comparing Figure 2.1 and Figure 2.2, we can see that neither the change in stock price nor the current day stock price as the data to be predicted shows a strong relationship with the number of new confirmed COVID cases.

There are two possible reasons for this low correlation.

1. The model created by the linear regression algorithm we used is not inherently suitable for predicting the relationship between stocks and the number of new confirmed COVID cases.
2. The stock price of Air Canada does not have a strong correlation with the number of new COVID-19 positive cases.

The Coefficient of determination is 0.130, which is not higher than 0.5, compared to the Coefficient of determination of 0.02 using the linear regression algorithm. The use of the SVM algorithm has brought about a significant improvement in the correlation. So, we can now conclude that the model created by the linear regression algorithm we used is not suitable for predicting the relationship between Air Canada's stock and the number of new COVID-19 positive cases.

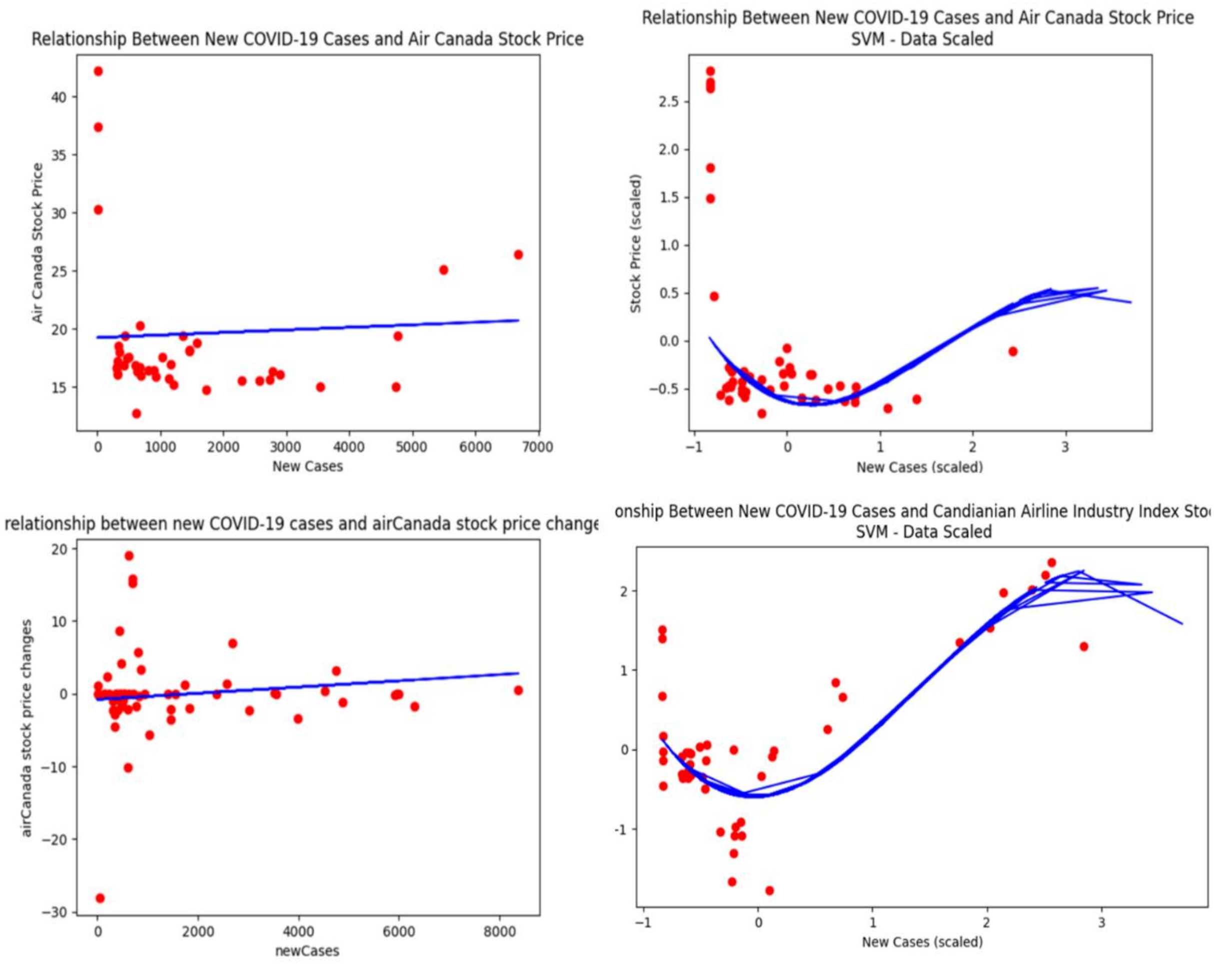

Figure 2

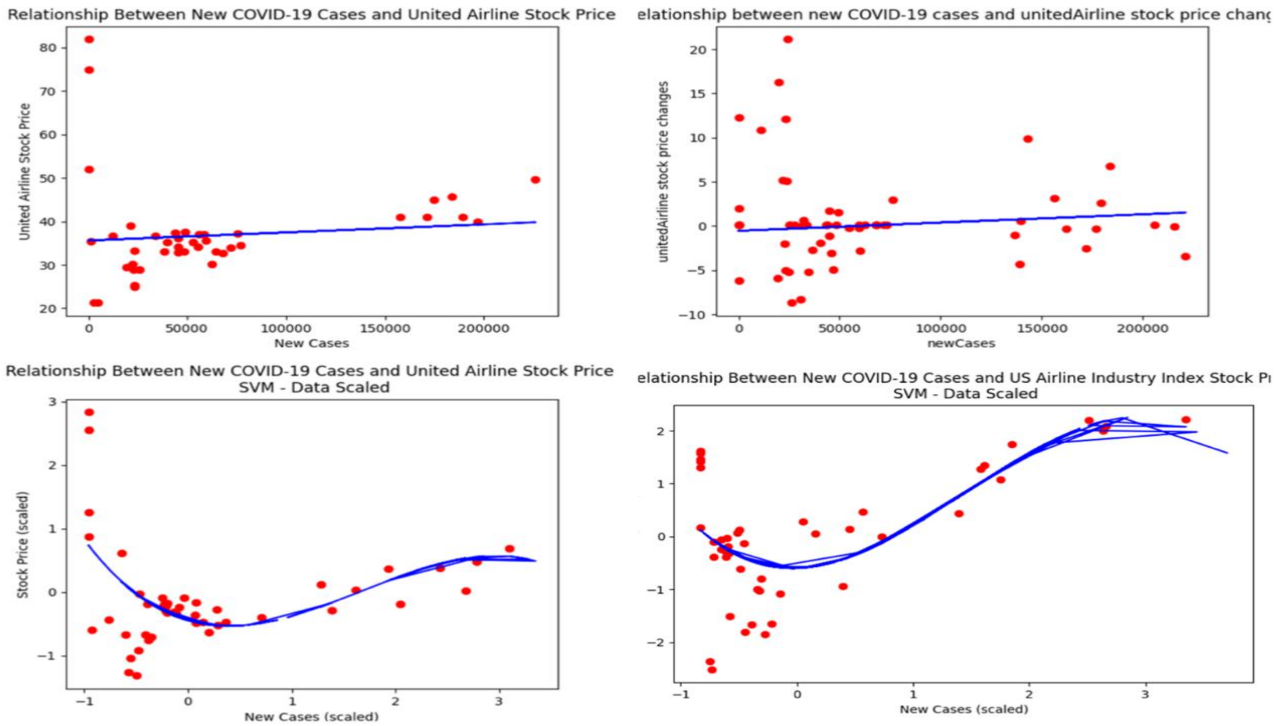

Figure 3.1-3.4, from light to right; up to down

By analyzing Figure 2.4 and its corresponding Coefficient of determination=0.703, we can find that the relationship between the number of new confirmed COVID cases and the stock price of the aviation industry is a strong correlation. The reason for this difference in correlation 1. between the total industry and the number of new confirmed COVID cases, and 2. between Air Canada and the number of new confirmed COVID cases, is that Air Canada, as the largest airline industry in Canada[19], shows stronger robustness to the change in stock price due to the COVID pandemic. However, for other airlines such as Cargo Jet, their stock prices will be more affected.

So, we can conclude that Air Canada has not been affected as much as expected in the face of the COVID pandemic and the restrictions on airline schedules. However, the Canadian airline industry as a whole has been hit hard.

4.1.2 United Airline Stock Price Data Analysis

In 4.1.1, we analyze Air Canada's stock and the Canadian airline industry's stock impact in the face of the COVID pandemic based on the differences between the resulting forecast models and the real data. In 4.1.2, the relationship between the stock price of United Airlines in the United States and the number of new confirmed COVID cases is discussed. First, please see Figure 3.

Each graph in Figure 3, and its corresponding coefficient of determination, in turn, is

- Figure 3.1: Relationship between new COVID-19 cases and United Airline stock price. coefficient of determination: -0.079
- Figure 3.2: Relationship between new COVID-19 cases and United Airline stock price changes. coefficient of determination: 0.009
- Figure 3.3: Relationship between new COVID-19 cases and United Airline stock price by SVM. coefficient of determination: 0.427
- Figure 3.4: Relationship between new COVID-19 cases and U.S. Airline industry Index stock price by SVM. Coefficient of determination: 0.570

Looking at Figure 3.1 and Figure 3.2, and the coefficient of determination values, as in 4.1.1, when using the Linear Regression algorithm for model building, the coefficient of determination still does not show a strong correlation, whether it is the stock price on the day or the stock price change as the predicted value. However, Figure 3.3, which uses the SVM algorithm, shows that the coefficient of determination is 0.427, which again shows that the predicted model created by the Linear Regression algorithm is not suitable for predicting the relationship between airline stocks and new cases.

By analyzing Figure 3.4 and its corresponding Coefficient of determination=0.570, we can find that the relationship between the number of new COVID-19 positive cases and the US airline industry's overall stock prices are correlated. However, it does not show a robust correlation. The possible reason for this phenomenon is that the East and West coasts of the United States, being the sea regions, are well connected and economically developed and have higher productivity levels. This may lead to the fact that despite the possible risk of infection, such long-distance airlines within the United States between the two coasts and between the east and west coasts and the interior are still the best choice for long-distance domestic travel. Combined with the fact that many airlines carry passengers and cargo, there may not be a large impact in terms of cargo transportation. The similarity between the predicted arcs of Figure 3.3 and Figure 3.4 is very high. The coefficient of determination differs by only 0.157, which shows that although American airlines are affected by the COVID pandemic, they are not very significant. This relatively low correlation for US Airlines maybe since larger companies were less affected by the COVID pandemic.

4.1.3 Section Conclusion

Comparing the data obtained for the airline industry in the United States and Canada, we can find that.

1. The Linear Regression algorithm is not suitable to find the relationship between the airline stocks and COVID before.
2. When using the SVM algorithm, the overall airline stocks in both countries show a strong correlation with the number of new confirmed COVID cases. And Canadian airline stocks received a more significant impact.
3. Unexpectedly, the impact on Air Canada does not show a high correlation. This may have a lot to do with Air Canada's being the largest airline in Canada and therefore has better robustness.

4.2 Telecom Industry

During the COVID Pandemic, countries imposed varying degrees of travel, party size, and mall size restrictions. Many universities in Canada and the U.S. have also changed their lecture format online. As offline activities are forced to decrease due to regulations, many things are being handled online, so we analyze the potential for growth in the telecommunications industry. In this section, we have chosen two stocks, Bell in Canada and T Mobil in the U.S., to analyze whether our idea is correct.

4.2.1 Bell Stock Price Data Analysis

Each graph in Figure 4, and its corresponding score, in turn, is：

- Figure 4.1: Relationship between new COVID-19 cases and Bell stock price. Coefficient of determination: -0.033
- Figure 4.2: Relationship between new COVID-19 cases and Bell stock price changes. Coefficient of determination: 0.001
- Figure 4.3: Relationship between new COVID-19 cases and Bell stock price by SVM. Coefficient of determination: 0.294
- Figure 4.4: Relationship between new COVID-19 cases and Canadian Telecom industry Index stock price by SVM. Coefficient of determination: 0.269

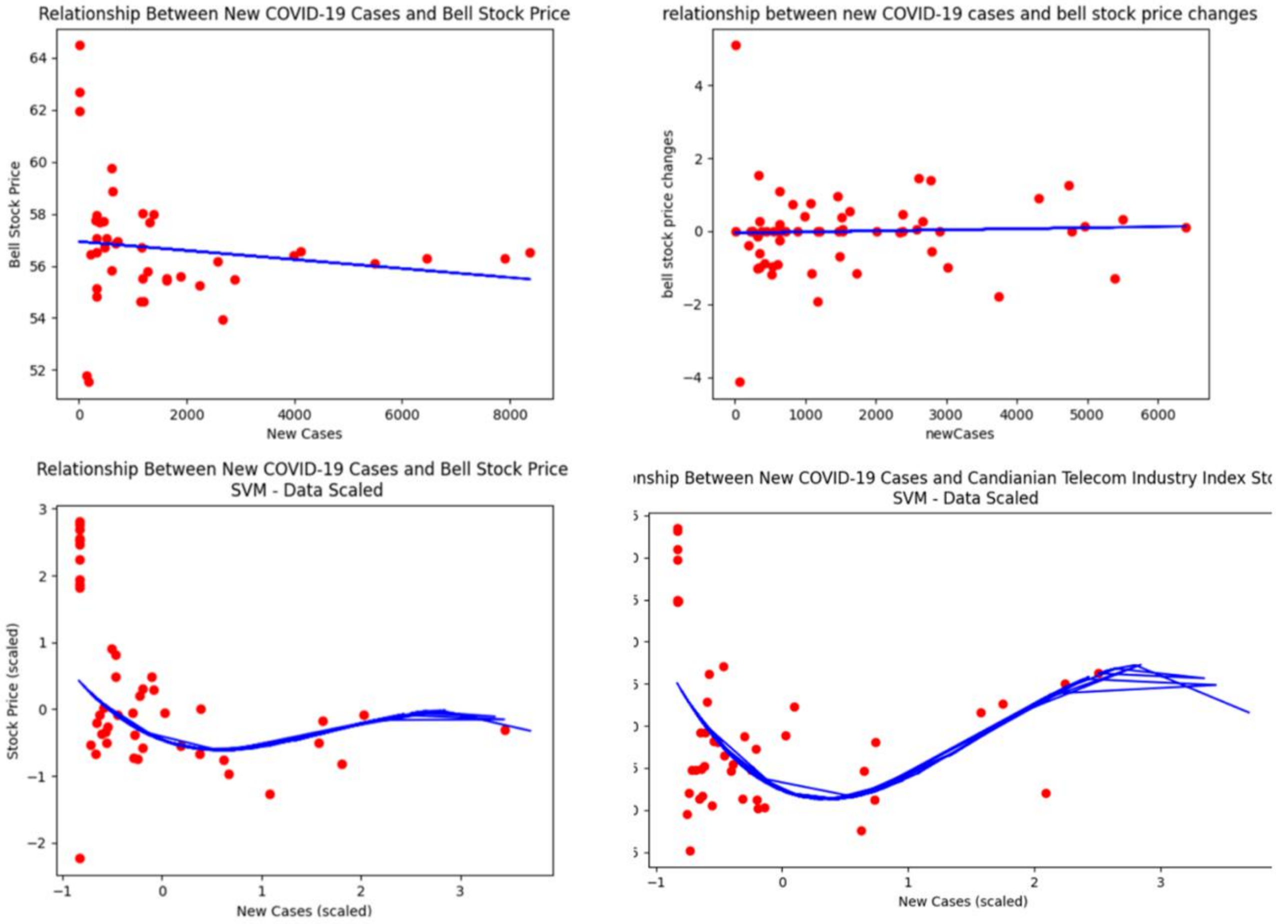

Figure 4.1-4.4

Figure IV.2.1-4.4 from light to right, up to down

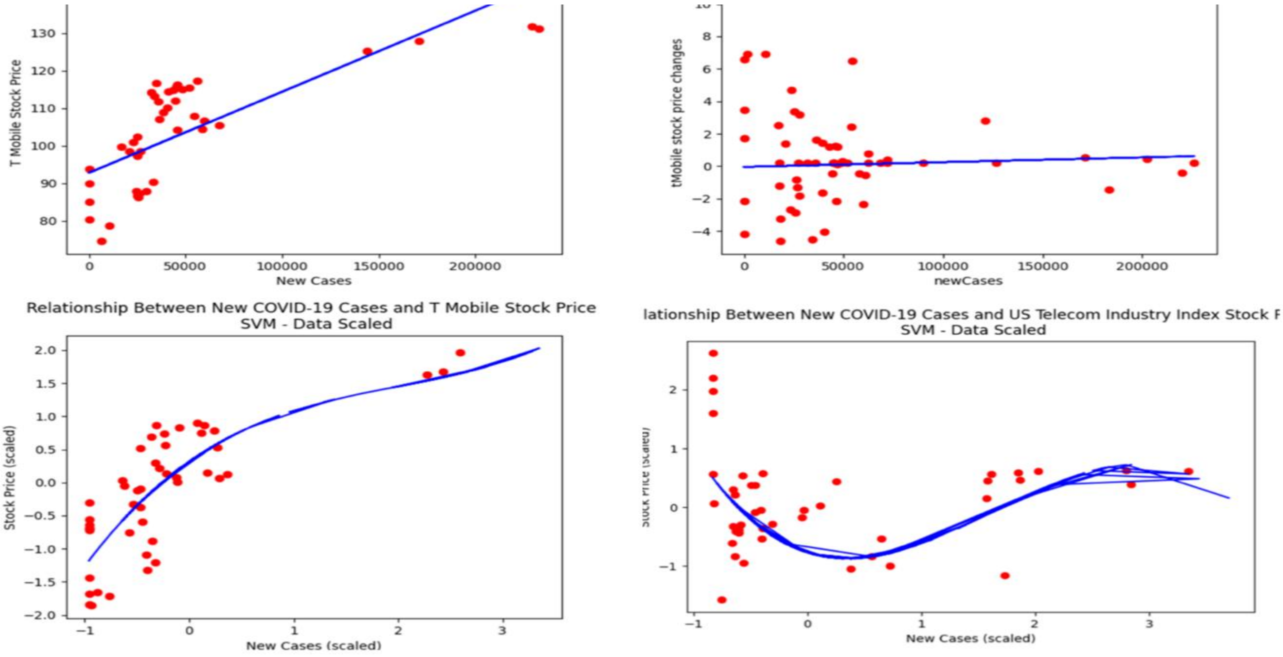

Figure 5.IV.1-5.4

First, comparing Figure 4.1 and Figure 4.2, we can see that Bell's stock price does not show a strong correlation with the number of new confirmed COVID cases, regardless of whether the change in stock price or the current day stock price is used as the data to be predicted. There are two possible reasons for this low correlation：

1. The model created by the linear regression algorithm we used is not inherently suitable for predicting the relationship between telecommunications class stocks and the number of new confirmed COVID cases.
2. The telecommunication class stock prices do not strongly correlate with the number of new confirmed COVID cases.

At this time, we can observe in conjunction with Figure 4.3 that although the coefficient of determination of 0.294 is not higher than 0.5. It has a significant change compared with the coefficient of determination value of 0.001 using the linear regression algorithm and the predicted arc. It is also closer to actual data points. Therefore, it can be concluded that using the SVM algorithm to build a predictive model is better.

By analyzing Figure 4.3 and Figure 4.4 and their corresponding $\text{Coefficient of determination}_{\text{F4.3}} = 0.294$ and $\text{Coefficient of determination}_{\text{F4.4}} = 0.269$, we can find that the relationship between the number of new confirmed COVID cases and the corresponding relationship of stock prices in the airline industry as a whole does not show a very obvious correlation. This is inconsistent with our initial guesses. The reason for the analysis may be because text messaging and phone calls are not used as the most dominant communication and office aid when working online. Another reason could be that due to the COVID pandemic, many companies were forced to lay off staff or even close down so that no opportunity was created to promote the telecommunication industry by conducting online offices and thus.

All in all, we can conclude that neither Bell nor the Canadian telecommunications industry as a whole has been hit much by the epidemic. There is no significant relationship between the stock price and the number of daily additions.

4.2.2 T Mobile Stock Price Data Analysis

Each graph in Figure 5, and its corresponding score, in turn, is：

- Figure 5.1: Relationship between new COVID-19 cases and T Mobile stock price. Coefficient of determination: 0.696
- Figure 5.2: Relationship between new COVID-19 cases and T Mobile stock price changes. Coefficient of determination: -0.039
- Figure 5.3: Relationship between new COVID-19 cases and T Mobile stock price by SVM. Coefficient of determination: 0.711
- Figure 5.4: Relationship between new COVID-19 cases and the US Telecom industry Index stock price by SVM. Coefficient of determination: 0.318

Looking at Figure 5.1 and Figure 5.2 and the corresponding Coefficient of determination values, we can see a significant difference between the results obtained using T Mobile stock price as the predicted value and using T Mobile stock price changes as the expected value.

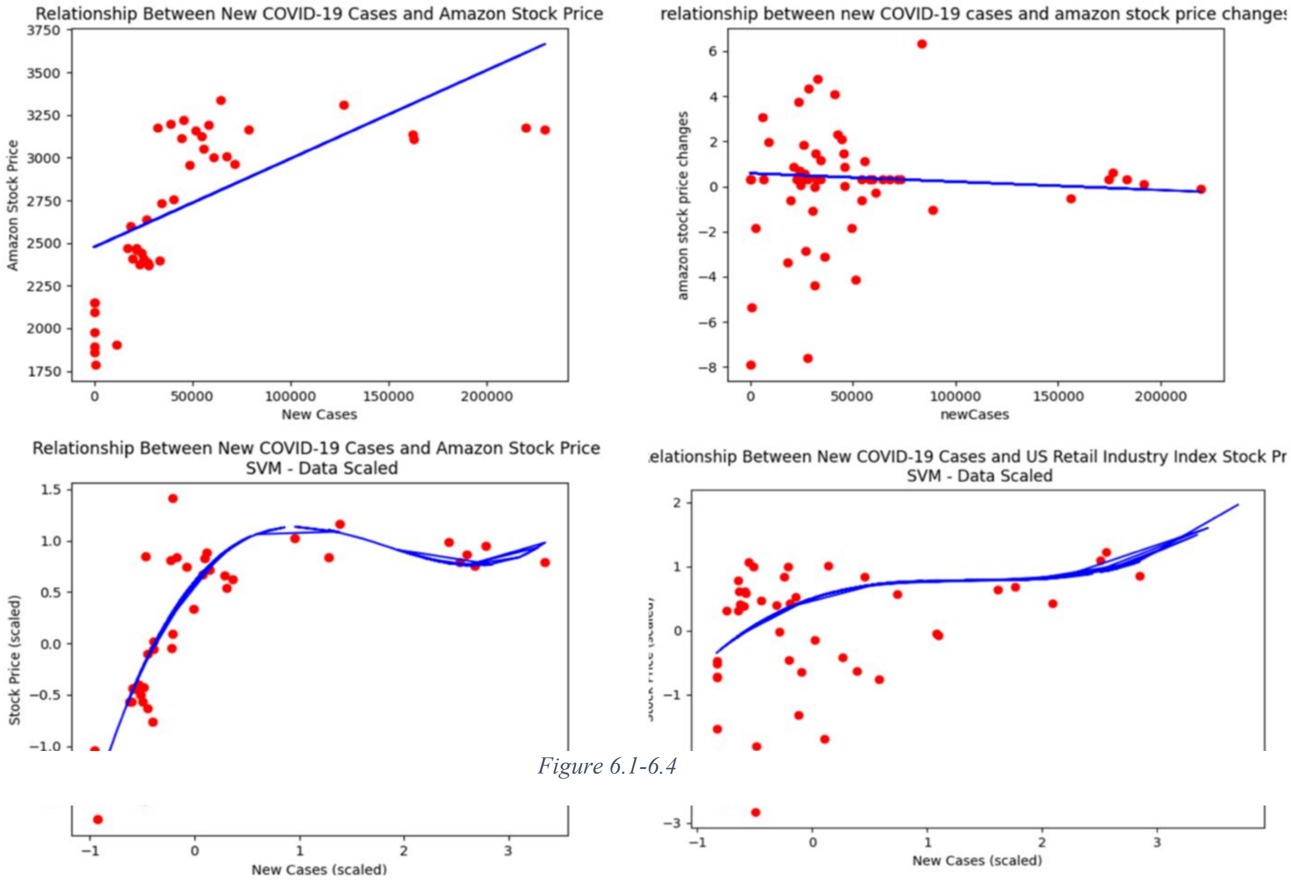

Figure 6.1-6.4

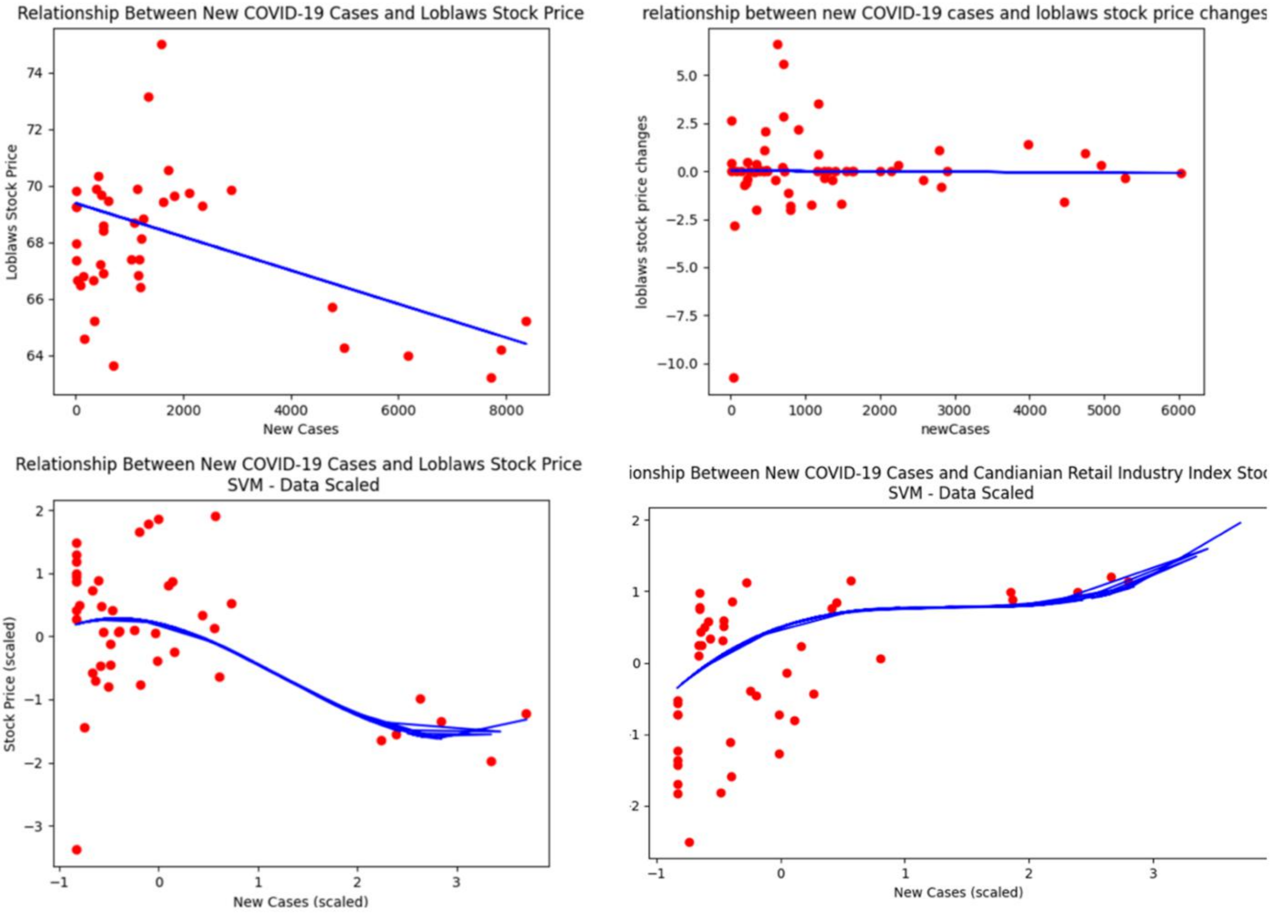

Figure7.1-7.4

If we look at Figure 5.3 using the SVM algorithm; we can see that Figure 5.1 and Figure 5.3 are very similar, with a high coefficient of determination values, which indicates that the COVID pandemic greatly influences T Mobile's stock price. However, combined with Figure 5.4, it can find that the US telecommunication industry as a whole was not influenced as much as T Mobile. This inconsistency between individual and overall trends is that T Mobile, as a multinational telecommunications company, has not only 100.3 million subscribers in the U.S.20, but also has a large subscriber base in many other countries around the world21. Therefore, in the face of the COVID outbreak, T Mobile was more affected than other telecommunications companies whose operations are more localized in the United States.

Using Figure 5, we can find that in general, the COVID pandemic does not show a strong relationship with the stock price of the telecom industry. Still, there is a strong correlation between T Mobile and the number of new confirmed COVID cases.

4.2.3 Section Conclusion

Summarizing and analyzing the data obtained from the forecasting models for the United States and Canada's communications industry. We can find that: the conclusions obtained from the building models using Linear Regression algorithm on stock price changes and the number of new confirmed COVID cases are not good and the telecommunications industry in both countries did not suffer much in the face of the COVID pandemic.

4.3 Retail Industry

This section uses SVM and Linear Regression algorithms to build predictive models to uncover the correlation between the COVID pandemic and the retail industry. We select two very representative data on the stock prices of Amazon in the U.S. and Loblaws in Canada for the analysis. The relationship between retail stock prices and the COVID pandemic is also discussed with the entire U.S. and Canadian markets.

4.3.1 Amazon Stock Price Data Analysis

Each graph in Figure 6, and its corresponding score, in turn, is：

- Figure 6.1: Relationship between new COVID-19 cases and Amazon stock price. Coefficient of determination: 0.286
- Figure 6.2: Relationship between new COVID-19 cases and Amazon stock price changes. Coefficient of determination: -0.039
- Figure 6.3: Relationship between new COVID-19 cases and Amazon stock price by SVM. Coefficient of determination: 0.687
- Figure 6.4: Relationship between new COVID-19 cases and the US Retail industry Index stock price by SVM. Coefficient of determination: 0.010

From the above results, we can learn that Amazon's stock price change is not related to COVID when using the Linear

Regression algorithm. There is a correlation between Amazon stock price and the number of new confirmed COVID cases when using the SVM algorithm.

Amazon's stock price change is shown in Figure 6.1. When using the Linear Regression algorithm to build the prediction model, we find that the correlation with the COVID pandemic is very low, with a value of 0.286, which shows that the COVID pandemic does not affect Amazon's business, perhaps because Amazon itself is an online business, so the change in Amazon's stock price is not affected much. However, when we use the SVM algorithm, as in Figure 6.3, we get a value of 0.688, at which point we can learn that the change in Amazon's stock price is related to the number of new confirmed COVID cases. It can also find that the SVM algorithm is more suitable for this case, probably because as the pandemic worsened, many people chose not to go to the mall but to shop online, and it is this situation that made Amazon's online business grow. At the same time, let Amazon's stock price to increase.

When we explore the Amazon stock price changes in Figure 6.2, we can find no correlation between stock price changes and the number of new confirmed COVID cases by Coefficient of determination = -0.039.

When we use the SVM algorithm to calculate the association between the retail industry and the COVID pandemic, as in Figure 6.4, we can get a value of 0.010. We can see that there is no significant correlation between the two. There is no significant correlation between the number of new confirmed COVID cases and the retail industry. This is probably because even if the COVID affects people, they still need to go to supermarkets, shopping malls, or online to buy essential goods and food. Therefore, the impact on the retail sector will be minimal.

We can learn that there is a correlation between Amazon stock price and the COVID pandemic by comparing the two algorithms and their application in different situations. But there is no significant association between the retail industry as a whole and the number of new confirmed COVID cases.

4.3.2 Loblaws Stock Price Data Analysis

Each graph in Figure 7, and its corresponding score, in turn, is：

- Figure 7.1: Relationship between new COVID-19 cases and Loblaws stock price. Coefficient of determination: 0.180
- Figure 7.2: Relationship between new COVID-19 cases and Loblaws stock price changes. Coefficient of determination: -0.029
- Figure 7.3: Relationship between new COVID-19 cases and Loblaws stock price by SVM. Coefficient of determination: 0.344
- Figure 7.4: Relationship between new COVID-19 cases and the Canada Retail industry Index stock price by SVM. Coefficient of determination: 0.353

Now let's analyze the correlation between the stock price of Loblaws in Canada and the COVID pandemic.

Using the Linear Regression algorithm to build the prediction model, the relationship between the stock price and the number of new confirmed COVID cases is calculated as in Figure 3.1. We can get the value 0.18. We can learn that the association between Loblaws' stock price and the COVID pandemic is not very big, probably because the pandemic in Canada is not as serious as that in the United States. Many people need to maintain their regular work life, so they still frequent Loblaws to buy food and other daily necessities. This can be seen from the predicted model constructed by the SVM algorithm, as in Figure 3.3, where we calculate a value of 0.344, which better proves the Linear Regression algorithm's results that the Loblaws share price is not strongly correlated with the pandemic.

Moreover, in Winnipeg, people can only buy non-lifestyle goods online because of the pandemic. For example, people want to buy computers on Bestbuy, but because of government regulations, people cannot go to Bestbuy malls to shop; they can only shop online. However, there is no blockade for the superstore under Loblaws, although the restrictions are strict for people. Therefore, the SVM and Linear Regression algorithms build prediction models that yield similar conclusions: Loblaws share prices are not significantly correlated with the pandemic changes.

When using the SVM algorithm to build a predictive model to explore the association between the overall retail industry and the COVID pandemic, as shown in Figure 3.4, we arrive at a Coefficient of determination = 0.353, from which we can find that the previous association between the two is also small. Because people's demand for a basic living has not changed, people can give up buying entertainment products. Still, People cannot stop buying food and other necessities, so people's consumption situation in this industry has not changed, so the change for profit is little affected by the COVID pandemic. Therefore, we can conclude that the correlation between the number of new confirmed COVID cases and the industry share price is not significant in the retail industry.

From the above analysis, we can see that the impact of the epidemic on the Canadian Loblaws stock or the Canadian retail stock price as a whole is not significant. Because people need to do basic life shopping, buy food, etc.

4.3.3 Section Conclusion

Summarizing and analyzing the data obtained from the retail industry prediction models in the United States and Canada. We can find that the conclusions on the changes in retail industry stock prices and the number of newly confirmed COVID cases obtained using the linear regression algorithm to establish models do not have much reference value. Popular, the retail industry in the two countries has not received much influence, whether it is a leading or ordinary enterprise. The reason for this situation may be that people use online or offline shopping to meet daily life needs.

VI. CONCLUSIONS

This project used the Linear Regression algorithm and the SVM algorithm to build a predictive model to find the relationship between the COVID-19 pandemic and stock prices. We have tried to dig out the relationship between stock price and the number of new COVID cases and the relationship between stock price changes and the number of new COVID cases. In the aviation industry, the telecommunications industry, and the retail industry, a total of 15 stock price data sets were collected for evaluation. The evaluation results show that we use the Linear Regression algorithm to find the relationship between stock price changes and the number of new additions, and we have not seen much. However, when using the SVM algorithm, we found that both the United States and Canada have affected the aviation industry. However, large airlines such as Canada's Air Canada and the United States' United Airline have not significantly impacted. The price of stocks in the telecommunications and offline retail industries has not been affected much. Still, companies like T Mobile that have user bases worldwide have been affected to a certain extent.

VII. Future Works

In order to improve the accuracy of our model, we need to refine our model. With the development of the pandemic, we are able to gather more data in the future to train our model. Future more, we can include more stocks into the industry index to have a more accurate industry index prediction model.

References

[1] Mikut, Ralf, and Markus Reischl. "Data mining tools." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1, no. 5 (2011): 431-443.

[2] 2Demšar, Janez, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina et al. Orange: data mining toolbox in Python." *the Journal of machine Learning research* 14, no. 1 (2013): 2349-2353.

[3] 3Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education India, 2016.

[4] 4Harrington, Peter. *Machine learning in action*. Manning Publications Co., 2012

[5] 5 Heymann, David L. "The international response to the outbreak of SARS in 2003." *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 359, no. 1447 (2004): 1127-1129.

[6] *Coronavirus disease (COVID-19)*. (2020). World Health Organization. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19

[7] He, P., Sun, Y., Zhang, Y., & Li, T. (2020). COVID–19's Impact on Stock Prices Across Different Sectors—An Event Study Based on the Chinese Stock Market. *Emerging Markets Finance and Trade*, *56*(10), 2198–2212. https://doi.org/10.1080/1540496x.2020.1785865

[8] Kumari, K., & Yadav, S. (2018). Linear regression analysis study. *Journal of the Practice of Cardiovascular Sciences*, *4*(1), 33. https://doi.org/10.4103/jpcs.jpcs_8_18

[9] Linear regression. (2020, December 17). Retrieved December 18, 2020, from https://en.wikipedia.org/wiki/Linear_regression

[10] Lane, David M. "When there is only one predictor variable, the prediction method is called simple regression" *Introduction to Statistics: An Interactive e-Book.* Chapter 14: Regression. p.462.

[11] Bloomenthal, A. (2020, August 29). How the Coefficient of Determination Works. Retrieved December 16, 2020, from https://www.investopedia.com/terms/c/coefficient-of-determination.asp

[12] Enders, F. B. (2020, May 26). *coefficient of determination | Interpretation & Equation*. Encyclopedia Britannica. https://www.britannica.com/science/coefficient-of-determination

[13] Saunders, L. J., Russell, R. A., & Crabb, D. P. (2012). The Coefficient of Determination: What Determines a UsefulR2Statistic? *Investigative Opthalmology & Visual Science*, *53*(11), 6830. p.6830.

[14] Gillen, D., & Lall, A. (2003). International transmission of shocks in the airline industry. *Journal of Air Transport Management*, *9*(1), 37–49. https://doi.org/10.1016/s0969-6997(02)00068-6

[15] Loh E. The impact of SARS on the performance and risk profile of airline stocks. Int. J. Transp. Econ. 2006;33(3):401- 422. https://www.jstor.org/stable/42747811?seq=1 -metadata_info_tab_contents

[16] Maneenop, S., & Kotcharin, S. (2020). The impacts of COVID-19 on the global airline industry: An event study approach. *Journal of Air Transport Management*, *89*, 101920. https://doi.org/10.1016/j.jairtraman.2020.101920

[17] Gandhi, R. (2018, July 5). *Support Vector Machine — Introduction to Machine Learning Algorithms*. Medium. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[18] E. (2020, December 13). *Canada-U.S. border closure extended to Jan. 21 as coronavirus cases soar, CBSA says*. Global News. https://globalnews.ca/news/7518359/coronavirus-canada-us-border-closure-extension/

[19] *Air Canada Corporate Profile*. (2020). Air Canada. https://www.aircanada.com/content/aircanada/ca/en/aco/home/about/corporate-profile.html

[20] Wikipedia contributors. (2020, December 22). *T-Mobile US*. Wikipedia. https://en.wikipedia.org/wiki/T-Mobile_US

[21] Wikipedia contributors. (2020a, December 13). *T-Mobile*. Wikipedia. https://en.wikipedia.org/wiki/T-Mobile

[22] Dr.Saed Sayad. Support Vector Machine – Regression (SVR). https://www.saedsayad.com/support_vector_machine_reg.htm