AN AUTOMATED DATA-DRIVEN PREDICTION OF PRODUCT PRICING BASED ON COVID-19 CASE NUMBER USING DATA MINING AND MACHINE LEARNING

Zhuoyang Han[1], Ang Li[2] and Yu Sun[3]

[1]University of California, Irvine, California, USA
[2]California State University, Long Beach, USA
[3]California State Polytechnic University, Pomona, USA

ABSTRACT

In early 2020, a global outbreak of Corona Disease Virus 2019 (Covid-19) emerged as an acute respiratory infectious Disease with high infectivity and incidence. China imposed a blockade on the worst affected city of Wuhan at the end of January 2020, and over time, covid19 spread rapidly around the world and was designated pandemic by the World Health Organization on March 11. As the epidemic spread, the number of confirmed cases and the number of deaths in countries around the world are changing day by day. Correspondingly, the price of face masks, as important epidemic prevention materials, is also changing with each passing day in international trade. In this project, we used machine learning to solve this problem. The project used python to find algorithms to fit daily confirmed cases in China, daily deaths, daily confirmed cases in the world, and daily deaths in the world, the recorded mask price was used to predict the effect of the number of cases on the mask price. Under such circumstances, the demand for face masks in the international trade market is enormous, and because the epidemic changes from day to day, the prices of face masks fluctuate from day to day and are very unstable. We would like to provide guidance to traders and the general public on the purchase of face masks by forecasting face mask prices.

KEYWORDS

Corona Virus, Machine Learning, Price Prediction, Linear Regression, Poly Regression, Data Cleaning

1. INTRODUCTION

Coronavirus [1] is a kind of RNA virus which exists widely in nature. It has the tropism of gastrointestinal tract, respiratory tract and nervous system. The coronavirus founded in December 2019 is named 2019-nCoV, which causes covid-19. The major media of COVID-19 are direct, aerosol, and contact. Direct transmission refers to the patient sneezes, speaking when the droplets were inhaled close to other people caused by infection; Aerosol transmission refers to the droplets in the air formed Aerosol, was inhaled after the infection; Contact infection is a kind of infection caused by droplets attached to the surface of articles and finally contacting the mucous membrane of eyes, mouth and nose through intermediary articles. Covid-19 is as transmissible as influenza, and because of its initial clinical manifestations of fever, dry cough and weakness, it may lead patients to mistakenly believe that they have the common cold, thus lowering their risk of infection, and delayed the best time for treatment. As a result, the virus continued to spread

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 275-283, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101420

around the world and grow rapidly in March, after an outbreak in China and the closure of the city in January.

Open problem: The price of the mask as an important epidemic prevention material should be related to one or more features. However, the relationship is unknown and different algorithms need to be tried with extensive experiments.

Solution: Machine learning [2] models were used to find the relationship between mask prices and features, which could then be applied to business situations such as buying at a low price or selling at a high price.

We concluded from our predictions that the price [3] of face masks would fall over the next four days. There is a strong correlation between mask prices and China daily cases for some time to come. By comparing the relationships between eight features and mask prices, we find out that the most closely linked function is the exponential function of China daily case. In the future, we believe using this method can effectively predict mask price and provide trade guidance for business and life demand.

The rest of the paper is organized as follows: Section 2 lists the key challenges to be solved in this problem scope. Section 3 details the solution, followed by presenting the experimental results in Section 4. Section 5 analyzes the related work and we conclude the paper in Section 6 with the future work summarized.

2. CHALLENGES

2.1. The most important global data on early infectious diseases are unreliable

Covid-19 is still controlled in China from the end of 2019 to February 2020, and the world has not started to provide large-scale testing of COVID-19 to patients, so early world data are lacking. Even if there were confirmed cases of influenza-like illness caused by coronavirus in various countries, or deaths due to symptoms caused by Covid-19, it would not be included in the statistics. As China is the world's first, worst and fastest outbreak of the disease, the confirmed figures for January and February are based on China.

2.2. China's Data May not be a Perfect Reflection of Global Models

Even if China's data were used as a sample of earlier data, given the varying levels of health system soundness in different countries around the world and the different measures proposed and implemented by health authorities for epidemic prevention, China's infectious diagnosis model does not necessarily fit all countries, especially underdeveloped countries with inadequate health systems and developed countries with slow response to epidemic prevention. Except for a few Asian countries such as China, Japan, South Korea and Singapore, most countries did not take emergency measures during this period, and thus may have contributed to the spread of the virus, so in terms of epidemiology, data for countries that have implemented measures are likely not to be exactly similar to data for countries that have not implemented measures effectively and comprehensively.

2.3. There is not Enough Data on Mask Prices

The data about the price of the mask was taken from amazon.com. I recorded the daily prices from January 21 to March 8 and collated them as data input into the algorithm. However, the data was interrupted on March 8 when Amazon announced the removal of the N95 mask, thus missing the mask prices in March and April. In this project, I used China's data from the early days of the epidemic, when the city of Wuhan was shut down from January to March and the country was on high alert, so the peak time coincided with most of the mask price data, so the lack of data from Amazon's announcement about the removal of the masks does not unduly affect the accuracy of the predictions.

3. SOLUTION

3.1. Overview of the Solution

The CDC first obtained daily figures for confirmed cases and deaths worldwide from January1st through April 11th. The features are then obtained through data cleaning, and a fitting algorithm is used to guess which model fits. The relationship between the number of cases and mask price was predicted by combining the mask price data as dependent variable and independent variable.

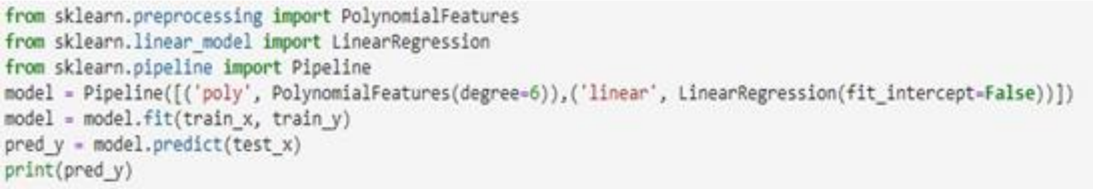
3.2. Machine Learning Model and Feature Selection

The following features have been identified in the data model construction:

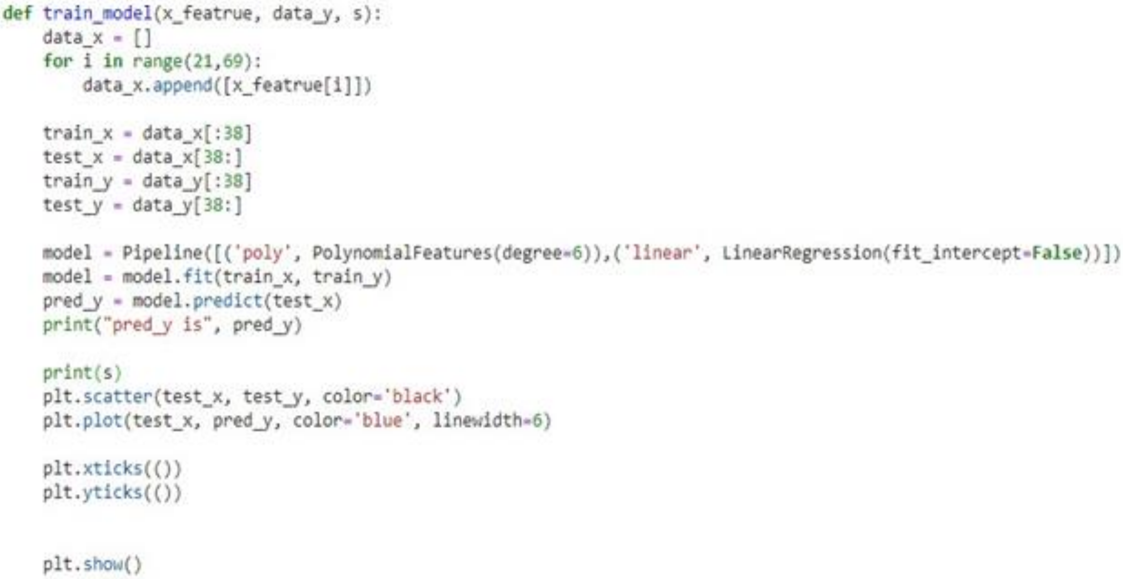
1) China daily case
2) China total case
3) China daily death case
4) China total death case
5) World daily case
6) World total case
7) World daily death case
8) World total death case
9) Mask price/5pcs

3.3. Training and Prediction

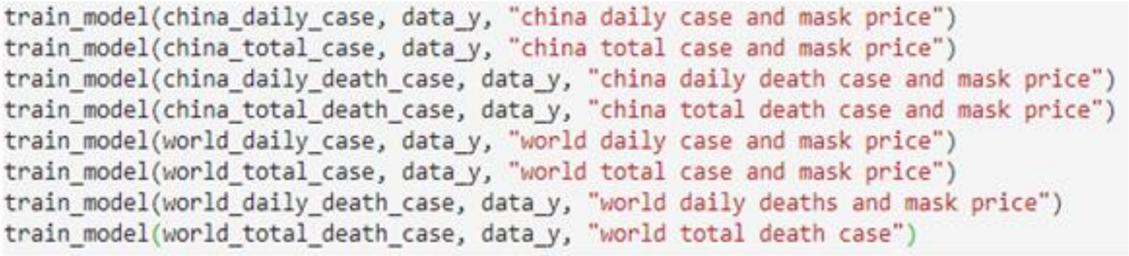
We used the machine learning library scikit-learn [4] to train and predict the model. The models we used in the project are linear regression [5] and polynomial regression. Linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. It is helpful to interpret data on a modular level, especially when we want to quantify cases and prices. Polynomial regression provides the best approximation between variables and is compatible for many functions [6].



```
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression
from sklearn.pipeline import Pipeline
model = Pipeline([('poly', PolynomialFeatures(degree=6)),('linear', LinearRegression(fit_intercept=False))])
model = model.fit(train_x, train_y)
pred_y = model.predict(test_x)
print(pred_y)
```



```
def train_model(x_featrue, data_y, s):
    data_x = []
    for i in range(21,69):
        data_x.append([x_featrue[i]])

    train_x = data_x[:38]
    test_x = data_x[38:]
    train_y = data_y[:38]
    test_y = data_y[38:]

    model = Pipeline([('poly', PolynomialFeatures(degree=6)),('linear', LinearRegression(fit_intercept=False))])
    model = model.fit(train_x, train_y)
    pred_y = model.predict(test_x)
    print("pred_y is", pred_y)

    print(s)
    plt.scatter(test_x, test_y, color='black')
    plt.plot(test_x, pred_y, color='blue', linewidth=6)

    plt.xticks(())
    plt.yticks(())


    plt.show()
```



```
train_model(china_daily_case, data_y, "china daily case and mask price")
train_model(china_total_case, data_y, "china total case and mask price")
train_model(china_daily_death_case, data_y, "china daily death case and mask price")
train_model(china_total_death_case, data_y, "china total death case and mask price")
train_model(world_daily_case, data_y, "world daily case and mask price")
train_model(world_total_case, data_y, "world total case and mask price")
train_model(world_daily_death_case, data_y, "world daily deaths and mask price")
train_model(world_total_death_case, data_y, "world total death case")
```

4. EXPERIMENT RESULTS

4.1. Comparison of Different Features

We first draw the plot of features based on dates, from 2020-01-01 to 2020-04-11.

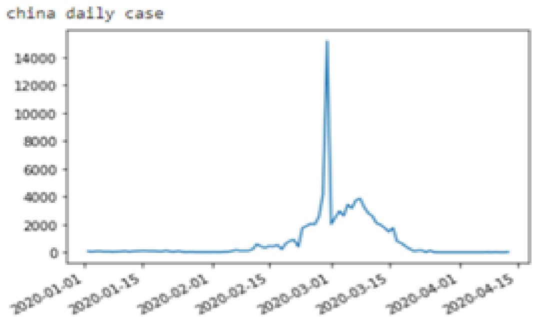


Figure 1. China daily case

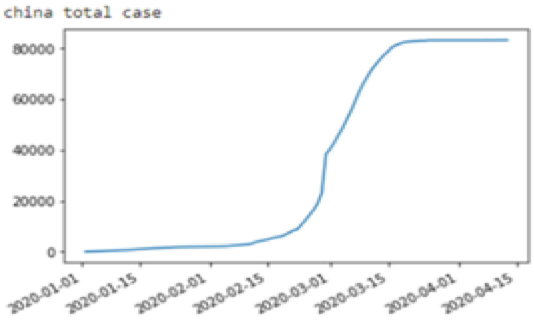


Figure 2. China total case

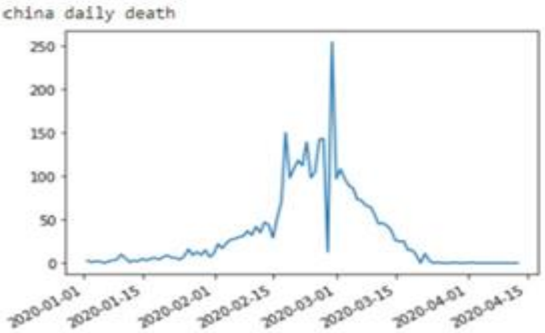


Figure 3. China daily death

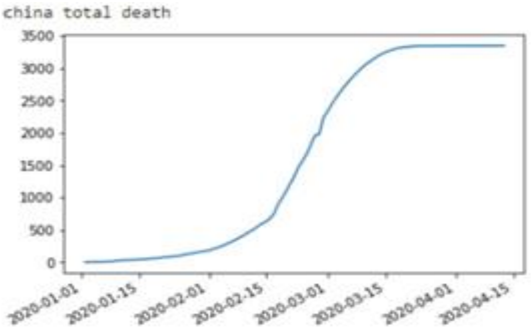


Figure 4. China total death

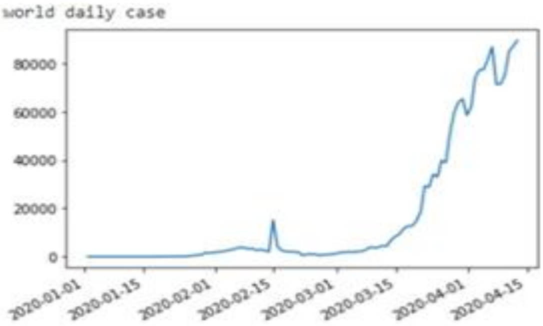


Figure 5. World daily case

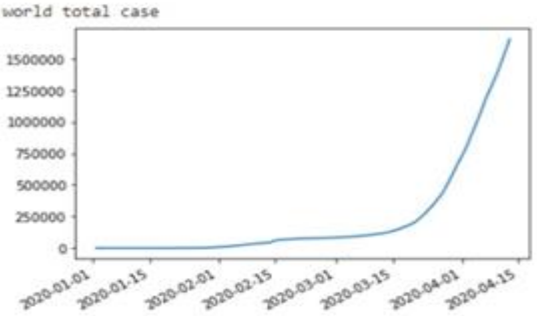


Figure 6. World total case

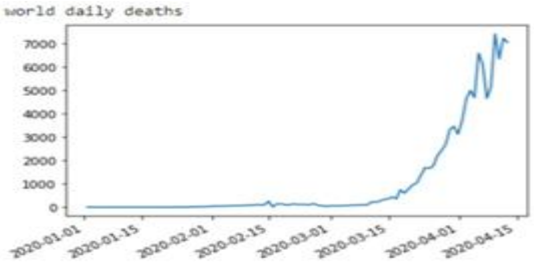


Figure 7. World daily deaths

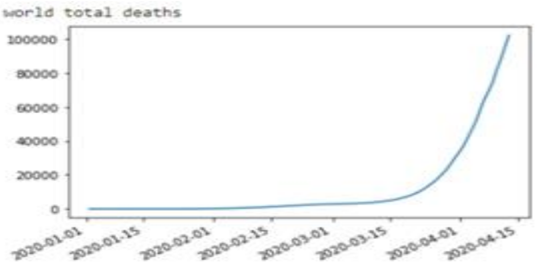


Figure 8. World total deaths

Figure 1, figure 3 show that China had an outbreak of daily confirmed cases and daily deaths in early March, when the figures peaked. China's total number of confirmed cases and deaths (Figure 2,4) and the world's total number of confirmed cases and deaths (Figure 6,8) both experienced a dramatic increase. The curve between the total number of confirmed cases and the number of deaths in China is s-shaped, which means that the epidemic has reached an inflection point, while the curve between the total number of confirmed cases and the number of deaths in the world is exponential[7]. This means that covid-19 is still highly contagious [8] in the world, and the number of patients is growing rapidly every day.

4.2. Comparison of Different Models of Training Features

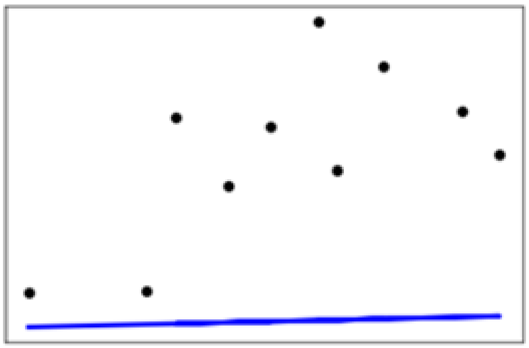


Figure 9.Linear Regression

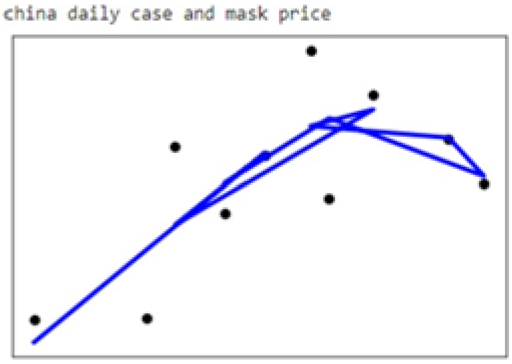


Figure 10. Polynomial regression

As can see from figures 9 and 10, linear regression is not ideal, but polynomial regression fits better.

4.3. Comparison of Prediction of all 8 Features

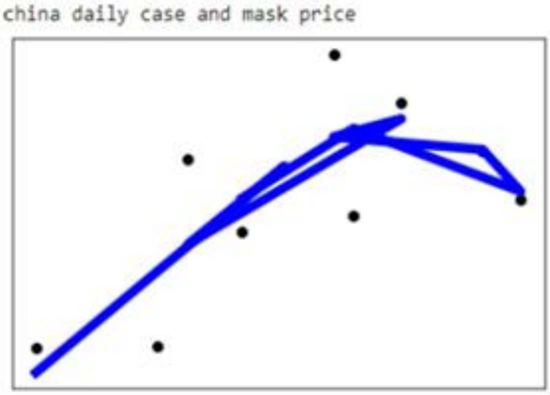


Figure 11. china daily case and mask price

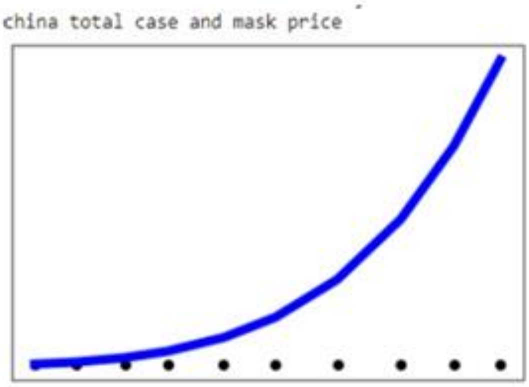


Figure 12. china total case and mask price

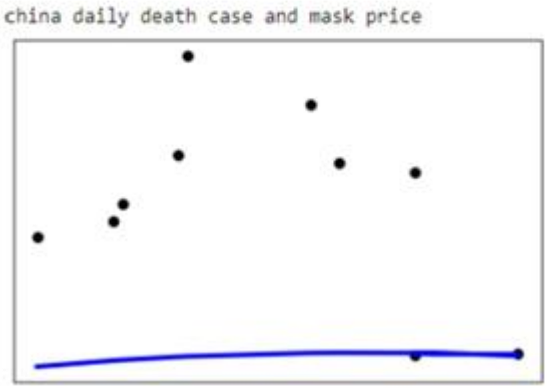


Figure 13. china daily deaths and mask price

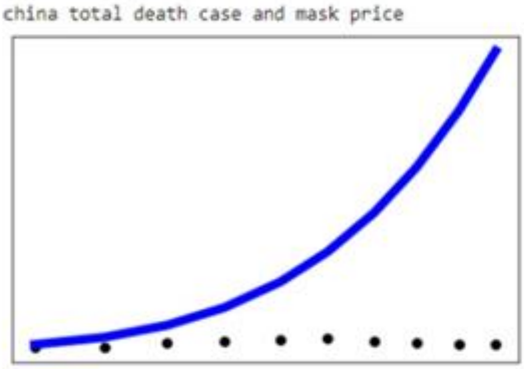


Figure 14. china total deaths and mask price

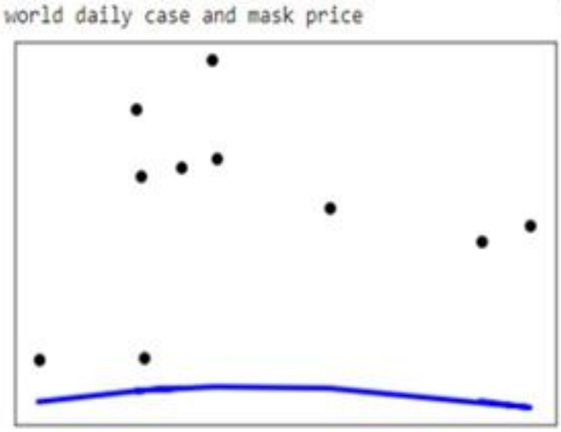


Figure 15. world daily case and mask price

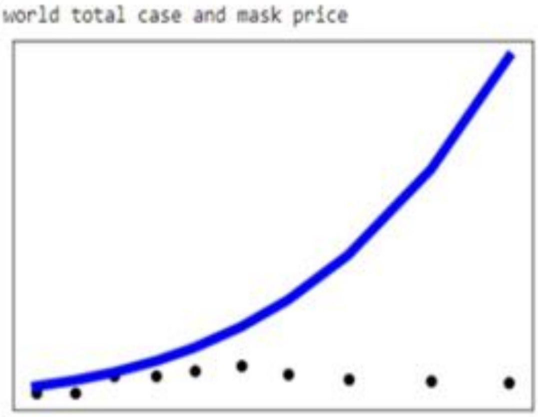


Figure 16. world total case and mask price

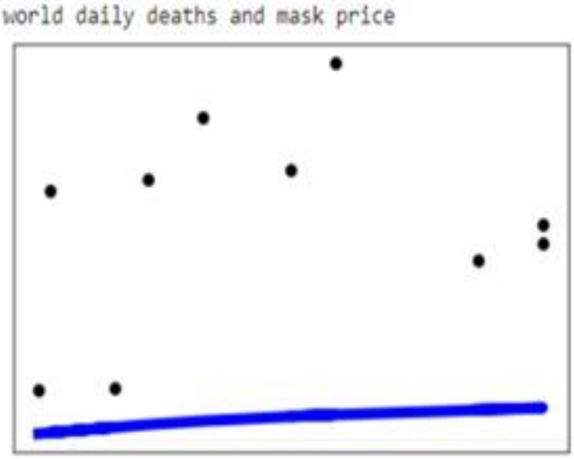


Figure 17. World daily death and mask price

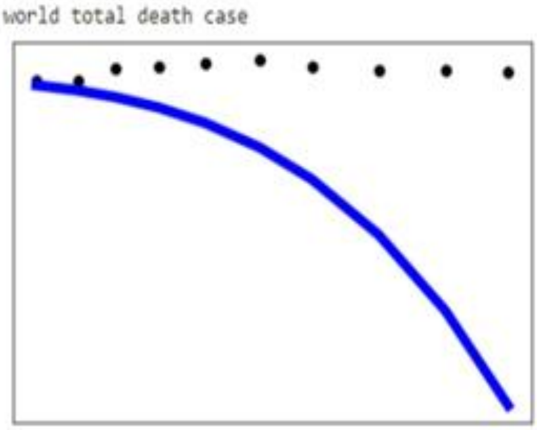


Figure 18. World total death and mask price

When Polynomial degree [9] was 6, Figure 1 was more accurate.

5. RELATED WORK

Liu, Y. et al [10] calculated reproduction number(R0) of the COVID-19 virus to find out that the ability of the virus to spread is higher than WHO expected, which can provide explanation to why the cases increased fast world-widely. Wu, Z. and McGoogan, M. J. [11] analyzed the emergent measures applied in Wuhan, China from January to March. They showed the whole timeline to explain the outbreak in China and the international impact. By comparison, we used machine learning to analyze and present the relationship between Chinese cases and World cases. Grasselli, G. et al [12] used linear and exponential models to estimate Italy's ICU demand [13]. In this project, we used linear and polynomial regression models to predict the world's mask prices. Fanelli, D. et al [14] analyzed the outbreak in China, Italy and France using a simple susceptible-infected-recovered-deaths model to indicate the relationships of situations in three countries. We focused on the infected and deaths number of China as it was the first country to break out, the first to block, and the first to reach the inflection point. Harrell FR Jr. et al [15] introduced the advantages of regression models in making accurate predictions than other methods.

6. CONCLUSIONS

We concluded from our predictions that the price of face masks [16] would fall over the next four days. There is a strong correlation between mask prices and China daily cases for some time to come. By training eight different features, we have eight different data, and know that the price of the most closely linked function is the exponential function.

The Algorithm can effectively predict the trend of mask price fluctuation, so we believe in the future using this method can effectively predict mask price and provide trade guidance for business and life demand.

REFERENCES

[1] Holmes, Kathryn V. "SARS-associated coronavirus." New England Journal of Medicine 348.20 (2003):1948-1951.

[2] Alpaydin, Ethem. Introduction to machine learning. MIT press,2020.

[3] Lee, Jae Won. "Stock price prediction using reinforcement learning." ISIE 2001. 2001 IEEE International Symposium on Industrial Electronics Proceedings (Cat. No. 01TH8570). Vol. 1. IEEE,2001.

[4] Seber, George AF, and Alan J. Lee. Linear regression analysis. Vol. 329. John Wiley & Sons,2012.

[5] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." Journal of machine learning research 12.Oct (2011):2825-2830.

[6] Pant, Ayush. "Introduction to Linear Regression and Polynomial Regression." Medium, Towards Data Science, 16 Jan. 2019, towardsdatascience.com/introduction-to-linear- regression-and-polynomial-regression-f8adc96f31cb.

[7] Maier, Benjamin F., and Dirk Brockmann. "Effective containment explains sub- exponential growth in confirmed cases of recent COVID-19 outbreak in Mainland China." arXiv preprint arXiv:2002.07572(2020).

[8] Dubé, C., et al. "Comparing network analysis measures to determine potential epidemic size of highly contagious exotic diseases in fragmented monthly networks of dairy cattle movements in Ontario, Canada." Transboundary and emerging diseases 55.9-10 (2008): 382- 392.

[9] Homer, Steven. "Minimal degrees for polynomial reducibilities." Journal of the ACM (JACM) 34.2 (1987):480-491.

[10] Liu, Ying, et al. "The reproductive number of COVID-19 is higher compared to SARS coronavirus." Journal of travel medicine(2020).

[11] Wu, Zunyou, and Jennifer M. McGoogan. "Characteristics of and important lessons from the

coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention." Jama(2020).

[12] Grasselli, Giacomo, Antonio Pesenti, and Maurizio Cecconi. "Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: early experience and forecast during an emergency response." Jama(2020).

[13] Villari, Paolo, et al. "Unusual genetic heterogeneity of Acinetobacter baumannii isolates in a university hospital in Italy." American journal of infection control 27.3 (1999):247-253.

[14] Fanelli, Duccio, and Francesco Piazza. "Analysis and forecast of COVID-19 spreading in China, Italy and France." Chaos, Solitons & Fractals 134 (2020):109761.

[15] Harrell Jr, Frank E., et al. "Regression models for prognostic prediction: advantages, problems, and suggested solutions." Cancer treatment reports 69.10 (1985):1071-1077.

[16] Moore, Rachael E., et al. "Nasal mask." U.S. Patent Application No.29/166,190.