Prediction of an epidemic with Machine Learning and Covid-19 Data

Fang Wenhui[1,a], Wang Yihui[2,b], Lu Zhipeng[3,c]

[1]University of California, Santa Barbara, California, United States, 93117.
[2]University of California, Santa Barbara, California, United States, 93117.
[3]Sichuan University, Chengdu, Sichuan.

Abstract. The human race has already overcome many epidemics such as smallpox, SARS, and Black Death with vaccines or cures. As the number of infected people climes up to 10 million due to the new coronavirus in 2020, the human race faced another public health challenge. Because of the strong infectivity of the new coronavirus, humans have not won this fight after half a year. During the times of defeating these viruses, humans sacrificed not only wealth but also lives. Apart from many tribulations, human race also has great development in on technology. Machine learning method was invented and applied in many fields such as robotics, healthcare, and medicine. Since the transmission of a virus is related to social factors such as the percentage of college degree, and population density, there is a model built in this article that only related to outside factors such as health insurance coverage to predict that when the climax of an epidemic will arrive by using machine learning techniques and data related to Covid-19. Since the model does not take transmissibility of one specific virus, this model can apply to any epidemics to forecast the peak with enough data.

1 INTRODUCTION

World Health Organization states that the transmission of Covid-19 between people is based on the contacts with droplets caused by infected people. Therefore, when scientists predict the progress of an epidemic, the contact rate between people is an important factor. One way to quantify the contact rate is the population density. According to Hami and his coworkers, the density only is the third most important factors in transmission. There are many other factors he measured such as the relationship between counties. Other factors they mentioned include the percentage of college educated, percentage of population aged 60, etc. In his research, due to the complexity of the pandemic spread, the factors such as health-related are not under consideration. [1] Machine Learning Models find the relationship in data and refine the models to make predictions. Since there are social factors affecting the transmission, the purpose of this project is to utilize machine learning models to build a model that can predict the progress of an epidemic based on social factors. Because this model does not rely on the characteristics of one specific virus, it can be used not only in Covid-19 epidemic, but also virus that appears in the future.

2 COVID-19

In December 2019, unknown pneumonia was discovered in Wuhan, China. Scientists soon identified it as a new coronavirus and named it Covid-19. Because of its strong infectivity, it spread to the whole world in a very short time. Without a specific medicine that can treat and prevent Covid-19 [2], after six months, the situation continued to deteriorate with faster growth speed of world cumulative cases. On June 29, the number of total infected in the world has reached ten million, and over fifty thousand people have lost their lives because of the pandemic. [3] With the economic stagnation, and the shortage of medical supplies, people are struggling and wondering when this disaster will come to an end.

According to WHO, the main route of transmission includes contacts with droplets containing virus created by infected people's sneezes or coughs directly and indirectly. When these droplets entered the respiration system, a healthy person has a high probability of infection. People with no symptoms can also spread Covid-19 virus to other people as well [4] With this route, the rate of contact plays a significant role in the spreading of Covid-19, and it is associated with social factors such as population density, unemployment rate.

Apart from infectivity, the rate of hospitalization is important to the spread of the disease. By the CDC's statistics, 19% of infected people were hospitalized in the United States. [5] Health care coverage, family income,

Email: [a]wfang@ucsb.edu, Email:[b]yihuiwang@ucsb.edu, Email: [c]2017141491001@stu.scu.edu.cn

and other social factors have influences on the hospitalization rate.

3 MACHINE LEARNING

With the removal of the intrinsic characteristics of Covid-19, a machine learning model was trained to predict when the peak of the plague will reach. In other words, the model outputs the time when the highest new cases will come only with outside factors such as population density and health care coverage. Machine learning is a method pertaining to artificial intelligence. The model is trained by a large number of data so that the model can gather experience and improve itself to output the predictions with limited information [6]. With the development of technology, the completeness, variance, and accuracy of data were much improved, and the machine learning method was applied in many fields.

4 METHODOLOGY

4.1 Study area

The data for training the model is from the United States. Despite that the U.S has the most people infected in the world, it has the most detailed data. The government and private organizations gather information by the unit of the city. All the numbers are public and easy to find. With the large size of the dataset we build, we can eliminate all the invalid data as precisely as possible without worry about scarcity of data for training, and it is easier to obtain accuracy with detailed data.

4.2 Data

To have detailed data, data are collected by the unit of the city. All the information concerned with Covid-19 is obtained from John Hopkins University. [7] The dataset includes the number of new cases and cumulative cases everyday for over 3000 cities. To make the prediction only related with social factors, relative statistics are proceeded by government and private organizations [8] and also collected by the unit of the city. Since they came from different resources and have different demarcations for places, unmatched data are considered invalid and disposed of. Furthermore, some districts have small infected numbers as a result of small populations, which do not have strong influence or representativeness for training models. These data are removed as well. At last, the dataset used for training is composed of more than 600 cities with corresponding data.

There are nine categories, including population density, health insurance, the median of household income, the median of resident ages, and other factors that possibly have an influence on the transmission.

4.3 Machine learning models

There is no specific machine learning model for all questions because of the variance of sizes and types for datasets. Therefore, we apply seven machine learning models and let them vote for the results to have the best estimation. Compared to the unsupervised learning type without output data provided, all models used in this research are classified as supervised learning with input and output data provided for models. [9] The seven models are Decision Trees, Random Forest, Logistic Regression, Naïve Bayes Classification, Support Vector Machines (SVM), Linear Discriminant Analysis (LDA), K-nearest Neighbors (KNN). The following information is a brief introduction to these models.

Decision Tree is one of the most used models in machine learning problems with easily-reading feedback and the capability of processing numerical and categorical data. [10] With the basic concept of Decision Tree, Random Forest utilized several decision trees. By averaging the results of decision trees, the Random Forest model can process datasets of large scale or with variables that are more than observations. [11] Probability plays an important role in predication, and many machine learning models such as Logistic Regression and Naïve Bayes are based on statistics. Logistic Regression predicts the probability of variables. The type of Logistic Regression we used is multinomial that can output many types[12]. Naïve Bayes Classification utilizes Bayes' Theorem related to probability and assumes the independence between predictors. "Naïve" represents the computer to determine a feature is independent with any other feature. [13] Support Vector Machine (SVM) can be used for both regression and classification with less computational power required, but a higher accuracy produced. [14] Linear Discriminant Analysis (LDA) is also a classification model, and it quantifies the differences between objects and assigns them into different categories or features. [15] The last one is K-Nearest Neighbors, and it classifies objects by testing the resemblance of objects within a small distance. [16]

4.4 Machine learning approach

The basic process is explained in figure 1. After the data collection and the elimination of invalid data, all the data was classified and compiled in a dictionary for convenience for further study when the model request specific data in Python. Also, labeling each object is also required in machine learning to underscore their characteristics. Since the purpose of this model is to predict when the peak will come, we need to quantize the meaning of peak. In this study, the labels are the duration of reaching the highest increase. To be specific, the label is obtained by subtraction the days when the first confirmed case appeared from the days of the highest daily increase, which is divided by 7 to have the label with units of weeks. After handling the dataset, the dataset was divided into two parts with the training set and test set, and the ratio of them is 9 to 1 correspondingly. After the complication of training 7 models, the program will let 7 models vote for the results and validate the result by comparing the predictions and the labels.

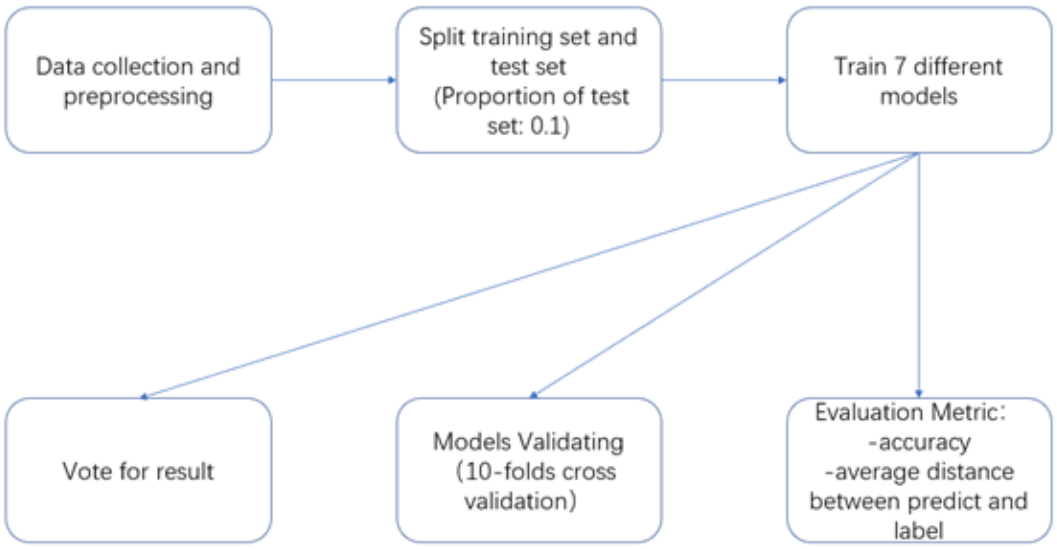

Fig. 1. Machine Learning Approach

5 RESULT AND DISCUSSION

Figure 2 compares the validation set and the predication. The average distance between labels (actual numbers of weeks) and predication is 0.9 weeks, which means that the probability that predication will have the same number (weeks) with the labels is 90%.

Because there is no particular machine learning to solve all the questions and the complexity of social factors, it combines 7 models in this research and let them vote for the results. In this way, the influence of one model's disadvantages can be best avoided. For example, Decision Trees can develop an overly complex model due to over-fitting problems caused by the excess of data. [10] To counteract this disadvantage, we introduce random forests to average the results of serval decision trees. Based on the purpose of avoiding defects of one specific model, the selection of other models follows the same rule. By the result, we obtain a good prediction with this method.

The model includes nine different aspects concerning nine social issues. Since the results indicate that the model has 90% correctness, by using the Covid-19 data, the result indicates that machine learning techniques can successfully find the relationship between social conditions and transmission of the virus. This model can give advice to the government of changing social conditions to shorten the length of the peak's arrival.

In a future study, more data can be obtained that are related to the transmission of the virus. For example, the number of physicians available for 1000 people can be taken into consideration. To consider factors that are hard to quantify, such as the enforcement of quarantine measurements, social distancing, and masks orders, we need to introduce medical knowledge such as the change in transmission rate and the diagnosis rate. Both will be infected by quarantine measures [17]. Also, because the data is collected in the United States, the data set for training can be aggrandized by adding other countries' data. Moreover, deep learning pertains to machine learning, but it permits software to learn and predict. We can explore this field with this project.

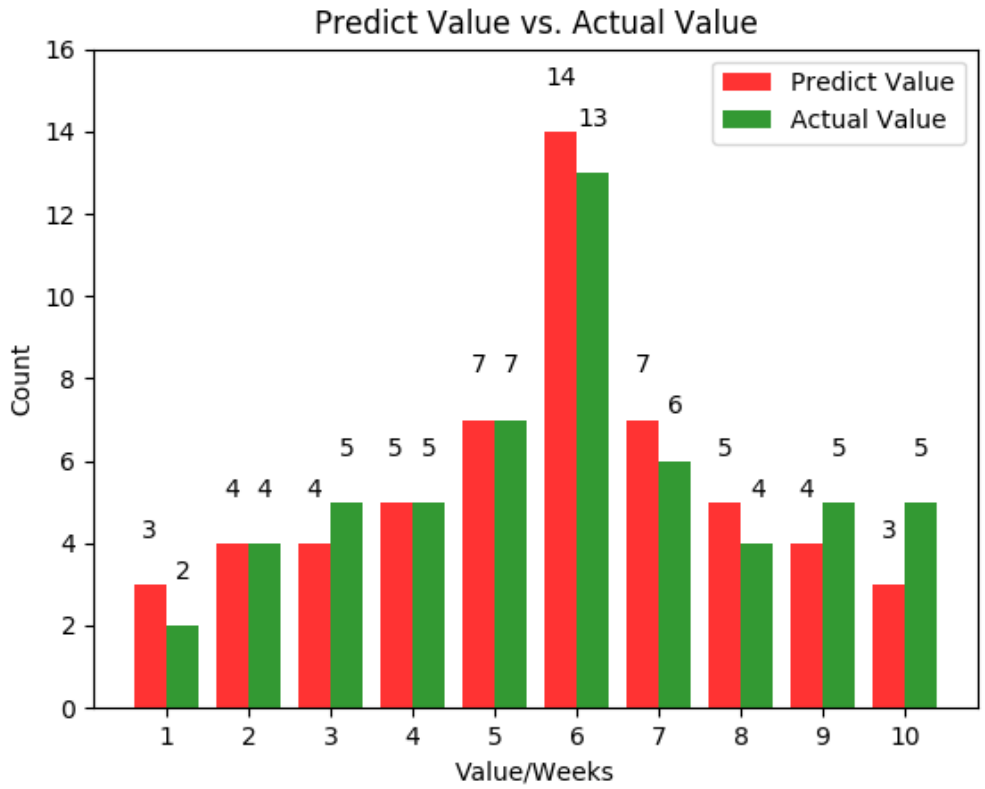

Fig. 2. Predict Value vs. Actual Value

6 CONCLUSION

Because there is linkage between social factors and the transmission of a virus, this research picked nine categories include population density, education, and other social factors, along with the new confirmed Covid-18 cases data, to train machine learning method. With this project, the model is shaped with these data, and it is expected to return the prediction of the epidemic peak time if it finds the relationship. Since this model does not utilize any categories that are specific characteristics of virus such as infectivity, mortality, and influences to different human races, the model is expected not only to be an application to Covid-19, but also any virus that will appear in the future.

The project incorporates seven models to counteract the disadvantages of each model to acquire higher accuracy. As the results present, the model successfully finds the linkage between transmission and social conditions with data in the United States. The results indicate that the predication will have 90% probability that the result will be one weak difference from the actual data. Despite the relatively high accuracy, the model still can be improved with more relevant data such as numbers of hospitals, and pre-existing health-issues including diabetes. Also, the dataset can be aggrandized with data from other countries. With more data, the accuracy of the model can be improved.

ACKNOWLEDGEMENT

Many people including my professors and classmates offered me valuable help in my thesis. Valuable biologic knowledge is given by Prof. Ottoman in Massachusetts Institute of Technology. He taught me once in an online course about data analysis in biology. Furthermore, I am grateful with my classmates with their assistance throughout the process of writing. I'd like to express my gratitude to all the people who not only give help in the writing but also offered emotional help.

REFERENCES

1. S. Hamidi, S. Sabouri, R Ewing. (2020) "Does Density Aggravate the Covid-19 Pandemic?" Journal of the American Planning Association, DOI: 10.1080/01944363.2020.1777891
2. WHO, (2020). *Mythbusters*. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters?gclid=EAIaIQobChMIv5Px68zT6gIVQ38rCh2IQwpaEAAYASAAEgJlD_D_BwE#medicines
3. Lisa Du, (2020), *Pandemic Tops 10 Million Cases, 500,000 Deaths as Momentum Grows.* https://www.bloomberg.com/news/articles/2020-06-28/global-covid-19-cases-hit-10-million-as-pandemic-gains-momentum
4. WHO, (2020). *Q&A: How is COVID-19 transmitted?*. https://www.who.int/news-room/q-a-detail/q-a-how-is-covid-19-transmitted

5. Centers for Disease Control and Prevention, 2020. *Clinical Care Guidance*. https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html
6. Bells, J., (2014) "What Is Machine Learning" in: Machine Learning, John Wiley & Sons, Inc. Indianapolis. pp. 2.
7. John Hopkins University, (2020). Coronavirus Resource Center. https://coronavirus.jhu.edu/us-map
8. City-Data, (2020). https://www.city-data.com/
9. Bells, J., (2014) "What Is Machine Learning" in: Machine Learning, John Wiley & Sons, Inc. Indianapolis. pp. 3.
10. Bells, J, (2014) "Wording with Decision Trees" in: Machine Learning, John Wiley & Sons, Inc. Indianapolis. pp. 46.
11. Biau, G., Scornet, E. (2016) A random forest guided tour. TEST 25, 197–227 https://doi.org/10.1007/s11749-016-0481-7
12. Tutorialspoint. Machine Learning - Logistic Regression. https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_logistic_regression.htm#:~:text=Logistic%20regression%20is%20a%20supervised,be%20only%20two%20possible%20classes.&text=Mathematically%2C%20a%20logistic%20regression%20model,as%20a%20function%20of%20X.
13. Anand Venkataraman., (2019) Naive Bayes for Machine Learning - From Zero to Hero. https://blog.floydhub.com/naive-bayes-for-machine-learning/#:~:text=And%20the%20Machine%20Learning%20%E2%80%93%20The,presence%20of%20any%20other%20feature.
14. Bells, J, (2014) "Support Vector Machines" in: Machine Learning, John Wiley & Sons, Inc. Indianapolis. pp. 139-140.
15. N. Mohanty, A. Lee-St. John, R. Manmatha, T.M. Rath, (2013) "Shape-Based Image Classification and Retrieval" in: Handbook of Statistics, Volume 31. pp. 249-267.
16. S. D. Jadhav and H. P. Channe, (2016) "Comparative study of K-NN, naive bayes and decision tree classification techniques," International Journal of Science and Research (IJSR), vol. 5, no. 1, pp. 1842–1845. https://www.ijsr.net/archive/v5i1/NOV153131.pdf
17. Weston C. Roda,Marie B. Varughese, Donlin Han, Michael Y. Li, (2020) "Why is it difficult to accurately predict the Covid-19 epidemic?" Infectious Disease Modelling 5 (2020) pp. 271-281