Applying Machine Learning to Predict COVID-19 Transmission and Mortality on a County-level Basis in the USA

Author: Otto Beall

Author Bio:

Otto Beall is a high school researcher from Plano, Texas. He is active in science fair, having competed at the Dallas Regional Science fair, the Texas Junior Academy of Science, and the Texas State Science fair. His project on developing a low-cost Braille embosser won grand prize in the junior physical division. He participates in Science Bowl, NAQT, and is an NHS officer of his local school chapter. His personal passions are computer science, electrical engineering, politics, and Civil War history. When he's not in the garage working on a hobby project or reading the latest edition of Science magazine, you may find him sailing or tending to his vegetable garden.

Abstract

Since the COVID-19 virus emerged in late 2019, the resulting pandemic has caused global economic disruption and taken millions of lives. The rampant spread of the disease, even among developed countries with advanced healthcare infrastructure, highlights the need to understand the dynamics of the COVID-19 pandemic on the community level, not just at the level of the individual patient. As the pandemic highlighted structural inequalities in our disease-prevention systems, it is of particular interest how the demographic features of communities affect their vulnerability to COVID-19. Machine learning offers a new approach to analyze patterns in publicly-available COVID-19 data, and in doing so, provides opportunities to improve prediction models and improve the equity of pandemic response.

While existing research has applied machine learning to diagnosing COVID or predicting the likelihood of an individual's mortality, the purpose of this project is to use machine learning to determine the link from socioeconomic and demographic factors to the rate of transmission and mortality of COVID in US communities.

This project used a dataset containing COVID-19 and demographic data (sourced from the New York Times COVID-19 database and the US Census Bureau, respectively) indexed at the county level. This data was used to train neural networks of varying complexity. Each model was optimized to predict either a county's transmission rate (cases as a portion of the total population) or mortality rat (deaths as a portion of the total population) when given a certain combination of demographic information about a county.

The simplest class of model with one hidden layer and plain inputs consistently outperformed both linear regression and more complex neural networks at predicting COVID-19 transmi

This research has broader implications that machine learning is a powerful tool in predicting the spread of infectious diseases in our communities, and yielded important information on the impact of demographic and socioeconomic factors on the COVID-19 pandemic.

Key Words: neural network, county-level, transmission rate, mortality rate, input layer, hidden layers, demographic inputs, training set, validation set, overfitting

Introduction

Rationale

In late 2019, an obscure strain of coronavirus emerged from a wet-seafood market in Southern China. Though the local flare-up initially went unnoticed by the world, the virus spread rapidly to all corners of the inhabited world, infecting hundreds of millions, deluging our overburdened health systems, inflicting a worldwide economic recession, and most tragically, taking the lives of five million human beings. In short, the COVID-19 pandemic dealt humanity the greatest—and most unexpected—global health challenge of the 21st century. It exposed significant flaws in our disease-control networks and persistent inequities in our society at large. Alarmingly, even developed countries like the United States largely failed to contain the spread of the disease. Moreover, the United States witnessed wide disparities in terms of COVID-19 transmission and mortality between communities of different socioeconomic statuses. The highly transmissive nature of COVID-19 necessitates an analysis of not just an individual patient's response to the disease but the characteristics of transmission and mortality within a community at large. The goal of this project is to use Machine Learning to understand the deeper relationship between communities' demographic factors and their vulnerability to COVID-19. By training supervised machine-learning models using publicly available demographic US Census data to predict the rate of COVID-19 transmission and the overall death rate from the disease, this project aimed to shed light on the relative importance of these demographic factors in determining a county's predicted COVID-19 case and death counts.

Literature Review

One group of researchers in Hungary attempted to use time-sequenced data to predict the course of future outbreaks. (Pinter et al., 2020) The goal of their research was to use advanced machine learning algorithms, including the Hybrid Multi-Layered Perceptron-Imperialist Competitive Algorithm and the Adaptive neuro-fuzzy inference system, to improve upon pre-existing Susceptible-Infected-Recovered models for predicting rates of COVID-19 transmission in Hungary. Their models, however, did not take demographic information of individuals or regions into account when predicting future case counts. One meta-analysis of Machine Learning applied to the COVID-19 pandemic highlighted the use of machine learning models to process x-ray images and CT-scans to diagnose individual patients (Bachtiger et al.,

2020, p. 1). Another meta-analysis cited the use of machine learning to process medical imaging, to identify clusters of symptoms as indicative of the disease, and to give personalized treatment plans based on demographic factors of the patients. (Kushwaha et al., 2020) However, little existing research has been done on applying Machine Learning to study the effect of demographic factors on the cumulative spread of COVID-19 in distinct geographic communities, particularly US counties.

Supervised Machine Learning

The type of Machine Learning used in this project is Supervised Machine Learning, in which models are trained on datasets to make predictions from a certain input. The training data for a Supervised Learning model must include input data (also known as labels) and output data. The model takes in a given set of inputs and returns an output value. Through repeated iterations of measuring its performance on the known data and tweaking its algorithm slightly to better match the known data, the model gradually improves its performance, so that it can then make predictions for data that it was not trained. For this project, the data comprised of the demographic data for each county (inputs) and the COVID-19 case and death counts as a portion of the population (outputs).

The simplest form a Supervised Machine Learning model is the linear regression. A linear regression model will multiply each of the inputs by a fixed value, then add these terms to a fixed constant to give its final result. When a linear regression model is trained to predict the case percentage of a certain county given only the portion of the county which is college educated and the portion of the county which has health insurance, it produces an equation of the form:

$$\textit{(Transmission Rate)} =$$
$$w_1 * \textit{(portion college educated)}$$
$$w_2 * \textit{(portion with health insurance)}$$
$$+b$$

To find the estimated transmission rate for a certain county, it receives the known values, multiplies each by a certain weight (or coefficient), and adds to the bias (or intercept) of the output node *(Fig. 1)*.

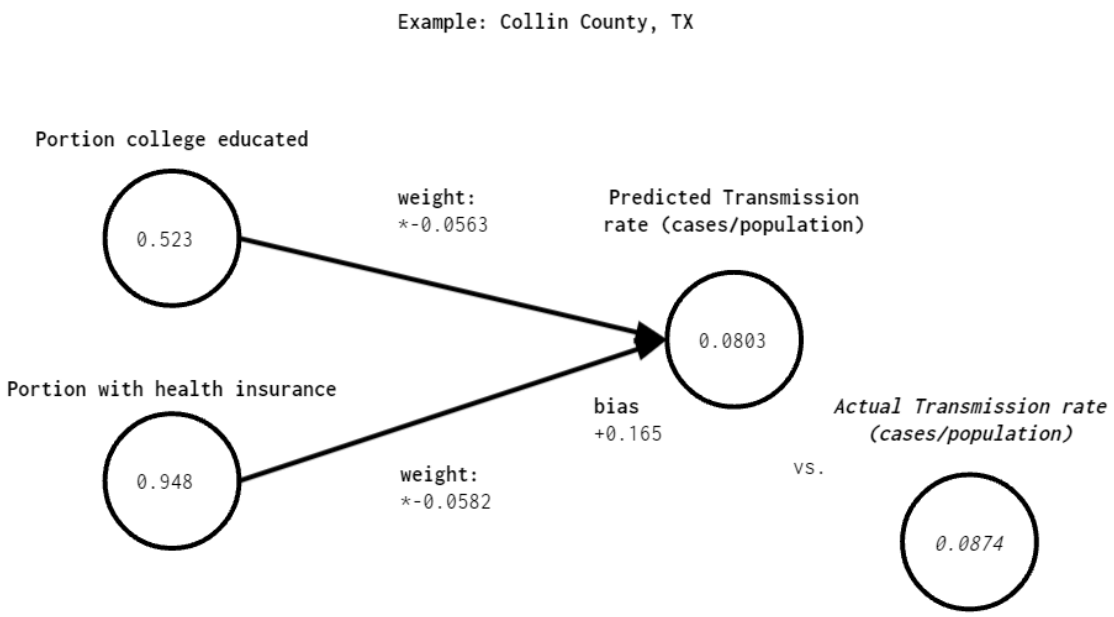

Figure 1: In this representation of a linear model, the output node adds weighted inputs to a predefined bias to produce a prediction.

The linear model, however, can only predict based off simple linear patterns in the data. Its simple structure can be extended to have multiple layers of nodes *(Fig. 2)*, each taking weighted inputs from previous layers, and adding on to their own biases in the same manner. This way, the hidden layers may process deeper patterns in the data which defy a simple linear or even polynomial regression to produce a more accurate prediction,

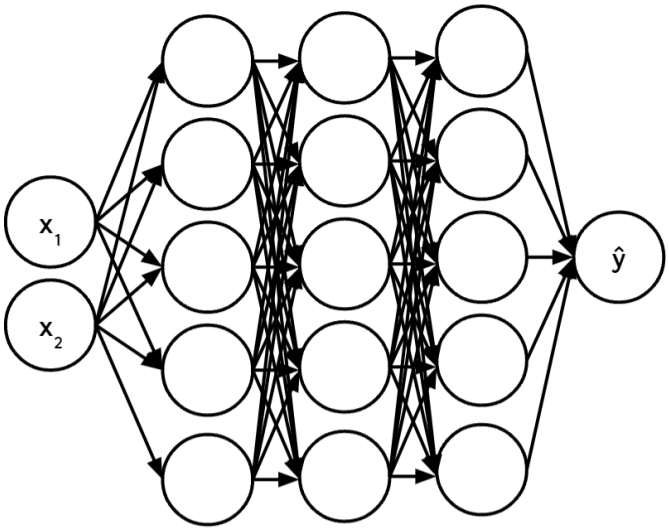

Figure 2: Diagram of a multi-layer model with three hidden layers

The weights and biases of this multi-layer neural network are initialized randomly, and through many iterations of testing, are gradually adjusted using an optimization algorithm known as gradient descent to minimize the deviation of the predictions from the actual values. In this project, multiple sizes of models trained with different combinations of demographic predictors are used to predict both case and death rates from COVID-19.

Methods

Data collection

For this project, a dataset was compiled from multiple US Census Bureau sources and the New York Times COVID database. The Census Bureau collects data on the total population of each county, the portion of each county's population which is non-Hispanic white, the portion of each county's population below the poverty rate, the portion of each county's population with a college degree, the portion of each county's population over 65, the portion of each county's population which has health insurance, and each county's land area (from which population density may be calculated. Each of these demographic factors was incorporated as a possible predictor in the overall dataset. Each row of data, representing one of over 3000 counties in the US, was indexed according to the Federal Information Processing Standard (FIPS) assigned county-level index number.

The cumulative COVID-19 case and death counts up to July 8th for every county (also indexed by the FIPS standard) were sourced from the New York Times COVID-19 database, which compiled information from state and local government health websites. The two values, divided by the overall population, gave the total transmission rate and mortality rate from COVID-19 in each county. These two values, ranging from 0 to 1, were the values which the models were trained to predict given only selected demographic data from a certain county.

Models Trained:

Each model is a neural network composed of one input layer with a node for each input, zero to four "hidden" layers with ten nodes each, and a single output node giving the predicted value. Each model may take in the unmodified input values (power one) or quadratic combinations of input values (power two), that is the plain values along with every input raised to the power two and every pairwise product of input values The models are created from the MultiLayerModel class (developed for this project), which makes use of the Sequential class from tensorflow.keras:

Each instance of this class has the attributes "layers," ranging from two to five, and the attribute "power," ranging from one to two. (Note: the number of layers only includes the number of hidden layers plus the one output layer, but not the input layer. For example, a two-layer model, has one hidden layer and one output layer.) In total, for this project, eight total neural network designs were tested.

Training and Validation:

Multiple neural networks with the same structure were each trained to predict a certain output from a certain combination of demographic inputs. Each individual model with one of the ten neural network designs were trained to predict either the transmission rate (total cases/population) or the mortality rate (total deaths/population), given either…

- All six demographic factors (portion college educated, poverty rate, portion white, portion over 65, population density, and portion with health insurance)
- All but one of the six demographic factors (six possible combinations)
- Every pair of demographic factors (twenty-one total combinations)

Additionally, for each of the combinations of inputs and outputs, a linear model was created (using the LinearModel class of the sklearn library) to serve as a control.

To test the accuracy of the models, they are given the demographic information of counties outside the dataset on which they were trained. At the beginning of each program, the county dataset is split randomly into a "training set," comprising 75% of the data, and a "validation set," comprising 25% of the data. The models are all trained on the counties in the training set, and then produce their predicted values (transmission or mortality rate) from the demographic information in the validation set. These values are compared against the true COVID-19 values for the counties in the validation set and the average error of every model is measured by the mean absolute error (the mean of the absolute differences between the predicted and actual values). Thus, if a complex model becomes "overfitted" to its training set—that is, too perceptive of random noise in the training set, rather than the general pattern —this will be reflected by a lower mean absolute error from the validation set.

Results:

(Raw Data in Appendices)

The measure of loss for this project was standardized mean absolute error, the mean absolute error of each model divided by the standard deviation of true values in the validation set.

All demographic factors

The simplest neural network, the two-layer one-power model, had the lowest standardized mean absolute error for predicting both COVID-19 transmission and mortality of US counties when given all six demographic factors *(fig. 3)*. The two-layer one-power model outperformed both the linear regression model and the more complex models which had more hidden layers or higher powers of inputs.

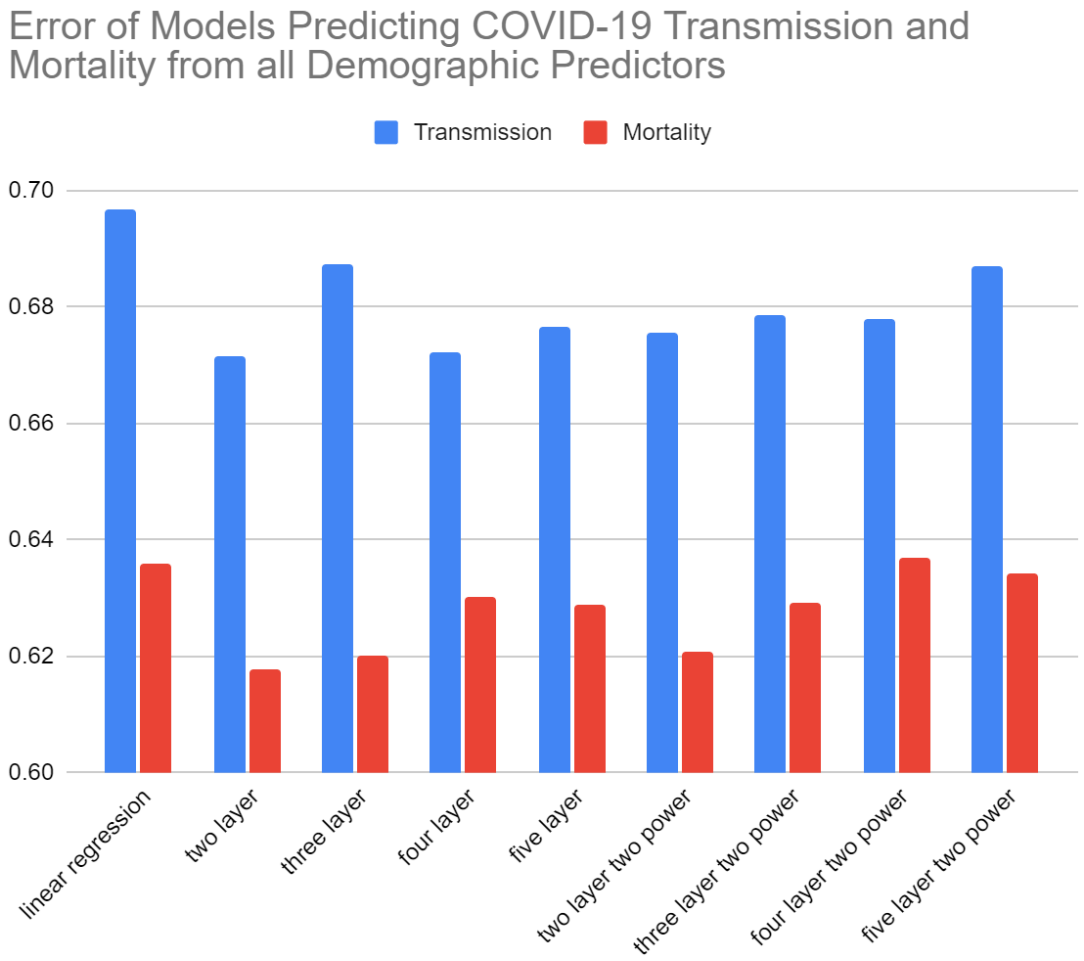

Figure 3: The error of different neural network structures in predicting either COVID-19 transmission or COVID-19 mortality in US counties. The x-axis indicates the structures of the models, ordered from left to right in increasing complexity and the y-axis indicates the resulting prediction error, in standardized mean absolute error.

All but One Factor

After determining that the two-layer one-power model was most accurate in predicting the COVID-19 data from all six demographic factors of US counties, one may examine how selectively removing factors from the model's training data affects the model's performance. If the error of the model in output increases significantly after a certain factor is removed, one may conclude that the factor played an important role in determining the output in question. For predicting COVID-19 transmission, removing the portion of the counties' populations over 65 years old from the model's training data caused the largest relative increase in the model's prediction error *(Fig.4)*. This suggests that the portion of a county's population which is over 65 plays the most major role in determining the county's COVID-19 transmission level. On the other hand, removing the portion of the counties' population which is college educated from the model's training data caused the largest relative increase in the model's prediction error. This suggests that the portion of a county's population which is college educated plays the most major role in determining the county's COVID-19 mortality rate.

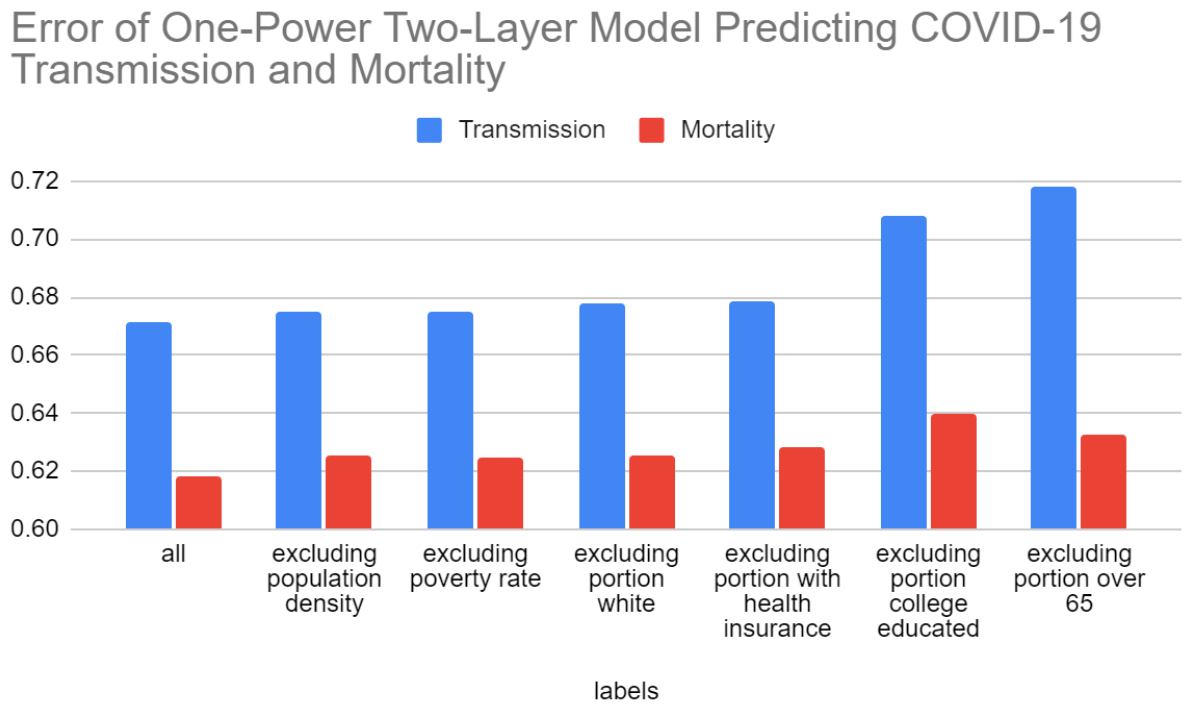

Figure 4: The error of the one-power two-layer model in predicting either COVID-19 transmission or COVID-19 mortality in US counties. The x-axis indicates the labels being excluded from the training data and the y-axis indicates the resulting prediction error, in standardized mean absolute error.

Pairwise Predictors

Averaged over all tests of pairwise predictors, the simple two-layer one-power model again performed best for predicting both mortality and transmission rates. The linear regression and the five-layer two-power models (the least and most complex models, respectively) performed the worst in both cases. On average, the best performing model, the two-layer one-power network, will predict, on average, 0.671 standard deviations from the true COVID-19 death count, and 0.729 standard deviations from the true COVID-19 case count, given two demographic factors of a given county.

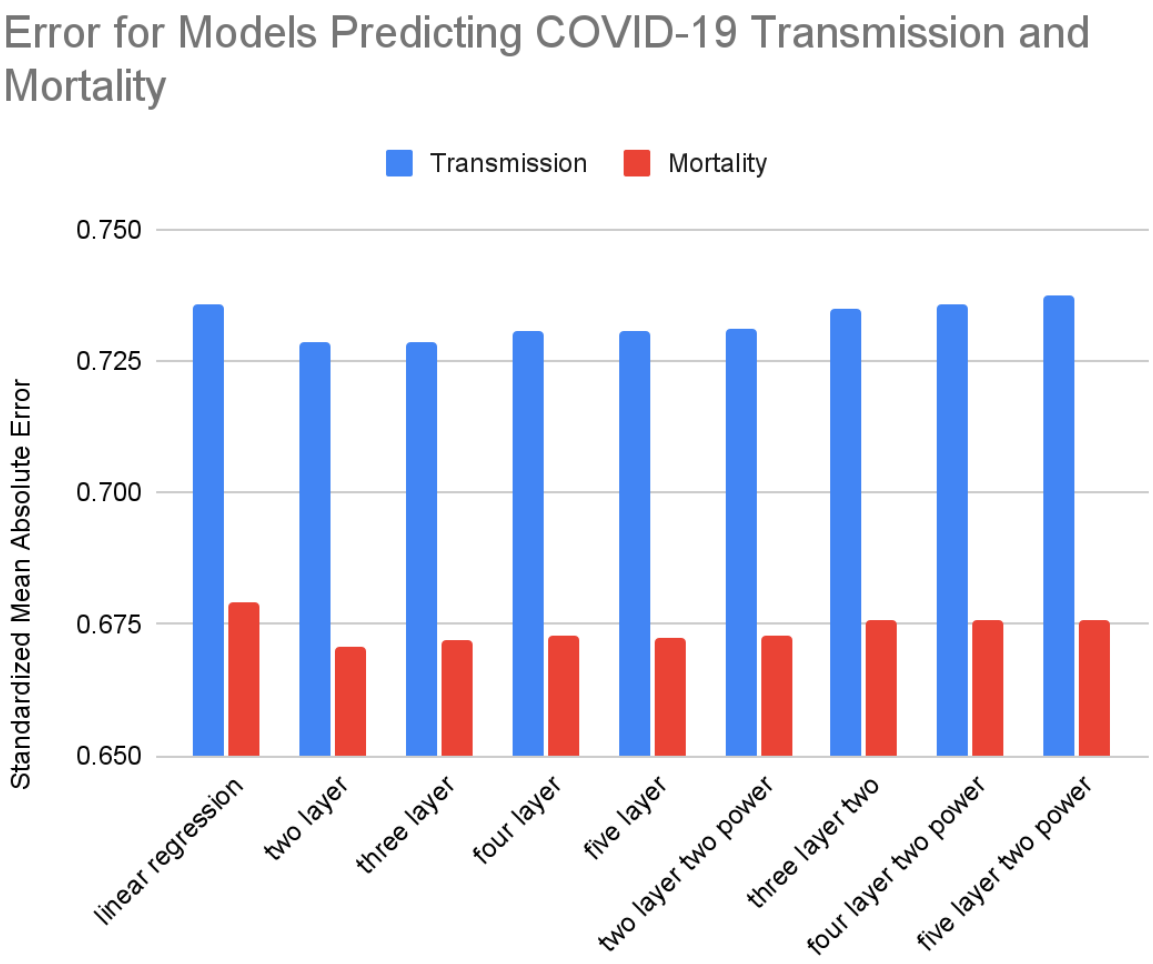

Figure 5: Error of models in predicting COVID-19 transmission or mortality averaged over all possible pairs of demographic predictors. The x-axis indicates the structures of the models, ordered from left to right in increasing complexity and the y-axis indicates the resulting prediction error, in standardized mean absolute error.

Analyzing these predictive factors just using the one-power two-layer model *(Fig. 6)* shows that the strongest pair of factors for predicting transmission rate of US counties consisted of the portion of the population with a college degree and the poverty rate. The strongest pair of factors for predicting transmission rate of US counties consisted of the portion of the population college educated and the

portion of the population over 65.

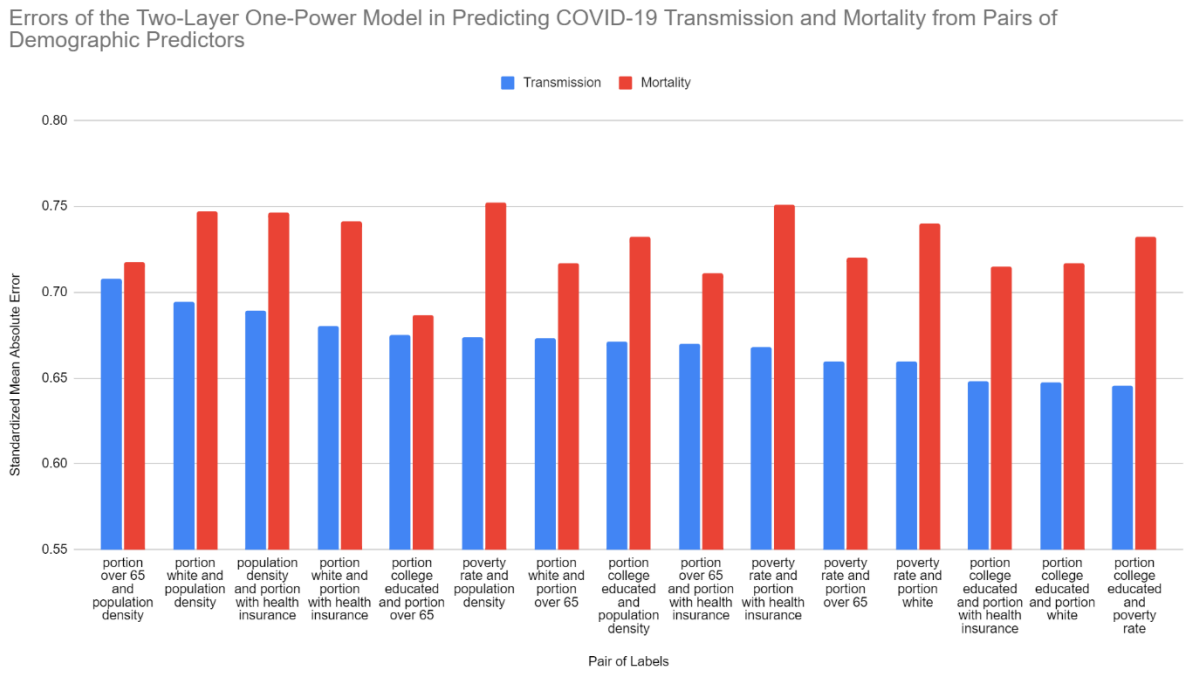

Figure 6: Error of the two-layer one-power models in predicting COVID-19 Transmission and Mortality . The x-axis contains all fifteen possible pairs of demographic predictors and the y-axis displays the resulting prediction error

Discussions and Conclusions

One of the most surprising results of the research is that the simplest neural network, which had only two layers and which received plain inputs, performed the best at predicting both COVID-19 transmission and mortality knowing all inputs and averaged over all pairwise combinations of inputs. The most likely reason why more complex models with more hidden layers and quadratic combinations of inputs did not perform better at detecting patterns in the data is overfitting of random noise. This occurs when the more complex models become attuned to random patterns in the training data, but these patterns are not reflective of general trends in the data. Thus, while these models may achieve higher accuracy on the training set, they have greater prediction errors on the validation set.

Visual evidence of overfitting is apparent when plotting the real values and predicted values different models on a scatterplot. On the following figures *(Figs. 7-9)*, the portion white is plotted on the x-axis and the COVID-19 mortality rate is plotted on the y-axis. The black points represent the true values of all the counties in the validation set, while the red points have y-axis values predicted by a certain model, which was trained knowing all demographic factors. The three graphs are sorted by the increasing complexity of model, starting from the linear model. The predicted values of the linear model follow a constricted linear cluster. The predicted values of the simple two-layer one-power model are more spread out, indicating that the neural network is more capable of predicting the full range of true values. However, in the complex five-layer two-power model shows clustering along parallel sinews which do not reflect any trends in the real data, suggesting that the model has overfitted to random noise in the training set. This suggest a reason why the simplest neural network was able to outperform the linear regression and the more complex neural networks in terms of prediction accuracy.

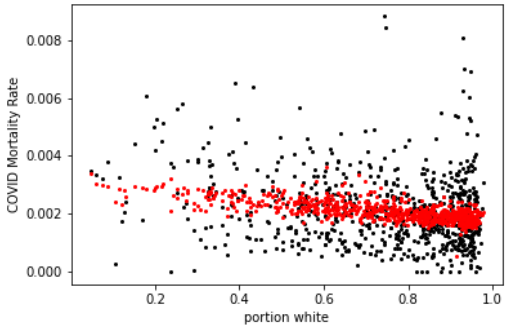

Figure 7: real and predicted values from a simple linear regression model

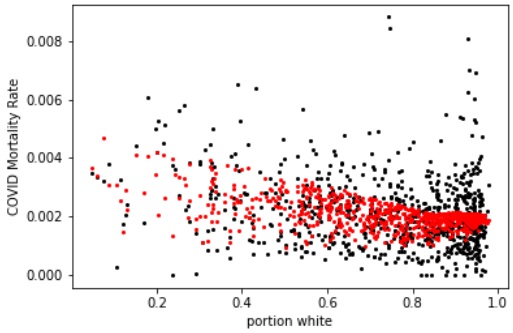

Figure 8: real and predicted values from the two-layer one-power model

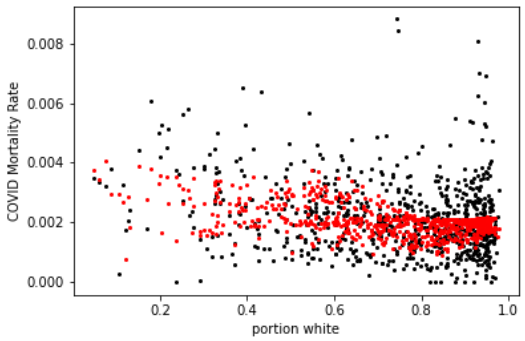

Figure 9: real and predicted values from the five layer-layer two-power model

The neural networks trained on the demographics and COVID-19 data yielded important insights into how demographic factors affect the vulnerability of communities to COVID-19. The results showed that the portion of the population over 65 had the most significant impact in determining the rate of COVID-19 transmission, and that the portion of the population with a college education had the most significant impact in determining the rate of COVID-19 mortality.

Notably, there is a negative correlation between the portion of the population over 65 and the transmission rate, suggesting that, although (or perhaps, because) senior citizens are at greater risk if they catch COVID-19, having a large portion of senior citizens in the population actually had a significant impact of reducing the disease's spread in the first place. Additionally, the portion of the counties' populations which were college educated had the most significant impact on the rate of COVID-19 mortality. This suggests that highly educated communities either had greater insulation from the disease (through remote work) or had habits and attitudes broadly conducive to limiting the rate at which people died of COVID-19.

These facts on their own may yield powerful information to scientists and policymakers on how to distribute medical resources to combat COVID-19. Additionally, the models themselves could also be used to estimate the relative vulnerability of communities as demographics change over time, or estimate

how a future pandemic would affect different communities unequally. There are limitations to this project, particularly how even the most powerful models would guess on average more than half a standard deviation from the true value. Moreover, the error values of the linear regression and neural networks were very close to one another, with judgements being made on miniscule differences in accuracy of the models. Yet, the varied trials offered self-consistent results, and with more advanced machine learning techniques, including regularization and adversarial neural networks, this research could yield a highly effective tool for predicting the impact of COVID-19 and other similar diseases.

Cited Literature

Bachtiger, P., Peters, N. S., & Walsh, S. L. (2020). Machine learning for COVID-19—Asking the right questions. *The Lancet Digital Health*, *2*(8), e391–e392. https://doi.org/10.1016/S2589-7500(20)30162-X

Kushwaha, S., Bahl, S., Bagha, A. K., Parmar, K. S., Javaid, M., Haleem, A., & Singh, R. P. (2020). Significant Applications of Machine Learning for COVID-19 Pandemic. *Journal of Industrial Integration and Management*, *05*(04), 453–479. https://doi.org/10.1142/S2424862220500268

Pinter, G., Felde, I., Mosavi, A., Ghamisi, P., & Gloaguen, R. (2020). COVID-19 Pandemic Prediction for Hungary; A Hybrid Machine Learning Approach. *Mathematics*, *8*(6). https://doi.org/10.3390/math8060890

Appendices

Pairwise Factors-Predicting Cases as a portion of Population-Standardized Mean Absolute Error

Pair of Demographic Labels	linear regression	two layer	three layer	four layer	five layer	two layer two power	three layer two power	four layer two power	five layer two power
poverty rate and population density	0.7546119856	0.7519390629	0.7527417016	0.7524641227	0.7521718973	0.7558475083	0.7548160539	0.7572025485	0.7518696708
poverty rate and portion with health insurance	0.7537842266	0.7510170826	0.7519795038	0.7519853195	0.7557599053	0.7553025807	0.7597028301	0.7549324633	0.7619531347
portion white and population density	0.7498149834	0.7471305971	0.7469034882	0.7461992983	0.7496019177	0.7499201356	0.7530464194	0.7539049311	0.7494353029
population density and portion with health insurance	0.7536154754	0.7483397625	0.7490381459	0.7468255195	0.7531360714	0.7584109529	0.7598223958	0.7634987558	0.7706538909
portion white and portion with health insurance	0.7522697837	0.7412938356	0.7374612433	0.7420449883	0.7442857458	0.7503548475	0.7545939799	0.7651609057	0.7682323498
poverty rate and portion white	0.7545886299	0.7403426419	0.740938712	0.7411195868	0.7449054884	0.7425327091	0.7407861269	0.7474674131	0.7470676543
portion college educated and population density	0.731423449	0.732286932	0.7307169815	0.7302903937	0.7315404209	0.7256208159	0.739025365	0.7273238246	0.7319963729
portion college educated and poverty rate	0.7352982854	0.7322161821	0.7315696564	0.7319147971	0.7326521956	0.7294155265	0.7386112954	0.7386709036	0.7329943789
poverty rate and portion over 65	0.7250525203	0.7197957606	0.7208974074	0.7199704089	0.7201077449	0.7236200322	0.7259751787	0.7245744467	0.7258997086
portion over 65 and population density	0.7238778974	0.7176313529	0.7192854214	0.7217254463	0.7228539069	0.7193418187	0.7222969371	0.7250594157	0.7254885873
portion white and portion over 65	0.7291052101	0.7171780573	0.7133628225	0.7266892623	0.7111694065	0.7137097059	0.7173286712	0.7150151625	0.7112922146
portion college educated and portion white	0.7240221781	0.717155727	0.7205207242	0.7135454073	0.7172436728	0.7180003779	0.7148909179	0.716125573	0.7186934357
portion college educated and portion with health insurance	0.7333085371	0.7150020323	0.713422002	0.7172751564	0.7223454106	0.7201800898	0.73960733	0.7252652289	0.7380786299
portion over 65 and portion with health insurance	0.727412301	0.7113912847	0.7134150509	0.7182101039	0.7121652965	0.7130890802	0.7147954301	0.7190544947	0.7328214887
portion college educated and portion over 65	0.6899691352	0.686819769	0.6888379057	0.700514258	0.6890853211	0.6900785097	0.6912778796	0.7023392825	0.6980608066
Average	0.7357836397	0.728502672	0.7287393845	0.7307182711	0.7306002934	0.7310283127	0.7349717874	0.7355730233	0.7376358418

Pairwise Factors-Predicting Deaths as a portion of Population-Standardized Mean Absolute Error

Pair of Demographic Predictors	linear regression	two layer	three layer	four layer	five layer	two layer two power	three layer two power	four layer two power	five layer two power
portion over 65 and population density	0.7342387179	0.7076197288	0.7074993225	0.7063901844	0.7086740247	0.7125807672	0.7108895796	0.709999562	0.7124238718
portion white and population density	0.7059515939	0.6945238661	0.6953566961	0.6894035767	0.6944703741	0.6948925801	0.6947809797	0.7046584624	0.6949139942
population density and portion with health insurance	0.6878272702	0.6895261892	0.6939583332	0.6892676175	0.6880159805	0.6888378911	0.6944843636	0.6957823106	0.6963575778
portion white and portion with health insurance	0.6859960785	0.6800364111	0.6819308541	0.6802954655	0.680934109	0.6815484416	0.6867712403	0.6878885571	0.6821231399
portion college educated and portion over 65	0.679793642	0.6748560578	0.6729907468	0.6715499277	0.6732449955	0.6715402307	0.6739546937	0.6743077169	0.6816784749
poverty rate and population density	0.6698056992	0.6739320723	0.6723065299	0.6743732003	0.6787822312	0.6818050409	0.6725773045	0.6736317444	0.6755998684
portion white and portion over 65	0.7015648058	0.6734847328	0.6759719241	0.6767556736	0.6755941831	0.6743850757	0.6780597012	0.6736777686	0.6769053878
portion college educated and population density	0.6727264043	0.6713274832	0.6690359985	0.6664277079	0.6663797647	0.6688000426	0.6763666863	0.6681247226	0.6705279188
portion over 65 and portion with health insurance	0.6863463969	0.6700265108	0.6682997856	0.6717275361	0.6688906903	0.6690822811	0.6709143696	0.6728545897	0.6781842256
poverty rate and portion with health insurance	0.6670589891	0.6677768343	0.6672434376	0.6735271147	0.6686852751	0.670993447	0.6780229282	0.6748780952	0.6819574263
poverty rate and portion over 65	0.6694568082	0.6598909467	0.6599256779	0.6581230092	0.6619967594	0.6568815973	0.6693634952	0.664814879	0.6657242094
poverty rate and portion white	0.6670898419	0.659875367	0.6675985065	0.6676679269	0.6685820875	0.6663504222	0.6710585159	0.6713263647	0.6751506578
portion college educated and portion with health insurance	0.6568950691	0.6483005227	0.6507966862	0.65758855	0.6508530013	0.6541823115	0.6557164773	0.66584496	0.6580632034
portion college educated and portion white	0.6612266899	0.6472573813	0.6498669384	0.6558217132	0.6489157803	0.6521445414	0.6513802494	0.6505450308	0.659365736
portion college educated and poverty rate	0.6517940934	0.6453479382	0.6484825924	0.6503865461	0.6492870837	0.6494797999	0.6526311325	0.6484861439	0.6482743745
Average	0.6791181399	0.6709188028	0.6720842687	0.672620382	0.6722204227	0.672900298	0.6757981145	0.6757880605	0.6758166711

All Factors and Excluding one at a time-Predicting Cases as a portion of Population-Standardized Mean Absolute Error

labels	linear regression	two layer	three layer	four layer	five layer	two layer two power	three layer two power	four layer two power	five layer two power
all	0.6967138818	0.6714857471	0.6872593861	0.6722724722	0.6765257678	0.6757726304	0.6786851986	0.6779573496	0.6869895433
excluding population density	0.6967783724	0.6749060574	0.6869793199	0.6806584506	0.6935033762	0.6899433494	0.6791339195	0.6772836267	0.6733400527
excluding poverty rate	0.6959673508	0.6752841616	0.6701622392	0.6697271761	0.6728607926	0.6741584932	0.677753278	0.6721646048	0.6751084058
excluding portion white	0.6954878907	0.6781454948	0.6797307914	0.6811047888	0.6926987005	0.6908494263	0.6829995776	0.6852724186	0.6806195744
excluding portion with health insurance	0.6934664961	0.6786586797	0.6806271341	0.6821916009	0.6713192247	0.6772622282	0.6743335477	0.679295458	0.6834887951
excluding portion college educated	0.7222081939	0.7078212477	0.7039749628	0.7036210837	0.7038764383	0.7025098376	0.7027503273	0.6955811533	0.7028399275
excluding portion over 65	0.7277522414	0.7177149097	0.7092784069	0.7157749172	0.7295694181	0.7175657302	0.723617204	0.7324898796	0.7178673454

All Factors and Excluding one at a time-Predicting Deaths as a portion of Population-Standardized Mean Absolute Error

labels	linear regressior	two layer	three layer	four layer	five layer	two layer two power	three layer two power	four layer two power	five layer two power
all	0.6358822774	0.6179353882	0.6201151623	0.6301880076	0.6288614494	0.6206339306	0.6291841776	0.6370271157	0.63424259
excluding poverty rate	0.6430382856	0.6245338853	0.6285255834	0.6242391883	0.6282946259	0.6357226091	0.6324125766	0.6320765195	0.63818506
excluding portion white	0.6441744741	0.625207264	0.6277949498	0.628225678	0.6324695681	0.6343290345	0.635445925	0.6439662274	0.6394724179
excluding population density	0.6377344272	0.625381315	0.6310513106	0.6335403881	0.6438958299	0.633005501	0.6336351489	0.6298554721	0.635376727
excluding portion with health insurance	0.6355291772	0.6285214382	0.6175842171	0.6273919588	0.6286188175	0.6329095115	0.6280796201	0.6302339398	0.6298503105
excluding portion over 65	0.637963804	0.6324862729	0.6318271815	0.6282728313	0.638648094	0.6353239397	0.6446581219	0.6458160365	0.6434176402
excluding portion college educated	0.6616448025	0.6396270743	0.6411111545	0.6351508149	0.6382490336	0.6438951651	0.6503364713	0.6588906262	0.6516273569

Code and training data:

https://drive.google.com/drive/folders/1_Q2FljoBMWTF21pSacdcojxlrJfhI4Kk?usp=sharing