Analyzing Machine Learning Models to Predict Covid-19 Deaths
By Nate Athorn, Vishal Madhav, Matthew Sirkot and Roger Yu
CMPSC 445 - Applied Machine Learning in Data Science

Abstract

COVID-19 has had a devastating impact on the lives and livelihoods of humans all over the globe. In Pennsylvania alone, the death toll has exceeded 26,000. In the past year, numerous predictive models have been created with the goal of informing policy decisions. These models have the greatest impact when they can be implemented efficiently on hardware all over the globe. To that end, our project compares low-cost Linear Regression and Neural Networks, built using case and death count data in Pennsylvania. In our analysis, we determined that low-cost models offer limited insight into the ongoing pandemic.

1. Introduction

The impact COVID-19 has had on Pennsylvanians cannot be underestimated. Since the virus began spreading through the commonwealth, businesses were ordered to close ("Gov. Wolf"), hospitals were pushed to their limits (Kaplan), and more than 26,000 residents lost their lives. Efficient and accurate models are necessary to inform residents and leaders, and develop mitigation strategies that can save lives. Accurate models have been developed; but they carry burdens of high computational costs and complex datasets. Simpler models may offer insight when computational power is constrained or when it is not feasible to collect or maintain large datasets.

In the literature we reviewed, we found that scientific study of modeling the spread and mortality of COVID-19 places a great emphasis on accuracy, while eschewing complexity concerns. Consider, for example, the SEIR (susceptible, exposed, infectious and recovered) model presented by the Institute for Health Metrics and Evaluation. This is a deterministic model that uses state-level epidemiological data to generate several scenarios, and describes a range of outputs (IHME). Researchers note that among other models released in June of 2020, their extremely thorough computations had the lowest median absolute percentage error at 20.2%. Around the same time, researchers from Portugal and Spain proposed a similar type of model that adds additional sub-groups, including hospitalized patients and so-called "super-spreaders." (Ndaïrou et al.) While their tool well models the initial outbreak in Wuhan, China, the authors suggest that further study is warranted into models that use more granular data, such as demographics like age and sex.

None of the work we surveyed addresses computational or dataset complexity. As COVID-19 is a global pandemic, we believe it is important that any nation make informed decisions, regardless of technological progress - something that may be possible with less expensive models. Furthermore, in the early stages of infection, policy makers must make decisions on very short notice. Reducing the complexity of the models used could offer useful insights during these critical moments. For that reason, we examine simplified models, taking the state of Pennsylvania as an example.

1.1 Contribution

This work is by Nate Athorn, Vishal Madhav, Matthew Sirkot and Roger Yu, students at the Pennsylvania State University. Athorn, Madhav and Sirkot wrote the code used in this project. The written report is authored by Sirkot and Yu.

2. Methods

Figure 1 depicts a model of our analysis. Using a dataset containing the daily number of tests given, new cases and deaths, we built simple Linear Regression and Neural Network models, attempting to predict deaths. We selected relatively simple datasets and models in order to investigate the feasibility of using low-cost computational models to predict COVID-19 deaths.

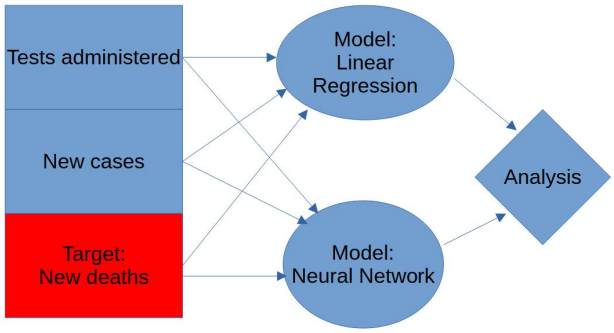

Figure 1:
diagram of analysis

2.1 Dataset description

The dataset we used is a simplified composite of datasets available from the Pennsylvania Department of Health (*Test Counts, Aggregate Cases, Aggregate Death Data)*. A portion of this raw data is presented in figures 2 through 4.

Date	New PCR Tests
03/01/2020	0
03/02/2020	0
03/03/2020	5
03/04/2020	2
03/05/2020	4
03/06/2020	16
03/07/2020	12
03/08/2020	17

Figure 2:
The number of COVID-19 tests administered in Pennsylvania

Jurisdiction	Date	New Cases	7-day Average New Cases	Cumulative cases	Population (2019)	New Case Rate	7-Day Average New Case Rate
Centre	03/05/2020	0		0	162385	0	
Pennsylvania	03/28/2021	2888	3861.4	1016221	12801989	22.6	30.2
Sullivan	10/30/2020	0	0.3	21	6066	0	4.7
Sullivan	10/02/2020	1	0.3	15	6066	16.5	4.7
Sullivan	11/03/2020	0	0.4	23	6066	0	7.1
Northumberland	08/17/2020	7	11.3	606	90843	7.7	12.4
Union	05/30/2020	0	0.7	58	44923	0	1.6
Union	07/08/2020	0	0.6	100	44923	0	1.3
Union	06/07/2020	1	2.3	74	44923	2.2	5.1
Union	06/14/2020	1	1.6	85	44923	2.2	3.5
Union	05/14/2020	1	0.6	36	44923	2.2	1.3
Union	05/01/2020	1	0.9	30	44923	2.2	1.9
Union	05/03/2020	0	0.9	30	44923	0	1.9
Union	07/16/2020	1	1.7	112	44923	2.2	3.8
Union	06/06/2020	1	2.1	73	44923	2.2	4.8
Cameron	03/21/2021	0	0.1	267	4447	0	3.2
Union	07/20/2020	1	1	115	44923	2.2	2.2

Figure 3:
New COVID-19 cases, organized by county

County Name	Date of Death	New Deaths	7-day Average New Deaths	Total Deaths	2019 Population	New Deaths Rate	7-day Average New Death Rate	Total Death Rate
Erie	03/10/2021	1	0.6	390	269728	0.37	0.21	144.59
Clarion	02/01/2021	0	0.4	78	38438	0	1.11	202.92
Northampton	03/18/2021	2	1.3	672	305285	0.66	0.42	220.12
Perry	01/11/2021	0	1.4	62	46272	0	3.09	133.99
Clarion	03/26/2021	0	0.1	88	38438	0	0.37	228.94
Fulton	03/22/2021	0	0	15	14530	0	0	103.23
Erie	02/23/2021	1	0.7	381	269728	0.37	0.26	141.25
Butler	11/21/2020	0	2.6	74	187853	0	1.37	39.39
Blair	03/08/2021	0	0.3	309	121829	0	0.23	253.63
Northampton	12/31/2020	6	4.9	465	305285	1.97	1.59	152.32
Centre	03/21/2021	0	0	213	162385	0	0	131.17
Philadelphia	05/20/2020	11	17.4	1393	1584064	0.69	1.1	87.94
Delaware	03/05/2021	1	0.9	1258	566747	0.18	0.15	221.97
Erie	01/07/2021	2	3.4	297	269728	0.74	1.27	110.11
Blair	03/14/2021	0	0.3	311	121829	0	0.23	255.28
Greene	02/13/2021	0	0	30	36233	0	0	82.8
Washington	01/26/2021	2	3.9	223	206865	0.97	1.86	107.8
Blair	02/10/2021	0	1.6	295	121829	0	1.29	242.14

Figure 4:
New COVID-19 deaths, organized by county

Figure 2 records the number of COVID-19 deaths administered per day in Pennsylvania. Figures 3 and 4 offer daily new cases and deaths, respectively. Figures 3 and 4 include additional analytical data that was excluded in order to simplify our model. These figures also organize cases and deaths by county; these were summarized in order to obtain figures for the state as a whole. Once those extraneous features were removed, a join of all 3 sets was computed to obtain only data that agrees on date. A sample of our composite dataset is presented in Figure 5.

Date	New PCR Tests	New Cases	New Deaths
2020-03-18	467	112	2
2020-03-19	1071	190	0
2020-03-20	1386	242	2
2020-03-21	1314	222	2
2020-03-22	1859	362	6
...	...	...	...
2021-04-18	25676	4286	64
2021-04-19	32365	10134	70
2021-04-20	51791	8888	30
2021-04-21	57373	8836	28
2021-04-22	61301	8056	6

Figure 5:
Composite dataset of COVID-19 tests, cases and deaths

2.2 Methodology

Two models are implemented: one using Linear Regression and another built using a Neural Network. The Linear Regression model is created using ordinary least squares regression, calculated with scaled data from the composite dataset. This model generates coefficients of -0.77 and 1.62.

Figure 6 comprises a diagram of the Neural Network we implemented. This model also takes scaled data from our composite dataset. As our intention is to explore simple, easy to implement models, our Neural Network comprises two input features; two hidden layers with 4 and 8 nodes, respectively; and one output. To further optimize performance, the hidden layers use a rectified linear unit activation function. The output node uses a sigmoidal function to ensure output values are between 0 and 1.

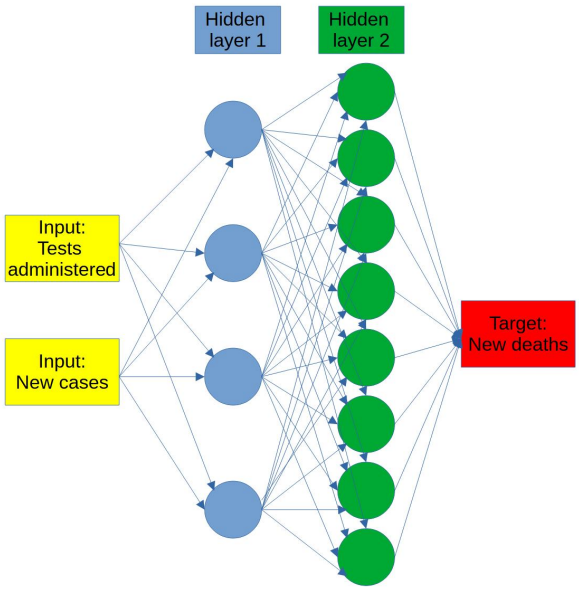

Figure 6: a diagram of the Neural Network used to predict deaths

2.3 Performance evaluation

For both models, values from our composite dataset were scaled from 0 to 1, relative to the minimum and maximum values for each feature. The data comprises 401 entries; 300 entries, or about 75%, are used as training data while the remaining 101 entries (25%) are used for validation. For reference, we calculate mean squared error, mean average error, and root-mean-square deviation in both models. We also calculate the coefficient of determination for the Linear Regression model.

3. Experimental Analysis

A few trends can be taken from the analysis. The first thing to note is that the mean squared error (MSE), mean average error (MAE) and root-mean-square deviation (RMSD) for the Linear Regression model are lower than the corresponding values from the Neural Network model, as shown in figure 7. However, the mean of the Neural Network is closer than the mean of the Linear Regression to the mean of the actual data, also shown in figure 7. An analysis of the graphs in figures 10, 11, and 12 indicates that the scatterplot for the Linear Regression more closely resembles the trends in the actual data, leading to a lower error, while the Neural Network more closely approximates the absolute values in terms of the number of cases, resulting in the mean values being closer. As the Linear Regression better approximates the actual data trends, it would be the better choice for approximating COVID-19 cases and deaths.

There are a few possible reasons why Linear Regression would approximate the actual cases better. As the model is being built from scratch, without any underlying assumptions, a Linear Regression could approximate the trends in the actual COVID-19 data more accurately than the Neural Network could, given the smoothness of the data. A dataset with more noise would be worse for a Linear Regression, as a trend would not be as easily derived from the data. The Neural Network model is also more complex and may not have gone through enough epochs to properly approximate the actual data, though the model loss plateauing (see figure 9) indicates that it would likely not get much more accurate with more epochs.

Major improvements can be made to our model. The MSE for our models is nearly half the mean, indicating that our models are not very accurate. The biggest way to improve our model is to ground it on a theoretical basis. More accurate models use an epidemiological model for the spread of disease to provide a basis for approximating COVID-19 spread. Additional assumptions can be made based on new knowledge about COVID-19 and reactions to it like lockdowns and testing. Machine learning can be used to refine the basic model with assumptions to more closely approximate the number of cases and deaths.

The code and datasets we utilized can be found here: https://github.com/matthewsirkot/Modeling-COVID-19

Model	MSE	MAE	RMSD	Mean	Mean (actual)
Neural Network	57.24	5465.11	73.93	122.82	117.7
Linear Regression	54.18	5175.6	71.94	135.29	117.7

Figure 7: mean, mean square error, mean average error, root-mean-square deviation of predictive models compared to mean of the actual values

Epoch	Training MSE	Training MAE	Validation MSE	Validation MAE
1	0.1268	0.3283	0.104	0.2873
2	0.1162	0.3115	0.1028	0.2858
3	0.1197	0.3179	0.1004	0.2843
...				
148	0.0304	0.1351	0.0332	0.1425
149	0.0293	0.1349	0.0332	0.1423
150	0.0276	0.1321	0.0333	0.142

Figure 8: selected training loss values by epoch for Neural Network

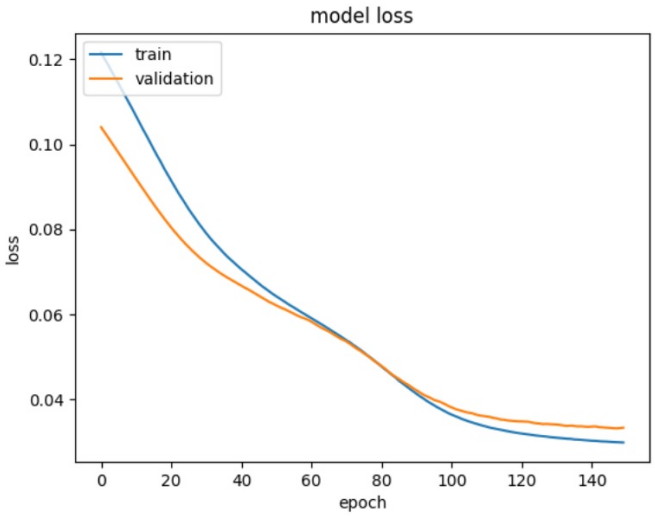

Figure 9: Neural Network model loss by epoch

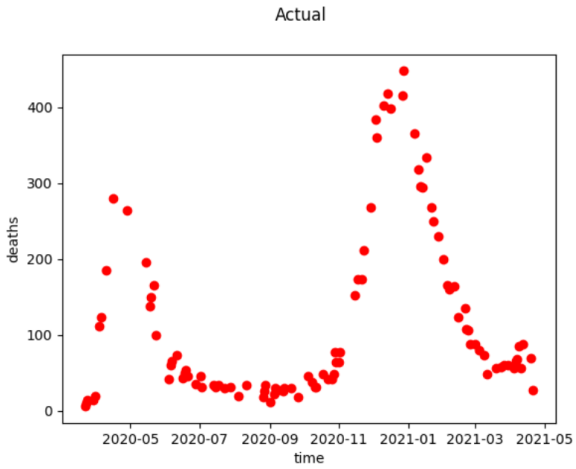

Figure 10: Validation data plotted

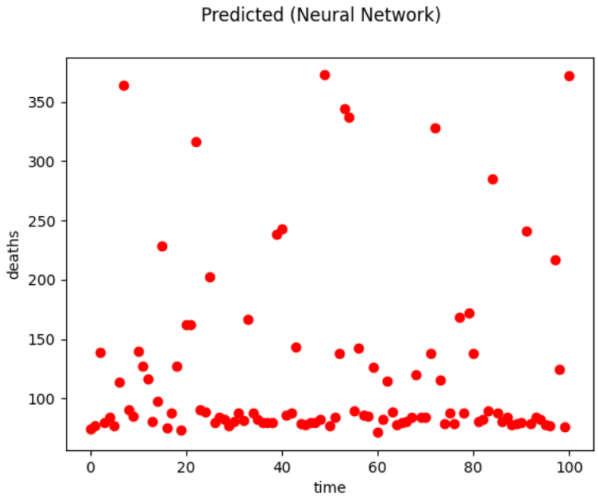

Figure 11: Neural Network predicted values

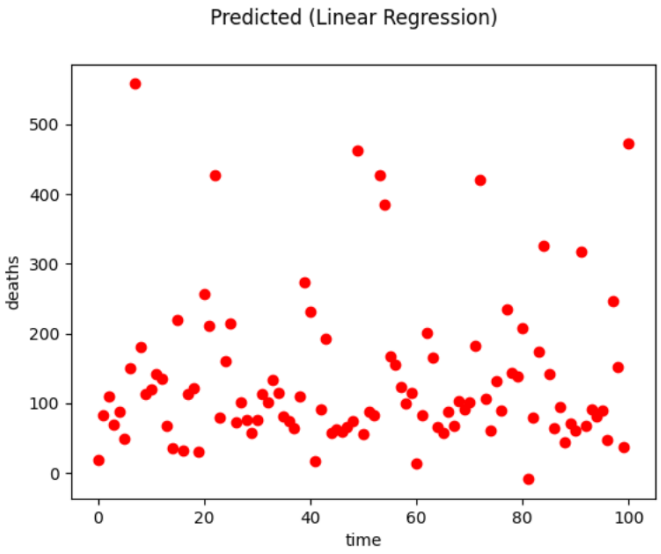

Figure 12: Linear Regression predicted data

4. Conclusion

We built predictive models for COVID-19 deaths in Pennsylvania using Linear Regression and Neural Network models. The linear regression was shown to be better for approximating COVID-19 cases and deaths, as indicated by its lower error values. This is likely due to the data being smooth, creating trends that a linear approximation is more capable of approximating. However, neither the neural network nor the linear regression are very accurate approximations of the data, as they both have high error values in proportion to the mean. Potential improvements to the model include basing it on an epidemiological model and taking assumptions based on various factors like lockdowns and mask-wearing into account.

Works cited

COVID-19 PCR Test Counts March 2020 - Current Statewide Health. Harrisburg, Pennsylvania: Pennsylvania Department of Health. https://data.pa.gov/Health/COVID-19-PCR-Test-Counts-March-2020-Current-Statew/r6ti-va88

COVID-19 Aggregate Cases Current Daily County Health. Harrisburg, Pennsylvania: Pennsylvania Department of Health. https://data.pa.gov/Health/COVID-19-Aggregate-Cases-Current-Daily-County-Heal/j72v-r42c

COVID-19 Aggregate Death Data Current Daily County Health. Harrisburg, Pennsylvania: Pennsylvania Department of Health. https://data.pa.gov/Health/COVID-19-Aggregate-Death-Data-Current-Daily-County/fbgu-sqgp

"Gov. Wolf officially orders the closure of 'non-life sustaining businesses' statewide." *abc27*, 19 March 2020. https://www.abc27.com/news/health/coronavirus/gov-wolf-officially-orders-the-closure-of-non-life-sustaining-businesses-statewide/

IHME COVID-19 Forecasting Team., Reiner, R.C., Barber, R.M. *et al.* Modeling Covid-19 scenarios for the United States. *Nat Med* 27, 94-105 (2021). https://doi.org/10.1038/s41591-020-1132-9

Kaplan, Seth. "Levine: Surging Hospitalizations 'of significant concern' for PA hospitals." *abc27,* 3 December 2020. https://www.abc27.com/news/health/coronavirus/coronavirus-pennsylvania/levine-surging-hospitalizations-of-significant-concern-for-pa-hospitals/

Ndaïrou, Faïçal *et al.* Mathematical modeling of COVID-19 transmission dynamics with a case study of Wuhan, *Chaos, Solitons & Fractals,* Volume 135, 2020. https://doi.org/10.1016/j.chaos.2020.109846.